THAP9 transposase cleaves DNA via conserved acidic residues in an RNaseH-like domain

Vasudha Sharma[1], Prachi Thakore[1] and Sharmistha Majumdar[1]*

1 Discipline of Biological Engineering, Indian Institute of Technology Gandhinagar, Gujarat, 382355, India

* Correspondence should be addressed to Sharmistha Majumdar.

Email: sharmistham@iitgn.ac.in

Abstract

The catalytic domain of most 'cut and paste' DNA transposases have the canonical RNase-H fold which is also shared by other polynucleotidyl transferases like retroviral integrases and the RAG1 subunit of V(D)J recombinase. The RNase-H fold is a mixture of beta sheets and alpha helices with three acidic residues (Asp, Asp, Glu/Asp - DDE/D) that are involved in metal-mediated cleavage and subsequent integration of DNA. Human THAP9 (hTHAP9), homologous to the well-studied Drosophila P-element transposase (DmTNP), is an active DNA transposase that, although domesticated, still retains the catalytic activity to mobilize transposons. In this study we have modelled the structure of hTHAP9 using the recently available cryo-EM structure of DmTNP as a template to identify an RNase-H like fold along with important acidic residues in its catalytic domain. Site-directed mutagenesis of the predicted catalytic residues followed by screening for DNA excision and integration activity, led to the identification of candidate Ds and Es in the RNaseH fold that appear to constitute the catalytic triad of hTHAP9.

Significance statement

Many DNA transposases execute DNA excision via a catalytic domain, which has a canonical RNase-H fold. Despite the similar nature of the catalytic domain, these transposases exhibit mechanistically different strategies of transposition. We identify a potential RNase-H fold in hTHAP9 with a conserved DDE motif required for cutting DNA. Additionally, we have found a residue, which when mutated, leads to an increase in hTHAP9's transposition activity. Such hyperactive transposase mutants can be exploited as tools in genome engineering and gene therapy. This study has helped widen our knowledge about the catalytic activity of a functionally uncharacterized transposon-derived gene in the human genome.

Introduction

Transposons are mobile DNA sequences that can move from one genomic location to another using autonomous or non-autonomous enzyme machinery(1). In the course of their propagation

in the host genome, these selfish genetic elements are known to disrupt host cellular functions by ectopic recombination events, insertion mutations and interference with gene regulation (2–4). In some cases, entire transposons or parts of the same are co-opted or 'domesticated' by the host to serve beneficial functions. Numerous genes have recruited promoters, enhancers, DNA binding domains, alternative splice sites, polyadenylation sites and cis-regulatory sequences from an ancestral DNA transposon (5) to make evolutionarily beneficial, functional host genes(2, 4).

A DNA transposon, also known as a 'cut and paste' transposon uses a transposase enzyme, often encoded by itself, to cut or excise itself from one location in the genome and paste or integrate it in a different genomic location, without the need of homology between donor and target DNA sequences(6). DNA transposons contribute to 3% of the human genome and around 50 human genes are derived from these elements(7). TE-derived genes may become immobile and exist as single copies by losing the hallmarks of transposons, i.e., terminal inverted repeats (TIRs) and target site duplications (TSDs)(5). RAG1/2, CENPB, SETMAR, hTHAP9, which are examples of human genes that have domains recruited from DNA transposons(2) have evolved diverse functions. For example, RAG1/2 mediates the DNA rearrangement reactions in V(D)J recombination; CENPB (Centromeric protein B) helps in chromatin assembly at centromeres during cell division; SETMAR is a DNA repair protein while the function of hTHAP9 is still unknown.

All seventeen eukaryotic cut-and-paste transposase superfamilies have a D, D, E (Asp, Asp, Glu) catalytic triad in the active sites of their respective transposases(8). The third residue in the triad can be either a D or E, hence it is represented as DDE/D motif. Topologically, the DDE/D triad is always present in an 'RNase-H like fold' which is a mixture of alpha helices and beta sheets(9). The conserved RNase-H fold has a characteristic secondary structure- β1-β2-β3-α1-β4-α2/3-β5-α4-α5, wherein, the first D of the triad is always present on β1, the second D on or just after β4, and the E/D either on or just before α4 (6). The RNase-H like fold is present in many transposases (Tn3, Tn5, Mu, hAT, Mariner, Sleeping beauty) and retroviral integrases (HIV-1 integrase, RSV integrase), in which the DDE/D motif catalyzes magnesium ion-dependent nucleophilic attack on DNA (9–11).

Wessler *et al* in 2011 had predicted the DDE motif in P superfamily; this has been confirmed in the recently solved cryo-EM structure of *Drosophila* P-element transposase (DmTNP) (8, 12). The P-element is a widely studied transposon in *Drosophila melanogaster that* invaded this species recently and is continuing to spread through the natural fly populations by horizontal gene transfer (13). It was discovered that P-element transposition is responsible for hybrid dysgenesis leading to male sterility in *Drosophila* (14). Ever since their discovery, P-elements have been used as a model to study transposition and as a genetic and genome engineering tool (15, 16). The DmTNP cryo-EM structure has revealed that the RNase H domain is located near the donor-target DNA junctions, with the catalytic residues coordinating a Mg^{2+} ion. GTP is not required for providing energy for the reaction; rather it possibly positions transposon DNA for catalysis by directly interacting with the terminal base of the cut transposon. Other unique features highlighted in the DmTNP transpososome structure are the looping out of the cut strand to maintain DNA integration and the bending of target DNA (12, 17).

The hTHAP9 protein is homologous (25% identical and 40% similar) to the DmTNP. It is present as a single copy in the human genome and lacks transposon hallmarks, i.e., TIRs and TSDs; this suggests that human THAP9 may have undergone molecular domestication and is

immobile, whereas its homolog in zebrafish (Pdre2) appears to be functional since it has intact TIRs and STIRs and multiple copies in the zebrafish genome (16, 18, 19). Although the function of hTHAP9 remains undiscovered, it is catalytically active and can cut DNA flanked by P element's TIRs in human as well as *Drosophila* cells (20). Since hTHAP9 has retained its ability to cut DNA, it is important to investigate the nature of its catalytic domain. Using secondary structure predictions, homology modeling and multiple sequence alignment, we identify an RNase-H like fold in hTHAP9. Further, we experimentally demonstrate that D304, D374 and E613 form the catalytic triad of hTHAP9 and are responsible for DNA excision.

Results

The catalytic domains of all known eukaryotic cut-and-paste transposases have a characteristic RNase-H-like fold with a conserved DDE/D motif. The well-studied Drosophila P element transposase (DmTNP) is a member of the P superfamily, which is one of the seventeen known cut-and-paste TE superfamilies (8, 12). hTHAP9, a homolog of DmTNP, was reported to be an active transposase that can cut P-element DNA flanked with intact inverted repeats (20). Thus, we decided to investigate if hTHAP9 had a RNaseH fold-like catalytic domain (8)

Prediction of secondary structure elements and RNase-H domain in hTHAP9 and its homologs. Many DDE/D transposases, including DmTNP and retroviral integrases have similar catalytic domains(21). The DDE transposases always have an RNase-H fold in their catalytic domain within which the first D lies in **β**1, the second D in or just after **β**4 and third E/D in **α**4 (8, 21).

Using PSIPRED (22), we predicted the secondary structure elements present in full length DmTNP, hTHAP9 and Pdre2. The presence of RNase-H fold was inspected visually by locating the pattern of **β** sheets and **α** helices in the PSIPRED output. The Ds and E that were found in beta sheets (**β**1 and **β**4) and alpha helices (**α**4), are in agreement with the characteristic structure of RNase-H fold in integrases and other DDE/D transposases.

In DmTNP, the reported residues of the catalytic triad, namely D230 is present in **β**1, D303 in **β**4 and E531 in **α**4 (Fig.1). For hTHAP9 and Pdre2, we have identified the catalytic triad by comparing the predicted secondary structure to the well-studied RNase-H fold structure, common in transposases (Rag1, Hermes, Mos1, DmTNP) and integrases (HIV1). We have identified D304 in **β**1, D374 in **β**4 and E613 in **α**4 in hTHAP9. Similarly, we found D367 in **β**1, D374 just after **β**4 and E679 in Pdre2 (Fig.1). The RNase-H folds in DmTNP, hTHAP9 and Pdre2 appear to be disrupted by an insertion domain between **β**5 and **α**4. In DmTNP, the insertion domain has been reported to carry a unique GTP binding domain(12) but the role of insertion domains in hTHAP9 and Pdre2 are unknown.

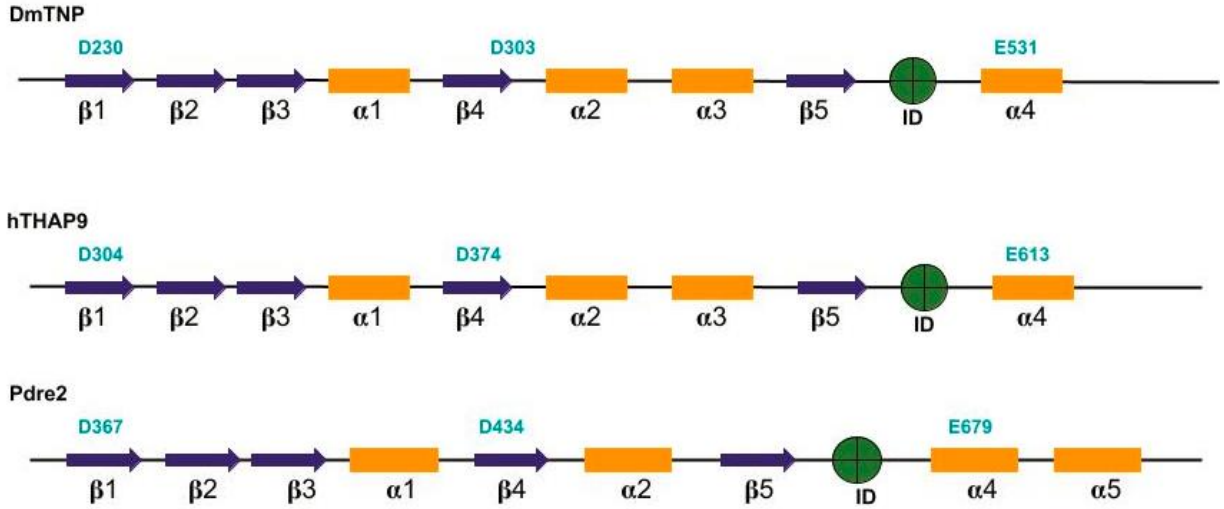

Fig 1: Predicted RNase-H fold in Drosophila, human and zebrafish THAP9. DDE motifs for each homolog are represented in bold (cyan) above its putative location. Blue arrows indicate beta sheets and yellow bars represent alpha helices. Insertion domains for each homolog [DmTNP(residues 339-528) (12), hTHAP9 (residues 415-604), Pdre2 (residues 475-670) are shown in green circles.

Homology modeling of hTHAP9 identified a putative RNaseH-like domain

To further investigate if the secondary structure elements predicted in hTHAP9 were indeed able to fold into an RNaseH-like domain, we made a 3D model of hTHAP9 using the recently solved DmTNP Cryo-EM structure (6p5a) as a template for homology modelling. SWISS-MODEL(23) interactive GUI was used for generating the hTHAP9 model with default parameters and the output structure was docked with magnesium ion using AUTODOCK(24). We then performed energy minimization for the docked structure using MD simulations for 100ns with GROMACS(25) and CHARMM 36 force field(26). The resultant energy minimized model was viewed in PyMol and analyzed using SAVES. 98.2% of the total residues were in allowed regions and only 0.5% in disallowed regions, as revealed by the Ramachandran plot analysis (Fig.S1).

As predicted in Fig.1, D374 and E613 are respectively present in a beta sheet (**β**4) and **α**-helix (**α**4) within the RNaseH fold (Fig 2). The two residues are also proximal to the magnesium ion, forming an active site. Although D304 appeared just after a beta sheet, it was found to be away from the other catalytic residues. As predicted by the SWISS-MODEL secondary structure, D304 lies in the loop region immediately after a **β** sheet of the final 3D model (Fig. S2). It is possible that due to low sequence similarity of hTHAP9 with the template, many predicted residues did not form a proper conformation and ended up in loops.

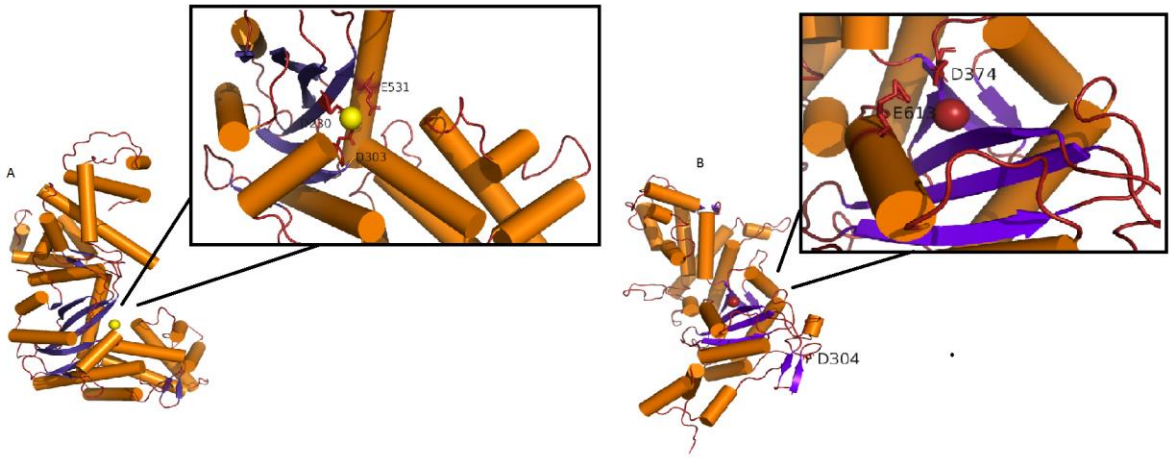

Fig 2: Catalytic triad shown in 3D structure of hTHAP9 built using homology modelling (A) DmTNP structure with beta sheets (blue arrows) and helices (orange rods) (PDB ID - 6P5A(12)). The active site is shown in the inset with DDE motif (D230, D303, E531) in close proximity to the Magnesium ion (light yellow). (B) Homology model of hTHAP9 with magnesium ion. Active site shown in the rectangle has D374 and E613 close to magnesium ion (red). D304 does not adopt any secondary structure but follows immediately after a beta sheet.

Identification of other conserved acidic residues in the catalytic domain of hTHAP9 and its homologs. We used multiple sequence alignment to examine the conserved acidic residues in *Drosophila* P element transposase (DmTNP), Human THAP9 (hTHAP9) and Zebrafish THAP9 (Pdre2). Many residues throughout the sequences, including several Ds and Es were conserved among the three homologs (Fig. 3, marked with black boxes).

Fig. 3: Sequence conservation in the putative RNaseH-domains of DmTNP, hTHAP9 and Pdre2. Amino acid sequences of DmTNP, hTHAP9 and Pdre2 aligned using Clustal Omega and visualized using MView. The Ds and Es that are conserved in the putative RNaseH-domains ofall three homologs are highlighted in black boxes and labeled (in bold on top of the sequence) with the corresponding DmTNP residue number. Coloring scheme of alignment is based on conserved physicochemical properties of amino acids (Colored by identity - Palette P1 of MView, refer to Supplementary Fig.1 for key)..

Signature string: Besides the canonical DDE motif, additional conserved residues in the RNaseH-like catalytic domain have been reported to form a signature string specific to each 'cut and paste' transposase superfamily (8). The residues in a signature string have similar approximate spacing in all the members of the superfamily. For the P superfamily, the signature string (Fig 4) first described by Wessler et al in 2011, carries the catalytic triad (D230, D303 and E531) as well as a [C/D](2)H motif (D339) downstream of the second D of the DDE triad. This motif is always right after β5 in the predicted secondary structure We have identified a similar pattern of conserved amino acids in hTHAP9 and Pdre2 by visually inspecting the sequences in the multiple sequence alignment results (Fig 4).

P-element superfamily	**D230**(~90)**D303**(4)[N/y/h](~30)**D339**(2)H(2)K(2)[R/y][N/t](~190)**E531**(2)[F/n](3)R
THAP9	**D304**(~70)**D374**(4)H(~30)**D414**(2)H(2)R(2)RN(~190)**E613**(1)F(4)R
PDRE2	**D367**(~70)**D434**(4)N(~30)**D474**(2)H(2)K(2)RN(~190)**E679**(2)F(3)R

Fig 4: Signature string of P-element superfamily in hTHAP9 and Pdre2. Signature string of P element superfamily (8) along with similar signature strings identified in human and zebrafish THAP9. Conserved Ds and Es are represented in bold and the numbers represent the spacing between conserved amino acid residues.

The conserved Ds and Es in the RNase-H fold of the three homologs are highlighted in Fig 3 and listed in Table 1.

DmTNP	**hTHAP9**
D230	
D252	D304
D303	D374
D339	D414
D444	D519
E531	E613
D604	D695
E676	E776

Table 1: List of conserved Asp and Glu residues (which were mutated) in RNase-H fold of DmTNP and hTHAP9. Residues of each homolog, that aligned with each other in Fig.3, were adjacent to each other in the same row

Site directed mutagenesis of conserved acidic residues in the catalytic domain of hTHAP9 and its homologs. To investigate the role of the conserved acidic residues (Table 1 and Fig 3), we mutated all the conserved Aspartate (D) residues to Cysteine (C) and the Glutamate (E) residue to Glutamine (Q). We avoided the conventional alanine(A) substitutions because the size of A is very small in comparison to D and E; replacement with a smaller amino acid may affect protein structure and thus it will be difficult to infer whether potential loss of activity is due to the absence of catalytic activity or structure disruption (27). Moreover, since Metnase is known to function using a DDN catalytic triad, we decided not to substitute D → N since it may not affect the catalytic activity of the transposase (28). Thus, in order to replace D or E with a similar-sized amino acid, we substituted them with a C or Q respectively. The

mutants cloned in pCDNA 3.1(+) include hTHAP9 (D304C, D374C, E776Q, E613Q, D414C, E519C, D695C) and DmTNP (D303C, D230C, D444C, D339C, D252C, E531Q, D604C, D676C).

Many conserved acidic residues in the catalytic domain of hTHAP9 and its homologs are important for DNA excision: To investigate the catalytic activities of the hTHAP9 and DmTNP mutants, we used a plasmid based DNA excision assay(20). In this assay, a reporter plasmid (pISP-2/Km, having 0.6 Kb insertion of P element DNA flanked by intact terminal inverted repeats disrupting the Kanamycin resistance gene) is co-transfected with a plasmid encoding for the transposase (Fig. 5). We ensured that all the mutant THAP9 proteins were expressed in human cells in approximately equal amounts (Fig. S4).

The excision activity of each mutant was measured by scoring transposase–induced transposon excision events in bacteria (29). We further confirmed P element excision, by sequencing (Fig. S6) or performing diagnostic PvuII restriction digestion of plasmid DNA excision products recovered from bacteria (Fig. S5). When the P element DNA (600 bp) is excised by the functional transposase followed by repair of the reporter plasmid, the PvuII restriction pattern of the recovered reporter plasmid shifted by 600bp. (Fig S5, Fig S6).

We observed that in DmTNP; D230C, D303C and E531Q exhibit lowest relative excision activity (~10% of the wildtype DmTNP activity; Fig 5). These residues have been predicted in previous reports to constitute the DDE motif in DmTNP (8, 12). Besides the DDE triad, we also found that D444 is extremely important for DmTNP transposition because the number of kanamycin resistant colonies observed in all replicates was zero (Fig 5, Table S1, Fig. S7). Moreover, D339 from the conserved D(2)H motif, which is a part of the signature string (12) (Fig 4) of DmTNP superfamily, also affects the excision activity when mutated (Fig 5).

In hTHAP9 (excision activity of wild type hTHAP9 was observed to be 70% of that of wt DmTNP), the relative excision activities were observed to be low for the D304C, D374C and E613Q mutants. This result is in agreement with our computational prediction of the catalytic triad. Surprisingly, the excision activity of D695C was enhanced in comparison to the wild type protein. (Table S1, Fig. S7)

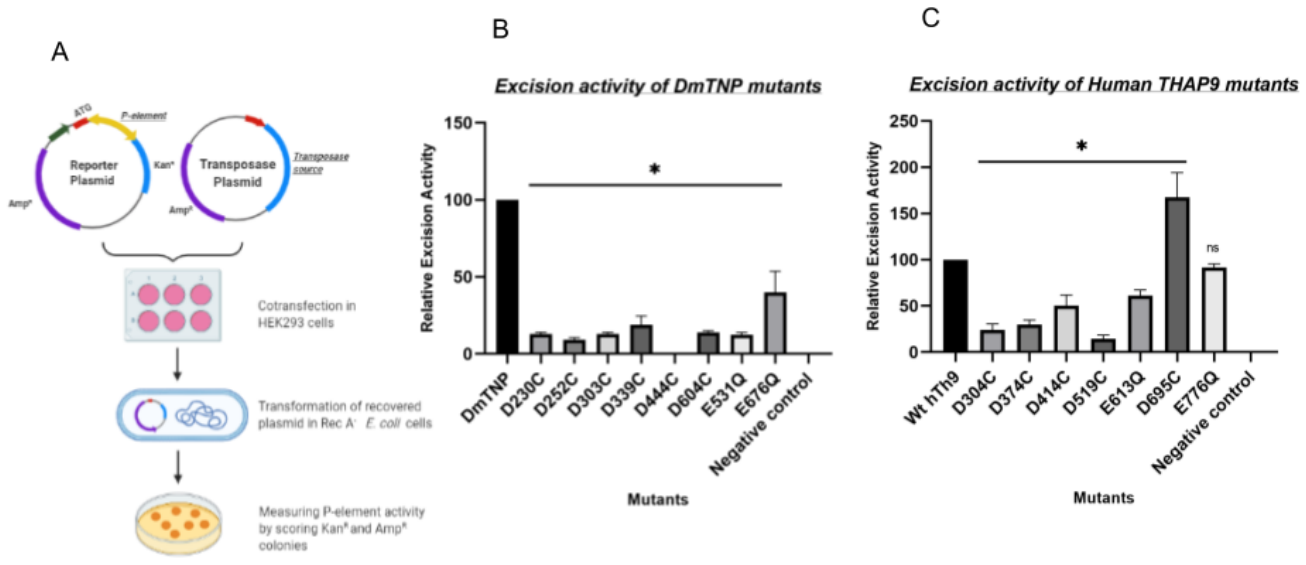

Fig 5: Conserved acidic residues within the signature string of DmTNP and hTHAP9 are important for DNA excision. (A) Schematic of transposon excision assay. Relative excision activities of (B) DmTNP mutants (compared to wild type DmTNP) (C) hTHAP9 mutants (compared to wild type hTHAP9). All the values are represented as a mean of 4 independent experiments (±SEM, n=4) p value <0.05. Negative control (pBluescript empty vector)

Many conserved acidic residues in the catalytic domain of hTHAP9 and its homologs are important for DNA integration: We next investigated whether the mutant transposases could carry out transposition of a marker gene flanked by P element terminal inverted repeats, from a plasmid into the human genome. The relative integration activity was calculated by screening and staining G418 resistant colonies(20).

In DmTNP, mutation of catalytic triad DDE residues, D230, D303 or E531 led to low DNA integration activity (Fig 6). The D444C mutant underwent complete loss of excision as well as integration function, suggesting its role in DNA-protein interaction during transposition (Fig 7). Consistent with the DNA excision results (Fig.5), mutations of D304, D374 and E613 in hTHAP9 resulted in low DNA integration activity. Surprisingly, the D695C mutant, which showed heightened excision assay (Fig.6) demonstrated low integration activity (Fig 6, Table S1, Fig. S7).

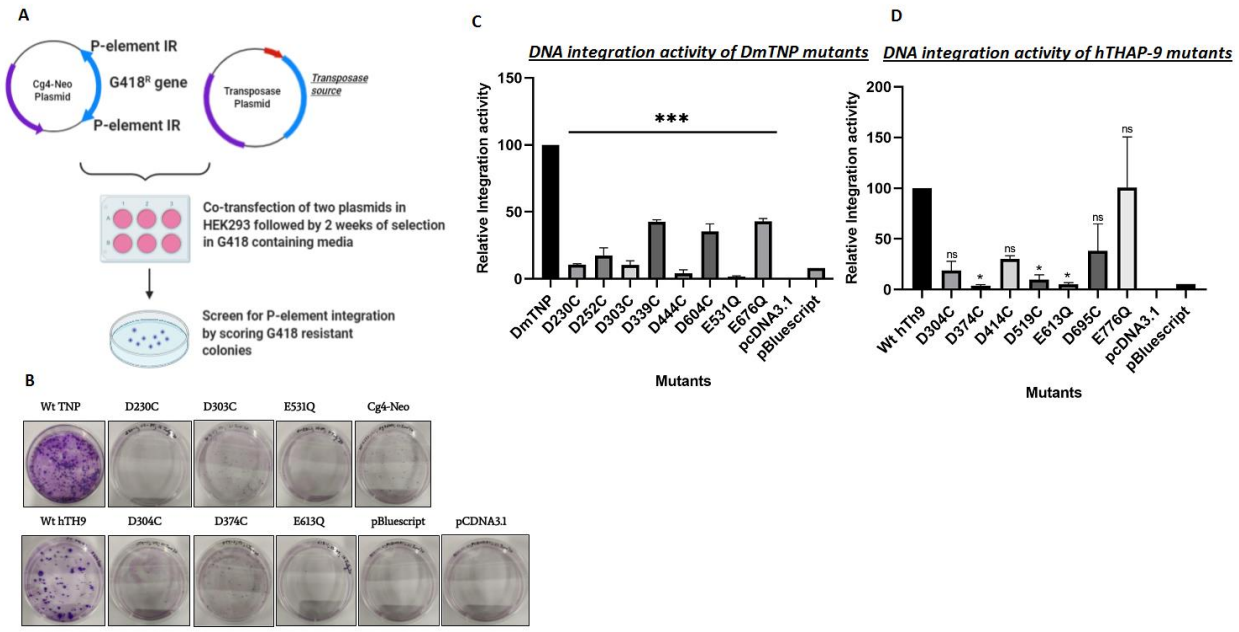

Fig 6: Conserved acidic residues within the signature string of DmTNP and hTHAP9 are important for DNA integration (A) Schematic of Transposon integration assay. (B) Crystal violet stained G418 resistant colonies of wildtype Drosophila and hTHAP9 along with DDE motif mutants and negative controls. Relative integration activities of (C) DmTNP mutants (compared to wild type DmTNP) (D) hTHAP9 mutants (compared to wild type hTHAP9). Empty vectors - pCDNA3.1 (+) and pBluescript were used as negative controls. All the values are represented as mean of two independent experiments performed in duplicates (±SEM) p value <0.05.

Thus, we identify the catalytic domain of hTHAP9 and demonstrate that it has a putative RNase-H like fold (like in DmTNP) with a conserved DDE motif made up of D304, D374 and E613 that constitute the catalytic triad (D230, D303 and E531 in DmTNP) and are important for DNA excision and integration.

Discussion

The architecture of most primitive proteins, as deduced by molecular fossils found in modern proteins, share a handful of folds namely P-loop, TIM β/α-barrel, Rossmann-fold domains, Ferredoxin-like, Flavodoxin-like and Ribonuclease H-like folds (30). The RNase-H fold, which is one of the most ancient and abundant protein folds derived from viruses, has been co-opted by diverse proteins in all kingdoms of life (transposases, integrases, resolvases, Piwi nucleases, CRISPR-associated enzymes like Cas9, RAG1 etc.) (30, 31) that are involved in a variety of cellular processes like DNA replication, recombination and repair, transposition, immune defense, splicing, RNA interference and CRISPR-Cas immunity (31).

The members of all seventeen eukaryotic cut-and-paste transposase superfamilies [include pogo, Mariner, Mutator, Harbinger, Merlin, PIF, Zator, Sola, Ginger, Transib, Sleeping beauty, piggybac, hAT (38–44)] carry the characteristic RNase-H fold (8),suggesting a common

evolutionary origin of the “cut-and-paste” transposition mechanism. (11, 36, 37). The RNase-H fold is also found in viral integrases (HIV-1, RSV (11)) that share strikingly similar structural properties to several transposases, few LTR transposons (Copia, Gypsy) and bacterial IS sequences (eg: IS1, IS3, IS6, IS30) (6, 45, 46), suggesting a common ancestor (11, 33, 34).

Although the function as well as overall architecture of RNase-H fold-containing proteins is very diverse, the architecture of their RNaseH-like catalytic core and the corresponding catalytic mechanism is highly conserved (32, 33). The RNase-H fold brings three catalytic residues (DDD/E) in close proximity which coordinate with two divalent metal ion cofactors (mostly Mg^{2+}) to form the active site for both DNA cleavage and strand transfer during transposition (6, 37). The conserved steps during catalysis include initial DNA cleavage by hydrolysis, formation of an oligomeric synaptic complex that carries out catalysis in trans (i.e. the enzyme subunit will bind one transposon end and cut the opposite end of the transposon DNA), the subsequent strand transfer that involves the nucleophilic attack on target DNA by the free 3’-OH at the transposon end and strong bending of target DNA (6).

RNase-H fold-containing proteins are multi-domain proteins which often carry other domains involved in site-specific DNA binding and multimerization in addition to the RNaseH-like catalytic domain. The DDE transposase family members studied till date exhibit striking differences in their transposition mechanisms (second strand cleavage, synaptic complexes that are formed are quite different; DmTNP, Tn5 and Mos1 are dimeric, Mu and PFV are tetrameric while Hermes transposase is octameric) (12, 21, 29, 34, 35).

Moreover, both the size of the overall domain as well as individual secondary structure elements of the RNaseH-like catalytic domain, varies significantly. This is because the RNase-H fold of many DDE/D transposases contain an insertion domain (6) often inserted between the fifth β-strand and fourth α-helix. The length and function of the insertion domain varies in different DDE transposases, as observed in the limited number of corresponding solved structures (Tn5, Hermes, Rag1, DmTNP) (11, 17, 21). Although the exact functions of these insertion domains have not been determined, it is observed that they often contain amino acids essential for transposition function. For example, in Tn5 transposase, the insertion domain is important for formation of a DNA hairpin at the transposon ends while the insertion domain of Hermes transposase carries a conserved tryptophan residue (Trp319) responsible for target joining. The insertion domain of DmTNP harbors an alpha-helical GTP binding domain essential for binding GTP as a prerequisite for successful DNA transposition (12, 17). It is also interesting to note that despite the dissimilarity in primary sequence and functions, the insertion domains of Hermes, Rag1, Tn5 and DmTNP are topologically similar.

Our study predicts the presence of an RNase-H fold in hTHAP9 (and Pdre2), which appears to be disrupted by an insertion domain between **β**5 and **α**4, similar to the insertion domain of DmTNP which binds GTP via GTP binding motifs (12, 17). Despite the disruption of the RNase-H fold, both DmTNP and hTHAP9 are able to cut the substrate DNA (20).

The DDE/D motif has been shown to be functionally relevant in several transposases by investigating catalytic activity after performing site-directed mutagenesis to substitute the essential DDE triad with other amino acids (6). Previous studies had defined the DDE motif in DmTNP by substituting each of the acidic residues in the triad with alanine and demonstrating that these mutants had significantly lower DNA excision activity (12). However, alanine is much smaller than D/E amino acids and may have disrupted the RNaseH fold. To validate that excision

is indeed decreasing because of loss of essential catalytic amino acids and not fold disruption, we replaced the motif with amino acids of comparable sizes. We observed that D230, D303, D339 (when mutated to Cys) and E531 (when mutated to Gln) fail to excise and integrate the target DNA (Fig 5, Fig 6). All these amino acids are the part of the signature string of DmTNP superfamily (Fig 3): D230, D303 and E531 constitute the catalytic triad while D339 is a part of the [C/D](2)H motif. Moreover, D339 marks the boundary of the insertion domain (residues 339-528) of DmTNP (12). Additionally, D444 is absolutely essential for the transposition function (Fig 5, Fig 6, Table S1, Fig. S7); it is interesting to note that in the recently solved DmTNP structure, D444 mediates binding to GTP that is coordinated to a magnesium ion (12).

In hTHAP9, DNA excision and integration were severely impaired when residues D304, D374 and E613 were mutated. This suggests that these three residues, which are conserved in its homologs (DmTNP, Pdre2) and are located on appropriate secondary structure elements in the predicted RNaseH fold, form the DDE motif in hTHAP9. Also, D414, which is a part of the [C/D](2)H motif of hTHAP9, is also important for DNA excision.

Moreover, it was observed that besides the catalytic triad, there were other conserved acidic residues (D519 in hTHAP9; D252, D604, E676 in DmTNP) in the signature string, which were important for DNA transposition (Table S1, Fig. S7). It is interesting to speculate about the exact roles of other conserved residues and motifs in the catalytic domain of hTHAP9 and its homologs. The recent cryo-EM structure of DmTNP(12) illustrates that the catalytic domain makes extensive contacts with both DNA and GTP during transposition. Thus, future biochemical and structural investigations of hTHAP9, may identify additional conserved residues involved in various steps of the transposition process.

The hTHAP9 D695C mutant displayed elevated transposase activity that was higher (1.4 times) than the wild type hTHAP9 activity (Fig 5, Table S1). However, its DNA integration activity was not enhanced to a similar extent (Fig 6). Hyperactive mutants of *Sleeping beauty* (SB100X) and *Piggy-Bac* (hyPBase) transposons are used as gene delivery vectors for stable expression of transgenes (47, 48). The hTHAP9 D695C mutant may also serve as an attractive genome engineering tool upon further investigations.

Conclusions

This study identifies the catalytic triad and highlights the critical importance of conserved acidic residues in the RNaseH-fold containing catalytic domain of human THAP9. The RNaseH fold of hTHAP9 was predicted using homology modeling and secondary structure predictions. Mutagenesis of conserved acidic residues in the catalytic triad (DDE/D) as well as the canonical signature string of the predicted RNaseH fold significantly impacted catalytic function i.e. DNA excision and subsequent integration during transposition. Moreover, we identified an Asp residue, which when mutated, is able to achieve heightened excision activity; such hyperactive transposases can be exploited as possible genome engineering tools. Thus, this study provides new insights towards elucidating the catalytic mechanism and unknown physiological roles of a relatively unexplored human transposase.

Materials and Methods:

Materials:

HEK293 cell lines obtained from Dr Virupakshi Soppina's lab at IIT Gandhinagar, DMEM (HyClone) supplemented with 10% Fetal Bovine Serum, TurboFect Transfection Reagent (ThermoFisher Scientific; R0531). pISP2/Km, Cg4-Neo, pBluescript vectors were generous gifts from Prof. Donald Rio, UC Berkeley.

G418 (Merck, 345810), GeneJET plasmid miniprep kit (Thermo Fisher Scientific; K0503), Ampicillin (Sigma; 10835242001), Kanamycin (Sigma; K1377), Restriction Enzyme PvuII-HF (NEB; R0151S), CutSmart Buffer (NEB; B7204S), rabbit anti-HA monoclonal antibody (Sigma; SAB1306169), mouse anti-GAPDH monoclonal antibody (Novus Biologicals; NB300-328SS), anti-rabbit (GENA934) and anti-mouse (GENA931) IgG HRP secondary antibodies.

Methods:

Multiple sequence alignment and Secondary structure prediction: hTHAP9 (Uniprot Identifier: Q9H5L6-1), DmTNP (Uniprot Identifier: Q7M3K2-1) and Pdre2 (18) amino acid sequences were aligned using Clustal omega (49) with default parameters and the results were viewed in MView (50). PSIPRED (22) was used with default parameters to predict the secondary structure of the three THAP9 homologs.

Homology Modelling, Docking and MD simulations: Homology modelling was performed using default parameters and target sequence-template option in freely available SWISS MODEL (23) with DmTNP (6p5a) structure as template. The output from SWISS MODEL was used to perform Magnesium ion docking using AUTODOCK Vina (24). The hTHAP9 homology model with magnesium ion docked in it was further used for MD simulations with GROMACS (25). The model was processed using CHARMM 36 force field (26) and further simulated for 100ns. The final structure was viewed and analyzed using PyMol (51). RMSD was calculated using GROMACS and Ramachandran plot was analyzed using SAVES software (52).

Site-directed mutagenesis: Primers for generating mutations in hTHAP9 and DmTNP coding sequences were designed using NEB Base changer tool™ and site directed mutagenesis was performed using the Q5 Site directed Mutagenesis kit (NEB E0554S) as per manufacturer's instructions. All mutant constructs were verified by Sanger sequencing at Scigenom, India.

***In vivo* P element Excision assay:** 0.5 X 10^6 HEK293 cells (grown in DMEM supplemented with 10% fetal bovine serum under standard tissue culture conditions) were plated per well of 6-well plates and transfected at 70-90% confluency using Turbofect transfection reagent. The standard transfection protocol of the manufacturer was followed to transfect 1 ug pISP2/Km reporter (20) plasmid (has 0.6kb P element insertion which disrupts the kanamycin resistance open reading frame) along with either 1 ug DmTNP or hTHAP9 (wild type or mutants separately) or pBluescript empty vector (negative control), in triplicates. After 48h, the cells were harvested in PBS to recover

and analyze plasmid DNA and protein expression. The plasmid DNA from the transfected cells was isolated using GeneJET plasmid miniprep kit following the standard protocol. The DNA was quantified and 1µg of total DNA was transformed into *Escherichia coli* DH5α™ ultracompetent cells (RecA-) and was plated on ampicillin and kanamycin+ampicillin plates. 1/10 of the cells were plated on ampicillin plates while 1/2 of the cells were plated on kanamycin plus ampicillin plates. After 24h, the colonies were counted.

The ratio of the number of kanamycin-ampicillin colonies to that of the ampicillin colonies gave the P element excision activity (20). The relative excision activity was calculated using the following formula:

Excision activity = (Number of Kan^R +Amp^R colonies)/ Total number of Amp^R colonies
Relative Excision activity = (Excision activity of mutant*100) / Excision activity of wild type transposase

Statistical significance of relative excision activity values were calculated using two-tailed non-parametric t-test followed by Mann-Whitney's test.

Plasmid DNA (pISP-2/Km reporter (20)) was isolated from the colonies obtained on the kanamycin plus ampicillin plates. pISP2/Km which is originally kanamycin sensitive, gives rise to kanamycin resistant colonies due to P element excision and subsequent repair. pISP2/Km is a 4483bp plasmid with a 600bp P element between 2240 and 2840 positions. PvuII restriction endonuclease cuts pISP2/Km at positions 2054, 3173, 3935 and 4173 (20).

1µg of isolated repaired reporter DNA was digested with PvuII as per the standard restriction digestion protocol. If P element excision has taken place, PvuII digestion of the repaired pISP2/Km reporter plasmid will give rise to a 0.5kb fragment. On the other hand, if excision has not occurred, a 1.1kb fragment (made up of 0.6kb P element and 0.5kb of kanamycin resistance gene) is observed with 3 other bands corresponding to 2kb, 0.75kb and 0.25kb.

To investigate the nature of donor site repair after P element cleavage, the isolated reporter plasmids were sequenced across the P element excision site using a primer (5'-GTTGTGTGGAATTGTGAGCGG-3') flanking the P element insertion site.

***In vivo* Integration assay**: Cg4-Neo reporter plasmid (20) (that carries an SV40 promoter-G-418R gene inside the Cg4 P element vector) was co-transfected (Turbofect), with the transposase expression plasmid (DmTNP or hTHAP9 wild type or mutants separately) or negative control (pBluescript or pCDNA3.1 (+) empty vector) in HEK293 cells (0.5 X 10^6 cells per well of a 6-well plate) at 70% confluency. A total of 2 ug of plasmid DNA was transfected per well: 50ng of CgNeo4, 1ug of the transposase plasmid/negative control and the amount was made upto 2ug with pBluescript empty vector. After 48 hours, the cells from each well were transferred to separate 10 cm dishes and were allowed to adhere for 24 hours at 37℃ in a CO_2 incubator. After 24 hours, the cells were grown in media supplemented with G418 (0.5 mg/ml) and selected for 2 weeks, after which the G-418-resistant colonies were fixed with methanol, stained with crystal violet and counted. The relative integration activity was computed as a percentage of activity of each mutant to the wild type DmTNP (20). Significant values were determined using a one-way ANOVA method followed by Dunn's test.

Protein Immunoblotting: The cell extracts from transfected cells were lysed and run on 10% denaturing SDS PAGE in running buffer (1X Tris- Glycine SDS). PageRuler™ Prestained Protein Ladder (10 to 180 kDa, ThermoFisher Scientific; 26616) was used as a molecular size marker and the samples were transferred to PVDF membrane by electroblotting using a standard gel transfer system. The membrane was blocked to remove nonspecific binding using 3% (w/v) skimmed milk in Tris-Buffered Saline with Tween-20 (TBST). Membranes were washed with 1X TBST and incubated overnight at 4°C with 1:2000 dilution of primary anti-HA antibody followed by incubation with 1:5000 dilution of HRP- coupled secondary antibodies for 2h at room temperature and detected using enhanced chemiluminescence (Pierce, PI32106) on a Bio- Rad Gel Documentation system.

Acknowledgements

We thank Althaf Shaik, Arthi Hariharan, Chinmaya Panda and Deepshikha Ghosh for their suggestions about homology modelling and MD simulations. We also thank Dr Virupakshi Soppina for sharing HEK293 cell lines for the experiments. We acknowledge funding from DBT - Ramalingaswami fellowship, SERB and IIT Gandhinagar.

References:

1. Craig, *et al.*, Eds., *Mobile DNA III* (American Society of Microbiology, 2015) https:/doi.org/10.1128/9781555819217 (October 8, 2020).
2. L. Sinzelle, Z. Izsvák, Z. Ivics, Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell. Mol. Life Sci. CMLS* **66**, 1073–1093 (2009).
3. C. Feschotte, Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).
4. D. Jangam, C. Feschotte, E. Betrán, Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet. TIG* **33**, 817–831 (2017).
5. C. Feschotte, E. J. Pritham, DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu. Rev. Genet.* **41**, 331–368 (2007).
6. A. B. Hickman, M. Chandler, F. Dyda, Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. *Crit. Rev. Biochem. Mol. Biol.* **45**, 50–69 (2010).
7. E. S. Lander, *et al.*, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
8. Y.-W. Yuan, S. R. Wessler, The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7884–7889 (2011).
9. M. Nowotny, S. A. Gaidamakov, R. J. Crouch, W. Yang, Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell* **121**, 1005–1016 (2005).
10. E. Asante-Appiah, A. M. Skalka, HIV-1 integrase: structural organization, conformational changes, and catalysis. *Adv. Virus Res.* **52**, 351–369 (1999).
11. P. A. Rice, T. A. Baker, Comparative architecture of transposase and integrase complexes. *Nat. Struct. Biol.* **8**, 302–307 (2001).
12. G. E. Ghanim, E. H. Kellogg, E. Nogales, D. C. Rio, Structure of a P element transposase–DNA complex reveals unusual DNA structures and GTP-DNA contacts. *Nat. Struct. Mol. Biol.* **26**, 1013–1022 (2019).
13. R. Kofler, T. Hill, V. Nolte, A. J. Betancourt, C. Schlötterer, The recent invasion of natural

Drosophila simulans populations by the P-element. *Proc. Natl. Acad. Sci.* **112**, 6659–6663 (2015).
14. , Evolution of hybrid dysgenesis determinants in Drosophila melanogaster (October 9, 2020).
15. M. Muñoz-López, J. L. García-Pérez, DNA Transposons: Nature and Applications in Genomics. *Curr. Genomics* **11**, 115–128 (2010).
16. S. Majumdar, D. C. Rio, P Transposable Elements in Drosophila and other Eukaryotic Organisms. *Microbiol. Spectr.* **3**, MDNA3-0004–2014 (2015).
17. P. A. Rice, Structure of the P element transpososome reveals new twists on the DD(E/D) theme. *Nat. Struct. Mol. Biol.* **26**, 989–990 (2019).
18. S. E. Hammer, S. Strehl, S. Hagemann, 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**, 833–844 (2005).
19. S. Hagemann, S. E. Hammer, The implications of DNA transposons in the evolution of P elements in zebrafish (Danio rerio). *Genomics* **88**, 572–579 (2006).
20. S. Majumdar, A. Singh, D. C. Rio, The Human THAP9 Gene Encodes an Active P-Element DNA Transposase. *Science* **339**, 446–448 (2013).
21. A. B. Hickman, F. Dyda, DNA Transposition at Work. *Chem. Rev.* **116**, 12758–12784 (2016).
22. D. W. A. Buchan, D. T. Jones, The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).
23. A. Waterhouse, *et al.*, SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
24. O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
25. Lindahl, Abraham, Hess, V. D. Spoel, GROMACS 2020.4 Manual (2020) https:/doi.org/10.5281/ZENODO.4054996 (October 9, 2020).
26. J. Huang, *et al.*, CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
27. J. D. Wright, C. Lim, A fast method for predicting amino acid mutations that lead to unfolding. *Protein Eng. Des. Sel.* **14**, 479–486 (2001).
28. H.-S. Kim, *et al.*, The DDN catalytic motif is required for Metnase functions in non-homologous end joining (NHEJ) repair and replication restart. *J. Biol. Chem.* **289**, 10930–10938 (2014).
29. E. L. Beall, D. C. Rio, Drosophila P-element transposase is a novel site-specific endonuclease. *Genes Dev.* **11**, 2137–2151 (1997).
30. B.-G. Ma, *et al.*, Characters of very ancient proteins. *Biochem. Biophys. Res. Commun.* **366**, 607–611 (2008).
31. K. Moelling, F. Broecker, G. Russo, S. Sunagawa, RNase H As Gene Modifier, Driver of Evolution and Antiviral Defense. *Front. Microbiol.* **8** (2017).
32. K. A. Majorek, *et al.*, The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res.* **42**, 4160–4179 (2014).
33. P. Rice, R. Craigie, D. R. Davies, Retroviral integrases and their cousins. *Curr. Opin. Struct. Biol.* **6**, 76–83 (1996).
34. S. P. Montaño, Y. Z. Pigli, P. A. Rice, The μ transpososome structure sheds light on DDE recombinase evolution. *Nature* **491**, 413–417 (2012).
35. R. J. Sarnovsky, E. W. May, N. L. Craig, The Tn7 transposase is a heteromeric complex in which DNA breakage and joining activities are distributed between different gene products. *EMBO J.* **15**, 6348–6361 (1996).

36. W. Yang, T. A. Steitz, Recombining the structures of HIV integrase, RuvC and RNase H. *Struct. Lond. Engl. 1993* **3**, 131–134 (1995).
37. N. Iv, H. Pb, DDE transposases: Structural similarity and diversity. *Adv. Drug Deliv. Rev.* **62**, 1187–1195 (2010).
38. L. Zhou, *et al.*, Transposition of hAT elements links transposable elements and V(D)J recombination. *Nature* **432**, 995–1001 (2004).
39. A. B. Hickman, *et al.*, Molecular architecture of a eukaryotic DNA transposase. *Nat. Struct. Mol. Biol.* **12**, 715–721 (2005).
40. H. M. Robertson, The Tcl-mariner superfamily of transposons in animals. *J. Insect Physiol.* **41**, 99–105 (1995).
41. X. Zhang, N. Jiang, C. Feschotte, S. R. Wessler, PIF- and Pong-Like Transposable Elements: Distribution, Evolution and Relationship With Tourist-Like Miniature Inverted-Repeat Transposable Elements. *Genetics* **166**, 971–986 (2004).
42. C. Feschotte, Merlin, a New Superfamily of DNA Transposons Identified in Diverse Animal Genomes and Related to Bacterial IS1016 Insertion Sequences. *Mol. Biol. Evol.* **21**, 1769–1780 (2004).
43. V. V. Kapitonov, J. Jurka, Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet. TIG* **23**, 521–529 (2007).
44. V. V. Kapitonov, J. Jurka, RAG1 Core and V(D)J Recombination Signal Sequences Were Derived from Transib Transposons. *PLOS Biol.* **3**, e181 (2005).
45. A. Muszewska, M. Hoffman-Sommer, M. Grynberg, LTR Retrotransposons in Fungi. *PLoS ONE* **6** (2011).
46. C. P, L. T, H. D, M. P, B. C, Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? *Genetica* **100**, 63–72 (1997).
47. L. Mátés, *et al.*, Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.* **41**, 753–761 (2009).
48. K. Yusa, L. Zhou, M. A. Li, A. Bradley, N. L. Craig, A hyperactive piggyBac transposase for mammalian applications. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1531–1536 (2011).
49. M. F, *et al.*, The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).
50. N. P. Brown, C. Leroy, C. Sander, MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics* **14**, 380–381 (1998).
51. *PyMOL The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.*
52. https://servicesn.mbi.ucla.edu/SAVES/.