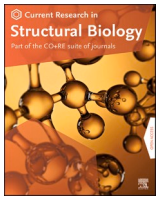
Current Research in Structural Biology 7 (2024) 100113

Contents lists available at ScienceDirect

Current Research in Structural Biology

journal homepage: www.journals.elsevier.com/current-research-in-structural-biology



Research Article

Bioinformatic analysis of THAP9 transposase homolog: conserved regions, novel motifs

Richa Rashmi [a], Chandan Nandi [a], Sharmistha Majumdar [a,*]

[a] Discipline of Biological Engineering, IIT Gandhinagar, Gandhinagar, Gujarat, India

ARTICLE INFO

Handling editor: A Wlodawer

Keywords:
Evolutionary analysis
Transposase gene
Protein sequence characterization

ABSTRACT

THAP9 is a transposable element-derived gene that encodes the THAP9 protein, which is homologous to the *Drosophila* P-element transposase (DmTNP) and can cut and paste DNA. However, the exact functional role of THAP9 is unknown. Here, we perform structure prediction, evolutionary analysis and extensive *in silico* characterization of THAP9, including predicting domains and putative post-translational modification sites. Comparison of the AlphaFold-predicted structure of THAP9 with the DmTNP CryoEM structure, provided insights about the C2CH motif and other DNA binding residues, RNase H-like catalytic domain and insertion domain of the THAP9 protein. We also predicted previously unreported mammalian-specific post-translational modification sites that may play a role in the subcellular localization of THAP9. Furthermore, we observed that there are distinct organism class-specific conservation patterns of key functional residues in certain THAP9 domains.

1. Introduction

Transposable elements (TEs) are DNA sequences that can move and duplicate within a genome (Su et al., 2020). They can cause mutations as well as increase genome size (Bourque et al., 2018). 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 (>25% homology) of the *Drosophila* P-element transposase (DmTNP) (Majumdar and Rio, 2015). The DmTNP protein mobilizes the P-element transposon, which is the causative agent for hybrid dysgenesis in *Drosophila*.

hTHAP9 belongs to the human THAP (Thanatos-associated protein) protein family, with twelve proteins (hTHAP0-hTHAP11). As shown in Fig. 1, all THAP proteins are characterized by a signature THAP domain, an 80–90 residue-long DNA-binding domain located at the N-terminal end of the protein (Sabogal et al., 2010), which consists of a signature C2CH (Cys-X2-4-Cys-X35-50-Cys-X2-His) zinc-coordinating motif. C2CH type zinc fingers are the second-most common zinc-coordinating DNA-binding motifs after the C2H2 type of zinc fingers (Roussigne et al., 2003; Bessière et al., 2008). Structural studies indicated that THAP domains have a conserved three-dimensional structure characterized by a β-α-β fold, with four loops L1-L4 connecting the β sheets and the helix (Bessière et al., 2008; Balakrishnan et al., 2009; Dejosez et al., 2010; Sabogal et al., 2010; Gervais et al., 2013). Additionally, most human THAP proteins also possess a conserved 'AVPTIF' motif at the C-terminal end of the THAP domain. THAP-domains participate in sequence-specific DNA binding via a bipartite recognition of adjacent major and minor grooves, wherein the β sheet interacts with the DNA major groove. In contrast, loop 4 (L4) interacts with the DNA minor groove via basic amino acid residues. However, THAP domains share very little sequence identity (~20 %) and DNA-binding specificity, i.e., they are involved in recognizing different DNA target sequences (Gervais et al., 2013).

In addition to the signature THAP domain, the coiled-coil domain and HBM (HCF1 Binding motif) are two other conserved elements found in most of the 12 human THAP proteins, highlighting their importance for this protein family (Lu et al., 1998; Freiman and Herr, 1997; Burkhard et al., 2001; Mazars et al., 2010; Sanghavi et al., 2019). The coiled-coil domain is strongly implicated in THAP protein oligomerization. However, in the case of THAP9, homo and hetero-oligomerization possibly occur via multiple interactions across the length of the protein (Sanghavi and Majumdar, 2021). The HBM has a consensus motif [(D/E)HXY] present in all human THAP proteins except THAP8, THAP10, and THAP12 (Dehaene et al., 2020). THAP proteins may use the HBM to extend their regulatory network, including the transcription co-regulator HCF-1. HCF-1 (also known as HCFC1) is a

* Corresponding author.
E-mail address: sharmistham@iitgn.ac.in (S. Majumdar).

https://doi.org/10.1016/j.crstbi.2023.100113
Received 1 September 2023; Received in revised form 10 November 2023; Accepted 13 November 2023
Available online 24 November 2023

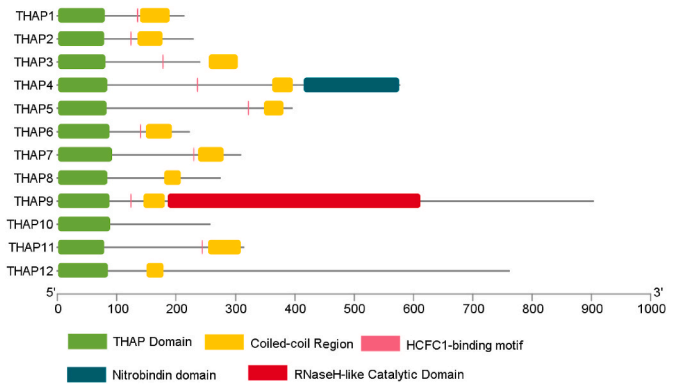


Fig. 1. Schematic representation of domain organization of human THAP family proteins. The plot has been generated using TBtools. The grey line represents the length of the individual proteins, and the colored boxes on the line represent the position of the domain in the protein.

regulatory protein involved in cell cycle progression, embryonic stem cell pluripotency, and stress response (Mazars et al., 2010; Parker et al., 2012; Zargar and Tyagi, 2012). However, surprisingly, the presence or lack of an HBM in a THAP protein does not determine whether it is actively interacting with HCF-1. For example, the THAP5 protein, which has the HBM, does not interact with HCF-1, while despite lacking HBM, THAP8 still interacts with HCF-1 (Dehaene, 2019, Dehaene et al., 2020). These results point toward the complex regulatory mechanisms of the THAP-protein family, resulting in their involvement in distinct cellular processes.

Human THAP proteins have been implicated in various human diseases. For example, THAP1 is involved in torsional dystonia and hemophilia (Richter et al., 2017), THAP5 is involved in heart diseases (Balakrishnan et al., 2009), THAP2 (Leite et al., 2013), THAP10 (De Souza Santos et al., 2008), and THAP11 (Parker et al., 2012) are involved in multiple cancers. However, the exact functional role of THAP9 remains a gap in our understanding.

The lack of detailed structural data for human THAP proteins has hindered studies of their structure-function relationships. The three-dimensional structure of the isolated THAP domain has been reported for a few THAP family proteins (Bessière et al., 2008; Liew et al., 2007; Bessière et al., 2008; Campagne et al., 2010; Sabogal et al., 2010). Reliable three-dimensional structures of proteins have recently been predicted by AlphaFold and deposited in the AlphaFold Protein Structure Database (Jumper et al., 2021; Varadi et al., 2022). In this study, we have compared the AlphaFold model of hTHAP9, highlighting similarities and differences with the solved structure of DmTNP (homolog).

Phylogenetic analysis of THAP9 can help us understand the evolution of the hTHAP9 gene and its associated protein. 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. This study identified previously unreported functional features in the THAP9 protein sequence, highly conserved in mammals. These include four adjacent motifs: N-glycosylation site, Protein kinase C (PKC) phosphorylation site, Leucine zipper domain, and Bipartite nuclear localization signal (NLS), which may play a role in the subcellular localization of THAP9. The study also revealed two N-myristoylation sites within the THAP domain.

2. Materials and methods

2.1. 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 (aves), 4 alligators, 7 turtles, 6 lizards, 4 amphibians, 120 mammals, and 1 lamprey (hyperoartia). We used the following five representative classes: amphibians, hyperoartia, reptiles, birds, and mammals for comparison and representation purposes.

2.2. Sequence alignment & phylogenetic analysis

A phylogeny-aware sequence alignment was performed for amino acid sequences using the CLUSTALW (Thompson et al., 1994) method present in MEGAX (Kumar et al., 2018) software using the guide trees generated from TimeTree (Kumar et al., 2017) (http://www.timetree.org/). We created multiple sequence alignments (MSA) separately for all vertebrate classes (birds, alligators, turtles, lizards, amphibians, and mammals) and combined all species. Following the protein MSA of the orthologs, the Coding DNA Sequences (CDS) were codon aligned using PAL2NAL (Suyama et al., 2006)) (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 (Letunic and Bork, 2019).

2.3. Protein sequence characterization

The domain architecture of hTHAP9 and orthologs was predicted using SMART (Letunic et al., 2021) (Simple Modular Architecture Research Tool, http://smart.embl-heidelberg.de/) in "batch mode" considering the option for including Pfam (El-Gebali et al., 2019) domains. ELM (Kumar et al., 2020) 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 (Bailey et al., 2009) and ScanProsite (de Castro et al., 2006). Disordered binding regions (DBR) (Ward et al., 2004) and secondary structure elements of the hTHAP9 protein were predicted using PSIPRED Workbench (Buchan and Jones, 2019), and areas with disorder scores greater than 0.6 were considered 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 (Wilkins et al., 1999) and EMBOSS Pepstats (Rice et al., 2000) (https://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats/). Finally, previously annotated SMART & Pfam domains and newly predicted domains of hTHAP9 and its orthologs were visualized using iTOL (Letunic and Bork, 2019).

2.4. Structural analyses

The structure of the human THAP9 protein was fetched from the AlphaFold Protein Structure Database (AF Database https://alphafold.ebi.ac.uk, Jumper et al., 2021; Varadi et al., 2022), which has been developed by the joint efforts of DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI). AlphaFold produces two metrics for assessing the predicted structures, (i) pLDDT (predicted score on Local Distance Difference Test) and (ii) PAE (Predicted Aligned Error). pLDDT estimates a per-residue confidence score on a scale from 0 to 100, which can assess confidence within a single domain (Mariani et al., 2013). These confidence scores are stored in the B-factor fields of the mmCIF and PDB files downloaded from the AlphaFold Database. These values are also used to color the residues of the 3D structure visualized in the database. Residues with a pLDDT score of 90 or more (highlighted in

blue) have a high model confidence level. Residues with a pLDDT score between 70 and 90 (highlighted in cyan) are also considered to be modeled well with a confident backbone prediction. Low and very low confidence corresponds to residues with a pLDDT score between 50 and 70 (highlighted with orange) and 0–50, respectively (Jumper et al., 2021; Tunyasuvunakool et al., 2021; Varadi et al., 2022). Intrinsically disordered regions probably correspond to regions with very low pLDDT scores (Akdel et al., 2021). Predicted Aligned Error (PAE) is shown as an interactive 2D-Plot in the AF database. If the predicted and actual structures are aligned at residue y, the PAE scores between residues x and y (x,y) indicate AlphaFold's expected position error at residue x. If the residues x and y belong to two different domains, then a low PAE score (represented in dark green) determines the well-defined relative positions of the two residues, while a high PAE score (represented in light green) highlights their uncertain and unreliable relative positioning (Jumper et al., 2021; Varadi et al., 2022). This plot can be used to identify different domains in the predicted structures.

The experimentally derived structures of DmTNP were downloaded from the Protein Data Bank (https://www.rcsb.org) (Berman et al., 2000). We compared the THAP domain of DmTNP with the predicted hTHAP9 structure using Chimera (https://www.rbvi.ucsf.edu/chimera/). We first superimposed the structures using the 'MatchMaker' tool (taking the DmTNP structure as reference), followed by creating structure-based sequence alignment using the 'Match- > Align' tool. Finally, we selected the catalytic domain (which included the insertion domain) of the hTHAP9 model predicted by Alphafold and structurally aligned it with the corresponding structure of DmTNP for comparison.

3. Results

3.1. Evolution of THAP9 through organisms

To study the possible functional role of THAP9, 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 homolog (Majumdar et al., 2013). Therefore, we used the same transcript variant for our analysis.

According to NCBI (Jan 2021), hTHAP9 protein has orthologs in 216 vertebrate organisms, including 74 birds (aves), 4 alligators, 7 turtles, 6 lizards, 4 amphibians, 120 mammals, and 1 lamprey. We downloaded the protein and coding DNA sequences for the same from NCBI. We began our analysis by aligning the protein sequences using CLUSTALW (Thompson et al., 1994) in a phylogeny-aware manner. We generated the time tree for the organisms using timetree (Kumar et al., 2017). Post-alignment, we filtered the diverse, low-quality, and partial sequences, after which we were left with 178 sequences (3 amphibians, 1 hyperoartia, 13 reptiles, 53 birds, 108 mammals). Later we realigned the sequences using CLUSTALW in a phylogeny aware manner (Thompson et al., 1994), followed by building Maximum likelihood Phylogenetic trees (Hall, 2013) with 100 bootstrap replicates. These parts were performed using options available in MEGAX (Kumar et al., 2018). We codon aligned the coding DNA sequences from the orthologs using PAL2NAL, taking corresponding aligned protein sequences as input. The resulting trees (Fig. 2) were drawn using the iTOL web server. Rooting trees to the midpoint perfectly aligns the lamprey (hyperoartia), amphibians, reptiles, birds, and mammals in the given order, thus suggesting that THAP9 gradually evolved over the course of evolution, wherein THAP9 orthologs from each class cluster separately, both at the level of DNA (Fig. 2B) as well as the corresponding amino acid sequence (Fig. 2A).

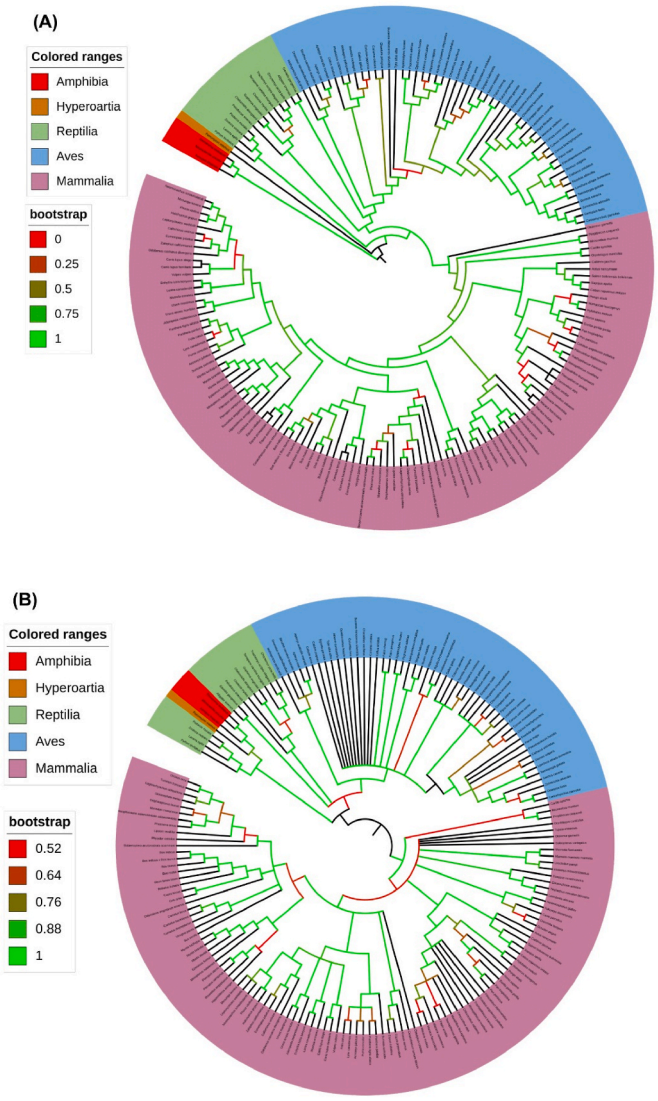


Fig. 2. Phylogenetic analysis of THAP9 orthologs. (A) Phylogenetic tree of aligned protein sequences of THAP9 orthologs. Protein sequences were aligned using CLUSTALW, followed by building Maximum likelihood Phylogenetic trees with 100 bootstrap replicates. **(B) Phylogenetic tree of aligned coding DNA sequences of THAP9 orthologs.** Coding DNA sequences of the THAP9 orthologs were codon aligned using PAL2NAL and Maximum likelihood Phylogenetic trees were generated using MEGAX with 100 bootstrap replicates. The trees were generated using iTOL, and the organism classes are marked in different colors (legend on the left). Bootstrap values are marked on the tree. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.2. Analysis of THAP9 protein orthologs suggests new functional motifs in mammals

Domain architecture: According to SMART & Pfam (El-Gebali et al., 2019), *hTHAP9* has an 89 residue-long THAP-type Zinc Finger domain at the N-terminal end followed by a 177 residue-long P-element transposase domain (Tnp_P_element, Pfam ID: PF12017) between residues 142 to 389 (Fig. 3 (a)). SMART also predicted two low complexity regions in the h*THAP9* protein sequence from 444 to 455 and 567 to 579. Most of the THAP9 orthologs have similar SMART & PFAM annotations. All the annotated THAP9 ortholog sequences have the Tnp_P_element domain. However, some species lack the THAP domain, for example, *Falco peregrinus* (aves) and *Alligator Sinensis* (Reptile) (Fig. 3 (a), full representation in Supplementary Fig. 1 (a)). There are no THAP9

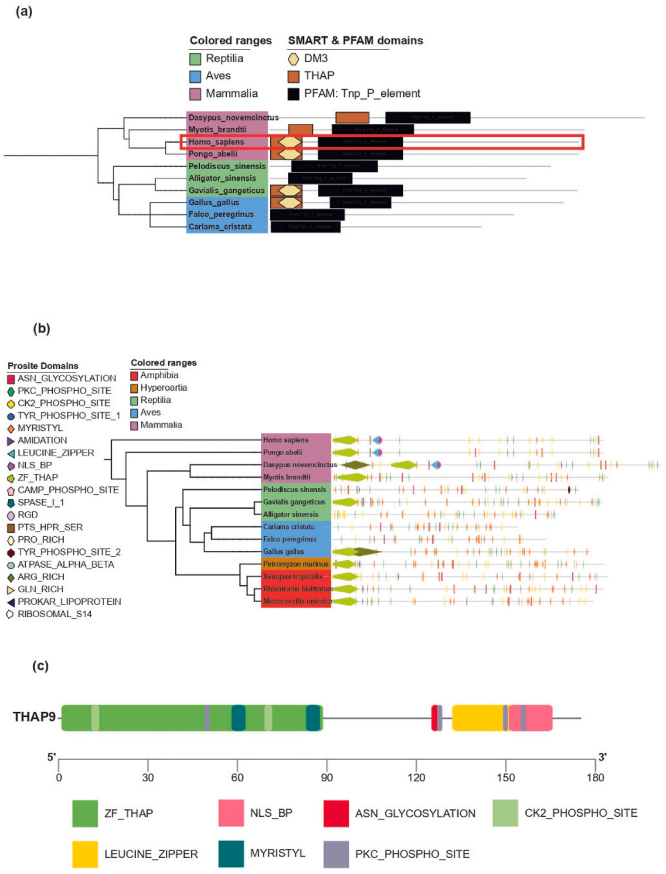


Fig. 3. **(a)** Previously annotated SMART & Pfam domains of THAP9 orthologs. Results were visualized on a phylogenetic tree using iTOL. Domain architectures were displayed using the data export feature of SMART. **(b)** New predicted functional features of THAP9 orthologs. using Eukaryotic Linear Motif (ELM) and ScanProsite. The tree was made using iTOL. **(c)** The architecture of N terminal (1–175) region of hTHAP9. Predicted N-myristoylation sites (PS00008) (i) "GAilCS" (residues 58–63), (ii) "GAvpSV" (residues 83–88). Predicted 4 adjacent motifs (i) "NYSL" N-glycosylation site (PS00001) (125–128) (ii) "SlK" Protein kinase C phosphorylation site (PS00005) (127–129) (iii) "LtigaekLaevqqmLqvskkrL" Leucine zipper pattern (PS00029) (132–153) (iv) "KRLISVKNYR-MIKKRK" Bipartite nuclear localization signal profile (PS50079) (151–166). This figure has been created using TBtools.

orthologs of Amphibians and Hyperoartia class with already annotated SMART or Pfam domains.

Functional Motifs: Various new functional features of THAP9 were predicted using the Eukaryotic Linear Motif (Kumar et al., 2020) (ELM) and ScanProsite (de Castro et al., 2006) (Fig. 3 (b), full representation in Supplementary Fig. 1 (b)).

3 N-myristoylation sites (Prosite ID: PS00008) were predicted in hTHAP9, namely (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 are located in the N terminal THAP domain. hTHAP9 may localize to the nucleus as it has a sequence-specific DNA binding THAP domain (Campagne et al. (2010); Sabogal et al. (2010)) as well as a predicted bipartite NLS (Sanghavi and Majumdar, 2021). However, myristoylation is a protein-lipid modification essential for cellular signaling, protein-protein interactions, and intracellular targeting of proteins to endomembrane or plasma membrane systems (Udenwobele et al., 2017). Thus, it is possible that THAP9 could also localize to cytosolic locations via selective myristoylation.

Moreover, the ELM and ScanProsite analysis also predicted the occurrence of 4 adjacent motifs located between 125 and 166, namely (i) "NYSL" N-glycosylation site (PS00001) from 125 to 128 (ii) "SlK" 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 (Fig. 3 (c), Supplementary Fig. 1 (b)). Multiple previously unreported functional motifs (Supplementary Table 2) were also predicted. These motifs can be explored further to understand the regulation and function of THAP9.

NLS: NLS sequences generally appear either as a single-stretch (monopartite) or as two clusters (bipartite) of basic residues separated by approximately ten amino acid residues, with the respective consensus sequences being (K/R)4–6 and (K/R)2 X10–12 (K/R)3 (Robbins et al., 1991). ELM and ScanProsite analysis found a 16 residue long bipartite NLS sequence "KRLISVKNYR-MIKKRK," in which 50% of the residues were basic, and the basic region was more concentrated in the C-Terminal end of the motif.

Bipartite nuclear localization signals (NLS) are sometimes located close to (e.g., HSF2) (Sheldon and Kingston, 1993) or within (e.g., SREBP-2 (Nagoshi et al., 1999)) a Leu-zipper domain. Moreover, the subcellular localization of a protein may be regulated by masking and unmasking of its NLS by local phosphorylation events (Nagoshi et al., 1999). It is tempting to speculate that the subcellular localization of hTHAP9 may be regulated by its Leucine zipper domain which contains a bipartite NLS as well as a highly conserved Protein kinase C phosphorylation site.

Phosphorylation sites: Several kinase phosphorylation sites (Protein kinase C & Casein kinase II) were predicted to be distributed across the length of the protein. ELM also predicted one Host Cell Factor-1 binding motif in hTHAP9 that has been previously annotated (Kumar et al., 2020; Sanghavi and Majumdar, 2021).

Other domains: When we looked for weak domain matches in THAP9, ScanProsite also predicted the presence of a Phosphatase tensin-type domain (PPASE_TENSIN) overlapping TnP_Pelement domain. Tumor suppressor protein PTEN is the best-characterized member of the PPASE_TENSIN family (Chu and Tarnawski, 2004). To further explore the weak presence of the PPASE_TENSIN (Prosite Entry: PS51181) domain in the 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.

The Phosphatase tensin-type domain is present in 38 protein sequences in the Uniprot database (The UniProt Consortium, 2019), including human PTEN. To look for conserved motifs in the predicted PPASE_TENSIN domain sequences in 105 THAP9 orthologs and the 38 proteins from Uniprot, we processed them through the MEME tool part of the MEME-Suite (Bailey et al., 2009) using the site distribution parameter as "one occurrence per sequence." We identified 15 highly conserved motifs, but the confidence score of these matches was not significant enough.

To further examine the diversification of THAP9 orthologs, we predicted conserved motifs using MEME (Bailey et al., 2009). We identified the distribution of the top 30 conserved motifs occurring at least once in all the sequences (Table 1). The motif composition appeared more conserved in mammals than other groups, suggesting functional relevance in mammals as well as functional diversification among other classes.

3.3. *Evolutionary and structural analysis of individual domains*

Although the THAP proteins have been studied using bioinformatic and biochemical methods, there are limited structural studies on these proteins and their domains. Alphafold has recently revolutionized structural biology by predicting highly accurate and reliable protein structures. Thus, we comprehensively analyzed the Alphafold-predicted hTHAP9 structure and compared it with the experimentally-derived structures of DmTNP. Further, to study the functional diversification

Table 1
Table of top 15 MEME motifs from THAP9 ortholog proteins.

Sr. No.	Motif (Regular Expression)	No. of orthologs (out of 178) in which motif is present
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

of THAP9 among different vertebrate classes (178 orthologs) in more detail, we studied the evolution of some previously reported functional domains and residues of functional relevance. We looked at the evolution of the THAP domain, L-Zipper domain, predicted insertion domain and RNase H-like catalytic domain (Sharma et al., 2021).

We started by searching the Protein Data Bank (PDB) for experimentally-derived structures of the well-characterized active transposase DmTNP which is homologous to the hTHAP9 protein. We found three entries (reported in Table 2) namely the DmTNP-THAP domain bound to donor DNA (PDB ID - 3KDE) and the full-length DmTNP strand transfer complex structure (PDB ID - 6PE2 & 6P5A) (Ghanim et al., 2019; Sabogal et al., 2010).

According to the solved structures, DmTNP consists of multiple domains (Fig. 4 (A)), which include the 88 residue long, zinc-coordinating DNA binding THAP domain located at the N terminal end, a Dimerization domain (residues 88–151) that includes a leucine zipper and a coiled-coil domain followed by a newly reported Helix-Turn-Helix domain (residues 151–200) which helps in binding donor DNA during DNA excision. Then there is an RNase H-like catalytic domain (residues 200–566), which brings together the Mg^{2+} coordinating DDE/D catalytic triad important for DNA cleavage during transposition. The DmTNP catalytic domain is disrupted by a GTP binding insertion domain (residues 339–528), which is required for assembling the protein into a higher-order nucleoprotein synaptic or paired-end complex (PEC). Finally, the C-terminal end of the protein has a C-Terminal Domain (residues 566–751) rich in basic residues and is responsible for binding target DNA via electrostatic interactions.

hTHAP9, our protein of interest, does not have a solved structure, but its domain organization has been predicted by various studies (Fig. 4 (B) (Sanghavi et al., 2019; Sharma et al., 2021). It has an 89 residue long DNA binding THAP domain located at its N-terminal end. This is followed by a ~40 amino acid long Leucine-rich domain, which has been reported to play a role in oligomerization and also carries an HBM (residues 123–126, consensus motif [(D/E)HXY]). A recent study suggested the possible colocalization of hTHAP9 with HCF-1. We have also predicted a Leucine Zipper domain in this region (Fig. 3c), which is highly conserved across the mammalian species. An RNase H-like domain-containing catalytic residues responsible for DNA excision has been predicted adjacent to the oligomerization domain. Like the catalytic domain of DmTNP, the RNase H fold of hTHAP9 is disrupted by an insertion domain (Sharma et al., 2021). In DmTNP, the insertion domain contains a unique GTP binding domain (Ghanim et al., 2019), but the role of the hTHAP9 insertion domain is still unknown. Moreover, the functional significance of the C-terminal end of hTHAP9 has not been characterized yet.

We fetched Alphafold (AF)-predicted structure of the human THAP9 protein from the database. Table 2 shows the comparison of the experimentally-derived structures DmTNP with the AF-predicted structure of hTHAP9. Fig. 5 shows the AF-predicted structures of the hTHAP9 protein colored according to the pLDDT score and corresponding PAE plots. As originally described (Roussigne et al., 2003), the AF models of hTHAP9 confirmed the presence of a THAP domain at the N-terminal end (Fig. 5 (C)). In addition to the THAP domain, the model exhibited two more structured regions, the previously predicted RNase H-like catalytic domain (Fig. 5 (D)) and a novel domain at the C-Terminal end of the protein (Fig. 5 (E)). Corroborating previous studies (Sanghavi et al., 2019), hTHAP9 also possessed an alpha-helical/coiled-coil region between residues 145–182, downstream of the THAP domain.

We then proceeded to perform a domain-by-domain analysis of hTHAP9 with DmTNP using both structural as well as protein sequence data.

THAP-Domain: THAP domain, the characteristic feature of the THAP family proteins, is a conserved 80–90 amino acid DNA binding domain located at the N-terminal end of the protein. It is a C2CH type Zinc Finger domain that can bind specific DNA sequences and has a

Table 2
Details of available experimental structures (from PDB) of DmTNP (1st 4 columns). The experimental structures of DmTNP were individually compared (RMSD) with the AlphaFold predicted structures of hTHAP9 and the hTHAP9 THAP domain (last column).

Protein	Domain/ complex	PDB ID	Type of Structure/ Resolution (A)	RMSD with AF Model (Å)
DmTNP	THAP domain-DNA Complex	3KDE	X-RAY DIFFRACTION 1.74 Å	1.004 (With hTHAP9 THAP domain)
	Strand Transfer complex	6PE2	ELECTRON MICROSCOPY 4 Å	1.785 (with hTHAP9)
	Strand Transfer complex	6P5A	ELECTRON MICROSCOPY 3.6 Å	1.039 (with hTHAP9)

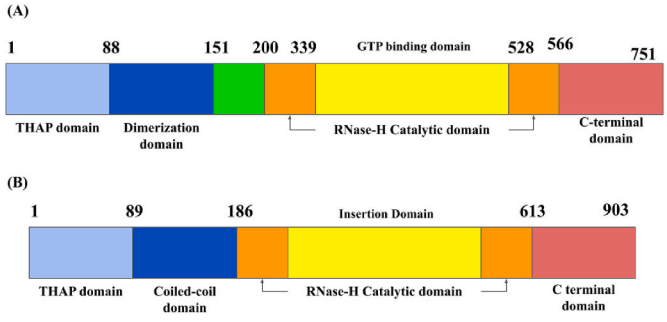


Fig. 4. Schematic representation of protein domain architecture of (A) DmTNP (B) hTHAP9.

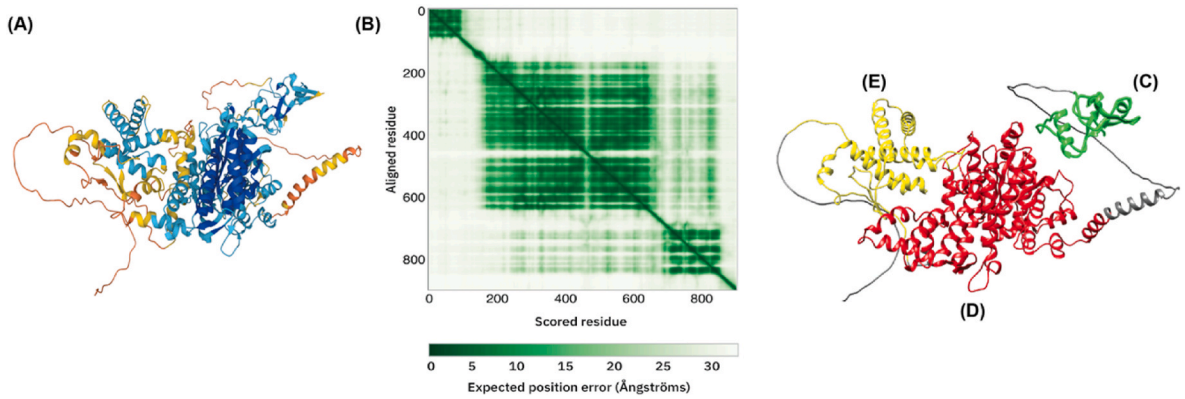


Fig. 5. AlphaFold-predicted structure of human THAP9 protein. (A) Alphafold structure of full-length hTHAP9 protein colored according to pLDDT score (Blue, cyan, orange, and yellow color represents very high (LDDT >90), high (90 > LDDT >70), low (70 > LDDT >50), and very low (LDDT <50) per-residue pLDDT scores, respectively). **(B)** PAE plot for the predicted structure. Low PAE score (dark green) for well-defined relative positions of the two residues; high PAE score (light green) for unreliable relative positioning (C)–(E) Individual domains of hTHAP9. **(C)** Green - THAP domain (domain boundaries: 1–89), **(D)** Red - RNase H-like catalytic domain (domain boundaries: 153–675), **(E)** Yellow - Predicted C-Terminal domain (domain boundaries: 681–865). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

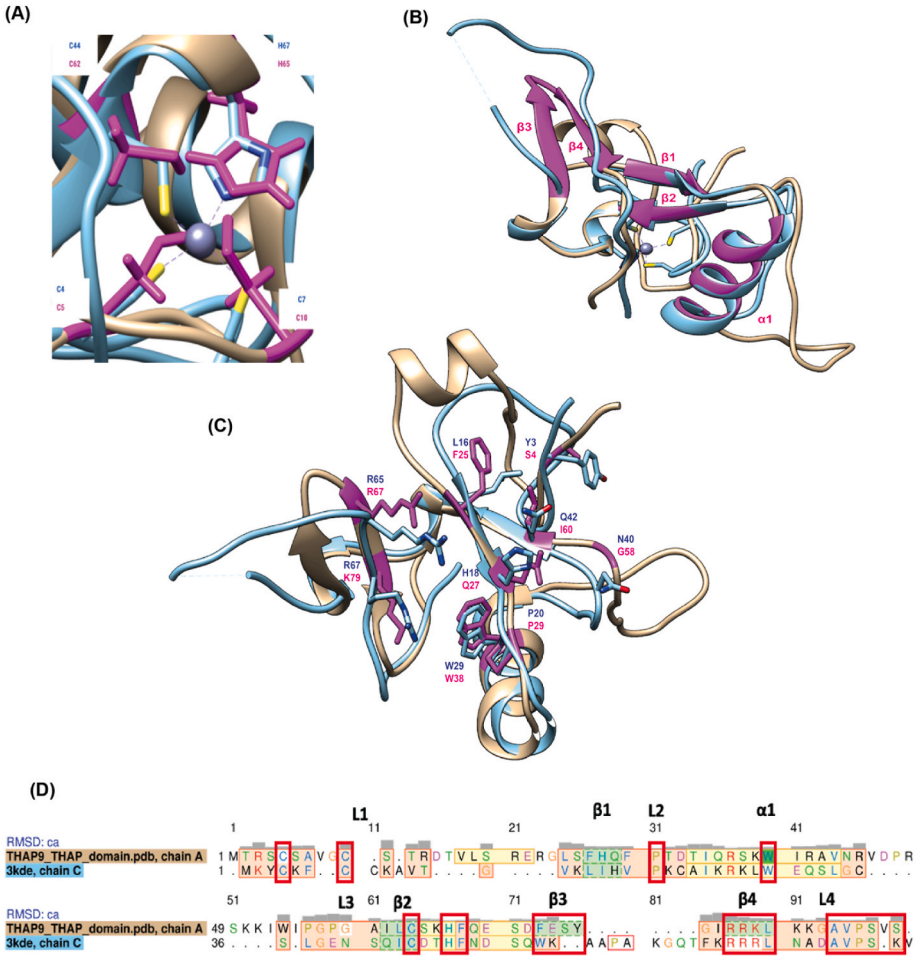


Fig. 6. Comparison between THAP domains of hTHAP9 (predicted) and DmTNP (PDB ID 3KDE chain C). Structural elements [hTHAP9 (tan), DmTNP (cyan)] and important residues of [hTHAP9 (magenta), DmTNP (blue) are highlighted. Structural superimposition of THAP domain (A) C2CH motif (B) β-α-β fold. hTHAP9 structural elements (magenta) include β1-α1-β2 and newly identified hairpin structure (β3 and β4) in loop 4 region (C) DNA-interacting residues in DmTNP superposed on corresponding residues in hTHAP9. (D) Structure-based sequence alignment of THAP domains was performed using Chimera v1.14. Residues in the α-helices (α1), β-strands (β1-β4) and loops (L1-L4) are indicated by yellow, green and orange respectively. Important residues like conserved C2CH, residues aligning with DNA interacting residues of DmTNP, newly identified hairpin structure in loop4 (in 5 THAP proteins), and other conserved residues are highlighted with red boxes. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

conserved secondary structure, namely a characteristic β-α-β fold with four loops (L1-L4) interconnecting the sheets and the helix (Campagne et al., 2010). L1-L4 are flexible regions that may contribute towards DNA-binding specificity. Structure-based multiple sequence alignment of the THAP domains of DmTNP, human THAP1, 2, 7, 9, and 11, and *C. elegans* CtBP by (Sabogal et al., 2010) reported the conservation of zinc-coordinating C2CH motif and base-specific DNA-binding residues (corresponding hTHAP9 residues are C5, C10, C62, H65, P29, W38, V85, and P86 (Figs. 6 and 8).

Comparison of the structures of the THAP domains of hTHAP9 (AF-predicted) and DmTNP crystal structure (3KDE chain C) determined that the root mean square deviations (RMSD) is high (1.004 Å) (Table 2). We

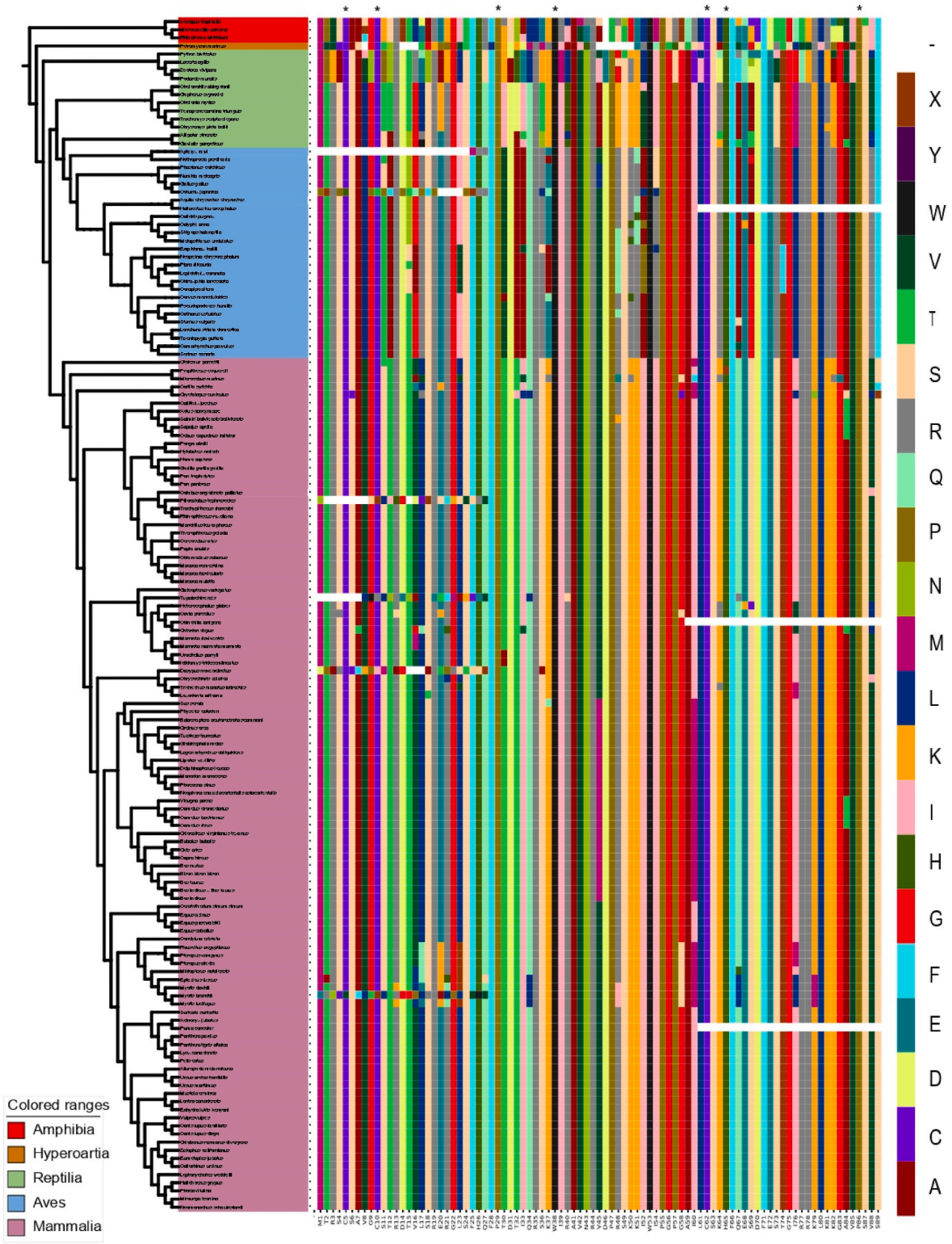


Fig. 7. Evolution of THAP domain in THAP9 orthologs. Conserved functional residues (C5, C10, C62, H65; P29, W38, V85, and P86 in hTHAP9) have been highlighted with a * mark at the top.

performed structural superimposition (Fig. 6 A, B, C) followed by structure-based sequence alignment (Fig. 6D) of the two structures using UCSF Chimera.

C2CH motif: The C2CH zinc-coordinating motif appeared to be structurally conserved between the two homologs (Fig. 6 (A)), with a perfect alignment of the zinc coordinating residues of the C2CH motif, wherein C4, C7, C44 and H67 (shown in blue) of DmTNP aligned with C5, C10, C62 and H65 (shown in magenta) of hTHAP9. It has been reported that mutations of these conserved residues may disrupt protein folding leading to disrupted DNA binding (Campagne et al., 2010; Sabogal et al., 2010).

β-α-β fold: As expected, the AF-predicted model of the hTHAP9-THAP domain consisted of the conserved β-α-β fold (β1-α1-β2) and four loops (L1-L4) (Fig. 6 (B)). Fig. 6 (D) shows the structure-based sequence alignment of the THAP domains of hTHAP9 and DmTNP with highlighted beta-sheets (green) and helices (yellow). The predicted loop regions (orange), namely L1, L2, L3, and L4 regions (low confidence, increased RMSD) may be flexible intrinsically disordered regions (Fig. 6 (D)). Interestingly, the predicted THAP domain structure of hTHAP9 also exhibited a novel hairpin formed by two newly identified β sheets (represented as β3 and β4 in Fig. 6 (B)) in the loop 4 region. Such hairpin structures have not been previously identified in well-studied THAP domains. Further experimental investigations can validate this novel structural element and probe its possible role in DNA interactions.

THAP domains have been reported to recognize and bind consensus DNA motifs (e.g.,'TXXGGGX (A/T)') via a bipartite mechanism (Sabogal et al., 2010). For example, the DmTNP THAP domain uses its N-Terminal β-sheet (residues H18, Q42) to bind to the DNA major groove and the 'RXR' sequence (where 'X' can be any amino acid; R65 and R67 in DmTNP) in its C-terminal Loop 4 for making specific contacts with the DNA minor groove. Further, the C-Terminal 'AVPTIF' motif is essential for placing the minor groove binding residues in optimal orientation for DNA binding. Moreover, DmTNP residues L16, N40 and Y3 are main chain contacting residues (Sabogal et al., 2010). It has been suggested that variations in the length and sequence of β-sheet and loop 4 may be responsible for sequence-specific DNA binding. We studied the conservation of these important DNA-binding residues in the Alphafold predicted structure of the hTHAP9-THAP domain. Fig. 6 (C) shows the alignment of the DNA interacting residues of DmTNP (blue) with corresponding hTHAP9 residues (magenta). It was observed that only P29, W38 and R77 (of hTHAP9) were conserved. Interestingly the AVPTIF motif lies on a previously unreported β-sheet (represented as β4 in Fig. 6 (B) & (D)) instead of loop 4 in hTHAP9.

Further, our evolutionary analysis showed that the THAP-domain is present in all amphibian (3) and hyperoartia (1) orthologs and in most orthologs in reptiles (12 out of 13), birds (26 out of 53), and mammals (105 out 108). The THAP domain shows high conservation within these organism classes (Fig. 7, Supplementary Table 3). In a few organisms, the THAP domain was truncated (N Terminal deletion: *Apteryx rowi, Piliocolobus tephrosceles, Tupaia Chinensis*; C Terminal deletion: *Haliaeetus leucocephalus, Chinchilla lanigera, Puma concolor*).

We investigated the conservation of the C2CH motif and base-specific DNA-binding residues across THAP9 orthologs (Figs. 6 and 7). It was observed that within the 147 orthologs containing THAP domains, C5 is conserved in 142 orthologs (mutated in 3 mammals and 2 aves), C10 is conserved in 141 orthologs (mutated in 4 mammals and 2 aves), C62 is conserved in 144 orthologs (mutated in 2 mammals and 1 aves), H65 is conserved in 143 orthologs (mutated in 3 mammals and 1 aves), V85 is conserved in 137 orthologs (mutated in 2 mammals, 1 aves, 4 reptiles, and 3 amphibians), P86 is conserved in 144 orthologs (mutated in 2 mammals and 1 aves). Interestingly, P29 and W38 are conserved in all 147 orthologs. Moreover, the residues predicted to interact with major and minor grooves of DNA showed distinct patterns within bats, camels, whales, rodents & lizards (Supplementary Table 3).

Leucine Zipper Domain: hTHAP9 has a ~40 amino acid long leucine-rich region located downstream of the DNA-binding THAP domain (Sanghavi et al., 2019). This was corroborated by our ScanProsite analysis which also predicted a highly conserved L-Zipper domain in the same region. Several DNA transposases like DmTNP (Ghanim et al., 2019), IS911 (Haren et al., 1998), and KP repressor (Lee et al., 1996) (inhibitor of DmTNP) form multimers using the leucine zipper region. In another study, it was observed that mutating the leucines (L90, L128, L132, L139, L146, and L153 to Ala, either individually or together) or deleting the leucine-rich predicted coiled-coil region in hTHAP9 did not disrupt homo-oligomerization (Sanghavi and Majumdar, 2021). This suggests that maybe THAP9 utilizes a multidomain mechanism for oligomerization. So, we decided to look at the evolution of the L-Zipper domain of THAP9 to get a further understanding of its conservation.

We observed that the L-Zipper Domain is not present in any of the amphibians, hyperoartia, and reptilian orthologs of THAP9. Moreover, it is only present in 14 out of 53 aves orthologs and 91 out of 108 mammalian orthologs. Thus, it is interesting to note that the L-Zipper domain appeared much later during vertebrate evolution and is highly conserved across mammals. Moreover, L132, L139, L146, and L153 (residue numbers correspond to hTHAP9, Supplementary Table 4, highlighted with * in Fig. 8) are conserved in all the orthologs in which L-Zipper is detected.

RNase H-like fold: A recent study (Sharma et al., 2021) predicted an RNase H-like fold in hTHAP9 using secondary structure predictions, homology modeling, and multiple sequence alignment. The RNase H-like fold is characterised by a catalytic triad (DDD/E residues) which coordinates a Mg^{2+} ion to create an active site essential for DNA cleavage and strand transfer during transposition in DmTNP (D230, D303 & E531 form the catalytic triad) and other DDE/D transposases (Hickman et al., 2010; Nesmelova and Hackett, 2010).

Fig. 9 (A) shows the structural superimposition of the catalytic domain of the AF-predicted hTHAP9 structure with DmTNP (6P5A chain A). Two of the three DmTNP catalytic residues, D303 and E531, align with acidic residues namely D374 and E613 of hTHAP9 while D230 aligns with a Lys instead. The DmTNP structure shows two negatively charged residues D230 and D303 coordinating the Mg^{2+} ion. But in hTHAP9, one of the negatively charged Aspartate (D) residues has been replaced by positively charged Lysine (K282) residue in the catalytic triad. It is tempting to speculate that this change of negatively charged Mg^{2+} coordinating residue to a positively charged residue may disrupt the coordination of Mg^{2+} ion essential for DNA excision and transposition. However, studies have reported that hTHAP9 is still catalytically active, suggesting the possibility of other interactions replacing the interaction between K282 and Mg^{2+}.

It has also been demonstrated that other acidic residues in the hTHAP9 catalytic domain, i.e. D304, D414, D519 and D695, have a role in DNA excision (Sharma et al., 2021). However, Fig. 9 (B), which highlights these residues, suggests that despite their impact on hTHAP9's catalytic activity, they do not appear to lie in the vicinity of the putative catalytic triad.

We then investigated the evolution of these important catalytic residues i.e. K282, D304, D374, D414, D519, E613, D695 & E776 (Fig. 10, Supplementary Table 5). Interestingly, while most of the catalytic residues were conserved across all THAP9 orthologs, D374 and D695 were only conserved across the mammalian orthologs and exhibited distinct class-specific substitutions in other classes. It is tempting to speculate that such class-specific conservation patterns of key catalytic residues play important roles in the modification of THAP9's function.

Insertion domain: The RNase H fold in hTHAP9 is disrupted by an insertion domain between residues S415 to T604 (Sharma et al., 2021). An insertion domain also disrupts the RNase H-like fold of DmTNP (Ghanim et al., 2019); this is responsible for binding GTP via GTP-binding motifs. We compared the structures of the insertion domains of DmTNP and hTHAP9 (Fig. 11).

Though the overall structure of the insertion domains (Fig. 11 (A)) look similar, the RMSD between the two domains is 1.8 A which suggests considerably high dissimilarity. The DmTNP GTP binding domain

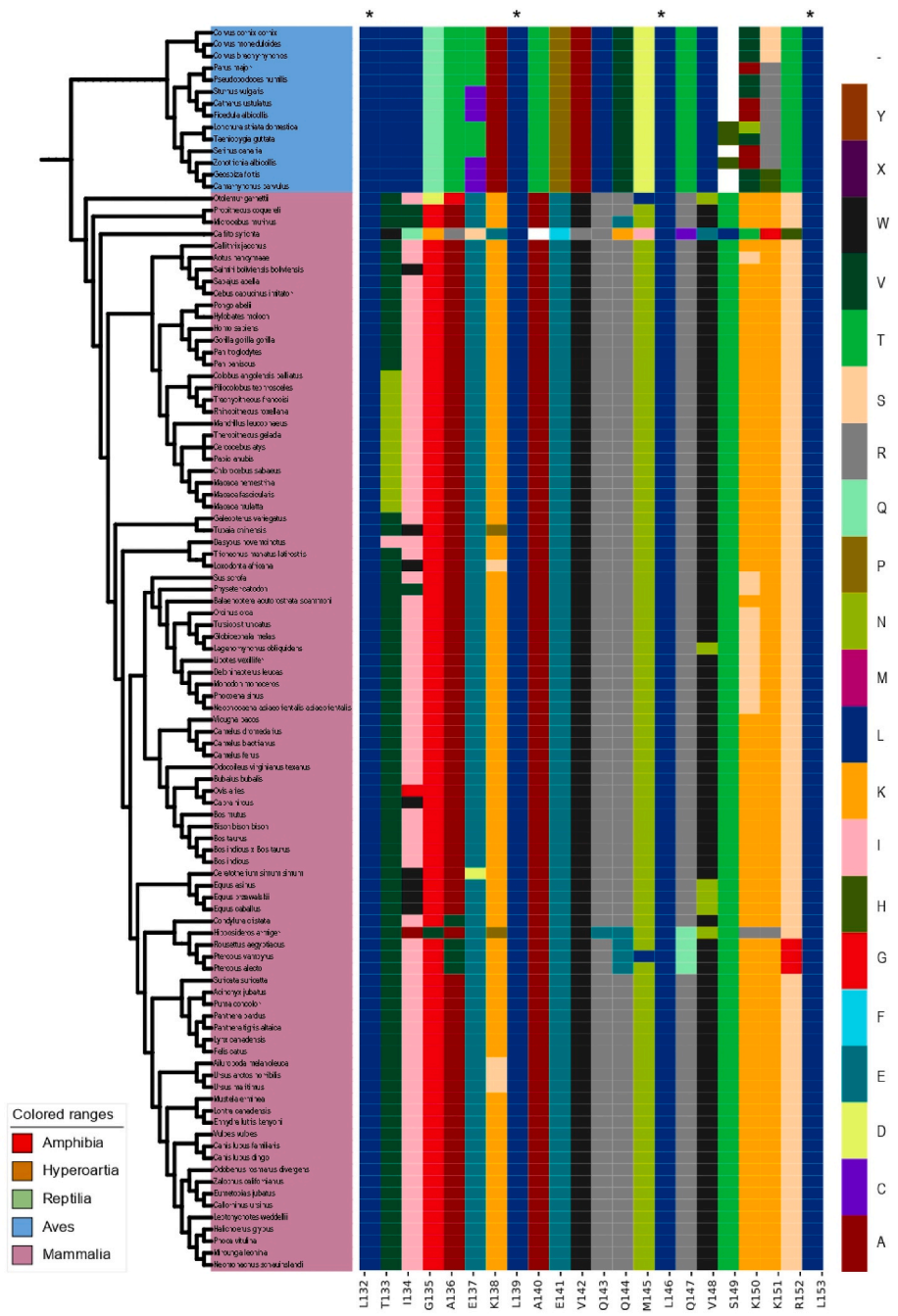


Fig. 8. Evolution of Leucine Zipper Domain in THAP9 orthologs. L132, L139, L146, and L153 (residue numbers correspond to hTHAP9) highlighted with *.

consists of five G-Boxes (G1-G5 boxes marked in Fig. 11(B) and (C)). Mutating residues in these G Boxes (W266G/Q268S/V352D/D353V) affected DmTNP's ability to bind GTP and carry on transposition (Mul and Rio, 1997). In Fig. 11 (B) and **(C)**, we mark the hTHAP9 region (magenta) that aligns with the DmTNP G-Boxes (yellow) and observed that all the DmTNP G-Boxes aligned well with hTHAP9 (conservation of secondary structure but not primary sequence), except the G4-box (loop in DmTNP, two helices separated by a loop in hTHAP9).

In the DmTNP structure, these G-Boxes do not bind GTP (yellow in Fig. 11 (B)). Instead, a different set of residues (K400, K385, V401, D444, F443, N447, D528) (highlighted in blue in Fig. 11 (B), marked with blue boxes in Fig. 11 (C)) interact with GTP during transposition (Ghanim et al., 2019). Structure-based sequence alignment of the DmTNP and hTHAP9 insertion domains demonstrate that the corresponding hTHAP9 residues are C416, Q480, S483, E484, S485, N522, R524, K532, Y605 and D610 (Fig. 11 (C)); of which only N522, D519, F518, D610 are conserved (align with N447, D444, F443 and D528 of DmTNP respectively, highlighted in magenta boxes in Fig. 11 (C)). Interestingly, D519 (hTHAP9, also highlighted in Fig. 11 (B)) as well as the corresponding D444 (in DmTNP) has been shown to play an important role in DNA excision and integration (Sharma et al., 2021). The rest of the residues, despite having similar secondary structures Fig. 11 (C), do not align well.

We then investigated the evolution of the predicted insertion domain of THAP9 (lies between S415 to T604 of hTHAP9) (Fig. 12, Supplementary Table 6). Interestingly, this domain exhibited weak conservation across orthologs.

C-Terminal Domain - In addition to the previously identified domains in hTHAP9, (i.e., THAP domain, L-Zipper domain, RNase H-like catalytic domain), we observed a novel structure at the carboxy-terminal end of

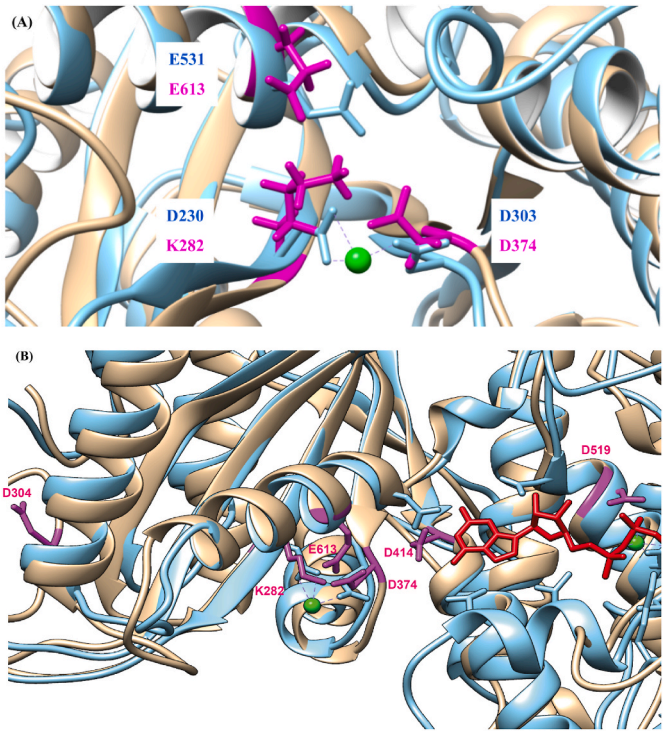


Fig. 9. Comparison between catalytic domains of hTHAP9 (predicted) and DmTNP (PDB ID 3KDE chain C). Structural elements [hTHAP9 (tan), DmTNP (cyan)] and important residues of [hTHAP9 (magenta), DmTNP (blue) are highlighted. Structural superimposition of (A) Catalytic triad (B) Catalytic domain highlighting other acidic residues (hTHAP9) important for DNA excision (Sharma et al., 2021). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

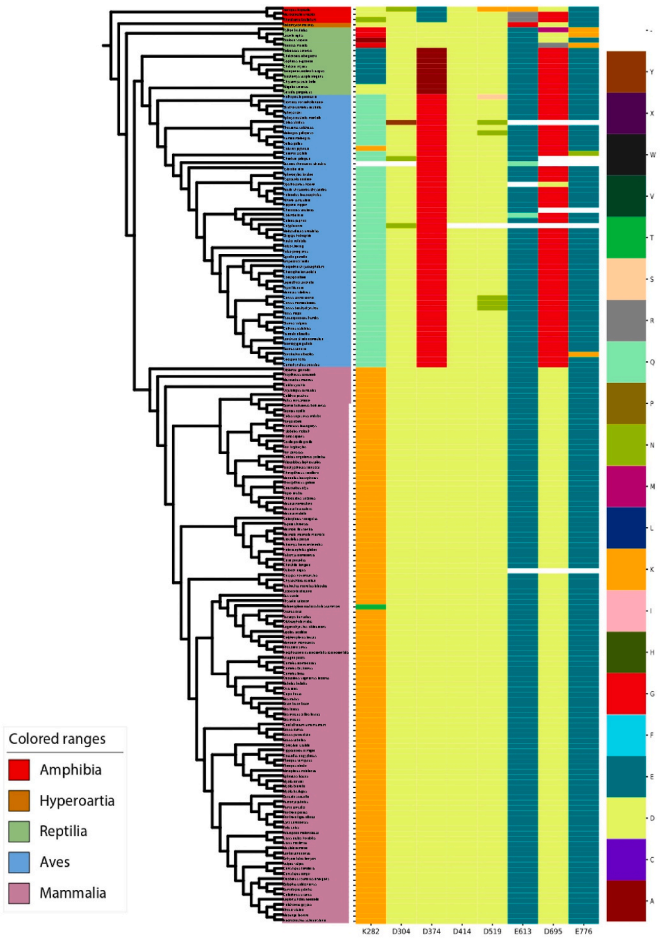


Fig. 10. Evolution of catalytic residues of the RNase H-like catalytic domain in THAP9 orthologs. D304, D374, D414, D519, E613, D695 & E776 are the important catalytic residues of hTHAP9.

the protein (Fig. 13). This C-Terminal Domain (between residues 681–865) has been observed for the first time in the AlphaFold predicted structure of hTHAP9. Interestingly, this region did not align with any DmTNP region during structural alignment. Given the highly organized structure of this novel domain (Fig. 5 (E)), it is tempting to speculate that it has a unique role that is specific to hTHAP9.

4. 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). THAP9 possesses a C2CH type DNA binding THAP domain which is shared between THAP domain-containing proteins (human THAP proteins, CDC14, CTBP1, Lin36, Lin15B, etc.), DmTNP, and zebrafish Pdre2 (Hammer et al., 2005; Hagemann and Hammer, 2006; Majumdar and Rio, 2015). P element-like transposable elements or P element transposase-like genes are present in other eukaryotes including the sea squirt Ciona, sea urchin, and hydra (Chapman et al., 2010; Kimbacher et al., 2009).

In this study, we conducted an evolutionary analysis and extensive *in silico* characterization of THAP9 and its orthologs from 178 organisms, using available THAP9 sequence data in NCBI. We observed that the orthologs of THAP9 had remarkable conservation across species. We also observed some distinct evolutionary patterns within the individual domains of THAP9. For instance, for most THAP9 domains, the sequence as well as order of occurrence within the protein, was highly conserved within the mammalian orthologs, thus implying their possible functional importance in mammals. It was also observed that while the zinc-coordinating C2CH motif and base-specific DNA-binding residues of the THAP domain were highly conserved, class-specific distinct patterns were observed in other domains. According to NCBI, in addition to the studied orthologs, THAP9 also has homologs in chimpanzees, Rhesus monkeys, dogs, cows, chickens, and frogs. In the future, it will be interesting to compare the THAP9 orthologs and homologs and also investigate the role of THAP9 in organisms that have both THAP9 homolog and ortholog sequences.

We observed some previously unreported functional features in the hTHAP9 protein sequence. These included a highly conserved pattern in mammalian orthologs consisting of 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). We speculate that this region may have a role in the subcellular localization of THAP9. We also identified several disordered binding regions (DBR) in hTHAP9 (Supplementary Table 1); the DBR region with the highest IUPRED (Dosztányi, 2018) score is between residues 1–180 and overlaps with the THAP domain and L-Zipper region. These regions warrant further investigation since DBRs are typically involved in protein-protein (Mészáros et al., 2012) or intramolecular interactions (Stein et al., 2009). Moreover, predicted short linear motifs (Supplementary Table 2) in hTHAP9 may be involved in binding regulatory molecules like casein kinase 1 and 2 or HCF1.

Since there is limited information available about the experimentally derived protein structures of THAP9 and THAP family proteins, we used the structural information in the AlphaFold database to acquire insights into the structure of human THAP9. As predicted from its primary sequence homology with DmTNP, the hTHAP9 structure appeared to have distinct domains namely a THAP domain at the N-Terminal end and an RNase H-like catalytic domain disrupted by an insertion domain.

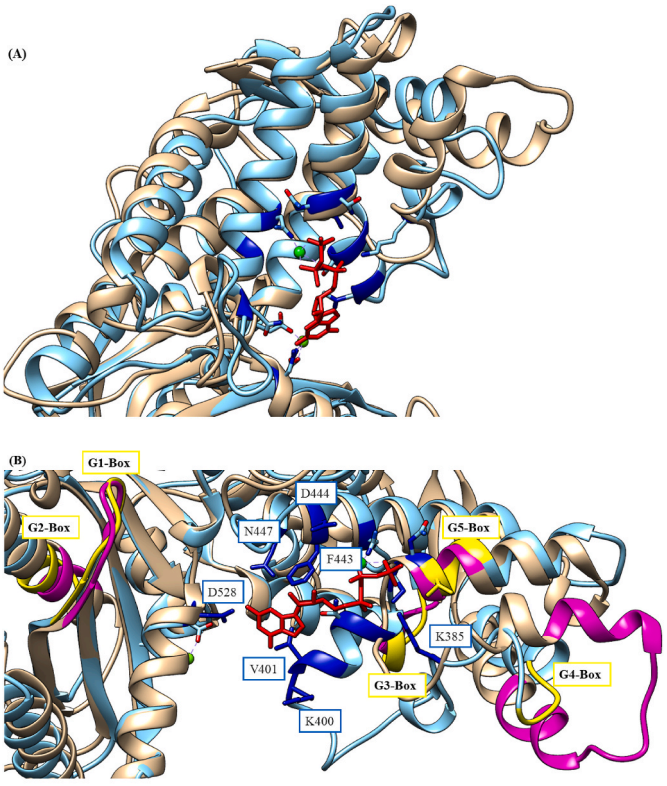


Fig. 11. Insertion Domains of hTHAP9 and DmTNP. hTHAP9 (AF-predicted, tan); DmTNP (PDB 6P5A, cyan; bound to GTP (red) and Mg2+ (green), GTP-interacting residues (blue, labeled in B) (A) Superimposed insertion domains of hTHAP9 and DmTNP (B) Comparison between GTP binding region of DmTNP and corresponding region in hTHAP9. Previously reported DmTNP G-Domain (G1-G5 Boxes, yellow); corresponding hTHAP9 region (magenta) (C) Structure-based sequence alignment of the two insertion domains; helix (yellow), loop (orange), G-Boxes (red boxes), GTP-interacting residues (blue boxes), corresponding hTHAP9 residues (pink boxes). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

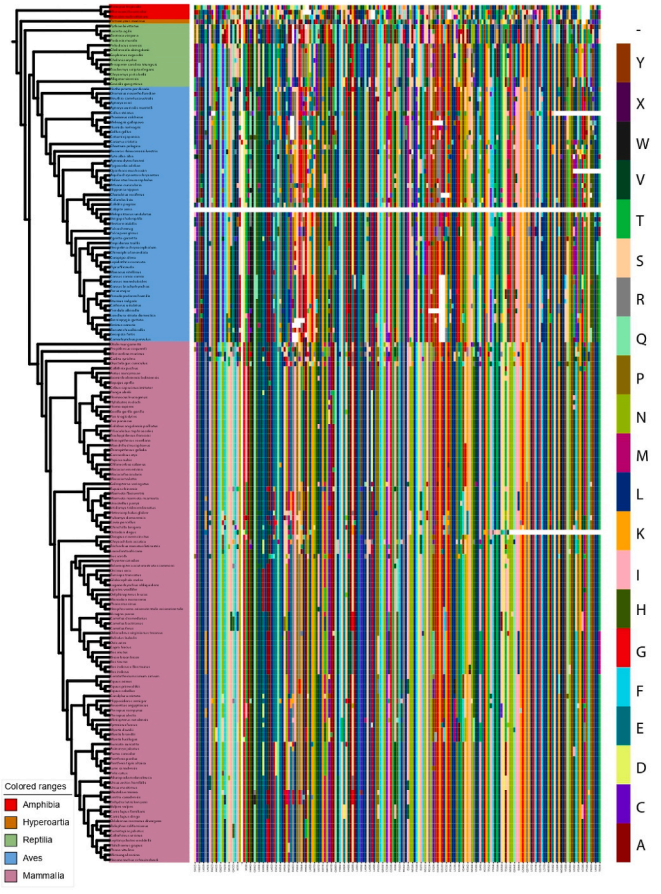


Fig. 12. Evolution of predicted insertion domain in THAP9 orthologs.

Fig. 13. Structure of C-Terminal domain of hTHAP9 predicted by Alphafold.

We observed that the hTHAP9 THAP domain has a novel hairpin structure in the loop 4 region; it is interesting to note that the corresponding region in DmTNP is essential for sequence specific DNA binding. Moreover, the carboxy-terminal region of hTHAP9 appears to have a distinct structural fold. Thus, our study shed light on structural and evolutionary facets of THAP9, a relatively less understood TE in the human genome, and set the stage for further studies to characterize its role.

Funding

SERB (Science and Engineering Research Board, Government of India) grant ECR/2016/000479 and DBT Ramalingaswami Fellowship BT/RLF/Re-entry/43/2013 and BT/PR16074/BID/7/569/2016, GSBTM (Gujarat State Biotechnology Mission), CISCO USA CG# 2207376, IIT Gandhinagar

CRediT authorship contribution statement

Richa Rashmi: Conceptualization, Methodology, Data curation, Investigation, Validation, Visualization, Writing – original draft. **Chandan Nandi:** Data curation, Investigation, Validation. **Sharmistha Majumdar:** Supervision, Conceptualization, Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sharmistha Majumdar reports financial support was provided by Indian Institute of Technology Gandhinagar.

Data availability

Shared link in manuscript

Appendix A Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crstbi.2023.100113.

References

Akdel, M., et al., 2021. A Structural Biology Community Assessment of AlphaFold 2 Applications. bioRxiv, p. 2021. https://doi.org/10.1101/2021.09.26.461876, 09.26.461876.

Bailey, T.L., et al., 2009. Meme suite: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208. https://doi.org/10.1093/nar/gkp335 (Web Server issue).

Balakrishnan, M.P., et al., 2009. 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 (2), H643–H653. https://doi.org/10.1152/ajpheart.00234.2009.

Berman, H.M., et al., 2000. The protein Data Bank. Nucleic Acids Res. 28 (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

Bessière, D., et al., 2008. Structure-function analysis of the THAP zinc finger of THAP1, a large C2CH DNA-binding module linked to Rb/E2F pathways. J. Biol. Chem. 283 (7), 4352–4363. https://doi.org/10.1074/jbc.M707537200.

Bourque, G., et al., 2018. Ten things you should know about transposable elements. Genome Biol. 19 (1), 199. https://doi.org/10.1186/s13059-018-1577-z.

Buchan, D.W.A., Jones, D.T., 2019. The PSIPRED protein analysis Workbench: 20 years on. Nucleic Acids Res. 47 (W1), W402–W407. https://doi.org/10.1093/nar/gkz297.

Burkhard, P., Stetefeld, J., Strelkov, S.V., 2001. Coiled coils: a highly versatile protein folding motif. Trends Cell Biol. 11 (2), 82–88. https://doi.org/10.1016/s0962-8924(00)01898-5.

Campagne, S., et al., 2010. Structural determinants of specific DNA-recognition by the THAP zinc finger. Nucleic Acids Res. 38 (10), 3466–3476. https://doi.org/10.1093/nar/gkq053.

Chapman, J.A., et al., 2010. The dynamic genome of Hydra. Nature 464 (7288), 592–596. https://doi.org/10.1038/nature08830.

Chu, E.C., Tarnawski, A.S., 2004. PTEN regulatory functions in tumor suppression and cell biology. Med. Sci. Mon. Int. Med. J. Exp. Clin. Res.: International Medical Journal of Experimental and Clinical Research 10 (10), RA235–241.

de Castro, E., et al., 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 34, W362–W365. https://doi.org/10.1093/nar/gkl124. Web Server issue).

De Souza Santos, E., et al., 2008. Silencing of LRRC49 and THAP10 genes by bidirectional promoter hypermethylation is a frequent event in breast cancer. Int. J. Oncol. 33 (1), 25–31.

Dehaene, H., 2019. THAP Proteins in the Transcriptional Control of Cell Proliferation. Université de Lausanne. Faculté de biologie et médecine. Available at: https://serval.unil.ch/notice/serval:BIB_02ABC667C655. (Accessed 3 March 2022).

Dehaene, H., et al., 2020. THAP11F80L cobalamin disorder-associated mutation reveals normal and pathogenic THAP11 functions in gene expression and cell proliferation. PLoS One 15 (1), e0224646. https://doi.org/10.1371/journal.pone.0224646.

Dejosez, M., et al., 2010. Ronin/Hcf-1 binds to a hyperconserved enhancer element and regulates genes involved in the growth of embryonic stem cells. Genes & Development 24 (14), 1479–1484. https://doi.org/10.1101/gad.1935210.

Dosztányi, Z., 2018. Prediction of protein disorder based on IUPred. Protein Sci. : A Publication of the Protein Society 27 (1), 331–340. https://doi.org/10.1002/pro.3334.

El-Gebali, S., et al., 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47 (D1), D427–D432. https://doi.org/10.1093/nar/gky995.

Freiman, R.N., Herr, W., 1997. Viral mimicry: common mode of association with HCF by VP16 and the cellular protein LZIP. Genes & Development 11 (23), 3122–3127. https://doi.org/10.1101/gad.11.23.3122.

Gervais, V., et al., 2013. NMR studies of a new family of DNA binding proteins: the THAP proteins. J. Biomol. NMR 56 (1), 3–15. https://doi.org/10.1007/s10858-012-9699-1.

Ghanim, G.E., et al., 2019. Structure of a P element transposase-DNA complex reveals unusual DNA structures and GTP-DNA contacts. Nat. Struct. Mol. Biol. 26 (11), 1013–1022. https://doi.org/10.1038/s41594-019-0319-6.

Hagemann, S., Hammer, S.E., 2006. The implications of DNA transposons in the evolution of P elements in zebrafish (Danio rerio). Genomics 88 (5), 572–579. https://doi.org/10.1016/j.ygeno.2006.06.010.

Hall, B.G., 2013. Building phylogenetic trees from molecular data with MEGA. Mol. Biol. Evol. 30 (5), 1229–1235. https://doi.org/10.1093/molbev/mst012.

Hammer, S.E., Strehl, S., Hagemann, S., 2005. Homologs of Drosophila P transposons were mobile in zebrafish but have been domesticated in a common ancestor of chicken and human. Mol. Biol. Evol. 22 (4), 833–844. https://doi.org/10.1093/molbev/msi068.

Haren, L., et al., 1998. Multiple oligomerisation domains in the IS911 transposase: a leucine zipper motif is essential for activity. J. Mol. Biol. 283 (1), 29–41. https://doi.org/10.1006/jmbi.1998.2053.

Hickman, A.B., Chandler, M., Dyda, F., 2010. Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. Crit. Rev. Biochem. Mol. Biol. 45 (1), 50–69. https://doi.org/10.3109/10409230903505596.

Jumper, J., et al., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kimbacher, S., et al., 2009. Drosophila P transposons of the urochordata Ciona intestinalis. Mol. Genet. Genom.: MGG 282 (2), 165–172. https://doi.org/10.1007/s00438-009-0453-7.

Kumar, S., et al., 2017. TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. 34 (7), 1812–1819. https://doi.org/10.1093/molbev/msx116.

Kumar, S., et al., 2018. Mega X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35 (6), 1547–1549. https://doi.org/10.1093/molbev/msy096.

Kumar, M., et al., 2020. ELM-the eukaryotic linear motif resource in 2020. Nucleic Acids Res. 48 (D1), D296–D306. https://doi.org/10.1093/nar/gkz1030.

Lee, C.C., Mul, Y.M., Rio, D.C., 1996. The Drosophila P-element KP repressor protein dimerizes and interacts with multiple sites on P-element DNA. Mol. Cell Biol. 16 (10), 5616–5622. https://doi.org/10.1128/MCB.16.10.5616.

Leite, K.R.M., et al., 2013. MicroRNA 100: a context dependent miRNA in prostate cancer. Clinics 68, 797–802. https://doi.org/10.6061/clinics/2013(06)12.

Letunic, I., Bork, P., 2019. Interactive Tree of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47 (W1), W256–W259. https://doi.org/10.1093/nar/gkz239.

Letunic, I., Khedkar, S., Bork, P., 2021. SMART: recent updates, new developments and status in 2020. Nucleic Acids Res. 49 (D1), D458–D460. https://doi.org/10.1093/nar/gkaa937.

Liew, C.K., et al., 2007. Solution structure of the THAP domain from Caenorhabditis elegans C-terminal binding protein (CtBP). J. Mol. Biol. 366 (2), 382–390. https://doi.org/10.1016/j.jmb.2006.11.058.

Lu, R., et al., 1998. The herpesvirus transactivator VP16 mimics a human basic domain leucine zipper protein, luman, in its interaction with HCF. J. Virol. 72 (8), 6291–6297. https://doi.org/10.1128/JVI.72.8.6291-6297.1998.

Majumdar, S., Rio, D.C., 2015. P transposable elements in Drosophila and other eukaryotic organisms. Microbiol. Spectr. 3 (2) https://doi.org/10.1128/microbiolspec.MDNA3-0004-2014. MDNA3-0004–2014.

Majumdar, S., Singh, A., Rio, D.C., 2013. The human THAP9 gene encodes an active P-element DNA transposase. Science (New York, N.Y.) 339 (6118), 446–448. https://doi.org/10.1126/science.1231789.

Mariani, V., et al., 2013. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29 (21), 2722–2728. https://doi.org/10.1093/bioinformatics/btt473.

Mazars, R., et al., 2010. The THAP-zinc finger protein THAP1 associates with coactivator HCF-1 and O-GlcNAc transferase: a link between DYT6 and DYT3 dystonias. J. Biol. Chem. 285 (18), 13364–13371. https://doi.org/10.1074/jbc.M109.072579.

Mészáros, B., Dosztányi, Z., Simon, I., 2012. Disordered binding regions and linear motifs–bridging the gap between two models of molecular recognition. PLoS One 7 (10), e46829. https://doi.org/10.1371/journal.pone.0046829.

Mul, Y.M., Rio, D.C., 1997. Reprogramming the purine nucleotide cofactor requirement of Drosophila P element transposase in vivo. EMBO J. 16 (14), 4441–4447. https://doi.org/10.1093/emboj/16.14.4441.

Nagoshi, E., et al., 1999. 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 (7), 2221–2233. https://doi.org/10.1091/mbc.10.7.2221.

Nesmelova, I.V., Hackett, P.B., 2010. DDE transposases: structural similarity and diversity. Adv. Drug Deliv. Rev. 62 (12), 1187–1195. https://doi.org/10.1016/j.addr.2010.06.006.

Parker, J.B., et al., 2012. A transcriptional regulatory role of the THAP11-HCF-1 complex in colon cancer cell function. Mol. Cell Biol. 32 (9), 1654–1670. https://doi.org/10.1128/MCB.06033-11.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet.: TIG (Trends Genet.) 16 (6), 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.

Richter, A., et al., 2017. In-depth characterization of the homodimerization domain of the transcription factor THAP1 and dystonia-causing mutations therein. J. Mol. Neurosci. 62 (1), 11–16. https://doi.org/10.1007/s12031-017-0904-2.

Robbins, J., et al., 1991. Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: identification of a class of bipartite nuclear targeting sequence. Cell 64 (3), 615–623. https://doi.org/10.1016/0092-8674(91)90245-T.

Roussigne, M., et al., 2003. The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. Trends Biochem. Sci. 28 (2), 66–69. https://doi.org/10.1016/S0968-0004(02)00013-0.

Sabogal, A., et al., 2010. THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves. Nat. Struct. Mol. Biol. 17 (1), 117–123. https://doi.org/10.1038/nsmb.1742.

Sanghavi, H.M., Majumdar, S., 2021. Oligomerization of THAP9 transposase via amino-terminal domains. Biochemistry 60 (23), 1822–1835. https://doi.org/10.1021/acs.biochem.1c00010.

Sanghavi, H.M., Mallajosyula, S.S., Majumdar, S., 2019. Classification of the human THAP protein family identifies an evolutionarily conserved coiled coil region. BMC Struct. Biol. 19 (1), 4. https://doi.org/10.1186/s12900-019-0102-2.

Sharma, V., Thakore, P., Majumdar, S., 2021. THAP9 transposase cleaves DNA via conserved acidic residues in an RNaseH-like domain. Cells 10 (6), 1351. https://doi.org/10.3390/cells10061351.

Sheldon, L.A., Kingston, R.E., 1993. Hydrophobic coiled-coil domains regulate the subcellular localization of human heat shock factor 2. Genes & Development 7 (8), 1549–1558. https://doi.org/10.1101/gad.7.8.1549.

Stein, A., et al., 2009. Dynamic interactions of proteins in complex networks: a more structured view. FEBS J. 276 (19), 5390–5405. https://doi.org/10.1111/j.1742-4658.2009.07251.x.

Su, W., Zuo, T., Peterson, T., 2020. Ectopic expression of a maize gene is induced by composite insertions generated through alternative transposition. Genetics 216 (4), 1039–1049. https://doi.org/10.1534/genetics.120.303592.

Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, W609–W612. https://doi.org/10.1093/nar/gkl315. Web Server issue).

The UniProt Consortium, 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47 (D1), D506–D515. https://doi.org/10.1093/nar/gky1049.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. 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 (22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673.

Tunyasuvunakool, K., et al., 2021. Highly accurate protein structure prediction for the human proteome. Nature 596 (7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1.

Udenwobele, D.I., et al., 2017. Myristoylation: an important protein modification in the immune response. Front. Immunol. 8, 751. https://doi.org/10.3389/fimmu.2017.00751.

Varadi, M., et al., 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 50 (D1), D439–D444. https://doi.org/10.1093/nar/gkab1061.

Ward, J.J., et al., 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. 337 (3), 635–645. https://doi.org/10.1016/j.jmb.2004.02.002.

Wilkins, M.R., et al., 1999. Protein identification and analysis tools in the ExPASy server. Methods Mol. Biol. 112, 531–552. https://doi.org/10.1385/1-59259-584-7:531.

Zargar, Z., Tyagi, S., 2012. Role of host cell factor-1 in cell cycle regulation. Transcription 3 (4), 187–192. https://doi.org/10.4161/trns.20711.