nature structural & molecular biology

ARTICLES

https://doi.org/10.1038/s41594-019-0319-6

Structure of a P element transposase–DNA complex reveals unusual DNA structures and GTP-DNA contacts

George E. Ghanim[1,2,6], Elizabeth H. Kellogg[2,5,6*], Eva Nogales[1,3,4] and Donald C. Rio[1,2*]

P element transposase catalyzes the mobility of P element DNA transposons within the *Drosophila* genome. P element transposase exhibits several unique properties, including the requirement for a guanosine triphosphate cofactor and the generation of long staggered DNA breaks during transposition. To gain insights into these features, we determined the atomic structure of the *Drosophila* P element transposase strand transfer complex using cryo-EM. The structure of this post-transposition nucleoprotein complex reveals that the terminal single-stranded transposon DNA adopts unusual A-form and distorted B-form helical geometries that are stabilized by extensive protein-DNA interactions. Additionally, we infer that the bound guanosine triphosphate cofactor interacts with the terminal base of the transposon DNA, apparently to position the P element DNA for catalysis. Our structure provides the first view of the P element transposase superfamily, offers new insights into P element transposition and implies a transposition pathway fundamentally distinct from other cut-and-paste DNA transposases.

Transposons are mobile genetic elements that move by a DNA rearrangement reaction using an element-encoded transposase and are ubiquitous among the genomes of all organisms. The *Drosophila* P element is one such well-characterized cut-and-paste DNA transposon that spread rapidly (within ~60 years) throughout wild populations of *Drosophila melanogaster* in the early to mid twentieth century[1]. In the late 1970s, mobilization of P elements within the *Drosophila* germline was identified as the causative agent of hybrid dysgenesis, a syndrome of aberrant genetic traits linked to mutation, chromosomal rearrangements and sterility[2]. After their initial discovery as mobile elements, P elements were engineered as a critically important tool for *Drosophila* molecular genetics and germline transformation[3]. P elements also served as a model system for understanding DNA repair mechanisms[4], the role of PIWI-interacting small RNA pathways that drive transposon adaption and limit transposon mobility[5,6], and for identifying RNA binding proteins as regulators of tissue-specific alternative splicing[7,8].

It is now appreciated that the N-terminal site-specific DNA-binding domain of P element transposase (TNP), termed a Thanatos-associated protein or THAP domain, is a very common C_2CH zinc-binding, DNA binding domain[9]. For example, in the human genome there are 12 THAP domain-containing genes. THAP9, in particular, displays extensive homology along the entire length of TNP and exhibits transposase activity upon *Drosophila* P elements[10]. However, the human THAP9 locus lacks the hallmarks of a mobile genetic element (that is, THAP9 is present as a single copy and lacks terminal inverted repeats (TIRs) and target site duplications (TSDs))[11,12]. The cellular function of THAP9 has yet to be identified.

The 2.9 kbp full-length P element transposon possesses 31 base pair (bp) TIRs, internal THAP domain binding sites, internal 11 bp inverted repeats (IIRs) and an encoded transposase gene[13–16] (Fig. 1a). The 5′ and 3′ P element transposon ends differ in the spacing between the THAP domain DNA-binding sites and the TIRs. Previous studies indicate that transposition is initiated by binding of a transposase tetramer to one P element end, followed by pairing of the transposon ends into what is termed a synaptic or paired end complex (PEC). Assembly of this higher-order nucleoprotein complex requires a guanosine triphosphate (GTP) cofactor[17,18] and is necessary for the subsequent DNA cleavage (excision) reaction in which the P element transposon is excised from flanking host DNA[19]. Like other transposable elements, 3′ cleavage occurs at the end of the P element DNA, but 5′ top strand cleavage occurs 17 bp within the P element 31 bp inverted repeats, generating atypically long 17-nucleotide 3′-single-stranded extensions at the transposon termini[19]. These staggered transposon ends are the substrate that transposase uses to integrate P element DNA into a target site.

The excised transposon–transposase nucleoprotein complex is termed the cleaved donor complex (CDC), which then locates, captures and integrates the transposon DNA into a target site elsewhere in the genome. Large-scale analysis of P element insertion sites revealed a preference for integration into a 14 bp palindromic target sequence motif (TSM) that contains the previously known 8 bp GC-rich target site, flanked by 3 bp AT-rich sequences[20]. Integration into the central portion of the TSM, followed by disassembly and host DNA repair, gives rise to the characteristic 8 bp direct TSD[13].

Among the characterized DNA transposases, TNP is mechanistically distinct in the requirement of GTP[21] and the unusually long staggered cleavage of the transposon termini[19]. To understand the mechanisms underlying the unique features of the P element superfamily, we prepared and characterized protein–DNA transposition complexes and used cryo-EM to determine the structure of the TNP strand transfer complex (STC) at 3.6 Å resolution. Our structure reveals a dimeric arrangement of the transposase protein intimately engaged with the transposon and target DNAs, providing the

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [2]California Institute for Quantitative Biosciences, University of California, Berkeley, Berkeley, CA, USA. [3]Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, CA, USA. [4]Molecular Biophysics and Integrative Bio-Imaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [5]Present address: Molecular Biology and Genetics Department, Cornell University, Ithaca, NY, USA. [6]These authors contributed equally: George E. Ghanim, Elizabeth H. Kellogg. *e-mail: ehk68@cornell.edu; don_rio@berkeley.edu

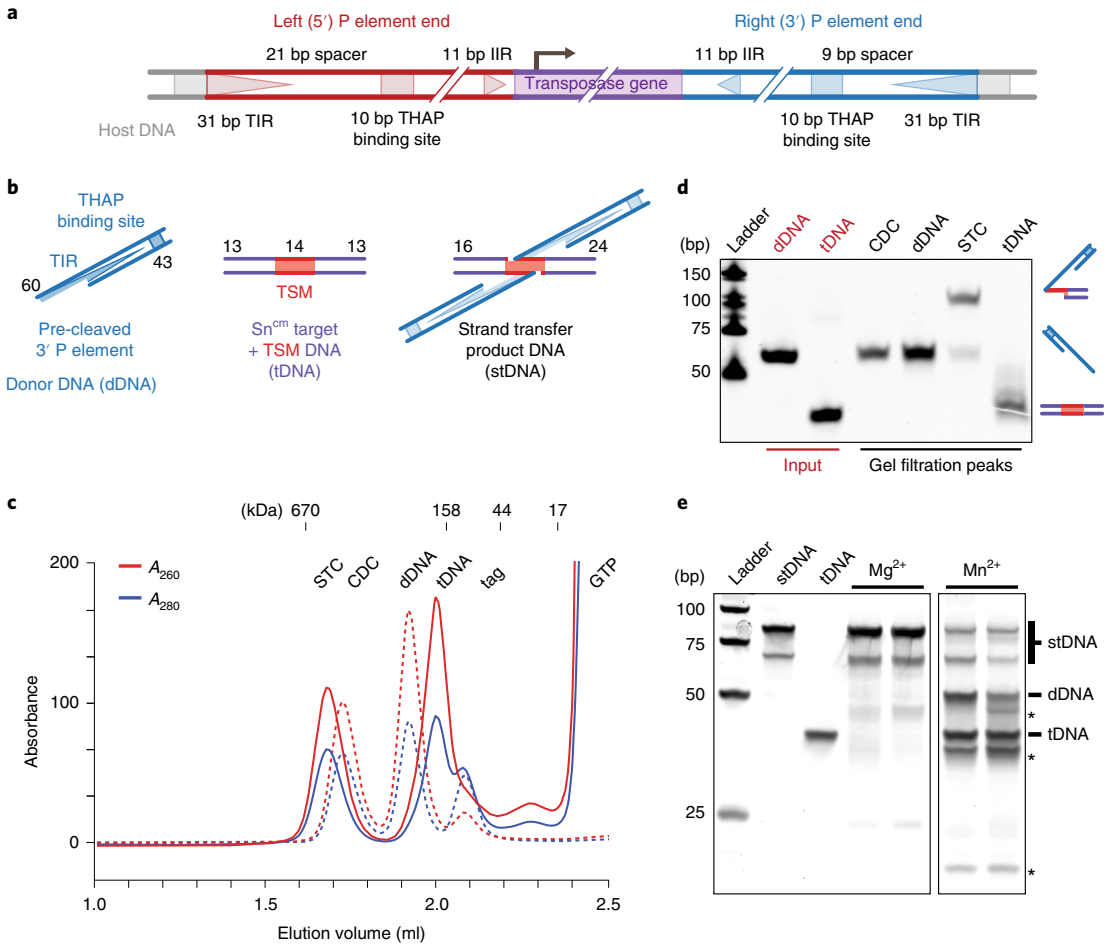

Fig. 1 | Reconstituted strand transfer complex represents the active form of TNP. a, Diagram of the full-length P element transposon depicting the differently spaced 5′ and 3′ ends. The 31 bp TIRs (triangles), 10 bp THAP domain binding site (squares), the 11 bp IIRs (triangles) and the TNP gene (purple) are indicated. The 5′ and 3′ P element ends are colored red and blue, respectively. Not drawn to scale. **b**, Schematic of the DNA substrates used. The nucleotide length of each strand is indicated (TSM, target sequence motif; dDNA, donor DNA; tDNA, target DNA; stDNA, strand transfer product DNA). Not drawn to scale. **c**, Cleaved donor complex (CDC) and strand transfer complex (STC) gel filtration elution profiles (CDC, dotted lines; STC, solid lines). Absorbance A_{260} and A_{280} is indicated in red and blue, respectively. Elution positions of mass standards (in kDa) are shown above. **d**, SYBR Gold stained urea PAGE of dDNA input, tDNA input and peak fractions from **c**. Schematics of DNAs are shown to the right. Input DNA standards are colored red. bp, base pairs of markers. **e**, SYBR Gold stained native PAGE gel of disintegration assay with strand transfer product DNA. The expected mobilities of the dDNA and tDNA products are indicated to the right. Unidentified bands are indicated with asterisks. The uncropped gel image is provided in Supplementary Data Set 1.

first detailed view of the P element product DNA–protein complex. Surprisingly, we find that the 17 nt DNA extension at the transposon ends is not simply single-stranded but base pairs in an unusual A-form DNA arrangement. To our knowledge, this DNA arrangement has not been observed in other nucleoprotein structures. In addition, we observe direct interactions between the guanine in GTP and the terminal guanosine residue of the transposon DNA that probably acts to position the reactive transposon DNA end into the active site, providing a rationale for the requirement for GTP. Our structure also reveals severe bending of the target DNA (tDNA) at the sites of transposition. Finally, we suggest a mechanism for pairing the differently spaced 5′ and 3′ P element ends during synaptic complex formation. Together, these results illuminate the unique features of P element transposition and how complex the interactions between transposase or integrase enzymes and their DNA substrates can be.

Results

Reconstituted STC represents the active form of TNP. Highly active samples of *Drosophila* TNP were prepared from baculovirus-infected Sf9 cells (see Methods). To assemble the STC, we first prepared the CDC by incubating TNP with a minimal pre-cleaved 3′ P element donor DNA (dDNA) end in the absence of GTP and Mg^{2+} (see Methods, Fig. 1b and Supplementary Fig. 1a). The STC was then prepared by incubating the CDC overnight at 30 °C with GTP, Mg^{2+} and an optimized tDNA derived from the *Drosophila singed* locus, a hotspot for P element transposition[20,22,23] (Fig. 1b and Supplementary Fig. 1b). Fractionation by size exclusion chromatography (SEC) of either the CDC or STC sample produced higher-order species with distinct elution profiles (Fig. 1c and Supplementary Fig. 1d). Analysis of the DNA from deproteinized SEC fractions revealed that the CDC fraction contained dDNA, while the STC fraction contained a slower-mobility species, resulting from strand transfer of the dDNA into the tDNA generating the strand transfer product DNA (stDNA) (Fig. 1d). The abundance of the slower-mobility stDNA species indicates that the CDC preparations are highly active for strand transfer.

To further improve STC sample homogeneity, we assembled TNP on a symmetric branched DNA substrate mimicking the product of a double-ended integration reaction, with the 3′ dDNA covalently attached to the target (Fig. 1b, stDNA and Supplementary Fig. 1c),

a strategy used for retroviral intasomes[24–27]. Particles in negative-stained electron micrographs of STC complexes assembled on stDNA were indistinguishable from authentically generated STC (Supplementary Fig. 1e). To assess the biological relevance of STC samples prepared this way, we exploited a property of transposases and retroviral integrases termed 'disintegration'[28–32]. In the presence of Mn^{2+}, transposase will reverse the transesterification reactions of strand transfer, liberating the dDNA and rejoining the tDNA strands to give products that resemble an unintegrated dDNA and a duplex tDNA[33]. In the presence of transposase, disintegration of the stDNA to dDNA and tDNA was observed with Mn^{2+}, but not with Mg^{2+} (Fig. 1e). Minor faster migrating bands were also observed (Fig. 1e, asterisks), and may arise from an alternate reversal foldback pathway that has been observed for *Mu* transposase[30] and retroviral integrases[34]. Reversal of strand transfer in the presence of Mn^{2+} demonstrates that, for the majority of complexes, the stDNA is properly positioned within the STC active site for catalytic nucleophilic attack, as would be expected in an authentic STC. We did attempt to generate asymmetric stDNA substrates with 5′ and 3′ P element ends, but this produced mixed 3′-3′, 3′-5′ and 5′-5′ samples, decreasing the homogeneity.

The STC structure is dimeric and reveals four domains in each monomer. Single particle cryo-EM data were collected on an Arctica microscope equipped with a K2 detector (Supplementary Fig. 2a). Computational processing and iterative refinement yielded a final reconstruction with fairly uniform local resolution, ranging between 3.5 and 4 Å (Table 1 and Supplementary Fig. 2b–g). Transposase can be divided into six structural domains (Fig. 2a,b), four of which could be modeled de novo. The N-terminal THAP DNA-binding domain and a majority of the following dimerization domain[35–38] are not resolved in the reconstruction due to flexibility. Thus, our model begins with the N-terminal DNA-binding helix-turn-helix domain (HTH; dark cyan), followed by a split catalytic RNase H domain (RNase H; orange) that is interrupted by a GTP-binding insertion domain (GBD; blue), and a C-terminal domain (CTD; red) (Fig. 2a–d and Supplementary Video 1). The linker between the RNase H domain and the C-terminal domain (residues 570–616) is not visible in the density map (Fig. 2c, left, white asterisks), consistent with the high probability for disorder in this region[39] (Supplementary Fig. 3a). However, the orientation of the sparse density at the beginning and end of this linker suggests that the depicted RNase H and C-terminal domains are connected to constitute a monomer (Fig. 2c, left, white asterisks, and Supplementary Fig. 3b,c). The last 17 residues of the C terminus are also not visible, again consistent with computational disorder predictions[39] (Supplementary Fig. 3a).

Our structure reveals that the STC adopts a dimeric assembly arranged with two-fold symmetry around the stDNA (Fig. 2c,d, Supplementary Fig. 3c and Supplementary Video 1). We note that 26 bp of the 40 bp tDNA and the first 23 bp out of 55 bp of each dDNA are not well resolved in the symmetrized reconstruction. Each monomer closely interacts with the pre-cleaved P element 31 bp TIR dDNAs. The two dDNAs adopt a 55° angle relative to each central duplex axis (Fig. 2c,d and Supplementary Video 1) and insert into the tDNA, separated by 8 bp (the characteristic TSD size). The tDNA is distorted and bent, as observed in other transposase and retroviral integrase structures[25,27,40–42] (Fig. 2c,d).

The catalytic RNase H domain adopts a canonical RNase H-like fold, similar to that found in other DDE (Asp-Asp-Glu) transposases and related retroviral integrases[43] (Fig. 2d, left). Notably, the α-helical GTP-binding domain is inserted into the RNase H fold, between the fifth β-strand and fourth α-helix. This location is amenable to insertions, as observed in several other transposases and transposase-like proteins (Supplementary Fig. 4a,b). We additionally identified densities that correspond to GTP and a coordinated magnesium ion in the GTP-binding domain (Fig. 2e).

Table 1 | Cryo-EM data collection, refinement and validation statistics

	STC-C2 (EMD-20254, PDB 6P5A)	STC-C1 (EMD-20321, PDB 6PE2)
Data collection and processing		
Magnification	35,000	35,000
Voltage (kV)	200	200
Electron exposure (e^{-} Å^{-2})	60	60
Defocus range (μm)	−1 to −3 μm	−1 to −3 μm
Pixel size (Å)	1.16	1.16
Symmetry imposed	C2	C1
Initial particle images (no.)	547,929	547,929
Final particle images (no.)	252,574	252,574
Map resolution (Å)/FSC threshold	3.6/0.143	3.9/0.143
Map resolution range (Å)	3–5	4–10
Refinement		
Initial model used	-	6P5A
Model resolution (Å)/FSC threshold	3.7/0.5	4/0.5
Model resolution range (Å)	-	-
Map sharpening *B* factor (Å^2)	100	100
Model composition		
Non-hydrogen atoms	11,956	12,753
Protein residues	1,120	1,148
Ligands	6	6
B factors (Å^2)		
Protein	47.13	185.97
Ligand	38.87	171.81
R.m.s. deviations		
Bond lengths (Å)	0.008	0.003
Bond angles (°)	0.52	0.532
Validation		
MolProbity score	1.22	1.56
Clashscore	4.46	6.4
Poor rotamers (%)	0%	0%
Ramachandran plot		
Favored (%)	98%	96.75%
Allowed (%)	2%	3.25%
Disallowed (%)	0%	0%

Within the RNase H domain, we identified the three catalytic acidic residues, D230 located on β1, D303 after β4 and E531 on α4 (Fig. 2f), in agreement with previous computational predictions[44]. Indeed, alanine substitution of any one of these acidic residues eliminates TNP excision activity in vivo[45], confirming their essential role for transposase catalytic activity (Supplementary Fig. 4c,d). The RNase H domains are located near the donor-target DNA junctions, with the catalytic residues coordinating a Mg^{2+} ion (Fig. 2f). However, the scissile phosphate of the tDNA at the donor-target junction is rotated out of the active site (Fig. 2g, cyan phosphate). Because this is a product complex, this rotation may have occured to prevent reversal of the integration reactions. A similar configuration

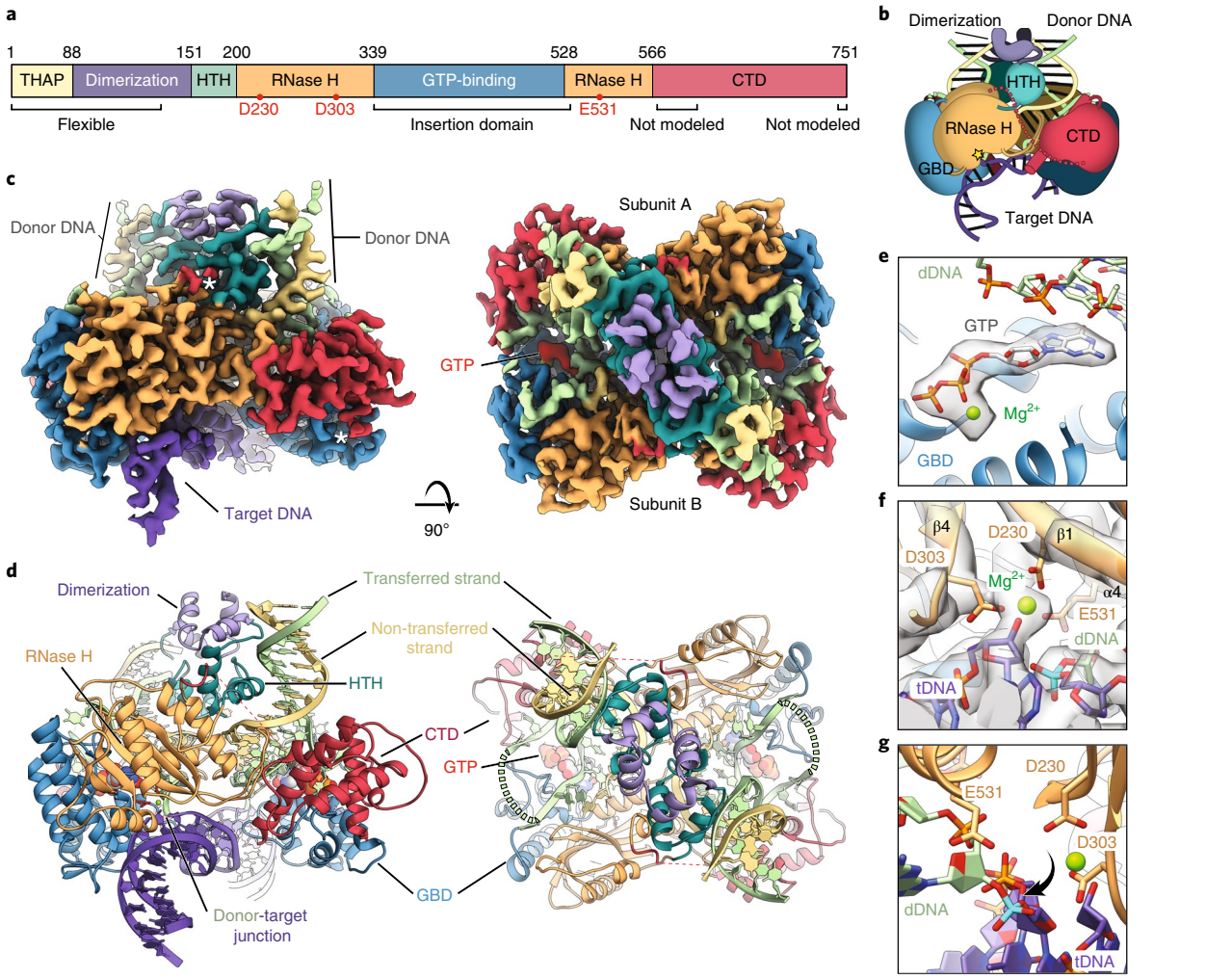

Fig. 2 | Structure of the *Drosophila* P element STC. a, Domain architecture of *Drosophila* TNP with the domain boundaries indicated by amino acid residue numbers. The RNase H catalytic residues are indicated as red dots. THAP, THAP DNA-binding domain (yellow); dimerization, leucine zipper dimerization domain (purple); HTH, helix-turn-helix domain (dark cyan); RNase H, RNase H-like catalytic domain (orange); GTP-binding, GTP-binding insertion domain (blue); CTD, C-terminal domain (red). **b**, Cartoon of the TNP STC. The catalytic site is indicated with a yellow star and domains are colored as in **a**. Domains of the other subunit are darkened (GBD, GTP-binding insertion domain). **c**, Side (left) and top (right) views of the cryo-EM reconstruction at 3.6 Å. Domains are colored as in **a** and GTP is colored red. White asterisks indicate the sparse density of the disordered RNase H-CTD linker. **d**, Side (left) and top (right) views of the TNP STC model (colored as in **c**, with domains indicated). Catalytic residues are colored red and unmodeled connections are shown as dashed lines (dashed green, dashed red). tDNA is shown in purple, the donor transferred strand in light green and the donor non-transferred strand in yellow. **e**, Close-up view of the GTP density. Only the density corresponding to GTP is shown for clarity. **f**, Close-up view of the RNase H catalytic residues. The density is as in **c**, with relevant residues labeled. The scissile phosphate is colored cyan. **g**, Close-up view showing the scissile phosphate rotation out of the RNase H active site. The view is similar to that in **f**, but rotated 90°. Density is omitted for clarity. The scissile phosphate is colored cyan.

of a donor-target DNA junction was observed with the prototype foamy virus retroviral integrase STC[42].

The three additional domains of TNP, a previously unrecognized HTH domain, the GTP-binding domain and the C-terminal domain, all participate in protein-DNA interactions. The HTH domain directly contacts the dDNA (Fig. 3b). The GTP-binding domain packs against the RNase H fold and extends a loop to contact the central region of the tDNA (see 'Altered tDNA structure stimulates transposition' section below). As with the GTP-binding domain we also observe protein-DNA contacts between the C-terminal domain and the stDNAs.

The dDNAs adopt an unusual, partially distorted structure with A- and B-form helices. An unusual feature of P element transposition is the staggered cleavage of the transferred and non-transferred strands at the P element ends, resulting in 17 nt 3′ single-stranded DNA (ssDNA) overhangs. We were able to place 12 of the 17 nt into our reconstruction. One unanticipated observation is the unusual configuration of the DNA at the P element ends. We observe that the 3′ region of the transferred strand base pairs with the 5′ portion of the non-transferred strand resulting in a short A-form DNA duplex (Fig. 3a and Supplementary Fig. 5a,b). The transferred strand is displaced from the non-transferred strand at nucleotide C_{-22} to accommodate the A-form duplex (Fig. 3a, schematic). This displaced transferred strand 'loops out' and is stabilized by numerous contacts from the C-terminal and GBD, including aromatic base-stacking interactions from Y721, F722, F384, Y629 (Figs. 3c and 4) and Y519 (Fig. 4).

To investigate the importance of base pairing between distant regions in the dDNA, we performed in vitro strand transfer assays with mutated dDNA substrates (Supplementary Fig. 5c). Mismatches introduced into the transferred strand that disrupt

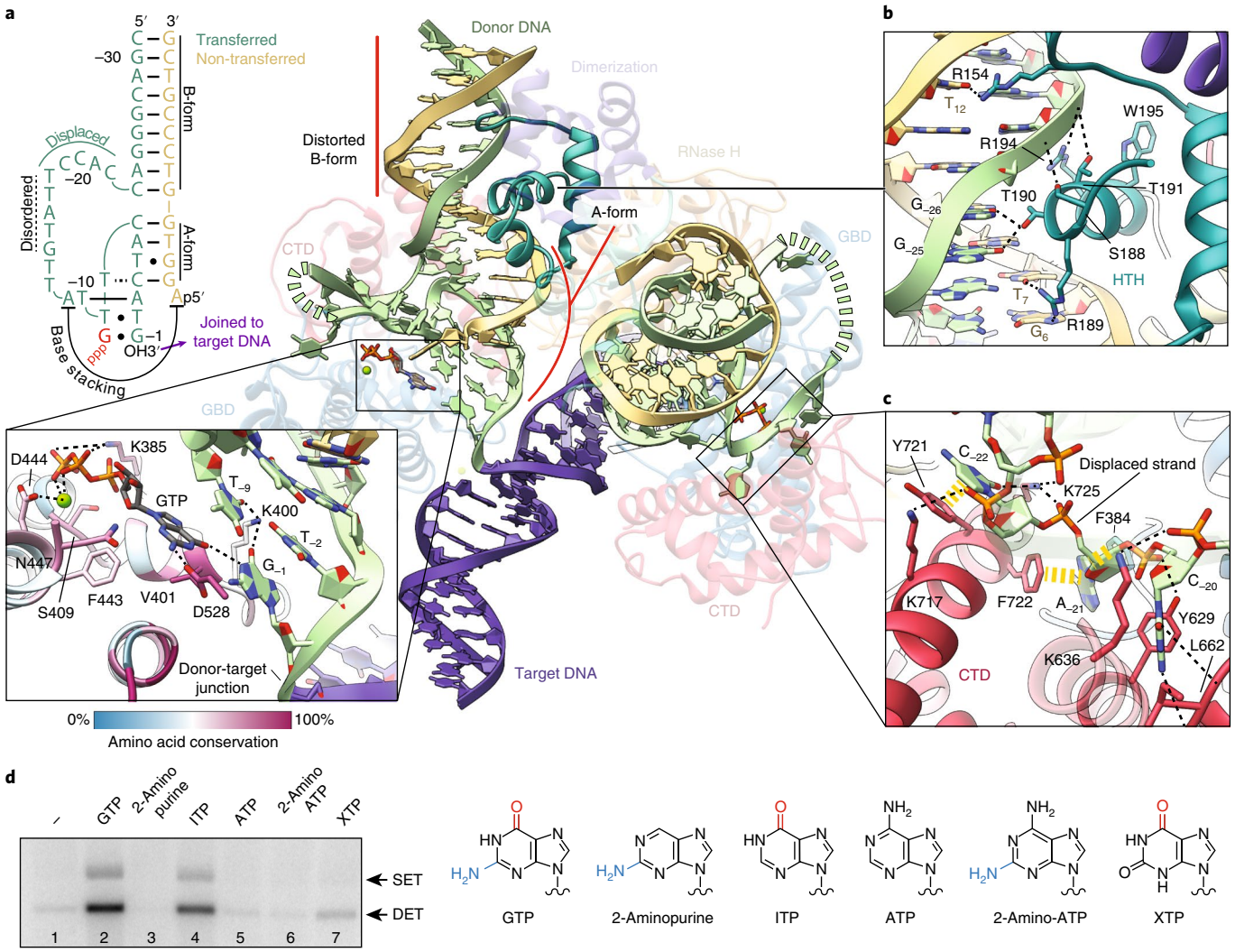

Fig. 3 | dDNA adopts a non-canonical geometry within the STC. a, Overview of dDNA structure within the STC. Distorted B-form and A-form regions of the dDNA are indicated. The transposase protein is faded out for clarity, with relevant domains labeled. The opposing RNase H domain was omitted for clarity. The disordered nucleotides of the transferred strand (−14 to −18) are marked by a dashed green line. Schematic of the secondary structure of dDNA TIR (top left). GTP is in red lettering. Watson–Crick base pairings are indicated by solid lines. Non-canonical base pairings are indicated by dots, or dotted lines. Nucleotides of the transferred strand are numbered −1 to −31, starting at the 3′ terminal guanosine. Inset: Close-up of the interaction between GTP, the GBD and dDNA (bottom). Inferred hydrogen-bonding and electrostatic interactions are shown as black dashed lines. Residues are colored by sequence conservation, following the coloring scheme shown in the scale bar. **b**, Close-up view of the HTH and dDNA contacts. Nucleotides are numbered as in **a**. **c**, Close-up view of the CTD and displaced transferred strand contacts. Aromatic base-stacking interactions are shown as yellow dashed lines. Inferred polar and hydrogen-bonding interactions are shown as black dashed lines. **d**, Strand transfer assay with different purine nucleoside triphosphate analogs. An agarose gel of a strand transfer assay is shown on the left, with the expected positions of single-ended integration (SET) and double-ended integration (DET). Nitrogenous base structures of the purine nucleoside triphosphates tested in this assay are shown on the right. C6 carbonyl groups and C2 amino groups are colored red and blue, respectively. The uncropped gel image is shown in Supplementary Data Set 1.

base pairing at the A-form duplex region decreased or eliminated strand transfer activity at nearly all positions (Supplementary Fig. 5c, lanes 2, 3, 5–8). Compensatory mutations on the non-transferred strand that restored base pairing were able to rescue or partially rescue strand transfer activity (Supplementary Fig. 5c, compare lanes 5–8 and 13–16, most prominently lanes 7, 8, 15 and 16). These results confirm the importance of base pairing between distant regions of the transferred and non-transferred strands for strand transfer activity.

Additional protein-DNA contacts occur via the HTH domain, which engages the dDNA at the 31 bp TIRs through a loop in the minor groove and an α-helix inserted into the major groove (Fig. 3b). Numerous backbone and base contacts are made by R154, S188, R189, T190, T191, R194 and W195 (Figs. 3b and 4). Of these, the positioning of R154, R189 and T190 leads us to infer that these side chains form base-specific hydrogen bonds with T_{12}, G_6 or T_7, and G_{-25} or G_{-26}, respectively (R154:T_{12}; R189:G_6 or T_7; T190:G_{-25} or G_{-26}).

Overall, we observe extensive protein-DNA contacts of a single subunit with both of the P element dDNAs. The depicted protein subunit in Supplementary Fig. 6 (left) is catalytically engaged with a P element end (red) through the RNase H (orange) and GTP-binding domains (not depicted). However, a 90° rotated view shows that the same subunit contacts the other P element end (blue) through the HTH domain, a long loop in the RNase H domain, and the C-terminal domain (Supplementary Fig. 6, right). Overall, the observed architecture supports a *trans*-catalysis mechanism, in which transposase binds to and holds one P element end but

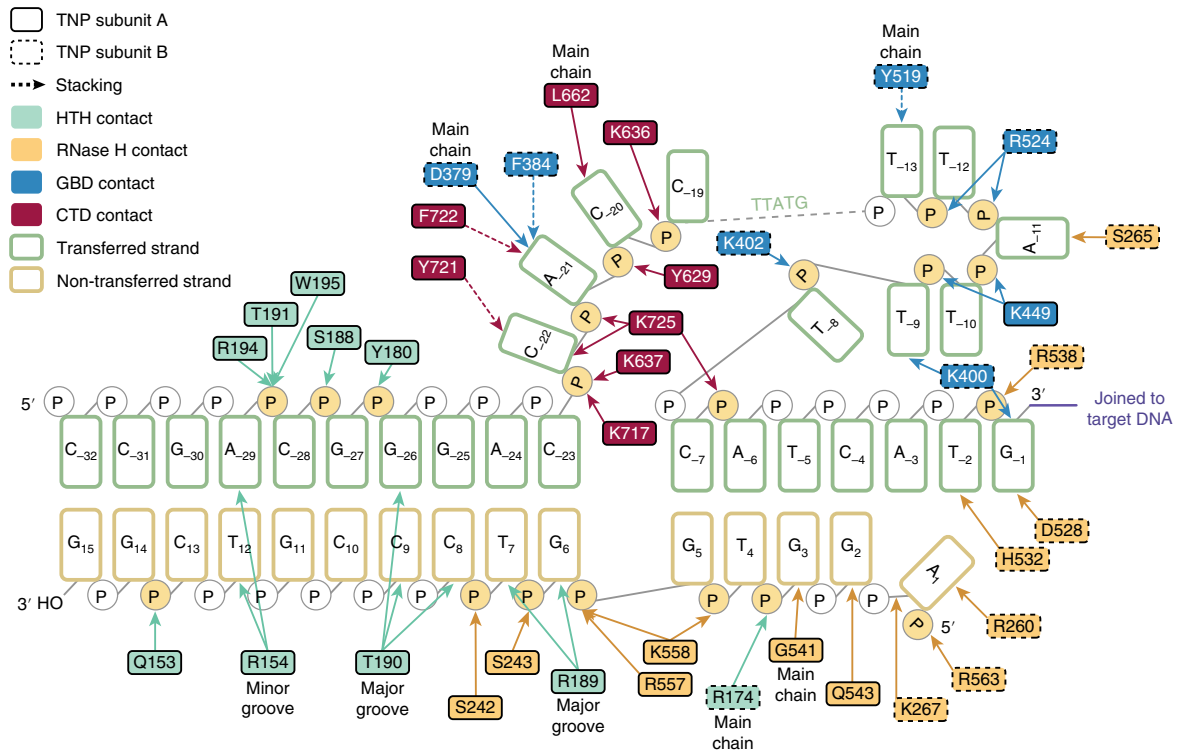

Fig. 4 | Each subunit makes extensive contacts with a single dDNA. Schematic representation of the inferred base-specific and backbone contacts between transposase and the dDNA. Nucleotides of the transferred strand (green outline) are numbered −1 to −32, starting at the 3′ terminal guanosine. Nucleotides of the non-transferred strand (gold outline) are numbered 1 to 15 starting at the 5′ adenosine. Amino acid residue numbers are indicated and outlined in a solid or dashed border to indicate transposase subunit A or transposase subunit B, respectively. Residues are colored according to domain (HTH, light cyan; RNase H, orange; GBD, blue; CTD, red). Direct contacts are shown as solid lines; aromatic base-stacking interactions are shown as dashed lines; major groove, minor groove and main chain contacts are indicated; interacting phosphates are highlighted in yellow.

catalyzes the strand transfer of the other end. This interlocking architecture probably acts as a checkpoint to ensure proper assembly of the nucleoprotein complex prior to catalysis of DNA integration.

The GTP cofactor interacts with the dDNA. TNP is unique in its requirement of a GTP cofactor for assembly of the PEC and the strand transfer reaction. We were able to identify densities that correspond to GTP and a coordinated magnesium ion (Fig. 2e). Comparison with similar resolution cryo-EM densities of other GTP-binding proteins supports our interpretation that the nucleotide density corresponds to GTP rather than GDP (Supplementary Fig. 7). Interestingly, residues that mediate GTP or metal binding (D528, K385, V401, S409, F443, D444 and N447) are conserved within members of the P element superfamily[44] (Fig. 3a, inset). We observe that GTP makes base-stacking interactions with the transferred strand (T_{-9}) and is probably hydrogen bonding with G_{-1} (the terminal dDNA nucleotide) through the GTP C6 carbonyl group. The interaction with GTP appears to alter the trajectory of the dDNA strand and may act to position the attacking 3′OH in the active site, explaining why GTP is required for strand transfer.

To investigate the interactions with GTP in the STC, we performed strand transfer assays with radiolabeled dDNAs and different purine nucleoside triphosphate analogs (Fig. 3d). Nucleotides that lacked a C6 carbonyl group did not support strand transfer activity (2-aminopurine, adenosine triphosphate (ATP), 2-amino-ATP, Fig. 3d, lanes 3, 5, 6). Conversely, inosine triphosphate (ITP) and to lesser extent xanthosine triphosphate (XTP), both of which carry the C6 carbonyl group, did support strand transfer activity, but not to the same level as GTP (Fig. 3d, lanes 2, 4, 7). This is probably due to differences in the substituents at the purine C2 position. Taken together, this experiment indicates that the purine C6 carbonyl group is critical for strand transfer activity, while the interaction between D528 and the C2 amino group probably facilitates nucleotide binding. These results and the structure support a model in which interactions with GTP act to position the dDNA for strand transfer and explain the specificity of GTP (GTP is the only nucleotide that can fully support the observed interactions at this stage of transposition).

Altered tDNA structure stimulates transposition. tDNA bending is a common feature among DDE transposases[40,41] and the related retroviral integrases[25,27,42]. Consistent with these findings, we observe substantial distortion of the tDNA within the P element STC (Fig. 5a). At each strand transfer site, the tDNA duplex exhibits a sharp ~55° bend away from the central axis (Fig. 5b). This distortion is accommodated over the AT-rich flanking sequences, which display a widened minor groove (Fig. 5b, green and Supplementary Fig. 8a). The central 8 bp GC-rich TSD duplex remains approximately B-form (Fig. 5, red).

The tDNA binds along a basic channel formed by the RNase H and GBD of each monomer (Supplementary Fig. 8b). Numerous residues from both the RNase H domain (K310, R538 and H546) and the GBD (H350, R394, Q399 and K487) are positioned to contact the phosphate backbone, probably stabilizing the observed tDNA conformation (Supplementary Fig. 8c). A loop from the GBD extends into the major groove of the 8 bp GC-rich central duplex to make phosphate (R394 and Q399) and base (S395 to G_6 and K398 to G_1) contacts (Fig. 5c). RNase H domain residues T306 and Y253 are positioned within the minor groove of the flanking AT-rich regions (Fig. 5d). T306 contacts T_{11} at the extremity of the TSM. Although Y253 is also positioned within

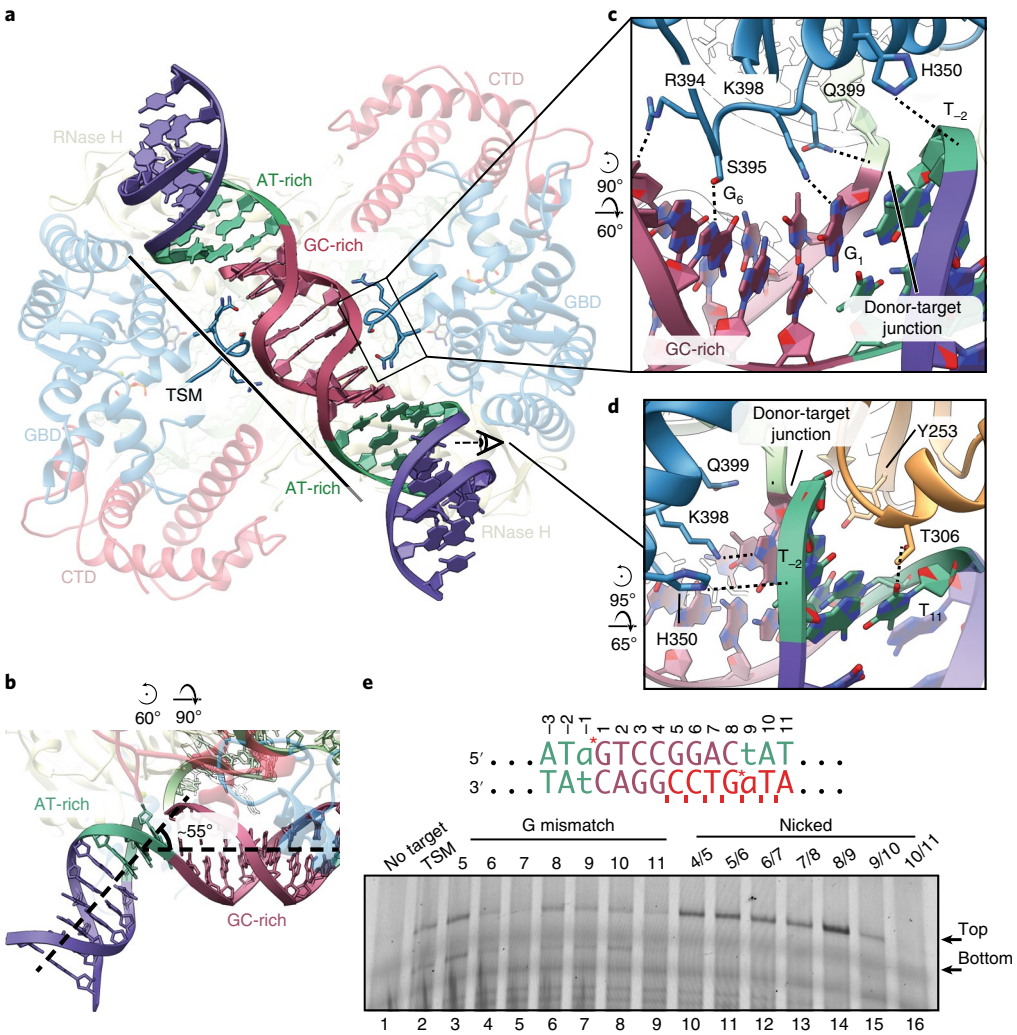

Fig. 5 | The tDNA is severely bent at AT-rich sites. a, Bottom view of the STC, highlighting the bent tDNA. AT-rich (green) and GC-rich (red) regions of the tDNA are indicated. The GBD loop that interacts with the tDNA is shown. The transposase protein is faded out for clarity with relevant domains labeled. All subsequent panel rotations are depicted with respect to **a**. **b**, Bend at flanking AT-rich sites. The bend is highlighted and dashed lines indicate the central axis of the DNA. The tDNA is colored as in **a**. **c**, Close-up view of the tDNA-GBD-loop interaction inferred from the atomic model. Site-specific interactions are indicated (S395:G_6, K398:G_1). Nucleotides are numbered as in **e**. **d**, Close-up view of tDNA-RNase H domain interaction inferred from the atomic model. Site-specific interactions are indicated (T306:T_{11}). A region of tDNA backbone has been made transparent for clarity. **e**, Denaturing PAGE gel of a transposition assay using mismatched or nicked tDNA substrates. The sequence of the TSM is shown above. Sites of transposition into the top and bottom strand are indicated with red asterisks (top strand, −1, 1; bottom strand, 8, 9). Nucleotide numbering corresponds to the top strand. G mismatches were introduced within the bottom strand at the indicated positions (red bases). Nicks were introduced into the bottom strand between the indicated positions (red ticks). Expected sizes of transposition into the top strand or bottom strand of the tDNA are indicated to the right of the gel. The transferred strand of the dDNA was fluorescently labeled at the 5′ position with a TAMRA dye. The uncropped gel image is shown in Supplementary Data Set 1.

the minor groove at the site of transposition, it does not appear to make direct base-specific contacts. This positioning may facilitate the observed widening of the minor groove or tDNA bending and thereby help position the scissile phosphate within the transposase active site. Finally, although the 17-residue C-terminal tail is not modeled, this region contains multiple basic residues and is ideally positioned to electrostatically interact with the tDNA (Supplementary Fig. 8d).

Although P element transposition is not site-specific, integration preferentially occurs into TSM or TSM-like sequences. In our structure, base-specific interactions between TNP and the tDNA are sparse, suggesting that the preference for the TSM is not achieved through direct sequence readout alone. Recent studies indicate that DNA flexibility and deformability play a critical role in transposase or integrase target site selection[46,47].

To investigate the effects of tDNA flexibility on transposase activity, we performed in vitro strand transfer assays with nicked or mismatched tDNA substrates. G mismatches or nicks were included along the bottom strand to introduce deformability and flexibility into specific regions of the tDNA duplex (Fig. 5e). Mismatches did not appreciably stimulate activity, but rather decreased activity in specific instances (Fig. 5e, lanes 4, 5 and 9). Mismatches at positions G_6 and T_{11} coincide with observed TNP-tDNA base interactions, and probably decrease affinity for the target DNA by disrupting these contacts or altering crucial duplex geometries. Notably, nicks along the bottom strand central GC-rich region increased strand transfer into the top strand of the target DNA. The greatest stimulation was observed with a nick positioned at the site of strand transfer, between nucleotides 8 and 9 on the bottom strand (Fig. 5e, lane 14). This is the same region that accommodates the highest

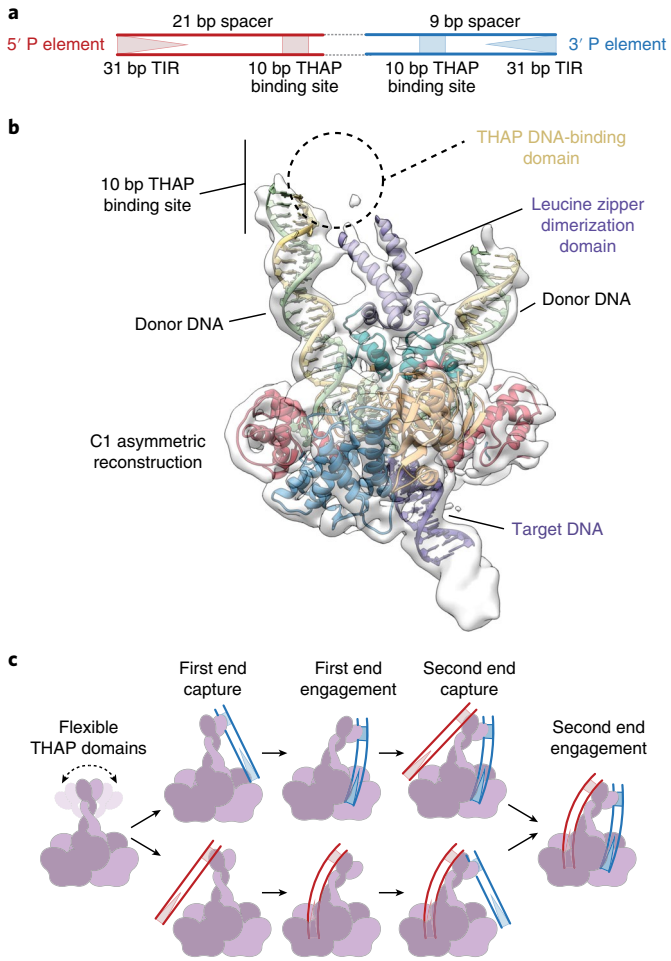

Fig. 6 | The unsymmetrized reconstruction suggests a mechanism for 5′ and 3′ P element end pairing. a, Diagram of a P element transposon depicting the differently spaced 5′ and 3′ ends. The 31 bp TIRs (triangles) and 10 bp THAP domain binding site (squares) are indicated. The 5′ and 3′ P element ends are colored red and blue, respectively. **b**, Unsymmetrized 3.9 Å reconstruction showing additional density near the N terminus. Additional dDNA and the leucine zipper dimerization domain were modeled into the density. The expected position of the THAP domain and the THAP domain binding site are indicated. **c**, Model for pairing of the 5′ and 3′ P element ends. The TNP protein (purple and light purple), 3′ P element transposon end (blue) and the 5′ P element transposon end (red) are represented as cartoons.

level of distortion within the tDNA duplex. Taken together, this supports a model in which the preference for the P element TSM is driven by a pattern of tDNA flexibility and is further enforced by the observed amino acid side chain-base interactions.

Unsymmetrized reconstruction suggests a mechanism for 5′ and 3′ P element end pairing. The 5′ and 3′ P element transposon ends differ in the spacing between the internal THAP domain DNA-binding site and the TIR (Fig. 6a). Furthermore, the 5′ end cannot substitute for the 3′ end during the initial stages of synaptic complex assembly before DNA cleavage[14,19]. These observations suggest that TNP engages differently with each P element end to ensure proper synaptic complex assembly. Our highest resolution reconstruction, in which two-fold symmetry was applied, did not resolve the N-terminal leucine zipper and THAP DNA-binding domains. However, an asymmetric, lower resolution reconstruction revealed additional density corresponding to the N-terminal leucine zipper (Fig. 6b and Supplementary Fig. 2g), while the THAP DNA-binding domain remains unresolved, probably due to flexibility. The additional 12 residues of the leucine zipper dimerization domain are oriented towards one of the 3′ P element dDNAs adjacent to the 10 bp TNP binding site. This asymmetry could accommodate and facilitate assembly of differently spaced 5′ and 3′ P element ends (Fig. 6c), reminiscent of the flexible nonamer binding domain in the RAG1–RAG2–12–23 RSS complex, which enforces the 12–23 rule of V(D)J recombination[19,48]. We propose that TNP pairs the P element ends by a mechanism analogous to that previously described for RAG1–RAG2 of V(D)J recombinase[49–51]; that is, when TNP engages with the 3′ P element end (9 bp spacer) there is an induced asymmetry, such that only the longer 5′ P element end (21 bp spacer) can span the distance between the THAP DNA-binding domain and the catalytic core. Conversely, when the transposase engages the longer 5′ P element end, the induced asymmetry will dictate that only the shorter 3′ P element end can fit between the THAP DNA-binding domain and the catalytic core. However, we note that the disorder at this region of the structure may be caused by the flexibility of the P element DNA ends, as well as by the use of two 3′ end dDNAs to assemble this complex.

Discussion

P elements are one of the best-studied eukaryotic DNA transposons and have revealed a wealth of insights into the mechanisms and regulation of DNA transposition, as well as fundamental cellular processes such as tissue-specific alternative splicing and DNA repair pathways. Among previously characterized DNA transposases, TNP is unique in at least two respects. First, GTP is required as a cofactor for the DNA pairing, cleavage and strand transfer stages of transposition. Second, the staggered cleavage of the transposon ends is atypical in length, resulting in a 17 nt 3′ single-stranded transposon DNA extension. Here, we provide the first three-dimensional view of the P element superfamily of eukaryotic DNA transposases, illuminating many mechanistic features.

Our structure reveals a complex nucleoprotein architecture and allows the unambiguous identification of the domain organization of TNP, including a HTH domain, a catalytic RNase H domain, a GBD and a highly charged C-terminal domain. The GBD is inserted into the RNase H catalytic domain. The location of this insertion domain is similar to other insertion domains found in bacterial Tn*5*, housefly *Hermes* and the jawed vertebrate V(D)J RAG1 enzymes (Supplementary Fig. 4a). In fact, some of the insertion domains share structural similarity (Supplementary Fig. 4b).

TNP is unique in using GTP as a non-hydrolyzed cofactor for both the cleavage and integration steps of transposition. Our data reveal that the guanine base of GTP interacts with the terminal transposon base, altering its trajectory from the A-form duplex and potentially directing the 3′OH toward the RNase H active site. This suggests that GTP is used to position the terminal transposon G-3′OH for catalysis, linking the requirement of the GTP cofactor to direct interactions with the terminal base of the transposon DNA, thereby providing a rationale for the requirement of GTP during strand transfer.

Previous studies with full-length P element ends indicated that a transposase tetramer acts at the early stages of transposition in forming synaptic PECs and CDCs[17,18]. However, we observed that the STC is dimeric. Assembly of the STC used minimal oligonucleotide dDNA substrates, rather than the two full-length ~150 bp P element ends. The longer P element ends include the 11 bp IIRs, which act as transpositional enhancers in vivo[14]. It is possible that a tetramer (or a dimer of dimers) initially assembles to pair the natural P element ends and activate the protein for dDNA cleavage. Once this complex excises the P element DNA and rearranges the terminal cleaved transposon ends, it is possible that loss of two

catalytic subunits occurs to form the dimeric complex, as we have observed, which captures a tDNA and performs strand transfer. Contributions to DNA-binding by non-catalytic subunits has been observed in both the bacteriophage *Mu* transposome[40] and the retroviral integrase structures[25,27] and is thought to occur in the octameric *Hermes* transposome[52].

Overall, our structure suggests that, during the early stages of transposition, when the THAP domains engage with the internal 10 bp transposase binding sites, that TNP acts to pair the two different P element ends in a manner reminiscent of the 12–23 rule imposed by the RAG1–RAG2 V(D)J recombinase[49–51]. The atypically long staggered cleavage and the arrangement of the dDNAs observed within the STC implies that P element transposition is mechanistically and fundamentally distinct from other cut-and-paste DNA transposases. That is, as transposition proceeds, large structural transitions and rearrangements must occur at the P element transposon ends to generate the distorted DNA conformations observed in the STC structure. Furthermore, GTP is required for pairing of the two P element ends prior to the DNA cleavage[17,18], indicating that GTP plays an additional role(s) at the early stages of transposition. Although the STC structure does not reveal the role of GTP in the initial stages of transposition or how it acts to 'gate' the proposed model for P element end pairing, collectively, these features further underscore the complexity inherent to this class of proteins. Future structural studies of early transposition intermediates should illuminate the mechanistic details involved in orchestrating these conformational changes to perform P element transposition.

Finally, only recently have the functional roles of the numerous repetitive-element derived sequences and genes within large eukaryotic genomes begun to be characterized[53]. For example, the human THAP9 gene encodes a functional TNP homolog that can mobilize *Drosophila* P element DNA in both *Drosophila* and human cells[3]. However, the natural DNA substrates and cellular functions of these TNP homologs are currently unknown. Our data provide a structural framework for understanding all future biochemical studies, not only of *Drosophila* TNP, but also of the related vertebrate TNP THAP9 homologs with as yet unidentified cellular functions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41594-019-0319-6.

Received: 16 April 2019; Accepted: 11 September 2019;
Published online: 28 October 2019

References

1. Kidwell, M. G. Horizontal transfer of P-elements and other short inverted repeat transposons. *Genetica* **86**, 275–286 (1992).
2. Engels, W. R. P elements in *Drosophila*. *Curr. Top. Microbiol. Immunol.* **204**, 103–123 (1996).
3. MajumdarS. & RioD. C. P transposable elements in *Drosophila* and other eukaryotic organisms. *Microbiol. Spectr.* **3**, MDNA3-0004-2014 (2015).
4. Sekelsky, J. DNA repair in *Drosophila*: mutagens, models and missing genes. *Genetics* **205**, 471–490 (2017).
5. Khurana, J. S. et al. Adaptation to P element transposon invasion in *Drosophila melanogaster*. *Cell* **147**, 1551–1563 (2011).
6. Teixeira, F. K. et al. piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. *Nature* **552**, 268–272 (2017).
7. Laski, F. A., Rio, D. C. & Rubin, G. M. Tissue specificity of *Drosophila* P element transposition is regulated at the level of mRNA splicing. *Cell* **44**, 7–19 (1986).
8. Siebel, C. W., Fresco, L. D. & Rio, D. C. The mechanism of somatic inhibition of *Drosophila* P-element pre-mRNA splicing: multiprotein complexes at an exon pseudo-5′ splice site control U1 snRNP binding. *Genes Dev.* **6**, 1386–1401 (1992).
9. Roussigne, M. et al. The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. *Trends Biochem. Sci.* **28**, 66–69 (2003).
10. Majumdar, S., Singh, A. & Rio, D. C. The human THAP9 gene encodes an active P-element DNA transposase. *Science* **339**, 446–448 (2013).
11. Quesneville, H., Nouaud, D. & Anxolabehere, D. Recurrent recruitment of the THAP DNA-binding domain and molecular domestication of the P-transposable element. *Mol. Biol. Evol.* **22**, 741–746 (2005).
12. Hammer, S. E. 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).
13. O'Hare, K. & Rubin, G. M. Structures of P transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. *Cell* **34**, 25–35 (1983).
14. Mullins, M. C., Rio, D. C. & Rubin, G. M. *cis*-acting DNA sequence requirements for P-element transposition. *Genes Dev.* **3**, 729–738 (1989).
15. Kaufman, P. D., Doll, R. F. & Rio, D. C. *Drosophila* P element transposase recognizes internal P element DNA sequences. *Cell* **59**, 359–371 (1989).
16. Rio, D. C., Laski, F. A. & Rubin, G. M. Identification and immunochemical analysis of biologically active *Drosophila* P element transposase. *Cell* **44**, 21–32 (1986).
17. Tang, M., Cecconi, C., Kim, H., Bustamante, C. & Rio, D. C. Guanosine triphosphate acts as a cofactor to promote assembly of initial P-element transposase–DNA synaptic complexes. *Genes Dev.* **19**, 1422–1425 (2005).
18. Tang, M., Cecconi, C., Bustamante, C. & Rio, D. C. Analysis of P element transposase protein-DNA interactions during the early stages of transposition. *J. Biol. Chem.* **282**, 29002–29012 (2007).
19. Beall, E. L. & Rio, D. C. *Drosophila* P-element transposase is a novel site-specific endonuclease. *Genes Dev.* **11**, 2137–2151 (1997).
20. Linheiro, R. S. & Bergman, C. M. Testing the palindromic target site model for DNA transposon insertion using the *Drosophila melanogaster* P-element. *Nucleic Acids Res.* **36**, 6199–6208 (2008).
21. Kaufman, P. D. & Rio, D. C. P element transposition in vitro proceeds by a cut-and-paste mechanism and uses GTP as a cofactor. *Cell* **69**, 27–39 (1992).
22. Roiha, H., Rubin, G. M. & O'Hare, K. P element insertions and rearrangements at the singed locus of *Drosophila melanogaster*. *Genetics* **119**, 75–83 (1988).
23. Hawley, R. S. et al. Molecular analysis of an unstable P element insertion at the singed locus of *Drosophila melanogaster*: evidence for intracistronic transposition of a P element. *Genetics* **119**, 85–94 (1988).
24. Yin, Z., Lapkouski, M., Yang, W. & Craigie, R. Assembly of prototype foamy virus strand transfer complexes on product DNA bypassing catalysis of integration. *Protein Sci.* **21**, 1849–1857 (2012).
25. Yin, Z. et al. Crystal structure of the Rous sarcoma virus intasome. *Nature* **530**, 362–366 (2016).
26. Ballandras-Colas, A. et al. A supramolecular assembly mediates lentiviral DNA integration. *Science* **355**, 93–95 (2017).
27. Passos, D. O. et al. Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. *Science* **355**, 89–92 (2017).
28. Chow, S. A., Vincent, K. A., Ellison, V. & Brown, P. O. Reversal of integration and DNA splicing mediated by integrase of human-immunodeficiency-virus. *Science* **255**, 723–726 (1992).
29. Melek, M. & Gellert, M. RAG1/2-mediated resolution of transposition intermediates: two pathways and possible consequences. *Cell* **101**, 625–633 (2000).
30. Au, T. K., Pathania, S. & Harshey, R. M. True reversal of Mu integration. *EMBO J.* **23**, 3408–3420 (2004).
31. Polard, P. et al. IS911-mediated transpositional recombination in vitro. *J. Mol. Biol.* **264**, 68–81 (1996).
32. Jonsson, C. B., Donzella, G. A. & Roth, M. J. Characterization of the forward and reverse integration reactions of the Moloney murine leukemia virus integrase protein purified from *Escherichia coli*. *J. Biol. Chem.* **268**, 1462–1469 (1993).
33. Beall, E. L. & Rio, D. C. Transposase makes critical contacts with, and is stimulated by, single-stranded DNA at the P element termini in vitro. *EMBO J.* **17**, 2122–2136 (1998).
34. Donzella, G. A., Jonsson, C. B. & Roth, M. J. Coordinated disintegration reactions mediated by Moloney murine leukemia virus integrase. *J. Virol.* **70**, 3909–3921 (1996).
35. Roussigne, M., Cayrol, C., Clouaire, T., Amalric, F. & Girard, J.-P. THAP1 is a nuclear proapoptotic factor that links prostate-apoptosis-response-4 (Par-4) to PML nuclear bodies. *Oncogene* **22**, 2432–2442 (2003).
36. Sabogal, A., Lyubimov, A. Y., Corn, J. E., Berger, J. M. & Rio, D. C. THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves. *Nat. Struct. Mol. Biol.* **17**, 117–U145 (2010).
37. Lee, C. C., Mul, Y. M. & Rio, D. C. The *Drosophila* P-element KP repressor protein dimerizes and interacts with multiple sites on P-element DNA. *Mol. Cell. Biol.* **16**, 5616–5622 (1996).

38. Lee, C. C., Beall, E. L. & Rio, D. C. DNA binding by the KP repressor protein inhibits P-element transposase activity in vitro. *EMBO J.* **17**, 4166–4174 (1998).
39. Dunker, A. K. et al. Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59 (2001).
40. Montaño, S. P., Pigli, Y. Z. & Rice, P. A. The Mu transpososome structure sheds light on DDE recombinase evolution. *Nature* **491**, 413–417 (2012).
41. MorrisE. R., GreyH., McKenzieG., JonesA. C. & RichardsonJ. M. A bend, flip and trap mechanism for transposon integration. *eLife* **5**, e15537 (2016).
42. Maertens, G. N., Hare, S. & Cherepanov, P. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* **468**, 326–329 (2010).
43. Hickman, A. B., Chandler, M. & Dyda, F. Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. *Crit. Rev. Biochem. Mol. Biol.* **45**, 50–69 (2010).
44. Yuan, Y.-W. & Wessler, S. R. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl Acad. Sci. USA* **108**, 7884–7889 (2011).
45. Beall, E. L. & Rio, D. C. *Drosophila* IRBP/Ku p70 corresponds to the mutagen-sensitive mus309 gene and is involved in P-element excision in vivo. *Genes Dev.* **10**, 921–933 (1996).
46. Fuller, J. R. & Rice, P. A. Target DNA bending by the Mu transpososome promotes careful transposition and prevents its reversal. *eLife* **6**, 257 (2017).
47. Wright, A. V. et al. Structures of the CRISPR genome integration complex. *Science* **357**, 1113–1118 (2017).
48. Rodgers, K. K. Riches in RAGs: revealing the V(D)J recombinase through high-resolution structures. *Trends Biochem. Sci.* **42**, 72–84 (2017).
49. Lapkouski, M., Chuenchor, W., Kim, M.-S., Gellert, M. & Yang, W. Assembly pathway and characterization of the RAG1/2-DNA paired and signal-end complexes. *J. Biol. Chem.* **290**, 14618–14625 (2015).
50. Kim, M.-S., Lapkouski, M., Yang, W. & Gellert, M. Crystal structure of the V(D)J recombinase RAG1–RAG2. *Nature* **518**, 507–511 (2015).
51. Ru, H. et al. Molecular mechanism of V(D)J recombination from synaptic RAG1–RAG2 complex structures. *Cell* **163**, 1138–1152 (2015).
52. Hickman, A. B. et al. Structural basis of hAT transposon end recognition by Hermes, an octameric DNA transposase from *Musca domestica*. *Cell* **158**, 353–367 (2014).
53. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).

Acknowledgements

We thank the Rio Lab members for help and advice. We are grateful to P. Grob, E. Montabana and D. Toso for help with cryo-EM data acquisition and for general microscope maintenance. We thank A. Chintangal for computational support. We are grateful to A. Ban and A. Zanghellini (Arzeda Corporation) for the gift of the codon-optimized P element gene. We thank F. Dimaio and O. Sobolev for advice on modeling with RosettaES and PHENIX, respectively. We thank J. Berger (JHUMS) for examining our DNA and protein modeling and for advice. We thank K. Collins, J. Berger, T.H.G. Nguyen and Y. Lee for critical reading of the manuscript. Work in the Rio Lab was supported by NIH grant R35GM118121. E.H.K. was supported by NIH grant no. K99GM124463. E.N. is an Investigator of the Howard Hughes Medical Institute.

Author contributions

D.C.R. and E.N. supervised the study. G.G. developed the transposase expression conditions and purification procedure, performed in vitro biochemical assays and in vivo assays and prepared complexes for imaging. E.H.K. performed negative-stain and cryo-EM specimen preparation, data collection and data processing. E.H.K. interpreted the protein density with feedback from E.N., G.G. and D.C.R. G.G. and E.H.K. built the DNA model into the map. E.H.K. built the protein model into the map and refined the structure. G.G. and D.C.R. wrote the initial manuscript. All authors contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41594-019-0319-6.

Correspondence and requests for materials should be addressed to E.H.K. or D.C.R.

Peer review information Beth Moorefield was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Cell lines. *Spodoptera frugiperda* (Sf9) cells were obtained from the UC Berkeley Tissue Culture facility and the *Drosophila* Schneider 2 (S2) cells were long-term Rio Lab stock. None of the cell lines used were authenticated. Sf9 cells tested negative for mycoplasma contamination. The S2 cell line was not tested for mycoplasma contamination.

Protein expression. To achieve high level expression and purification of TNP for structural determination, we generated complete *Drosophila* codon-optimized baculovirus expression constructs with two tandem N-terminal solubility tags. *Drosophila* codon-optimized His_8-MBP-TEV protease cleavage site TNP was provided by Arzeda. *Drosophila* codon-optimized SUMO* sequence was ordered as a geneblock from Integrated DNA Technologies and cloned in place of the TEV protease cleavage site to generate His8-MBP-SUMO* (HMS*) TNP. The 5′ untranslated region was replaced with a lobster tropomyosin cDNA leader sequence[54] by PCR, and the resulting fragment was cloned into pFastBacDual expression vector (Invitrogen), downstream of the polyhedron promoter. The expression vectors were used to make recombinant baculoviruses based on the protocol established in the Bac-to-Bac Baculovirus Expression System (Invitrogen) using EmBacY cells[55]. A 10 ml volume of high titer baculovirus stock was used to infect 1 l of *S. frugiperda* (Sf9) cells at a density of 1.0×10^6 cells ml^{-1}. Cells were cultured in paddle flasks in TNM-FH/10% FBS/1× penicillin/streptomycin (Gibco). Infected cells were incubated for 72 h (27 °C) before harvesting by centrifugation. Harvested cell pellets were washed with PBS and snap-frozen in liquid nitrogen for later purification.

Protein purification. Cell pellets were thawed on ice, disrupted in 35 ml lysis buffer (25 mM HEPES-KOH pH 7.6, 400 mM KCl, 400 mM $(NH_4)_2SO_4$, 50 mM NaF, 1 mM EDTA, 0.01% NP-40, 1 mM DTT, 1 mM PMSF, 1× protease inhibitor cocktail), briefly sonicated, then clarified by centrifugation. Polyethylenimine was added to the supernatants dropwise to a final concentration of 0.1%, incubated for 10 min on ice with stirring, then ultracentrifuged at 160,000*g* for 30 min. Supernatants were supplemented with solid L-arginine HCl (final concentration of 140 mM), then filtered through a 0.22 µm syringe filter before application to 5 ml of pre-equilibrated dextrin Sepharose resin (GE Healthcare) using a peristaltic pump for 2 h. The resin was washed three times with 10 column volumes (CVs) of wash buffer (25 mM HEPES-KOH pH 7.6, 400 mM KCl, 500 mM L-arginine HCl, 1 mM EDTA, 0.01% NP-40, 1 mM DTT, 1 mM PMSF). Protein was eluted in batch three times with one CV elution buffer (wash buffer + 10% glycerol, 50 mM maltose). The eluted protein was dialyzed overnight into low-salt buffer (25 mM HEPES-KOH pH 7.6, 100 mM $(NH_4)_2SO_4$, 1 mM EDTA, 0.01% NP-40, 10% glycerol, 1 mM DTT, 1 mM PMSF), then loaded onto a 5 ml HiTrap heparin HP column (GE Healthcare) pre-equilibrated in heparin buffer (25 mM HEPES-KOH pH 7.6, 100 mM $(NH_4)_2SO_4$, 5 mM $MgCl_2$, 0.01% NP-40, 10% glycerol, 1 mM DTT, 1 mM PMSF) and eluted with a linear gradient of 100 mM to 1,000 mM $(NH_4)_2SO_4$ over five CVs. Peak fractions were concentrated to 24 µM to 72 µM using a Spin-X UF 20 10k MWCO (Corning), and stored on ice until complex formation.

DNA preparation. DNA oligonucleotides were purchased from Integrated DNA Technologies or synthesized in house on a 392 DNA and RNA synthesizer (Applied Biosystems), and were purified using denaturing PAGE (urea–PAGE). DNA substrates were prepared by mixing the appropriate ssDNA oligonucleotides in 20 mM HEPES-KOH, pH 7.6, 25 mM KCl, 10 mM $MgCl_2$, incubating at 95 °C for 5 min and slow-cooling to room temperature. Radiolabeled substrates were prepared by labeling with T4 polynucleotide kinase (USB) and [γ-^{32}P]-ATP (Perkin Elmer) and annealing with a slight excess of the unlabeled strands. The DNA substrates used in this study are listed in Supplementary Table 1.

Strand transfer complex assembly. For assembly of the STC, a mixture containing 24 µM HMS* TNP, 12.6 µM strand transfer product DNA, 6 µM SUMOstar protease (LifeSensors) and 2 mM GTP was dialyzed against low-salt buffer (25 mM HEPES-KOH pH 7.6, 100 mM KCl, 10 mM Mg $(OAc)_2$, 10 µM $ZnSO_4$, 0.5% zwittergent 3–08, 0.5 mM tris(2-carboxyethyl)phosphine (TCEP)) at 4 °C overnight. After dialysis, a white precipitate was observed that could not be solubilized by the addition of salt[25,56]. The mixture was centrifuged to remove precipitates. Soluble TNP DNA complexes were incubated at 25–30 °C for 1 h before purification through SEC (Superose 6 Increase 3.2-30, GE Healthcare) running with SEC buffer (25 mM HEPES-KOH pH 7.6, 100 mM KCl, 10 mM Mg $(OAc)_2$, 10 µM $ZnSO_4$ and 0.5 mM TCEP), before immediately proceeding to cryo-EM sample vitrification.

Disintegration assay. Approximately 9 µg of HMS* TNP (65 pmol monomer) was preincubated with 2 pmol strand transfer product DNA and incubated at room temperature for 20 min in a total volume of 10 µl disintegration buffer (25 mM HEPES-KOH pH 7.6, 5% glycerol, 10 µM $ZnSO_4$, 0.05% zwittergent 3-08, 0.5 mM TCEP). Reactions were initiated by the addition of SUMOstar protease and either 10 mM $MgCl_2$ or $MnCl_2$, then incubated overnight at room temperature. Reactions were terminated by the addition of 10 µl 20× STOP buffer (85 mM EDTA, 5% SDS), then incubated at 37 °C for 2 h with 0.1 mg ml^{-1} proteinase K. A 2 µl sample of each deproteinized reaction product was resolved by electrophoresis on 6% native polyacrylamide gel and visualized by SYBR Gold staining (Thermo Fisher Scientific).

Strand transfer assays. Strand transfer assays with plasmid target were largely performed as previously described[33]. Briefly, 250 ng HMS* TNP (1.8 pmol monomer) was preincubated with 0.4 pmol of radiolabeled minimal pre-cleaved 3′ dDNA for 20 min on ice, in a total volume of 6 µl HGED buffer (25 mM HEPES-KOH pH 7.6, 20% glycerol, 1 mM EDTA, 1 mM EGTA, 0.5 mM DTT, 100 µg ml^{-1} BSA). The reaction was initiated by the addition of 14 µl of 0.35× HGED buffer, 5 mM Mg $(OAc)_2$, 2 mM GTP and 100 ng Bluescript tetrameric target plasmid DNA, then incubated at 30 °C for 2 h. Reactions were terminated by the addition of 1.5 µl of 20× STOP buffer, then incubated at 37 °C for 30 min with 0.1 mg ml^{-1} proteinase K. Reaction products were analyzed by electrophoresis on 0.7% agarose gel, dried and visualized by phosphorimaging. Strand transfer assays in Fig. 3d, were performed as described but with 5 µM of either GTP, ATP, ITP (Jena Bioscience), XTP (TriLink Biotechnologies), 2-aminopurine (TriLink Biotechnologies) or 2-amino-ATP (TriLink Biotechnologies).

Strand transfer assays with 60 bp duplexed targets were performed as follows: ~1.2 µg HMS* TNP (~8.5 pmol monomer) was preincubated with 20 pmol of 5-carboxytetramethylrhodamine (5-TAMRA) labeled minimal pre-cleaved 3′ dDNA for 20 min on ice, in a 20 µl volume of strand transfer assay buffer (25 mM HEPES-KOH pH 7.6, 35 mM KCl, 20% glycerol, 1 mM EDTA, 1.0 mM DTT, 100 µg ml^{-1} BSA, 10 mM Mg $(OAc)_2$, 2 mM GTP). Reactions were initiated by the addition of 5 pmol of tDNAs, then incubated at 30 °C for 2 h. Reactions were terminated by the addition of 1.5 µl of 20× STOP buffer (85 mM EDTA, 5% SDS), then incubated at 37 °C for 30 min with 0.1 mg ml^{-1} proteinase K. A 22 µl volume of deionized formamide and 2 µl 100 mM NaOH were added, boiled for 5 min, then 6 µl of each sample was resolved on a 10% denaturing polyacrylamide gel protected from light. Gels were visualized using a Typhoon imager (GE Healthcare).

In vivo excision assay. Assays were performed in triplicate, essentially as previously described[16,45]. Briefly, 3.0×10^6 *Drosophila* Schneider 2 cells were transfected with 2 µg pISP-2–Km reporter plasmid and either 0.5 µg empty plasmid (pBSKS (+)pAc) or transposase source (pBSKS (+)pAc-TNP), using Effectene transfection reagent (QIAGEN). At 24 h after transfection, cells were washed with PBS, then harvested for immunoblot analysis and plasmid DNA recovery. Plasmid DNA was recovered as previously described[16], resuspended in 10 µl TE buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA), then 1 µl was used to transform RecA$^-$ *Escherichia coli* strain AG1574[21] with a BioRad Gene Pulser as described by the manufacturer. Cells were grown for 1.5 h at 37 °C with shaking, then plated onto Luria broth plates containing either 100 µg ml^{-1} of ampicillin (1 µl of a 1:1,000 dilution) or 100 µg ml^{-1} of ampicillin and 50 µg ml^{-1} of kanamycin (50 µl undiluted cells). Colonies were allowed to develop for 16 h at 37 °C, then counted.

Cryo-electron microscopy sample vitrification and data collection. Samples were vitrified using a Mark IV vitrobot (FEI). A 4 µl volume of concentrated STC complex was applied to a Quantifoil 1.2/1.3 UltraAuFoil grid after being plasma cleaned (Solarus) for 10 s in air. After 30 s incubation, the sample was blotted using a blot force of 8 pN and a blot time of 6 s. Images were collected on an Arctica scope (Thermo Fisher) using a K2 detector (Gatan) using SerialEM[57]. During data collection, the stage was tilted by 40° to circumvent preferential orientation[58]. A total of 1,857 micrographs were collected during a three-day period with a nominal defocus range of −1 to −3 µm. Dose-fractionated movies were collected with a total dose of 60 electrons and 10 s per movie. Please see Table 1 for additional details.

Image processing. After motion correction with MotionCor2[59] and particle-picking using Gautomatch, an initial per-micrograph contrast transfer function (CTF) estimation and a subsequent per-particle CTF estimation were carried out using GCtf[60]. Ab initio model generation using cryoSPARC[61] with three classes resulted in one highly populated class (60% of particles) and two 'junk' classes. The selected particles (253,209) were exported to RELION-3.0[62] and an initial refinement in an ~4 Å reconstruction. Subsequent rounds of automatic refinement, followed by per-particle CTF refinement and Bayesian polishing, were iterated until convergence (Supplementary Fig. 2c) and resulted in the final 3.6 Å reconstruction. The reconstruction has a relatively uniform resolution, with the highest resolution in the core of the complex estimated to be 3.3 Å (Supplementary Fig. 2g). The alignment parameters from this final C2 reconstruction were then refined without imposing symmetry (C1) resulting in an overall 3.9 Å structure (masked half-map), which matches the phase-randomized FSC estimate (Supplementary Fig. 2f).

De novo model building. An initial Cα trace and the initial sequence register were built manually using Coot[63]. Subsequent rounds of refinement using RosettaES[64] filled in loops and rebuilt regions that were incorrect. The model for the nucleic acid was generated using Coot and refined with PHENIX[65]. The model for GTP was taken from the highest resolution available structure containing GTP (PDB ID 4GMU, 1.2 Å resolution). A rigid body fit, followed by rotation around the α-phosphate group, resulted in the modeled ligand. Geometry minimization was

performed using PHENIX with constraints on the starting coordinates to improve model ideality. The r.m.s.d. difference between input and minimized atomic models is ~0.1 Å r.m.s.d. The calculated final model-map FSC (0.5 cutoff) was 3.7 Å.

Map and model visualization. Maps were visualized in Chimera[66] and all model illustrations were prepared using either Chimera or ChimeraX[67].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Atomic models are available through the Protein Data Bank with accessions codes 6P5A (C2) and 6PE2 (C1); cryo-EM reconstructions are available through the EMDB with accession codes EMD-20254 (C2) and EMD-20321 (C1).

References

54. Sano, K.-I., Maeda, K., Oki, M. & Maéda, Y. Enhancement of protein expression in insect cells by a lobster tropomyosin cDNA leader sequence. *FEBS Lett.* **532**, 143–146 (2002).
55. Trowitzsch, S., Bieniossek, C., Nie, Y., Garzoni, F. & Berger, I. New baculovirus expression tools for recombinant protein complex production. *J. Struct. Biol.* **172**, 45–54 (2010).
56. Ballandras-Colas, A. et al. Cryo-EM reveals a novel octameric integrase structure for betaretroviral intasome function. *Nature* **530**, 358–361 (2016).
57. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
58. Tan, Y. Z. et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* **14**, 793–796 (2017).
59. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
60. Zhang, K. Gctf: real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
61. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
62. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, 163 (2018).
63. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
64. Frenz, B., Walls, A. C., Egelman, E. H., Veesler, D. & DiMaio, F. RosettaES: a sampling strategy enabling automated interpretation of difficult cryo-EM maps. *Nat. Methods* **14**, 797–800 (2017).
65. Adams, P. D. et al. The Phenix software for automated determination of macromolecular structures. *Methods* **55**, 94–106 (2011).
66. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
67. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).

nature research

Corresponding author(s): Elizabeth Kellogg and Donald Rio

Last updated by author(s): Aug 15, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SerialEM version 3.6 was used to collect images.
Data analysis	Motioncorr version 2 was used to align raw movies, Gautomatch version 0.56 was used to pick particles, and GCTF version 1.0 was used to estimate micrograph CTF. CryoSPARC version 1 was used for ab-initio cryo-EM reconstruction and RELION-3.0 (beta version) was used for subsequent refinement to high-resolution. Coot version 0.8.9.1 was used for manual model building, RosettaES (version 1) was used for rebuilding loop regions and PHENIX version 1.14 was used for final refinement of the protein-nucleic acid complex. PHENIX was used to compute model-map and half-map FSC as well as model statistics. Gel images were analyzed using ImageJ v1.51.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We have included a data availability statement in the manuscript. Atomic models are available through the Protein Data Bank (PDB) with accessions codes 6P5A (C2) and 6PE2 (C1); cryo-EM reconstructions are available through the EMDB with accession codes EMD-20254 (C2) and EMD-20321 (C1).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size here is the number of particle images used in the final reconstruction. Because smaller datasets resulted in lower resolution (4 Å) cryo-EM reconstructions and were improved by imposing symmetry (indicated that more data was needed), we then collected a larger dataset containing double the number of particles. This larger dataset was sufficient to achieve the resolution we needed for model building (3.6 Å) as assessed by the half-map FSC and contained features appropriate for a map of this resolution.
Data exclusions	41% of the initial particle set was excluded from 3D refinement. These particles were assessed to be artifacts resulting from the particle picking program. During ab-initio model generation we sorted particles into one of three 3D classes. Only one resembled the particles in the micrographs, the other two contained no structural features. Mapping these particle picks back onto the aligned micrographs confirmed that these discarded particles were false positives from the particle picking algorithm.
Replication	Biochemical purification and activity assays are all replicated successfully.
Randomization	Randomization was not relevant to this study.
Blinding	Blinding was not relevant to this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Rabbit anti-P element transposase (Rio Lab), rabbit anti-HRP48 (Rio Lab), Anti-rabbit (Bio-Rad Laboratories Inc., Cat. #170-6515, Lot 350003248)
Validation	Rabbit anti-P element transposase and anti-HRP48 serum were raised and antigen affinity purified in the Rio Lab, and validated by Western blots of transfected and non-transfected cell lysates or against recombinant protein.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Sf9 cells were obtained from UC Berkeley Tissue Culture facility and the S2 cells were long-term Rio Lab stock.
Authentication	None of the cell lines used were authenticated.
Mycoplasma contamination	Sf9 cells tested negative for mycoplasma contamination. The S2 cell line was not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study.