Evolutionarily conserved regions of THAP9 transposase reveal new motifs for subcellular localization and post-translational modification.

Richa Rashmi, Chandan Nandi, Sharmistha Majumdar

IIT Gandhinagar

Abstract

THAP9 is a transposable element-derived gene that encodes the THAP9 protein. It is a homolog of Drosophila P-element transposase (DmTNP). hTHAP9 belongs to the THAP (Thanatos-associated protein) protein family in humans, which has twelve proteins (hTHAP0-hTHAP11). Human THAP proteins have a ~90 amino acid long conserved THAP domain located at the N-terminal end of the protein. The THAP domain is a C2CH type Zinc Finger domain that binds specific DNA sequences and has a conserved secondary structure, namely a characteristic β-α-β fold. THAP family proteins are known to be involved in various human diseases like torsional dystonia, hemophilia, heart diseases, and multiple cancers. Also, the THAP9-AS1 gene, which is the neighboring gene in the 5' upstream region of THAP9, is overexpressed in pancreatic cancer. But the exact functional role of THAP9 and its association with various diseases remains a significant gap in our understanding. Here, we present the evolutionary analysis and extensive *in silico* characterization, including predicting domains and putative post-translational modification sites for THAP9 to understand its function better. Our studies identified some previously unreported functional features in the THAP9 protein sequence, including four adjacent motifs following the THAP domain: N-glycosylation site, Protein kinase C (PKC) phosphorylation site, Leucine zipper domain, and Bipartite nuclear localization signal (NLS). These motifs, which are highly conserved in mammals, may play a role in the subcellular localization of THAP9. The study also revealed two N-myristoylation sites within the THAP domain, pointing towards the post-translational modification of THAP9. We also observed that the THAP9 has evolved under a strong pervasive purifying selection, yielding high conservation of THAP9. Furthermore, investigation of THAP9 expression profiles in various cancer and matched normal datasets demonstrated overexpression and underexpression in Testicular cancers and Thymic Epithelial Tumors respectively. These findings make Testicular cancer and Thymic Epithelial Tumor good candidates for further examination on the role of THAP9 in cancer.

Keywords

Evolutionary analysis; Transposase gene; Role in cancers, Protein sequence characterization;

Introduction

Transposable elements (TEs) are DNA sequences that can move and duplicate within a genome[1]. They can cause various mutations as well as an increase in genome size[2]. Thus, TEs, which constitute 25% - 50% of various genomes, play a significant role in evolution.

Many genes are derived from transposable elements. Human THAP9, which encodes the hTHAP9 protein, is one such gene. The hTHAP9 protein is a homolog of the Drosophila P-element transposase (DmTNP) (more than 25% homology)[3]. The DmTNP protein mobilizes the P-element transposon, which is the causative agent for Hybrid Dysgenesis in Drosophila.

hTHAP9 belongs to the THAP (Thanatos-associated protein) protein family in humans, which has twelve proteins (hTHAP0-hTHAP11). All human THAP proteins are characterized by an 80-90 amino acid residue-long conserved THAP domain located at the N-terminal end of the protein[4]. The THAP domain is a C2CH type Zinc Finger domain that can bind specific DNA sequences and has a conserved secondary structure, namely a characteristic β-α-β fold with four loops (L1-L4) flanking interconnecting the sheets and the helix[5].

Many THAP family proteins are known to be involved in various human diseases. THAP1 is involved in torsional dystonia and hemophilia[6], THAP5 is involved in heart diseases[7], THAP2[8], 10[9], and 11[10] are involved in multiple cancers. Also, the THAP9-AS1 gene, which is the neighboring gene in the 5' upstream region of THAP9, is overexpressed in pancreatic cancer[11]. Moreover, according to NCBI & the Protein Atlas, THAP9 has high expression in the brain, ovary, prostate, testis & thyroid (https://www.ncbi.nlm.nih.gov/gene/79725, https://www.proteinatlas.org/ENSG00000168152-THAP9/tissue). But the exact functional role of THAP9 and its association with various diseases remains a significant gap in our understanding.

Phylogenetic analysis and evolutionary tests of hTHAP9 can help us understand the evolution and the type and extent of natural selection for the hTHAP9 gene and associated protein. In the course of evolution, negative or purifying selection is used to eliminate unfavorable and harmful mutations[12]. On the other hand, the positive or diversifying selection promotes molecular adaptation by promoting favorable mutations[12]. Here, we present the evolutionary analysis and extensive in silico characterization, including predicting domains and putative post-translational modification sites for THAP9 and its orthologs. Our studies identified some previously unreported functional features in the THAP9 protein sequence, including four adjacent motifs: N-glycosylation site, Protein kinase C (PKC) phosphorylation site, Leucine zipper domain, and Bipartite nuclear localization signal (NLS). These motifs, which are highly conserved in mammals, may play a role in the subcellular localization of THAP9. The study also revealed two N-myristoylation sites within the THAP domain. We also observed that the THAP9 has evolved under a strong pervasive purifying selection.

Materials and Methods

Identification of orthologs.

THAP domain containing 9 (*THAP9*) transcript variant 1 (RefSeq: NM_024672.6) from *Homo sapiens* was used to identify the orthologs from the Eukaryotic Genome Annotation pipeline of the NCBI database (National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov/) using a combination of protein sequence similarity and local synteny information for the available sequences. *hTHAP9* orthologs are present in 216 organisms including 74 birds, 4 alligators, 7 turtles, 6 lizards, 4 amphibians, 120 mammals, and 1 lamprey. We used the following five representative classes: amphibians, hyperoartia, reptiles, birds, and mammals for comparison and representation purposes.

Sequence alignment & Phylogenetic analysis

A phylogeny-aware sequence alignment was performed for amino acid sequences using the CLUSTALW[13] method present in MEGAX[14] software using the guide trees generated from TimeTree[15] (http://www.timetree.org/). We created multiple sequence alignments (MSA) separately for all the vertebrate classes (birds, alligators, turtles, lizards, amphibians, and mammals) and a combined alignment for all species. Following the protein MSA of the orthologs, the Coding DNA Sequences (CDS) were codon aligned using PAL2NAL[16] (http://www.bork.embl.de/pal2nal/). Divergent sequences which created alignment errors were identified and removed manually.

The codon alignments and the protein alignments were used to generate the maximum-likelihood tree using MEGAX with 100 bootstrap replicates, and the tree was visualized using iTOL[17].

Analysis for Selective Pressure

To characterize the selection patterns on the alignments of THAP9, we used various tools from the HyPhy package[18] implemented on the Galaxy webserver[19]. With the user-friendly graphical interfaces, new fast methods, and multi-tasking abilities, HyPhy provides better analysis power to users than the PAML[20] software package. The Galaxy web-server version of HyPhy provides additional advantages like it does not require any local installation, and we can automate the whole process by creating a workflow. The approximately-maximum-likelihood phylogenetic trees generated using FASTTREE[21] also implemented on Galaxy webserver were selected with the HyPhy model selection tool along with the codon alignment file. For the HyPhy software package analysis, the default set parameters of p-values/posterior probabilities were considered. FEL[22] (Fixed Effect Likelihood) and FUBAR[23] (Fast Unbiased Bayesian AppRoximation) were utilized to identify sites evolving under pervasive diversifying and pervasive purifying selection pressures. MEME[24] (Mixed Effects Model of Evolution) was used to identify episodically diversifying sites. FEL is an ML-based model which assumes constant selection pressure on each site. The FUBAR model estimates site-by-site ω values while MEME allows the distribution of ω to vary over sites and additionally from branch to branch. Significance was assessed by posterior probability > 0.9 (FUBAR) and p-value < 0.1 (MEME) and p-value < 0.1 (FEL)[25].

Protein sequence characterization

The domain architecture of hTHAP9 and its orthologs was predicted using SMART[26] (Simple Modular Architecture Research Tool, http://smart.embl-heidelberg.de/) in “batch mode” considering the option for including Pfam[27] domains. ELM[28] Prediction tool (Eukaryotic Linear Motif, http://elm.eu.org/) was used to identify short linear motifs using the fasta file containing hTHAP9 & the ortholog protein sequence as input. Furthermore, the previously unidentified conserved motifs and functional domains were searched using Meme-Suite[29] and ScanProsite[30], respectively. Disordered binding regions (DBR)[31] and secondary structure elements of the hTHAP9 protein were predicted using PSIPRED Workbench[32], and areas with disorder scores greater than 0.6 were considered as disordered. The physiochemical attributes of the hTHAP9 protein sequence such as molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index, and grand

average of hydropathy (GRAVY) were computed using Expasy ProtParam[33] and EMBOSS Pepstats[34] (https://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats/). Finally, previously annotated SMART & Pfam domains, as well as newly predicted domains of hTHAP9 and its orthologs, were visualized using iTOL.

Distribution of SNPs and mutations in cancer

The 1000 Genomes project[35] data using Ensembl[36] (reference genome: GRCh38.p13) was searched for missense mutations of hTHAP9 (ENSG00000168152). PolyPhen-2[37] scores given on the Ensembl result table were used to classify the SNPs into “benign,” “possibly damaging,”, “probably damaging”. PolyPhen-2 predicts the effect of an amino acid substitution on the structure and function of a protein using sequence homology, Pfam annotations, 3D structures from PDB were available, and a number of other databases and tools (including DSSP, ncoils, etc)[38]. Cancer-related missense mutations in THAP9 were identified using cBioPortal[39] for Cancer Genomics (http://www.cbioportal.org/). PolyPhen2 scores for these mutations were calculated by submitting their genomic coordinates (for GRCh37 reference genome) to the Ensembl Variant Effect Predictor tool[40] (https://www.ensembl.org/Tools/VEP).

Cancer expression databases and web tools

To get an idea about the expression profile of the *hTHAP9 gene* in various cancers, we used the expression data from TCGA[41] (https://cancergenome.nih.gov) and GTEx[42] (https://gtexportal.org/home). Cancer and matched normal datasets from the above databases were accessed and plotted using the GEPIA[43] web server (http://gepia.cancer-pku.cn).

Results

Evolution of THAP9 through organisms

To study the possible functional role of hTHAP9, we conducted an extensive characterization of its protein sequence and investigated its evolution. hTHAP9 protein is encoded by the gene with the same name and is found on chromosome 4 in humans. Transcript variant 1 of hTHAP9 is known to encode the transposase protein[44]. Therefore, we used the same transcript variant for our analysis.

According to NCBI, hTHAP9 protein has orthologs present in 216 organisms, including 74 birds, 4 alligators, 7 turtles, 6 lizards, 4 amphibians, 120 mammals, and 1 lamprey. We downloaded the protein & coding DNA sequences for the same from NCBI. We began our analysis by aligning the protein sequences using CLUSTALW[13] in a phylogeny-aware manner. We created the time tree for the organisms using timetree[15]. After alignment, we filtered the diverse, low-quality, and partial sequences, after which we were left with 175 sequences. Later we realigned the sequences using CLUSTALW[13], followed by building Maximum likelihood Phylogenetic trees[45] with 100 bootstrap replicates. Both these parts were performed using options available in MEGAX[14]. We also codon aligned the coding sequences from the orthologs using PAL2NAL[16], which also takes corresponding aligned protein sequences as input. The

resulting trees were drawn using the iTOL[17] web server. Rooting trees to the midpoint perfectly aligns the Lampey, amphibians, reptiles, birds & mammals in the given order showing the evolution of THAP9 protein with the evolution of the organisms(**Figure 1**).

To further examine the diversification of THAP9 & its ortholog proteins, we identified their conserved motifs using MEME[29]. We identified the distribution of the top 30 conserved motifs occurring at least once in all the sequences (**Figure 2; Table 1**). The motif composition looked highly conserved in mammals compared to other groups implied its functional relevance in mammals. Future studies should examine the function of the distinctive motifs. The different motif composition among different classes also implies functional diversification among different vertebrate classes.

One of the ways we can study how particular proteins have changed over time in the course of evolution is to study their gene, the corresponding codons, and the amino acid sequence across generations. Mainly it concerns observing and predicting the rates and impacts of single nucleotide mutations, selection pressure, and development of new genes and protein products.

There are 20 amino acids in nature which are coded by 61 codons, making them degenerate[46]. Codons are the connective bridge between the coding gene sequence and the translated protein sequence, allowing them to be essential in evolution. Random mutations or selective pressure that mutates gene sequences often do not translate to any change in amino acids at the protein level; such mutations are called Synonymous mutations[47]. In contrast, if a mutation in the coding sequence leads to a change in the translated amino acid, it is called a Non-synonymous/silent mutation[48]. The rates of these two types of mutations are often denoted as dS and dN, respectively, and their ratio is denoted as (omega); =dN/dS. Thus the value acts as an indication of how natural selection has acted on the protein. Sites under diversifying/positive selection have a higher rate of non-synonymous mutations and have a value greater than 1. If the site has a higher rate of synonymous mutations, it is said to be under Purifying/negative selection and bears a value less than 1. The tools used in our analysis implied mathematical models to measure the values for each codon site across the length of the protein by comparing it with the orthologs[49].

To explore variations in selective pressures across THAP9, we used the HyPhy[18] software package available at Galaxy[19] web-server. Our analysis was aimed at the detection of selection at individual codons. The nature of selection at individual codons can highlight the residues interacting with pathogenic molecules[50]. We used HyPhy-FEL [22](Fixed Effect Likelihood), HyPhy-FUBAR[23] (Fast Unconstrained Bayesian AppRoximation), and HyPhy-MEME[24] (Mixed Effect Likelihood Model) to the codon-aligned sequence file from PAL2NAL[16]. All three methods employ a distinct algorithmic approach to infer selection. FEL uses maximum likelihood to fit a codon model to each site, thereby estimating a value for dN and dS at each site. FUBAR uses the Bayesian approach to selection inference. It uses a priori specified grid of dN and dS values (typically 20×20), spanning the range of negative, neutral, and positive selection regimes, whose likelihoods can be pre-computed and used throughout the analysis. MEME employs a mixed-effects maximum likelihood approach, where it infers two ω rate classes for each site and corresponding weights representing the probability that the site evolves under each rate class at a given branch[25].

FEL and FUBAR identified codon sites that evolved under the influence of pervasive (consistently across the entire phylogeny) diversifying and purifying selection, with p-value < 0.1 and posterior probability > 0.9, respectively. FEL predicted 8 diversifying sites and 514 purifying selected sites (**Figure 3 (a)**). Similarly, FUBAR predicted 1 diversifying selected sites and 618 purifying sites (**Figure 3 (b)**). By combining the results, we can say that THAP9 had been subjected to a strong negative selection, with one common positively selected site at 831 reported by both FEL and FUBAR. MEME identifies site(s) in a gene that is subjected to both pervasive or episodic (only on a single lineage or subset of lineages) diversifying selection[25]. 197 codon sites distributed across the hTHAP9 were observed to be evolved under episodic diversifying selection through this analysis (**Figure 3 (c)**). By combining the results from all 3 methods, we can say that THAP9 has evolved under a pervasive purifying and episodic diversifying selection, with pervasive diversifying selection occurring at individual sites.

Analysis of the *hTHAP9* protein sequence and its orthologs suggests new functional motifs in mammals.

Aiming to perform a thorough characterization of the THAP9 protein sequence, we studied its physicochemical properties and domain architecture in humans and its orthologs. *Physicochemical properties*: We predicted the Physicochemical properties using Expasy ProtParam[33] and EMBOSS Pepstats[34]. hTHAP9 protein is a 903 amino acid long protein with a Molecular weight of 103410.79. The isoelectric point (pI) of hTHAP9 is 9.4853. The instability index of hTHAP9 is 38.49. Proteins having an instability index of less than 40 are considered stable. The GRAVY value of the hTHAP9 is -0.179. The lower range of GRAVY indicates a better interaction between protein and water. The amino acid composition of hTHAP9 is shown in **Figure 4**.

Domain architecture: According to SMART[26] & Pfam[27], *hTHAP9* has an 89 residue-long THAP-type Zinc Finger domain at N-Terminal followed by a 177 residue-long P-element transposase domain (Tnp_P_element, Pfam ID: PF12017) between residues 142 to 389 (**Figure 5**). SMART also predicted two low complexity regions in the h*THAP9* protein sequence from 444 to 455 and 567 to 579. Most of the orthologs of hTHAP9 have similar SMART & PFAM annotations. Though some species lack the presence of THAP domain, for example, *Falco Peregrinus* (Aves) and *Alligator Sinensis* (Reptile), all the annotated THAP9 ortholog sequences have Tnp_P_element domain in them (**Figure 5,** Full representation is in **Supplementary Figure 1(a)**). There are no THAP9 orthologs of Amphibians and Hyperoartia class with already annotated SMART or Pfam domains. A recent study conducted by our lab also predicted an RNase-H-like fold in hTHAP9 using secondary structure predictions, homology modeling, and multiple sequence alignment. The study experimentally confirmed that D304, D374, and E613 are the catalytic residues of hTHAP9 and are responsible for DNA excision[51].

Motifs: The analysis revealed various new functional features using the Eukaryotic Linear Motif[28] (ELM) and ScanProsite[30] (**Figure 6,** full representation in **Supplementary Figure 1(b)**). We found 3 N-myristoylation sites (PS00008) in hTHAP9, which are (i) "GAilCS" located between residues 58 to 63, (ii) "GAvpSV" located between residues 83 to 88, and (iii) "GVsvTK" located between residues 679 to 684.

The first two N-myristoylation sites overlap with the N terminal THAP domain. The THAP domain is a C2CH type Zinc finger domain, involved in sequence-specific DNA binding[52]. Thus, it is possible that hTHAP9 may localize to the nucleus but could also reside temporarily in cytosolic locations that require myristoylation[53]. Myristoylation is a protein-lipid modification. Attachment of lipid moieties is essential for intracellular trafficking of proteins and for generating the final native structure of a protein. The attachment of myristic acid to the N-terminus is catalyzed by N-myristoyl transferase (NMT), a prosurvival protein, which uses myristoyl-coenzyme A (CoA) as a substrate. Although myristoylation typically occurs cotranslationally, there is evidence of NMT-mediated posttranslational modification of proteins as well[54].

Moreover, the ELM and ScanProsite analysis also predicted the occurrence of 4 adjacent motifs located between 125 to 166. The order of the domains is (i) "NYSL" N-glycosylation site (PS00001) from 125 to 128 (ii) "SIK" Protein kinase C phosphorylation site (PS00005) from 127 to 129 (iii) “LtigaekLaevqqmLqvskkrL” Leucine zipper pattern (PS00029) from 132 to 153 (iv) “KRLISVKNYR-MIKKRK" Bipartite nuclear localization signal profile (PS50079) from 151 to 166. This pattern is highly conserved in mammals (**Figure 7, Supplementary Figure 1(b)**).

NLS: A similar pattern was observed by Sheldon et.al[55] where they demonstrated that the presence of bipartite nuclear localization signal (NLS) immediately surrounding the N-terminal zipper domain required for nuclear localization in HSF2. It also suggested that the zipper domains can regulate subcellular localization by regulating the Bipartite NLS domain. More recent literature talks about two types of mechanisms for regulating NLS[56], with one involving the regulation of masking and unmasking of NLS by phosphorylation. Based on the above results, we can speculate that this highly conserved phosphorylation site & Leucine zipper domain with Bipartite NLS domain are essential for regulating the subcellular localization of THAP9.

NLS sequences generally appear either as a single-stretch or as two clusters of basic residues separated by approximately ten amino acid residues, with the respective consensus sequences of the monopartite and bipartite basic NLS's (K/R)4–6 and (K/R)2 X10–12(K/R)3[57]. ELM and ScanProsite analysis found a 16 residue long bipartite NLS protein sequence "KRLISVKNYR-MIKKRK," with 50% of the residues being Basic and the Basic region is more concentrated in the C-Terminal of the domain.

The analysis also predicted several kinase phosphorylation sites (Protein kinase C & Casein kinase II) distributed all over the protein. ELM also predicted one Host Cell Factor-1 binding motif in hTHAP9, also annotated in literature[51].

Other domains: When we looked for weak domain matches in THAP9, ScanProsite also predicted the presence of a Phosphatase tensin-type domain profile (PPASE_TENSIN). Tumor suppressor protein PTEN is the best-characterized member of the PPASE_TENSIN family[58]. To further explore the weak presence of the PPASE_TENSIN (Prosite Entry: PS51181) domain in the human THAP9 protein, we looked for this domain in all the THAP9 ortholog sequences using ScanProsite (Option 3). The PPASE_TENSIN domain was identified in 105 orthologs. This domain is already known to be present in 38 protein sequences in the Uniprot database[59], including human PTEN, a very well-studied tumor suppressor gene. To look for conserved

motifs between PPASE_TENSIN domain sequences in THAP9 ortholog sequences and the 38 protein sequences from the Uniprot, we processed them through the MEME tool part of the MEME-Suite[29] using site distribution parameter as "one occurrence per sequence." We identified 15 highly conserved motifs between these sequences, but the confidence score of these matches was not significant enough.

Besides, we found multiple other previously unreported functional motifs (**Supplementary Table 1**), which suggest tight regulatory control of *THAP9* protein stability and provide speculations towards possible functions. These potential new structural features reinforce the view that THAP9 is a multifunctional protein with many interaction partners such as DNA, Kinases, NMT.

We also characterized *THAP9* from the ortholog sequences to see the evolution of conserved motifs and functional features of THAP9 protein. These families show very high conservation within themselves (**Supplementary Figure 1(b)**).

The study of Missense mutation revealed a low concentration of pathogenic mutations in *THAP9.*

We wanted to investigate further the possibility of the role of THAP9 in disease. No disease-specific mutation has been reported in the case of the THAP9 gene yet. So, we examined the missense SNPs reported in the THAP9 gene on the Ensembl Portal[36] (which uses the 1000 Genomes project[35] data). PolyPhen-2[38] scores given on the Ensembl result table classified these SNPs into three categories: 1) score from 0 to 0.444 as “benign,” 2) scores from 0.453 to 0.906 as “possibly damaging,” 3) scores from 0.909 to 1 as “probably damaging.”

According to Ensembl data, ~79% of missense mutations were “non-pathogenic,” and ~21% of them were “pathogenic” in nature (**Table 2, Figure 8**).

Moreover, we also analyzed the missense mutations found in THAP9 in different cancer samples using cBioPortal[39] (**Table3, Figure 9**). Following the Ensembl trend,~74% of missense mutations from CBioPortal were “non-pathogenic,” and ~27% of them were “pathogenic” in nature. No significant hotspots for mutations in cancer were observed as the mutations were distributed throughout the THAP9 sequence.

The only missense SNP that has a global MAF > 0.05 (minor allele frequency) was rs897945 (MAF $= 0.3253$), making it a candidate for a derived allele (DA). In the *THAP9* gene, rs897945 yields a $G \rightarrow T$ nucleotide substitution which leads to Leucine to Phenylalanine amino acid change at position 299 located on the Tnp_P_element (Pfam ID: PF12017) domain (**Figure 10**). According to the literature, this SNP may have a role in Atopy and Allergic Rhinitis in a Singapore Chinese Population[60].

THAP9 shows an altered expression in the case of Testicular & Thymic Epithelial Tumors

To investigate the gene expression profile of hTHAP9 in various cancers, we used the GEPIA[43] tool, which used cancer and matched normal from TCGA[41] and GTEx[42] portal. The results showed overexpression of THAP9 in the case of testicular cancer and reduced expression of THAP9 in the case of Thymic Epithelial Tumors (THYM) (**Figure 11**). From this, we can speculate that THAP9 may have a role to play in cancers like its neighboring gene THAP9-AS1 and other THAP family proteins like THAP1 & 11. Moreover, it is also reported in the literature that THAP9 shows 5-fold upregulation in TBM (Tuberculous Meningitis) patients co-infected with HIV as compared to patients with TBM[61].

Discussion

Human *THAP9* (*hTHAP9*) is a transposable element-derived gene that encodes the *hTHAP9* protein, which is a homolog of Drosophila P-element transposase (*DmTNP*). In an attempt to get more insight into the possible role of *hTHAP9* we performed an extensive computational evolutionary analysis. The characterization of the *hTHAP9* protein sequence suggested several new functional features of *hTHAP9*. *THAP9* orthologs were identified in 216 organisms using the NCBI database The sequence alignment and evolutionary analysis showed high conservation among the *THAP9* orthologs. Evolutionary tests also predicted more no. of negatively selected sites compared to the few positively selected sites. Negative selection is generally observed in the Housekeeping and essential genes whereas Positive selection is mostly observed in genes associated with immunity, reproduction, and sensory systems. Moreover, we also identified cancer-associated missense mutations in *THAP9*.

The representative species birds, reptiles, amphibians, and mammals show a very similar domain architecture containing primarily THAP domain & transposase domain suggesting that these domains are important for the key functions of *THAP9*. The predicted and highly conserved PPASE_TENSIN domain in *THAP9* suggests its possible role in cancer. This and be further studied after solving the structure of the domain.

The THAP domain is known to bind DNA. Crystal structure of THAP domain binding with DNA is available of human THAP1 & Drosophila P-element transposase. In this study, we predicted the THAP domain of human *THAP9*. This model can further be used for MD simulations to study its interaction with DNA.

We also identified a cancer-related mutation in THAP which can be further studied in various cancers using gene expression analysis. GEPIA suggests altered expression of *THAP9* in the case of Testicular cancer. These results suggest the need for further investigation of the role of *THAP9* in cancers.

References

1. Su, W., Zuo, T. & Peterson, T. Ectopic Expression of a Maize Gene Is Induced by Composite Insertions Generated Through Alternative Transposition. *Genetics* **216**, 1039–1049 (2020).
2. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol.* **19**, 199 (2018).
3. Majumdar, S. & Rio, D. C. P transposable elements in Drosophila and other eukaryotic organisms. *Microbiol. Spectr.* **3**, MDNA3-0004–2014 (2015).
4. Sabogal, A., Lyubimov, A. Y., Corn, J. E., Berger, J. M. & Rio, D. C. THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves. *Nat. Struct. Mol. Biol.* **17**, 117–123 (2010).
5. Campagne, S., Saurel, O., Gervais, V. & Milon, A. Structural determinants of specific DNA-recognition by the THAP zinc finger. *Nucleic Acids Res.* **38**, 3466–3476 (2010).
6. Richter, A. *et al.* In-depth Characterization of the Homodimerization Domain of the Transcription Factor THAP1 and Dystonia-Causing Mutations Therein. *J. Mol. Neurosci.* **62**, 11–16 (2017).
7. Balakrishnan, M. P. *et al.* THAP5 is a human cardiac-specific inhibitor of cell cycle that is cleaved by the proapoptotic Omi/HtrA2 protease during cell death. *Am. J. Physiol. Heart Circ. Physiol.* **297**, H643-653 (2009).
8. Leite, K. R. M. *et al.* MicroRNA 100: a context dependent miRNA in prostate cancer. *Clinics* **68**, 797–802 (2013).
9. De Souza Santos, E., De Bessa, S. A., Netto, M. M. & Nagai, M. A. Silencing of LRRC49 and THAP10 genes by bidirectional promoter hypermethylation is a frequent event in breast cancer. *Int. J. Oncol.* **33**, 25–31 (2008).
10. Parker, J. B., Palchaudhuri, S., Yin, H., Wei, J. & Chakravarti, D. A Transcriptional Regulatory Role of the THAP11–HCF-1 Complex in Colon Cancer Cell Function. *Mol. Cell.*

Biol. **32**, 1654–1670 (2012).

11. Li, N. *et al.* lncRNA THAP9-AS1 Promotes Pancreatic Ductal Adenocarcinoma Growth and Leads to a Poor Clinical Outcome via Sponging miR-484 and Interacting with YAP. *Clin. Cancer Res.* **26**, 1736–1748 (2020).
12. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. & Clark, A. G. Recent and ongoing selection in the human genome. *Nat. Rev. Genet.* **8**, 857–868 (2007).
13. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
14. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
15. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
16. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
17. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
18. Pond, S. L. K., Poon, A. F. Y. & Frost, S. D. W. Estimating selection pressures on alignments of coding sequences. in *The Phylogenetic Handbook* (eds. Lemey, P., Salemi, M. & Vandamme, A.-M.) 419–490 (Cambridge University Press, 2009). doi:10.1017/CBO9780511819049.016.
19. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44**, W3–W10 (2016).
20. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13**, 555–556 (1997).

21. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).
22. Kosakovsky Pond, S. L. & Frost, S. D. W. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Mol. Biol. Evol.* **22**, 1208–1222 (2005).
23. Murrell, B. *et al.* FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Mol. Biol. Evol.* **30**, 1196–1205 (2013).
24. Murrell, B. *et al.* Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLOS Genet.* **8**, e1002764 (2012).
25. Spielman, S. J. *et al.* Evolution of Viral Genomes: Interplay Between Selection, Recombination, and Other Forces. in *Evolutionary Genomics: Statistical and Computational Methods* (ed. Anisimova, M.) 427–468 (Springer, 2019). doi:10.1007/978-1-4939-9074-0_14.
26. Letunic, I., Khedkar, S. & Bork, P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.* **49**, D458–D460 (2021).
27. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
28. ELM—the eukaryotic linear motif resource in 2020 | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/48/D1/D296/5611669.
29. Bailey, T. L. *et al.* MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
30. de Castro, E. *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* **34**, W362–W365 (2006).
31. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*

337, 635–645 (2004).

32. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).
33. Wilkins, M. R. *et al.* Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol. Clifton NJ* **112**, 531–552 (1999).
34. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
35. Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nat. Methods* **9**, 459–462 (2012).
36. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
37. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
38. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7.20 (2013).
39. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2**, 401–404 (2012).
40. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
41. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. Poznan Pol.* **19**, A68-77 (2015).
42. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
43. Tang, Z. *et al.* GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **45**, W98–W102 (2017).
44. Majumdar, S., Singh, A. & Rio, D. C. The Human THAP9 Gene Encodes an Active P-Element DNA Transposase. *Science* **339**, 446–448 (2013).
45. Hall, B. G. Building Phylogenetic Trees from Molecular Data with MEGA. *Mol. Biol. Evol.* **30**,

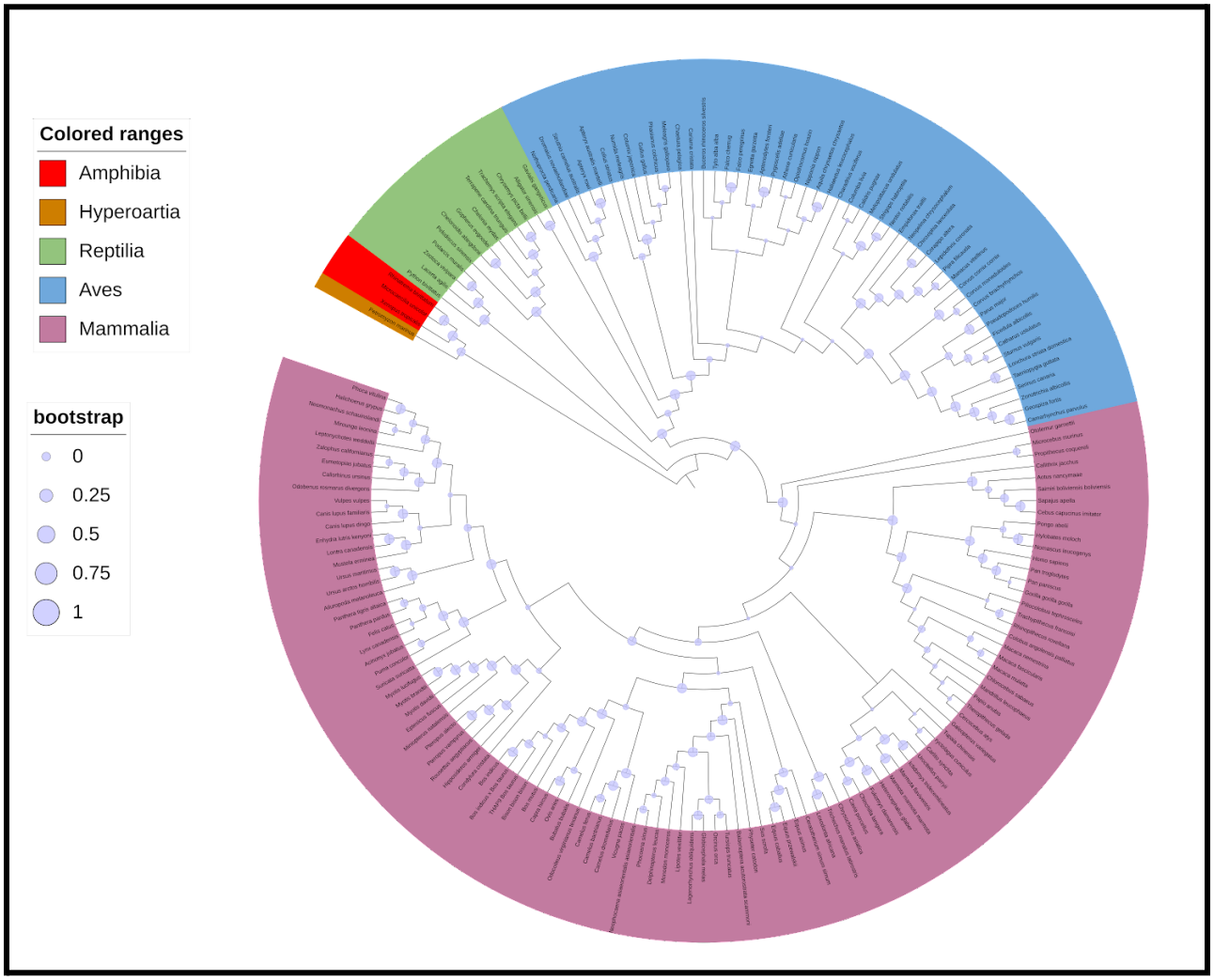
1229–1235 (2013).

46. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the genetic code: the universal enigma. *Iubmb Life* **61**, 99–111 (2009).
47. Genetic Mutation | Learn Science at Scitable. http://www.nature.com/scitable/topicpage/genetic-mutation-1127.
48. Chu, D. & Wei, L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. *BMC Cancer* **19**, 359 (2019).
49. Kryazhimskiy, S. & Plotkin, J. B. The Population Genetics of dN/dS. *PLOS Genet.* **4**, e1000304 (2008).
50. Kumar, S., Chera, J. S., Vats, A. & De, S. Nature of selection varies on different domains of IFI16-like PYHIN genes in ruminants. *BMC Evol. Biol.* **19**, 26 (2019).
51. Sharma, V., Thakore, P. & Majumdar, S. THAP9 Transposase Cleaves DNA via Conserved Acidic Residues in an RNaseH-Like Domain. *Cells* **10**, 1351 (2021).
52. Roussigne, M. *et al.* The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. *Trends Biochem. Sci.* **28**, 66–69 (2003).
53. Podell, S. & Gribskov, M. Predicting N-terminal myristoylation sites in plant proteins. *BMC Genomics* **5**, 37 (2004).
54. Udenwobele, D. I. *et al.* Myristoylation: An Important Protein Modification in the Immune Response. *Front. Immunol.* **8**, 751 (2017).
55. Sheldon, L. A. & Kingston, R. E. Hydrophobic coiled-coil domains regulate the subcellular localization of human heat shock factor 2. *Genes Dev.* **7**, 1549–1558 (1993).
56. Nagoshi, E., Imamoto, N., Sato, R. & Yoneda, Y. Nuclear Import of Sterol Regulatory Element–binding Protein-2, a Basic Helix-Loop-Helix–Leucine Zipper (bHLH-Zip)–containing Transcription Factor, Occurs through the Direct Interaction of Importin β with HLH-Zip. *Mol. Biol. Cell* **10**, 2221–2233 (1999).

57. Robbins, J., Dilwortht, S. M., Laskey, R. A. & Dingwall, C. Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: Identification of a class of bipartite nuclear targeting sequence. *Cell* **64**, 615–623 (1991).
58. PTEN regulatory functions in tumor suppression and cell biology. *Medical Science Monitor* https://www.medscimonit.com/download/index/idArt/11797.
59. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
60. Andiappan, A. K. *et al.* Genome-Wide Association Study for Atopy and Allergic Rhinitis in a Singapore Chinese Population. *PLOS ONE* **6**, e19719 (2011).
61. Kumar, G. S. S. *et al.* Gene Expression Profiling of Tuberculous Meningitis Co-infected with HIV. *J. Proteomics Bioinform.* **5**, 235–244 (2012).

Figure 1: Phylogenetic analysis of THAP9 orthologs. Phylogenetic analysis of the THAP9 orthologs perfectly aligns the Lampey, amphibians, reptiles, birds & mammals in the given order showing the evolution of THAP9 protein with the evolution of the organisms. The trees were generated using iTOL and the organisms' classes are marked in different colors with the legend given on the left side. **(a) Phylogenetic tree of aligned protein sequences of THAP9 orthologs.** Protein sequences were aligned using CLUSTALW, followed by building Maximum likelihood Phylogenetic trees with 100 bootstrap replicates. Bootstrap values are marked on the tree and the legend is given on the left side. **(b) Phylogenetic tree of aligned coding DNA sequences of THAP9 orthologs.** Coding DNA sequences of the THAP9 orthologs were codon aligned using PAL2NAL, which also takes corresponding aligned protein sequences as input. Finally, the Maximum likelihood Phylogenetic trees were generated using MEGAX with 100 bootstrap replicates. Bootstrap values are marked on the tree and the legend is given on the left side.

(a)



(b)

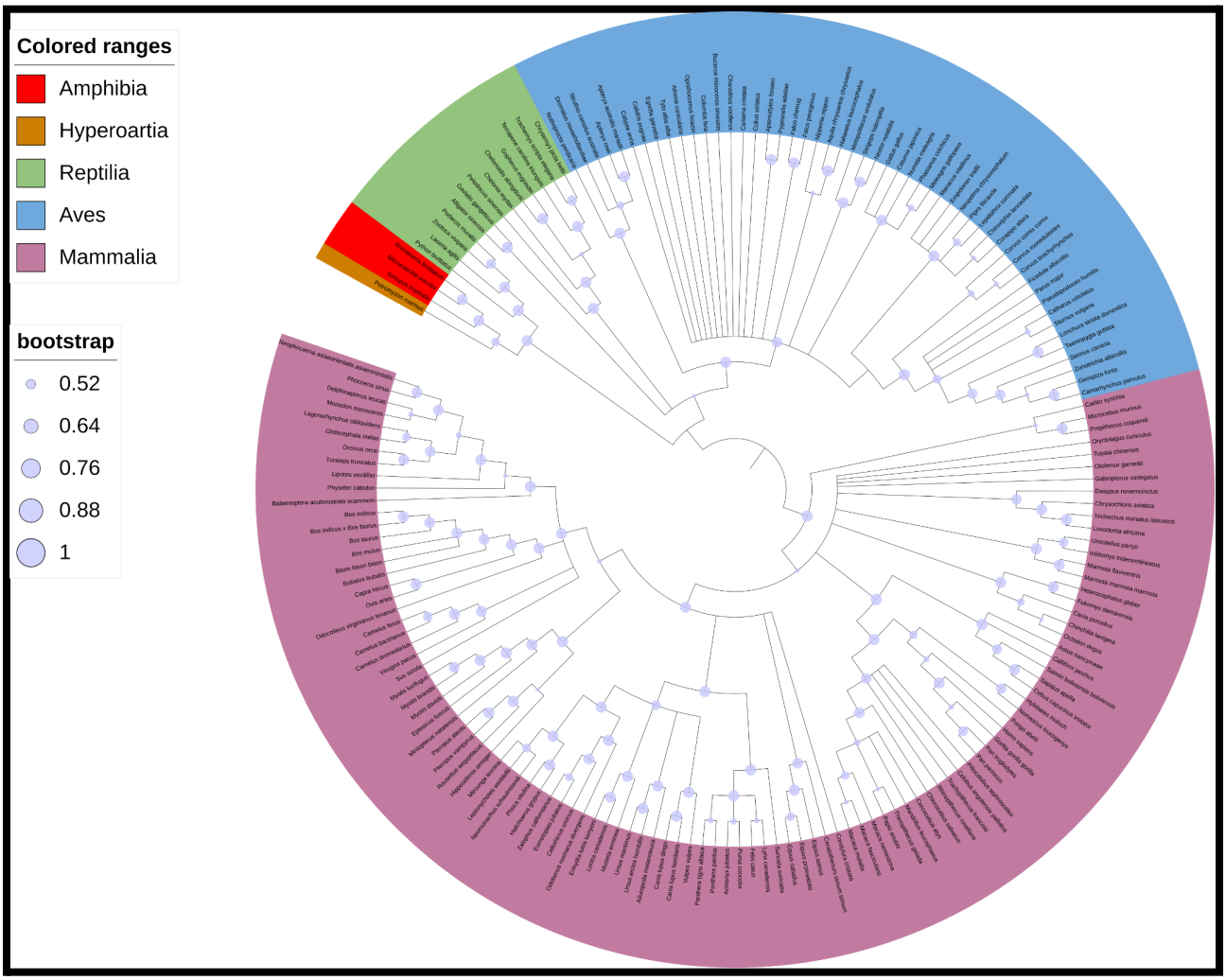

Colored ranges
Amphibia
Hyperoartia
Reptilia
Aves
Mammalia
bootstrap
0.52
0.64
0.76
0.88
1

Figure 2: Top 30 MEME Motifs from THAP9 Ortholog Proteins. Schematic distribution of top 30 conserved motifs predicted in protein sequences of THAP9 orthologs. Motif analysis of 173 sequences was performed using MEME-suite. The phylogenetic tree was constructed using the Maximum likelihood method available in MEGAX. The trees were generated using iTOL and the organisms' classes are marked in different colors with the legend given on the left side.

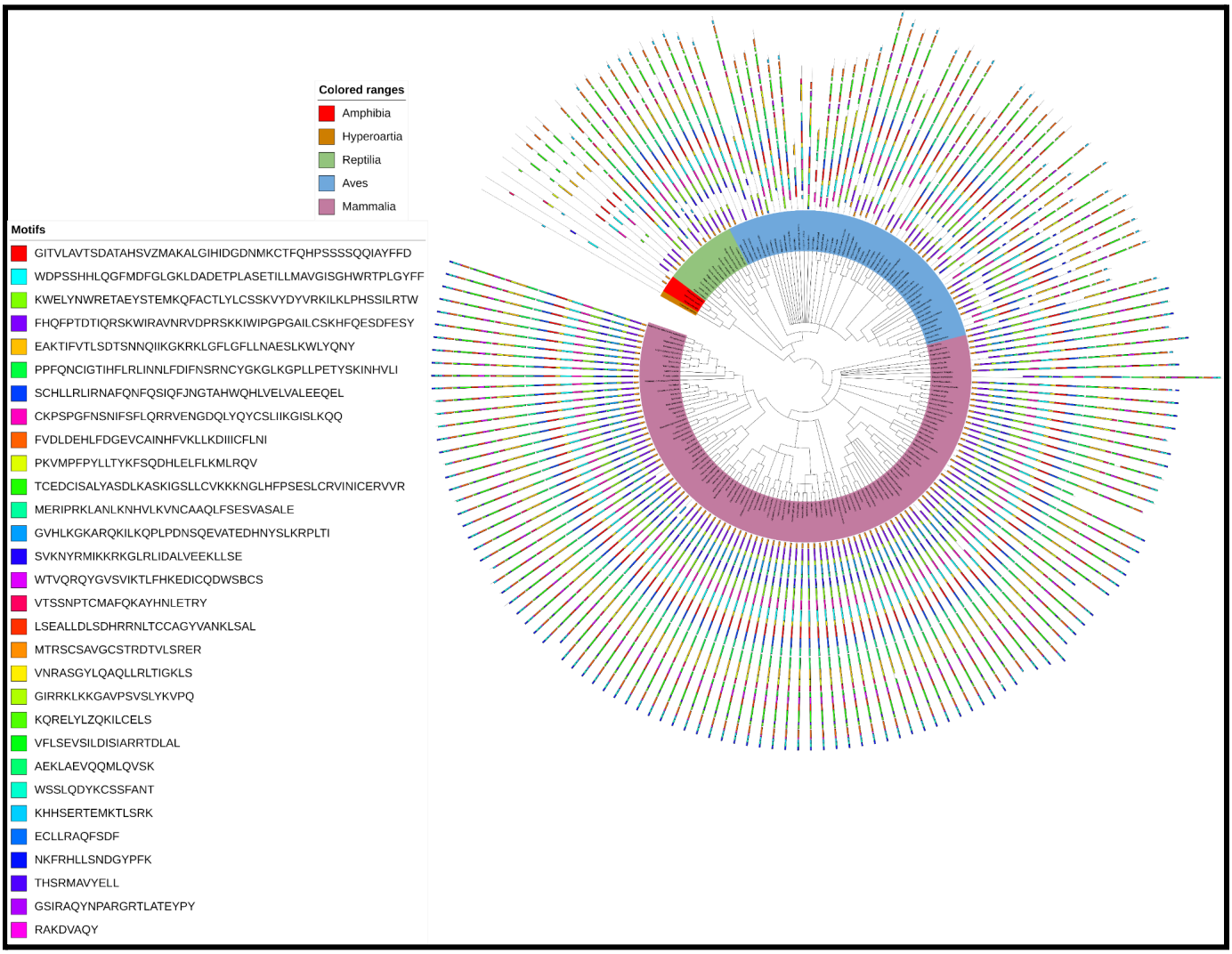


Figure 3: Analysis of Selective Pressure Applied on THAP9 gene using. Three independent analyses were performed using HyPhy-FEL (Fixed Effect Likelihood), HyPhy-FUBAR (Fast Unconstrained Bayesian AppRoximation), and HyPhy-MEME (Mixed Effect Likelihood Model) packages. The HyPhy package was run on the Galaxy webserver. FEL is an ML-based model which assumes constant selection pressure on each site. The FUBAR model estimates site-by-site ω values while MEME allows the distribution of ω to vary over sites and additionally from branch to branch. Significance was assessed by posterior probability > 0.9 (FUBAR) and p-value < 0.1 (MEME) and p-value < 0.1 (FEL).**(a)** FEL predicted 8 diversifying sites and 514 purifying selected sites. **(b)** FUBAR predicted 1 diversifying selected site and 618 purifying sites. **(c)** MEME observed 197 codon sites, distributed across the hTHAP9, to be evolved under episodic diversifying selection. By combining the results from all 3 methods, we can say that THAP9 has evolved under a pervasive purifying and episodic diversifying selection, with one common episodic diversifying selection site at 831.

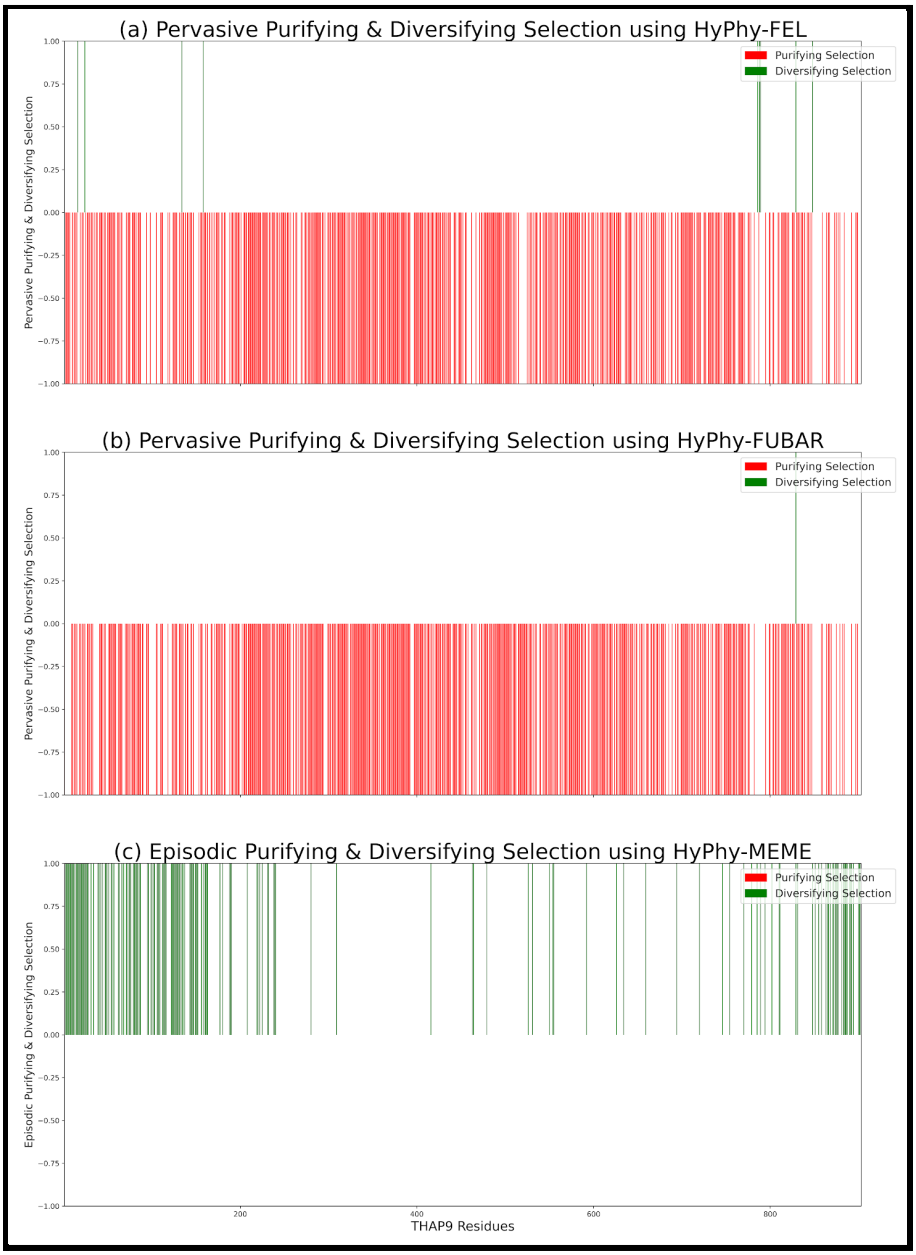


Figure 4: Amino Acid composition of human THAP9 protein.

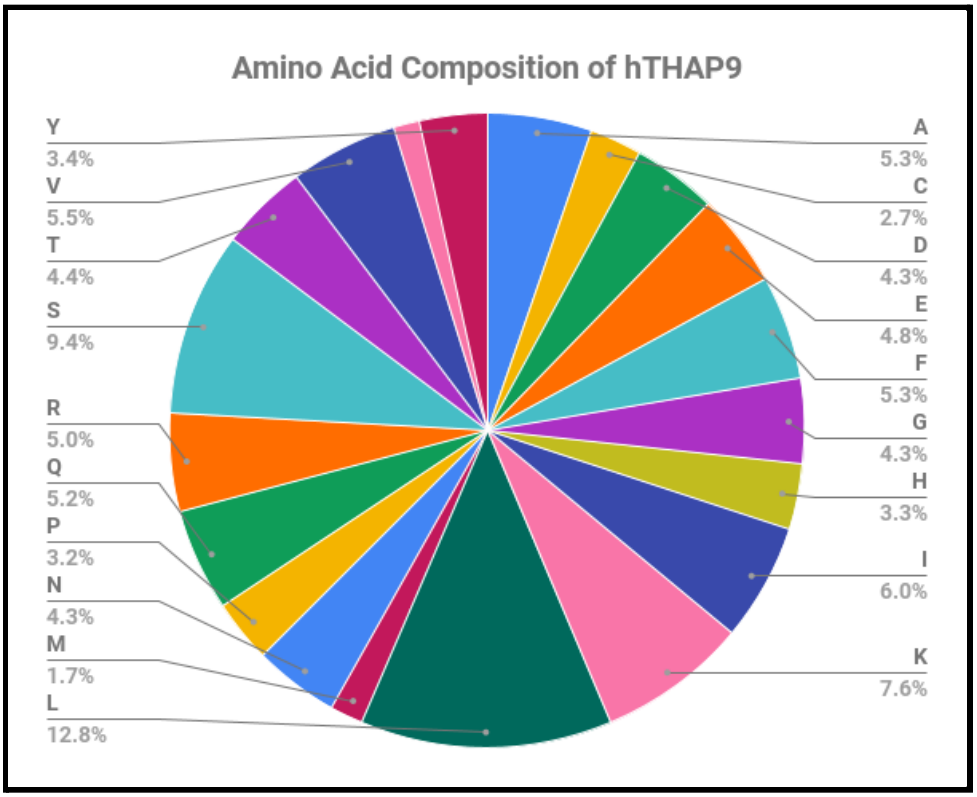


Figure 5: Previously annotated SMART & Pfam domains of human THAP9 and its orthologs. SMART was queried for THAP9 in "batch mode" including the Pfam domains. Results were visualized on a phylogenetic tree using iTOL. Displaying the resulting domain architectures in iTOL was performed using the data export feature of SMART which allows users to simply display domain architecture of results on a phylogenetic tree. Most of the orthologs of hTHAP9 have similar SMART & PFAM annotations. Though some species lack the presence of THAP domain, for example, Falco Peregrinus (Aves) and Alligator Sinensis (Reptile), all the annotated THAP9 ortholog sequences have Tnp_P_element domain in them (Full tree representation is in **Supplementary Figure 1 (a)**). There are no THAP9 orthologs of Amphibians and Hyperoartia class with already annotated SMART or Pfam domains.

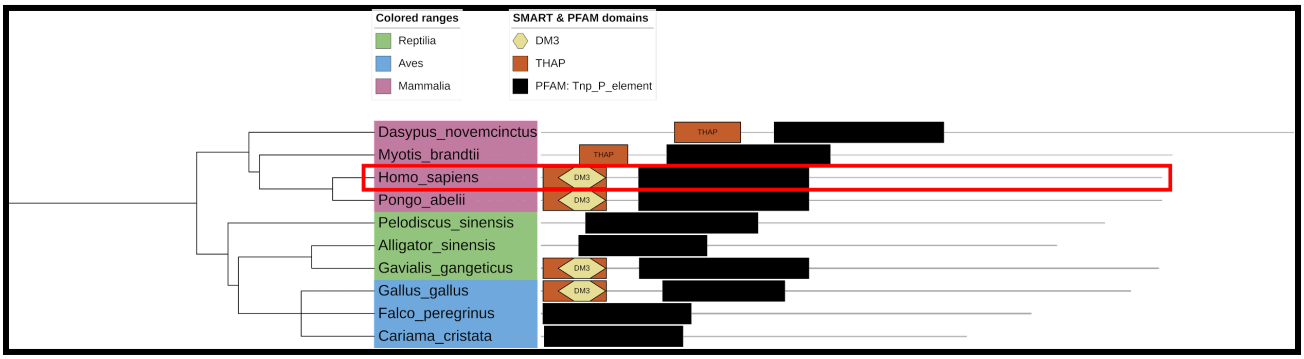


Figure 6: New predicted domains of human THAP9 and its orthologs. The analysis revealed various new functional features using the Eukaryotic Linear Motif (ELM) and ScanProsite (full representation in **supplementary Figure 1 (b)**).

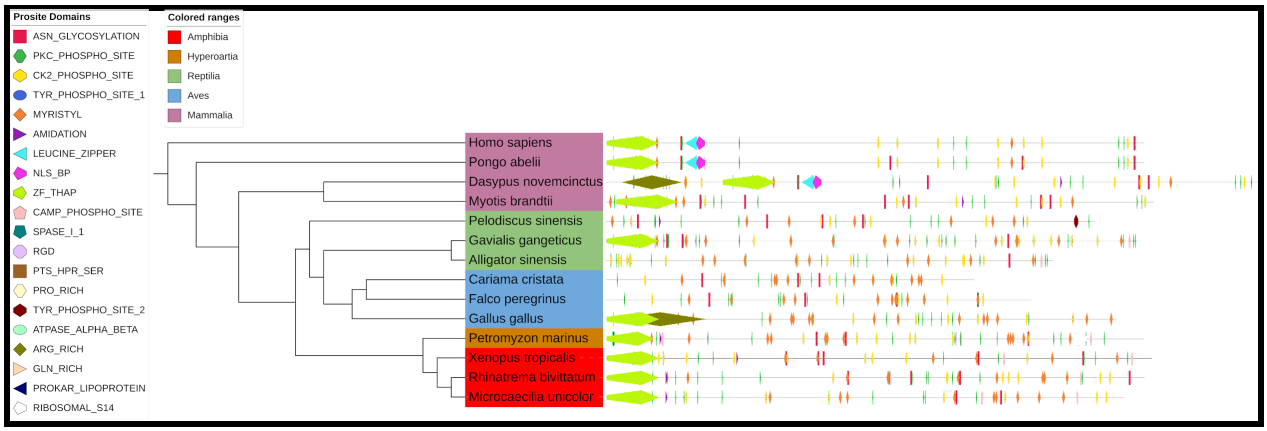


Figure 7: Architecture of N terminal (1-175) region hTHAP9. We found 3 N-myristoylation sites (PS00008) in hTHAP9, which are (i) "GAilCS" located between residues 58 to 63, (ii) "GAvpSV" located between residues 83 to 88, and (iii) "GVsvTK" located between residues 679 to 684. Moreover, the ELM and ScanProsite analysis also predicted the occurrence of 4 adjacent motifs located between 125 to 166. The order of the domains is (i) "NYSL" N-glycosylation site (PS00001) from 125 to 128 (ii) "SIK" Protein kinase C phosphorylation site (PS00005) from 127 to 129 (iii) “LtigaekLaevqqmLqvskkrL” Leucine zipper pattern (PS00029) from 132 to 153 (iv) “KRLISVKNYR-MIKKRK" Bipartite nuclear localization signal profile (PS50079) from 151 to 166. This pattern is highly conserved in mammals (**supplementary figure 1 (b)**)

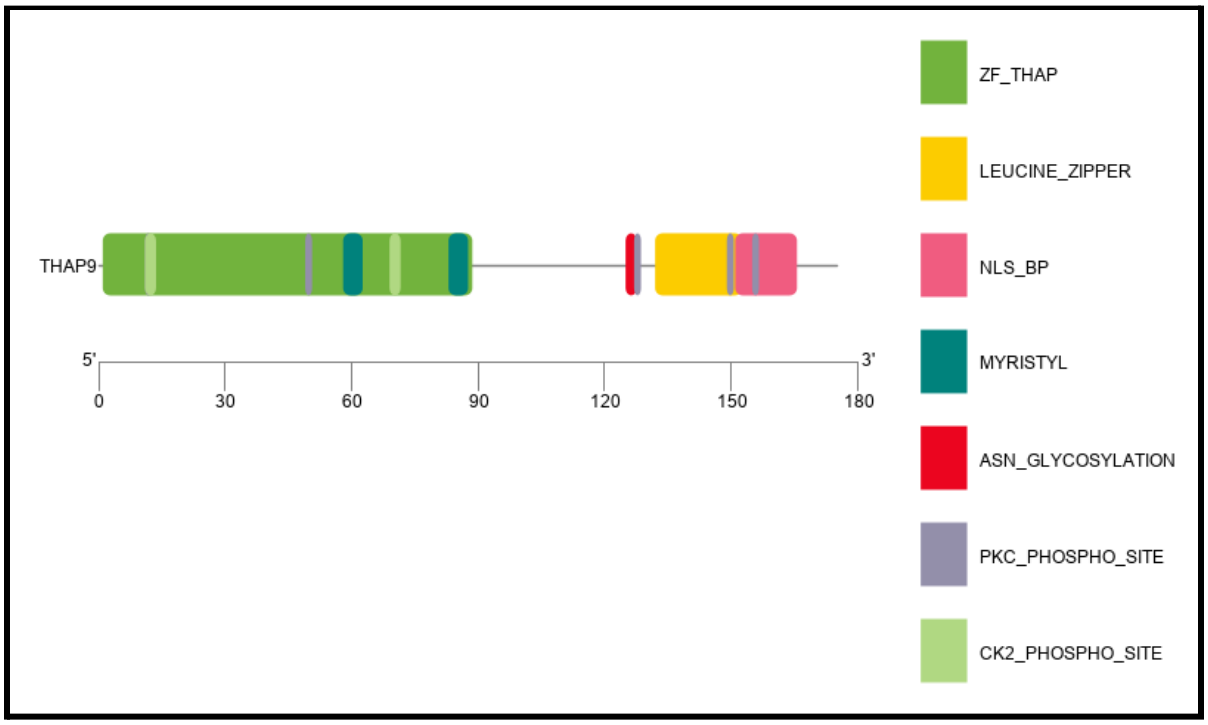


Figure 8: PolyPhen classification for Missense Mutations from Ensembl. PolyPhen classification of 820 Missense SNPs of human THAP9 gene from Ensembl into benign, possibly- damaging, and probably damaging categories.

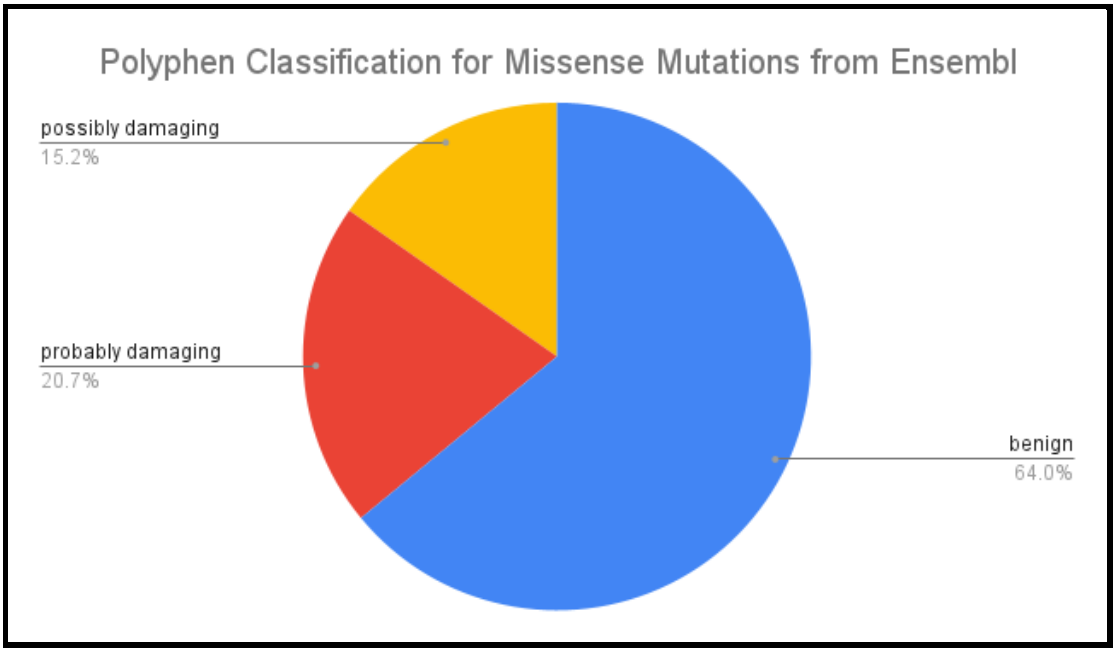


Figure 9: PolyPhen classification for Missense Mutations from CBio Portal. PolyPhen classification of 272 Missense SNPs of human THAP9 gene from cBio Portal into benign, possibly- damaging, and probably damaging categories.

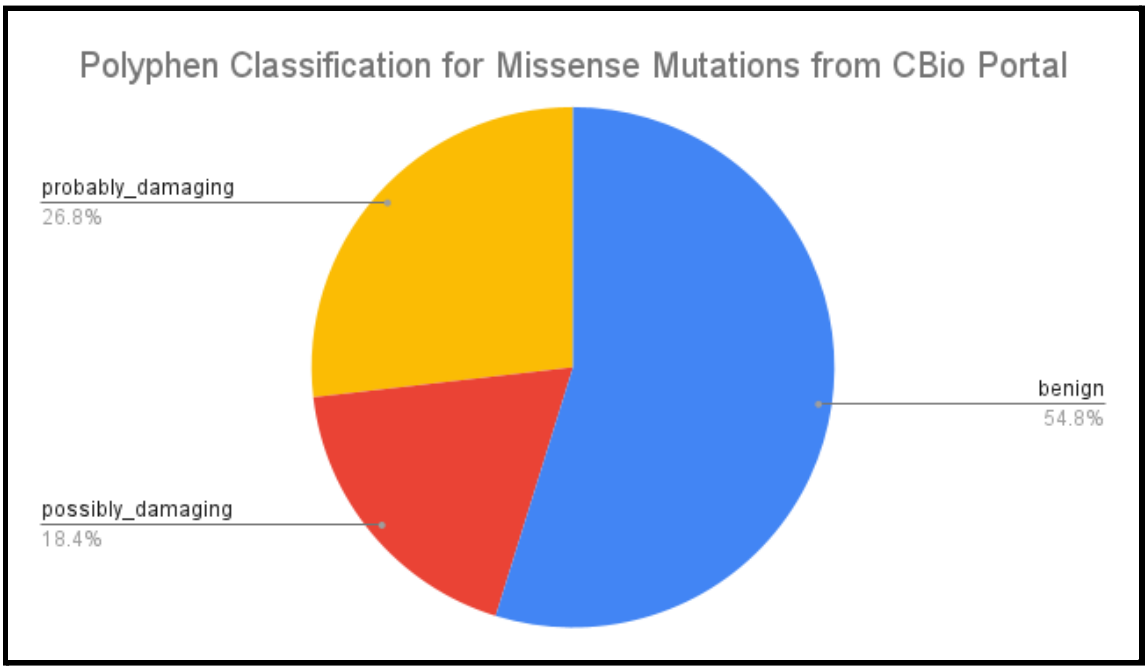


Figure 10: Genome Variation Viewer view of rs897945. rs897945 yields a G→T nucleotide substitution which leads to Leucine to Phenylalanine amino acid change at position 299 located on the Tnp_P_element (Pfam ID: PF12017) domain in hTHAP9 protein.

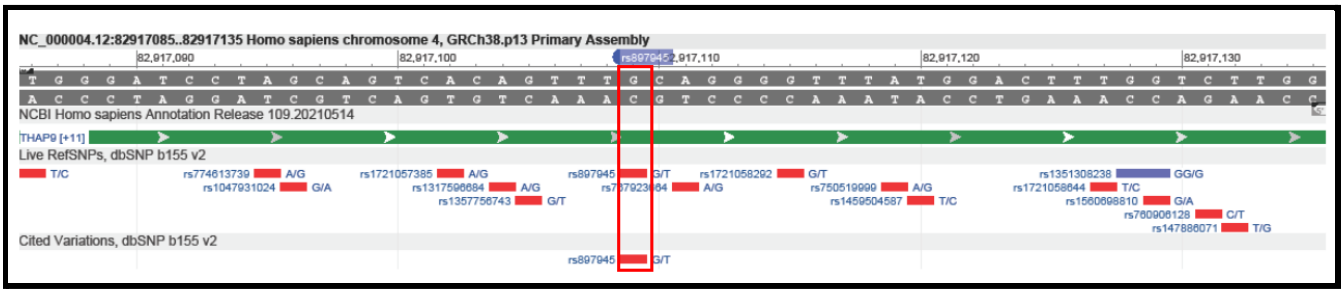
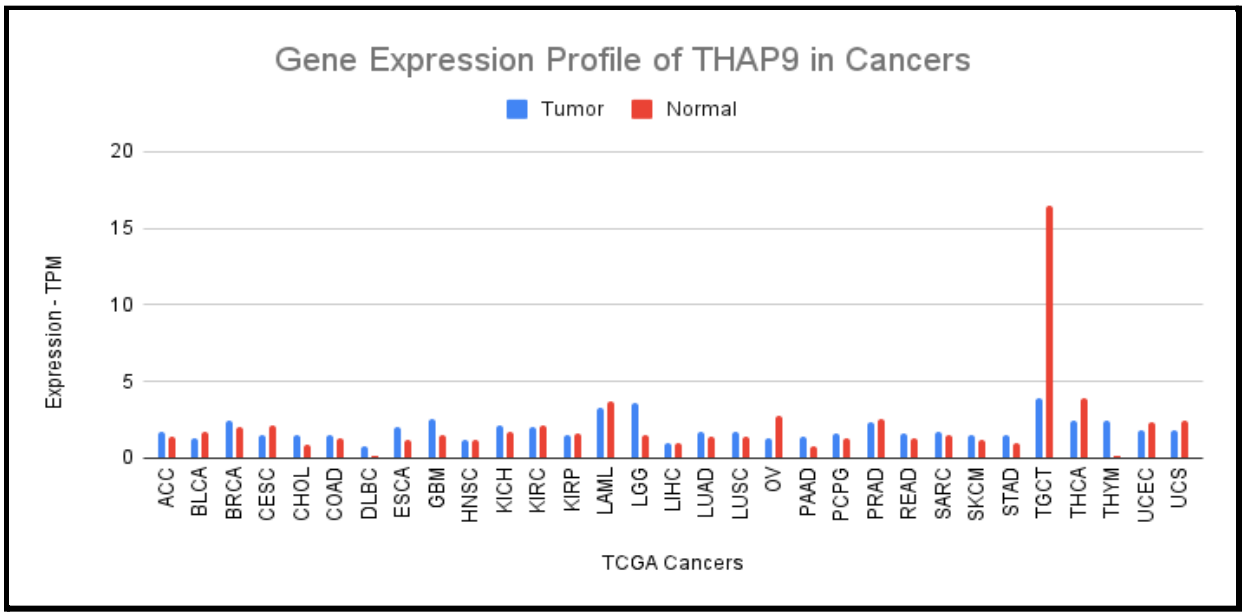
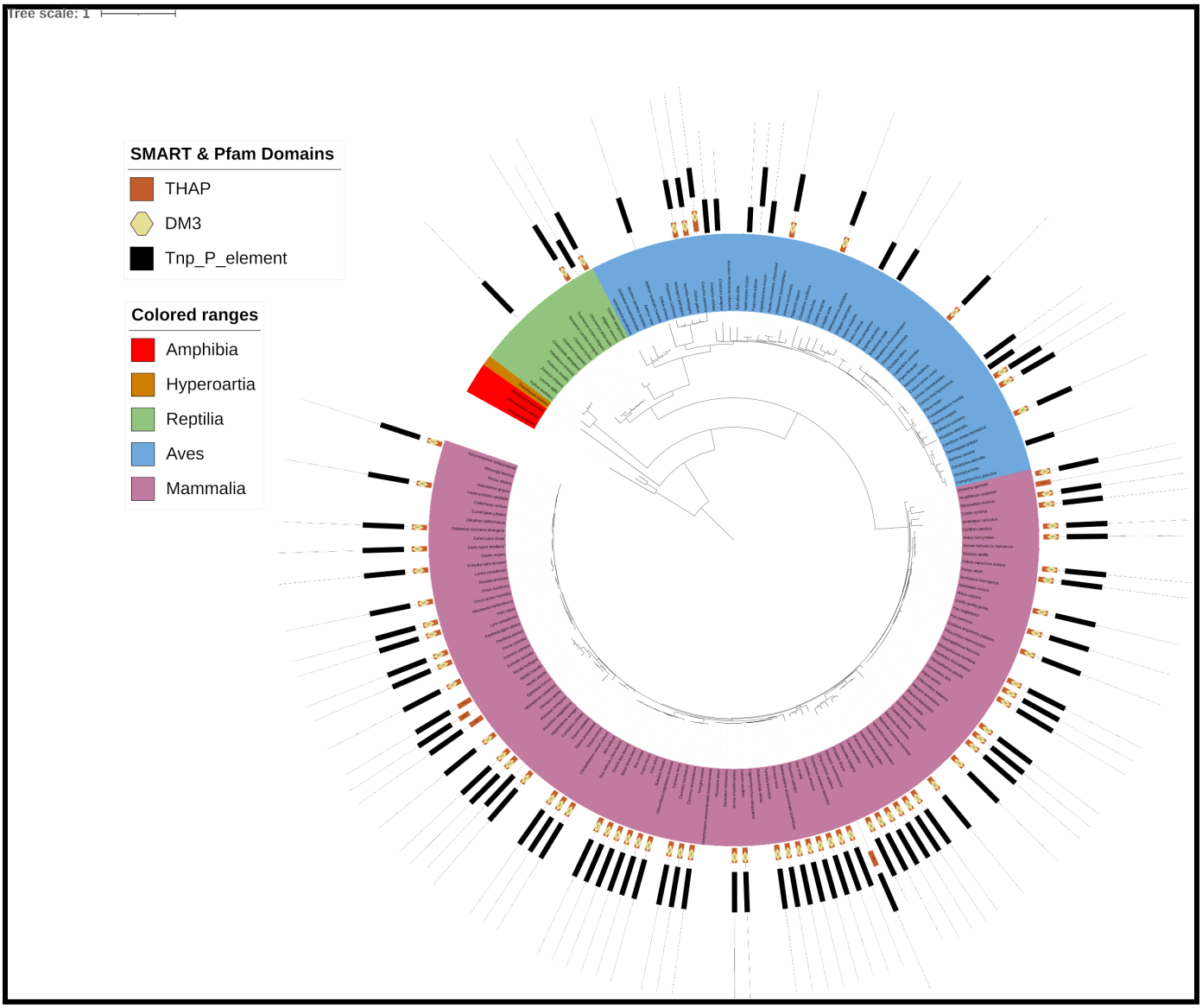


Figure 11: Gene Expression Profile of THAP9 in Cancers. The pan-cancer expression profile of THAP9 has been plotted using the GEPIA tool, which used cancer and matched normal from TCGA and GTEx portal. Pan-cancer view of THAP9 expression showed overexpression of THAP9 in Testicular Cancer (TGCT) and reduced expression of THAP9 in Thymic Epithelial Tumors (THYM). ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical and endocervical cancers; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LAML, acute myeloid leukemia; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; STES, stomach and esophageal carcinoma; TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma; UVM, uveal melanoma.



Supplementary Figure 1: (a) Full tree representation of previously annotated SMART & Pfam domains of human THAP9 and its orthologs. SMART was queried for THAP9 in "batch mode" including the Pfam domains. Results were visualized on a phylogenetic tree using iTOL. Displaying the resulting domain architectures in iTOL was performed using the data export feature of SMART which allows users to simply display domain architecture of results on a phylogenetic tree. Most of the orthologs of hTHAP9 have similar SMART & PFAM annotations. Some species lack the THAP domain but all the annotated THAP9 ortholog sequences have the Tnp_P_element domain in them. There are no THAP9 orthologs of Amphibians and Hyperoartia class with already annotated SMART or Pfam domains. **(b) Full tree representation of new predicted domains of human THAP9 and its orthologs.** The analysis revealed various new functional features using the Eukaryotic Linear Motif (ELM) and ScanProsite.

(a)



(b)

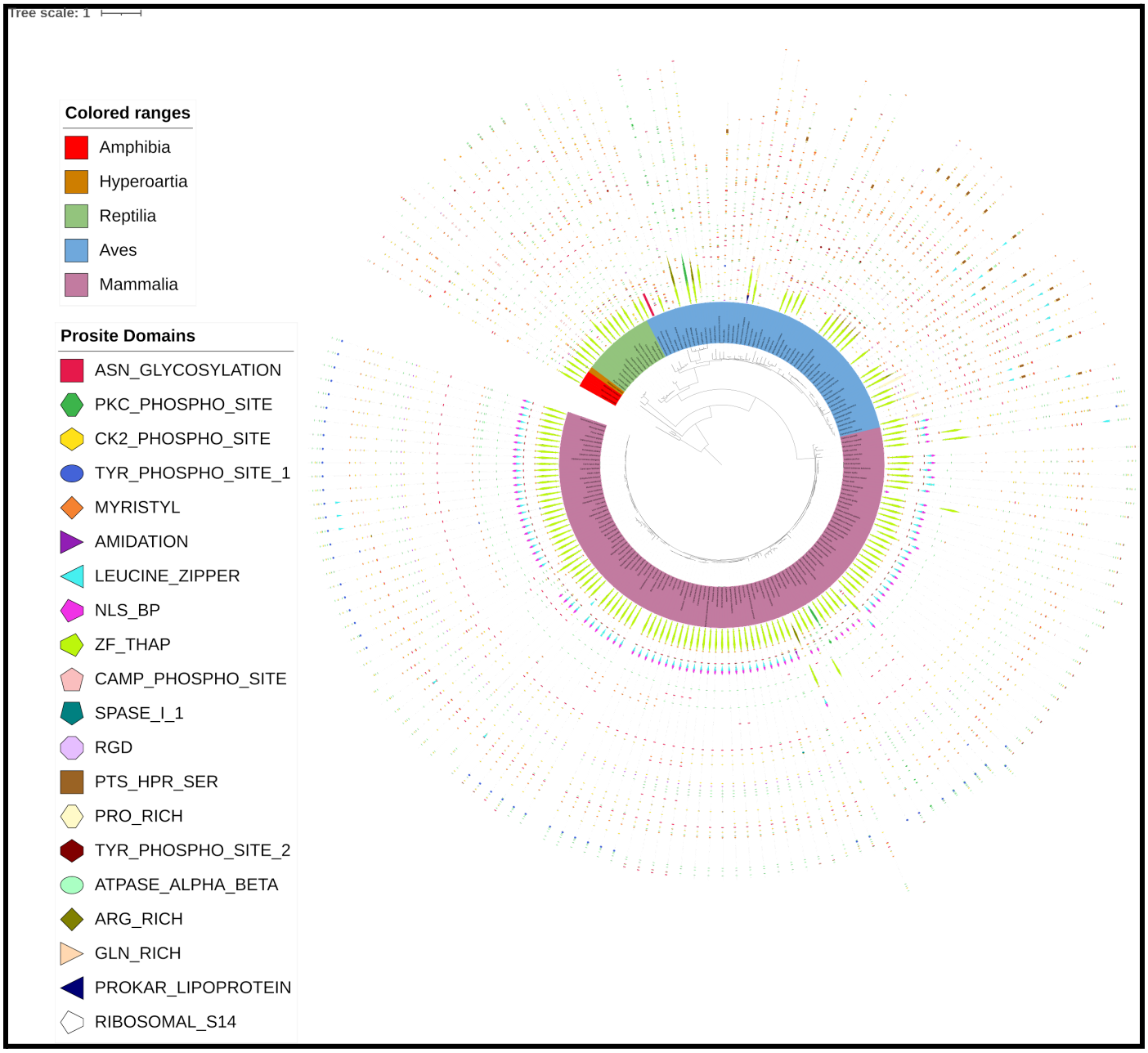

Tree scale: 1
Colored ranges
Amphibia
Hyperoartia
Reptilia
Aves
Mammalia
Prosite Domains
ASN_GLYCOSYLATION
PKC_PHOSPHO_SITE
CK2_PHOSPHO_SITE
TYR_PHOSPHO_SITE_1
MYRISTYL
AMIDATION
LEUCINE_ZIPPER
NLS_BP
ZF_THAP
CAMP_PHOSPHO_SITE
SPASE_I_1
RGD
PTS_HPR_SER
PRO_RICH
TYR_PHOSPHO_SITE_2
ATPASE_ALPHA_BETA
ARG_RICH
GLN_RICH
PROKAR_LIPOPROTEIN
RIBOSOMAL_S14

Table 1: Table of Top 15 MEME Motifs from THAP9 Ortholog Proteins

Sr. No.	Motif (Regular Expression)	No. of sites
1	GITVLAVTS[DG]ATAH[SG][VA][QE][MT]A[KR]ALGI[HR]ID[GP]D[NR][MI][KQ]CTFQHP[SP][SG]S[SA][QH][QS]IAYFFD	173
2	WDP[SQ][ST][HQ][HRS]L[QT]GF[MV]D[FL]G[LA]G[KI]LDADE[TA]PLASE[TA][IV][LI]LMAVGI[SF][GS][HP]W[RT][TA]PLGYFF	173
3	[KP]WEL[YH][NS]WR[EQ][TM]AEYS[TP]EM[KR]QFACTL[YH]L[CY][SH]SK[VA]YDY[VL]RKIL[KP]LPH[SP]S[IS]L[RT][TN]W	155
4	FHQFPTDTIQRSKWIRAVNRVDPRSKKIWIPGPGA[IM]LCS[KR]HF[QA]ESDFESY	144
5	EAK[TS]IFVTL[ST]D[TS]S[NG][NR][QY]IIK[GS]KRKLGFL[GS]FLLNAESLKWLYQNY	172
6	[PA][PS]FQNC[IS]GT[IV][HK]F[LV]RL[IM][NS]NL[FC]D[IV]F[NH][SG]RN[CP]YGKGLKGPLL[PA]E[TN][YF][SN]KIN[HR][VL][LF]I	152
7	SCH[LA]L[RQ]LIRNA[FL]Q[NC]FQ[SK]I[QE][FW][IL]N[GD]TAHWQH[LV]VEL[VA]AL[EGQ]EQ[ER][LV]	168
8	[CN][KE][PA][SA][PA]GF[NS][SN][ND][IV]F[SL]FLQ[REQ][RK]VE[NR]G[DE]Q[LA]YQYC[SA]L[IML][IV][KQ][GD][IMV][SP]L[KQ][QK]Q	171
9	F[VP][DT]L[DNT][EKN]HLFDGE[VL]C[AI][IN]NH[FY][VT]KL[LV]K[DE]I[IT][IR]C[FY]L[NK]I	165
10	P[KE][VG][MT]P[FS][PH][YH]LLTY[KA]FS[QL]D[HP]LELFL[KR][MA]L[RQ]Q[VA]	168
11	[TQ]CE[DA]C[IL][SAT][AS]L[YF][AE]SD[LE][KS][AR][SL][KR][IC]GS[LV]L[CY][VI]KK[KL][NG]G[LV][HS][FL]PS[EA]S[LV][CH][RQH][VI]I[NS]I[CS]E[RQ]V[VL]R	167
12	[MC][ES][RP][IL][PS][RGS][KR]L[AG][NS]L[KE][NS][HY][VH]LK[VM]NCA[AT]QLFSE[SG]VA[SD]ALE	164
13	GVHLKGKARQKILKQPLPDNSQEVATEDHNYSLK[RT]PLTI	107
14	S[VA]KNYRMIKKRKGLRLIDALVEEKLLSE	136
15	WTVQRQYGVSV[IT]KTLFH[KE]E[DG]ICQDWS[ND]CS	106

Table 2: Distribution of Missense Mutations from Ensembl

Missense Mutations from Ensembl		
Polyphen Mutation Class	**Number of Mutations**	**Percentage**
Benign	525	64.03
Possibly Damaging	125	15.24
Probably Damaging	170	20.73
Total	820	

Table 3: Distribution of Missense Mutations from CBio Portal.

Missense Mutations from CbioPortal		
Polyphen Mutation Class	**Number of Mutations**	**Percentage**
Benign	149	54.8
Possibly Damaging	50	18.4
Probably Damaging	73	26.8
Total	272	

Supplementary Table 1: Previously unreported functional motifs of hTHAP9from ELM.

Elm Name	Positions	Elm Description	Cell Compartment	Pattern	Probability
LIG_HCF-1_HBM_1	123-126	The DHxY Host Cell Factor-1 binding motif (HBM) interacts with the N-terminal kelch propeller domain of the cell cycle regulator HCF-1	nucleus, Transcription factor complex,	[DE]H.Y	

			viral tegument		
CLV_NRD_NRD_1	164-166 [A] 230-232 [A] 265-267 [A] 665-667 [A] 712-714 [A] 869-871 [A]	N-Arg dibasic convertase (NRD/Nardilysin) cleavage site (X-|-R-K or R-|-R-X).	extracellular, Golgi apparatus, cell surface	(.RK)|(RR[^KR])	7.47E-03
CLV_PCSK_KEX2_1	151-153 [A] 164-166 [A] 265-267 [A] 665-667 [A] 712-714 [A]	Yeast kexin 2 cleavage site (K-R-|-X or R-R-|-X).	extracellular, Golgi apparatus	[KR]R.	7.97E-03
CLV_PCSK_PC1ET2_1	151-153 [A] 164-166 [A]	NEC1/NEC2 cleavage site (K-R-|-X).	extracellular, Golgi apparatus, Golgi membrane	KR.	3.90E-03
CLV_PCSK_PC7_1	160-166 [A]	Proprotein convertase 7 (PC7, PCSK7) cleavage site (R-X-X-X-[RK]-R-|-X).	extracellular, Golgi apparatus, Golgi membrane	R...[KR]R.	5.09E-04
CLV_PCSK_SKI1_1	106-110 [A] 160-164 [A] 231-235 [A] 232-236 [A] 423-427 [A] 529-533 [A] 551-555 [A] 620-624 [A] 684-688 [A]	Subtilisin/kexin isozyme-1 (SKI1) cleavage site ([RK]-X-[hydrophobic]-[LTKF]-|-X).	endoplasmic reticulum lumen, endoplasmic reticulum,	[RK].[AILMFV][LTKF].	6.82E-03

	819-823 [A] 834-838 [A] 892-896 [A]		Golgi apparatus, extracellular		
DEG_APCC_DBOX_1	164-172 [A]	An RxxL-based motif that binds to the Cdh1 and Cdc20 components of APC/C thereby targeting the protein for destruction in a cell cycle dependent manner	nucleus, cytosol	.R..L..[LIVM].	7.68E-04
DEG_SPOP_SBC_1	402-406 [A] 624-628 [A] 885-889 [A]	The S/T rich motif known as the SPOP-binding consensus (SBC) of the MATH-BTB protein, SPOP, is present in substrates that undergo SPOP/Cul3-dependant ubiquitination.	nuclear speck, nucleus, Cul3-RING ubiquitin ligase complex	[AVP].[ST][ST][ST]	9.38E-04
DOC_CKS1_1	128-133 [A] 333-338 [A]	Phospho-dependent motif that mediates docking of CDK substrates and regulators to cyclin-CDK-bound Cks1.	cytosol, nucleus	[MPVLIFWYQ].(T)P..	1.99E-03
DOC_CYCLIN_RxL_1	160-172 [A] 163-172 [A] 229-240 [A] 816-825 [A] 889-898 [A]	Both fungal and mammalian S-phase Cyclin/CDK complexes recognize specific RxL docking motifs in their target proteins.	nucleus, membrane, cytosol, centriole, Cajal body	(.|([KRH].{0,3}))[^EDWNSG][^D][RK][^D]L.{0,1}[FLMP].{0,3}[EDST]	4.21E-03

DOC_CYCLIN_yCln2_LP_2	128-134 [A] 531-537 [A]	The budding yeast G1/S cyclins Cln1 and 2 bind a specific leucine- and proline-rich (LP) docking motif on G1-specific target proteins.	cytosol, nucleus	(L[MLIV]PP)|(((L[LMIV]PA)|(L..P))[ILMVAFYW].[MLIVFHPAY])	1.07E-04
DOC_MAPK_gen_1	150-156 [A] 160-170 [A] 163-170 [A] 164-170 [A] 567-574 [A] 569-577 [A] 644-653 [A] 665-671 [A] 745-755 [A] 756-763 [A] 757-763 [A] 777-787 [A] 834-843 [A]	MAPK interacting molecules (e.g. MAPKKs, substrates, phosphatases) carry docking motif that help to regulate specific interaction in the MAPK cascade. The classic motif approximates (R/K)xxxx#x# where # is a hydrophobic residue.	nucleus, cytosol	[KR]{0,2}[KR].{0,2}[KR].{2,4}[ILVM].[ILVF]	4.32E-03
DOC_MAPK_MEF2A_6	360-368 [A] 834-843 [A]	A kinase docking motif that mediates interaction towards the ERK1/2 and p38 subfamilies of MAP kinases.	nucleus, cytosol, Transcription factor complex	[RK].{2,4}[LIVMP].[LIV].[LIVMF]	2.58E-03
DOC_MAPK_RevD_3	651-667 [A]	Reverse (C to N direction) of the classical MAPK docking motif ELM:DOC_MAPK_gen_1 with an often extended linker region of the bipartite motif.	cytosol, nucleus	[LIVMPFA].[LIV].{1,2}[LIVMP].{4,6}[LIV]..[RK][RK]	1.81E-04

DOC_PP1_RVXF_1	569-575 [A]	Protein phosphatase 1 catalytic subunit (PP1c) interacting motif binds targeting proteins that dock to the substrate for dephosphorylation. The motif defined is [RK]{0,1}[VI][^P][FW].	nucleus, protein phosphatase type 1 complex, cytosol	..[RK].{0,1}[VIL][^P][FW].	8.30E-04
DOC_PP1_SILK_1	239-244 [A]	Protein phosphatase 1 catalytic subunit (PP1c) interacting motif that often cooperates with and is located N-terminal to the RVXF motif to dock proteins to PP1c.	nucleus, protein phosphatase type 1 complex, cytosol	.[GS]IL[KR][^DE]	3.78E-05
DOC_PP2A_B56_1	199-205 [A] 447-453 [A]	Docking site required for the regulatory subunit B56 of PP2A for protein dephosphorylation.	nucleus, kinetochore, chromosome, centromeric region, cytosol	([LMFYWIC]..I.E)|(L..[IVLWC].E).	1.46E-03
DOC_PP2B_LxvP_1	234-237 [A] 496-499 [A]	Docking motif in calcineurin substrates that binds at the interface of the catalytic CNA and regulatory CNB subunits.	cytosol, calcineurin complex, nucleus	L.[LIVAPM]P	2.30E-03

DOC_PP4_FxxP_1	399-402 [A]	The FxxP-like docking motif recognized by the EVH1 domains of the PPP4R3 regulatory subunits of the PP4 holoenzyme	nucleus, nucleoplasm	F..P	1.73E-03
DOC_USP7_MATH_1	113-117 [A] 237-241 [A] 316-320 [A] 370-374 [A] 402-406 [A] 764-768 [A] 885-889 [A]	The USP7 MATH domain binding motif variant based on the MDM2 and p53 interactions.	nucleus	[PA][^P][^FYWIL]S[^P]	1.24E-02
DOC_USP7_UBL2_3	102-106 [A] 567-571 [A]	The USP7 CTD domain binding motif variant based on the ICP0 and DNMT1 interactions	nucleus	K...K	3.74E-03
DOC_WW_Pin1_4	127-132 [A] 249-254 [A] 312-317 [A] 332-337 [A] 625-630 [A]	The Class IV WW domain interaction motif is recognised primarily by the Pin1 phosphorylation-dependent prolyl isomerase.	cytosol, nucleus	...([ST])P.	1.54E-02
LIG_14-3-3_CanoR_1	152-157 [A] 189-199 [A] 460-466 [A] 777-787 [A] 870-879 [A]	Canonical Arg-containing phospho-motif mediating a strong interaction with 14-3-3 proteins.	nucleus, internal side of plasma membrane, cytosol	R[^DE]{0,2}[^DEPG]([ST])(([FWYLMV].)|([^PRIKGN]P)|([^PRIKGN].{2,4}[VILMFWYP]))	4.48E-03

LIG_Actin_WH2_2	498-514 [A] 732-747 [A] 822-839 [A]	The WH2 motif is of variable length (16-19 amino acids) binding to the hydrophobic cleft formed by actin's subdomains 1 and 3. At the N-terminus it forms an alpha-helix followed by a flexible loop stabilised upon actin binding.	cytosol	[^R]..((.[ILMVF])|([ILMVF].))[^P][^P][ILVM].{4,7}L(([KR].)|(NK))[VATIGS]	6.60E-04
LIG_APCC_ABBA_1	386-391 [A] 811-816 [A]	Amphipathic motif that is involved in APC/C inhibition by binding of CDH1/CDC20. In metazoan cyclin A, the motif also acts as a degron, enabling the cyclin's degradation in prometaphase.	spindle pole, nucleus, cytosol	[ILVMF].[ILMVP][FHY].[DE]	3.84E-04
LIG_BRCT_BRCA1_1	251-255 [A] 256-260 [A] 430-434 [A] 750-754 [A] 887-891 [A]	Phosphopeptide motif which directly interacts with the BRCT (carboxy-terminal) domain of the Breast Cancer Gene BRCA1 with low affinity	nucleus, BRCA1-BARD1 complex	.(S)..F	1.91E-03
LIG_BRCT_BRCA1_2	750-756 [A]	Phosphopeptide motif which directly interacts with the BRCT (carboxy-terminal) domain of the Breast Cancer Gene BRCA1 with high affinity.	nucleus, BRCA1-BARD1 complex	.(S)..F.K	1.17E-04
LIG_Clathr_ClatBox_1	447-451 [A] 493-497 [A]	Clathrin box motif found on cargo adaptor proteins, it interacts with the beta propeller structure located at the N-terminus of Clathrin heavy chain.	cytosol, Golgi apparatus, cytoskeleton,	L[IVLMF].[IVLMF][DE]	3.41E-04

			clathrin-coated endocytic vesicle, Golgi trans-face		
LIG_deltaCOP1_diTrp_1	194-202 [A]	Tryptophan-based motifs enable targeting of the tethering and (dis)assembly factors to the C-terminal mu homology domain (MHD) of the coatomer subunit delta, delta-COP.	endoplasmic reticulum membrane, Golgi membrane, COPI coated vesicle membrane, cytosol, COPI vesicle coat, transport vesicle	[DE]{1,3}.{0,2}W.{1,6}[WF]	5.01E-04
LIG_EF_ALG2_ABM_2	251-255 [A]	This isoform-unspecific ALG-2-binding motif binds to the EF hand domains of the proapoptotic Ca2+-binding ALG-2 protein in a Ca2+-dependent manner.	cytosol, nucleus, cytoplasmic side of endoplasmic reticulum membrane, cytoplasmic side of membrane	P.P.{0,1}GF	2.42E-05

LIG_eIF4E_1	274-280 [A] 348-354 [A]	Motif binding to the dorsal surface of eIF4E.	cytosol	Y....L[VILMF]	1.89E-04
LIG_FHA_1	182-188 [A] 216-222 [A] 319-325 [A] 550-556 [A] 681-687 [A]	Phosphothreonine motif binding a subset of FHA domains that show a preference for a large aliphatic amino acid at the pT+3 position.	nucleus	..(T)..[ILV].	8.66E-03
LIG_FHA_2	554-560 [A] 729-735 [A]	Phosphothreonine motif binding a subset of FHA domains that have a preference for an acidic amino acid at the pT+3 position.	nucleus, Replication fork	..(T)..[DE].	8.29E-03
LIG_GBD_Chelix_1	486-494 [A] 722-730 [A]	Amphipatic alpha helix that binds the GTPase-binding domain (GBD) in WASP and N-WASP.	extracellular exosome, vesicle membrane, cytosol, actin cytoskeleton, cell-cell junction	[ILV][VA][^P][^P][LI][^P][^P][^P][LM]	9.79E-05
LIG_HCF-1_HBM_1	123-126	The DHxY Host Cell Factor-1 binding motif (HBM) interacts with the N-terminal kelch propeller domain of the cell cycle regulator HCF-1	nucleus, Transcription factor complex, viral tegument	[DE]H.Y	5.07E-05

LIG_KLC1_WD_1	290-299 [A]	This short WD or WE motif is found in cargo proteins and mediates kinesin-1-dependent microtubule transport by binding to the KLC TPR region.	cytosol, cytoskeleton	[LMTAFSRI][^KRG]W[DE].{3,5}[LIVMFPA]	3.54E-04
LIG_LIR_Apic_2	398-402 [A]	Apicomplexa specific variant of the canonical LIR motif that binds to Atg8 protein family members to mediate processes involved in autophagy.	cytosol, cytoplasmic side of late endosome membrane	[EDST].{0,2}[WFY]..P	3.37E-03
LIG_LIR_Gen_1	224-230 [A] 257-263 [A] 507-513 [A] 538-543 [A] 539-543 [A] 552-557 [A]	Canonical LIR motif that binds to Atg8 protein family members to mediate processes involved in autophagy.	cytosol, cytoplasmic side of late endosome membrane	[EDST].{0,2}[WFY][^RKPG][^PG][ILV]	3.06E-03
LIG_LIR_Nem_3	194-200 [A] 224-230 [A] 257-263 [A] 507-513 [A] 538-543 [A] 539-543 [A] 552-557 [A] 888-894 [A] 889-894 [A]	Nematode-specific variant of the canonical LIR motif that binds to Atg8 protein family members to mediate processes involved in autophagy.	cytosol, cytoplasmic side of late endosome membrane	[EDST].{0,2}[WFY]..[ILVFY]	6.36E-03
LIG_NRBOX	489-495 [A] 574-580 [A] 725-731 [A]	The nuclear receptor box motif (LXXLL) confers binding to nuclear receptors.	nucleus	[^P]L[^P][^P]LL[^P]	2.63E-04

LIG_PCNA_yPIPBox_3	256-266 [A]	The PCNA binding motifs include the PIP Box, PIP degron and the APIM motif, and are found in proteins involved in DNA replication, repair, methylation and cell cycle control. This is the variant for the yeast PIPbox	mismatch repair complex, nucleus, nuclear replication fork, delta DNA polymerase complex	([KR].{0,6}[QN].[^FHWY][LIVM][^P][^PFWYMLIV][FYLMWV][FYLMWVI])|([QN].[^FHWY][LIVM][^P][^PFWYMLIV][FYLMWV][FYLMWVI].{0,6}[KR])	5.58E-04
LIG_Pex3_1	490-501 [A]	LxxLLxxxLxxF motif is located in N-terminus of Pex19 receptors that are responsible for docking to Pex3 docking factor at cis side of peroxisomal membrane.	cytosol, peroxisome, peroxisome membrane, peroxisomal membrane, glycosome, glycosome membrane, glycosomal membrane	L..LL...L..F	5.24E-07

LIG_PTB_Apo_2	512-519 [A] 587-594 [A] 894-901 [A]	These phosphorylation-independent motifs bind to Dab-like PTB domains. Binding is not driven by contacts at the 0 or FY position, but instead is dependent upon the large number of hydrophobic and hydrogen bond contacts between motif and domain.	cytosol, internal side of plasma membrane, integrin, receptor complex, cytoplasmic membrane-bounded vesicle	(.[^P].NP.[FY].)|(.[ILVMFY].N..[FY].)	3.11E-04
LIG_PTB_Phospho_1	894-900 [A]	This phosphorylation-dependent motif binds to Shc-like and IRS-like PTB domains. The pTyr is positioned within a highly basic-charged anchoring pocket. A hydrophobic residue -5 (compared to pY) increases the affinity of the interaction.	internal side of plasma membrane, cytosol, integrin, receptor complex, cytoplasmic membrane-bounded vesicle	(.[^P].NP.(Y))|(.[ILVMFY].N..(Y))	1.35E-04
LIG_Rb_pABgroove_1	512-520 [A]	The LxxLFD motif binds in a deep groove between pocket A and pocket B of the Retinoblastoma protein	nucleus, Transcription factor complex, Rb-E2F complex	..[LIMV]..[LM][FY]D.	2.48E-05

LIG_REV1ctd_RIR_1	338-348 [A]	Several DNA repair proteins interact with the C-terminal domain of the Rev1 translesion synthesis scaffold through the Rev1-Interacting Region RIR motif that is centred around two neighbouring Phe residues.	nucleoplasm, nucleus	..FF[^P]{0,2}[KR]{1,2}[^P]{0,4}	5.35E-04
LIG_SH2_CRK	159-163 [A] 227-231 [A] 276-280 [A] 540-544 [A]	CRK family SH2 domain binding motif.	actin cytoskeleton, cytosol	(Y)[^EPILVFYW][^HDEW][PLIV][^DEW]	1.52E-03
LIG_SH2_GRB2like	588-591 [A] 638-641 [A]	GRB2-like Src Homology 2 (SH2) domains binding motif	Early endosome, cytosol	(Y)([EDST]|[MLIVAFYHQW])N.	3.18E-04
LIG_SH2_NCK_1	540-544 [A]	NCK Src Homology 2 (SH2) domain binding motif	plasma membrane, actin cytoskeleton, cytosol, host cell plasma membrane	(Y)[DESTNA][^GWFY][VPAI][DENQSTAGYFP]	8.29E-04
LIG_SH2_PTP2	591-594 [A]	SH-PTP2 and phospholipase C-gamma Src Homology 2 (SH2) domains binding motif.	cytosol	(Y)[IV].[VILP]	2.45E-04
LIG_SH2_SRC	227-230 [A]	Src-family Src Homology 2 (SH2) domains binding motif.	cytosol	(Y)[QDEVAIL][DENPY	8.73E-04

				HI][IPVGAHS]	
LIG_SH2_STAP1	208-212 [A] 740-744 [A]	STAP1 Src Homology 2 (SH2) domain Class 2 binding motif	plasma membrane	(Y)[DESTA][^GP][^GP][ILVFMWYA]	1.03E-03
LIG_SH2_STAT3	645-648 [A] 798-801 [A]	YXXQ motif found in the cytoplasmic region of cytokine receptors that bind STAT3 SH2 domain.	cytosol	(Y)..Q	7.98E-04
LIG_SH2_STAT5	220-223 [A] 229-232 [A] 276-279 [A] 339-342 [A] 348-351 [A] 411-414 [A] 492-495 [A] 591-594 [A] 601-604 [A] 721-724 [A] 798-801 [A]	STAT5 Src Homology 2 (SH2) domain binding motif.	cytosol	(Y)[VLTFIC]..	3.30E-03
LIG_SH3_3	531-537 [A]	This is the motif recognized by those SH3 domains with a non-canonical class I recognition specificity	plasma membrane, focal adhesion, cytosol	...[PV]..P	1.32E-02
LIG_SUMO_SIM_anti_2	319-326 [A] 320-326 [A] 321-326 [A]	Motif for the antiparallel beta augmentation mode of non-covalent binding to SUMO protein.	PML body, nucleus, nuclear body	[DEST]{1,10}.{0,1}[VIL][DESTVILMA][VIL][VILM]	2.35E-03

				.[DEST]{0,5}	
LIG_SUMO_SIM_par_1	367-374 [A] 446-452 [A] 491-497 [A] 492-497 [A] 702-710 [A] 703-710 [A] 704-710 [A]	Motif for the parallel beta augmentation mode of non-covalent binding to SUMO protein.	PML body, nucleus, nuclear body	[DEST]{0,5}.[VILPTM][VIL][DESTVILMA][VIL].{0,1}[DEST]{1,10}	4.55E-03
LIG_TRAF2_1	449-452 [A]	Major TRAF2-binding consensus motif. Members of the tumor necrosis factor receptor (TNFR) superfamily initiate intracellular signaling by recruiting the C-domain of the TNFR-associated factors (TRAFs) through their cytoplasmic tails.	cytosol	[PSAT].[QE]E	4.30E-03
LIG_TYR_ITIM	337-342 [A]	ITIM (immunoreceptor tyrosine-based inhibitory motif). Phosphorylation of the ITIM motif, found in the cytoplasmic tail of some inhibitory receptors (KIRs) that bind MHC Class I, leads to the recruitment and activation of a protein tyrosine phosphatase.	cytosol	[ILV].(Y)..[ILV]	2.99E-04
LIG_UBA3_1	170-178 [A] 280-287 [A] 353-360 [A] 787-793 [A] 840-849 [A]	UBA3 adenylation domain binding motif variant based on the UBE2M and UBE2F interactions.	nucleus	[ILM][ILMF].{1,2}[ILM].{0,4}K	1.20E-03

LIG_WD40_WDR5_VDV_2	127-134 [A] 130-134 [A] 137-142 [A] 149-156 [A] 183-190 [A] 184-190 [A] 185-190 [A] 229-236 [A] 362-368 [A] 363-368 [A] 372-377 [A] 373-377 [A] 374-377 [A] 379-385 [A] 414-421 [A] 415-421 [A] 483-489 [A] 484-489 [A] 485-489 [A] 491-498 [A] 492-498 [A] 495-498 [A] 507-513 [A] 552-557 [A] 591-596 [A] 608-614 [A] 610-614 [A] 654-658 [A] 657-663 [A] 660-663 [A] 678-682 [A] 702-708 [A] 703-708 [A] 721-728 [A] 731-738 [A] 733-738 [A] 734-738 [A] 740-746 [A] 742-746 [A] 743-746 [A]	Fungi-specific variant of the WDR5-binding motif that binds to a cleft between blades 5 and 6 of the WD40 repeat domain of WDR5, opposite of the Win motif-binding site, to mediate assembly of histone modification complexes.	nucleus, histone methyltransferase complex	[EDSTY].{0,4}[VIPLA][TSDEKR][ILVA]	4.68E-02

LIG_WRC_WIRS_1	259-264 [A]	WRC interacting receptor sequence (WIRS) is a highly conserved and widespread interaction motif that is employed by diverse membrane proteins to recruit the WRC to initiate the dynamic rearrangements of the actin cytoskeleton.	cytosol, actin cytoskeleton, protein-containing complex	[FYILMV].[TS]F(G|[^P]).	5.80E-04
MOD_CK1_1	127-133 [A] 181-187 [A] 295-301 [A] 373-379 [A] 403-409 [A] 485-491 [A] 558-564 [A] 627-633 [A] 654-660 [A] 697-703 [A] 860-866 [A]	CK1 phosphorylation site	cytosol, nucleus	S..([ST])...	1.70E-02
MOD_CK2_1	453-459 [A] 485-491 [A] 697-703 [A]	Casein kinase 2 (CK2) phosphorylation site	protein kinase CK2 complex, nucleus, cytosol	...([ST]).E	1.46E-02
MOD_GlcNHglycan	206-211 [A] 207-211 [A] 208-211 [A] 345-348 [A] 476-479 [A] 487-490 [A] 726-729 [A] 806-810 [A] 807-810 [A]	Glycosaminoglycan attachment site	extracellular, Golgi apparatus	[ED]{0,3}.(S)[GA].	1.79E-02

MOD_GSK3_1	202-209 [A] 237-244 [A] 254-261 [A] 312-319 [A] 369-376 [A] 476-483 [A] 549-556 [A] 553-560 [A] 601-608 [A] 623-630 [A] 678-685 [A] 724-731 [A] 744-751 [A] 860-867 [A] 880-887 [A]	GSK3 phosphorylation recognition site	cytosol, nucleus	...([ST]). ..[ST]	2.68E-02
MOD_N-GLC_1	124-129 [A] 885-890 [A]	Generic motif for N-glycosylation. It was shown that Trp, Asp, and Glu are uncommon before the Ser/Thr position. Efficient glycosylation usually occurs when ~60 residues or more separate the glycosylation acceptor site from the C-terminus.	extracellular, Golgi apparatus, endoplasmic reticulum	.(N)[^P][ST]..	5.02E-03
MOD_N-GLC_2	773-775 [A]	Atipical motif for N-glycosylation site. Examples are Human CD69, which is uniquely glycosylated at typical (Asn-X-Ser/Thr) and atypical (Asn-X-Cys) motifs, beta protein C	extracellular, Golgi apparatus, endoplasmic reticulum	(N)[^P]C	2.97E-04
MOD_NEK2_1	146-151 [A] 215-220 [A] 221-226 [A] 280-285 [A] 354-359 [A] 395-400 [A] 412-417 [A] 482-487 [A] 557-562 [A] 623-628 [A]	NEK2 phosphorylation motif with preferred Phe, Leu or Met in the -3 position to compensate for less favorable residues in the +1 and +2 position.	centrosome, Ndc80 complex, condensed nuclear chromosome	[FLM][^P][^P]([ST])[^DEP][^DE]	9.80E-03

	659-664 [A] 744-749 [A] 884-889 [A]		outer kinetochore, cytosol, nucleus		
MOD_NEK2_2	504-509 [A]	NEK2 phosphorylation motif with specific set of residues in the +1 and +2 position to compensate for less favorable residues in the -3 position.	centrosome, Ndc80 complex, condensed nuclear chromosome outer kinetochore, cytosol, nucleus	[WYPCAG][^P][^P]([ST])[IFCVML][KRHYF]	1.30E-03
MOD_PIKK_1	113-119 [A] 403-409 [A] 476-482 [A]	(ST)Q motif which is phosphorylated by PIKK family members.	nucleus, cytosol	...([ST])Q..	9.23E-03
MOD_PK_1	152-158 [A]	Phosphorylase kinase phosphorylation site	cytosol	[RK]..(S)[VI]..	9.42E-04
MOD_PKA_2	202-208 [A] 343-349 [A] 354-360 [A]	Secondary preference for PKA-type AGC kinase phosphorylation.	cytosol, nucleus, cAMP-dependent protein kinase complex	.R.([ST])[^P]..	9.46E-03

MOD_Plk_1	124-130 [A] 659-665 [A]	Ser/Thr residue phosphorylated by the Plk1 kinase	centralspindlin complex, nucleus, spindle, gamma-tubulin complex, midbody, cytosol, kinetochore, spindle midzone, nuclear condensin complex, cleavage furrow, nucleoplasm, microtubule organizing center	.[DNE][^PG][ST]([FYILMVW]..)|([^PEDGKN][FWYLIVM]).)	7.67E-03
MOD_Plk_2-3	118-124 [A]	Ser/Thr residue phosphorylated by Plk2 and Plk3	nucleus, nucleolus, cytosol,	[DE]..([ST])[EDILMVFWY](([DE].)|(.[DE]))	2.18E-03

			SCF ubiquitin ligase complex, centrosome, centriole		
MOD_Plk_4	215-221 [A] 237-243 [A] 275-281 [A] 318-324 [A] 549-555 [A] 654-660 [A] 682-688 [A] 748-754 [A]	Ser/Thr residue phosphorylated by Plk4	nucleus, cytosol, SCF ubiquitin ligase complex, cleavage furrow, centriole, gamma-tubulin ring complex, centriolar satellite, pericentriolar material	..[^IRFW]([ST])[ILMVFWY][ILMVFWY].	6.02E-03
MOD_ProDKin_1	127-133 [A] 249-255 [A] 312-318 [A] 332-338 [A] 625-631 [A]	Proline-Directed Kinase (e.g. MAPK) phosphorylation site in higher eukaryotes.	cytosol, nucleus	...([ST])P..	1.54E-02

MOD_SUMO_for_1	195-198 [A]	Motif recognised for modification by SUMO-1	nucleus, PML body	[VILMAFP](K).E	1.91E-03
MOD_SUMO_rev_2	172-180 [A] 206-215 [A] 207-215 [A] 209-215 [A] 391-398 [A] 393-398 [A] 742-747 [A] 860-868 [A] 861-868 [A] 874-882 [A] 875-882 [A]	Inverted version of SUMOylation motif recognized for modification by SUMO-1	PML body, nucleus	[SDE].{0,5}[DE].(K).{0,1}[AIFLMPSTV]	1.28E-02
TRG_ENDOCYTIC_2	159-162 [A] 227-230 [A] 276-279 [A] 339-342 [A] 540-543 [A] 638-641 [A]	Tyrosine-based sorting signal responsible for the interaction with mu subunit of AP (Adaptor Protein) complex	plasma membrane, clathrin-coated endocytic vesicle, cytosol	Y..[LMVIF]	2.59E-03
TRG_ER_diArg_1	265-267 [A]	The di-Arg ER retention motif is defined by two consecutive arginine residues (RR) or with a single residue insertion (RXR). The motif is completed by an adjacent hydrophobic/arginine residue which may be on either side of the Arg pair.	cytosol, endoplasmic reticulum membrane, integral protein, ER-Golgi transport vesicle membrane, endoplasmic reticulum membrane,	([LIVMFYWPR]R[^YFWDE]{0,1}R)|(R[^YFWDE]{0,1}R[LIVMFYWPR])	5.37E-03

			Golgi-ER transport vesicle membrane, rough endoplasmic reticulum, endoplasmic reticulum cisterna		
TRG_LysEnd_APsAcLL_1	183-188 [A] 275-280 [A] 414-419 [A]	Sorting and internalisation signal found in the cytoplasmic juxta-membrane region of type I transmembrane proteins. Targets them from the Trans Golgi Network to the lysosomal-endosomal-melanosomal compartments. Interacts with adaptor protein (AP) complexes	cytosol, Endocytic vesicle	[DERQ]...L[LVI]	2.76E-03
TRG_NLS_Bipartite_1	150-168 [A] 151-168 [A]	Bipartite variant of the classical basically charged NLS.	nucleus, Nuclear pore, NLS-dependent protein nuclear import complex	[KR][KR].{7,15}[^DE]((K[RK])|(RK))(([^DE][KR])|([KR][^DE]))[^DE]	2.59E-04
TRG_NLS_MonoCore_2	162-167 [A]	Monopartite variant of the classical basically charged NLS. Strong core version.	nucleus, Nuclear pore, NLS-dependent protein nuclear	[^DE]((K[RK])|(RK))[KRP][KR][^DE]	2.35E-04

			import complex		
TRG_NLS_MonoExtN_4	160-167 [A] 163-168 [A]	Monopartite variant of the classical basically charged NLS. N-extended version.	nucleus, Nuclear pore, NLS-dependent protein nuclear import complex	(([PKR].{0,1}[^DE])|([PKR]))((K[RK])|(RK))(([^DE][KR])|([KR][^DE]))[^DE]	1.28E-03
TRG_Pf-PMV_PEXEL_1	178-182 [A] 795-800 [A] 802-806 [A]	Plasmodium Export Element, PEXEL, is a trafficking signal for protein cleavage by PMV protease and export from Plasmodium parasites to infected host cells.	endoplasmic reticulum, host cell cytoplasm, extracellular, host cell outer membrane, pathogen-containing vacuole membrane	(R.[LI].[EDQ])|(R.L..[EDQ])|(K.L.E)	2.16E-03
CLV_NRD_NRD_1	77-79 [A]	N-Arg dibasic convertase (NRD/Nardilysin) cleavage site (X-|-R-K or R-|-R-X).	extracellular, Golgi apparatus, cell surface	(.RK)|(RR[^KR])	7.47E-03
CLV_PCSK_KEX2_1	77-79 [A]	Yeast kexin 2 cleavage site (K-R-|-X or R-R-|-X).	extracellular,	[KR]R.	7.97E-03

			Golgi apparatus		
CLV_PCSK_SKI1_1	78-82 [A]	Subtilisin/kexin isozyme-1 (SKI1) cleavage site ([RK]-X-[hydrophobic]-[LTKF]-|-X).	endoplasmic reticulum lumen, endoplasmic reticulum, Golgi apparatus, extracellular	[RK].[AILMFV][LTKF].	6.82E-03
DOC_MAPK_DCC_7	81-90 [A] 82-90 [A]	A kinase docking motif mediating interaction towards the ERK1/2 and p38 subfamilies of MAP kinases	cytosol, Transcription factor complex, nucleus	[RK].{2,4}[LIVP]P.[LIV].[LIVMF]|[RK].{2,4}[LIVP].P[LIV].[LIVMF]	4.51E-04
DOC_MAPK_MEF2A_6	81-90 [A] 82-90 [A]	A kinase docking motif that mediates interaction towards the ERK1/2 and p38 subfamilies of MAP kinases.	nucleus, cytosol, Transcription factor complex	[RK].{2,4}[LIVMP].[LIV].[LIVMF]	2.58E-03
DOC_PP1_RVXF_1	19-26 [A]	Protein phosphatase 1 catalytic subunit (PP1c) interacting motif binds targeting proteins that dock to the substrate for dephosphorylation. The motif defined is [RK]{0,1}[VI][^P][FW].	nucleus, protein phosphatase type 1 complex,	..[RK].{0,1}[VIL][^P][FW].	8.30E-04

			cytosol		
DOC_USP7_MATH_1	84-88 [A] 86-90 [A]	The USP7 MATH domain binding motif variant based on the MDM2 and p53 interactions.	nucleus	[PA][^P][^FYWIL]S[^P]	1.24E-02
LIG_14-3-3_CanoR_1	21-26 [A]	Canonical Arg-containing phospho-motif mediating a strong interaction with 14-3-3 proteins.	nucleus, internal side of plasma membrane, cytosol	R[^DE]{0,2}[^DEPG]([ST])(([FWYLMV].)|([^PRIKGN]P)|([^PRIKGN].{2,4}[VILMFWYP]))	4.48E-03
LIG_BRCT_BRCA1_1	62-66 [A]	Phosphopeptide motif which directly interacts with the BRCT (carboxy-terminal) domain of the Breast Cancer Gene BRCA1 with low affinity	nucleus, BRCA1-BARD1 complex	.(S)..F	1.91E-03
LIG_CSL_BTD_1	52-55 [A]	The motif mediates the interaction between a Notch-like protein and the transcription factor CSL by placing two amino acids (W and P) into a hydrophobic pocket of the BTD domain of CSL.	nucleus, Transcription factor complex, chromatin	[AFILMPTVW]W[FHILMPSTVW]P	1.28E-04
LIG_FHA_1	28-34 [A]	Phosphothreonine motif binding a subset of FHA domains that show a preference for a large aliphatic amino acid at the pT+3 position.	nucleus	..(T)..[ILV].	8.66E-03

LIG_LIR_Apic_2	89-94 [A]	Apicomplexa specific variant of the canonical LIR motif that binds to Atg8 protein family members to mediate processes involved in autophagy.	cytosol, cytoplasmic side of late endosome membrane	[EDST].{0,2}[WFY]..P	3.37E-03
LIG_LIR_Nem_3	24-28 [A] 68-74 [A] 69-74 [A] 70-74 [A]	Nematode-specific variant of the canonical LIR motif that binds to Atg8 protein family members to mediate processes involved in autophagy.	cytosol, cytoplasmic side of late endosome membrane	[EDST].{0,2}[WFY]..[ILVFY]	6.36E-03
LIG_SH2_CRK	91-95 [A]	CRK family SH2 domain binding motif.	actin cytoskeleton, cytosol	(Y)[^EPILVFYW][^HDEW][PLIV][^DEW]	1.52E-03
LIG_WD40_WDR5_VDV_2	36-41 [A] 87-90 [A]	Fungi-specific variant of the WDR5-binding motif that binds to a cleft between blades 5 and 6 of the WD40 repeat domain of WDR5, opposite of the Win motif-binding site, to mediate assembly of histone modification complexes.	nucleus, histone methyltransferase complex	[EDSTY].{0,4}[VIPLA][TSDEKR][ILVA]	4.68E-02
MOD_CK2_1	66-72 [A]	Casein kinase 2 (CK2) phosphorylation site	protein kinase CK2 complex, nucleus, cytosol	...([ST]).E	1.46E-02

MOD_GlcNHglycan	5-8 [A]	Glycosaminoglycan attachment site	extracellular, Golgi apparatus	[ED]{0,3}.(S)[GA].	1.79E-02
MOD_GSK3_1	8-15 [A] 29-36 [A] 66-73 [A]	GSK3 phosphorylation recognition site	cytosol, nucleus	...([ST]).. [ST]	2.68E-02
MOD_NEK2_1	1-6 [A]	NEK2 phosphorylation motif with preferred Phe, Leu or Met in the -3 position to compensate for less favorable residues in the +1 and +2 position.	centrosome, Ndc80 complex, condensed nuclear chromosome outer kinetochore, cytosol, nucleus	[FLM][^P][^P]([ST])[^DEP][^DE]	9.80E-03
MOD_NEK2_2	86-91 [A]	NEK2 phosphorylation motif with specific set of residues in the +1 and +2 position to compensate for less favorable residues in the -3 position.	centrosome, Ndc80 complex, condensed nuclear chromosome outer kinetochore, cytosol, nucleus	[WYPCAG][^P][^P]([ST])[IFCVML][KRHYF]	1.30E-03

MOD_PKA_2	12-18 [A]	Secondary preference for PKA-type AGC kinase phosphorylation.	cytosol, nucleus, cAMP-dependent protein kinase complex	.R.([ST])[^P]..	9.46E-03
MOD_PKB_1	19-27 [A]	PKB Phosphorylation site	cytosol, nucleus	R.R..([ST])[^P]..	6.03E-04
MOD_Plk_4	12-18 [A] 86-92 [A]	Ser/Thr residue phosphorylated by Plk4	nucleus, cytosol, SCF ubiquitin ligase complex, cleavage furrow, centriole, gamma-tubulin ring complex, centriolar satellite, pericentriolar material	..[^IRFW]([ST])[ILMVFWY][ILMVFWY].	6.02E-03

TRG_Cilium_RVxP_2	44-48 [A]	The RVxPx motif is located in the cytoplasmatic tails of vesicular cargoes. It allows the interaction with proteins that permit the vesicle budding from the trans-Golgi-network and its posterior transport to the plasma membrane of the cilia	cilium, cytosol	RV.P.	2.73E-04
TRG_ER_diArg_1	76-78 [A]	The di-Arg ER retention motif is defined by two consecutive arginine residues (RR) or with a single residue insertion (RXR). The motif is completed by an adjacent hydrophobic/arginine residue which may be on either side of the Arg pair.	cytosol, endoplasmic reticulum membrane, integral protein, ER-Golgi transport vesicle membrane, endoplasmic reticulum membrane, Golgi-ER transport vesicle membrane, rough endoplasmic reticulum, endoplasmic reticulum cisterna	([LIVMFYWPR]R[^YFWDE]{0,1}R)|(R[^YFWDE]{0,1}R[LIVMFYWPR])	5.37E-03

TRG_NLS_MonoExtC_3	77-82 [A]	Monopartite variant of the classical basically charged NLS. C-extended version.	nucleus, Nuclear pore, NLS-dependent protein nuclear import complex	[^DE]((K[RK])|(RK))(([^DE][KR])|([KR][^DE]))(([PKR])|([^DE][DE]))	7.25E-04
TRG_NLS_MonoExtN_4	77-82 [A]	Monopartite variant of the classical basically charged NLS. N-extended version.	nucleus, Nuclear pore, NLS-dependent protein nuclear import complex	(([PKR].{0,1}[^DE])|([PKR]))((K[RK])|(RK))(([^DE][KR])|([KR][^DE]))[^DE]	1.28E-03