A proteome-wide Domain-centric Perspective on Protein Phosphorylation

The domain context of phosphorylation

Antonio Palmeri, Gabriele Ausiello, Fabrizio Ferrè, Manuela Helmer-Citterich and Pier Federico Gherardini#*

Antonio Palmeri (antonio.palmeri@uniroma2.it)

Gabriele Ausiello (gabriele.ausiello@uniroma2.it)

Fabrizio Ferrè (ferre@uniroma1.it)

Manuela Helmer-Citterich (citterich@uniroma2.it)

Pier Federico Gherardini (pier.federico.gherardini@uniroma2.it)

Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica snc, 00133 Rome, Italy.

present address: Baxter Laboratory for Stem Cell Biology, Department of Microbiology & Immunology, Stanford University School of Medicine, Stanford, California, USA.

*: to whom correspondence should be addressed, citterich@uniroma2.it

Summary

Phosphorylation is a widespread post-translational modification that modulates the function of a large number of proteins. Here we show that a significant proportion of all the domains in the human proteome are significantly enriched or depleted in phosphorylation events. A substantial improvement in phosphosites prediction is achieved by leveraging this observation, which has not been tapped by existing methods.

Phosphorylation sites are often not shared between multiple occurrences of the same domain in the proteome, even when the phosphoacceptor residue is conserved. This is partly due to different functional constraints acting on the same domain in different protein contexts. Moreover, by augmenting domain alignments with structural information, we were able to provide direct evidence that phosphosites in protein-protein interfaces need not be positionally conserved, likely because they can modulate interactions simply by sitting in the same general surface area.

Introduction

Phosphorylation, the most widespread protein post-translational modification, is an important regulator of protein function. The addition of phosphate groups on serine, threonine and tyrosine residues can modulate the activity of the target protein by inducing complex conformational changes, by modifying protein electrostatics, and by regulating domain-peptide interactions, as in 14-3-3 or SH2 domains, that specifically recognize phosphorylated residues. The standard experimental technique for the high-throughput identification of phosphorylation sites is mass spectrometry(1).

Phosphorylation is catalyzed by protein kinases, a family that in humans comprises ~540 members(2, 3). It is well understood that these enzymes recognize specific sequence motifs in their substrates (4, 5). Accordingly the sequence around the phosphorylation site is undisputedly the most important feature for phosphosite prediction (6, 7). However the "context", in a broad sense, where these motifs occur is also important as sequence alone is not enough to achieve the observed specificity of phosphorylation. Therefore, several

studies have characterized multiple aspects of phosphosites such as their preference for loops and disordered regions (reviewed in (8)), or the tendency of phosphoserines and phosphothreonines to occur in clusters (9), and these features have been used to improve the performance of phosphosite predictors(6, 7, 10-12). Moreover placing kinases and substrates in the context of protein interaction networks has been shown to improve the prediction of phosphorylation by specific kinases (13).

Perhaps one of the most puzzling observations when looking at the phosphoproteome as a whole, is the fact that a large proportion of phosphorylation sites is poorly conserved. This has led to various hypotheses. First some sites may represent non-functional, possibly low-stoichiometry, phosphorylation events that are picked up because of the sensitivity of mass-spectrometry (14, 15). Indeed functionally characterized sites and those matching known kinase motifs are more conserved on average(15-17). However, although in biology function often equates with conservation, there could be genuinely functional fast-evolving phosphosites, that are responsible for species-specific differences in signaling and regulation. Moreover in some cases, especially in the regulation of protein-protein interactions, the exact position of the phosphosites may be unimportant (18, 19).

Here we explore the issues of "context" and "conservation" of phosphorylation sites from the perspective of protein domains. To this end we assembled a comprehensive database of phosphosites from publicly available sources and studied their proteome distribution with respect to the location and identity of protein domains. We focus on the human phosphoproteome because it has been very well characterized in a multitude of low- and high-throughput experiments, thus providing the opportunity for a comprehensive, proteome-wide, study. In particular the issues we want to address are the following:

1) Are specific domain types preferentially phosphorylated? Or conversely are some domains specifically depleted of phosphorylation sites?
2) Can the domain context be used to improve the prediction of phosphorylation sites?
3) What is the conservation pattern of phosphosites when looking at multiple instances of the same domain in the proteome?

Results and discussion

Dataset Composition

We collected 65239 human phosphorylation sites from the Phospho.ELM(20), PhosphositePlus(21), UniProt(22), and PHOSIDA(7) databases (see Methods). We then identified PFAM(23) domains in all the human proteome and investigated the distribution of phosphorylation sites with respect to these protein domains. 6710 sites were either located in proteins with no PFAM domain, or they were assigned to multiple domains because of overlaps in the domain definitions. These sites were discarded and not considered further. We expect extracellular phosphorylation sites to be both more rare, and under-represented in databases due to experimental setup. In order to eliminate this bias we constructed an "intra-cellular proteome" by predicting the cellular localization of all the proteins in our dataset (see Methods). Without this step, domains which are predominantly extracellular would appear as depleted in phosphorylation. Our final dataset comprises 48252 phosphorylation sites, of which 29837 are serines, 9479 threonines and 8936 tyrosines. These sites map to 6880 proteins (~56% of the predicted human intracellular proteome). The average number of phosphosites per protein is 7 and almost 19% of phosphoproteins contain more than 10 phosphoresidues (Supplementary Figure 1). 12% of the sites were identified in low-throughput experiments, while the remainder was derived from high-throughput datasets.

The majority of phosphorylatable residues (i.e. Ser/Thr/Tyr) are in N-/C-terminal regions, accounting for 53% of the total, while 26% are in domains and 21% in Inter-Domain Regions (IDRs). However both IDRs and N-/C-terminal regions have the highest phosphorylation density (5.6%)—i.e. the proportion of phosphorylatable residues that are actually phosphorylated— compared to domains (3.6%).

Interestingly sites from high-throughput experiments are preferentially located outside protein domains (76% vs 64% of low throughput sites, Chi-square test $p < 2.2e-16$). Protein regions outside globular domains are more exposed to solvent and therefore more likely to be recognized by kinases. Recently a number of

authors have suggested that a proportion of phosphorylation sites may result from random encounters between kinase and substrate and have no functional meaning(15), representing only the "noise" in the system. In general we can assume that sites from low-throughput experiments are more likely to have a functional meaning, since they were derived from studies investigating single sites of interest. The observed enrichment would therefore lend support to this hypothesis as mass-spectrometry could be picking up low-stoichiometry non-functional phosphorylation events, which are more likely to happen in highly accessible regions of the protein (i.e. outside globular domains).

The domain context of protein phosphorylation

In order to explore the domain-context of phosphorylation we investigated whether specific domain types are significantly enriched/depleted in phosphorylation, i.e. whether phosphorylation specifically modulates certain domains. We estimated the average propensity of each residue type (Ser/Thr/Tyr) to be phosphorylated by pooling all the domain types together and calculating the ratio of phosphorylated residues to the total number of phosphoacceptor residues in the proteome. If a specific domain is not more/less frequently phosphorylated than the average domain, we expect this ratio, when calculated for a single domain type, to be similar to the value obtained by pooling all the domains together. Conversely domains enriched/depleted in phosphorylation will display a higher/lower propensity. The significances of these differences in propensity were evaluated with a statistical test (see Methods).

Following this analysis we obtained, for each residue type, a list of domains significantly enriched or depleted in phosphorylation. 151 domains were significantly enriched in pSer phosphorylation and 33 were depleted (see Table 1a,b); 55 were enriched in pThr and 11 depleted(see Table 2a,b). Finally, for pTyr, we found 39 domains enriched and 8 depleted (see Tables 3a,b). We observed that the significantly enriched domain types represent 6% of the 3131 domain types and the significantly depleted domain types are 1.1%. If we consider the total number of domain instances in the proteome (i.e. accounting for multiple copies of the same domain), the significantly enriched and depleted domains respectively represent 12% and 17% of the total. This difference is mainly due to pS and pT, as shown in table 4.

The different types of kinase domains represent a specific case that deserves to be discussed separately. These proteins often participate in signaling pathways where phosphorylation is used by upstream kinases to regulate the activity of downstream ones. Therefore this protein family is both responsible for phosphorylation, and finely modulated by it. Furthermore there is a clear tendency for these domains to be phosphorylated on the same type of residues for which they catalyze the reaction, especially for the Tyr-Kinase domain. Indeed, the Protein Kinase domain (which includes mostly Ser/Thr kinases) shows a significant enrichment for pSer and pThr, while it is not enriched for pTyr. Similarly the Protein Tyrosine Kinase domain is significantly enriched for pTyr, but not for pSer and pThr.

The domain showing the highest mean propensity across all the phospho-modifications is the Paxillin Family domain. This domain is found in adaptor proteins and the phosphosites act as docking sites for other proteins(24). Some of the other interesting domains which have been described as highly modulated by phosphorylation are Core histone H2A/H2B/H3/H4, which reflects the role of phosphorylation in regulating the cellular response to DNA damage, histone turnover and chromatin architecture and oncogenesis(25, 26). The Ubiquitin family domain is also massively targeted by phosphorylation on all three residue types. Phosphorylation has been reported to influence the degradation of proteins, preventing it via ubiquitination(27-30). Another evidence of the cross-talk between ubiquitination and phosphorylation, is represented by *phosphodegrons*—phosphorylation sites recognized by ubiquitin ligases. They serve as markers for the destruction of inhibitors of cyclin-dependent kinases at the initiation of DNA replication (31-37).

In order to evaluate the actual number of different protein families in which a domain appears we calculated the number of paralogy groups (i.e. as opposed to the actual number of proteins) having a specific domain. As shown by the size of the red bubbles in figures 1.a-c, many depleted domains are widespread in the proteome and occur in a large number of different families. Conversely, as one moves to regions of higher propensity, the domains are more restricted to specific protein families (small bubbles). A notable exception is the kinase domain. The y-axis represents the fraction of domain instances that are phosphorylated. Interestingly this figure is very variable, even for domains with comparable propensities.

As stated above, our dataset only contains proteins predicted to be intracellular. Accordingly we do not find among the depleted domains those occurring predominantly in extracellular proteins or in the extracellular

portion of membrane proteins (e.g. Cadherin, Fibronectin type III, EGF-like, Immunoglobulin). On one hand extracellular proteins are less phosphorylated in general. On the other hand, depending on the experimental setup, these sites might be completely left out of the analysis. For instance if the data have been collected in cell cultures and the medium is discarded, then obviously secreted proteins will not appear in the data. Recently a number of works have investigated the phosphoproteome of several body fluids(38-40), but certainly these sites have received much less attention than the ones in intracellular proteins.

Inter Domain Regions significantly enriched/depleted in phosphorylation

80% of pSer, 69% of pThr and 54% of pTyr map outside protein domains (similar figures were reported in (21)). For all three residue types the phosphosites are preferentially located in Inter Domain Regions (IDRs) compared with non-modified residues of the same type (all Fisher's exact tests p-values < 2.2e-16). The increase in preference is more evident for pSer followed by pThr and pTyr (data not shown).

In order to include these sites in the analysis we analyzed the distribution of phosphorylations with respect to the identity of the two domains flanking the Inter Domain Region (IDR). Thus, similarly to what we did for sites located in domains, we determined whether each IDR is enriched or depleted in phosphorylation. In defining the IDR we did not take into account the ordering of the two domains, as this would excessively reduce the cardinality of each case. As we did with domains, we excluded from the analysis extracellular IDRs. There are 179 IDRs enriched in pSer and 76 depleted, while for pThr 59 are enriched and 11 are depleted. For pTyr there are the 39 enriched IDRs and only 5 depleted, consistent with the observation that pTyr is less often found in IDRs, compared with pSer/pThr (tables 5-7).

Interestingly almost all pTyr-enriched IDRs involve at least one protein-protein interaction domain (SH2, SH3, WW, PDZ, PX, etc.). These are very likely to be high-density regions where multiple signals are integrated.

11 IDRs are simultaneously enriched for pSer, pThr and pTyr. The IDR with the highest phosphopropensity for all phospho-modifications is flanked by the domains DNA gyrase/topoisomerase IV, subunit A and DTHCT (NUC029) region.

Figure 2 shows the propensity of significantly enriched/depleted IDRs, together with the propensities of the flanking domains (for clarity only IDRs with at least 10 occurrences are shown, moreover the figure does not include regions between a domain and the N/C-term of the protein).

There are a number of cases where the propensity of the IDR is different from that of the flanking domains. This is represented by small blue dots, indicating low propensity of the flanking domain, and high propensity of the IDR. For instance, even though the SH3 domain has a low domain propensity, the IDRs that are combinations of SH3 with Spectrin and with SH2, have a very high IDR propensity for pSer and pTyr respectively, thus suggesting that IDRs flanking the SH3 domain are highly modulated by phosphorylation.

Using domain information for the improvement of phosphosite prediction

Following the observation that phosphorylation is influenced by the domain context, we next tested whether this information could be used to improve the prediction of phosphosites. The rationale for this is that it would seem desirable to assign a higher score to sites predicted in a domain that is enriched in phosphorylation and converseley reduce the score of those predicted in a depleted domain.

We used a machine-learning approach based on Support Vector Machines (SVM) to build three predictors, one each for pSer, pThr and pTyr.

For each residue type we built two predictors, one including only the sequence around the phosphosite (in standard orthogonal binary encoding), and the other including also the phosphorylation propensity of the domain or IDR. For Ser, the predictor with all the features obtained an AUC of 0.72, 2% higher than the sequence-only predictor (see Table 8). For Tyr the inclusion of the domain features affords an improvement of 7%, reaching an AUC of 0.66. For Thr we observe an improvement of 4%, reaching 0.72. It must be noted

that we do not include in any way the information on the identity of the domain, as we only gave the propensity in input to the SVM.

Clearly more elaborate encoding schemes are possible, based for instance on the domain signature of the protein(41). However such an encoding would bias more and more the predictor towards recognizing specific families of proteins (defined by their domain signatures), thus providing an unfair advantage. Moreover our encoding is general enough to be applied to any protein irrespective of whether its domain composition is unique, or any domain is present at all.

The reason for this improvement lies in the fact that the two sources of information – the sequence of the peptide and the domain propensity – are completely independent yet they are both related to the probability of a site being phosphorylated.

We want to stress that the dataset does not include extracellular proteins, as we filtered out predicted extracellular domains. Therefore the improvement afforded by the domain information is not trivially due to the fact that the predictor is down-scoring extracellular proteins.

Conservation of phosphorylation sites in different instances of the same domain

Different domains vary considerably in the conservation of their phosphorylation sites. Both for pSer/Thr and pTyr a number of domains have a very small number of highly conserved phosphorylation sites. The fact that several of these domains have extremely low propensities means that, even though the alignment column is conserved, a small number of residues is actually phosphorylated (at least in the conditions tested in the experiments from which our dataset is derived). Therefore these residues represent either cases where the phosphorylation has a functional effect that is specific to a limited number of the proteins containing the domain, or possibly non-functional phosphorylation events.

We analyzed the proportion of phosphorylatable residues that are actually phosphorylated in the alignment columns containing at least one phosphosite and at least ten phosphorylatable residues. Interestingly, for a large number of columns, 77% for Ser, 81% for Thr, 67% for Tyr, this proportion is less than 10%.

Undoubtedly this is partly due to the fact that, by aligning all the copies of a given domain in the human proteome, we are comparing domain instances that are located in proteins with different functions and different regulation. We therefore repeated the analysis by grouping together domains contained in proteins belonging to the same family (see Methods). The proportion of sites with a ratio of phosphorylated/phosphorylatable less than 10% decreases, reaching 48% for Ser, 56% for Thr and 27% for Tyr. However these figures still represent a serious *caveat* against the practice of inferring the phosphorylation of a site on the basis of the observation that the same site is phosphorylated in another domain of the same family. Moreover these results indicate that, inside protein domains, Tyr phosphorylation is more conserved than Ser/Thr.

Phosphorylation and protein-protein interfaces

A number of reports have shown that phosphorylation sites are often not conserved in position, although sometimes different phosphorylation sites are clustered in the same region of the alignment(42). In these cases the exact position of the phosphorylation site may not be important as long as the same region of the protein is phosphorylated. It has been proposed that this phenomenon preferentially occurs at protein-protein interfaces, where phosphorylation of any residue in a given surface region may regulate the formation of the complex (19). Accordingly Tan et al.(18) observed that proteins displaying this pattern of phosphosites conservation are enriched in protein- and DNA-binding annotations and frequently interact with other proteins. We therefore set out to verify this hypothesis with our dataset. We mapped all the phosphosites of a domain on a reference domain structure and clustered the sites according to the geodesic distance between the residues on the surface of the protein (i.e. the distance "walking" along the surface and not "cutting" through it).

Each cluster therefore represents a set of phosphosite positions in the domain that are located in the same surface region, although not necessarily close in sequence.

We next calculated for each cluster of phosphosites the average conservation and the proportion of phosphorylation sites that are located in a protein-protein interface. Interestingly, we found a negative correlation between these two variables so that clusters of phosphosites localized in interface regions tend to be less conserved (Kendall's correlation test $p < 9.4e-9$, Wilcoxon test between the two extreme bins in figure 3 $p < 9.3e-8$).

This observation provides direct and independent evidence in support of the hypothesis that clusters of non-positionally conserved phosphosites modulate protein-protein interactions. Figure 4 shows four examples of surface clusters of poorly conserved phosphoresidues that have a good overlap with protein-protein interface regions. A visual inspection of the alignments shows how distant in sequence phosphosites belonging to the same surface cluster can be, which clearly precludes the identification of these cases by sequence analysis only.

We briefly discuss four examples of phosphorylation sites that are not positionally conserved, but cluster together on the domain structure and are also located in protein-protein interaction interfaces. The first example involves the Variant SH3 domain, a signaling module involved in domain-ligand interactions. We mapped the phosphosites clusters to the Variant SH3 domain of the protein DOCK2 in a structure that describes its interaction with ELMO1(43). Figure 4A shows the remarkable overlap between the phospho-cluster in orange and the surface region of DOCK2 that interacts with the C-terminal Proline-rich region of ELMO1.

The family of apolipoprotein B messenger RNA-editing enzyme catalytic (APOBEC) proteins deaminates mRNA and single-stranded DNA(44)(see Figure 4B). Ser38 of Activation-induced cytidine deaminase (Q9GZX7)(45) and Ser47/72 of APOBEC-1 (P41238) (46) have been characterized as modulators of the enzymatic activity of the respective proteins. The interface shown in the picture is from the structure of APOBEC-2(47), for which no phosphosites are present in our dataset. The structure is a homotetramer and many other APOBEC enzymes have been reported to form multimers. Interestingly this raises the possibility that the phosphosites in this surface cluster may modulate the activity of these proteins by affecting their oligomerization, even though they are not positionally conserved.

Phosphorylation of the SH2 domain can tune its affinity for phosphotyrosine substrates, and can also affect the localization of SH2-containing proteins(48-52). Figure 4C shows different phospho-clusters mapped on the surface of Grb2 in a homodimeric complex.

The last example (figure 4D) involves the RNA recognition motif domain here mapped on the structure of the U1 small nuclear ribonucleoprotein A in complex with the *E. coli* ThiM riboswitch(53). The different phospho-clusters map to distinct interaction surfaces and they may modulate the affinity of the protein for RNA as well as other proteins.

Conclusions

In this work we provide a proteome-wide assessment of the relationship between protein domains and phosphorylation in the human intracellular proteome. Our results show that 7% of the domain types in the proteome are significantly enriched or depleted in phosphorylation. Interestingly we found that a number of these domains, such as Ankyrin repeats, zinc fingers and WD, constitute a significant fraction of all the domain instances in the human proteome.

We showed that the information about the domain composition of a protein and the specific domain or IDR in which a putative phosphosite is located can be used to improve the prediction of phosphorylation sites. This information was coded as a *propensity* value defined as the proportion of domains or IDRs of each type that are phosphorylated in the training set. We achieved a 2%, 4% and 7% improvement in the prediction of pSer, pThr and pTyr respectively, when compared with a predictor using sequence information only. This improvement is comparable to those reported in other studies including features such as conservation, secondary structure, disorder and local amino acid composition (6, 7, 10, 11). Importantly, the domain propensity value we use represents orthogonal information, while features such as disorder and secondary structure are already quite effectively captured in the sequence data. Our method does not explicitly encode the domain composition of the protein, which would bias the predictor too much towards the recognition of known examples, and is general enough to be applied to any protein.

We also used our dataset of domain alignments to study the conservation of phosphorylation sites. There are conflicting reports in the literature about this issue with some authors reporting phosphorylation sites as more conserved(37, 54), or not (15, 17, 55) than corresponding non-modified residues. The issue is undoubtedly confounded by the different criteria used to score conservation and also by the over-representation of phosphorylation sites in disordered regions. However when the analysis is restricted to sites which are likely to be functional then a conservation signal definitely emerges (15-17). These considerations notwithstanding, the possibility that non-conserved sites represent species-specific differences in regulation must not be ruled out. Whatever the answer to this question the inclusion of conservation provides only a very modest increment in phosphorylation site prediction(7). This could be explained by the fact that, in testing over a complete dataset, one is also trying to predict non-conserved, possibly non-functional phosphorylation sites.

By augmenting domain alignments with structural information we were able to provide a novel and direct evidence to the notion that phosphorylation sites that regulate protein-protein interfaces need not be positionally conserved (18, 19). This mechanism can explain a portion, though obviously not all, of the observed "non-conservation". Moreover we observe that sites from high-throughput experiments are more likely to be located outside protein domains. We can use as proxy for functionality the fact that a phosphosite has been identified in a low-throughput study, as these sites are extremely likely to be functional, while the others may not be. The regions outside domains are more solvent accessible and therefore more likely to be recognized by protein kinases possibly resulting in a higher-proportion of non-functional phosphorylation events.

In terms of conservation of the phosphorylation event (i.e. as opposed to simply the phosphoacceptor residue), even after grouping together paralogous proteins, 48% of pSer, 56% for pThr- and 27% of pTyr-containing domain alignment columns are phosphorylated on less than 10% of the phosphorylatable residues. Even though any dataset is necessarily incomplete, this observation should elicit caution when using sequence conservation to transfer phosphosites between different proteins. This is especially true in light of

the fact that these figures refer to alignments of domains, that are more likely to be correct than those of unstructured regions.

In conclusion, our work offers a new perspective on proteome-wide studies of phosphorylation. By studying the distribution of phosphorylation sites with respect to protein domains we were able to derive an informative measure for phosphosite prediction that is independent from other features commonly used for this task. Finally we showed that phosphosites in protein-protein interfaces need not be positionally conserved and shed new light on a number of other issues pertaining to their general characteristics.

Methods

We collected human phosphorylation sites from the following databases: Phospho.ELM, PhosphositePlus, UniProt, PHOSIDA. All the phosphorylation sites were mapped on UniProt sequences, checking for the identity of a 10-residue window centered on the phosphosite. Phosphosites on different isoforms were mapped on the UniProt reference isoform using the program water from the EMBOSS package. HMMs for the identification of protein domains were downloaded from the PFAM database, selecting only the PFAM-A entries. The human proteome was scanned against this collection of HMMs using the pfam_scan.pl program.

Phosphorylation propensity of domains and Inter Domain Regions

We first estimated an average phosphorylation propensity by pooling all the domain types together and calculating the ratio of phosphorylated residues to the total number of phosphorylatable residues in the proteome. If a specific domain is not more/less phosphorylated than the average domain we expect this ratio, when calculated for a single domain type, to be similar to the value obtained by pooling all the domains together. The difference between these two proportions can be quantified with a Fisher test, i.e. by asking what would be the probability of obtaining the observed phospho/non-phospho domain counts for a specific

domain type if its probability of phosphorylation was equal to the overall phosphorylation propensity. The p-values were adjusted for multiple testing by controlling the False Discovery Rate. We considered a domain as significantly enriched/depleted in phosphorylation when the adjusted p-value was less than 0.05.

We performed a similar procedure for Inter Domain Regions (IDR) defined as the sequence regions lying between two domains of a given type, or a single domain and the N/C-term of the protein (irrespective of the ordering). Thus we obtained an overall propensity for IDRs that was compared with the propensity of each specific IDR in order to identify IDRs enriched/depleted in phosphorylation.

Extracellular Domains Filtering in Phosphorylation Dataset

Extracellular domains are expected to be depleted in phosphorylation and they were excluded from the dataset to avoid introducting biases..To this end, we first predicted signal peptides in the whole proteome using SignalP. Thereafter we predicted Transmembrane segments with TOPcons single, after removing the predicted signal peptides from the sequences. All the proteins having a signal peptide were discarded, while the other proteins containing TM sequences were considered for further filtering. We tested the reliability of these predictions using the set of proteins from SwissProt annotated with the GO-term "cell". 13253 proteins have the GO-term cell and are intracellular according to our procedure. Only 705 have the GO-term cell but are not correctly predicted. 5171 are predicted to be intracellular, e.g. do not possess a signal peptide or have a TM region, but are not annotated with the GO-term cell. Despite the good reliability of signal peptides and transmembrane segments predictions, these predictors are not perfect. Therefore we calculated an "intracellular propensity" of each domain/IDR, as the ratio between the number of residues predicted as intracellular and the total number of the domain/IDR residues, and we used this measure to discard non-intracellular domains/IDRs. The majority of domains score either 1 or 0 on this intracellular propensity, highlighting the sharp distinction between the two classes. We concluded that a threshold of 0.7 was not overly stringent, while at the same time allowing us to eliminate from the dataset almost all the domains annotated as extracellular.

Domains alignments and conservation scores

The sequence ranges corresponding to each domain were aligned on the Hidden Markov Model (HMM) describing the domain using HMMER 3.0(56). These alignments were then used to map the phosphorylation sites and interface residues (see below) derived from each sequence on a common domain reference. We used two different measures of conservation throughout the paper. The conservation of the alignment columns was calculated by taking a sequence as reference and then counting the percentage of residues in each column having a BLOSUM62 substitution score >= 1 with the residue in the reference sequence. In order to compare different alignments we normalized this conservation scores by calculating the empirical percentile of the conservation of each column with respect to the distribution of all the columns in the alignment. The percentile was then used as conservation score. To obtain a single value for each alignment column we calculated the average of all the conservation scores obtained when using each sequence in turn as reference. The conservation of the phosphorylation event was defined as the proportion of phosphorylatable residues that are actually phosphorylated in each column containing at least one phosphosite. Only columns with at least ten aligned sequences were considered in this analysis.

Paralogs identification

We used EnsemblCompara GeneTrees(57) to obtain the paralogy relationships between all the human genes. We considered both the homology relationships within_species_paralogs and other_paralogs. Therefore we clustered the proteins in paralogy groups, according to the relationships contained in the Gene Trees, and obtained 3203 paralogous protein clusters. The number of paralogy groups in which a domain appears provides an estimate of the number of different families (i.e. as opposed to single proteins) in which each domain is present.

Structure-based surface clustering of phosphorylation sites

In order to obtain a representative structure for each domain we used all the sequences in the alignment to perform a BLAST search against the Protein Data Bank. We then extracted the sequences from the matching structure files and identified the domain boundaries using pfam_scan.pl. The phosphorylation sites from each sequence were projected on all the sequences in the domain alignment. We selected as representative the structure that provided the highest coverage in terms of phosphosite positions and domain sequence.

In order to cluster the phosphorylation sites we calculated the geodesic distance between all the pairs of residues in the structure corresponding to a phosphosite-containing column of the domain alignment. We used UCSF Chimera(58) to calculate the molecular surface of the protein, described as a triangle mesh. In order not to assign buried phosphosites to the surface of the protein each site was associated with the closest surface vertex if its distance from it was less than 7.5 Angstroms. A visual inspection of a large number of cases showed that this procedure is effective in assigning phosphosites to surface vertices, while at the same time discarding buried sites. The geodesic distance is then defined as the shortest path in a graph where the nodes represent the vertices and the edges connect adjacent vertices and are weighted according to their distance in Angstroms. The shortest weighted path was calculated using an implementation of Dijkstra's algorithm(59). The resulting matrix of residue distances was clustered using affinity propagation(60, 61).

Protein Interfaces and Phosphorylation

The dataset of interface residues was derived by collecting all the pairs of different chains in the same PDB structure that could both be mapped on Uniprot. We only retained pairs of chains that had the same relative orientation in both the asymmetric and biological units. We consider two residues (one from each chain) as interacting if their distance is less than 0.5 Angstrom plus the sum of their Van Der Waals radii. Interfaces consisting of less than 5 residues on either chain were discarded. The interface residues were mapped on the corresponding domain alignment using the Uniprot accessions.

Phosphosite Predictor

We built predictors for pSer, pThr and for pTyr. All predictors are SVM-based classifiers. The training and testing procedures were written in R, using the R package LiblineaR.

The datasets for each predictor were derived as follows. We extracted a window of -5/+5 residues around the phosphorylation sites in our dataset to obtain the positive set. The negative set was derived by extracting the same -5/+5 window around all the phosphorylatable aminoacids in the proteome, after excluding known phosphosites. We used 90% of the positive set for training and the remainder for the testing. We used a 50% sequence identity threshold to reduce the redundancy between the training and test sets (both positives and negatives) and also within each of the two sets. We then resized the negative training and test sets in order to have an equal number of negatives and positives.

For each residue type we built two predictors, one including only the sequence around the phosphosite (in standard orthogonal binary encoding), and the other including the information related to the domain composition of the protein, simply encoded as the domain propensity.

Acknowledgements

We thank Prof. Gianni Cesareni for critically reading the manuscript, Alessio Colantoni for providing the dataset of protein-protein interfaces, Prof. Gianpaolo Scalia Tomba for help with the statistical analysis and Luca Parca for helpful discussion.

This work was supported by the EPIGEN flagship project and PRIN 2010 (prot. 20108XYHJS_006 to M.H.C.).

Authors' contributions

AP performed the computational work, analyzed the data and drafted the manuscript; GA, FF and MHC participated in the analysis of data and contributed to the manuscript; PFG conceived and coordinated the study, analyzed the data and wrote the manuscript. All authors read and approved the final manuscript.

Conflict of Interest

The authors declare that they have no competing interests.

References

1. Stirnimann, C. U., Petsalaki, E., Russell, R. B., and Müller, C. W. (2010) WD40 proteins propel cellular networks. *Trends Biochem. Sci.* 35, 565–574

2. Cohen, P. (2000) The regulation of protein function by multisite phosphorylation--a 25 year update. *Trends Biochem. Sci.* 25, 596–601

3. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science* 298, 1912–1934

4. Bridges, D., and Moorhead, G. B. G. (2004) 14-3-3 proteins: a number of functions for a numbered protein. *Science's STKE* 2004, re10

5. Miller, M. L., Jensen, L. J., Diella, F., Jørgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S. A., Bordeaux, J., Sicheritz-Ponten, T., Olhovsky, M., Pasculescu, A., Alexander, J., Knapp, S., Blom, N., Bork, P., Li, S., Cesareni, G., Pawson, T., Turk, B. E., Yaffe, M. B., Brunak, S., and Linding, R. (2008) Linear motif atlas for phosphorylation-dependent signaling. *Science Signaling* 1, ra2

6. Gao, J., Thelen, J. J., Dunker, A. K., and Xu, D. (2010) Musite: a tool for global prediction of general and kinase-specific phosphorylation sites. *Molecular & Cellular Proteomics*,

7. Gnad, F., Ren, S., Cox, J., Olsen, J. V., Macek, B., Oroshi, M., and Mann, M. (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* 8, R250

8. Via, A., Diella, F., Gibson, T. J., and Helmer-Citterich, M. (2011) From sequence to structural analysis in protein phosphorylation motifs. *Front. Biosci.* 16, 1261–1275

9. Schweiger, R., and Linial, M. (2010) Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. *Biol Direct* 5, 6

10. Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., and Dunker, A. K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* 32, 1037–1049

11. Palmeri, A., Gherardini, P. F., Tsigankov, P., Ausiello, G., Späth, G. F., Zilberstein, D., and Helmer-Citterich, M. (2011) PhosTryp: a phosphorylation site predictor specific for parasitic protozoa of the family trypanosomatidae. *BMC Genomics* 12, 614

12. Moses, A. M., Hériché, J.-K., and Durbin, R. (2007) Clustering of phosphorylation site recognition motifs

can be exploited to predict the targets of cyclin-dependent kinase. *Genome Biol.* 8, R23

13. Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A. T. M., Jørgensen, C., Miron, I. M., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J. G., Samson, L. D., Woodgett, J. R., Russell, R. B., Bork, P., Yaffe, M. B., and Pawson, T. (2007) Systematic discovery of in vivo phosphorylation networks. *Cell* 129, 1415–1426

14. Lienhard, G. E. (2008) Non-functional phosphorylations? *Trends Biochem. Sci.* 33, 351–352

15. Landry, C. R., Levy, E. D., and Michnick, S. W. (2009) Weak functional constraints on phosphoproteomes. *Trends Genet.* 25, 193–197

16. Ba, A. N. N., and Moses, A. M. (2010) Evolution of characterized phosphorylation sites in budding yeast. *Mol. Biol. Evol.* 27, 2027–2037

17. Beltrao, P., Albanèse, V., Kenner, L. R., Swaney, D. L., Burlingame, A., Villén, J., Lim, W. A., Fraser, J. S., Frydman, J., and Krogan, N. J. (2012) Systematic functional prioritization of protein posttranslational modifications. *Cell* 150, 413–425

18. Tan, C. S. H., Jørgensen, C., and Linding, R. (2010) Roles of "junk phosphorylation" in modulating biomolecular association of phosphorylated proteins? *Cell Cycle* 9, 1276–1280

19. Serber, Z., and Ferrell, J. E. (2007) Tuning bulk electrostatics to regulate protein function. *Cell* 128, 441–444

20. Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., and Diella, F. (2011) Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res.* 39, D261–7

21. Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* 40, D261–70

22. UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 40, D71–5

23. Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012) The Pfam protein families database. *Nucleic Acids Res.* 40, D290–301

24. Schaller, M. D. (2001) Paxillin: a focal adhesion-associated adaptor protein. *Oncogene* 20, 6459–6472

25. Musselman, C. A., Lalonde, M.-E., Côté, J., and Kutateladze, T. G. (2012) Perceiving the epigenetic landscape through histone readers. *Nat Struct Mol Biol* 19, 1218–1227

26. Singh, R. K., and Gunjan, A. (2011) Histone tyrosine phosphorylation comes of age. *Epigenetics* 6, 153–160

27. Ju, D., Xu, H., Wang, X., and Xie, Y. (2007) Ubiquitin-mediated degradation of Rpn4 is controlled by a phosphorylation-dependent ubiquitylation signal. *Biochim. Biophys. Acta* 1773, 1672–1680

28. Lin, H.-K., Wang, L., Hu, Y.-C., Altuwaijri, S., and Chang, C. (2002) Phosphorylation-dependent ubiquitylation and degradation of androgen receptor by Akt require Mdm2 E3 ligase. *EMBO J* 21, 4037–4048

29. Welcker, M., Orian, A., Jin, J., Grim, J. E., Grim, J. A., Harper, J. W., Eisenman, R. N., and Clurman, B. E. (2004) The Fbw7 tumor suppressor regulates glycogen synthase kinase 3 phosphorylation-dependent c-Myc protein degradation. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9085–9090

30. Yada, M., Hatakeyama, S., Kamura, T., Nishiyama, M., Tsunematsu, R., Imaki, H., Ishida, N., Okumura, F., Nakayama, K., and Nakayama, K. I. (2004) Phosphorylation-dependent degradation of c-Myc is mediated by the F-box protein Fbw7. *EMBO J* 23, 2116–2125

31. la Cova, de, C., and Greenwald, I. (2012) SEL-10/Fbw7-dependent negative feedback regulation of LIN-45/Braf signaling in C. elegans via a conserved phosphodegron. *Genes Dev.* 26, 2524–2535

32. Feldman, R. M., Correll, C. C., Kaplan, K. B., and Deshaies, R. J. (1997) A complex of Cdc4p, Skp1p, and Cdc53p/cullin catalyzes ubiquitination of the phosphorylated CDK inhibitor Sic1p. *Cell* 91, 221–230

33. Jackson, P. K. (2003) Ubiquitinating a phosphorylated Cdk inhibitor on the blades of the Cdc4 beta-propeller. *Cell* 112, 142–144

34. Lyons, N. A., Fonslow, B. R., Diedrich, J. K., Yates, J. R., and Morgan, D. O. (2013) Sequential primed kinases create a damage-responsive phosphodegron on Eco1. *Nat Struct Mol Biol* 20, 194–201

35. Rossi, M., Duan, S., Jeong, Y.-T., Horn, M., Saraf, A., Florens, L., Washburn, M. P., Antebi, A., and Pagano, M. (2013) Regulation of the CRL4(Cdt2) ubiquitin ligase and cell-cycle exit by the SCF(Fbxo11) ubiquitin ligase. *Mol. Cell* 49, 1159–1166

36. Skowyra, D., Craig, K. L., Tyers, M., Elledge, S. J., and Harper, J. W. (1997) F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. *Cell* 91, 209–219

37. Minguez, P., Parca, L., Diella, F., Mende, D. R., Kumar, R., Helmer-Citterich, M., Gavin, A.-C., van Noort, V., and Bork, P. (2012) Deciphering a global network of functionally associated post-translational modifications. *Mol Syst Biol* 8, 599

38. Bahl, J. M. C., Jensen, S. S., Larsen, M. R., and Heegaard, N. H. H. (2008) Characterization of the human cerebrospinal fluid phosphoproteome by titanium dioxide affinity chromatography and mass spectrometry. *Anal. Chem.* 80, 6308–6316

39. Carrascal, M., Gay, M., Ovelleiro, D., Casas, V., Gelpí, E., and Abian, J. (2010) Characterization of the human plasma phosphoproteome using linear ion trap mass spectrometry and multiple search engines. *J. Proteome Res.* 9, 876–884

40. Stone, M. D., Chen, X., McGowan, T., Bandhakavi, S., Cheng, B., Rhodus, N. L., and Griffin, T. J. (2011) Large-scale phosphoproteomics analysis of whole saliva reveals a distinct phosphorylation pattern. *J. Proteome Res.* 10, 1728–1736

41. Liu, Y., and Tozeren, A. (2010) Modular composition predicts kinase/substrate interactions. *BMC Bioinformatics* 11, 349

42. Tan, C. S. H., Bodenmiller, B., Pasculescu, A., Jovanovic, M., Hengartner, M. O., Jørgensen, C., Bader, G. D., Aebersold, R., Pawson, T., and Linding, R. (2009) Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Science Signaling* 2, ra39

43. Hanawa-Suetsugu, K., Kukimoto-Niino, M., Mishima-Tsumagari, C., Akasaka, R., Ohsawa, N., Sekine, S.-I., Ito, T., Tochio, N., Koshiba, S., Kigawa, T., Terada, T., Shirouzu, M., Nishikimi, A., Uruno, T., Katakai, T., Kinashi, T., Kohda, D., Fukui, Y., and Yokoyama, S. (2012) Structural basis for mutual relief of the Rac guanine nucleotide exchange factor DOCK2 and its partner ELMO1 from their autoinhibited forms.

Proc. Natl. Acad. Sci. U.S.A. 109, 3305–3310

44. Conticello, S. G., Thomas, C. J. F., Petersen-Mahrt, S. K., and Neuberger, M. S. (2005) Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Mol. Biol. Evol.* 22, 367–377

45. Pham, P., Smolka, M. B., Calabrese, P., Landolph, A., Zhang, K., Zhou, H., and Goodman, M. F. (2008) Impact of phosphorylation and phosphorylation-null mutants on the activity and deamination specificity of activation-induced cytidine deaminase. *J. Biol. Chem.* 283, 17428–17439

46. Chen, Z., Eggerman, T. L., and Patterson, A. P. (2001) Phosphorylation is a regulatory mechanism in apolipoprotein B mRNA editing. *Biochem. J.* 357, 661–672

47. Prochnow, C., Bransteitter, R., Klein, M. G., Goodman, M. F., and Chen, X. S. (2007) The APOBEC-2 crystal structure and functional implications for the deaminase AID. *Nature* 445, 447–451

48. Comb, W. C., Hutti, J. E., Cogswell, P., Cantley, L. C., and Baldwin, A. S. (2012) p85α SH2 domain phosphorylation by IKK promotes feedback inhibition of PI3K and Akt in response to cellular starvation. *Mol. Cell* 45, 719–730

49. Couture, C., Songyang, Z., Jascur, T., Williams, S., Tailor, P., Cantley, L. C., and Mustelin, T. (1996) Regulation of the Lck SH2 domain by tyrosine phosphorylation. *J. Biol. Chem.* 271, 24880–24884

50. Huang, H., Li, L., Wu, C., Schibli, D., Colwill, K., Ma, S., Li, C., Roy, P., Ho, K., Songyang, Z., Pawson, T., Gao, Y., and Li, S. S.-C. (2008) Defining the specificity space of the human SRC homology 2 domain. *Molecular & Cellular Proteomics* 7, 768–784

51. Kaneko, T., Huang, H., Zhao, B., Li, L., Liu, H., Voss, C. K., Wu, C., Schiller, M. R., and Li, S. S.-C. (2010) Loops govern SH2 domain specificity by controlling access to binding pockets. *Science Signaling* 3, ra34

52. Stover, D. R., Furet, P., and Lydon, N. B. (1996) Modulation of the SH2 binding specificity and kinase activity of Src by tyrosine phosphorylation within its SH2 domain. *J. Biol. Chem.* 271, 12481–12487

53. Kulshina, N., Edwards, T. E., and Ferré-D'Amaré, A. R. (2010) Thermodynamic analysis of ligand binding and ligand binding-induced tertiary structure formation by the thiamine pyrophosphate riboswitch. *RNA* 16, 186–196

54. Malik, R., Nigg, E. A., and Körner, R. (2008) Comparative conservation analysis of the human mitotic phosphoproteome. *Bioinformatics* 24, 1426–1432

55. Jiménez, J. L., Hegemann, B., Hutchins, J. R. A., Peters, J.-M., and Durbin, R. (2007) A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database. *Genome Biol.* 8, R90

56. Eddy, S. R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform* 23, 205–211

57. Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19, 327–335

58. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 25, 1605–1612

59. Csardi, G., and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal Complex Systems*,

60. Bodenhofer, U., Kothmeier, A., and Hochreiter, S. (2011) APCluster: an R package for affinity propagation clustering. *Bioinformatics* 27, 2463–2464

61. Frey, B. J., and Dueck, D. (2007) Clustering by passing messages between data points. *Science* 315, 972–976

Figure legends

Figure 1: relationship between phosphorylation propensity of the domain – the proportion of ser/thr/tyr that are phosphorylated – and the proportion of domain copies that are phosphorylated in at least one domain instance. Each circle represents a domain. The size of the point is proportional to the number of different paralogy groups in which a domain is found. (a): pSer (b): pThr, (c): pTyr. For clarity only domains with at least ten occurrences in the proteome are shown.

Figure 2: phosphorylation propensity of IDRs and their flanking domains. The IDR propensity is calculated as the fraction of residues that are phosphorylated on all the phosphorylatable residues in each specific IDR. The color indicates the propensity of the IDR. The figure is symmetric in color because the propensity does not take into account the ordering of the two domains. Once a point is located, the size of the plotting symbol shows the propensity of the domain on the x-axis. By moving to the symmetric point along the diagonal it is possible to determine the propensity of the other domain defining the IDR. (a): pSer, (b): pThr, (c): pTyr. For clarity only significant IDRs (p-value < 0.05) with at least ten occurrences are shown for pSer, while for pThr and for pTyr only IDRs with at least 5 occurrences. Moreover the plot does not contain IDRs where one of the two domains is the N/C-term.

Figure 3: relationship between interface propensity - the proportion of interface residues for each structural cluster of phosphorylation sites and the average phosphosite conservation. Interface propensity was binned in five equally-spaced intervals. Points indicate outliers.

Figure 4: Example structures showing the overlap between clusters of poorly conserved phosphorylation sites and protein-protein interface regions. All the domains are represented as white molecular surfaces with phosphosites colored according to their original cluster. The interacting structures are represented as transparent ribbons. The phosphosites in the domain alignment are highlighted with the same color used in the structure representation. A: Variant SH3 domain in a homodimeric complex (PDB: 3a98). B: APOBEC-like N-terminal domain from Probable C->U-editing enzyme APOBEC-2 in a homotetrameric complex (PDB: 2nyt). C: SH2 domain (PDB: 1fyr). D: RNA recognition motif (PDB: 3k0j).

Table Legends

Table 1.a-b. Domains enriched/depleted in Ser phosphorylation (p-value < 0.05): domain types significantly (p-value < 0.05) enriched (table 1a) and depleted (table 1b) in Ser phosphorylation, sorted by p-value (for clarity only the first 15 rows are showed, the full data is available in Supplementary Table S2). Length indicates the length of the alignment. Adjusted p-value is the p-value for the Fisher test (see Methods), after False Discovery Rate correction. pSer Propensity is the proportion of serine residues in the alignment that are phosphorylated. Ser Content is the proportion of serine residues in the alignment.

Table 2.a-b: Domains enriched/depleted in Thr phosphorylation (p-value < 0.05): domain types significantly (p-value < 0.05) enriched/depleted in Thr phosphorylation, sorted by p-value (for clarity only the first 10 rows are showed, the full data is available in Supplementary Table S2). See the legend of table 1 for a description of the columns.

Table 3.a-b: Domains enriched/depleted in Tyr phosphorylation (p-value < 0.05): domain types significantly (p-value < 0.05) enriched/depleted in Tyr phosphorylation, sorted by p-value (for clarity only the first 8 rows are

showed, the full data is available in Supplementary Table S2). See the legend of table 1 for a description of the columns.

Table 4. Abundance of the significantly enriched/depleted domain types/copies in the Human Proteome: this table details the proportion of all the domain types that are significantly enriched/depleted in phosphorylation (Significantly Enriched Domain Types, Significantly Depleted Domain Types). We also calculated the proportion of all the domain occurrences in the proteome that belong to an enriched/depleted type (Significantly enriched domains, Significantly depleted domains). Pospho Propensity is the overall proportion of all Ser/Thr or Tyr that are phosphorylated.

Table 5.a-b: Inter Domain Regions enriched/depleted in Ser phosphorylation (p-value < 0.05): IDR types significantly (p-value < 0.05) enriched (table 5.a)/depleted (table 5.b) in Ser phosphorylation, sorted by p-value (for clarity only the first 15 rows are showed, the full data is available in Supplementary Table S3). Average Length is the average length of the IDR instances. Adjusted p-value is the p-value for the Fisher test (see Methods), after False Discovery Rate correction. Ser Content is the proportion of serine residues in the alignment. pSer Propensity is the proportion of serines in IRD occurrences that are actually phosphorylated.

Table 6.a-b: Inter Domain Regions enriched/depleted in Thr phosphorylation (p-value < 0.05): IDR types significantly (p-value < 0.05) depleted in Ser/Thr phosphorylation, sorted by p-value (for clarity only the first 6 rows are showed, the full data is available in Supplementary Table S3). See the legend of table 5 for a description of the columns.

Table 7.a-b: Inter Domain Regions enriched/depleted in Tyr phosphorylation (p-value < 0.05): IDR types significantly (p-value < 0.05) depleted in Ser/Thr phosphorylation, sorted by p-value (for clarity only the first 4 rows are showed, the full data is available in Supplementary Table S3). See the legend of table 5 for a description of the columns.

Table 8. Predictor Performances: performance of the pSer/pThr and pTyr predictors when using the sequence only, or including the additional domain feature.

Tables

Table 1a

Name	Description	Paralogy groups	Length	Adjusted P-value	Ser Content	pSer Propensity
K167R	K167R (NUC007) repeat	1	111	1.68E-38	1.10	0.49
Synaptobrevin	Synaptobrevin	2	79	6.22E-11	0.87	0.42
Histone	Core histone H2A/H2B/H3/H4	14	72	1.01E-16	1.11	0.20
BEX	Brain expressed X-linked like family	3	132	1.21E-05	0.76	0.18
Tubulin_C	Tubulin C-terminal domain	5	125	1.32E-08	0.86	0.17
Linker_histone	linker histone H1 and H5 family	3	70	2.22E-03	1.64	0.14
ubiquitin	Ubiquitin family	13	65	1.21E-05	0.82	0.13
Band_3_cyto	Band 3 cytoplasmic domain	2	244	1.30E-05	1.26	0.13
HSP90	Hsp90 protein	1	307	8.67E-05	0.89	0.12
Vinculin	Vinculin family	1	351	4.29E-05	1.00	0.11
RRM_6	RNA recognition motif (a.k.a. RRM, RBD, or RNP domain)	19	68	9.48E-03	0.74	0.09
HMG_box	HMG (high mobility group) box	11	67	2.60E-02	0.78	0.09
Tubulin	Tubulin/FtsZ family, GTPase domain	5	212	3.82E-04	1.07	0.09
HSP70	Hsp70 protein	2	514	4.83E-03	0.86	0.07
Pkinase	Protein kinase domain	89	255	2.93E-05	0.84	0.05

Table 1.b

Name	Description	Paralogy groups	Length	Adjusted P-value	Ser Content	pSer Propensity
PMG	PMG protein	4	150	2.8E-04	3.55	0.00
DUF1220	Repeat of unknown function (DUF1220)	1	65	3.4E-08	1.95	0.00
DENN	DENN (AEX-3) domain	3	188	1.2E-02	1.25	0.00
PI-PLC-X	Phosphatidylinositol-specific phospholipase C, X domain	4	138	4.0E-02	1.20	0.00
PLAT	PLAT/LH2 domain	5	107	2.2E-02	0.92	0.00
RhoGEF	RhoGEF domain	20	178	2.4E-09	0.89	0.00
BACK	BTB And C-terminal Kelch	14	98	8.0E-04	0.89	0.00
PDEase_I	3'5'-cyclic nucleotide phosphodiesterase	4	235	4.3E-03	0.84	0.00
RabGAP-TBC	Rab-GTPase-TBC domain	8	198	1.7E-06	0.81	0.00
PI3_PI4_kinase	Phosphatidylinositol 3- and 4-kinase	5	233	3.1E-02	0.71	0.00
BTB	BTB/POZ domain	32	104	6.2E-11	1.06	0.00
Oxysterol_BP	Oxysterol-binding protein	3	332	1.1E-02	1.13	0.00
MAGE	MAGE family	3	166	1.1E-02	0.77	0.00
SEA	SEA domain	7	96	8.1E-05	1.15	0.00
NACHT	NACHT domain	7	163	2.7E-02	0.98	0.00

Table 2.a

Name	Description	Paralogy groups	Length	Adjusted P-value	Thr Content	pThr Propensity
K167R	K167R (NUC007) repeat	1	111	7.1E-51	2.29	0.38
Histone	Core histone H2A/H2B/H3/H4	14	72	1.7E-09	0.97	0.15
ubiquitin	Ubiquitin family	13	65	4.2E-08	1.44	0.12
GTP_EFTU_D2	Elongation factor Tu domain 2	6	71	2.3E-02	1.33	0.10
Tubulin_C	Tubulin C-terminal domain	5	125	1.6E-03	1.30	0.09
HSP70	Hsp70 protein	2	514	3.4E-09	1.27	0.08
Pkinase	Protein kinase domain	89	255	1.1E-72	0.87	0.08
Tubulin	Tubulin/FtsZ family, GTPase domain	5	212	4.3E-06	1.37	0.08
Vinculin	Vinculin family	1	351	1.4E-02	1.10	0.07
Pkinase_Tyr	Protein tyrosine kinase	25	260	1.5E-02	0.79	0.04

Table 2.b

Name	Description	Paralogy groups	Length	Adjusted P-value	Thr Content	pThr Propensity
SPRY	SPRY domain	20	117	3.72E-04	1.03	0.00
RhoGEF	RhoGEF domain	20	178	1.80E-03	0.81	0.00
RabGAP-TBC	Rab-GTPase-TBC domain	8	198	2.78E-02	0.73	0.00
SEA	SEA domain	7	96	8.64E-04	1.89	0.00
HAD	haloacid dehalogenase-like hydrolase	3	433	2.77E-02	1.34	0.00
Hydrolase	Ankyrin repeats (3 copies)	6	81	1.94E-09	0.91	0.00
Hydrolase_3	Homeobox domain	1	57	1.00E-03	1.37	0.00
Ank_2	Neurotransmitter-gated ion-channel transmembrane region	66	168	2.77E-02	1.39	0.00
Homeobox	BTB/POZ domain	40	104	3.26E-02	0.90	0.00
Neur_chan_memb	Myosin head (motor domain)	9	623	1.50E-02	0.98	0.01

Table 3.a

Name	Description	Paralogy groups	Length	Adjusted P-value	Tyr Content	pTyr Propensity
Globin	Globin	2	104	6.02E-03	0.48	0.42
Linker_histone	linker histone H1 and H5 family	3	70	6.79E-03	0.91	0.35
Tubulin_C	Tubulin C-terminal domain	5	125	9.60E-10	0.90	0.34
Myosin_tail_1	Myosin tail	5	772	2.07E-08	0.21	0.30
Myosin_TH1	Cofilin/tropomyosin-type actin-binding protein	2	120	5.87E-05	1.32	0.29
Cofilin_ADF	Core histone H2A/H2B/H3/H4	4	72	1.21E-11	1.41	0.27
Histone	Ubiquitin family	14	65	1.63E-03	0.59	0.23
ubiquitin	Protein tyrosine kinase	13	260	4.63E-65	1.19	0.22

Table 3.b

Name	Description	Paralogy groups	Length	Adjusted P-value	Tyr Content	pTyr Propensity
DUF1220	Repeat of unknown function (DUF1220)	1	65	1.3E-03	1.36	0.00
adh_short	short chain dehydrogenase	13	162	4.7E-02	0.63	0.00
HECT	HECT-domain (ubiquitin-transferase)	9	305	3.7E-05	1.36	0.00
Hormone_recep	Ligand-binding domain of nuclear hormone receptor	11	180	3.1E-02	0.74	0.00
BTB	BTB/POZ domain	32	104	2.1E-05	1.09	0.01
SPRY	SPRY domain	20	117	5.7E-04	1.47	0.01
Kinesin	Kinesin motor domain	9	323	1.9E-02	1.00	0.02
UCH	Ubiquitin carboxyl-terminal hydrolase	12	417	3.7E-05	1.27	0.02

Table 4

Phospho Residue	Phospho Propensity	Significantly Enriched Domain Types	Significantly Depleted Domain Types	Significantly Enriched Domains	Significantly Depleted Domains
P-Ser	0.035	0.048	0.011	0.064	0.15
P-Thr	0.023	0.018	0.0035	0.050	0.078
P-Tyr	0.055	0.012	0.0026	0.066	0.033

Table 5.a

Name 1	Name 2	Description 1	Description 2	Average Length	Adjusted P-value	Ser Content	pSer Propensity
C1_1	Pkinase	Phorbol esters/diacylglycerol binding domain (C1 domain)	Protein kinase domain	78	4.84E-12	1.19	0.52
PH	PH	PH domain	PH domain	106	1.03E-07	1.44	0.25
Pkinase	CNH	Protein kinase domain	CNH domain	502	1.87E-23	1.03	0.25
PHD	Bromodomain	PHD-finger	Bromodomain	105	5.75E-05	1.46	0.24
CAP_GLY	CAP_GLY	CAP-Gly domain	CAP-Gly domain	102	6.56E-06	2.12	0.23
Ran_BP1	Ran_BP1	RanBP1 domain	RanBP1 domain	176	6.73E-06	1.21	0.23
TPR_1	TPR_2	Tetratricopeptide repeat	Tetratricopeptide repeat	186	6.05E-04	0.88	0.22
dsrm	dsrm	Double-stranded RNA binding motif	Double-stranded RNA binding motif	79	2.67E-02	0.87	0.22
Bromodomain	Bromodomain	Bromodomain	Bromodomain	130	5.80E-07	1.00	0.22
WW	WW	WW domain	WW domain	157	2.22E-05	1.16	0.21
CH	LIM	Calponin homology (CH) domain	LIM domain	270	2.08E-05	1.26	0.18
BAR	SH3_2	BAR domain	Variant SH3 domain	167	1.34E-02	0.93	0.18
GTF2I	GTF2I	GTF2I-like repeat	GTF2I-like repeat	121	7.09E-03	1.19	0.17
KH_1	KH_1	KH domain	KH domain	115	2.58E-03	1.02	0.17
C1_1	MA3	MIF4G domain	MA3 domain	188	3.58E-02	0.91	0.16

Table 5.b

Name 1	Name 2	Domain 1	Domain 2	Average Length	Adjusted P-value	Ser Content	pSer Propensity
DUF1053	Guanylate_cyc	Domain of Unknown Function (DUF1053)	Adenylate and Guanylate cyclase catalytic domain	226	6.00E-03	0.00	0
NACHT	LRR_6	NACHT domain	Leucine Rich repeat	420	7.48E-12	0.00	0
PLAT	PKD_channel	PLAT/LH2 domain	Polycystin cation channel	500	9.09E-06	0.00	0
Na_Ca_ex	Na_Ca_ex	Sodium/calcium exchanger protein	Sodium/calcium exchanger protein	223	3.35E-02	0.00	0
zf-C2H2	zf-C2H2_6	Zinc finger, C2H2 type	C2H2-type zinc finger	129	2.39E-02	0.00	0
zf-C3HC4	SPRY	Zinc finger, C3HC4 type (RING finger)	SPRY domain	242	3.14E-02	0.00	0
Ank_2	Ank_2	Ankyrin repeats (3 copies)	Ankyrin repeats (3 copies)	117	3.15E-03	0.00	0
PYRIN	NACHT	PAAD/DAPIN/Pyrin domain	NACHT domain	105	3.28E-02	0.00	0
Homeobox	Homeobox	Homeobox domain	Homeobox domain	155	1.13E-03	0.00	0
Calx-beta	Calx-beta	Calx-beta domain	Calx-beta domain	298	2.54E-05	0.00	0
SEA	SEA	SEA domain	SEA domain	105	4.15E-16	0.00	0
WW	HECT	WW domain	HECT-domain (ubiquitin-transferase)	152	2.39E-02	0.00	0
LRR_6	LRR_6	Leucine Rich repeat	Leucine Rich repeat	119	3.60E-05	0.00	0
Beach	WD40	Beige/BEACH domain	WD domain, G-beta repeat	144	8.75E-03	0.00	0
DUF1220	DUF1220	Repeat of unknown function (DUF1220)	Repeat of unknown function (DUF1220)	195	3.05E-03	0.00	0

Table 6.a

Name 1	Name 2	Domain 1	Domain 2	Average Length	Adjusted P-value	Thr Content	pThr Propensity
TPR_2	TPR_1	Tetratricopeptide repeat	Tetratricopeptide repeat	186	8.08E-07	1.26	0.26
Ran_BP1	GRIP	RanBP1 domain	GRIP domain	240	5.82E-04	1.42	0.16
PBD	Pkinase	P21-Rho-binding domain	Protein kinase domain	233	2.11E-03	1.22	0.15
GTF2I	GTF2I	GTF2I-like repeat	GTF2I-like repeat	121	3.84E-02	1.12	0.14
BAR	SH3_2	BAR domain	Variant SH3 domain	167	3.83E-02	1.66	0.13
Pkinase	CNH	Protein kinase domain	CNH domain	502	5.82E-04	0.80	0.12

Table 6.b

Name 1	Name 2	Domain 1	Domain 2	Average Length	Adjusted P-value	Thr Content	pThr Propensity
NACHT	LRR_6	NACHT domain	Leucine Rich repeat	420	2.00E-03	0.92	0.00
SEA	SEA	SEA domain	SEA domain	105	6.34E-09	2.90	0.00
KRAB	zf-H2C2_2	KRAB box	Zinc-finger double domain	153	1.54E-19	0.94	0.00
KRAB	zf-C2H2	KRAB box	Zinc finger, C2H2 type	150	4.88E-02	0.88	0.00
Ion_trans	Ion_trans	Ion transport protein	Ion transport protein	183	2.89E-02	0.87	0.01
WD40	WD40	WD domain, G-beta repeat	WD domain, G-beta repeat	135	7.72E-07	1.16	0.01

Table 7.a

Name 1	Name 2	Domain 1	Domain 2	Average Length	Adjusted P-value	Tyr Content	pTyr Propensity
SH2	SH2	SH2 domain	SH2 domain	160	3.75E-10	2.31	0.58
SH3	SH3_2	SH3 domain	Variant SH3 domain	118	9.89E-03	0.66	0.47
PH	SH2	PH domain	SH2 domain	143	2.75E-02	0.86	0.46
Pkinase	CNH	Protein kinase domain	CNH domain	502	9.98E-09	0.76	0.38

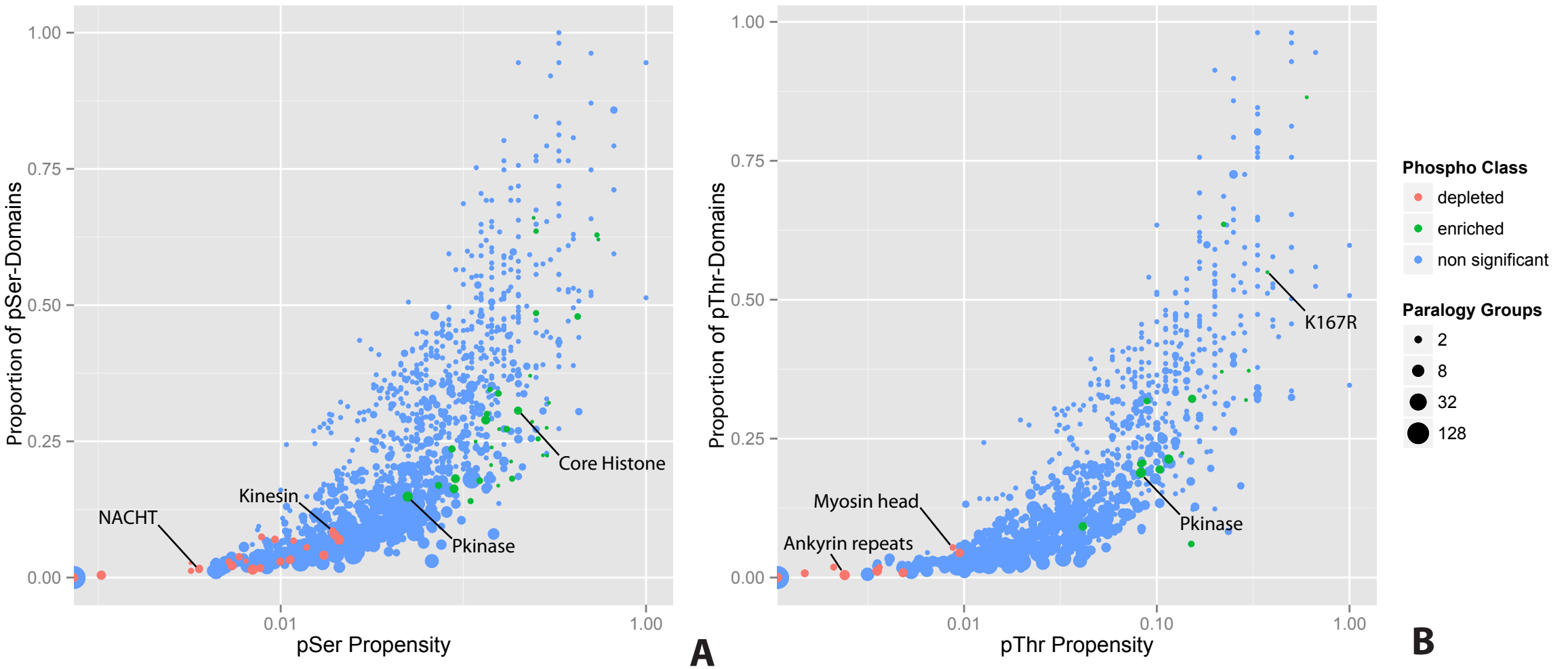
Table 7.b

Name 1	Name 2	Domain 1	Domain 2	Average Length	Adjusted P-value	Tyr Content	pTyr Propensity
SEA	SEA	SEA domain	SEA domain	105	5.5E-03	1.50	0.00
zf-H2C2_2	zf-H2C2_2	Zinc-finger double domain	Zinc-finger double domain	194	4.7E-02	0.93	0.01
KRAB	zf-H2C2_2	KRAB box	Zinc-finger double domain	153	8.5E-09	1.27	0.02
WD40	WD40	WD domain, G-beta repeat	WD domain, G-beta repeat	135	1.4E-03	1.21	0.02

Table 8

Res	Sequence	Sequence + Domain Context
Y	**0.58**	**0.66**
T	**0.68**	**0.72**
S	**0.70**	**0.73**

Figure 1


1.00
0.75
0.50
0.25
0.00
Proportion of pSer-Domains
0.01
1.00
pSer Propensity
NACHT
Kinesin
Pkinase
Core Histone
A
1.00
0.75
0.50
0.25
0.00
Proportion of pThr-Domains
0.01
0.10
1.00
pThr Propensity
Myosin head
Ankyrin repeats
Pkinase
K167R
B
Phospho Class
depleted
enriched
non significant
Paralogy Groups
2
8
32
128

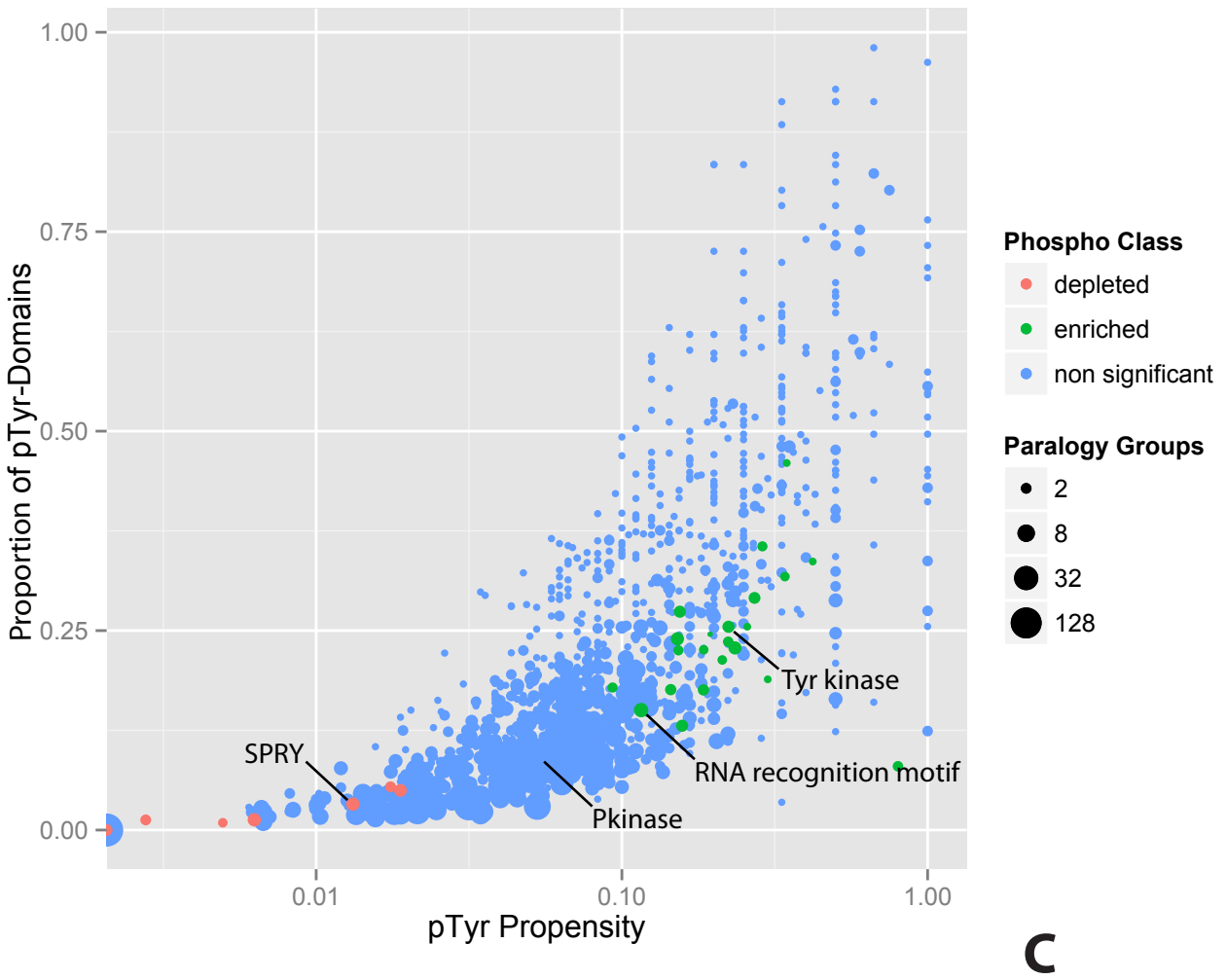
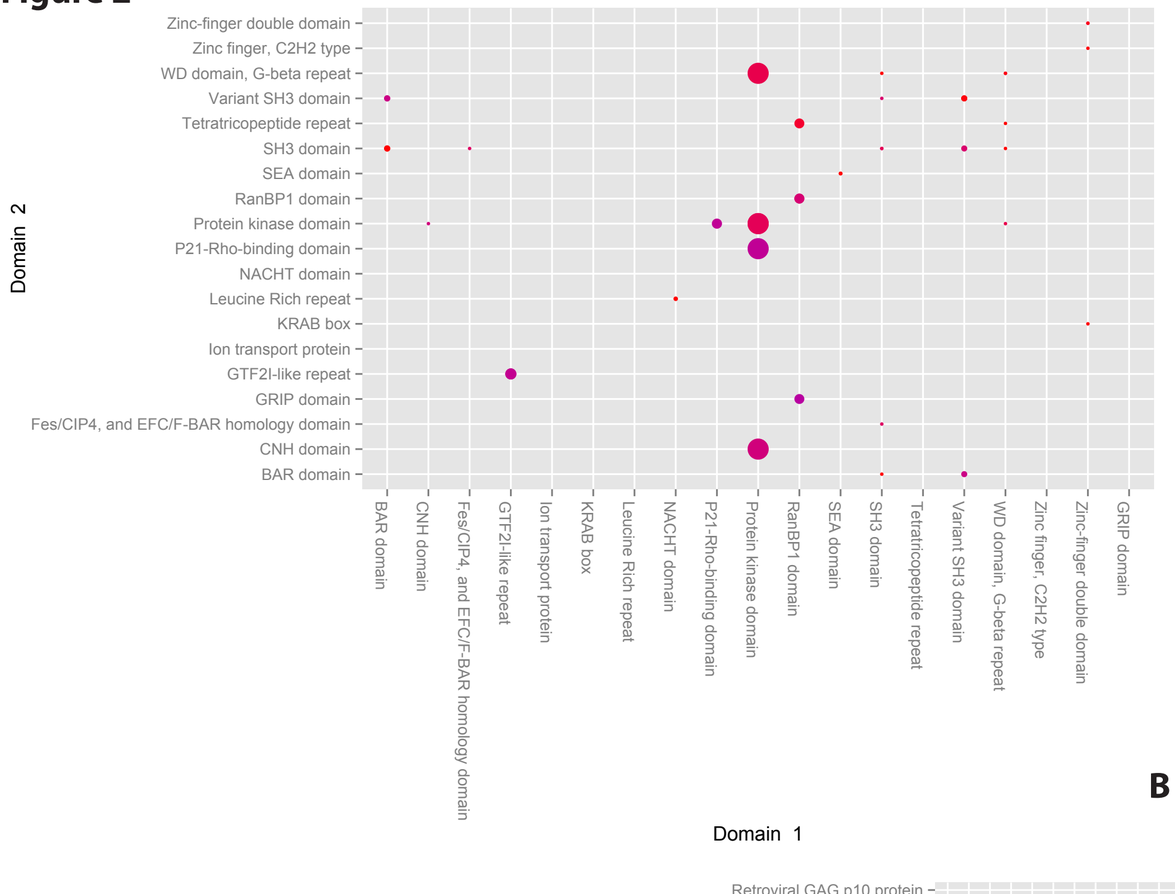
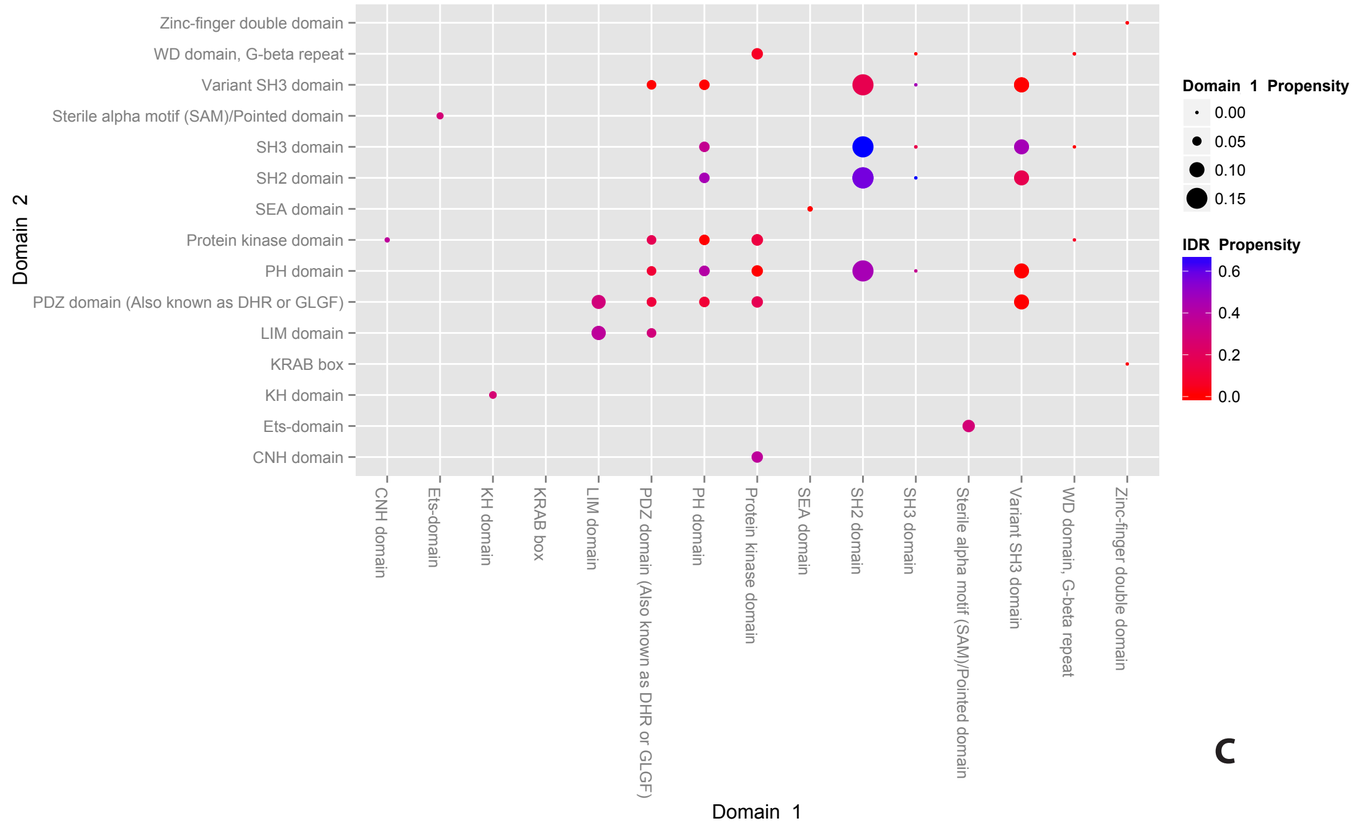

1.00
0.75
0.50
0.25
0.00
Proportion of pTyr-Domains
0.01
0.10
1.00
pTyr Propensity
SPRY
Pkinase
RNA recognition motif
Tyr kinase
C
Phospho Class
depleted
enriched
non significant
Paralogy Groups
2
8
32
128

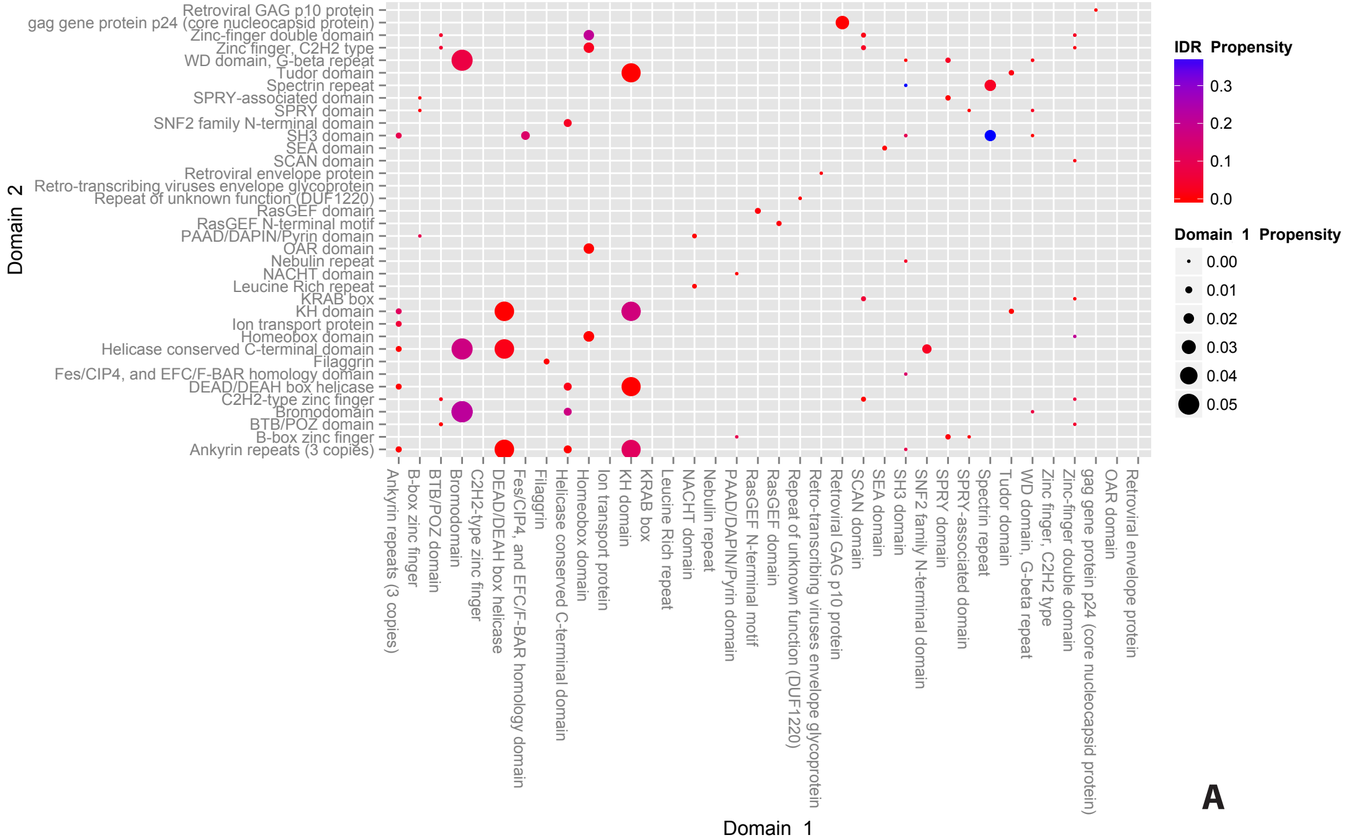
Figure 2


Zinc-finger double domain
Zinc finger, C2H2 type
WD domain, G-beta repeat
Variant SH3 domain
Tetratricopeptide repeat
SH3 domain
SEA domain
RanBP1 domain
Protein kinase domain
P21-Rho-binding domain
NACHT domain
Leucine Rich repeat
KRAB box
Ion transport protein
GTF2I-like repeat
GRIP domain
Fes/CIP4, and EFC/F-BAR homology domain
CNH domain
BAR domain
Domain 2
BAR domain
CNH domain
Fes/CIP4, and EFC/F-BAR homology domain
GTF2I-like repeat
Ion transport protein
KRAB box
Leucine Rich repeat
NACHT domain
P21-Rho-binding domain
Protein kinase domain
RanBP1 domain
SEA domain
SH3 domain
Tetratricopeptide repeat
Variant SH3 domain
WD domain, G-beta repeat
Zinc finger, C2H2 type
Zinc-finger double domain
GRIP domain
Domain 1

B


Zinc-finger double domain
WD domain, G-beta repeat
Variant SH3 domain
Sterile alpha motif (SAM)/Pointed domain
SH3 domain
SH2 domain
SEA domain
Protein kinase domain
PH domain
PDZ domain (Also known as DHR or GLGF)
LIM domain
KRAB box
KH domain
Ets-domain
CNH domain
Domain 2
CNH domain
Ets-domain
KH domain
KRAB box
LIM domain
PDZ domain (Also known as DHR or GLGF)
PH domain
Protein kinase domain
SEA domain
SH2 domain
SH3 domain
Sterile alpha motif (SAM)/Pointed domain
Variant SH3 domain
WD domain, G-beta repeat
Zinc-finger double domain
Domain 1
Domain 1 Propensity
0.00
0.05
0.10
0.15
IDR Propensity
0.6
0.4
0.2
0.0

C


Retroviral GAG p10 protein
gag gene protein p24 (core nucleocapsid protein)
Zinc-finger double domain
Zinc finger, C2H2 type
WD domain, G-beta repeat
Tudor domain
Spectrin repeat
SPRY-associated domain
SPRY domain
SNF2 family N-terminal domain
SH3 domain
SEA domain
SCAN domain
Retroviral envelope protein
Retro-transcribing viruses envelope glycoprotein
Repeat of unknown function (DUF1220)
RasGEF domain
RasGEF N-terminal motif
PAAD/DAPIN/Pyrin domain
OAR domain
Nebulin repeat
NACHT domain
Leucine Rich repeat
KRAB box
KH domain
Ion transport protein
Homeobox domain
Helicase conserved C-terminal domain
Filaggrin
Fes/CIP4, and EFC/F-BAR homology domain
DEAD/DEAH box helicase
C2H2-type zinc finger
Bromodomain
BTB/POZ domain
B-box zinc finger
Ankyrin repeats (3 copies)
Domain 2
Ankyrin repeats (3 copies)
B-box zinc finger
BTB/POZ domain
Bromodomain
C2H2-type zinc finger
DEAD/DEAH box helicase
Fes/CIP4, and EFC/F-BAR homology domain
Filaggrin
Helicase conserved C-terminal domain
Homeobox domain
Ion transport protein
KH domain
KRAB box
Leucine Rich repeat
NACHT domain
Nebulin repeat
PAAD/DAPIN/Pyrin domain
RasGEF N-terminal motif
RasGEF domain
Repeat of unknown function (DUF1220)
Retro-transcribing viruses envelope glycoprotein
Retroviral GAG p10 protein
SCAN domain
SEA domain
SH3 domain
SNF2 family N-terminal domain
SPRY domain
SPRY-associated domain
Spectrin repeat
Tudor domain
WD domain, G-beta repeat
Zinc finger, C2H2 type
Zinc-finger double domain
gag gene protein p24 (core nucleocapsid protein)
OAR domain
Retroviral envelope protein
Domain 1
IDR Propensity
0.3
0.2
0.1
0.0
Domain 1 Propensity
0.00
0.01
0.02
0.03
0.04
0.05

A

Figure 3

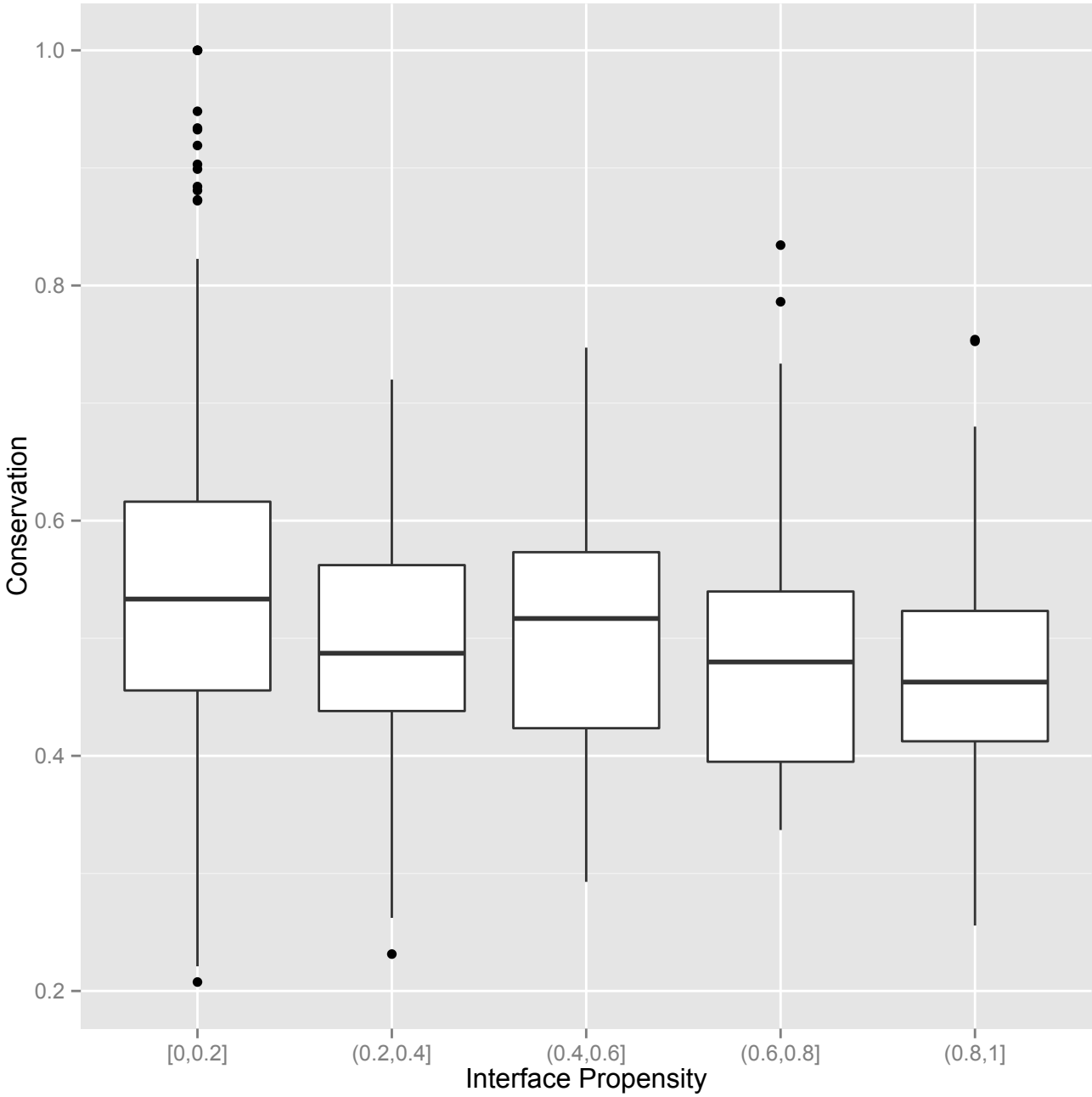
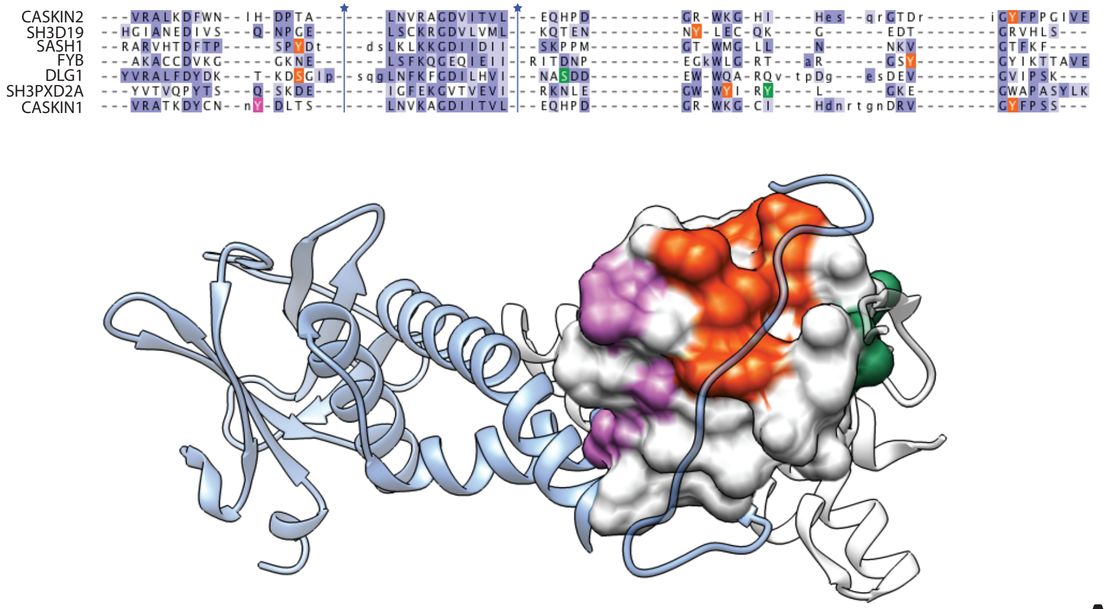
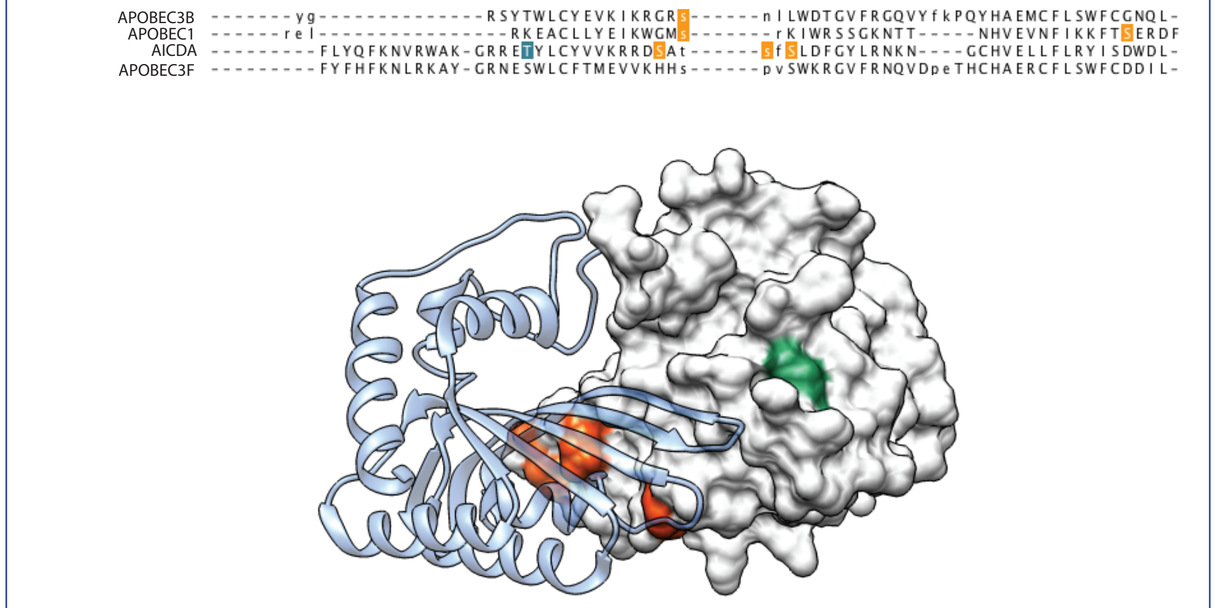
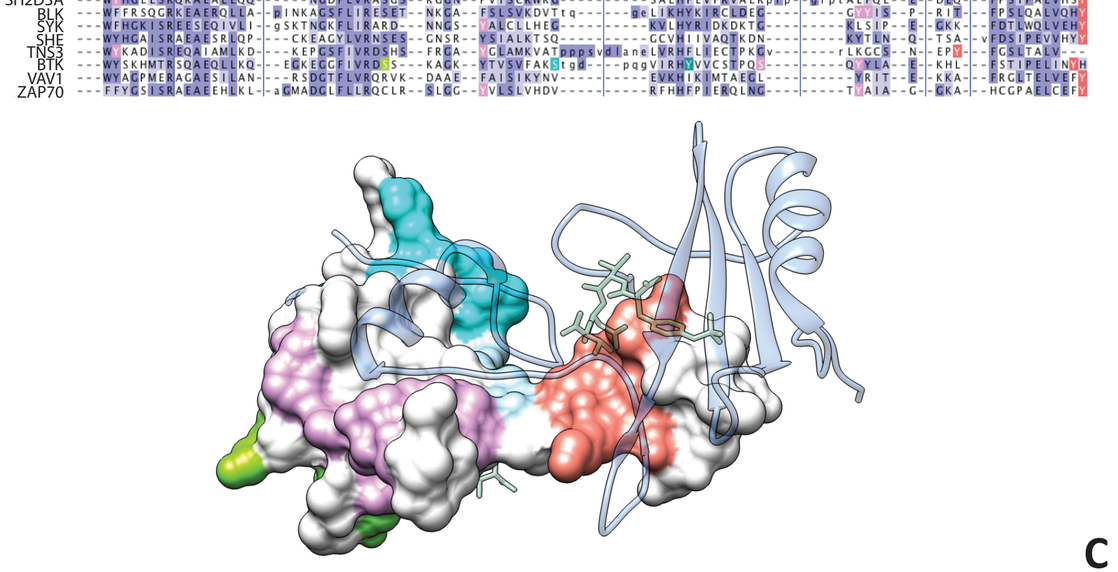
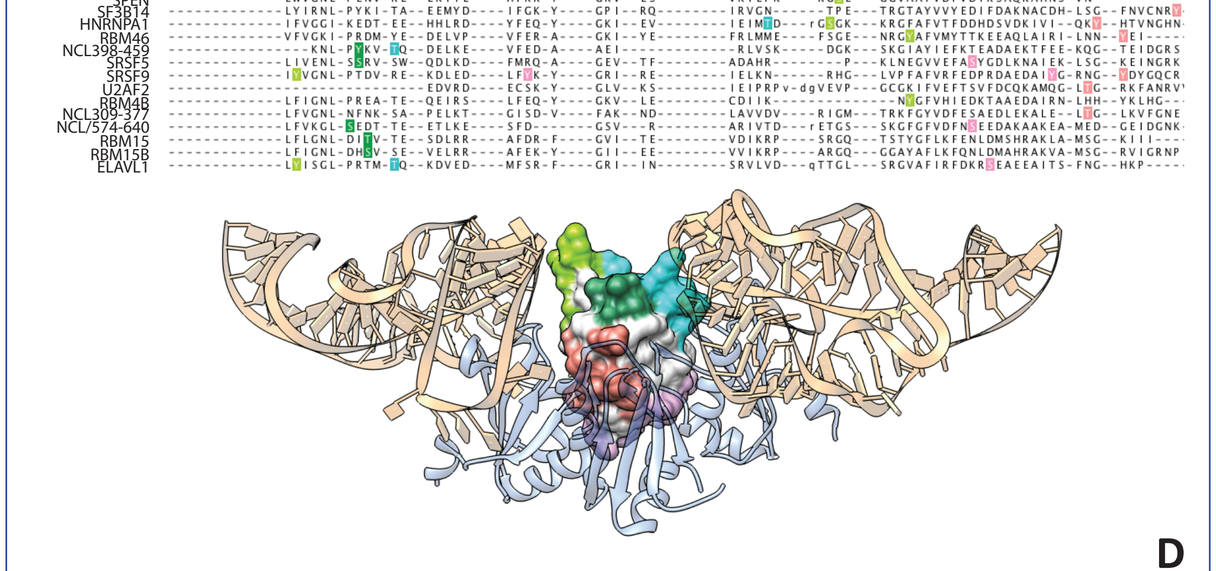

1.0
0.8
0.6
0.4
0.2
Conservation
[0,0.2]
(0.2,0.4]
(0.4,0.6]
(0.6,0.8]
(0.8,1]
Interface Propensity

Figure 4


CASKIN2
SH3D19
SASH1
FYB
DLG1
SH3PXD2A
CASKIN1
A


APOBEC3B
APOBEC1
AICDA
APOBEC3F
B


STAT3
BMX
SHB
PIK3R1
SH2D3A
BLK
SYK
SHE
TNS3
BTK
VAV1
ZAP70
C


HNRNPD
EIF4H
RBM47
SRSF4
SPEN
SF3B14
HNRNPA1
RBM46
NCL398-459
SRSF5
SRSF9
U2AF2
RBM48
NCL309-377
NCL/574-640
RBM15
RBM15B
ELAVL1
D