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

DNA transposons, also known as 'cut and paste' transposons, are DNA sequences that can be excised from one position in the genome by an autonomously encoded transposase, and randomly or specifically integrate into a different position using host cell DNA repair machinery. The catalytic domain of most 'cut and paste' transposases have the canonical RNase-H fold and are similar to the nuclease domain of retroviral integrases. The RNase-H fold is a mixture of beta sheets and alpha helices with three acidic residues (Asp, Asp, Glu/Asp - DDE/D) in its active site that are coordinated with Mg++/ Zn++ion and are involved in DNA cleavage. Human THAP9 is a 'cut and paste' DNA transposase homologous to the well studied Drosophila P-element transposase. It is present as a single copy in the human genome and exhibits an approximately 40% sequence-based similarity with P-element transposase. The lack of transposon hallmarks, i.e., TIRs and TSDs in THAP-9 suggests that it may have undergone molecular domestication. But its function in humans is yet to be discovered. Interestingly, THAP9 can cut P-element DNA flanked with inverted repeats in both human and Drosophila cells. In this study we have modelled the structure of human THAP9 using the recently available cryo-EM structure of P-element transposase as a template to identify the RNase-H like fold in its catalytic domain. Secondary structure predictions and protein modeling were used to predict a putative RNase H fold, along with acidic residues within the fold that might be involved in DNA excision by THAP9. Site-directed mutagenesis of the predicted catalytic residues followed by screening for DNA excision and integration activity, has led to the identification of candidate Ds and Es in the RNaseH fold that could be a part of the catalytic triad in human THAP9.

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 have been 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 transposon to make evolutionarily beneficial, functional host genes [3,4].

Almost all domesticated genes have derived sequences from an ancestral DNA transposon [5]. 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 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, THAP9 are examples of human transposon-derived genes which have domains recruited from DNA transposons [3]. Recombination activating proteins (RAG1/2) mediate the DNA rearrangement reactions in V(D)J recombination; Centromeric protein B (CENPB) helps in chromatin assembly at centromeres during cell division; SETMAR is a DNA repair protein and the function of THAP9 is still unknown.

Out of 19 DNA transposon superfamilies identified till date, 11 are known to 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]. RNase-H like fold is common among transposases and retroviral integrases (Tn3, Tn5, Mu, hAT, Mariner, Sleeping beauty, HIV-1 integrase, RSV integrase), where the DDE/D motif catalyzes magnesium ion-dependent nucleophilic attack on the DNA strands [9–11]. The DDE transposons family studied till date exhibit striking differences in their transposition mechanisms. However, the initial step of nucleophilic attack by a water molecule close to the transposon end remains the same [12].

Wessler et al in 2011 had predicted the DDE motif in P-element superfamily which has been confirmed in the recently solved cryo-EM structure of P-element transposase(DmTNP) in Drosophila [8,13]. P-element is a widely studied transposon in *Drosophila melanogaster that* invaded this species of fly in the late 1940s and is continuing to spread through the natural fly populations by horizontal gene transfer[14]. It was discovered that P-element is responsible for hybrid dysgenesis (male sterility) in Drosophila [15]. Ever Since its discovery, P-element has been used as a model to study transposition and also has been used as a genetic engineering tool [16,17]. The cryo-EM structure has revealed distinct properties of P-element transposase and transposon interactions including GTP binding and staggering cuts on the DNA. Most importantly, it shows that the GTP is not required for providing energy for the reaction, rather it has a stoichiometric role where it assists in arranging DNA in proper conformation for catalysis by interacting with the terminal base of the cut transposon. Another unique feature highlighted in the P-element transpososome structure is the looping out of the cut strand to maintain the DNA integration [13,18].

The human THAP9 protein is homologous (25% identical and 40% similar) to the Drosophila P element transposase beyond its DNA binding domain. It is immobile in humans and exists as a single copy gene, whereas its homolog in zebrafish (Pdre2) appears to be functional with intact TIRs and STIRs [19]. Besides carrying an intact full-length Pdre2 gene, zebrafish genome harbors multiple truncated copies of Pdre2, similar to the P-element gene in *Drosophila*, signalling towards the mobility of Pdre2 [17,19,20]. Although the function of THAP9 remains undiscovered, it is known to be catalytically active and can cut DNA flanked by P element's TIRs in humans as well as Drosophila cells [21]. THAP9 has retained its ability to cut DNA which makes it important to investigate the nature of its catalytic domain. We have observed a RNase-H like fold in THAP9 with D304, D374 and E613 as the catalytic residues responsible for excising the substrate DNA.

Results

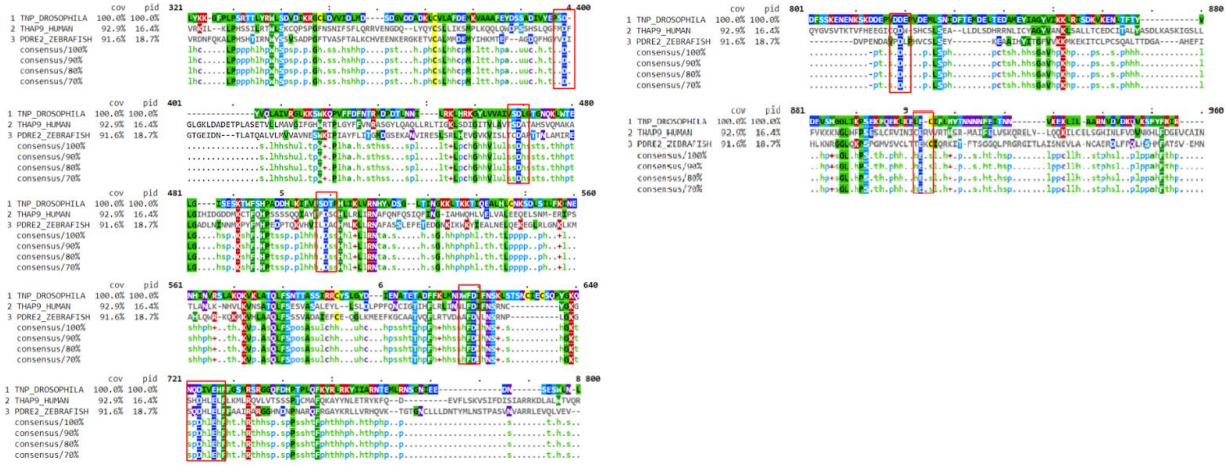

Fig1: Multiple sequence alignment of amino acid sequences of human THAP9, Drosophila THAP9 (P-element) and zebrafish THAP9 (Pdre2) aligned using Clustal Omega and visualization using MView. The conserved Ds and Es are highlighted in red boxes. Different groups of amino acid consensus are highlighted in various colours.

We used multiple sequence alignment to examine the conserved acidic residues in Drosophila P-element transposase (DmTNP), Human THAP9 (hTHAP9) and Zebrafish THAP9 (Pdre2). The three homologs were aligned using Clustal Omega and the alignment results were analysed using MView. Multiple conserved residues were identified throughout the sequences. Besides several Ds and Es, there are many other residues that are conserved among the three homologs. Many of these conserved residues form a 'signature string' of P element superfamily which was first identified by Wessler et al in 2011.

Signature string: Besides DDE motif, the additional conserved residues in the transposase sequence have been reported to form a signature string specific to each 'cut and paste' transposon superfamily [8]. The residues in a signature string have similar approximate spacing in all the members of the superfamily. For the P-element superfamily, the signature string was first described

by Wessler et al in 2011. Using this signature string, we have identified a similar pattern of conserved amino acids in human THAP9 and Pdre2 (Fig 2).

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 2: Signature string of P-element superfamily (Wessler 2011) alongwith the 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 amino acids in DmTNP that are involved in bending of the DNA while forming a complex were also observed to be conserved in hTHAP9 [13]. Most of the amino acids involved in DNA bending are not identical in hTHAP9 but the nature of the substituted amino acids are similar to the DmTNP. R194, W195, S242, S243, K398, K400, Y519, D528, R563, K636, K637, L662, F722 are few residues in DmTNP DNA bending that are conserved identically in hTHAP9 and Pdre2. The conserved Ds and Es in the three homologs are highlighted in red boxes and enlisted in Table 1.

Dm-TNP	**THAP9**
D230	D304
D252	D374
D303	D414
D339	D519
D444	D695
D604	E613
E531	E776
E676	-

Table 1: List of conserved Asp and Glu residues mutated in THAP9 and P-element transposase

Secondary structure and RNase-H domain

The catalytic domain of P element transposase resembles many other transposases and retroviral integrases [12] and belongs to DDE/D class of transposases. The DDE transposases always have RNase-H fold in their catalytic domain with first D in **β**1, second D in or just after **β**4 and last third E/D in **α**4 [8,12].

Using PSIPRED we predicted the secondary structure of P-element transposase, THAP9 and Pdre2. Presence of RNase-H fold was inspected visually and Ds and E were found in beta

sheets (**β**1 and **β**4) and alpha helices(**α**4), in agreement to the characteristic structure of RNase-H fold in integrases and other DDE/D transposases. In DmTNP, the reported residues of the triad, D230 is present in **β**1, D303 in **β**4 and E531 in **α**4. For THAP9 and Pdre2, we have identified the catalytic triad by comparing the predicted secondary structure to the well studied RNase H fold. We have identified D304 in **β**1, D374 in **β**4 and E613 in **α**4 in THAP9 transposases. Similarly, we found D367 in **β**1, D374 just after **β**4 and E679 in Pdre2 (Fig.3). The RNase-H fold in DmTNP, THAP9 and Pdre2 appears 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 [13] but the role of insertion domains in THAP9 and Pdre2 are unknown.

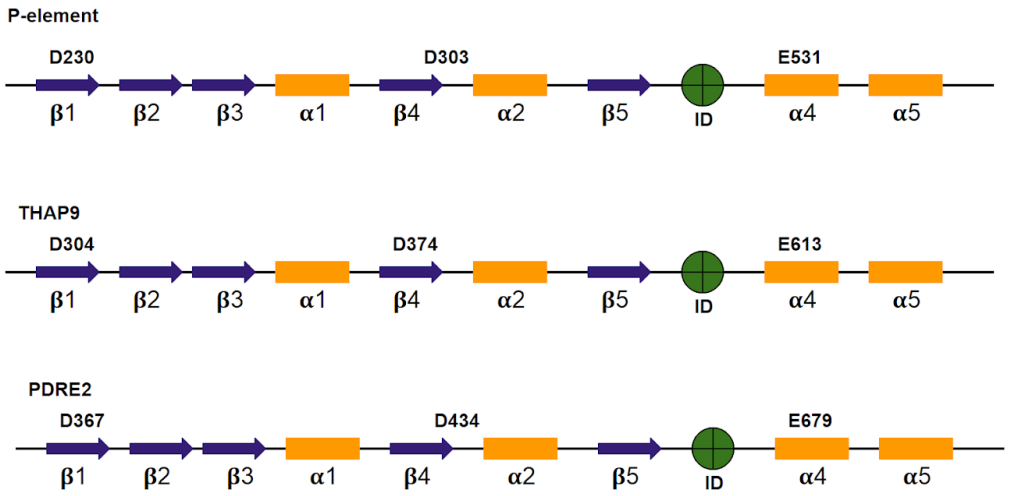

Fig 3: Predicted RNase-H fold in Drosophila, human and zebrafish THAP9. DDE motifs for each homolog are represented in bold above its putative location. Blue arrows indicate beta sheets and yellow bars represent alpha helices. Insertion domains for each homolog are shown in green circles.

THAP9 protein model using DmTNP as the template

Before proceeding to the functional characterization of the candidate residues, we tried to make a 3D model of THAP9 using DmTNP structure (6p5a) as a template for homology modelling. SWISS-MODEL interactive GUI was used for generating THAP9 model with default parameters and the output structure was docked with magnesium ion using AUTODOCK. We then performed energy minimization for the docked structure using MD simulations for 100ns with GROMACS and CHARMM 36 force field.

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).

D374 and E613 are present in their corresponding beta sheet and alpha helix within the RNase H fold (Fig 4). The two residues are also proximal to the magnesium ion, forming an active site. Although D304 appeared just after a beta sheet but it was found to be away from the other catalytic

residues. This might be the result of structure based homology predicted by SWISS-MODEL before producing the final 3D model of the protein, where D304 lies in the loop region of the predicted structure (Fig S2).

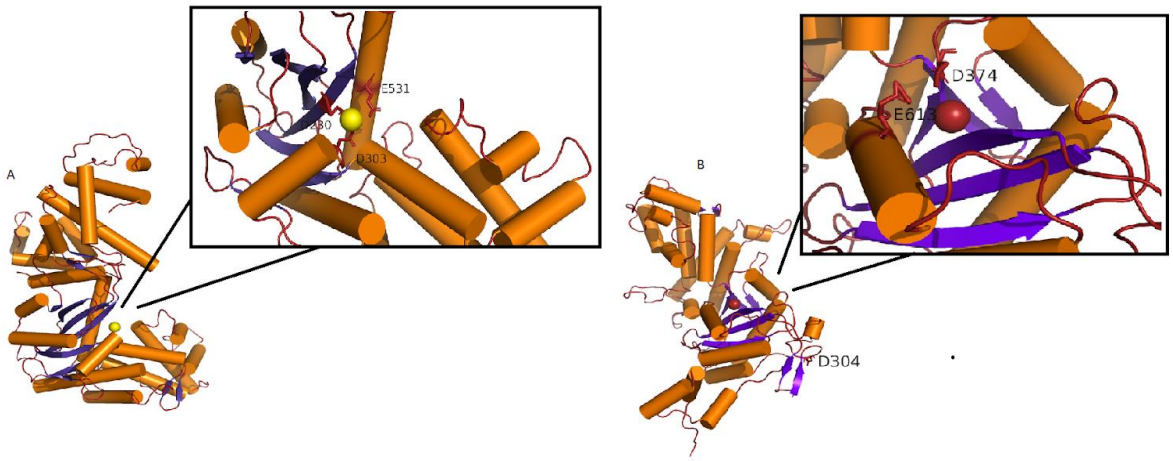

Fig 4: (A) Protein structure of Drosophila THAP9 with beta sheets represented as blue arrows and helices as orange rods ([13], PDB ID - 6P5A). The active site is shown in the inset with DDE motif (D230, D303, E531) in close proximity to the Magnesium ion (shown in light yellow). (B) Homology model of human THAP9 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.

Relative Excision activity

We used site-directed mutagenesis to convert Aspartic acid (D) to Cysteine (C) and Glutamic acid (E) 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 will eventually modulate the protein structure and it will become difficult to infer whether the loss of activity is due to the absence of catalytic activity or structure disruption. We also took into account the experimental exchangeability of the source and destination amino acids, which was higher for D → C and E → Q mutations than D → N and E → C respectively [22].

To investigate the catalytic nature of each mutant, we used a plasmid based DNA excision assay [21]. In this aasay, a reporter plasmid (pISP-2/Km) with DmTNP insertion flanked by terminal inverted repeats disrupting the Kanamycin resistance gene is co-transfected with a plasmid encoding for mutant THAP9 (Fig 5). We ensured that all the mutant THAP9 plasmids are expressed in the mammalian cells in approximately equal amounts (Fig. S3).

The excision activity of each mutant is measured indirectly by scoring the transposition events in the form of proceeding DNA repair reaction that leads to restoration of Kanamycin gene in the reporter plasmid. The reporter plasmid is isolated from mammalian cells post 48 hours of transfection and grown in bacteria plated on kanamycin and ampicillin resistant media[23]. The mutants are expected to exhibit a reduced excision capability in comparison to the wild type THAP9. We further ensured excision of P-element by each mutant, using PvuII restriction digestion sites in the reporter plasmid (Fig S4). The restriction digestion pattern of the reporter plasmid changed after successful excision of P-element in the plasmid (Fig S5).

We observed that D230C, D303C, E531Q exhibit lowest relative excision activity in DmTNP(Fig 5). These residues have been predicted in previous reports to constitute the DDE motif in P-element [8,13]. Besides the DDE triad we also found that D444 is extremely important for the DmTNP transposition because the number of kanamycin resistant colonies observed in all replicates was zero (Fig 5). We also found that D339 which is a part of the signature string of P-element superfamily also affects the excision activity when mutated (Fig 2).

In THAP9, the relative excision activities were observed to be low in D304, D374 and E613 mutants. This result is in agreement with our computational prediction of the catalytic triad. D695 emerged as a surprising mutation where the excision activity of THAP9 is enhanced in comparison to the wild type protein.

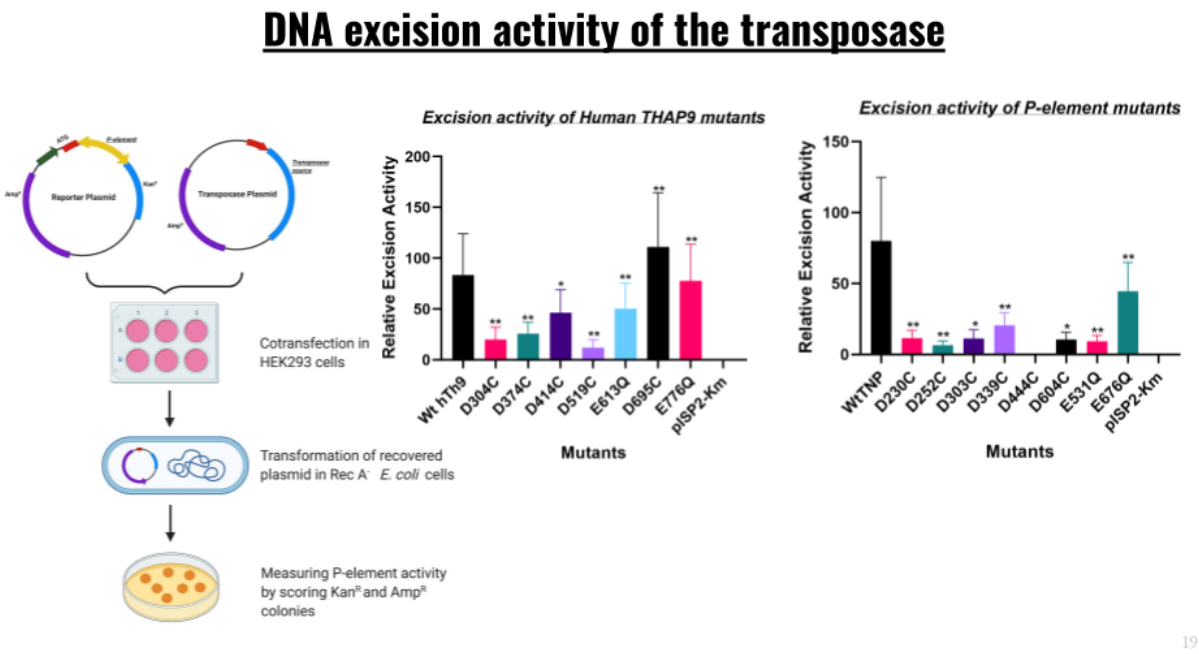

Fig 5: (A) Schematic of transposon excision assay. (B) Relative excision activities of Drosophila THAP9 mutants as compared to the wild type. (C.) Relative excision activities of THAP9 mutants in comparison to wild type THAP9. All the values are represented as mean of 3 independent experiments performed in triplicates (SEM) p value <0.05. No colonies were observed in negative control, i.e., pISP2-Km and D444C mutant in P-element.

Relative integration activity

To study whether the mutant transposases can mobilize the excised DNA into the mammalian genome, we used another plasmid based assay. Here we co-transfected the HEK293 cells with the mutant transposase expression plasmid and a Cg4-Neo plasmid carrying a G418 resistance gene flanked by P-element's terminal inverted repeats. Upon effective transposition, the G418 gene integrates in the mammalian genome. The relative integration activity is calculated by screening and staining G418 resistant HEK293 colonies [21].

In DmTNP, the substituted mutations in D230, D303, E531 led to fewer transposition events thereby leading to less number of G418 resistant HEK293 colonies and low DNA integration activity (Fig 7). D444 mutation caused a complete loss of excision as well as integration function in

DmTNP suggesting an important role in DNA-protein interaction during transposition (Fig 7). Consistent with the DNA excision results, D304, D374 and E613 mutations in THAP9 resulted in low DNA integration activity and fewer HEK293 colonies. D695 mutation which showed hyperactivity in excision assay showed low integration activity (Fig 7).

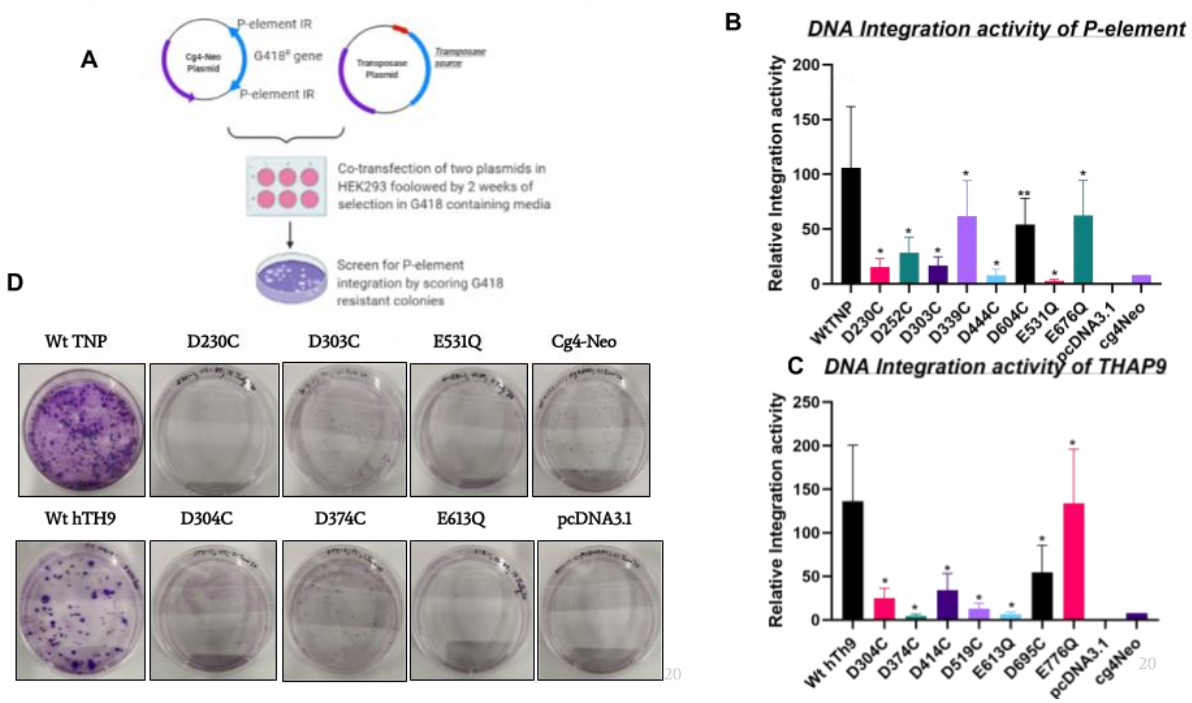

Fig 6: (A) Schematic of Transposon integration assay. (B) Relative integration activities of Drosophila THAP9 mutants and wild type clone. (C.) Relative excision activity of human THAP9 mutants and wild type transposase. All the values are represented as mean of two independent experiments performed in duplicates (SEM) p value <0.05. (D) Crystal violet stained G418 resistant colonies of Drosophila and Human THAP9 along with the DDE motif mutants and only vector transfections.

Discussion

The RNase-H fold is known to be one of the most ancient protein folds from viruses that has incorporated in versatile enzymatic systems in all kingdoms of life [24,25]. It is found majorly in DNA cutting enzymes like transposases, resolvases, repair enzymes, immune defense enzymes, Piwi nucleases, splicing enzymes and CRISPR enzymes [26]. In viruses, RNase-H fold is found in integrases that share strikingly similar structural and mechanistic properties to several transposases, suggesting a common ancestor [11,27,28]. The fold seems to be co-opted in different proteins required to break and bend the target DNA with minor differences in the mode of action, number of oligomers required to cut being the most common difference [28].

DDE/D transposases are also a branch of this huge RNase-H superfamily [11,29,30]. Several 'Cut & Paste' transposons are classified as DDE/D transposons based on their nuclease/catalytic domain. The RNase-H fold of these transposases carry an insertion domain unlike the conventional RNase-H fold to account for the variations related to the transposition mechanisms in

different classes of the transposases [12]. Out of 19 'Cut & Paste' transposon families, 11 carry the RNase-H fold in their carboxy-terminal regions [8]. The RNase-H fold brings the three catalytic residues (DDD/E) in close proximity which further coordinate with a metal ion cofactor to form the active site for the DNA breakage reaction in the process of transposition [12,30]. DDE/D motif has been shown to be functionally relevant in several transposases by using site-directed mutagenesis and replacing the essential DDE triad with other amino acids. A wide range of transposases including, *pogo, Mariner, Mutator, Harbinger, Merlin, PIF, Zator, Sola, Ginger, Transib, Sleeping beauty, piggybac, bacterial IS sequences, hAT* and LTR retrotransposons have been reported to contain the DDE motif for effective catalysis of DNA strand breakage [31–37].

The data in this study suggests the presence of RNase-H fold in THAP9. However, the fold seems to be disrupted due to the insertion domain between **β**5 and **α**4, similar to the insertion domain of P-element transposase which has been proven to carry GTP binding motifs important for the transposition reaction [8,13]. Despite the disruption of the RNase-H fold, both DmTNP and THAP9 are able to cut the substrate DNA [21].

For identifying the RNase-H fold in THAP9, we used secondary structure prediction to find the characteristic mix of beta sheets and alpha helices (β1-β2-β3-α1-β4-α2/3-β5-α4-α5). We further used multiple sequence alignment of the THAP9 homologs to check the conserved Ds and Es present in the identified fold (Fig1, Fig2).To further find if these residues form a catalytic pocket, we modelled THAP9 structure using P-element transposase template (6P5A; [13]. Due to low sequence similarity with the template, many predicted residues did not form a proper conformation and ended up in loops. D374 and E613 were found to be in expected secondary structures and close to the magnesium ion (Fig 4).

D304, D374 and E613 show the least DNA excision and integration when mutated suggesting that the three residues form DDE motif in human THAP9. Besides the low excision activity exhibited by the additional Ds in the signature string (Fig3), we also observed lower excision activity of D519C which implies that there are additional residues important for the transposition. D695C mutation displayed elevated transposase activity, more than the wild type transposase (Fig 5). However, it failed to integrate the DNA to a similar extent, as displayed in DNA excision (Fig 6). It might serve as an attractive gene engineering tool if probed and optimized further.

Previous studies had reported the DDE motif in P-element transposase by alanine substitutions but the size of alanine is very small in comparison to D/E amino acid [13]. To validate that the excision is indeed happening because of loss of the essential amino acids and not fold disruption, we replaced the motif with amino acids of comparable sizes. We observed that D230, D303, D339 and E531 fail to excise and integrate the target DNA when mutated. All these amino acids are the part of the signature string of P-element superfamily (Fig 3). Additionally, the plasmid based excision and integration assays revealed that D444 is absolutely essential for the transposition function P-element. We did not observe any bacterial or HEK293 colonies in the assays when D444 was mutated (Fig 5, Fig 6). This suggests an important role of D444 amino acid in the formation of the transpososome while cutting the DNA.

Beside the DDE motif, most of the identified DDE transposases have a characteristic 'signature' string conserved for each superfamily of transposons [8]. We have observed P-element superfamily's signature string in THAP-9 and Pdre-2 using sequence alignment (Fig 2), corroborating the fact that THAP-9 and Pdre-2 belong to the same family. The location of the insertion domain in both THAP-9 and Pdre-2 is similar to what have been observed in other transposases and transposase-like proteins, i.e., between β5 and α4 [13,18]. The length of insertion domain varies in different families of DDE transposases whose structure has been identified till date (Tn5, Hermes, Rag1, P-element) [11,18]. Although the exact function of the insertion domain has not been determined yet but it is known to carry essential amino acids for transposition function. In Tn5 transposase the insertion domain is important for formation of 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 P-element harbors a GTP binding domain essential to drive transposition reaction, however it does not have the characteristic G-fold of GTP binding proteins [13,18]. Also, the structure of the insertion domain of Hermes, Rag1, Tn5 and P-element are topologically similar which is quite intriguing [18].

Conclusions

The work suggests the critical importance of the acidic residues in the RNase H fold of the catalytic domain of human THAP9. Identification of these residues obtained through the computational data provided further understanding of the insertion domain and its function in excision activity. Our observation clearly shows that the mutations in these sequences, particularly those identified as the signature strings transversely impacts the excision/ catalytic function. This in conjunction with the *in vitro* data, highlights the significance of the DDE motif with a clear classification of key residues necessary for transposition. Also, these amino acids are crucial for the integration activity as the mutants depicted a higher loss of function in this case. It also led to the compelling discovery of a novel residue (695) which when mutated is able to achieve heightened excision activity and can be exploited further for downstream applications. Therefore, the work provides new fundamental insights in an attempt of decoding the function of the human THAP9 gene.

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; no. R0531).

Plasmids used: pISP2/Km, Cg4-Neo, Human mutants cloned in pCDNA 3.1(+)- WT_hTHAP9 (positive control), D304C, D374C, E776Q, E613Q, D414C, E519C, D695C, Drosophila mutants cloned in pCDNA 3.1(+)- WT_DmTNP(positive control), D303C, D230C, D444C, D339C, D252C, E531Q, D604C, D676C, pBluescript empty vector.

G418 (Merck; no. 345810), GeneJET plasmid miniprep kit (ThemoFisher Scientific; no. K0503), NEB Q5 Site directed Mutagenesis kit (NEB E0554S), *Escherichia coli* DH5α™ ultracompetent cells, Ampicillin (Sigma; no. 10835242001), Kanamycin (Sigma; no. K1377), Luria-Bertani (LB) broth, Luria-Bertani (LB) agar, Restriction Enzyme PvuII-HF (New England Biolabs; no. R0151S), CutSmart Buffer (New England Biolabs; no. B7204S), PVDF western blotting membrane (Sigma), PageRuler™ Prestained Protein Ladder, 10 to 180 kDa (ThermoFisher Scientific; no. 26616), Rabbit anti- HA Tag monoclonal antibody (Sigma; no. SAB1306169), Mouse anti- GAPDH Tag monoclonal antibody (Novus Biologicals; no. NB300-328SS), Rabbit IgG HRP secondary antibody (GENA934), Mouse IgG HRP secondary antibody (GENA931).

Methods

Multiple sequence alignment and Secondary structure prediction

THAP9, P-element and Pdre2 amino acid sequences were aligned using clustal omega with default parameters and the results were viewed in MView. PSIPRED 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 target sequence-template option in freely available SWISS MODEL with P-element transposase (6p5a) structure as template. Default parameters were used for homology modelling. The output from SWISS MODEL was used to perform Magnesium ion docking using AUTODOCK Vina. The THAP9 homology model with magnesium ion docked in it was further used for MD simulations with GROMACS. The model was processed using CHARMM 36 force field and further simulated for 100ns. The final structure was viewed and analyzed using PyMol. RMSD was calculated using GROMACS and Ramachandran plot was analyzed using SAVES software.

Site -directed mutagenesis:

Primers for generating mutation in THAP-9 and P-element coding sequence were designed using NEB Base changer tool™ and site directed mutagenesis was performed as per manufacturer's instructions.

***In vivo* Excision assay:**

HEK293 cells were grown in DMEM and 0.5 X 106 cells were plated per well in standard 6-well plates and were incubated at 37℃ for 24 h in a CO2 incubator. Transfection of the adherent cells was carried out at 70-90% confluency using Turbofect transfection reagent, pISP2/Km Reporter plasmid along with Drosophila P element transposase and hTHAP plasmid DNA mutants separately in triplicates. The standard transfection protocol of the manufacturer was followed. After a period of 48h, the cells were harvested to recover and analyze plasmid DNA, genomic DNA and protein expression. The plasmid DNA from the transfected cells was isolated through 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 and was plated on ampicillin

and kanamycin plus ampicillin plates. 1/10 of the cells were plated on ampicillin plates while 1/5 of the cells were plated on kanamycin plus ampicillin plates. After 24h, the colonies were screened, counted and computed. The ratio of the number of kanamycin-ampicillin colonies to that of the ampicillin colonies gave the excision activity and further the relative excision activity was identified through computational calculations.

The colonies from the kanamycin plus ampicillin plates were selected and the plasmid DNA was isolated. The presence of these colonies is due to the pISP2/Km Reporter plasmid which is originally kanamycin sensitive, but the excision of P element and resultant repair disrupts its ORF leading to kanamycin resistant colonies. The isolated DNA was quantified and 1µg of the total DNA was digested with PvuII as per the standard restriction digestion protocol. The pISP2/Km is a 4483bp plasmid with the presence of P element of size 600bp between 2240 and 2840 positions. PvuII restriction endonuclease cuts at 4 nucleotide positions 2054, 3173, 3935 and 4173 respectively. If excision has taken place then the PvuII digestion of the repaired pISP2/Km reporter plasmid will show a 0.5kb band. In case that the event of excision has not occurred, a 1.1kb band is observed with intact P element (0.6kb) and kanamycin resistance gene (0.5kb) with 3 other bands corresponding to 2kb, 0.75kb and 0.25kb each.

In vivo Integration assay

Cg4-Neo reporter plasmid was co-transfected with the transposase expression plasmid in HEK293 cells in a 6-well plate when the cells reached ~70% confluency. A total of 2 ug of DNA was transfected: 50ng of CgNeo4, 1ug of the transposase plasmid and the amount was made upto 2ug with pBluescript empty vector. Post 48 hours of transfection, the cells were transferred to 10 cm dishes and were allowed to adhere for 24 hours at 37 degrees in a CO2 incubator. After 24 hours, the cells were grown in media supplemented with G418 (0.5 mg/ml) and selected for 2 weeks. After 2 weeks the colonies were fixed with methanol and stained with crystal violet. The relative integration activity was computed as a percentage of activity of each mutant to the wild type transposase.

Protein Immunoblotting

The cell extracts from transfected cells were lysed and denaturing PAGE was carried out using 10% SDS gel in the running buffer consisting of 1X Tris- Glycine SDS. PageRuler™ Prestained Protein Ladder was used as a molecular size marker and the samples were transferred to the PVDF membrane by electroblotting using a standard gel transfer system. The membrane was blocked to remove non specific binding using 3% (w/v) skimmed milk in Tris-Buffered Saline with Tween-20 (TBST). Subsequent washes with 1X TBST were carried out and incubation with primary anti- HA antibody (1: 2000) was done overnight at 4 degree in a rocking shaker. After probing with secondary antibody (1:5000), the blot was developed using a luminescent substrate ECL, and the protein was detected using Bio- Rad Gel Documentation system.

Supplementary Section

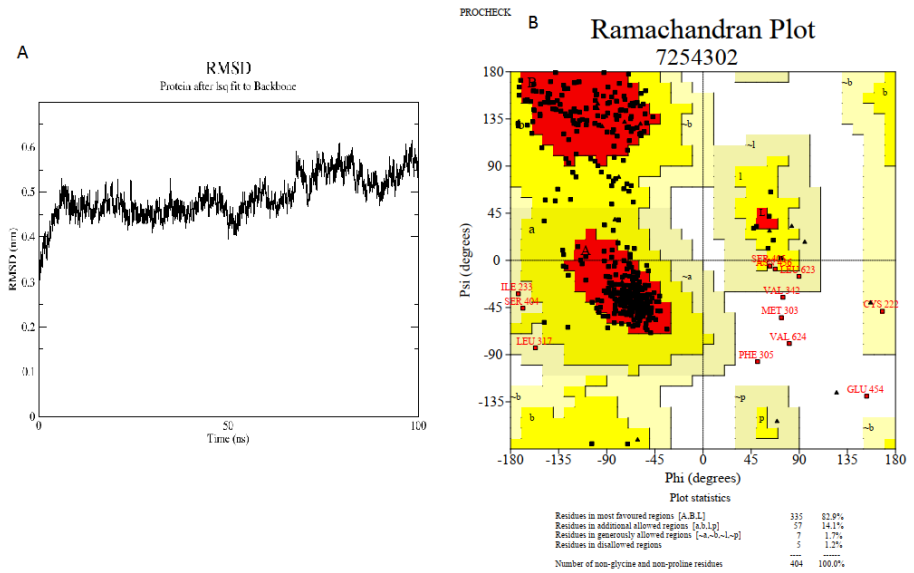

Fig S1: RMSD plot and Ramachandran plot after MD simulations.

Fig S2: SWISS MODEL secondary structure alignment with P-element template. D304 highlighted in black box does not align to any structure and end up in a loop in final structure.

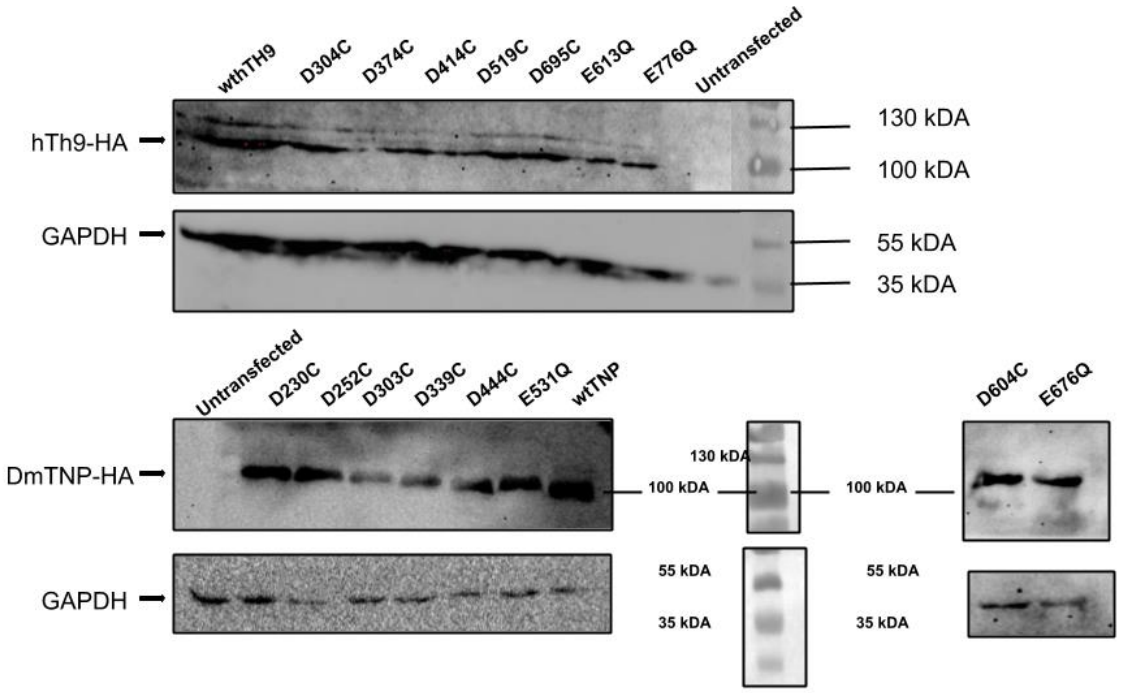

Fig S3: Immunoblots of P-element and THAP9 in transposon excision assay. The immunoblotting was performed using anti-HA antibody.

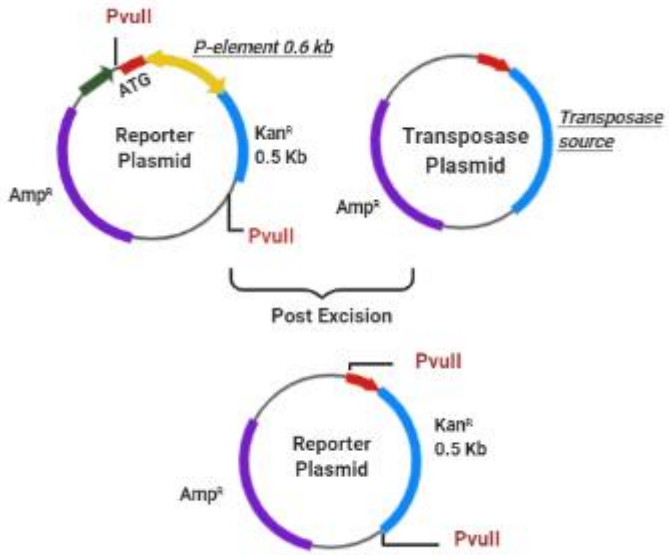

Fig S4 Schematic representation of PvuII digestion assay of recovered reporter plasmid

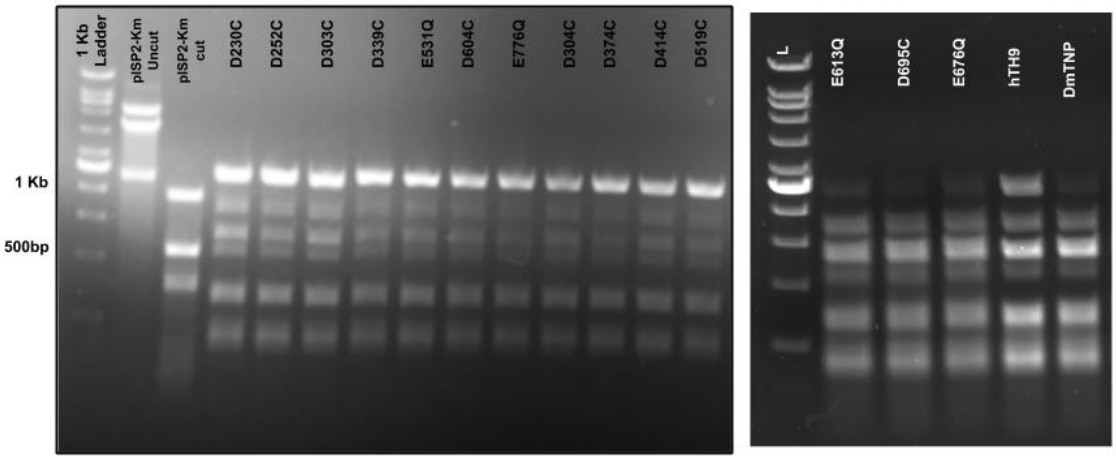

Fig S5 Agarose gel image of reporter plasmid after restriction digestion with PvuII enzyme

Acknowledgements: Althaf, Chinmay Panda, Deepshikha

References

1. Richardson, J. M., Colloms, S. D., Finnegan, D. J. & Walkinshaw, M. D. Molecular Architecture of the Mos1 Paired-End Complex: The Structural Basis of DNA Transposition in a Eukaryote. *Cell* **138**, 1096–1108 (2009).

2. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).

3. Sinzelle, L., Izsvák, Z. & Ivics, Z. Molecular domestication of transposable elements: From detrimental parasites to useful host genes. *Cell. Mol. Life Sci.* **66**, 1073–1093 (2009).

4. Jangam, D., Feschotte, C. & Betrán, E. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends Genet.* **33**, 817–831 (2017).

5. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* **41**, 331–368 (2007).

6. Hickman, A. B., Chandler, M. & Dyda, F. *Integrating prokaryotes and eukaryotes: DNA transposases in light of structure*. *Critical Reviews in Biochemistry and Molecular Biology* vol. 45 (2010).

7. Lander, E. S. *et al.* Erratum: Initial sequencing and analysis of the human genome: International Human Genome Sequencing Consortium (Nature (2001) 409 (860-921)). *Nature* **412**, 565–566 (2001).

8. Yuan, Y. W. & Wessler, S. R. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7884–7889 (2011).

9. Nowotny, M., Gaidamakov, S. A., Crouch, R. J. & Yang, W. Crystal structures of RNase H bound to an RNA/DNA hybrid: Substrate specificity and metal-dependent catalysis. *Cell* **121**, 1005–1016 (2005).

10. Asante-Appiah, E. & Skalka, A. M. HIV-1 integrase: structural organization, conformational changes, and catalysis. *Adv. Virus Res.* **52**, 351–369 (1999).

11. Rice, P. A. & Baker, T. A. Comparative architecture of transposase and integrase complexes. *Nat. Struct. Biol.* **8**, 302–307 (2001).

12. Hickman, A. B. & Dyda, F. DNA Transposition at Work. *Chem. Rev.* **116**, 12758–12784 (2016).

13. Ghanim, G. E. *et al.* HHS Public Access. **26**, 1013–1022 (2020).

14. Kofler, R., Hill, T., Nolte, V., Betancourt, A. J. & Schlötterer, C. The recent invasion of natural Drosophila simulans populations by the P-element. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6659–6663 (2015).

15. Kidwell, M. G. Evolution of hybrid dysgenesis determinants in Drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.* **80**, 1655–1659 (1983).

16. Munoz-Lopez, M. & Garcia-Perez, J. DNA Transposons: Nature and Applications in Genomics. *Curr. Genomics* **11**, 115–128 (2010).

17. Majumdar, S., Rio, D. C. & Berkeley, B. *P Transposable Elements in Drosophila and other Eukaryotic Organisms*. *Mobile DNA III* vol. 3 (2015).

18. Rice, P. A. Structure of the P element transpososome reveals new twists on the DD(E/D) theme. *Nat. Struct. Mol. Biol.* **26**, 989–990 (2019).

19. Hammer, S. E., Strehl, S. & Hagemann, S. 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).

20. Hagemann, S. & Hammer, S. E. The implications of DNA transposons in the evolution of P elements in zebrafish (Danio rerio). *Genomics* **88**, 572–579 (2006).

21. Majumdar, S., Singh, A. & Rio, D. C. The human THAP9 gene encodes an active P-element DNA transposase. *Science (80-.).* **339**, 446–448 (2013).

22. Yampolsky, L. Y. & Stoltzfus, A. The exchangeability of amino acids in proteins. *Genetics* **170**, 1459–1472 (2005).

23. Beall, E. L. & Rio, D. C. Drosophila P-element transposase is a novel site-specific endonuclease. *Genes Dev.* **11**, 2137–2151 (1997).

24. Ma, B. G. *et al.* Characters of very ancient proteins. *Biochem. Biophys. Res. Commun.* **366**, 607–611 (2008).

25. Majorek, K. A. *et al.* The RNase H-like superfamily: New members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res.* **42**, 4160–4179 (2014).

26. Moelling, K., Broecker, F., Russo, G. & Sunagawa, S. RNase H As gene modifier, driver of evolution and antiviral defense. *Front. Microbiol.* **8**, 1–20 (2017).

27. Rice, P., Craigie, R. & Davies, D. R. Retroviral integrases and their cousins. *Curr. Opin. Struct. Biol.* **6**, 76–83 (1996).

28. Montaño, S. P., Pigli, Y. Z. & Rice, P. A. The Mu transpososome structure sheds light on DDE recombinase evolution. *Nature* **491**, 413–417 (2012).

29. Yang, W. & Steitz, T. A. Recombining the structures of HIV integrase, RuvC and RNase H. *Structure* **3**, 131–134 (1995).

30. Nesmelova, I. V. & Hackett, P. B. DDE transposases: Structural similarity and diversity. *Adv. Drug Deliv. Rev.* **62**, 1187–1195 (2010).

31. Zhou, L. *et al.* Transposition of hAT elements links transposable elements and V(D)J recombination. *Nature* **432**, 995–1001 (2004).

32. Hickman, A. B. *et al.* Molecular architecture of a eukaryotic DNA transposase. *Nat. Struct. Mol. Biol.* **12**, 715–721 (2005).

33. Robertson, H. M. The Tcl-mariner superfamily of transposons in animals. *J. Insect Physiol.* **41**, 99–105 (1995).

34. Zhang, X., Jiang, N., Feschotte, C. & Wessler, S. R. PIF- and Pong-Like Transposable Elements: Distribution, Evolution and Relationship with Tourist-Like Miniature Inverted-Repeat Transposable Elements. *Genetics* **166**, 971–986 (2004).

35. Feschotte, C. 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).

36. Kapitonov, V. V. & Jurka, J. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet.* **23**, 521–529 (2007).

37. Kapitonov, V. V. & Jurka, J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol.* **3**, 0998–1011 (2005).