A quantitative intracellular peptide binding assay reveals recognition determinants and context dependence of short linear motifs

Mythili S. Subbanna[1*], Matthew J. Winters[1*], Mihkel Örd [2,3], Norman E. Davey[3], and Peter M. Pryciak[1†]

[1] Department of Biochemistry and Molecular Biotechnology, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA
[2] University of Cambridge, Cancer Research UK Cambridge Institute, Robinson Way, Cambridge CB2 0RE, UK
[3] Division of Cancer Biology, The Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK

* Equal contribution

† Correspondance: peter.pryciak@umassmed.edu

Running title: SIMBA: systematic intracellular motif binding analysis

Keywords: peptides, intrinsically disordered protein, protein motif, ligand-binding protein, protein-protein interaction, high-throughput screening (HTS), cyclin-dependent kinase (CDK), phosphorylation, yeast, mutagenesis

ABSTRACT

Transient protein-protein interactions play key roles in controlling dynamic cellular responses. Many examples involve globular protein domains that bind to peptide sequences known as Short Linear Motifs (SLiMs), which are enriched in intrinsically disordered regions of proteins. Here we describe a novel functional assay for measuring SLiM binding, called Systematic Intracellular Motif Binding Analysis (SIMBA). In this method, binding of a foreign globular domain to its cognate SLiM peptide allows yeast cells to proliferate by blocking a growth arrest signal. A high-throughput application of the SIMBA method involving competitive growth and deep sequencing provides rapid quantification of the relative binding strength for thousands of SLiM sequence variants, and a comprehensive interrogation of SLiM sequence features that control their recognition and potency. We show that multiple distinct classes of SLiM-binding domains can be analyzed by this method, and that the relative binding strength of peptides in vivo correlates with their biochemical affinities measured in vitro. Deep mutational scanning provides high-resolution definitions of motif recognition determinants and reveals how sequence variations at non-core positions can modulate binding strength. Furthermore, mutational scanning of multiple parent peptides that bind human tankyrase ARC or YAP WW domains identifies distinct binding modes and uncovers context effects in which the preferred residues at one position depend on residues elsewhere. The findings establish SIMBA as a fast and incisive approach for interrogating SLiM recognition via massively parallel quantification of protein-peptide binding strength in vivo.

INTRODUCTION

The proper function of cells depends on an enormous number of interactions between different proteins [1]. Interactions that are weak and transient are particularly important in controlling molecular events that are rapid and dynamic. Many of these interactions are mediated by peptide sequences known as Short Linear Motifs (SLiMs), which by definition do not form stable tertiary structures and instead are enriched in intrinsically disordered regions of proteins [2]. SLiMs bind to globular folded domains in their partners [2-5] and they can be recognized by a wide variety of modular protein domain families, with well-known examples including SH3, WW, and PDZ domains. Over 200 distinct families of globular SLiM-binding domains are known [2], and their binding to cognate SLiM peptides can control subcellular targeting, assembly of multi-protein complexes, and recognition of substrate proteins by modifying enzymes (e.g., kinases, phosphatases, ubiquitin ligases, etc.). SLiMs play key roles in signal transduction pathways and the control of protein stability [1, 5], and their gain or loss can drive the evolution of regulatory networks [6, 7]. Furthermore, SLiM-mediated interactions can be mimicked by viruses and other pathogens to co-opt host cell functions [8-11], they can be targets for drugs [12], and they may contribute to human diseases when dysregulated [13].

Major questions about SLiM function remain unresolved because of limitations in understanding how SLiM peptide sequences dictate their recognition (Fig 1A). "Consensus" residues, meaning those that are shared across most identified binding peptides, often play a central role in controlling the affinity of SLiMs for their relevant binding partners and specificity that minimizes off-target binding [3]. However, more-variable residues in adjacent non-core positions can also contribute substantially [14]. In addition, sequence features that affect the structural conformation of the peptide can modulate binding energetics without contributing to the domain-peptide interface [15]. Efforts to understand these binding determinants would benefit from improved methods for SLiM discovery and characterization. Current estimates suggest that roughly one third of the human proteome is intrinsically disordered [16, 17] and contains over 100,000 SLiMs [4, 18], with the majority yet to be identified. These SLiMs are estimated to fall into roughly 350 distinct classes, with over 4000 experimentally validated examples [2]. While some SLiM classes have been studied extensively, most have only a few known examples, and hence they lack accurate definitions of the range of functionally permissible sequences [2, 19]. Established consensus motifs usually lack information about which deviations from the consensus are functionally tolerated and they overlook contributions of flanking positions, which limits their utility for discovering novel motifs and for predicting the effects of polymorphisms or disease mutations. Furthermore, it is becoming increasingly evident that variation in the binding strength of SLiMs can tune the magnitude or timing of regulatory events, such as during cell cycle transitions and in the control of protein phosphorylation or degradation [20-24]. These findings highlight a critical need for comprehensive and quantitative approaches to define how variations in SLiM sequences affect their binding strength, specificity, and functional potency.

A key attribute of SLiMs is that their short length (~ 3-11 residues) and limited contact interface (e.g., only 3-4 residues buried in the binding pocket of their partner domains) means that their binding affinities are often relatively weak (e.g., K_D ~ 1-500 μM) and rapidly dissociating [3, 25, 26]. Consequently, they mediate interactions that are inherently transient and dynamic. While this makes them well-suited for many important physiological functions (e.g., rapidly reversible interactions among signaling proteins), it can hinder their discovery by methods that rely on stronger binding, and it can complicate efforts to distinguish functional motifs from non-binding sequences. For example, tens of thousands of interactions in the human proteome have been identified via yeast two-hybrid screens, or by affinity purification followed by mass spectrometry [27-30], but in those screens SLiM-mediated interactions are statistically underrepresented, likely due to their weak affinities [5, 31]. Therefore, additional methods are needed to focus more specifically on interrogating SLiM-mediated interactions. Current approaches fall into two categories [32]: SLiM discovery (unbiased screens to identify new binding motifs), and SLiM characterization (analysis of sequence features that govern specific binding).

For the first category, SLiM discovery, high-throughput display-based technologies (phage display, bacterial display, and mRNA display) allow screening of peptide libraries that can exceed 10^8 distinct sequences [33-40]. Despite providing a wealth of information on sequences that can bind a given bait domain, these approaches do not always yield robust information about the relative binding strength of the captured sequences or about features that prohibit binding [33]. The SLiM sequence features required for binding can be inferred indirectly from the most common residues in these sequences; while useful, these inferences are not comprehensive, as they provide no information about whether any residues absent from the identified sequences are compatible with binding, which is critical for full understanding of the binding determinants and for the accuracy of predictive modeling [41]. Importantly, to obtain affinity information about candidate SLiMs identified via screening, it is typical to conduct follow-up tests using low-throughput biophysical assays [19, 34, 42], which can introduce a bottleneck in post-screening stages. Therefore, alternate high-throughput methods to validate captured sequences and quantify their binding strengths would be valuable.

The second category, SLiM characterization, requires the initial identification of one or more binding peptides, which then can be probed for the sequence features that control their binding. These binding determinants can be defined comprehensively by using saturation mutagenesis, in which each position in a peptide is systematically varied to all possible amino acids. One common procedure for interrogating SLiM sequences uses “SPOT” arrays, in which peptides are synthesized in situ on a solid support [43, 44]. This method has the advantage of being accessible to many experimental labs as well as the option to incorporate residue modifications (e.g., phosphorylation) at specific peptide positions. However, the binding measurements are often only semi-quantitative, and they involve washing steps that disrupt equilibrium and bias results toward the strongest binders [3, 45]. A recent method, “MRBLE-pep”, uses peptide variants attached to spectrally coded beads to measure binding to purified domains in vitro [45]. This method can accurately quantify binding affinities for hundreds of peptides in parallel, although it requires specialized equipment and reagents as well as purified protein domains; moreover, the costs and effort scale in direct proportion to the numbers of domains to be tested, which can become prohibitive for larger-scale efforts in which the number of domains and peptide contexts range from dozens to thousands. Several additional approaches using functional assays in living cells have provided detailed characterization of peptides that act as kinase docking motifs [20, 46], transactivation domains [47] and degrons [48], although none has yet been generalized to permit their application to other SLiM-binding domain families.

All currently available methods for SLiM analysis have their own advantages and disadvantages [32]. Ideal approaches would integrate several key features: (i) a quantitative readout that directly correlates with biophysical properties (e.g., binding affinity) or functional characteristics (e.g., stability); (ii) scalability in terms of the number of domains and peptides that can be tested; and (iii) generalizability to accommodate a diverse range of domain and peptide classes. In addition, assaying interactions within a cellular environment can be desirable to replicate the conditions and effects of crowding found in the cytoplasmic milieu [49]. Therefore, to provide a new approach that can complement existing methods, help circumvent bottlenecks, and fill key gaps in knowledge, we have devised a high-throughput method that allows for systematic and quantitative analyses of SLiM binding using an in vivo assay. The approach is called “SIMBA”, for Systematic Intracellular Motif-Binding Analysis. By combining deep mutational scanning

(DMS) with a competitive growth assay in which SLiM binding confers a growth advantage, we can quantify the relative binding strength of thousands of motif variants simultaneously and thereby define the rules of SLiM recognition with high accuracy. The work described here validates SIMBA methodology and initiates several downstream applications to demonstrate its feasibility and utility. We apply the approach to refine our understanding of three well-studied motif families by providing unique insights into the positive and negative contributions of non-core residues to binding strength, the impact of motif sequence context on residue preferences, and the interdependence of preferences at distinct motif positions. The findings highlight the potential for the SIMBA approach to substantially illuminate our current understanding of protein-protein interactions by providing a new analytical tool that can define recognition rules for large numbers of SLiMs and their binding domains.

RESULTS

Origin and Development of SIMBA Methodology

In general overview, the SIMBA method is an intracellular assay in which binding between a protein domain and a SLiM peptide regulates signaling through a yeast growth arrest pathway. In this system, cyclin dependent kinase (CDK) can block the growth arrest response by phosphorylating a protein in the MAPK-dependent signaling pathway (Fig 1B). To do so, the CDK requires one of its cyclins to recognize the target protein via a SLiM “docking site” (Fig 1B), which promotes dynamic, multi-site phosphorylation of the target [50, 51]. Because this phosphorylation inhibits the signaling protein [52], it allows cells to grow in the presence of the arrest signal (yeast mating pheromone). Previously, we exploited this antagonism of growth arrest to define the sequence requirements of docking peptides for binding to the native yeast cyclin, Cln2 [20]. Here, we have adapted this system to monitor binding between foreign globular domains and their SLiMs. To achieve this, we fuse the foreign SLiM-binding domain to yeast Cln2, and then use its cognate foreign SLiM peptide to replace the Cln2 docking motif in the CDK substrate (Fig 1C). As a result, binding of the foreign domain to its SLiM can drive substrate phosphorylation and block the growth arrest signal in a manner that reflects their interaction strength.

In the basic procedure, yeast cells contain two constructs. One encodes the foreign domain fused to Cln2, expressed from a galactose-inducible promoter. The other encodes the signaling protein that hosts the SLiM sequence to be tested; this protein is a chimeric signaling molecule, Ste20$^{\mathrm{Ste5PM}}$, that can be inhibited by CDK phosphorylation and can tolerate the insertion of peptide sequences [20, 50]. To monitor the effects of SLiM binding, expression of the domain-Cln2 fusion is induced, and then pheromone is added to activate the growth arrest pathway. There are two options for a quantitative readout. The first uses low-throughput assays of signaling, involving either transcriptional reporters or western blots that detect phosphorylation of a MAPK in the arrest signaling pathway, to quickly test small numbers of domain-peptide pairs. The second uses high-throughput assays, involving competitive growth of mixed cell populations, to screen libraries containing thousands of different SLiM sequences (Fig 1D). Here, stronger SLiM binding confers faster growth [20], and deep sequencing is used to analyze the rates of enrichment or depletion for all SLiMs in the population (Fig 1E), which are then converted to scores of relative binding strength (see Methods). It is worth emphasizing that, while our findings below and in subsequent sections indicate that in vivo strength correlates with binding affinity, the in vivo scores cannot be converted directly to a biophysical parameter (such as K_D).

Modular SIMBA components permit the study of diverse domains and peptides

A broadly applicable assay for peptide binding must have sufficient modularity to accommodate a diverse range of domain-peptide pairs. Initial tests using low-throughput assays showed that several distinct classes of SLiM-binding domain could be analyzed by the SIMBA method. We tested 6 SLiM-binding domains of 4 distinct structural types (Fig 2A, top), each of which was fused to the N-terminus of the yeast cyclin Cln2: an SH3 domain from the yeast protein Abp1 [53]; an Ankyrin Repeat Cluster (ARC) domain from human Tankyrase 2 (TNKS2^{ARC4}) [54]; two WW domains from human YAP1 and NEDD4 [55, 56]; and two SWIB domains from human MDM2 and MDM4 [57, 58]. These domains were chosen because each had multiple known binding peptides with a range of affinities. Binding of these domains to their cognate SLiMs was readily detected by virtue of enabling the hybrid cyclin-CDK complex to block yeast pheromone signaling, as measured by MAPK phosphorylation or induction of a transcriptional reporter. For example, tests using the TNKS2^{ARC4} domain illustrated several key features (Figs 2B, S1A-C). First, the TNKS2^{ARC4} domain did not perturb the ability of Cln2 to recognize its own docking site ("LP_Ste5"), which provides a useful control for functionality of the fusion protein as well as an internal standard for binding strength. Second, the TNKS2^{ARC4} domain imparted the ability to inhibit signaling proteins containing its cognate SLiMs, which were not inhibited by Cln2 alone. Third, the magnitude of inhibition conferred by each SLiM peptide in vivo reflected their binding affinities measured previously in vitro. Fourth, the foreign SLiMs did not affect the levels of the recipient substrate protein but they did alter its electrophoretic mobility in cells expressing the TNKS2^{ARC4}-Cln2 fusion (Fig 2B), consistent with inhibitory phosphorylation by the hybrid cyclin-CDK complex. Analogous results were obtained with the 5 other SLiM binding domains (Fig S1D-F), which are summarized in Fig 2A (bottom). Altogether, these results validate the utility of SIMBA as a modular and generalizable approach for detecting foreign domain-SLiM interactions and assessing their relative binding strengths.

The tests above showed that the SIMBA system can distinguish interactions with K_D's in the 5-100 μM range but starts to saturate for stronger affinities. To increase resolving power for these stronger interactions, we reduced the levels of the domain-cyclin fusion protein by using a weaker promoter to drive its expression. Specifically, we replaced the strong galactose-inducible promoter (P_{GAL1}) with weakened versions, P_{GALL} and P_{GALS} (Fig S1G). Indeed, interaction strengths of peptides that bind the MDM2^{SWIB} domain with affinities in the 0.05-5 μM range were resolved better when the domain-Cln2 fusion was expressed from the weaker P_{GALL} promoter rather than the full-strength P_{GAL1} promoter (Fig 2C). Thus, simple adjustments of the system's sensitivity can permit resolution of affinities spanning roughly three orders of magnitude (K_D = 0.05-100 μM). In the following sections, we will present the results from high-throughput experiments that validate the utility of the SIMBA approach for screening large numbers of domain-peptide pairs as well as for revealing unforeseen determinants of binding strength and specificity.

SIMBA allows comprehensive screening of binding strength and preferences in vivo

To confirm that the functional strength in vivo reflects biochemical binding affinity, we established a test case involving the TNKS2^{ARC4} domain. We chose this domain because its binding affinity had been measured in vitro for almost 200 SLiM peptides, including a set of all 152 single-site substitutions in an 8-residue peptide (RSPPDGQS) derived from human 3BP2 [54]. Thus, we used SIMBA to measure binding strengths for this same set of single-site variants (Fig 3A), and then compared the results from our in vivo assay to those from the prior in vitro measurements. This comparison revealed a good agreement for the full set of 153 variants (i.e., WT plus 152 mutants) (Figs 3B-C). Moreover, the SIMBA results recapitulated the distinct categories of selectivity at individual peptide positions that were previously seen in vitro (Figs 3C, S2A); examples include an exclusive requirement for Arg at p1, broad tolerance with a continuum of binding strengths at p2, and unique intolerance for a Pro residue at p7. We also mutagenized two extra residues on either side of the 8-residue motif (i.e., p-2, p-1, p9, and p10) and confirmed that these flanking regions have minimal influence on binding (Figs 3A-B).

Because all peptide-encoding plasmids are pooled together and assayed simultaneously during the competitive growth, it is minimal extra work to increase the number of peptides tested by several fold. Therefore, we performed DMS on five other parent peptides concurrently with the 3BP2 peptide (Fig S2B-C), each of which had been assayed individually in vitro [54]. These peptides included 3 native sequences from other proteins (NUMA1, MCL1, AXIN1), a previously defined variant with optimum residues at each position (RPGopt), and a weakened version of the 3BP2 peptide with a Glu substitution at p5 (3BP2-E5). In total, binding was measured in parallel for 1374 total peptides. The sequence preferences averaged across 5 motifs (excluding the 3BP2-E5 mutant) resembled that obtained from the 3BP2 motif alone (Fig 3B), although we observed context-specific features that will be described below. None of the peptides showed strong preferences at positions flanking the core 8-residue motif (Fig S2C). For the AXIN1 peptide, binding affinities were measured previously for Ala mutations at 9 positions (p-1 to p8) [54]; as we found with the 3BP2 variants, these AXIN1 mutations affected binding strength similarly in vivo and in vitro (Fig 3D). The score distributions from the 6 parent peptides reflected their expected binding strengths. Namely, those with stronger affinities (Fig S2B) gave broader variant score distributions and larger (in absolute value) negative scores (Fig S2D), indicating that the wild-type peptide is further from non-functional. Also, as described previously [20], stronger peptides more closely approach the maximum possible score achieved by mutants containing termination codons (Fig S2D, red), which have the greatest selective advantage because they eliminate the protein that mediates the growth arrest signal. When the parent peptide scores were normalized to these maxima, their in vivo strengths correlated with their previously measured affinities (r = 0.93) (Fig 3E). Thus, the good agreement between in vivo and in vitro binding strengths is observed when testing either mutant variants of individual peptides or groups of distinct parent peptides.

Finally, we asked if the SIMBA results might be influenced by the regional polypeptide context in which the SLiM peptide was embedded, by comparing results of inserting them at two distinct locations in the recipient protein – either on the N-terminal or C-terminal side of the phosphorylated region (Fig 3F). The results in the two location contexts were strongly correlated (Figs 3G, 3I, S2E). Therefore, the peptide motifs are functionally autonomous, with binding specificities that are independent of the surrounding polypeptide context. We also compared results when expressing the TNKS2^{ARC4} fusion protein from the full-strength (P_{GAL1}) versus a weakened (P_{GALS}) promoter (Fig S2A), and we found robust agreement (Figs 3H-I, S2F-G). Even with the full-strength promoter, strong peptides that were indistinguishable in low-throughput transcription or MAPK phosphorylation assays, such as 3BP2 (5 μM) and RPGopt (0.6 μM) (see Figs 2B, S1A), showed resolved binding strengths in the competitive growth assay (Figs 3E, S2D), and both reductions and increases in binding strength were detectable for mutant variants of each peptide (Fig S2C-D). Thus, sub-micromolar affinities of peptides do not preclude discovery of their binding determinants. We speculate that the resolution of strong binders is improved in the growth assay because the longer timespan of the experiment allows small differences to be reinforced and compounded. Collectively, our findings show that SIMBA can serve as an accurate gauge of relative biochemical affinity, that it reveals the same motif sequence preferences as would be observed in vitro with purified components, and that these preferences are independent of the host protein context of the SLiM or the expression level of its partner domain.

Local context effects revealed by comparing multiple parent peptides

As mentioned earlier, we performed DMS for multiple TNKS2^{ARC4}-binding peptides. When their residue preferences were grouped by position, it became evident that some parent peptides had strikingly different preferences at p4 (Fig 4A-B). Namely, for three peptides (3BP2, RPGopt, and NUMA1), Pro and Gly were equally the most favored residues at p4 (Fig 4C), in agreement with the previous in vitro results using the 3BP2 peptide [54]. In contrast, for the peptides from MCL1 and AXIN1, Gly was among the most disfavored residues at p4 (Fig 4C). These two peptides also showed a unique preference for Pro at p2 (Fig 4A), and both are inherently Pro-rich (Fig 4B), suggesting that they might be predisposed to form a type II polyproline helix (PPII helix). This left-handed structure possesses a restricted conformational flexibility that can reduce the entropic cost of binding [59, 60], and hence we hypothesize that this energetic benefit is disrupted when a Gly residue is introduced. In support of this interpretation, the main-chain trajectories for all peptides co-crystallized with TNKS2^{ARC4} show nearly identical left-handed topology from p1 through p5 [54], compatible with a PPII helix (Fig 4D-E).

Another notable context effect at p4 was the unusual preference for Val or Ile in the AXIN1 peptide (Fig 4A). Because Val at p4 had been observed to be disfavored in the 3BP2 context [54], it was assumed that the same would be true in the AXIN1 context; hence, to explain strong binding by the AXIN1 peptide, it was hypothesized that the negative impact of Val at p4 was offset by the presence of a favored Glu residue at p8 [54]. Surprisingly, however, our experiments revealed that Val at p4 is not suboptimal in the AXIN1 context and instead it is one of the two most favored residues (Fig 4A). Conceivably, because p4 in AXIN1 is flanked on both sides by Pro residues, Val or Ile might maintain the predisposition toward PPII structure while improving packing against the hydrophobic p4-binding pocket in TNKS2^{ARC4}. Similarly, the native residues at p5 for MCL1 (Ile) and AXIN1 (Pro) are not as disfavored in each of these parent peptides as they are in the 3BP2 peptide context, and instead they are favored as strongly as the more-common Asp residue (Fig 4A). In contrast to these context-dependent differences, it is noteworthy that the requirement for Gly at p6 and the severe intolerance for Pro at p7 were observed in all parent peptides (Fig S2C). Collectively, our findings illustrate the informative value of performing DMS analysis on multiple parent peptides, as it can reveal context-dependent preferences that would not be suspected from analysis of any single peptide motif. Such benefits emerge readily from the SIMBA methodology due to its ability to analyze thousands of peptide sequences simultaneously.

Defining SLiM recognition rules for the SWIB domain from human MDM2

As an additional test case, we used SIMBA to characterize the sequence preferences of the MDM2^{SWIB} domain, which drives ubiquitin-mediated degradation of the tumor suppressor p53 and is a target of anticancer compounds designed to block its peptide-binding pocket [61, 62]. We performed DMS on two human MDM2^{SWIB} binding peptides, one from p53 and another from NUMB, that have a shared core motif (FxxxWxxL) but different affinities (Fig S1E). The SIMBA results revealed similar preference patterns for both peptides (Fig 5A), including strong selectivity for hydrophobic residues at the three core positions p5, p9, and p12 (FxxxWxx[LIVMF]) that engage a deep hydrophobic cleft in the MDM2^{SWIB} domain [57, 63]. Additional preferences were evident at the non-core positions, most obviously for bulky aromatic or nonpolar residues at p8. There was a clear difference in the optimization of the two peptides, as binding to the NUMB peptide could be strengthened by numerous mutations (Fig 5B), particularly by replacing suboptimal residues with preferred residues at several non-core positions (p4, p8, p13, p15) (Fig 5A). In contrast, the p53 peptide contains preferred residues at p4 and p8, providing a potential explanation for why, despite identical core motifs, the binding affinity of the p53 peptide is roughly ten-fold stronger than that of the NUMB peptide (Fig S1E) [33]. The data also revealed strong negative preferences that likely relate to peptide conformation rather than binding contacts between the peptide and domain. Namely, in both peptides, Pro and Gly were largely disfavored from p7 to p13, consistent with their propensity to disrupt the α-helical conformation of the bound peptide [57, 63]. Interestingly, p6 tolerates Pro, despite being within the α-helical region. At p13, which immediately follows the core motif, Pro was the most disfavored residue in both peptides (Fig 5A). This is notable given that Pro is the wild-type residue at this position in the p53 peptide, and its replacement with Ala strengthens p53 binding affinity in vitro [15, 64, 65]. In the NUMB peptide, a Pro mutation at p13 results in a severe loss of MDM2^{SWIB} binding, emphasizing that the context of the starting peptide influences its robustness to mutation.

The effect on MDM2^{SWIB} binding affinity has been measured for alanine substitutions at 12 positions along the p53 peptide [65]. A comparison of these affinities with their corresponding SIMBA scores showed excellent agreement (r = 0.91) over a range of 3-4 orders of magnitude in K_D (Fig 5C), giving further confirmation that SIMBA provides an accurate measure of relative binding strength. Compared to alanine scanning, DMS data provide a more nuanced definition of how binding is dictated by sequence. In particular, we observed that non-core positions can substantially modulate binding strength (Fig 5D), but this effect can be overlooked by Ala-scanning for two reasons: (i) the effect of Ala is often more neutral compared to other residues; and (ii) the effect of Ala depends on the favorability of the wild-type residue being replaced (e.g., Ala at p8 is more disruptive in p53 than in NUMB because the wild-type residue is more favorable in p53 [Leu] than in NUMB [Gln]). The range of residue tolerance at each position, as revealed by these DMS measurements, can be summarized using the Gini coefficient as a measure of inequality (Fig 5E). It shows a pattern that is largely similar for both the p53 and NUMB peptides and is also consistent with a pattern of "not permitted" substitutions (defined as > 3-fold change in IC_{50}) observed in previous mutagenic scanning of an optimized MDM2-binding peptide [66]. Finally, a comparison to previous phage display data for the MDM2^{SWIB} domain [33] (Fig 5A, bottom) similarly illustrates how SIMBA results can provide more refined discrimination between the degrees of preference at distinct positions, as well as information about substitutions that result in the loss of binding (which is absent in phage display results). Altogether, these analyses of MDM2^{SWIB} binding preferences reinforce the utility of the SIMBA method for providing accurate, high-resolution definitions of motif binding determinants and for understanding the sequence basis for differences in peptide binding strength.

Specificity determinants and context effects in WW domain-binding peptides

To further explore how SIMBA methodology could be used to interrogate SLiM preferences and context dependence, we studied two different WW domains that bind peptides with a common core motif, [LP]PxY, or "PY peptides". We chose the first WW domain from human YAP (YAP^{WW1}) and the third WW domain from human NEDD4 (NEDD4^{WW3}), to allow comparison with prior information about their peptide ligands, binding affinities, and structural details of peptide-domain contacts [55, 56, 67-73]. Notably, statistical analyses of peptides captured by these two domains in phage-display experiments suggested there were correlations between residue preferences at different positions in the peptide [74]. That is, the most favored residue at a given peptide position might depend on the identity of residues at other positions. We sought to conduct direct empirical tests of such contingent preferences by performing DMS on multiple parental peptides that differ in their starting sequence context.

Initial tests using low-throughput SIMBA assays and a small number of peptides (Fig S1F) confirmed that WW domain binding was readily detected for higher-affinity peptides (K_D ~ 2-7 μM) and was weak or undetectable for lower-affinity peptides (K_D ~ 50-180 μM). For subsequent high-throughput assays and DMS, we chose 22 parental peptides from three categories (Fig S3A): (i) 15 natural peptides reported to bind one or both WW domains; (ii) 3 mutationally-optimized peptides (UGR2, UGR1, UGR1-LYG) that bind the NEDD4^{WW3} domain with strong, sub-micromolar affinity; and (iii) 4 peptides representing the consensus sequences of possible motif sub-classes (PYcon1-4) that were suggested by the statistical analyses noted above [74]. We used the competitive growth assay to test binding of these 22 parental peptides to each WW domain (Fig 6A-B). To compare their binding strengths against each other (rather than comparing mutant variants with a single wild-type sequence), we calculated z-scores that express the magnitude of enrichment of all peptides relative to a set of “nonbinder” control sequences (see Methods). The two domain fusions showed similar scores for the control Cln2-binding motifs (LP peptides) (Fig 6A), which provide internal standards for their binding capacities. Peptides derived from viral capsid proteins were weak binders (Fig 6B), consistent with prior in vitro data (Fig S3A). Many of the other peptides were recognized by both WW domains, but the rank order of binding strength was clearly distinct for the two domains, and some peptides showed striking specificity for one domain over the other (Figs 6A-B). Of note, those with the strongest specificity for YAP^{WW1} have a Pro residue immediately following the core Tyr residue (Fig 6B). From here forward, we will refer to this core Tyr position as p0 and denote other positions by their distance before (p-2, p-1, etc.) or after (p+1, p+2, etc.).

For DMS, we mutagenized twelve of these parental peptides at 12 positions each, including the core motif plus flanking residues on both sides (Fig S3B). Collectively, these experiments assayed binding of each WW domain to 2748 peptide sequences. The averaged sequence preferences from all parent peptides (Fig 6C) matched the expected pattern at the core positions (i.e., [LP]PxY), while also suggesting milder preferences at surrounding positions. However, these averages obscured distinct context-dependent patterns that became evident when the preferences were grouped by position (Fig 6D). Especially striking was a bifurcation of YAP^{WW1}-binding peptides into two classes, which we designated as types 1 and 2 (Fig 6E), that show distinct requirements at p-3: type 1 tolerates either Leu or Pro at p-3, whereas type 2 strongly favors Pro over Leu. Type 2 motifs also showed a unique preference for Pro at non-core positions p-4, p-1 and p+1 (Fig 6D-E). Because the parental type 2 peptides are especially Pro-rich (Fig 6F), their partiality toward Pro suggests a hypothesis similar to the one raised earlier for TNKS2^{ARC4} peptides. Namely, the type 2 sequences are likely predisposed to form a PPII conformation that is adopted by peptides in the bound state [68, 71-73, 75], which reduces the entropic cost of binding, and hence substitutions that disrupt this PPII propensity are disfavored.

On the C-terminal side of the core motif, type 1 peptides preferred acidic residues at p+2 as well as nonpolar residues at p+3 and p+4 (Fig 6D). These preferences fit with prior structural studies that identified contacts between C-terminal non-polar residues and a hydrophobic pocket on the NEDD4^{WW3} domain (near the Tyr-binding pocket) [68, 71, 72, 76]. In such cases, a right-handed helical turn immediately following the p0 Tyr allows a nonpolar side chain at p+3 or p+4 to occupy the hydrophobic pocket (Figs 6G, S3C), and acidic side chains at p+2 can stabilize the helix by making intrapeptide hydrogen bonds back toward p-1. Similar C-terminal preferences were strikingly absent in type 2 peptides (Fig 6D-E), perhaps because the Pro at p+1 might prevent the helical turn and/or occupy the hydrophobic pocket itself, as seen in the complex of YAP^{WW1} with the Lats_PY2 peptide [73] (Fig S3C). Indeed, for YAP^{WW1}, Pro was the most favored p+1 residue in type 2 peptides but the least favored in type 1 peptides (Fig 6D). For type 1 peptides, YAP^{WW1} and NEDD4^{WW3} showed subtle differences in their favored nonpolar residues at p+3 (Fig 6D). Specifically, YAP^{WW1} showed a notable tolerance for the aromatic residue Trp whereas NEDD4^{WW3} generally preferred aliphatic residues (I/L/V/A/C), perhaps reflecting differing composition of their hydrophobic pockets. The inclusion of Cys with these nonpolar residues is consistent with prior findings that it interacts favorably with hydrophobic membrane or protein environments [77-79], and that it ranks as nonpolar in several hydrophobicity scales [80].

Contingent preferences implicate contributions from intrapeptide interactions

We also noticed that preferences at p-1 were influenced by peptide context (Fig 6D). For example, the most favored p-1 residues in several peptides were Ser or Thr while in other peptides they were Asp or Ala. Furthermore, in the Smad7 context there was an unusual preference for aliphatic residues (I/L/V). We reasoned that these preferences might be influenced by the C-terminal flanking sequence (C-flank), because peptides that adopt a C-terminal helical turn can form intrapeptide interactions in which the p+2 side chain projects back toward p-1 (Figs 6G, S3C) [68, 70, 72], and the Smad7 peptide adopts an atypical hairpin-like conformation that places the Pro residue at p+4 near p-1 (Fig S3C) [67]. To address this possibility, we appended seven different C-flank sequences to the core motifs from three parental type 1 peptides, and then tested the p-1 preferences for all 21 combinations (Fig 7A). Although a few of these hybrid peptides bound YAP^{WW1} too weakly to be informative, the majority indicated that the C-flanks had clear effects on p-1 preferences of both WW domains. Namely, with some exceptions, the general trends were that two of the C-flanks (PYcon3, aENaC) imposed a p-1 preference for Asp, four others (Comm, AmotPY1, ARRDC3, UGR2) imposed a preference for Ser or Thr, and one (Smad7) imposed a distinct preference for Pro. Notably, the C-flanks that imposed the Asp preference at p-1 have a Thr residue at p+2, whereas those that imposed the Ser/Thr preference have a Glu residue at p+2 (Fig 7A). In light of the aforementioned intrapeptide interactions between p+2 and p-1 seen in some structures, we tested if changing the residue identity at either position altered preferences at the other (Fig 7B). Although the effects were not as strongly determinative as when entire C-flanks were swapped, we found that the identity of the p+2 residue did alter the rank of p-1 preferences in several motif contexts (Fig 7B, left). Namely, for three type 1 motifs (UGR2, aENaC, PYcon3), the presence of Thr rather than Glu at p+2 led to an increased preference at p-1 for Asp rather than Ser/Thr. In contrast, in the type 2 motif from Smad7, the p+2 residue had little effect on its unusual p-1 preferences. In reciprocal tests (Fig 7B, right), the identity of the p-1 residue influenced p+2 preferences modestly for two motifs (ARRDC3, Comm), in which the presence of Asp rather than Ser at p-1 led to an increased preference for Thr at p+2. Such influences were more variable for the aENaC motif and absent for the strong UGR2 motif. Collectively, our observations reveal pair-wise contingent preferences between two positions (p-1 and p+2) that are further influenced by the surrounding context. Because p-1 and p+2 residues remain solvent accessible in type 1 peptides (Fig S3D), rather than buried as part of the peptide-domain interface, their coordinated effects on binding likely signify a role in intrapeptide interactions that stabilize the bound conformation.

Curiously, type 1 peptides showed a substantial preference for Trp at p+1 (Fig 7D-E), a position that normally faces away from the WW domain in peptides with the C-terminal helical turn (Fig 6G). We considered the possibility that a Trp at p+1 provides an alternative way to favorably occupy the hydrophobic pocket, which could thereby preclude the role for nonpolar residues in the C-terminal helix. However, replacing p+1 with Trp did not alter the preferences for nonpolar residues at p+3 or p+4 displayed by NEDD4^{WW3}, or those at p+4 displayed by YAP^{WW1}, although it did noticeably shift the p+3 preferences shown by YAP^{WW1} (Fig 7C). These results suggest that the contributions to binding strength conferred by Trp at p+1 and nonpolar residues at p+3/p+4 are additive, not competitive. This finding favors an alternative explanation in which the p+1 Trp helps stabilize the type 1 C-terminal helix, perhaps via an aromatic-aromatic interaction [81] with the adjacent p0 Tyr. Indeed, structural predictions using AlphaFold [82] suggest that these Trp side chains lie next to the Tyr in a perpendicular geometry (Fig 7D) that is common for aromatic-aromatic interactions [81]. Altogether, our interrogations of context-dependent preferences reveal that different peptide conformations and domain pocket characteristics can contribute to a remarkable variety of subtly distinct binding modes that are not adequately described by average preferences or those of any singular motif.

Preferences at non-core positions influence predictions of binding strength

Consensus sequences describe the minimal required features of binding motifs, but due to their relatively low complexity there can be multitudes of matching sequences that are not easily distinguishable. For the consensus sequence [LP]PxY, there are 1730 matches in the intrinsically disordered regions of the human proteome. To rank their potential for binding the YAP^{WW1} or NEDD4^{WW3} domains, we used the comprehensive SIMBA data to derive position-specific scoring matrices (PSSMs) that quantitatively weight the preference for every possible residue at each position in the extended motif. Then, these PSSM values were summed over the length of the motif for each of the 1730 human sequences. This process was performed separately for each WW domain and using PSSMs derived from all motifs combined versus from only type 1 or only type 2 motifs. Each PSSM dispersed the 1730 matches into a broad distribution (Fig 8, "hits"), which is primarily due to variation at non-core positions, as the core consensus describes only 2 distinct sequences (i.e., LPxY and PPxY). For comparison, we also calculated PSSM sums for the parental motifs that had been tested in SIMBA experiments (Fig 8, "tested"). The strong binders among these tested motifs were generally in the top quartile of PSSM sum distributions. Interestingly, for YAP^{WW1} the type 1 PSSM gave very high scores for the strong type 1 motifs but low scores for the strong type 2 motifs, whereas the type 2 PSSM gave the reverse pattern. These type-specific results predict that only sequences yielding PSSM sums within the top 10% of the distribution would show binding strengths comparable to those tested in our experiments. For NEDD4^{WW3}, the different PSSMs all gave high scores to the strongest binders, yet we again observed that some of the type 1 motifs were underestimated by the type 2 PSSM and vice-versa (Fig 8, “tested”). Overall, these results illustrate that quantitative PSSMs derived from SIMBA data can help distinguish strong binders from bulk consensus matches, and they also emphasize the importance of using PSSMs specific for each distinct binding mode to achieve the best assessment of candidate sequences. By extension, without prior knowledge of distinct binding modes, evidence for their existence could emerge from disagreements between observed binding strengths and predictions that are based on a single PSSM.

DISCUSSION

In this study we have described and validated a new method, SIMBA, for performing systematic and quantitative analyses of SLiM-mediated binding to multiple distinct domains. This comprehensive and quantitative analysis provides a fast, easy, and low-cost alternative to traditional biochemical assays of protein-peptide binding. The power of the approach lies in its ability to rapidly quantify the binding strength of thousands of variant motifs in parallel and thereby reveal the specificity determinants for the queried domain. Further advantages include that it monitors binding in an in vivo setting, it can discern variations in binding strength over a broad range, and it is amenable to both low-throughput and high-throughput assays.

Previous studies on docking peptides for cyclins and MAPKs provide precedents that competitive growth assays in yeast can be used for fine-scale measurement of peptide binding strength [20, 46]. SIMBA now generalizes these strategies in a way that can be applied to a broad range of different types of domains and peptides, without depending on their normal biological functions. We envision that this method will be of primary utility in two categories of investigation. The first is to define the sequence determinants of SLiM binding strength and specificity for a given domain or group of related domains. The second is to compare the relative binding strengths of large sets of candidate SLiM peptides. Such candidates might emerge from other screens (e.g., phage display hits) or they might be potential binding peptides identified in the proteome as matches to an existing “consensus” sequence for a specific domain. In either case, SIMBA allows dozens to thousands of such candidates to be analyzed quickly and easily.

SIMBA provides a valuable complement, though not a substitute, to other methods for determining biochemical affinity. Its measurements of binding strengths are relative rather than absolute. For many purposes, such as for defining the sequence preferences of SLiM-binding domains, this relative binding information of mutant variants is sufficient, and it is not necessary to know the affinity of the wild-type motif. In cases where there are pre-existing measurements of K_D values for some peptides that bind a given domain, their inclusion in SIMBA experiments allows them to serve as benchmarks against which to compare the many other peptides being tested.

The findings in test cases described here clearly demonstrate how systematic analyses of SLiM sequence requirements can illuminate unanticipated complexities in binding determinants even for well-studied examples such as TNKS2^{ARC4}, YAP^{WW1}, and NEDD4^{WW3}. Although we had primarily expected to confirm previously defined binding rules for these domains, in each case we also obtained evidence for different binding modes and contingent relationships between different motif positions that could not have been predicted based on prior knowledge. These observations illustrate the intrinsic discovery potential of the SIMBA approach that emerges readily from its ease of performing massively parallel analyses. Namely, gathering empirical binding results for large numbers of domain-peptide pairs can provide unexpected insights that would be missed in more limited analyses (e.g., those focusing on singular peptide contexts).

The SIMBA method is fundamentally a variation of the classic yeast two-hybrid system [83, 84], except the interaction being tested controls a phosphorylation reaction rather than transcription. In principle, either method could be used to investigate SLiM binding determinants, and the most crucial advances for large-scale interrogation arise from the combination of competitive growth and deep sequencing to allow massively parallel analyses. Nevertheless, SIMBA offers several additional benefits that are notable. (i) It can detect weak binding in the 1-100 μM range that includes many physiologically relevant SLiM interactions [3]. (ii) The interaction controls multisite substrate phosphorylation that yields a continuum of regulatory strength [52], which makes it sensitive to binding strength over a broad range. (iii) The phosphorylation reaction in vivo is dynamic and rapidly reversible [85], which can help ensure that the measured output reflects binding at equilibrium. (iv) Because the fused cyclin (Cln2) can still recognize its own docking site, it provides an internal standard against which to compare the strength of binding mediated by the fused domain. (v) The effects of the SLiM interaction are under acute control, as they require induced expression of the cyclin fusion and an external stimulus, and hence there is no chronic selection pressure to skew population distributions before the experiment begins. (vi) It allows adjustments in the required binding strength (by using different strength promoters) so that multiple affinity ranges can be monitored.

Because SIMBA monitors enrichment or depletion of all sequences in the tested population, the quality of data is similar for both functional and nonfunctional motifs. This aspect is a notable contrast to capture-based approaches, such as phage display, that provide only indirect inferences about any sequences that were not captured. Dependable quantification of a broad range of binding strengths, from strong binders to non-binding sequences, as well as identification of prohibited residue substitutions, will be of great value in developing accurate ranking and filtering criteria for future efforts at in silico prediction of binding motifs. High quality, comprehensive binding data should also be useful for structural modeling and for predicting pathogenic impacts of sequence variants. For example, quantitative data from large numbers of peptides, including abundant examples of non-binders, can provide training information to improve computational predictions of protein complexes and their binding affinities [86-88]. Separately, they can provide empirical tests of binding effects of cancer-associated mutations in the COSMIC database [89] and improve the confidence in pathogenicity predictions generated by AlphaMissense [90], especially for variants in disordered regions of proteins that lack the structural features used as a main input of that algorithm.

Finally, in addition to providing mechanistic insights into binding specificity, SIMBA could also accelerate development of useful tools for basic biochemical research or synthetic biology. For example, the ability to rapidly identify variant peptides with distinct binding strengths could allow the design of degron tags, enzyme docking sites, or localization anchors whose effects on recipient target proteins can be finely tuned over a broad range. Separately, SIMBA approaches could also assist with identifying competitive peptide inhibitors of SLiM-binding domains. Such peptide inhibitors could serve as drug proxies or guides for peptidomimetic compounds [91] and could also be used to pre-screen the effects of target inhibition in human cell lines [92]. In applications currently underway, we have already applied the SIMBA method to over thirty other SLiM-binding domains, and we have extended the approach to include unbiased screens for discovery of new binding motifs. Thus, we expect that SIMBA, due to its ability to systematically define SLiM recognition rules and functional potency en masse in vivo, will be a versatile tool for numerous areas of future investigation.

EXPERIMENTAL PROCEDURES

Yeast Strains and Growth Conditions

Standard procedures were used for growth and genetic manipulations of yeast [93, 94]. Unless indicated otherwise, cells were grown at 30˚C in yeast extract/peptone medium with 2% glucose (YPD), or in synthetic complete medium (SC) lacking histidine and/or uracil with 2% glucose or raffinose. Yeast strains and plasmids are listed in Tables S1 and S2, respectively. To construct plasmids encoding fusions of Cln2 to foreign SLiM-binding domains, the domain sequences were amplified by PCR or obtained as synthesized gene fragments, and then they were inserted as BamHI-XmaI fragments between the GST and CLN2 sequences of a P_{GAL1}-GST-CLN2 construct [50]. All such constructs used for competitive growth assays harbored a full-length CLN2 fragment (residues 1-545), whereas many initial constructs used for low-throughput signaling assays contained a truncated fragment (residues 1-372). Related constructs harbored weakened promoters [95] called P_{GALS} (-415 to -5) and P_{GALL} (-431 to -5), which were amplified by PCR and inserted in place of the full-strength promoter P_{GAL1} (-675 to -5). We note that the expression levels were reproducibly greater with P_{GALS} than with P_{GALL}, which is the reverse of the rank strength reported previously [96]. Strains used for competitive growth assays were constructed by integrating P_{GAL}-GST-domain-CLN2 constructs into the genome at the *HIS3* locus of strain PPY2617; PCR was used to confirm single copy integration. These strains were transformed with plasmids that encode derivatives of a chimeric signaling protein, Ste20^{Ste5PM}, that harbor different SLiMs [20].

Signaling Assays

For low-throughput assays of SLiM binding, we used methods similar to those described previously [20, 50] to monitor the ability of domain-Cln2 fusion constructs to inhibit pheromone signaling, either by using a transcriptional reporter (*FUS1-lacZ*) or by immunoblotting for phosphorylated MAPK (Fus3). Plasmids encoding Ste20^{Ste5PM} chimeras with various SLiMs were either (i) co-transformed with a domain-Cln2 fusion plasmid into a *STE5-8A ste20Δ* strain (PPY2617); or (ii) transformed into a similar strain containing an integrated copy of a relevant domain-Cln2 fusion plasmid. Measurement of *FUS1-lacZ* transcriptional induction and MAPK activation followed methods described previously [20, 97]. Cell cultures were grown to exponential phase in SC liquid medium containing 2% raffinose and lacking histidine and/or uracil. Cyclin expression was induced by adding 2% galactose for 90 minutes. For *FUS1-lacZ* assays, cells were treated with pheromone (50 nM) for 45 min, and β-galactosidase activity was measured as described previously [97]. To assay MAPK activation, cells were treated with pheromone (50 nM) for 15 min, whole cell lysates were prepared as described previously [98], and then proteins were resolved by SDS–PAGE and transferred to PVDF membranes (Immobilon-P; Millipore). Primary antibodies used include mouse anti-phospho-p44/42 (1:1000, Cell Signaling Technology #9101), rabbit anti-phospho-p44/42 (1:5000, Cell Signaling Technology #4370), rabbit anti-myc (1:200, Santa Cruz Biotechnologies #sc-789), and rabbit anti-G6PDH (1:100000, Sigma #A9521). Horseradish peroxidase-conjugated secondary antibody was goat anti-rabbit (1:3000, Jackson ImmunoResearch #111-035-144). Enhanced chemilluminescent detection used a BioRad Clarity substrate (#170-5060). Densitometry was performed using ImageJ software. Quantified measurements of signaling outputs using either assay are plotted relative to samples in which a control peptide (SNGNGSGSNGN) [20] is present in the Ste20^{Ste5PM} chimera. Fold inhibition is the difference in signal between the test peptide and the control peptide, as a fraction of the control signal.

Library Construction and Competitive Growth Assays

The construction of mutant libraries and the performance of bulk growth competition assays were based on procedures described previously [20]. SLiM variants were tested in the context of a chimeric Ste20$^{\text{Ste5PM}}$ signaling protein, in which the membrane-binding domain from Ste5 and its flanking CDK phosphorylation sites replace the native membrane-binding domain in Ste20, and SLiMs are placed at either the N or C terminal side of the Ste5 fragment [20, 50].

To construct libraries of SLiM variants, SLiM-encoding sequences were amplified by PCR using oligonucleotides as templates and primers that anneal to flanking linker sequences (Gly Gly Ser Gly). The template oligonucleotides were designed such that single codons were randomized (NNN) to generate all 64 nucleotide variants. All templates for a given motif were pooled in equal amounts and then amplified by 10 or 12 PCR cycles (98˚ for 10 sec; 56˚ for 20 sec; 72˚ for 6 sec) using different primer sets for insertion at the N-terminus (ggagtgacgtcGGAGGTAGTGGT and ctcacgctagcTCCAGATCCACC) or the C-terminus (actatacgcgtGGAGGTAGTGGT and tctttgcatgcTCCAGATCCACC). Column-purified PCR products were digested with restriction enzymes (AatII and NheI for insertion at N terminus; MluI and SphI for insertion at C terminus), treated with calf intestinal phosphatase, and then ligated (16 hrs at 18˚, then 10 min at 65˚) into the appropriately digested vector (pPP4375 or pPP4745). The ligation products were transformed into E. coli (XL-10 Gold Ultracompetent Cells; Agilent/Agilent Technologies) and plated on LB+Amp plates. Colonies (exceeding the number of variant sequences per library by > 10-fold) were harvested by adding LB+Amp liquid and gently agitating with a glass rod spreader, and then plasmid DNA was prepared from the suspension of pooled colonies. The isolated DNAs were checked by Sanger sequencing to verify that all four nucleotides were comparably represented at each position in the randomized codon. To construct plasmids with individual peptide motifs (e.g. wild type), the oligonucleotides used as PCR templates were unique sequences without codon randomization, or for some early constructions the insert fragments were obtained by annealing complementary oligonucleotides as described previously [20].

For competitive growth assays, a solution with equimolar amounts of each individual codon library was prepared. A mixture of control plasmids containing wild-type SLiMs plus unrelated and random sequences was spiked into the solution (making up 1-2.5% of the total) to create the final pool, which was transformed into yeast strains containing the relevant P_{GAL}-domain-CLN2 fusions. After the transformation procedure, the bulk of the transformation mixture was split and plated onto two -URA plates, plus a diluted aliquot (1%) was plated onto a third -URA plate. Colonies on the diluted plate were counted to ensure that the number of transformants exceeded the number of mutant variants in the library by at least 10-fold. Colonies on the concentrated plates were suspended in 10 mL of –URA/raffinose liquid medium, washed twice with 20 mL -URA/raffinose, then diluted into 50 mL -URA/raffinose to yield a density of ~ 4×10^6 cells/ml (OD_{660} ~ 0.3). These cultures were incubated in a shaking water bath for 4 hrs, then diluted back to OD_{660} ~ 0.01 and incubated overnight for ~16 hrs. The cultures were diluted back again (to OD_{660} ~ 0.6 in 50 mL), incubated for 1.5-2.5 hrs, and treated with 2% galactose for 75 min (in a volume of 70 mL) to induce cyclin expression. At this time (t = 0), an aliquot (~ 38 mL, 3×10^8 cells) was harvested, and the remaining culture was treated with pheromone (500 nM) and returned to incubation. Cells were diluted with fresh medium (including galactose and pheromone) after the first 8 hr and subsequently after every 12 hr to maintain an OD_{660} below 1 (in a volume of 50 mL). Additional aliquots (~ 20 mL, 3×10^8 cells) were harvested at 8, 20, 32 and 44 hr. Harvested cells were collected by centrifugation (5 min., 3200 rpm, room temp), washed with 10 mL sterile water, resuspended in 1 mL sterile water, and transferred to 1.5 mL microcentrifuge tubes. These suspensions were centrifuged, the supernatants were aspirated, and the pellets were frozen using liquid nitrogen and stored at -80˚C.

DNA Preparation and Deep Sequencing

Plasmids were isolated from yeast cell pellets following previously described methods [20, 99]. DNA was purified using the Zymo Research ZR Plasmid Miniprep Kit (#D4015). Frozen cell pellets were thawed and suspended in 200 µl of solution P1 and treated with 7 µL Zymolyase (0.2 units/µL; Zymo Research #E1005) at 37˚C for 1.5 hr, with vortexing approximately every 20 mins to disperse clumps, before proceeding with the remaining steps. Samples of plasmid DNA (4 µL) were subjected to PCR (17 cycles, 50 µL total volume) with primers that include standard P5 and P7 sequences for binding to Illumina flow cells during next generation sequencing (NGS). The forward primer included the P5 sequence, followed by an Illumina sequencing primer binding site, a 6-nucleotide bar code, and an upstream plasmid-annealing sequence; the reverse primer included a P7 sequence, followed by a 6-nucleotide i7-index sequence, an i7 sequencing primer binding site, and a downstream plasmid-annealing sequence. Aliquots (5 µL) of the PCR products were run in 1 % agarose gels to confirm the presence of the desired product. For any timepoint samples with low product yield, PCRs were repeated using NotI-digested plasmid and/or increased PCR cycles (19-25). Gel band intensities were quantified by densitometry (ImageJ), and then equal amounts of products from all timepoints for a given strain were pooled, purified using Zymo Spin I columns (Zymo Research #C1003-250), and eluted in 30 µL of 10 mM TRIS-HCl, pH 8. The concentration of the eluted products was measured using NanoDrop spectrophotometry, and then equal amounts of samples from all strains were pooled, gel-purified in triplicate from a 1% gel (NEB Monarch DNA Gel Extraction Kit #T1020L), eluted in 20 µL, and the triplicate eluates were combined. The final concentration was measured by Qubit fluorometry and the product size distribution was verified by Fragment Analyzer (Agilent) before being sent for Illumina-based NGS (paired-end sequencing, 150 bp) by a commercial vendor (Novogene).

Sequencing Data Analysis

NGS data were de-multiplexed by strain and timepoint using bar code and index identifiers. To compare mutant variants versus wild-type SLiM sequences, we used Enrich2 software [100] to obtain read counts for all sequence variants. Then, we added 0.5 counts to any variants depleted to 0 counts by the 32-hr timepoint (to prevent Enrich2 from ignoring fully-depleted variants), and used Enrich2 to calculate fitness scores that describe the rate of change of mutant variants compared to the wild-type sequence [20, 100]. All scores reported here were calculated from 32-hr time courses.

To compare groups of sequences where there was no single wild-type reference standard, such as when comparing multiple WW domain binding peptides to each other, we calculated their enrichment relative to a set of non-binder control sequences. For this, we used 2FAST2Q [101], a program in Python, to count the occurrence of sequences from all strains and timepoints, converted each value to a frequency by dividing by the total counts, and normalized each frequency relative to the starting (t=0) frequency. We calculated a raw score for each timepoint as the log2 of the relative frequency divided by the time in hours (raw score = log2[relative frequency]/time). For subsequent calculations we used the median raw score from all timepoints for a given sequence. We defined a set of non-binder controls as all 236 missense variants of the core Tyr codon in four [LP]PxY motifs, plus 9 other non-[LP]PxY sequences. Then, we calculated z-scores as (X- µ)/σ, where X is the raw score of a test sequence, µ is the mean raw score of the non-binder set, and σ is the standard deviation of the non-binder raw scores. Hence, this z-score represents the number of standard deviations that a test sequence differs from the mean of non-binders.

To create logos of sequence preferences, we transformed Enrich2 scores or z-scores into a preference metric that expresses the bias for each residue at a given motif position relative to the set of all possible residues, following procedures described previously [20]. For DMS analyses of wild-type motifs, the raw Enrich2 score for each amino acid variant was normalized to the lowest (= 0) and highest (= 1) scores in a given motif array: (normalized score) = (raw score – minimum) ÷ (mean terminator score – minimum). Then, the normalized scores were converted to a frequency metric by dividing each by the sum of all scores at the same position: (frequency) = (normalized score) ÷ (column sum). Finally, the frequency metric was converted to a preference metric by subtracting 0.05, so that a neutral preference is represented by zero, favored residues are positive, and disfavored residues are negative: (preference score) = (frequency – 0.05). These preference scores were used to generate sequence logos via a web-based tool (http://slim.icr.ac.uk/visualisation/index.html). For experiments that probe how positional preferences in WW domain binding peptides are affected by altered peptide context (such as swapping C-terminal flanking sequences), the calculated z-scores for all variants in the control strain (unfused Cln2) were subtracted from the corresponding scores in the test strains (WW-Cln2 fusion). These Cln2-subtracted z-scores for all non-terminator variants were normalized separately for each position that was individually interrogated. Then, these normalized scores were converted to frequency metrics, preference scores, and sequence logos as described above. The plotted results exclude peptides for which the Cln2-subtracted z-score was less than 1. To generate a preference logo from in vitro affinities of 3BP2 variants, we first calculated the inverse of published K_D values [54], assigned an inverse value of zero to variants with unquantifiable K_D's, and then used these inverse values to calculate frequency and preference metrics as described above.

PSSMs for scoring matches to the [LP]PxY consensus sequence were derived as follows. First, as described above, normalized scores were calculated for each WW-binding motif. These values were then transformed to a difference PSSM [20] by subtracting the average of all residues at a given position from the value of each residue at that position: (difference score) = (residue score) – (position average). Then, the difference scores were averaged for subsets of motifs, as listed in Figure 6D: all motifs (sequences 0-9 for NEDD4^{WW3} or 1-8 for YAP^{WW1}), type 1 motifs (sequences 1-8 for NEDD4^{WW3} or 1-4 for YAP^{WW1}), and type 2 motifs (sequences 9 and 0 for NEDD4^{WW3} or 5-8 for YAP^{WW1}). Finally, to give identical boundaries for all PSSMs, the averaged difference scores were further transformed such that positive values were normalized to the maximum (= 1) and negative values were normalized to the minimum (= -1). Human proteome sequences matching the [LP]PxY consensus were obtained using the SLiMSearch algorithm [102]. For each sequence, the corresponding boundary-normalized PSSM values for each residue at each position were summed across 10 motif positions (xx[LP]PxYxxxx) to obtain a predicted score (PSSM sum). PSSM sums were calculated for reference sequences in the same way.

For competitive growth assays of SLiM variants, fitness scores and standard errors were calculated by Enrich2 software [100]. Other statistical analyses, including calculation of means, medians, SD, SEM, Pearson's correlation (r), and Gini index, were performed using Microsoft Excel. The numbers of biological replicates are described in the Figure Legends.

Structure representations, analysis, and predictions

Illustrations based on prior crystallography or NMR structural data were generated using PyMOL software and original Protein Data Bank (PDB) coordinates. Calculations of buried surface area of peptide residues bound to WW domains were performed using PISA, an online protein interface analysis tool (https://www.ebi.ac.uk/msd-srv/prot_int/pistart.html). Structural predictions of peptide-domain complexes were generated by AlphaFold 2 [82], implemented using ColabFold [103] and UCSF Chimera X software [104]; illustrations were created in PyMOL.

Data availability

All necessary data are available in the submitted manuscript or supporting materials. Large datasets derived from competitive growth assays and next generation sequencing, including variant counts and calculated enrichment scores, have been deposited at Mendeley Data: (https://doi.org/10.17632/nghf59hf4s.1).

Supporting information

This article contains supporting information.

Funding and additional information

This work was funded by grants from the NIH to P.M.P. (R01GM057769 and R01GM145795). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. N.E.D. and M.O are funded by a Cancer Research UK Senior Cancer Research Fellowship (C68484/A28159) and by a UKRI grant EP/X042065/1 and Boehringer Ingelheim Fonds travel grant to M.O.

Conflict of interest

The authors declare that they have no conflicts of interest with the contents of this article.

REFERENCES

1. Pawson, T., and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. Science *300*, 445-452.
2. Kumar, M., Michael, S., Alvarado-Valverde, J., Zeke, A., Lazar, T., Glavina, J., Nagy-Kanta, E., Donagh, J.M., Kalman, Z.E., Pascarelli, S., et al. (2024). ELM-the Eukaryotic Linear Motif resource-2024 update. Nucleic Acids Res *52*, D442-D455.
3. Ivarsson, Y., and Jemth, P. (2019). Affinity and specificity of motif-based protein-protein interactions. Curr Opin Struct Biol *54*, 26-33.
4. Tompa, P., Davey, N.E., Gibson, T.J., and Babu, M.M. (2014). A million peptide motifs for the molecular biologist. Mol Cell *55*, 161-169.
5. Van Roey, K., Uyar, B., Weatheritt, R.J., Dinkel, H., Seiler, M., Budd, A., Gibson, T.J., and Davey, N.E. (2014). Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem Rev *114*, 6733-6778.
6. Davey, N.E., Cyert, M.S., and Moses, A.M. (2015). Short linear motifs - ex nihilo evolution of protein regulation. Cell Commun Signal *13*, 43.
7. Goldman, A., Roy, J., Bodenmiller, B., Wanka, S., Landry, C.R., Aebersold, R., and Cyert, M.S. (2014). The calcineurin signaling network evolves via conserved kinase-phosphatase modules that transcend substrate identity. Mol Cell *55*, 422-435.
8. Chemes, L.B., de Prat-Gay, G., and Sanchez, I.E. (2015). Convergent evolution and mimicry of protein linear motifs in host-pathogen interactions. Curr Opin Struct Biol *32*, 91-101.
9. Davey, N.E., Trave, G., and Gibson, T.J. (2011). How viruses hijack cell regulation. Trends Biochem Sci *36*, 159-169.
10. Meszaros, B., Samano-Sanchez, H., Alvarado-Valverde, J., Calyseva, J., Martinez-Perez, E., Alves, R., Shields, D.C., Kumar, M., Rippmann, F., Chemes, L.B., et al. (2021). Short linear motif candidates in the cell entry system used by SARS-CoV-2 and their potential therapeutic implications. Sci Signal *14*.
11. Via, A., Uyar, B., Brun, C., and Zanzoni, A. (2015). How pathogens use linear motifs to perturb host cell networks. Trends Biochem Sci *40*, 36-48.
12. Corbi-Verge, C., and Kim, P.M. (2016). Motif mediated protein-protein interactions as drug targets. Cell Commun Signal *14*, 8.
13. Uyar, B., Weatheritt, R.J., Dinkel, H., Davey, N.E., and Gibson, T.J. (2014). Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? Mol Biosyst *10*, 2626-2642.
14. Palopoli, N., Gonzalez Foutel, N.S., Gibson, T.J., and Chemes, L.B. (2018). Short linear motif core and flanking regions modulate retinoblastoma protein binding affinity and specificity. Protein Eng Des Sel *31*, 69-77.
15. Borcherds, W., Theillet, F.X., Katzer, A., Finzel, A., Mishall, K.M., Powell, A.T., Wu, H., Manieri, W., Dieterich, C., Selenko, P., et al. (2014). Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells. Nat Chem Biol *10*, 1000-1002.
16. Dunker, A.K., Silman, I., Uversky, V.N., and Sussman, J.L. (2008). Function and structure of inherently disordered proteins. Curr Opin Struct Biol *18*, 756-764.
17. Fukuchi, S., Hosoda, K., Homma, K., Gojobori, T., and Nishikawa, K. (2011). Binary classification of protein molecules into intrinsically disordered and ordered segments. BMC Struct Biol *11*, 29.
18. Davey, N.E., Van Roey, K., Weatheritt, R.J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T.J. (2012). Attributes of short linear motifs. Mol Biosyst *8*, 268-281.

19. Gibson, T.J., Dinkel, H., Van Roey, K., and Diella, F. (2015). Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. Cell Commun Signal *13*, 42.
20. Bandyopadhyay, S., Bhaduri, S., Ord, M., Davey, N.E., Loog, M., and Pryciak, P.M. (2020). Comprehensive Analysis of G1 Cyclin Docking Motif Sequences that Control CDK Regulatory Potency In Vivo. Curr Biol *30*, 4454-4466 e4455.
21. Davey, N.E., and Morgan, D.O. (2016). Building a Regulatory Network with Short Linear Sequence Motifs: Lessons from the Degrons of the Anaphase-Promoting Complex. Mol Cell *64*, 12-23.
22. Hertz, E.P.T., Kruse, T., Davey, N.E., Lopez-Mendez, B., Sigurethsson, J.O., Montoya, G., Olsen, J.V., and Nilsson, J. (2016). A Conserved Motif Provides Binding Specificity to the PP2A-B56 Phosphatase. Mol Cell *63*, 686-695.
23. Lu, D., Girard, J.R., Li, W., Mizrak, A., and Morgan, D.O. (2015). Quantitative framework for ordered degradation of APC/C substrates. BMC Biol *13*, 96.
24. Roy, J., Li, H., Hogan, P.G., and Cyert, M.S. (2007). A conserved docking site modulates substrate affinity for calcineurin, signaling output, and in vivo function. Mol Cell *25*, 889-901.
25. Gianni, S., and Jemth, P. (2017). How Fast Is Protein-Ligand Association? Trends Biochem Sci *42*, 847-849.
26. Zhou, H.X. (2012). Intrinsic disorder: signaling via highly specific but short-lived association. Trends Biochem Sci *37*, 43-48.
27. Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. Nature *545*, 505-509.
28. Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., et al. (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell *162*, 425-440.
29. Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402-408.
30. Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212-1226.
31. Neduva, V., and Russell, R.B. (2006). Peptides mediating interaction networks: new leads at last. Curr Opin Biotechnol *17*, 465-471.
32. Davey, N.E., Simonetti, L., and Ivarsson, Y. (2023). The next wave of interactomics: Mapping the SLiM-based interactions of the intrinsically disordered proteome. Curr Opin Struct Biol *80*, 102593.
33. Benz, C., Ali, M., Krystkowiak, I., Simonetti, L., Sayadi, A., Mihalic, F., Kliche, J., Andersson, E., Jemth, P., Davey, N.E., et al. (2022). Proteome-scale mapping of binding sites in the unstructured regions of the human proteome. Mol Syst Biol *18*, e10584.
34. Davey, N.E., Seo, M.H., Yadav, V.K., Jeon, J., Nim, S., Krystkowiak, I., Blikstad, C., Dong, D., Markova, N., Kim, P.M., et al. (2017). Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome. FEBS J *284*, 485-498.
35. Ivarsson, Y., Arnold, R., McLaughlin, M., Nim, S., Joshi, R., Ray, D., Liu, B., Teyra, J., Pawson, T., Moffat, J., et al. (2014). Large-scale interaction profiling of PDZ domains through proteomic peptide-phage display using human and viral phage peptidomes. Proc Natl Acad Sci U S A *111*, 2542-2547.
36. Kruse, T., Benz, C., Garvanska, D.H., Lindqvist, R., Mihalic, F., Coscia, F., Inturi, R., Sayadi, A., Simonetti, L., Nilsson, E., et al. (2021). Large scale discovery of coronavirus-host factor

protein interaction motifs reveals SARS-CoV-2 specific mechanisms and vulnerabilities. Nat Commun *12*, 6761.

37. Seo, M.H., Nim, S., Jeon, J., and Kim, P.M. (2017). Large-Scale Interaction Profiling of Protein Domains Through Proteomic Peptide-Phage Display Using Custom Peptidomes. Methods Mol Biol *1518*, 213-226.
38. Tonikian, R., Zhang, Y., Boone, C., and Sidhu, S.S. (2007). Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries. Nat Protoc *2*, 1368-1386.
39. Hwang, T., Parker, S.S., Hill, S.M., Grant, R.A., Ilunga, M.W., Sivaraman, V., Mouneimne, G., and Keating, A.E. (2022). Native proline-rich motifs exploit sequence context to target actin-remodeling Ena/VASP protein ENAH. Elife *11*.
40. Li, A., Voleti, R., Lee, M., Gagoski, D., and Shah, N.H. (2023). High-throughput profiling of sequence recognition by tyrosine kinases and SH2 domains using bacterial peptide display. Elife *12*.
41. Teyra, J., Sidhu, S.S., and Kim, P.M. (2012). Elucidation of the binding preferences of peptide recognition modules: SH3 and PDZ domains. FEBS Lett *586*, 2631-2637.
42. Tonikian, R., Zhang, Y., Sazinsky, S.L., Currell, B., Yeh, J.H., Reva, B., Held, H.A., Appleton, B.A., Evangelista, M., Wu, Y., et al. (2008). A specificity map for the PDZ domain family. PLoS Biol *6*, e239.
43. Atwater, J., Mattes, D.S., Streit, B., von Bojnicic-Kninski, C., Loeffler, F.F., Breitling, F., Fuchs, H., and Hirtz, M. (2018). Combinatorial Synthesis of Macromolecular Arrays by Microchannel Cantilever Spotting (microCS). Adv Mater *30*, e1801632.
44. Frank, R. (2002). The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports--principles and applications. J Immunol Methods *267*, 13-26.
45. Nguyen, H.Q., Roy, J., Harink, B., Damle, N.P., Latorraca, N.R., Baxter, B.C., Brower, K., Longwell, S.A., Kortemme, T., Thorn, K.S., et al. (2019). Quantitative mapping of protein-peptide affinity landscapes using spectrally encoded beads. Elife *8*.
46. Shi, G., Song, C., Torres Robles, J., Salichos, L., Lou, H.J., Lam, T.T., Gerstein, M., and Turk, B.E. (2023). Proteome-wide screening for mitogen-activated protein kinase docking motifs and interactors. Sci Signal *16*, eabm5518.
47. Erijman, A., Kozlowski, L., Sohrabi-Jahromi, S., Fishburn, J., Warfield, L., Schreiber, J., Noble, W.S., Soding, J., and Hahn, S. (2020). A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning. Mol Cell *78*, 890-902 e896.
48. Zhang, Z., Sie, B., Chang, A., Leng, Y., Nardone, C., Timms, R.T., and Elledge, S.J. (2023). Elucidation of E3 ubiquitin ligase specificity through proteome-wide internal degron mapping. Mol Cell *83*, 3377-3392 e3376.
49. Cohen, R.D., and Pielak, G.J. (2017). A cell is more than the sum of its (dilute) parts: A brief history of quinary structure. Protein Sci *26*, 403-413.
50. Bhaduri, S., and Pryciak, P.M. (2011). Cyclin-specific docking motifs promote phosphorylation of yeast signaling proteins by G1/S Cdk complexes. Curr Biol *21*, 1615-1623.
51. Bhaduri, S., Valk, E., Winters, M.J., Gruessner, B., Loog, M., and Pryciak, P.M. (2015). A docking interface in the cyclin Cln2 promotes multi-site phosphorylation of substrates and timely cell-cycle entry. Curr Biol *25*, 316-325.
52. Strickfaden, S.C., Winters, M.J., Ben-Ari, G., Lamson, R.E., Tyers, M., and Pryciak, P.M. (2007). A mechanism for cell-cycle regulation of MAP kinase signaling in a yeast differentiation pathway. Cell *128*, 519-531.
53. Stollar, E.J., Garcia, B., Chong, P.A., Rath, A., Lin, H., Forman-Kay, J.D., and Davidson, A.R. (2009). Structural, functional, and bioinformatic studies demonstrate the crucial role of

an extended peptide binding site for the SH3 domain of yeast Abp1p. J Biol Chem *284*, 26918-26927.

54. Guettler, S., LaRose, J., Petsalaki, E., Gish, G., Scotter, A., Pawson, T., Rottapel, R., and Sicheri, F. (2011). Structural basis and sequence rules for substrate recognition by Tankyrase explain the basis for cherubism disease. Cell *147*, 1340-1354.
55. Adler, J.J., Heller, B.L., Bringman, L.R., Ranahan, W.P., Cocklin, R.R., Goebl, M.G., Oh, M., Lim, H.S., Ingham, R.J., and Wells, C.D. (2013). Amot130 adapts atrophin-1 interacting protein 4 to inhibit yes-associated protein signaling and cell growth. J Biol Chem *288*, 15181-15193.
56. Iglesias-Bexiga, M., Palencia, A., Corbi-Verge, C., Martin-Malpartida, P., Blanco, F.J., Macias, M.J., Cobos, E.S., and Luque, I. (2019). Binding site plasticity in viral PPxY Late domain recognition by the third WW domain of human NEDD4. Sci Rep *9*, 15076.
57. Kussie, P.H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A.J., and Pavletich, N.P. (1996). Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. Science *274*, 948-953.
58. Popowicz, G.M., Czarna, A., and Holak, T.A. (2008). Structure of the human Mdmx protein bound to the p53 tumor suppressor transactivation domain. Cell Cycle *7*, 2441-2443.
59. Kay, B.K., Williamson, M.P., and Sudol, M. (2000). The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. FASEB J *14*, 231-241.
60. Zarrinpar, A., Bhattacharyya, R.P., and Lim, W.A. (2003). The structure and function of proline recognition domains. Sci STKE *2003*, RE8.
61. Burgess, A., Chia, K.M., Haupt, S., Thomas, D., Haupt, Y., and Lim, E. (2016). Clinical Overview of MDM2/X-Targeted Therapies. Front Oncol *6*, 7.
62. Shangary, S., and Wang, S. (2009). Small-molecule inhibitors of the MDM2-p53 protein-protein interaction to reactivate p53 function: a novel approach for cancer therapy. Annu Rev Pharmacol Toxicol *49*, 223-241.
63. Pazgier, M., Liu, M., Zou, G., Yuan, W., Li, C., Li, C., Li, J., Monbo, J., Zella, D., Tarasov, S.G., et al. (2009). Structural basis for high-affinity peptide inhibition of p53 interactions with MDM2 and MDMX. Proc Natl Acad Sci U S A *106*, 4665-4670.
64. Crabtree, M.D., Borcherds, W., Poosapati, A., Shammas, S.L., Daughdrill, G.W., and Clarke, J. (2017). Conserved Helix-Flanking Prolines Modulate Intrinsically Disordered Protein:Target Affinity by Altering the Lifetime of the Bound Complex. Biochemistry *56*, 2379-2384.
65. Li, C., Pazgier, M., Li, C., Yuan, W., Liu, M., Wei, G., Lu, W.Y., and Lu, W. (2010). Systematic mutational analysis of peptide inhibition of the p53-MDM2/MDMX interactions. J Mol Biol *398*, 200-213.
66. Bottger, A., Bottger, V., Garcia-Echeverria, C., Chene, P., Hochkeppel, H.K., Sampson, W., Ang, K., Howard, S.F., Picksley, S.M., and Lane, D.P. (1997). Molecular characterization of the hdm2-p53 interaction. J Mol Biol *269*, 744-756.
67. Aragon, E., Goerner, N., Xi, Q., Gomes, T., Gao, S., Massague, J., and Macias, M.J. (2012). Structural basis for the versatile interactions of Smad7 with regulator WW domains in TGF-beta Pathways. Structure *20*, 1726-1736.
68. Bobby, R., Medini, K., Neudecker, P., Lee, T.V., Brimble, M.A., McDonald, F.J., Lott, J.S., and Dingley, A.J. (2013). Structure and dynamics of human Nedd4-1 WW3 in complex with the alphaENaC PY motif. Biochim Biophys Acta *1834*, 1632-1641.
69. Castillo, F., Corbi-Verge, C., Murciano-Calles, J., Candel, A.M., Han, Z., Iglesias-Bexiga, M., Ruiz-Sanz, J., Kim, P.M., Harty, R.N., Martinez, J.C., et al. (2022). Phage display identification of nanomolar ligands for human NEDD4-WW3: Energetic and dynamic implications for the development of broad-spectrum antivirals. Int J Biol Macromol *207*, 308-323.

70. Henry, P.C., Kanelis, V., O'Brien, M.C., Kim, B., Gautschi, I., Forman-Kay, J., Schild, L., and Rotin, D. (2003). Affinity and specificity of interactions between Nedd4 isoforms and the epithelial Na+ channel. J Biol Chem *278*, 20019-20028.
71. Kanelis, V., Bruce, M.C., Skrynnikov, N.R., Rotin, D., and Forman-Kay, J.D. (2006). Structural determinants for high-affinity binding in a Nedd4 WW3* domain-Comm PY motif complex. Structure *14*, 543-553.
72. Qi, S., O'Hayre, M., Gutkind, J.S., and Hurley, J.H. (2014). Structural and biochemical basis for ubiquitin ligase recruitment by arrestin-related domain-containing protein-3 (ARRDC3). J Biol Chem *289*, 4743-4752.
73. Verma, A., Jing-Song, F., Finch-Edmondson, M.L., Velazquez-Campoy, A., Balasegaran, S., Sudol, M., and Sivaraman, J. (2018). Biophysical studies and NMR structure of YAP2 WW domain - LATS1 PPxY motif complexes reveal the basis of their interaction. Oncotarget *9*, 8068-8080.
74. Gfeller, D., Butty, F., Wierzbicka, M., Verschueren, E., Vanhee, P., Huang, H., Ernst, A., Dar, N., Stagljar, I., Serrano, L., et al. (2011). The multiple-specificity landscape of modular peptide recognition domains. Mol Syst Biol *7*, 484.
75. Kanelis, V., Rotin, D., and Forman-Kay, J.D. (2001). Solution structure of a Nedd4 WW domain-ENaC peptide complex. Nat Struct Biol *8*, 407-412.
76. Rheinemann, L., Thompson, T., Mercenne, G., Paine, E.L., Peterson, F.C., Volkman, B.F., Alam, S.L., Alian, A., and Sundquist, W.I. (2021). Interactions between AMOT PPxY motifs and NEDD4L WW domains function in HIV-1 release. J Biol Chem *297*, 100975.
77. Iyer, B.R., and Mahalakshmi, R. (2019). Hydrophobic Characteristic Is Energetically Preferred for Cysteine in a Model Membrane Protein. Biophys J *117*, 25-35.
78. Nagano, N., Ota, M., and Nishikawa, K. (1999). Strong hydrophobic nature of cysteine residues in proteins. FEBS Lett *458*, 69-71.
79. Wimley, W.C., and White, S.H. (1996). Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol *3*, 842-848.
80. Waibl, F., Fernandez-Quintero, M.L., Wedl, F.S., Kettenberger, H., Georges, G., and Liedl, K.R. (2022). Comparison of hydrophobicity scales for predicting biophysical properties of antibodies. Front Mol Biosci *9*, 960194.
81. Madhusudan Makwana, K., and Mahalakshmi, R. (2015). Implications of aromatic-aromatic interactions: From protein structures to peptide models. Protein Sci *24*, 1920-1933.
82. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583-589.
83. Chien, C.T., Bartel, P.L., Sternglanz, R., and Fields, S. (1991). The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. Proc Natl Acad Sci U S A *88*, 9578-9582.
84. Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature *340*, 245-246.
85. Repetto, M.V., Winters, M.J., Bush, A., Reiter, W., Hollenstein, D.M., Ammerer, G., Pryciak, P.M., and Colman-Lerner, A. (2018). CDK and MAPK Synergistically Regulate Signaling Dynamics via a Shared Multi-site Phosphorylation Region on the Scaffold Protein Ste5. Mol Cell *69*, 938-952 e936.
86. Agarwal, V., and McShan, A.C. (2024). The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins. Nat Chem Biol *20*, 950-959.
87. Yin, S., Mi, X., and Shukla, D. (2024). Leveraging machine learning models for peptide-protein interaction prediction. RSC Chem Biol *5*, 401-417.
88. Motmaen, A., Dauparas, J., Baek, M., Abedi, M.H., Baker, D., and Bradley, P. (2023). Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. Proc Natl Acad Sci U S A *120*, e2216697120.

89. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res *47*, D941-D947.
90. Cheng, J., Novati, G., Pan, J., Bycroft, C., Zemgulyte, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science *381*, eadg7492.
91. Corbi-Verge, C., Garton, M., Nim, S., and Kim, P.M. (2017). Strategies to Develop Inhibitors of Motif-Mediated Protein-Protein Interactions as Drug Leads. Annu Rev Pharmacol Toxicol *57*, 39-60.
92. Nim, S., Jeon, J., Corbi-Verge, C., Seo, M.H., Ivarsson, Y., Moffat, J., Tarasova, N., and Kim, P.M. (2016). Pooled screening for antiproliferative inhibitors of protein-protein interactions. Nat Chem Biol *12*, 275-281.
93. Rothstein, R. (1991). Targeting, disruption, replacement, and allele rescue: integrative DNA transformation in yeast. Methods Enzymol *194*, 281-301.
94. Sherman, F. (2002). Getting started with yeast. Methods Enzymol *350*, 3-41.
95. Mumberg, D., Muller, R., and Funk, M. (1994). Regulatable promoters of Saccharomyces cerevisiae: comparison of transcriptional activity and their use for heterologous expression. Nucleic Acids Res *22*, 5767-5768.
96. Janke, C., Magiera, M.M., Rathfelder, N., Taxis, C., Reber, S., Maekawa, H., Moreno-Borchart, A., Doenges, G., Schwob, E., Schiebel, E., et al. (2004). A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. Yeast *21*, 947-962.
97. Winters, M.J., and Pryciak, P.M. (2018). Analysis of the thresholds for transcriptional activation by the yeast MAP kinases Fus3 and Kss1. Mol Biol Cell *29*, 669-682.
98. Pope, P.A., Bhaduri, S., and Pryciak, P.M. (2014). Regulation of cyclin-substrate docking by a G1 arrest signaling pathway and the Cdk inhibitor Far1. Curr Biol *24*, 1390-1396.
99. Hietpas, R., Roscoe, B., Jiang, L., and Bolon, D.N. (2012). Fitness analyses of all possible point mutations for regions of genes in yeast. Nat Protoc *7*, 1382-1396.
100. Rubin, A.F., Gelman, H., Lucas, N., Bajjalieh, S.M., Papenfuss, A.T., Speed, T.P., and Fowler, D.M. (2017). A statistical framework for analyzing deep mutational scanning data. Genome Biol *18*, 150.
101. Bravo, A.M., Typas, A., and Veening, J.W. (2022). 2FAST2Q: a general-purpose sequence search and counting program for FASTQ files. PeerJ *10*, e14041.
102. Krystkowiak, I., and Davey, N.E. (2017). SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. Nucleic Acids Res *45*, W464-W469.
103. Mirdita, M., Schutze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat Methods *19*, 679-682.
104. Meng, E.C., Goddard, T.D., Pettersen, E.F., Couch, G.S., Pearson, Z.J., Morris, J.H., and Ferrin, T.E. (2023). UCSF ChimeraX: Tools for structure building and analysis. Protein Sci *32*, e4792.
105. Hagemans, D., van Belzen, I.A., Moran Luengo, T., and Rudiger, S.G. (2015). A script to highlight hydrophobicity and charge on protein surfaces. Front Mol Biosci *2*, 56.

Abbreviations

The abbreviations used are: SLiM, short linear motif; SIMBA, systematic intracellular motif binding analysis; DMS, deep mutational scanning; CDK, cyclin dependent kinase; MAPK, mitogen activated protein kinase; PSSM, position-specific scoring matrix; NGS, next generation sequencing; ARC, ankyrin repeat cluster; GST, glutathione S-transferase.

FIGURE LEGENDS

Figure 1. Overview of the SIMBA system
(A) Generic diagram emphasizing how SLiM sequence variants can differ in binding strength.
(B) Top, CDK activity associated with the cyclin Cln2 blocks signaling through a pheromone-stimulated growth arrest pathway. Bottom, a chimeric signaling protein, Ste20^{Ste5PM}, contains a plasma membrane binding domain (PM) and flanking phosphorylation sites from Ste5 joined to signaling domains from Ste20. A SLiM docking site for Cln2 promotes CDK phosphorylation of Ste20^{Ste5PM} at multiple sites, which inhibits signaling.
(C) The role of cyclin docking can be replaced using foreign SLiMs and their binding domains, such that the foreign interaction allows cells to escape the growth arrest signal.
(D) Libraries of Ste20^{Ste5PM} derivatives harboring large numbers of SLiM variants are introduced into yeast cells and analyzed by competitive growth.
(E) The rate of SLiM enrichment or depletion is an indicator of binding strength.

Figure 2. Test cases and low-throughput assays of SLiM binding via the SIMBA system
(A) Top, tested domains and their motif types. Bottom, summary of results presented in Figures 2B and S1D-F. The plot shows functional strength measured by SIMBA in vivo vs. previously measured affinities in vitro. Error bars are omitted for clarity. See also Figure S1C.
(B) Cln2 and a TNKS2^{ARC4}-Cln2 fusion were coexpressed with Ste20^{Ste5PM} derivatives harboring known TNKS2^{ARC4}-binding peptides, and then pheromone signaling was assayed. Top, representive blots; SLiM binding blocks activation of the MAPK Fus3 (P-MAPK) and promotes a mobility shift of the myc-tagged Ste20^{Ste5PM} substrate. Bottom, quantification of results (mean ± SD; n = 2 [Cln2] or n = 6 [TNKS2^{ARC4}-Cln2]), compared with in vitro binding affinity [54]. See also Figure S1A-B.
(C) Improved resolution of strong interactions by expressing the MDM2^{SWIB} domain from a weaker promoter (P_{GALL}) vs. the full strength promoter (P_{GAL1}). Bars, mean ± SD (n = 4). See also Figure S1C.

Figure 3. Interrogation of TNKS2^{ARC4}-binding peptides by SIMBA
(A) Heatmap showing effects of all single-position substitutions in the 3BP2 motif. Colors denote enrichment (red) or depletion (blue) relative to the wild-type motif, calculated as log scores by Enrich2 software (see Methods). The asterisk denotes a termination codon. The data average four independent experiments using P_{GAL1}-driven TNKS2^{ARC4}-Cln2 (two each of N-site and C-site libraries).
(B) Logos showing TNKS2^{ARC4} sequence preferences observed for the 3BP2 motif in vitro [54] and in vivo (this study), plus the average in vivo results from 5 distinct motifs (see also Figure S2B-D). For clarity, only positive preferences are shown (see Methods).
(C) Plots of SIMBA scores vs. K_D for all 3BP2 variants as well as representative individual positions. Yellow, wild-type; pink, missense variants. SIMBA scores are mean ± SEM (n = 4). K_D values [54] are the mean of technical duplicates (n = 1 ± SE of the regression fit); K_D values that were not quantifiable in vitro were assigned values of 200 μM to allow inclusion in the plot. See also Figure S2A.

(D) SIMBA scores vs. K_D values [54] for the wild-type AXIN1 peptide plus 9 Ala mutants, plotted as in panel (C). See also Figure S2B-D.
(E) Comparison of STOP-normalized SIMBA scores to in vitro K_D [54] for 6 wild-type peptides. See also Figure S2B-D.
(F) Diagram of N-site and C-site locations for inserting SLiM sequences.
(G) Correlation of SIMBA scores for variants of the RPGopt motif in N-site vs. C-site locations. Data are the mean ± range (n = 2). See also Figure S2E.
(H) Correlation of SIMBA scores for RPGopt motif variants when the TNKS2^{ARC4}-Cln2 fusion was expressed from different strength promoters (P_{GAL1} vs. P_{GALS}). Data are the mean ± SEM (n = 4; two each of N-site and C-site libraries). See also Figure S2F.
(I) Summary of Pearson correlation coefficients (r) for the indicated pairwise comparisons. See also Figures S2E-F.

Figure 4. Local context affects residue preferences at some positions
(A) Logo comparing the residue preferences at each peptide position in the context of 5 different parent peptides that bind TNKS2^{ARC4} (labeled i-v; see panel [B]).
(B) Sequence and K_D of parent peptides (i-v) analyzed in panel (A). Yellow, Pro residues flanking p4 in MCL1 and AXIN1. Blue, residues that are atypically preferred in MCL1 and AXIN1.
(C) Preference ranks of Pro and Gly at p4 in 5 different parent motifs.
(D) Similar trajectories and p4 contacts of 3BP2 and MCL1 peptides bound to TNKS2^{ARC4} [54]. PDB IDs: 3twr, 3twu.
(E) Rotated view of the TNKS2^{ARC4}-bound MCL1 peptide, showing left-handed trajectory of poly-proline sequence from p2 to p4

Figure 5. Determinants of binding strength for MDM2^{SWIB} peptide ligands
(A) Top, heatmaps of DMS results for p53 and NUMB peptides (mean, n = 2, for each peptide). Middle, sequence preference logos derived from the SIMBA scores. Bottom, logo from the ProP-PD database (https://slim.icr.ac.uk/proppd/) [33] showing residue frequencies in peptides captured by MDM2^{SWIB} phage display; this logo is shown twice to facilitate comparison with the SIMBA preferences in each of the two peptide contexts above.
(B) Distribution of SIMBA scores for all missense variants (black circles) and the average STOP codon (red X symbols) at the 15 positions in each of the 2 parent motifs.
(C) SIMBA scores vs. K_D values [65] for the wild-type p53 peptide (yellow) and 12 Ala mutants (pink). K_D values that were not quantifiable in vitro were assigned as 200 μM to allow inclusion in the plot.
(D) Plot of SIMBA scores for Ala residues (large, orange-filled circles) vs. all other residues (small, unfilled circles) at each position in the p53 and NUMB peptides.
(E) Bars show Gini coefficients, calculated from normalized SIMBA scores, for each position in the p53 and NUMB peptides. Black dots connected by a grey line show previous scoring of the number of unpermitted substitutions at each position of an optimized MDM2^{SWIB}-binding peptide (MPR$\underline{F}$MDY$\underline{W}$EG$\underline{L}$N) [66].

Figure 6. Specificity determinants and context effects in WW domain-binding peptides
(A) Scatterplot of z-scores (median, n = 6) for peptides tested for YAP^{WW1} and NEDD4^{WW3} binding.
(B) Sequences and binding z-scores (median, n = 6) for 22 peptides, sorted by YAP^{WW1} binding strength. The core [LP]PxY motif is colored red, and Pro residues at p+1 are colored blue.
(C) Logos showing average sequence preferences for each WW domain.
(D) Logos comparing preferences at each peptide position in the context of multiple parent peptides (identified at right for each WW domain). At bottom, plum and tan bars mark the distinct type 1 and 2 preference patterns summarized in panel (E).

(E) Logos showing distinct YAP^{WW1} preferences for type 1 and 2 peptides.
(F) Sequences of the type 1 and 2 peptides that contribute to the logos in panel (E). Yellow highlighting and blue font mark Pro residues at non-core positions in type 2 peptides.
(G) AARDC3 peptide bound to the NEDD4^{WW3} domain (PDB ID: 4n7h) [72]. Hydrophobic regions on the NEDD4^{WW3} surface are colored yellow [105].

Figure 7. Contingency relationships between distinct peptide positions
(A) Logos of p-1 preferences in 21 different peptide contexts. Seven 4-residue C-flank sequences were appended to each of three core motifs (...[LP]PxY), and then the p-1 codons were randomized and tested for binding YAP^{WW1} and NEDD4^{WW3}. For clarity, the logos show only the most preferred residues, and they omit data from peptides for which the difference in z-score between the tested WW-Cln2 fusion and unfused Cln2 was less than 1.
(B) Logos showing preferences at p-1 when the p+2 residue is varied (left), or preferences at p+2 when the p-1 residue is varied (right).
(C) Logos showing preferences at p+3 or p+4 when the p+1 residue is varied. Data in panels A-C are derived from the median synonym z-scores from 3 independent experiments.
(D) Predicted conformations of peptides bound to YAP^{WW1}. Bottom, prediction for the UGR2 peptide, showing 3 side chains for reference (-2P, 0Y, +1S). Above, analogous predictions for peptides with p+1 Trp substitutions (+1W), aligned with the bottom structure and hiding the YAP^{WW1} domain for clarity. The Trp groups are predicted to lie next to the p0 Tyr groups (0Y) with their planar faces in a perpendicular geometry.

Figure 8. Stratification of [LP]PxY matches by PSSM scoring
All 1730 matches to the [LP]PxY consensus sequence ("hits") in disordered regions of the human proteome were scored using 6 different PSSMs. For each of 2 WW domains, 3 PSSMs were derived from the DMS results: one from all motifs (average), one from type 1 motifs only, and one from type 2 motifs only. The PSSM values for individual residues were summed across 10 motif positions (xx[LP]PxYxxxx) to calculate a predicted score (PSSM sum) for all sequences. For all hits, the distribution of sums obtained using each PSSM is shown; pink lines denote median and quartile values. For all motifs tested by SIMBA ("tested"), PSSM sums were calculated similarly, and their symbol sizes are proportional to their observed SIMBA scores (see Figure 6B).

SUPPORTING INFORMATION

Supporting Figure Legends

Figure S1. Low-throughput assays of SLiM binding to six domains

(A) List of tested TNKS2^{ARC4}-binding peptides. K_D values are from [54].
(B) The experiment shown in Figure 2B was repeated using the transcriptional reporter, FUS1-lacZ (mean ± SD; n = 4).
(C) Comparison of the fold inhibition in FUS1-lacZ vs. P-MAPK assays, for all peptides in panel A. Both assays show affinity-dependent inhibition, but the P-MAPK assay detects weak binding with greater sensitivity. Fold inhibition (mean ± SD, n = 4) is the difference in signal between the test peptide and the control, as a fraction of the control signal. K_D values are mean ± SE [54].
(D) SLiM recognition by the Abp1^{SH3} domain, assayed using the FUS1-lacZ reporter. Bars, mean ± SD (n = 3). K_D values from [53].
(E) SLiM recognition by MDM2^{SWIB} and MDM4^{SWIB} domains, assayed as in Figure 2B. Bars, mean ± range (n = 2). K_D values from [33, 65].
(F) SLiM recognition by YAP^{WW1} and NEDD4^{WW3} domains, assayed as in Figure 2B. Bars, mean ± range (n = 2). Note that K_D values from distinct studies [55, 56] might not be directly comparable.
(G) Protein levels of GST-tagged Cln2 fusions expressed from promoters of different strengths (P_{GAL1}, P_{GALS}, P_{GALL}). Expression was induced with galactose for 105 min. G6PDH served as a loading control.

Figure S2. SIMBA interrogation of multiple TNKS2^{ARC4}-binding peptides

(A) SIMBA scores vs. K_D for all 3BP2 variants as well as variants at each of 8 individual positions, plotted as in Figure 3C, with the TNKS2^{ARC4}-Cln2 fusion driven by P_{GAL1} or P_{GALS}. All SIMBA scores are mean ± SEM (n = 4).
(B) Sequence and K_D of each motif subjected to DMS and SIMBA. Residues in the 8 core positions are colored where identical to (blue) or different from (orange) the 3BP2 sequence.
(C) Heatmaps of DMS results from six parent peptides, plotted as in Figure 3A.
(D) Distribution of SIMBA scores for all missense variants (black circles) and the average STOP codon (red X symbols) at the 12 positions in each of the 6 parent motifs.
(E) Correlation of SIMBA scores for variants inserted a the N-site vs. C-site locations, plotted as in Figure 3G, for six parent peptides.
(F) Correlation of SIMBA scores when the TNKS2^{ARC4}-Cln2 fusion was expressed from P_{GAL1} vs. P_{GALS}, plotted as in Figure 3F, for six parent peptides.
(G) Sequence preference logos, averaged from 5 motifs (excluding 3BP2-E5), comparing when the TNKS2^{ARC4}-Cln2 fusion was expressed from P_{GAL1} vs. P_{GALS}.

Figure S3. DMS data and characteristics of WW domain-binding peptides

(A) WW domain peptides used in this study, and their reported binding affinities. The consensus peptides (PYcon1-4) were designed based on “multiple PWM” sequences from Figure S1B of Ref [74], which were proposed as improved predictors of binding specificity compared to single PWMs.
(B) Heatmaps of DMS results for 12 peptides combined with 2 WW domains. Data are the mean of 3 independent replicate experiments.
(C) Conformations of WW domain bound peptides in published structures. The orientation is comparable to that in Figure 6G. PDB IDs: 4n7h, 2m3o, 2ez5, 5ydx, 2ltw.
(D) Plot of buried surface area (BSA) for residues in WW-bound peptides, expressed as a percentage of accessible surface area in the unbound peptide. Data are from 4 NEDD4^{WW3} complexes (all type 1) and 2 YAP^{WW1} complexes (type 2). Peptide names include PDB IDs.

Supporting Tables

Table S1. Yeast strains used in this study

Name	Relevant genotype *	Source
PPY2551	*MAT***a** *bar1Δ STE5-8A ste20Δ::kanMX6 HIS3::P*$_{GAL1}$*-GST-CLN2*	Ref. [20]
PPY2617	*MAT***a** *bar1Δ STE5-8A ste20Δ::kanMX6 FUS1::FUS1-lacZ::LEU2*	This study
PPY2622	*MAT***a** *bar1Δ STE5-8A ste20Δ::kanMX6 FUS1::FUS1-lacZ::LEU2 HIS3::P*$_{GAL1}$*-GST-MDM2*^{SWIB}*-CLN2*	This study
PPY2623	*MAT***a** *bar1Δ STE5-8A ste20Δ::kanMX6 FUS1::FUS1-lacZ::LEU2 HIS3::P*$_{GAL1}$*-GST-TNKS2*^{ARC4}*-CLN2*	This study
PPY2625	*MAT***a** *bar1Δ STE5-8A ste20Δ::kanMX6 FUS1::FUS1-lacZ::LEU2 HIS3::P*$_{GALL}$*-GST-MDM2*^{SWIB}*-CLN2*	This study
PPY2628	*MAT***a** *bar1Δ STE5-8A ste20Δ::kanMX6 FUS1::FUS1-lacZ::LEU2 HIS3::P*$_{GALS}$*-GST-MDM2*^{SWIB}*-CLN2*	This study
PPY2629	*MAT***a** *bar1Δ STE5-8A ste20Δ::kanMX6 FUS1::FUS1-lacZ::LEU2 HIS3::P*$_{GALL}$*-GST-TNKS2*^{ARC4}*-CLN2*	This study
PPY2631	*MAT***a** *bar1Δ STE5-8A ste20Δ::kanMX6 FUS1::FUS1-lacZ::LEU2 HIS3::P*$_{GALS}$*-GST-TNKS2*^{ARC4}*-CLN2*	This study
PPY2708	*MAT***a** *bar1Δ STE5-8A ste20Δ::kanMX6 FUS1::FUS1-lacZ::LEU2 HIS3::P*$_{GAL1}$*-GST-CLN2*	This study
PPY2727	*MAT***a** *bar1Δ STE5-8A ste20Δ::kanMX6 FUS1::FUS1-lacZ::LEU2 HIS3::P*$_{GAL1}$*-GST-YAP*^{WW1}*-CLN2*	This study
PPY2732	*MAT***a** *bar1Δ STE5-8A ste20Δ::kanMX6 FUS1::FUS1-lacZ::LEU2 HIS3::P*$_{GAL1}$*-GST-NEDD4*^{WW3}*-CLN2*	This study

* All strains in W303 background (*ADE2 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1*)

Table S2. Plasmids used in this study

SUBSTRATE fusion plasmids

Name	Description	Peptide Motif Insert	Source
pPP4365	CEN URA3 myc13-(LP_Ste20 motif)-(Ste5 1-85)-(Ste20 312-939)	LP_Ste20 = VSLDDPIQFTA	Ref. [20]
pPP4366	CEN URA3 myc13-(LP_Ste5 motif)-(Ste5 1-85)-(Ste20 312-939)	LP_Ste5 = SPLLPPFGLSY	Ref. [20]
pPP4367	CEN URA3 myc13-(LP_Sic1 motif)-(Ste5 1-85)-(Ste20 312-939)	LP_Sic1 = EVLLPPSRPTS	Ref. [20]
pPP4368	CEN URA3 myc13-(LP_Whi5 motif)-(Ste5 1-85)-(Ste20 312-939)	LP_Whi5 = MPLLPPTTPKS	Ref. [20]
pPP4369	CEN URA3 myc13-(LP_Lam5 motif)-(Ste5 1-85)-(Ste20 312-939)	LP_Lam5 = KQLGPPFEHAS	Ref. [20]
pPP4375	CEN URA3 myc13-(nonLP_control motif)-(Ste5 1-85)-(Ste20 312-939)	nonLP = SNGNGSGSNGN	Ref. [20]
pPP4518	CEN URA3 myc13-(Amot_PY1 motif)-(Ste5 1-85)-(Ste20 312-939)	Amot_PY1 = NNEELPTYEEAK	This study
pPP4519	CEN URA3 myc13-(Amot_PY2 motif)-(Ste5 1-85)-(Ste20 312-939)	Amot_PY2 = HRGPPPEYPFKG	This study
pPP4521	CEN URA3 myc13-(PR_Ark1 motif)-(Ste5 1-85)-(Ste20 312-939)	PR_Ark1 = KKTKPTPPPKPSHLK	This study
pPP4522	CEN URA3 myc13-(PR_Prk1 motif)-(Ste5 1-85)-(Ste20 312-939)	PR_Prk1 = KSRPPRPPPKPLHLR	This study
pPP4523	CEN URA3 myc13-(PY_53BP2 motif)-(Ste5 1-85)-(Ste20 312-939)	PY_53BP2 = YPPYPPPPYPSG	This study
pPP4524	CEN URA3 myc13-(PY_Marburg motif)-(Ste5 1-85)-(Ste20 312-939)	PY_Marburg = MQYLNPPPYADH	This study
pPP4525	CEN URA3 myc13-(PY_Rabies motif)-(Ste5 1-85)-(Ste20 312-939)	PY_Rabies = DLWLPPPEYVPL	This study
pPP4526	CEN URA3 myc13-(REctrl motif)-(Ste5 1-85)-(Ste20 312-939)	REctrl = GSGSRLKPRGGN	This study
pPP4531	CEN URA3 myc13-(p53 motif)-(Ste5 1-85)-(Ste20 312-939)	p53 = SQETFSDLWKLLPEN	This study

pPP4532	CEN URA3 myc13-(p53-W23A motif)-(Ste5 1-85)-(Ste20 312-939)	p53-W23A = SQETFSDLAKLLPEN	This study
pPP4533	CEN URA3 myc13-(p53-P27A motif)-(Ste5 1-85)-(Ste20 312-939)	p53-P27A = SQETFSDLWKLLAEN	This study
pPP4534	CEN URA3 myc13-(NUMB motif)-(Ste5 1-85)-(Ste20 312-939)	NUMB = PVDPFEAQWAALENK	This study
pPP4536	CEN URA3 myc13-(3BP2 motif)-(Ste5 1-85)-(Ste20 312-939)	3BP2 = LQRSPPDGQSFR	This study
pPP4537	CEN URA3 myc13-(3BP2-E5 motif)-(Ste5 1-85)-(Ste20 312-939)	3BP2-D5E = LQRSPPEGQSFR	This study
pPP4538	CEN URA3 myc13-(3BP2-A5 motif)-(Ste5 1-85)-(Ste20 312-939)	3BP2-D5A = LQRSPPAGQSFR	This study
pPP4539	CEN URA3 myc13-(RPGopt motif)-(Ste5 1-85)-(Ste20 312-939)	RPGopt = LAREAGDGEEAR	This study
pPP4540	CEN URA3 myc13-(NUMA1 motif)-(Ste5 1-85)-(Ste20 312-939)	NUMA1 = LPRTQPDGTSVP	This study
pPP4541	CEN URA3 myc13-(MCL1 motif)-(Ste5 1-85)-(Ste20 312-939)	MCL1 = VARPPPIGAEVP	This study
pPP4649	CEN URA3 myc13-(3BP2-Q5 motif)-(Ste5 1-85)-(Ste20 312-939)	3BP2-D5Q = LQRSPPQGQSFR	This study
pPP4650	CEN URA3 myc13-(3BP2-S3 motif)-(Ste5 1-85)-(Ste20 312-939)	3BP2-P3S = LQRSSPDGQSFR	This study
pPP4651	CEN URA3 myc13-(3BP2-M3 motif)-(Ste5 1-85)-(Ste20 312-939)	3BP2-P3M = LQRSMPDGQSFR	This study
pPP4652	CEN URA3 myc13-(3BP2-A4 motif)-(Ste5 1-85)-(Ste20 312-939)	3BP2-P4A = LQRSPADGQSFR	This study
pPP4653	CEN URA3 myc13-(AXIN1 motif)-(Ste5 1-85)-(Ste20 312-939)	AXIN1 = APRPPVPGEEGE	This study
pPP4654	CEN URA3 myc13-(AXIN1-P2A motif)-(Ste5 1-85)-(Ste20 312-939)	AXIN1-P2A = APRAPVPGEEGE	This study
pPP4655	CEN URA3 myc13-(AXIN1-P3A motif)-(Ste5 1-85)-(Ste20 312-939)	AXIN1-P3A = APRPAVPGEEGE	This study
pPP4656	CEN URA3 myc13-(AXIN1-V4A motif)-(Ste5 1-85)-(Ste20 312-939)	AXIN1-V4A = APRPPAPGEEGE	This study
pPP4657	CEN URA3 myc13-(AXIN1-P5A motif)-(Ste5 1-85)-(Ste20 312-939)	AXIN1-P5A = APRPPVAGEEGE	This study
pPP4745	CEN URA3 myc13-(Ste20 1-86)-(Ste5 1-85)-(C3RE control motif)-(Ste20 312-939)	C3RE = DVCTPR	This study
pPP4747	CEN URA3 myc13-(Ste20 1-86)-(Ste5 1-85)-(LP_Ste5 motif)-(Ste20 312-939)	LP_Ste5 = SPLLPPFGLSYT	This study
pPP4748	CEN URA3 myc13-(Ste20 1-86)-(Ste5 1-85)-(Cnon2 control motif)-(Ste20 312-939)	Cnon2 = GSNGGSGSNGGS	This study
pPP4749	CEN URA3 myc13-(Ste20 1-86)-(Ste5 1-85)-(Cctrl control motif)-(Ste20 312-939)	Cctrl = GGSGGSGGSGGS	This study
pPP4750	CEN URA3 myc13-(Ste20 1-86)-(Ste5 1-85)-(LP_Ste20 motif)-(Ste20 312-939)	LP_Ste20 = VSLDDPIQFTRV	This study
pPP4751	CEN URA3 myc13-(Ste20 1-86)-(Ste5 1-85)-(LP_Lam5 motif)-(Ste20 312-939)	LP_Lam5 = KQLGPPFEHASN	This study
pPP4762	CEN URA3 myc13-(3BP2 motif)-(Ste5 1-85)-(Ste20 312-939)	3BP2 = ggsgLQRSPPDGQSFRggsg	This study
pPP4763	CEN URA3 myc13-(3BP2-D5E motif)-(Ste5 1-85)-(Ste20 312-939)	3BP2-D5E = ggsgLQRSPPEGQSFRggsg	This study
pPP4764	CEN URA3 myc13-(RPGopt motif)-(Ste5 1-85)-(Ste20 312-939)	RPGopt = ggsgLAREAGDGEEARggsg	This study
pPP4765	CEN URA3 myc13-(NUMA1 motif)-(Ste5 1-85)-(Ste20 312-939)	NUMA1 = ggsgLPRTQPDGTSVPggsg	This study
pPP4766	CEN URA3 myc13-(MCL1 motif)-(Ste5 1-85)-(Ste20 312-939)	MCL1 = ggsgVARPPPIGAEVPggsg	This study
pPP4767	CEN URA3 myc13-(AXIN1 motif)-(Ste5 1-85)-(Ste20 312-939)	AXIN1 = ggsgAPRPPVPGEEGEggsg	This study
pPP4768	CEN URA3 myc13-(Ste20 1-86)-(Ste5 1-85)-(3BP2 motif)-(Ste20 312-939)	3BP2-WT = ggsgLQRSPPDGQSFRggsg	This study
pPP4769	CEN URA3 myc13-(Ste20 1-86)-(Ste5 1-85)-(NUMA1 motif)-(Ste20 312-939)	NUMA1 = ggsgLPRTQPDGTSVPggsg	This study
pPP4836	CEN URA3 myc13-(p53 motif)-(Ste5 1-85)-(Ste20 312-939)	p53 = ggsgSQETFSDLWKLLPENggsg	This study
pPP4837	CEN URA3 myc13-(NUMB motif)-(Ste5 1-85)-(Ste20 312-939)	NUMB = ggsgPVDPFEAQWAALENKggsg	This study

CYCLIN fusion plasmids

Name	Description	Binding Domain Insert	Source
pPP3573	CEN HIS3 Pgal1-GST-CLN2(1-372) Tcyc1	none	This study
pPP4517	CEN HIS3 Pgal1-GST-(YAP_WW1)-CLN2(1-372) Tcyc1	YAP_WW1 = aa 164-227	This study
pPP4520	CEN HIS3 Pgal1-GST-(Abp1_SH3)-CLN2(1-372) Tcyc1	Abp1_SH3 = aa 535-592	This study
pPP4527	CEN HIS3 Pgal1-GST-(NEDD4_WW3)-CLN2(1-372) Tcyc1	NEDD4_WW3 = aa 834-878	This study
pPP4542	CEN HIS3 Pgal1-GST-(TNKS2_ARC4)-CLN2(1-372) Tcyc1	TNKS2_ARC4 = aa 488-649	This study
pPP4543	CEN HIS3 Pgal1-GST-(MDM2_SWIB)-CLN2(1-372) Tcyc1	MDM2_SWIB = aa 17-109	This study
pPP4544	CEN HIS3 Pgal1-GST-(MDM4_SWIB)-CLN2(1-372) Tcyc1	MDM4_SWIB = aa 24-108	This study
pPP4575	integrating HIS3 Pgal1-GST-(MDM2_SWIB)-CLN2(1-545) Tcyc1	MDM2_SWIB = aa 17-109	This study
pPP4576	integrating HIS3 Pgal1-GST-(TNKS2_ARC4)-CLN2(1-545) Tcyc1	TNKS2_ARC4 = aa 488-649	This study
pPP4643	integrating HIS3 PgalL-GST-(MDM2_SWIB)-CLN2(1-545) Tcyc1	MDM2_SWIB = aa 17-109	This study
pPP4644	integrating HIS3 PgalS-GST-(MDM2_SWIB)-CLN2(1-545) Tcyc1	MDM2_SWIB = aa 17-109	This study
pPP4659	integrating HIS3 PgalL-GST-(TNKS2_ARC4)-CLN2(1-545) Tcyc1	TNKS2_ARC4 = aa 488-649	This study
pPP4660	integrating HIS3 PgalS-GST-(TNKS2_ARC4)-CLN2(1-545) Tcyc1	TNKS2_ARC4 = aa 488-649	This study
pPP4866	integrating HIS3 Pgal1-GST-(YAP_WW1)-CLN2(1-545) Tcyc1	YAP_WW1 = aa 164-227	This study
pPP4867	integrating HIS3 Pgal1-GST-(NEDD4_WW3)-CLN2(1-545) Tcyc1	NEDD4_WW3 = aa 834-878	This study

Figure 1

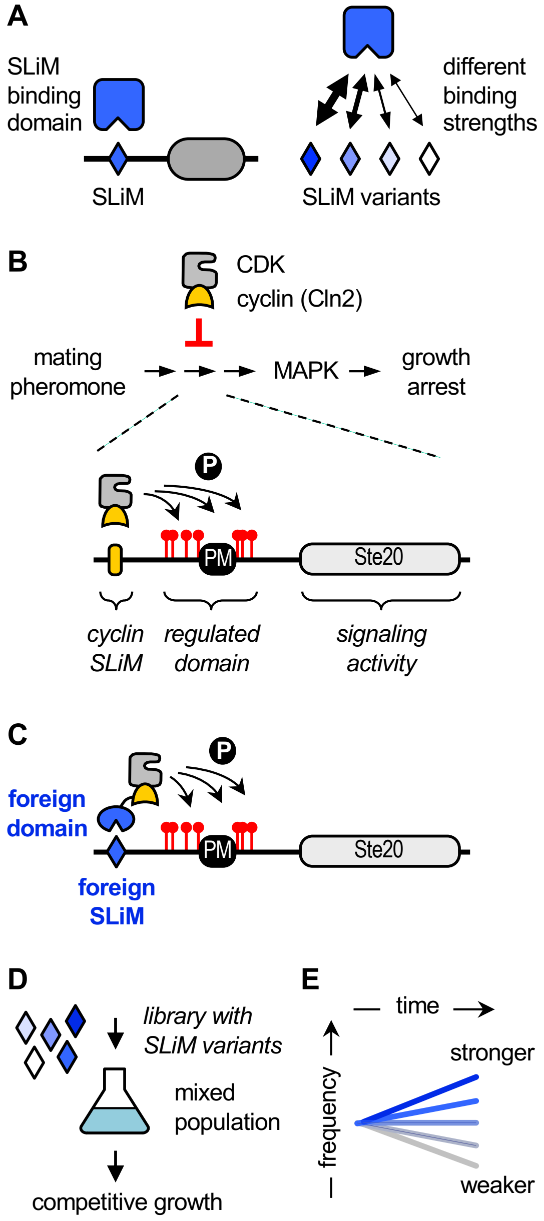


Figure 1. Overview of the SIMBA system

(**A**) Generic diagram emphasizing how SLiM sequence variants can differ in binding strength. (**B**) Top, CDK activity associated with the cyclin Cln2 blocks signaling through a pheromone-stimulated growth arrest pathway. Bottom, a chimeric signaling protein, Ste20^{Ste5PM}, contains a plasma membrane binding domain (PM) and flanking phosphorylation sites from Ste5 joined to signaling domains from Ste20. A SLiM docking site for Cln2 promotes CDK phosphorylation of Ste20^{Ste5PM} at multiple sites, which inhibits signaling. (**C**) The role of cyclin docking can be replaced using foreign SLiMs and their binding domains. (**D**) Libraries of Ste20^{Ste5PM} derivatives harboring large numbers of SLiM variants are introduced into yeast cells and analyzed by competitive growth. (**E**) The rate of SLiM enrichment or depletion is an indicator of binding strength.

Figure 1

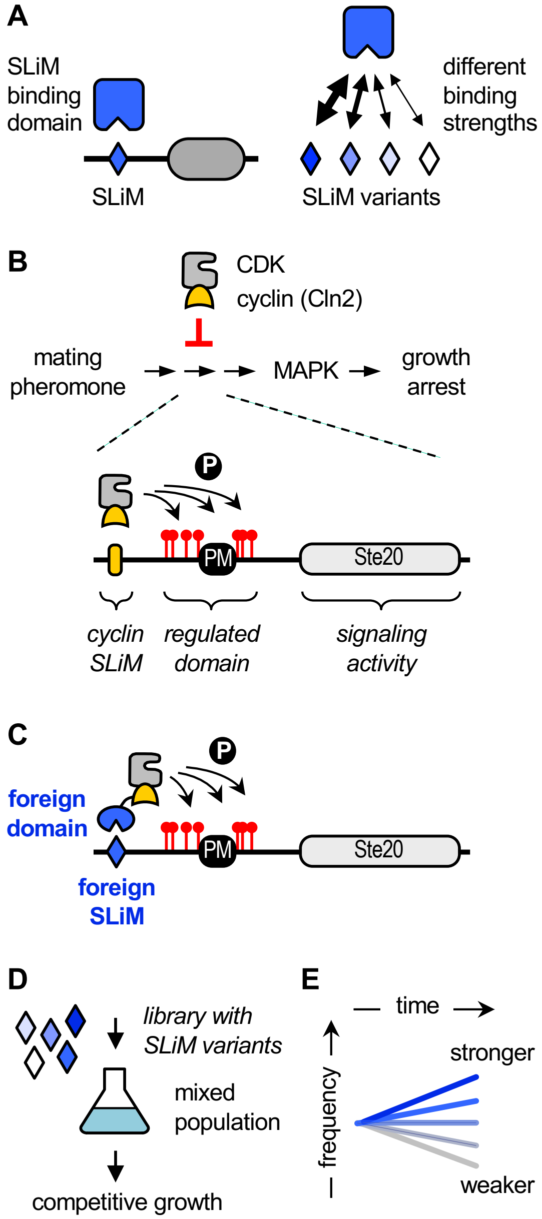
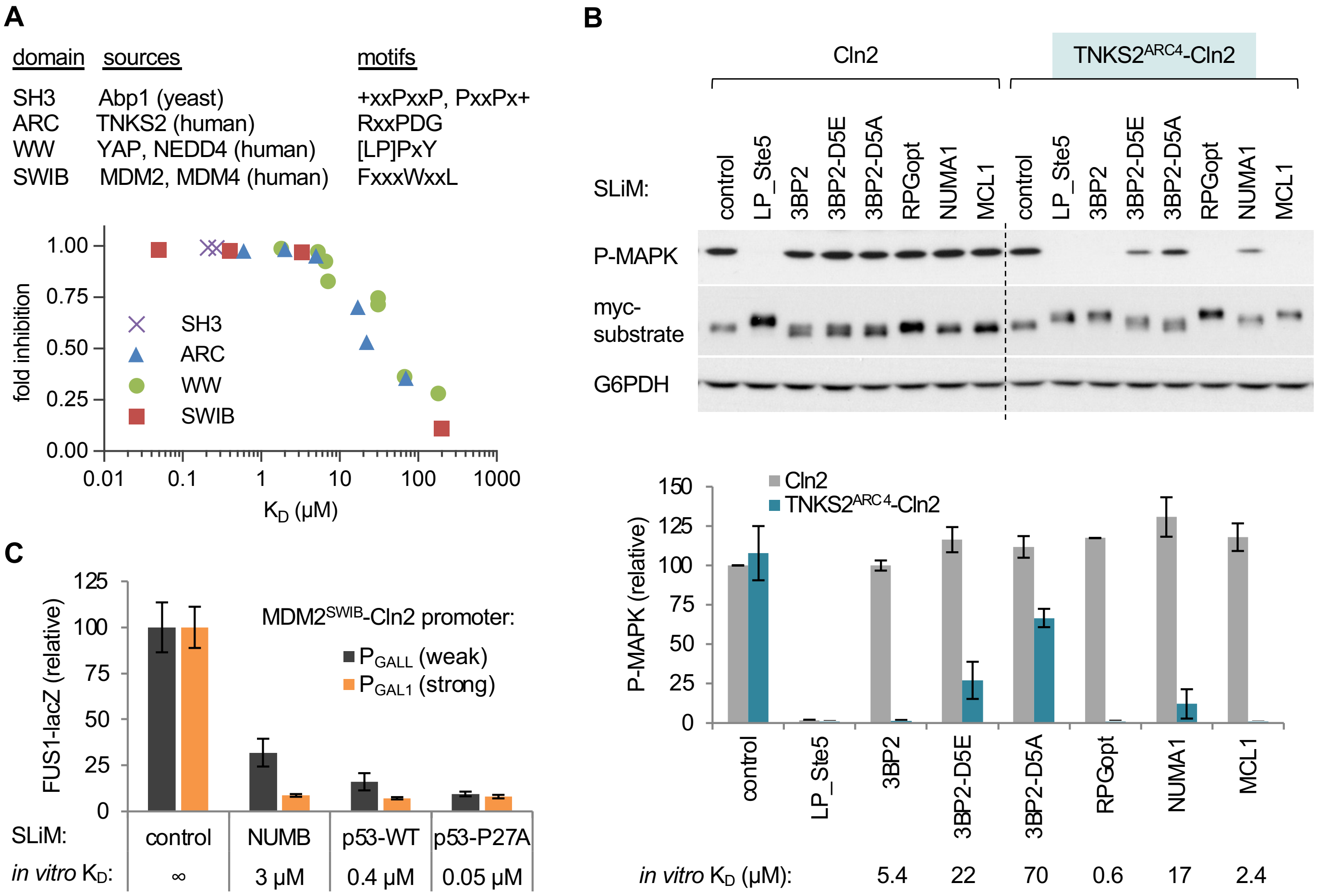


Figure 1. Overview of the SIMBA system

(**A**) Generic diagram emphasizing how SLiM sequence variants can differ in binding strength. (**B**) Top, CDK activity associated with the cyclin Cln2 blocks signaling through a pheromone-stimulated growth arrest pathway. Bottom, a chimeric signaling protein, Ste20^{Ste5PM}, contains a plasma membrane binding domain (PM) and flanking phosphorylation sites from Ste5 joined to signaling domains from Ste20. A SLiM docking site for Cln2 promotes CDK phosphorylation of Ste20^{Ste5PM} at multiple sites, which inhibits signaling. (**C**) The role of cyclin docking can be replaced using foreign SLiMs and their binding domains, such that the foreign interaction allows cells to escape the growth arrest signal. (**D**) Libraries of Ste20^{Ste5PM} derivatives harboring large numbers of SLiM variants are introduced into yeast cells and analyzed by competitive growth. (**E**) The rate of SLiM enrichment or depletion is an indicator of binding strength.

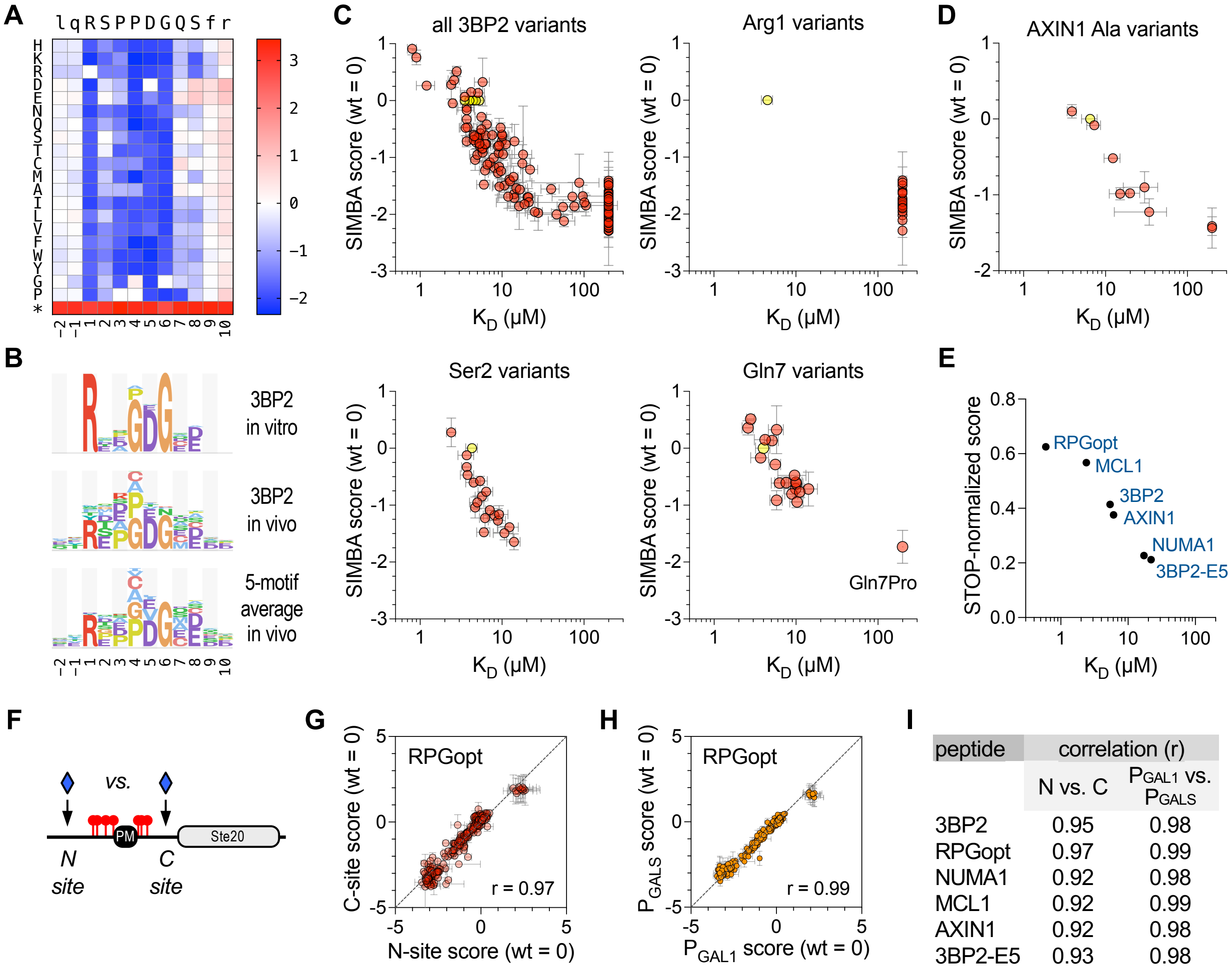
Figure 2



domain	sources	motifs
SH3	Abp1 (yeast)	+xxPxxP, PxxPx+
ARC	TNKS2 (human)	RxxPDG
WW	YAP, NEDD4 (human)	[LP]PxY
SWIB	MDM2, MDM4 (human)	FxxxWxxL

Figure 2. Test cases and low-throughput assays of SLiM binding via the SIMBA system
(**A**) Top, tested domains and their motif types. Bottom, summary of results presented in Figures 2B and S1D-F. The plot shows functional strength measured by SIMBA in vivo vs. previously measured affinities in vitro. Error bars are omitted for clarity. See also Figure S1C. (**B**) Cln2 and a TNKS2^{ARC4}-Cln2 fusion were coexpressed with Ste20^{Ste5PM} derivatives harboring known TNKS2^{ARC4}-binding peptides, and then pheromone signaling was assayed. Top, representive blots; SLiM binding blocks activation of the MAPK Fus3 (P-MAPK) and promotes a mobility shift of the myc-tagged Ste20^{Ste5PM} substrate. Bottom, quantification of results (mean ± SD; n = 2 [Cln2] or n = 6 [TNKS2^{ARC4}-Cln2]), compared with in vitro binding affinity [54]. See also Figure S1A-B. (**C**) Improved resolution of strong interactions by expressing the MDM2^{SWIB} domain from a weaker promoter (P_{GALL}) vs. the full strength promoter (P_{GAL1}). Bars, mean ± SD (n = 4). See also Figure S1C.

Figure 3

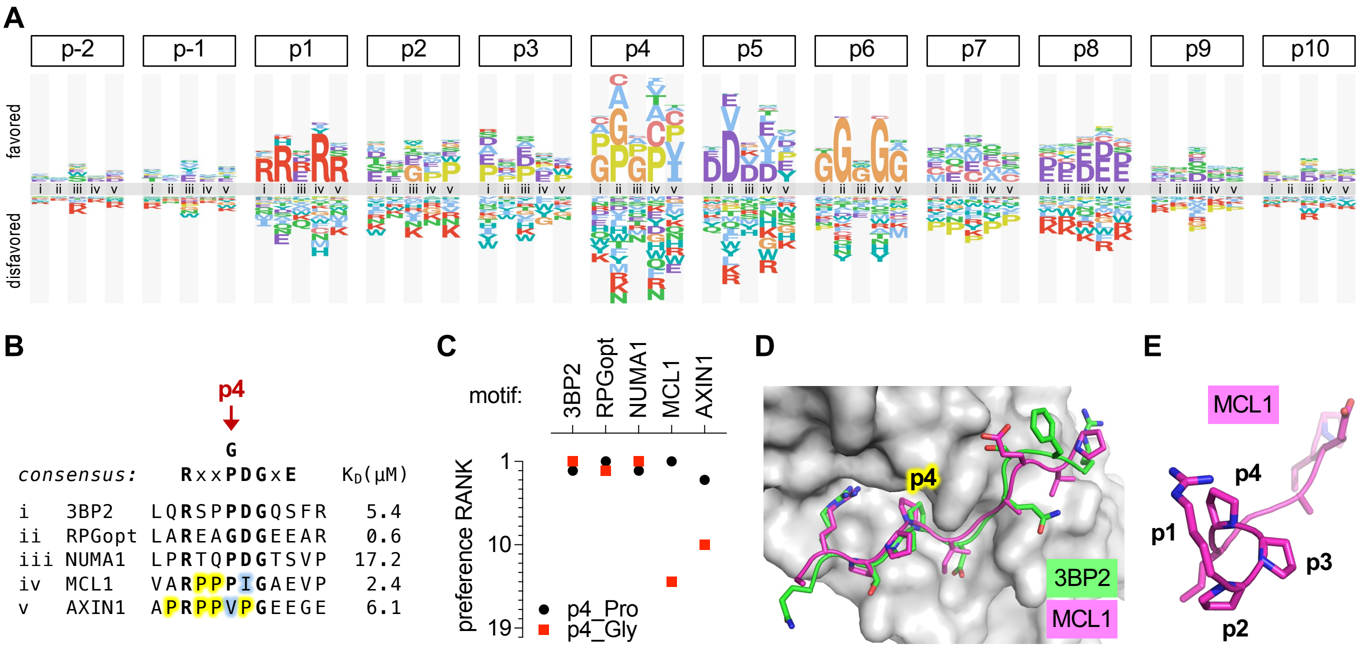


peptide	correlation (r)	
	N vs. C	P_{GAL1} vs. P_{GALS}
3BP2	0.95	0.98
RPGopt	0.97	0.99
NUMA1	0.92	0.98
MCL1	0.92	0.99
AXIN1	0.92	0.98
3BP2-E5	0.93	0.98

Figure 3. Interrogation of TNKS2^{ARC4}-binding peptides by SIMBA

(**A**) Heatmap showing effects of all single-position substitutions in the 3BP2 motif. Colors denote enrichment (red) or depletion (blue) relative to the wild-type motif, calculated as log scores by Enrich2 software (see Methods). The asterisk denotes a termination codon. The data average four independent experiments using P_{GAL1}-driven TNKS2^{ARC4}-Cln2 (two each of N-site and C-site libraries). (**B**) Logos showing TNKS2^{ARC4} sequence preferences observed for the 3BP2 motif in vitro [54] and in vivo (this study), plus the average in vivo results from 5 distinct motifs (see also Figure S2B-D). For clarity, only positive preferences are shown (see Methods). (**C**) Plots of SIMBA scores vs. K_D for all 3BP2 variants as well as representative individual positions. Yellow, wild-type; pink, missense variants. SIMBA scores are mean ± SEM (n = 4). K_D values [54] are the mean of technical duplicates (n = 1 ± SE of the regression fit); K_D values that were not quantifiable in vitro were assigned values of 200 μM to allow inclusion in the plot. See also Figure S2A. (**D**) SIMBA scores vs. K_D values [54] for the wild-type AXIN1 peptide plus 9 Ala mutants, plotted as in panel (C). See also Figure S2B-D. (**E**) Comparison of STOP-normalized SIMBA scores to in vitro K_D [54] for 6 wild-type peptides. See also Figure S2B-D. (**F**) Diagram of N-site and C-site locations for inserting SLiM sequences. (**G**) Correlation of SIMBA scores for variants of the RPGopt motif in N-site vs. C-site locations. Data are the mean ± range (n = 2). See also Figure S2E. (**H**) Correlation of SIMBA scores for RPGopt motif variants when the TNKS2^{ARC4}-Cln2 fusion was expressed from different strength promoters (P_{GAL1} vs. P_{GALS}). Data are the mean ± SEM (n = 4; two each of N-site and C-site libraries). See also Figure S2F. (**I**) Summary of Pearson correlation coefficients (r) for the indicated pairwise comparisons. See also Figures S2E-F.

Figure 4



consensus:		**R**xx**PDG**x**E**	K$_D$(μM)
i	3BP2	LQ**R**SP**PDG**QSFR	5.4
ii	RPGopt	LA**R**EA**GDG**EEAR	0.6
iii	NUMA1	LP**R**TQ**PDG**TSVP	17.2
iv	MCL1	VA**R**PPPI**G**AEVP	2.4
v	AXIN1	AP**R**PPVP**G**EEGE	6.1

Figure 4. Local context affects residue preferences at some positions

(**A**) Logo comparing the residue preferences at each peptide position in the context of 5 different parent peptides that bind TNKS2^{ARC4} (labeled i-v; see panel [B]). (**B**) Sequence and K$_D$ of parent peptides (i-v) analyzed in panel (A). Yellow, Pro residues flanking p4 in MCL1 and AXIN1. Blue, residues that are atypically preferred in MCL1 and AXIN1. (**C**) Preference ranks of Pro and Gly at p4 in 5 different parent motifs. (**D**) Similar trajectories and p4 contacts of 3BP2 and MCL1 peptides bound to TNKS2^{ARC4} [54]. PDB IDs: 3twr, 3twu. (**E**) Rotated view of the TNKS2^{ARC4}-bound MCL1 peptide, showing left-handed trajectory of poly-proline sequence from p2 to p4.

Figure 5

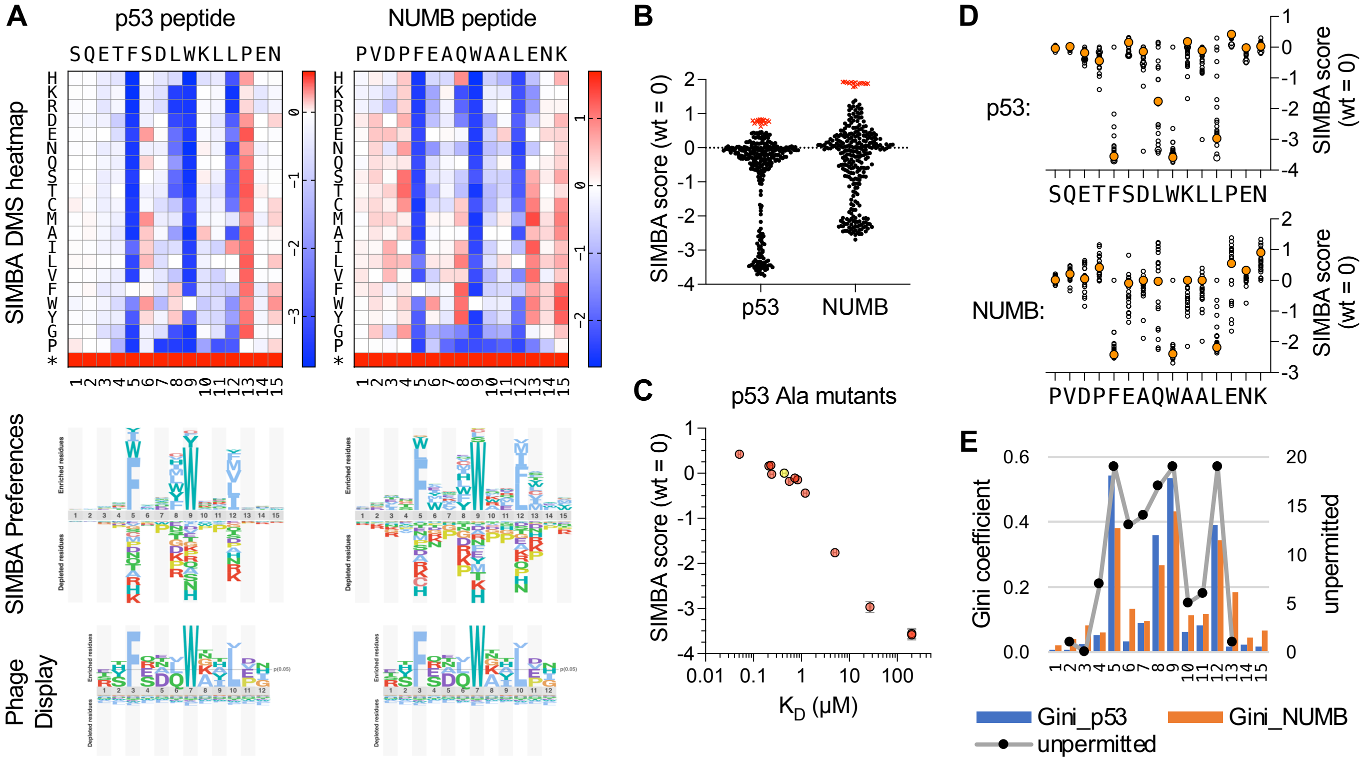
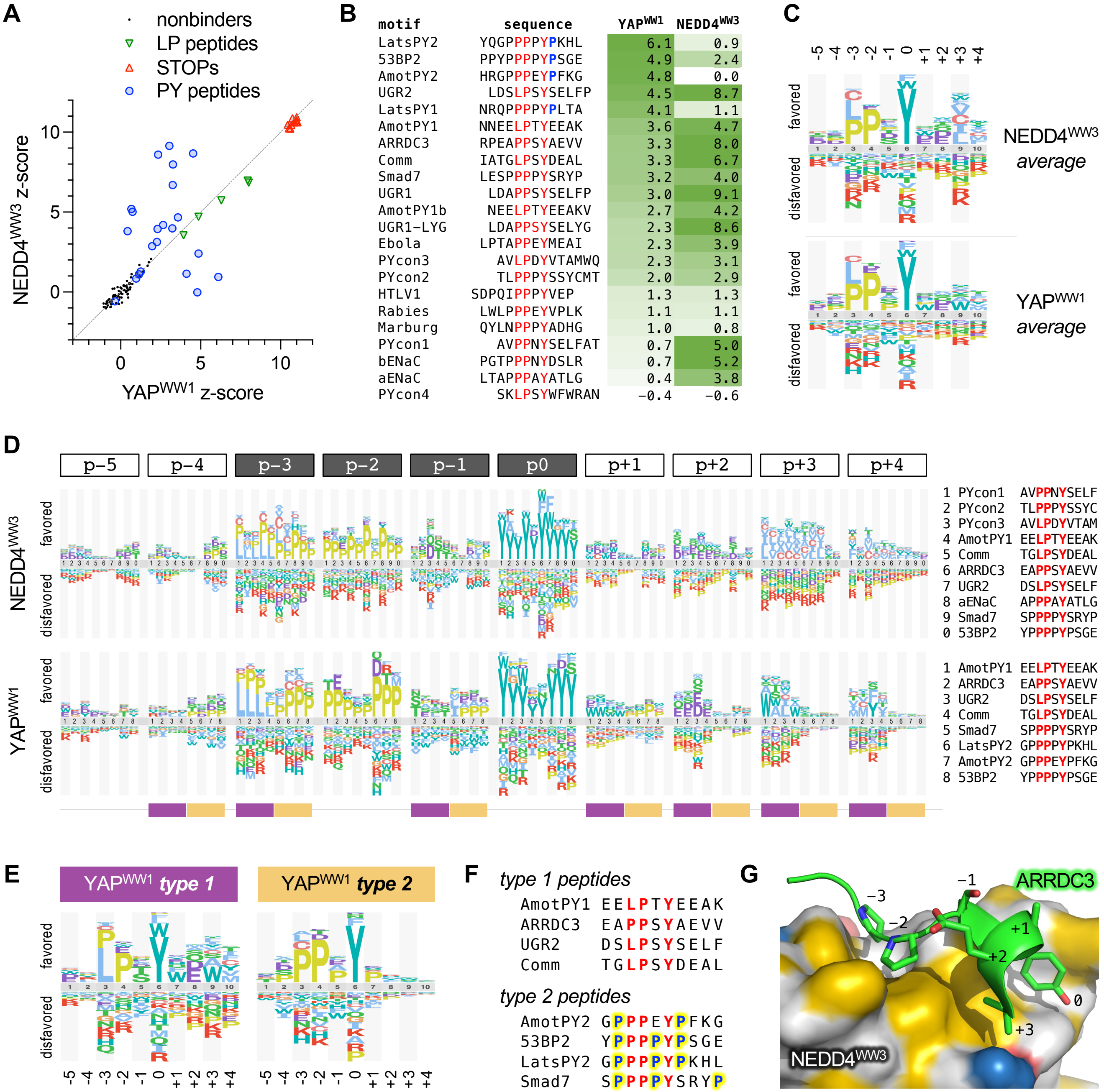


Figure 5. Determinants of binding strength for MDM2^{SWIB} peptide ligands

(**A**) Top, heatmaps of DMS results for p53 and NUMB peptides (mean, n = 2, for each peptide). Middle, sequence preference logos derived from the SIMBA scores. Bottom, logo from the ProP-PD database (https://slim.icr.ac.uk/proppd/) [33] showing residue frequencies in peptides captured by MDM2^{SWIB} phage display; this logo is shown twice to facilitate comparison with the SIMBA preferences in each of the two peptide contexts above. (**B**) Distribution of SIMBA scores for all missense variants (black circles) and the average STOP codon (red X symbols) at the 15 positions in each of the 2 parent motifs. (**C**) SIMBA scores vs. K_D values [65] for the wild-type p53 peptide (yellow) and 12 Ala mutants (pink). K_D values that were not quantifiable in vitro were assigned as 200 μM to allow inclusion in the plot. (**D**) Plot of SIMBA scores for Ala residues (large, orange-filled circles) vs. all other residues (small, unfilled circles) at each position in the p53 and NUMB peptides. (**E**) Bars show Gini coefficients, calculated from normalized SIMBA scores, for each position in the p53 and NUMB peptides. Black dots connected by a grey line show previous scoring of the number of unpermitted substitutions at each position of an optimized MDM2^{SWIB}-binding peptide (MPRFMDYWEGLN) [66].

Figure 6



B

motif	sequence	YAP^{WW1}	NEDD4^{WW3}
LatsPY2	YQGPPPPYPKHL	6.1	0.9
53BP2	PPYPPPPYPSGE	4.9	2.4
AmotPY2	HRGPPPEYPFKG	4.8	0.0
UGR2	LDSLPSYSELFP	4.5	8.7
LatsPY1	NRQPPPPYPLTA	4.1	1.1
AmotPY1	NNEELPTYEEAK	3.6	4.7
ARRDC3	RPEAPPSYAEVV	3.3	8.0
Comm	IATGLPSYDEAL	3.3	6.7
Smad7	LESPPPPYSRYP	3.2	4.0
UGR1	LDAPPSYSELFP	3.0	9.1
AmotPY1b	NEELPTYEEAKV	2.7	4.2
UGR1-LYG	LDAPPSYSELYG	2.3	8.6
Ebola	LPTAPPEYMEAI	2.3	3.9
PYcon3	AVLPDYVTAMWQ	2.3	3.1
PYcon2	TLPPPYSSYCMT	2.0	2.9
HTLV1	SDPQIPPPYVEP	1.3	1.3
Rabies	LWLPPPEYVPLK	1.1	1.1
Marburg	QYLNPPPYADHG	1.0	0.8
PYcon1	AVPPNYSELFAT	0.7	5.0
bENaC	PGTPPPNYDSLR	0.7	5.2
aENaC	LTAPPPAYATLG	0.4	3.8
PYcon4	SKLPSYWFWRAN	-0.4	-0.6

F

type 1 peptides

```
AmotPY1  EELPTYEEAK
ARRDC3   EAPPSYAEVV
UGR2     DSLPSYSELF
Comm     TGLPSYDEAL
```

type 2 peptides

```
AmotPY2  GPPPEYPFKG
53BP2    YPPPPYPSGE
LatsPY2  GPPPPYPKHL
Smad7    SPPPPYSRYP
```

Figure 6. Specificity determinants and context effects in WW domain-binding peptides

(**A**) Scatterplot of z-scores (median, n = 6) for peptides tested for YAP^{WW1} and NEDD4^{WW3} binding. (**B**) Sequences and binding z-scores (median, n = 6) for 22 peptides, sorted by YAP^{WW1} binding strength. The core [LP]PxY motif is colored red, and Pro residues at p+1 are colored blue. (**C**) Logos showing average sequence preferences for each WW domain. (**D**) Logos comparing preferences at each peptide position in the context of multiple parent peptides (identified at right for each WW domain). At bottom, plum and tan bars mark the distinct type 1 and 2 preference patterns summarized in panel (E). (**E**) Logos showing distinct YAP^{WW1} preferences for type 1 and 2 peptides. (**F**) Sequences of the type 1 and 2 peptides that contribute to the logos in panel (E). Yellow highlighting and blue font mark Pro residues at non-core positions in type 2 peptides. (**G**) AARDC3 peptide bound to the NEDD4^{WW3} domain (PDB ID: 4n7h) [71]. Hydrophobic regions on the NEDD4^{WW3} surface are colored yellow [104].

Figure 7

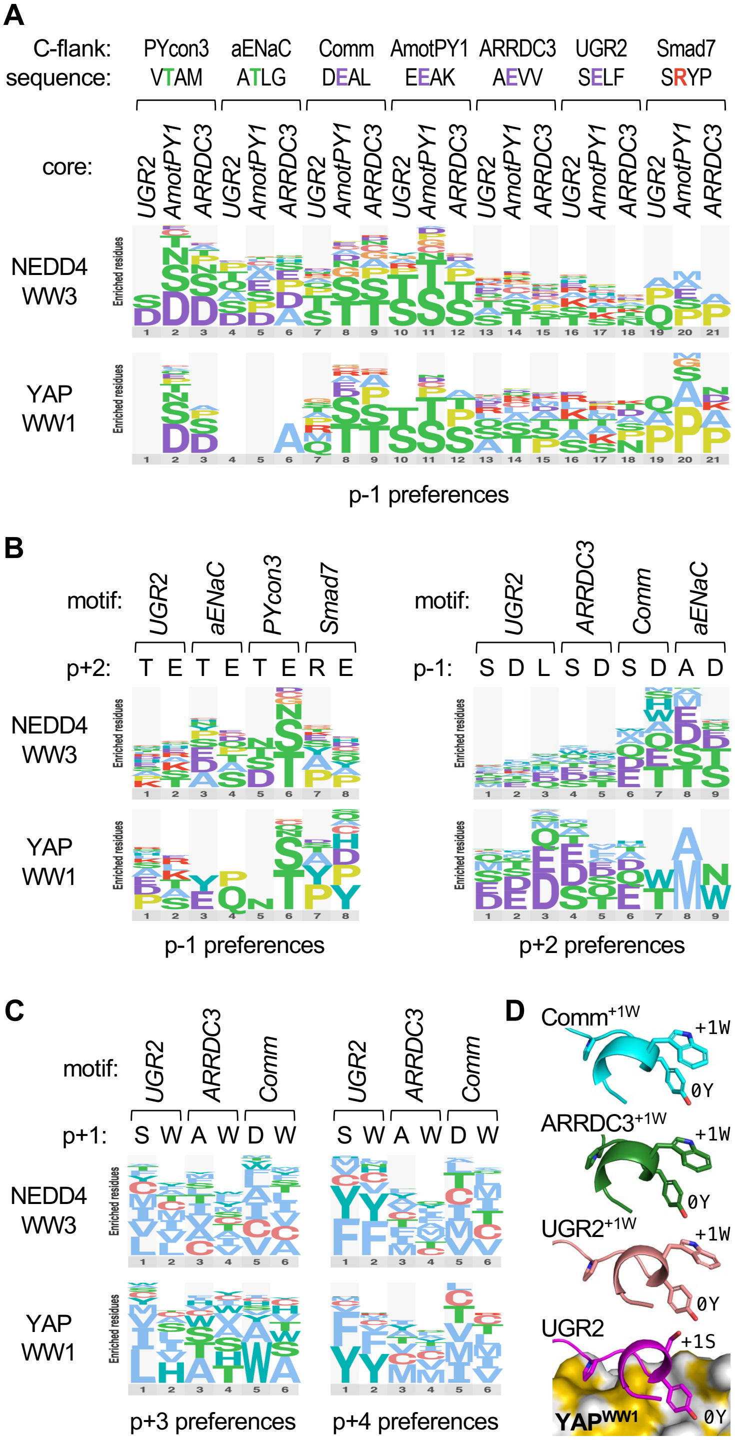


Figure 7. Contingency relationships between distinct peptide positions

(**A**) Logos of p-1 preferences in 21 different peptide contexts. Seven 4-residue C-flank sequences were appended to each of three core motifs (...[LP]PxY), and then the p-1 codons were randomized and tested for binding YAP^{WW1} and NEDD4^{WW3}. For clarity, the logos show only the most preferred residues, and they omit data from peptides for which the difference in z-score between the tested WW-Cln2 fusion and unfused Cln2 was less than 1. (**B**) Logos showing preferences at p-1 when the p+2 residue is varied (left), or preferences at p+2 when the p-1 residue is varied (right). (**C**) Logos showing preferences at p+3 or p+4 when the p+1 residue is varied. Data in panels A-C are derived from the median synonym z-scores from 3 independent experiments. (**D**) Predicted conformations of peptides bound to YAP^{WW1}. Bottom, prediction for the UGR2 peptide, showing 3 side chains for reference (-2P, 0Y, +1S). Above, analogous predictions for peptides with p+1 Trp substitutions (+1W), aligned with the bottom structure and hiding the YAP^{WW1} domain for clarity. The Trp groups are predicted to lie next to the p0 Tyr groups (0Y) with their planar faces in a perpendicular geometry.

Figure 8

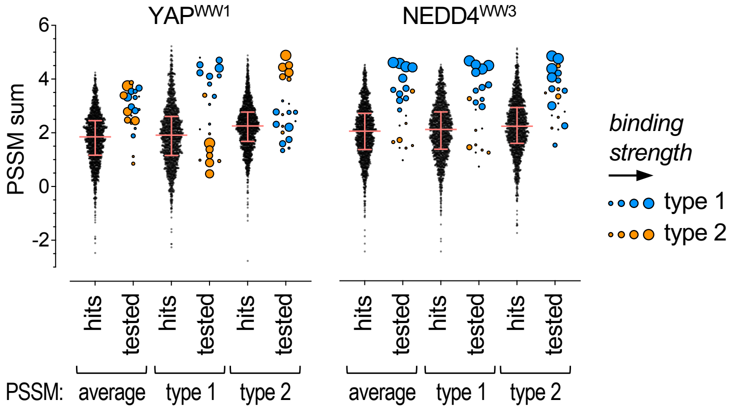
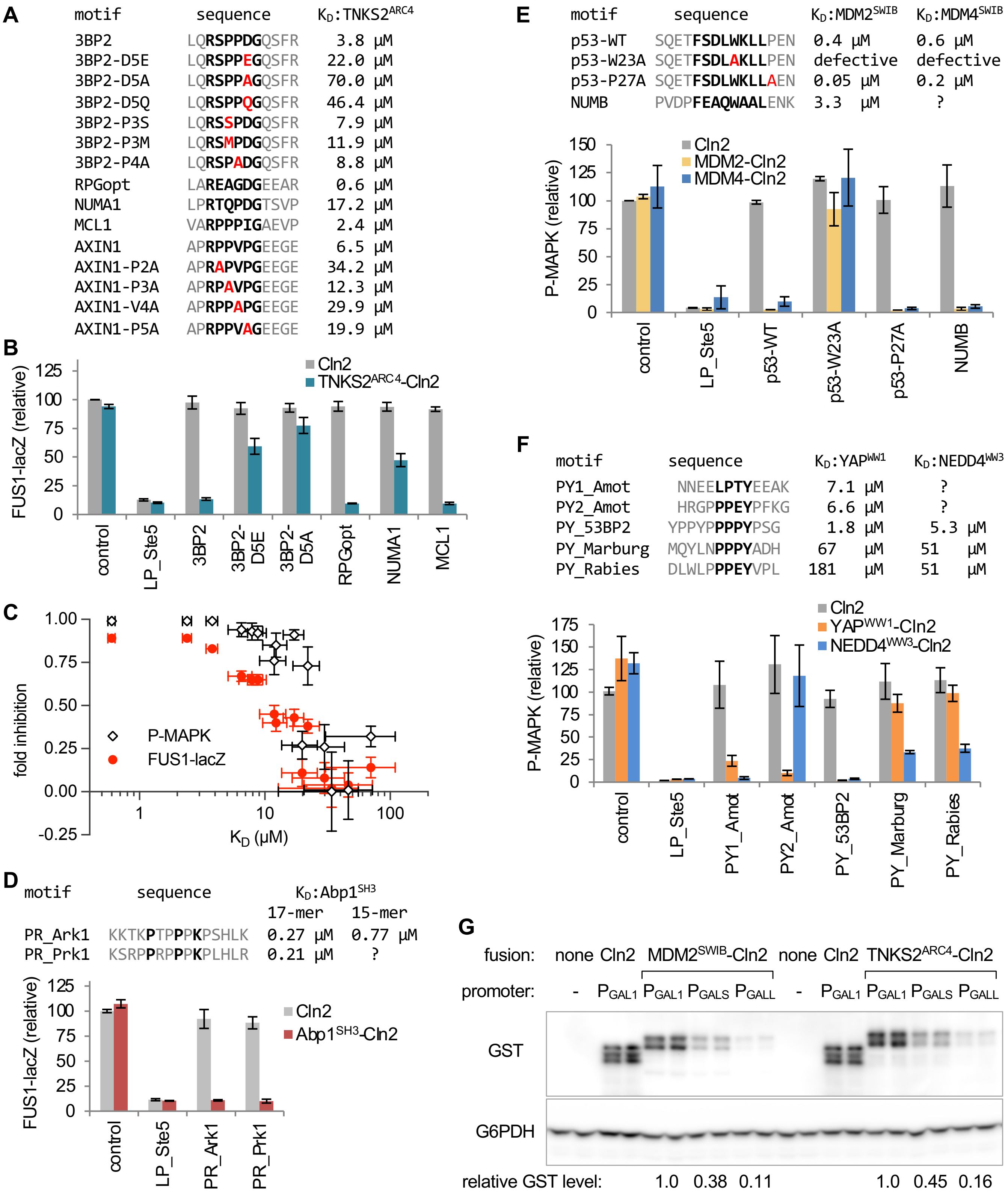


Figure 8. Stratification of [LP]PxY matches by PSSM scoring

All 1730 matches to the [LP]PxY consensus sequence ("hits") in disordered regions of the human proteome were scored using 6 different PSSMs. For each of 2 WW domains, 3 PSSMs were derived from the DMS results: one from all motifs (average), one from type 1 motifs only, and one from type 2 motifs only. The PSSM values for individual residues were summed across 10 motif positions (xx[LP]PxYxxxx) to calculate a predicted score (PSSM sum) for all sequences. For all hits, the distribution of sums obtained using each PSSM is shown; pink lines denote median and quartile values. For all motifs tested by SIMBA ("tested"), PSSM sums were calculated similarly, and their symbol sizes are proportional to their observed SIMBA scores (see Figure 6B).

Figure S1



A

motif	sequence	K_D:TNKS2^{ARC4}
3BP2	LQ**RSPPDG**QSFR	3.8 µM
3BP2-D5E	LQ**RSPPEG**QSFR	22.0 µM
3BP2-D5A	LQ**RSPPAG**QSFR	70.0 µM
3BP2-D5Q	LQ**RSPPQG**QSFR	46.4 µM
3BP2-P3S	LQ**RSSPDG**QSFR	7.9 µM
3BP2-P3M	LQ**RSMPDG**QSFR	11.9 µM
3BP2-P4A	LQ**RSPADG**QSFR	8.8 µM
RPGopt	LA**REAGDG**EEAR	0.6 µM
NUMA1	LP**RTQPDG**TSVP	17.2 µM
MCL1	VA**RPPPIG**AEVP	2.4 µM
AXIN1	AP**RPPVPG**EEGE	6.5 µM
AXIN1-P2A	AP**RAPVPG**EEGE	34.2 µM
AXIN1-P3A	AP**RPAVPG**EEGE	12.3 µM
AXIN1-V4A	AP**RPPAPG**EEGE	29.9 µM
AXIN1-P5A	AP**RPPVAG**EEGE	19.9 µM

D

motif	sequence	K_D:Abp1^{SH3}	
		17-mer	15-mer
PR_Ark1	KKTK**P**TP**P**P**K**PSHLK	0.27 µM	0.77 µM
PR_Prk1	KSRP**P**RP**P**P**K**PLHLR	0.21 µM	?

E

motif	sequence	K_D:MDM2^{SWIB}	K_D:MDM4^{SWIB}
p53-WT	SQET**FSDLWKLL**PEN	0.4 µM	0.6 µM
p53-W23A	SQET**FSDLAKLL**PEN	defective	defective
p53-P27A	SQET**FSDLWKLL**AEN	0.05 µM	0.2 µM
NUMB	PVDP**FEAQWAAL**ENK	3.3 µM	?

F

motif	sequence	K_D:YAP^{WW1}	K_D:NEDD4^{WW3}
PY1_Amot	NNEE**LPTY**EEAK	7.1 µM	?
PY2_Amot	HRGP**PPEY**PFKG	6.6 µM	?
PY_53BP2	YPPYP**PPPY**PSG	1.8 µM	5.3 µM
PY_Marburg	MQYLN**PPPY**ADH	67 µM	51 µM
PY_Rabies	DLWLP**PPEY**VPL	181 µM	51 µM

Figure S1. Low-throughput assays of SLiM binding to six domains

(**A**) List of tested TNKS2^{ARC4}-binding peptides. K_D values are from [54]. (**B**) The experiment shown in Figure 2B was repeated using the transcriptional reporter, FUS1-lacZ (mean ± SD; n = 4). (**C**) Comparison of the fold inhibition in FUS1-lacZ vs. P-MAPK assays, for all peptides in panel A. Both assays show affinity-dependent inhibition, but the P-MAPK assay detects weak binding with greater sensitivity. Fold inhibition (mean ± SD, n = 4) is the difference in signal between the test peptide and the control, as a fraction of the control signal. K_D values are mean ± SE [54]. (**D**) SLiM recognition by the Abp1^{SH3} domain, assayed using the FUS1-lacZ reporter. Bars, mean ± SD (n = 3). K_D values from [53]. (**E**) SLiM recognition by MDM2^{SWIB} and MDM4^{SWIB} domains, assayed as in Figure 2B. Bars, mean ± range (n = 2). K_D values from [33, 65]. (**F**) SLiM recognition by YAP^{WW1} and NEDD4^{WW3} domains, assayed as in Figure 2B. Bars, mean ± range (n = 2). Note that K_D values from distinct studies [55, 56] might not be directly comparable. (**G**) Protein levels of GST-tagged Cln2 fusions expressed from promoters of different strengths (P_{GAL1}, P_{GALS}, P_{GALL}). Expression was induced with galactose for 105 min. G6PDH served as a loading control.

Figure S2

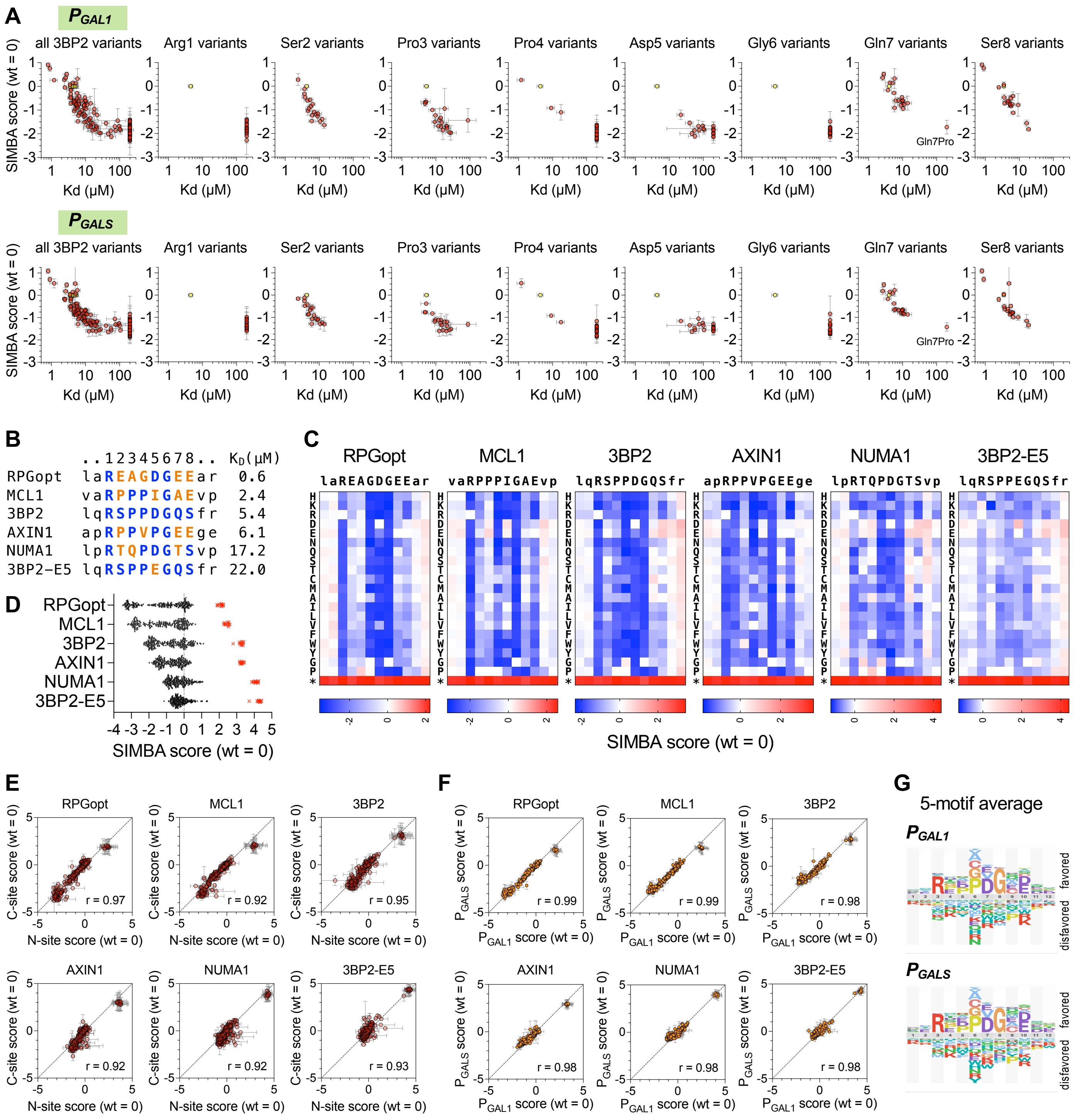


Figure S2. SIMBA interrogation of multiple TNKS2^{ARC4}-binding peptides

(**A**) SIMBA scores vs. K_D for all 3BP2 variants as well as variants at each of 8 individual positions, plotted as in Figure 3C, with the TNKS2^{ARC4}-Cln2 fusion driven by P_{GAL1} or P_{GALS}. All SIMBA scores are mean ± SEM (n = 4). (**B**) Sequence and K_D of each motif subjected to DMS and SIMBA. Residues in the 8 core positions are colored where identical to (blue) or different from (orange) the 3BP2 sequence. (**C**) Heatmaps of DMS results from six parent peptides, plotted as in Figure 3A. (**D**) Distribution of SIMBA scores for all missense variants (black circles) and the average STOP codon (red X symbols) at the 12 positions in each of the 6 parent motifs. (**E**) Correlation of SIMBA scores for variants inserted a the N-site vs. C-site locations, plotted as in Figure 3G, for six parent peptides. (**F**) Correlation of SIMBA scores when the TNKS2^{ARC4}-Cln2 fusion was expressed from P_{GAL1} vs. P_{GALS}, plotted as in Figure 3F, for six parent peptides. (**G**) Sequence preference logos, averaged from 5 motifs (excluding 3BP2-E5), comparing when the TNKS2^{ARC4}-Cln2 fusion was expressed from P_{GAL1} vs. P_{GALS}.

Figure S3

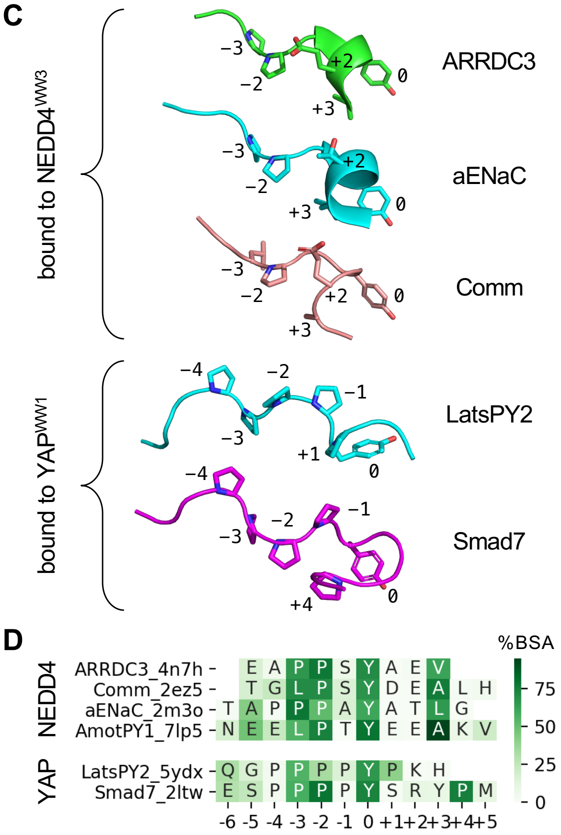
A

Peptide Category	Motif	Sequence *	Reported K_D (μM) †		Refs. ‡
			YAP^{WW1}	NEDD4^{WW3}	
(i) natural	AmotPY1	NNEELPTYEEAK	7.1	NR	1
	AmotPY1b	NEELPTYEEAKV	7.1	NR	1
	AmotPY2	HRGPPPEYPFKG	6.6	NR	1
	Smad7	LESPPPPYSRYP	6.9	8.0	2
	Lats1PY1	NRQPPPPYPLTA	15	NR	3
	Lats1PY2	YQGPPPPYPKHL	4.2	NR	3
	aENaC	LTAPPPAYATLG	NR	10	4
	bENaC	PGTPPPNYDSLR	NR	14	4
	Comm	IATGLPSYDEAL	NR	3.1	5
	ARRDC3	RPEAPPSYAEVV	NR	3.3	6
	53BP2	PPYPPPPYPSGE	1.8	5.3	7
	Marburg	QYLNPPPYADHG	67	51	7
	Rabies	LWLPPEYVPLK	181	51	7
	Ebola	LPTAPPEYMEAI	750	76–147	7,8
	HTLV1	SDPQIPPPYVEP	61	308	7
(ii) optimized	UGR2	LDSLPSYSELFP	NR	0.15	8
	UGR1	LDAPPSYSELFP	NR	0.15	8
	UGR1–LYG	LDAPPSYSELYG	NR	0.92	8
(iii) consensus	PYcon1	AVPPNYSELFAT	NR	NR	
	PYcon2	TLPPPYSSYCMT	NR	NR	
	PYcon3	AVLPDYVTAMWQ	NR	NR	
	PYcon4	SKLPSYWFWRAN	NR	NR	

* Sequences tested in this study; some sequences tested in previous studies differ in length or boundaries.

† Values from different studies and methods may not be directly comparable. NR, not reported.

‡ References – PMIDs: (1)23564455; (2)22921829; (3)29487715; (4)12654927; (5)16531238; (6)24379409; (7)31636332; (8)35257734



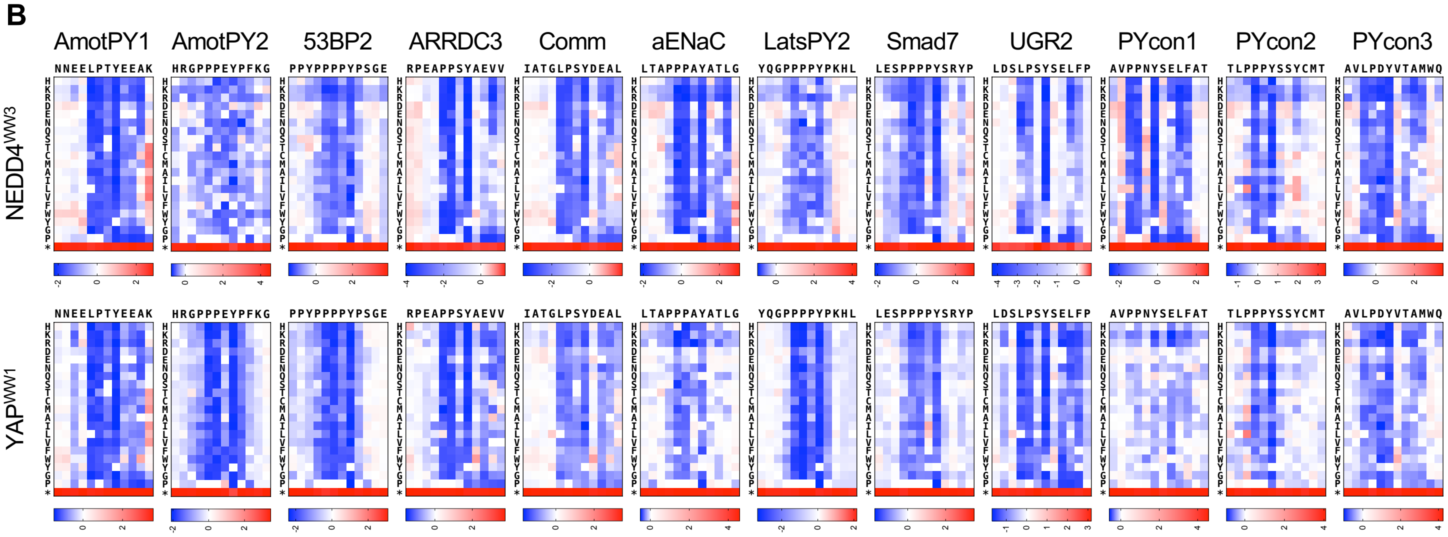


Figure S3. DMS data and characteristics of WW domain-binding peptides

(**A**) WW domain peptides used in this study, and their reported binding affinities. The consensus peptides (PYcon1-4) were designed based on "multiple PWM" sequences from Figure S1B of Ref [74], which were proposed as improved predictors of binding specificity compared to single PWMs. (**B**) Heatmaps of DMS results for 12 peptides combined with 2 WW domains. Data are the mean of 3 independent replicate experiments. (**C**) Conformations of WW domain bound peptides in published structures. The orientation is comparable to that in Figure 6G. PDB IDs: 4n7h, 2m3o, 2ez5, 5ydx, 2ltw. (**D**) Plot of buried surface area (BSA) for residues in WW-bound peptides, expressed as a percentage of accessible surface area in the unbound peptide. Data are from 4 NEDD4^{WW3} complexes (all type 1) and 2 YAP^{WW1} complexes (type 2). Peptide names include PDB IDs.