Structural basis of seamless excision and specific targeting by piggyBac transposase

Qiujia Chen1, Wentian Luo2,3, Ruth Ann Veach2,3, Alison B. Hickman1, Matthew H. Wilson2,3,4 & Fred Dyda1

The piggyBac DNA transposon is used widely in genome engineering applications. Unlike other transposons, its excision site can be precisely repaired without leaving footprints and it integrates specifically at TTAA tetranucleotides. We present cryo-EM structures of piggyBac transpososomes: a synaptic complex with hairpin DNA intermediates and a strand transfer complex capturing the integration step. The results show that the excised TTAA hairpin intermediate and the TTAA target adopt essentially identical conformations, providing a mechanistic link connecting the two unique properties of piggyBac. The transposase forms an asymmetric dimer in which the two central domains synapse the ends while two C-terminal domains form a separate dimer that contacts only one transposon end. In the strand transfer structure, target DNA is severely bent and the TTAA target is unpaired. In-cell data suggest that asymmetry promotes synaptic complex formation, and modifying ends with additional transposase binding sites stimulates activity.
Transposons are mobile genetic elements that can move from one position to another in the genome or between host genome and foreign DNA. They make up a large percentage of most eukaryotic genomes, where they have played important roles in genome evolution and the establishment of novel cellular functions and pathways. Although their movement has been severely constrained in humans, largely through inactivation, transposition has been linked to the development of specific diseases.

DNA transposons, most notably Sleeping Beauty (SB) and piggyBac (pB), have been intensively exploited as tools for genome engineering and therapeutic applications. Among eukaryotic DNA transposons, only pB is known to specifically integrate at TTAA sites and to exhibit the property of seamless excision whereby the genomic gap produced when the transposon is excised can be repaired precisely, without need for any DNA repair. pB transposases are mobile genetic elements that can move from one position to another in the genome or between segments relative to LE35 (Fig. 1b, in white and green, respectively), a puzzling arrangement. pB transposase called hyPBase and one mutant variant have been developed through random mutation including CAR T-cell engineering, CRISPR/Cas-mediated gene therapy, and human induced pluripotent stem cells (iPSC) engineering. Useful variants have been developed through random mutation including a hyperactive pB transposase called hyPBase as well as one that can excise pB but cannot integrate it.

Although pB (Fig. 1b) was isolated from the cabbage looper moth Trichoplusia ni over 30 years ago, no structural information is available that would explain its excision and targeting properties or provide a rational basis to understand how mutations affect its activities. pB contains a single open-reading frame (ORF) encoding its 594 amino-acid transposase (PB) (Fig. 1c) flanked by terminal inverted repeats (TIRs) (Fig. 1b). The pB TIRs are asymmetric; sufficient for certain in vitro and in vivo activities are the 35-nt left end TIR (LE TIR or LE35) and 63-nt right end TIR (RE TIR or RE63). Although LE35 and RE63 have short sequences and repeats in common (shown in blue, green, and purple in Fig. 1b), the relative organization of these differs and RE63 has a 17-nt insertion and a sequence duplication between segments relative to LE35 (Fig. 1c).

**Fig. 1 Overview of piggyBac transposition.** a Mechanism of piggyBac (pB) transposition. Hydrolisis by the piggyBac transposase (PB) liberates the 3′-OH on the DNA strand that will be integrated. This is followed by transesterification in which this 3′-OH attacks flanking DNA four nt from the transposon end forming a DNA hairpin. For subsequent integration, PB opens the hairpin at each end in another hydrolytic step, leaving a four nt TTAA overhang; this is followed by a second transesterification step that joins each end to target DNA. At the empty donor site (inset), complementary DNA strands allow for seamless repair. LE left end, RE right end, TIR terminal inverted repeat, TSD target site duplication.

b Schematic of pB transposon flanked by TTAA, and sequence and organization of the LE and RE TIRs. NTS non-transferred strand, TS transferred strand. c Domain organization of PB. The catalytic domain contains the conserved DDD motif (in red). NTD N-terminal domain. Dimerization and DNA-binding domain (DDBD). CRD C-terminal cysteine-rich domain. Gray indicates disordered regions in the structure. d Overall structures of the synaptic hairpin DNA (SNHP) complex and strand transfer complex (STC).
PB is a cut-and-paste DNA transposase, proposed to be a member of the RNaseH-like or retroviral integrase superfamily. PB-catalyzed transposition proceeds through a series of hydrolysis and transesterification steps that generate an excised intermediate in which DNA hairpins protect the transposon ends (Fig. 1a). The reaction sequence differs from those of other eukaryotic DNA transposases, such as members of the Tcl1/ mariner and hAT superfamilies, or the Transib precursor to RAG1 of the V(D)J recombination system, which is identical to the prokaryotic IS4 family of insertion sequences and transposons such as Tn5. Despite this shared reaction pathway, however, only PB precisely targets TTAA sequences.

Motivated by PB’s unique properties and its importance in established and emerging applications, we have determined two ab initio single-particle cryo-electron microscopy (cryo-EM) structures of PB in complex with DNA: a synaptic complex that contains the hairpinned DNA intermediates (synaptic hairpin complex (SNHP)) and a strand transfer complex (STC) in which two TIRs have been integrated into target DNA. The structures explain the basis of seamless excision and targeting, and together with biochemical and in-cell transposition data, suggest a transposition model that requires asymmetric TIRs in order to form the active synaptic complex in vivo.

Results

Mammalian-expressed PB is active in vitro. PB expressed and purified from mammalian cells forms dimers as judged by SEC-MALS (Supplementary Fig. 1a). In vitro, as previously demonstrated for PB expressed in *E. coli*, it catalyzed hairpin formation and resolution with oligonucleotide TIR substrates (Supplementary Fig. 1b). It also catalyzed double-ended integration of pre-cleaved transposon ends into a supercoiled (SC) pUC19 target plasmid (Supplementary Fig. 1c). For example, incubation with a pre-cleaved 35 bp LE TIR oligonucleotide (LE35) and pUC19 generated products consistent with single-end (SE) and double-end (DE) integration (Supplementary Fig. 1c). Under assay conditions with a protein:TIR ratio of 1:2, plasmid integration was less efficient when a pre-cleaved 63 bp RE TIR (RE63) was used, or when LE35 and RE63 were mixed at equal molar ratio (Supplementary Fig. 1c).

Present in both LE35 and RE63 is a palindromic-like 19-bp internal repeat (the purple box sequence in Fig. 1b) that is the binding site for the C-terminal cysteine-rich domain (CRD) of PB. The CRD is required for PB activity and has been proposed to be the driver of TIR binding. When we probed the effect of sequentially truncating RE63 to remove the CRD-binding site (Supplementary Fig. 1d), we were surprised to observe increasing stimulation of integration into either linearized or SC pUC19 as the RE was shortened to RE24. Further deletion to RE16 largely abolished integration.

To confirm that mammalian-expressed PB integrated with its expected target site specificity, we generated a linear mini-transposon in which LE35 and RE63 TIRs flanked a Kan resistance gene, and used this as a substrate for in vitro integration into SC pUC19. The reaction products were purified and transformed into *E. coli*, allowing us to select for Amp+Kan+ colonies corresponding to integrated mini-transposons. Sequencing confirmed that only TTAA tetranucleotides had been targeted: from 10 sequenced colonies, we detected four mini-transposon integration events corresponding to insertion at TTAA sequences at bp 635–638, bp 1568–1571, bp 1582–1585, and bp 2646–2649 of pUC19.

Cryo-EM structure determination. For structural studies, we investigated the biophysical properties of a variety of PB/TIR complexes. In most cases, analytical size-exclusion chromatography (SEC) revealed monodisperse elution profiles with overlapping peaks. However, when PB was complexed with LE35 hairpin intermediates (Supplementary Fig. 2a) in the presence of Ca2+ (which does not catalyze hairpin opening, Supplementary Fig. 1b), it was possible to isolate a monodisperse population containing SNHP (Supplementary Fig. 2b). We also assembled a STC that contained PB, Ca2+, and 37 bp LE TIRs covalently joined to target DNA (Supplementary Fig. 3). As DNase I footprinting data suggested that PB interacts with 4–5 bp of flanking DNA beyond its TTAA target, target DNA included 11 bp on either side of the central TTAA (Supplementary Fig. 3a). To avoid base-pairing during substrate annealing between the target sequence and the four nt TTAA overhang derived from hairpin opening (Fig. 1a), we substituted the overhang with the mutated sequence CCGG (Supplementary Fig. 3a). It has been shown that the four nt overhang is not required for TTAA target site selection in vitro, nor do mutated flanking sequences affect specific integration into TTAA sites in vivo. Our attempts to prepare stable monodisperse complexes containing one LE TIR and one RE TIR have not yet been successful.

We collected single-particle cryo-EM data on both complexes. For each, particle projections showed a wide distribution of different orientations in the micrographs and a number of distinct two-dimensional (2D) classes (Supplementary Fig. 4a, b). For the SNHP complex, we applied a mask around the CRD for masked three-dimensional (3D) classification which improved the resolution of the reconstructed density map to 3.66 Å (Supplementary Fig. 4a, c). For the STC, we used two rounds of 3D classification and one round of masked 3D classification, resulting in a 3.47 Å density map (Supplementary Fig. 4b, c) with better map quality when compared with the SNHP complex, most likely due to the additional stabilizing effect of the target DNA. While parts of the maps were clearly two-fold symmetric, the maps as a whole had no rotational symmetry (Supplementary Fig. 4a, b). Therefore, all processing was done without applying symmetry.

An NMR structure has been determined for the CRD (PDB 5LME) but as there was no known homologous structure available for the rest of PB, ab initio atomic models were built into Phenix-sharpened potential density maps. For the SNHP complex, we used Phenix’s map to model module and fragments were connected manually. The NMR model of the CRD dimer was consistent with the potential density in the region and was used as a starting model. We used density-guided rebuilding tools of Rosetta to complete and verify the trace and register of the model that was subsequently refined using Rosetta and Phenix. The STC model was built based on the SNHP model. Representative regions of the potential density maps and model fits are shown in Supplementary Figs. 5 and 6. The final models contain residues 117–594, consistent with the prediction by the PONDR server that the N-terminal 110 residues are largely intrinsically disordered.

Overall structures of PB transposase complexes. In both the SNHP and STC complexes, an asymmetric PB dimer synapses two approximately parallel TIRs (Fig. 1d). As predicted, the catalytic domain (residues 263–457, separated by an insertion domain from residues 372 to 433) possesses the fold of the RNaseH-like superfamily of transposases. After the fifth β-strand of the catalytic domain, a three β-stranded insertion domain interrupts the RNaseH-like fold (Fig. 2a and Supplementary Fig. 7), the same topological location as in all other DDE/D Exit transposases featuring insertion domains. The convergence of residues 117–263 and 457–535 forms a structurally unique

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-020-17128-1 | www.nature.com/naturecommunications
all-α-helical domain (denoted Dimerization and DNA-binding domain, DDBD) knitting the protein together and which interacts with TIR bp 7–16 (Fig. 2b). The C-terminal end of the DDBD is connected to the CRD through an extended linker that exhibits weak density. In both complexes, the DDBD, catalytic and insertion domains collaborate to synapse two LE TIRs and direct the scissile phosphates to the active sites comprised of D268, D346, and D447 where a Ca$^{2+}$ ion is bound by the carboxylates of D268 and D346 (Supplementary Fig. 6a, b). A bridging oxygen from the scissile phosphate also forms part of the Ca$^{2+}$ ligand sphere, suggesting that PB follows the paradigm of the two metal-ion-dependent chemical mechanism$^{51,52}$. In the SNHP complex, the horseshoe-shaped dimer formed by the DDBD, catalytic and insertion domains creates a wide channel whose sides are lined with β-hairpin loops largely contributed by the insertion domain (Fig. 1d). In the STC, this channel is filled with the target DNA (Fig. 1d). The SNHP and STC complexes are very similar and can be superimposed at 1.22 Å rmsd over 933 α-carbon positions. Upon target DNA binding, the channel formed by the catalytic and insertion domains narrows by about 3 Å as the two halves of the transpososome close down upon it, although we cannot rule out the possibility that this observation was influenced by the selection of the 3D classes.

**CRD binding generates asymmetry and bends one TIR.** The DDBDs, catalytic and insertion domains, and the first 16 bp of the LE TIRs bound by them obey two-fold symmetry (Fig. 1d and Supplementary Fig. 4) (extending also to the target DNA in the STC), yet overall the complexes are markedly asymmetric as the two CRDs form a dimer (Fig. 3) that binds bp 17–33 of only one of the LE TIRs (corresponding to the 19-bp palindromic internal repeat, in purple in Fig. 1b). There is no detectable density around the equivalent palindromic internal repeat sequence of the other LE TIR and, despite the weaker DNA density (presumably due to the lack of the stabilizing effect of bound CRDs), it is clear that it is not bent (Supplementary Fig. 4). The CRD monomers refined against the cryo-EM maps are consistent with the cross-brace Zn finger structure determined by NMR$^{43}$, although not identical due to side chain rearrangements at the hydrophobic CRD dimer interface (Fig. 3c).
CRD binding causes a ~40° bend in the TIR due to the insertion of the CRDs into two adjacent widened major grooves, accompanied by a significant narrowing of the intervening minor groove centered around bp 25 (Fig. 3a, b). The purple box of LE TIR DNA contains a canonical A-tract (bp 22–25) that appears to display propeller twist and has a narrowed minor groove, features previously noted for A-tracts. The CRD dimer is approximately two-fold symmetric with a rotation axis perpendicular to the DNA and centered at the minor groove in the middle of the CRD-binding site (Fig. 3b). At the level of the CRD monomer, the observed mode of DNA binding is broadly consistent with the model derived from NMR including the identification of Y558, R567, and K569 as crucial residues for specific recognition, but the NMR model did not predict the severe DNA bend induced by CRD dimerization.

TIR recognition close to the transposon end. Although the DDBDs interact with both TIRs, most of the protein/DNA interactions are in trans, such that the interactions of one monomer with a TIR direct its end into the active site of the other monomer (Fig. 2b). Such trans arrangements are common in transpososome assemblies. The parallel TIR configuration is in sharp contrast with the arrangement within the Tn5 transpososome (Fig. 2c) that catalyzes the same reaction sequence. Remarkably, the two transpososomes are organized in completely different ways, reflecting two very different modes of transposase dimerization, with the Tn5 TIRs almost antiparallel and essentially straight (Fig. 2c). However, the two transposases do share the feature of an entirely β-stranded insertion domain (Supplementary Fig. 7). While the DD(E/D) catalytic and parts of the insertion domains of PB and the Tn5 transposase can be aligned (1.99 Å rmsd over 182 α-carbons), the sequence identity in this region is only 9%.

In both PB complexes, the catalytic domain interacts with the minor groove closest to the transposon end (bp 2–6) through a rich set of interactions, including two α-helices linked by G444 and G445 (α9–α10 in Supplementary Fig. 7) that turn and follow the curve of the minor groove. The α10 helix carries D447 that together with D268 and D346 form the catalytic triad (Supplementary Fig. 6a, b). One turn away from the tip of the transposon, the DDBD binds the TIR in trans, widening the major groove (bp 9–14) through the insertion of three different parts of the protein (i.e., α4, α5, and a loop between α2 and α3) (Fig. 2b). In addition, the two adjacent, narrowed minor grooves (bp 5–8 and bp 15–18) are also recognized by elements of the DDBD (Fig. 2b). There is also a curious arrangement in which three carboxylates (D197, D215, and D218) coordinate a divalent metal ion (Ca2+) that is also coordinated by several bases in the widened major groove (Fig. 2b).

Compared to the PB transposase, that of Tn5 has additional β-strands, and β7 and β8 of Tn5 form a β-hairpin that lies deep in the major groove just adjacent to the DNA hairpin (Fig. 2a and Supplementary Fig. 7). For Tn5, mutations in this β-hairpin affect all steps of transposition. The same role in PB is served by an omega loop between the first and the second β-strands of the catalytic domain that interacts with the hairpin and hairpin-proximal major groove through R275, Y283, and K290 in trans (Fig. 4b).

SNHP complex and DNA hairpin recognition. In the SNHP complex, all the nts of the transposon end are base-paired whereas none of the four nts of the hairpin loop (numbered as A-1, A-2, T-3, T-4), derived from the flanking TSD DNA, are base-paired (Supplementary Fig. 5c). Multiple elements contribute to the stabilization of the hairpin. Backbone phosphates of the hairpin loop interact with Y439 from the catalytic domain in cis, and Y283 from the omega loop in trans (Figs. 4b and 5). There is also a base-specific interaction between N440 and A-2 in cis (Figs. 4b and 5). In addition, the methyl group of T-3 is in a hydrophobic pocket formed by V414/L416/Y439 of the insertion domain in cis.

Although A-1 is completely flipped out, there do not appear to be aromatic stacking interactions with PB side chains. This is in contrast to the hairpin loop of Tn5 in which the collaboration of two tryptophan residues is needed to form the tight hairpin on each transposon end, one pushing out a thymine base from the non-transferred strand (NTS) and preventing its return to the helix and another stacking against it once it is flipped out. This difference is likely because in Tn5, the tightest possible hairpin is formed by the attack of the transferred strand (TS) terminal 3'-OH on the terminal nt of the NTS directly opposite. In PB, the TS terminal 3'-OH instead attacks flanking DNA, four nt from the transposon end of the opposite strand (Fig. 1a), and the structure suggests that this is long enough that a hairpin can readily form. The active site area of PB is more open when compared to that of Tn5, consistent with the ability to accommodate a longer hairpin loop and to allow the
conformation change of the NTS that might be required to bring the scissile phosphate to the active site. The 4 nt long hairpin loop is stabilized by a set of interactions with the omega loop in trans, which are absent in Tn5.

**STC complex reveals target DNA recognition and integration.** In the STC, all nts of the transposon DNA are base-paired as are the target nts except for the bps that comprise the specific TTAA tetranucleotide (numbered as T0T, T1T, A2T, A3T), all four of which are unpaired (Supplementary Figs. 5d and 6c). It appears that specificity for the TTAA target sequence arises from its conformation as an ssDNA segment. The two ssDNA segments are stabilized by an elaborate network of protein/DNA and DNA/DNA interactions (Figs. 4c and 5) organized by the omega loop. Remarkably, relative to the hairpin loop seen in the SNHP complex, many of the same protein residues are involved and the conformations of T0T, T1T, and A2T are essentially
superimposable on T-4, T-3, and A-2 of the hairpin DNA, respectively (Fig. 4d). The major change is with A37 (equivalent to A-1 in the hairpin loop) which is not flipped out but, rather, turned back towards the other target bases and H-bonded through its N1 to N3 of T07 (Figs. 4c, d and 5) and stacked against Y283. To allow the reaction to proceed from the hairpin-bound state to one poised to capture target, it appears that hairpin resolution is followed by the movement of the resulting flap out of the active site. Then, a drastically distorted and unpaired target TAA tetranucleotide can be bound. As the structures reveal, the key to PB transposition is that the backbones of the TAA tetranucleotide in both the hairpin and target DNA adopt a very similar conformation (Fig. 4d). The role of the PB protein is therefore to enforce this conformation at both steps of the reaction.

The density is poor for the tetranucleotide flap CCGG after the first two nucleotides, presumably due to flexibility. Minimal interactions with the flap are consistent with observations that the flap is not necessary for PB reactions in vitro35 (Supplementary Fig. 1c, d), and that the exact 4 bp sequence of flanking region (A5T, Fig. 5b). We observe interactions between several amino acid residues and backbone phosphates for five bp beyond the target TAA, and the side chains of R372 and R376 are inserted into the minor groove of the flanking region (A5T–G8T, Fig. 5b).

The STC structure suggests that PB achieves target specificity by stabilizing the strand transfer product through the ssDNA form of TAA. This unusual mode of transposon target selection is in line with the lack of a region of PB recognizable in the
structure as a potential dsDNA TTAA target recognition domain. Although the role of the disordered N-terminal region remains to be established, the predicted isoelectric point of residues 1–117 is 4.7, making it highly unlikely that this region contributes to DNA binding.

PB integration occurs symmetrically, and the configuration of the dsDNA regions of the target that flank TTAA suggests that it must be severely bent before integration (Fig. 4e). Bent target DNA is a common feature of most DNA transpososome and intasome structures40,49,56,60–64. Strikingly, a unique aspect of the PB STC is that integration occurs at staggered target DNA phosphates across the minor groove (Fig. 4e). In all other known transpososome structures, integration has invariably been observed to occur across the major groove, as for example in the PFV intasome that also integrates with a four bp TSD (Fig. 4f, PDB 4E7L40). In PFV, the strand transfer sites (C, Fig. 4f, top) are easily accessed by bending the target DNA to widen the major groove to fulfill the distance between active sites. In the PB STC, however, the strand transfer sites (T, Fig. 4e, top) are opposite the bound PB dimer. Simply bending the target DNA is not sufficient to allow the target scissile phosphodiester bonds to reach the active sites that are located deep within the catalytic domains. Instead, to fit the orientation of the active site DDD motif, coordinated metals, and the 3′-OH of the TS during the strand transfer reaction, TTAA tetrancotides are melted and drastically distorted (Fig. 4e). Interestingly, the omega loop of PB interacts with target DNA, facilitating the melting of TTAA (Fig. 4e, bottom). The equivalent loops in the PFV intasome (Fig. 4f, bottom) and STC structures of other transposases and integrases have no contact with target DNA. Altogether, the unique aspects of PB target DNA recognition and integration are driven by the location of the active sites within the dimer, the interactions of the omega loops, and the relative positions of the target scissile bonds.

**pB transposition in cells.** We used colony count and excision assays in cultured human cells using pB and pB transposon derivatives as a proxy for in vivo transposition65. Excision analysis uses PCR to amplify re-joined transposon plasmid ends recovered from transfected cells indicating transposon excision has occurred. The colony count assay involves excision of a neomycin resistance transposon from transfected cells indicating transposon excision has occurred. However, if the pB transposition in cells

Discussion

The results here reveal that the distinguishing features of pB transposition, seamless excision and tetrancotide targeting, are the consequences of a structural echo between flanking hairpin recognition and TTAA target recognition. The STC structure suggests that the reason pB almost invariably integrates at a TTAA tetrancotide44,66 is a combination of structural features that include a dramatic target bend, a deformation that causes the TTAA to unpair, and a dense network of interactions that specifically recognize TTAA in single-stranded form. It may be that pB identifies its specific target site by randomly sampling DNA sequences until it encounters one that can be bent enough to unpair its center four bps and that it can subsequently stabilize in ssDNA form. Strand transfer occurs across the minor groove, in contrast with other known transpososomes or intasomes40,49,56,60–64,67 (Supplementary Fig. 8), and the location of active sites in the dimer mandates the 4-bp offset in transfectionerification positions (Fig. 4e). After specific integration and repair, pB is therefore flanked by 4-bp TSDs with a TTAA sequence (Fig. 1a). Crucially, only a subset of the interactions that recognize the single-stranded TTAA target is used to bind and stabilize the hairpin intermediate that form on the transposon ends upon generation of the DSB. At excision, the transposase does not require a specific flanking sequence44, but it does ensure that a hairpin is formed. However, if the flanking sequence is a TTAA, which is the TSD from the previous transposition, upon excision, this will generate the 4-nt TTAA hairpins on the transposon ends. This chain of structurally linked reactions is what leaves behind complementary 4-nt overhangs at the site of genomic excision and allows for seamless repair. Other transposons do not exhibit seamless excision because there is no coupling between how the
DSBs are generated at transposon ends during excision and how target is recognized. For example, the SB transposon is excised with 5′ recesses because cleavage on the NTS is 3-nr within the transposon itself. The resulting 3-nr CAG overhangs at the donor site are not complementary and when the DSB is repaired, an excision footprint is left that also includes one extra copy of the 5′ flanking sequence at the integration site are not complementary and when the DSB is repaired, an excision footprint is left that also includes one extra copy of the 5′ flanking sequence at the integration sites.

The structural view of PB transposition is consistent with data from sequences be TTAA; non-TTAA flanked versions are integrated with comparable efficiency relative to WT, but the TTAA targeting specificity remains. Also, as shown here (Fig. 6a), PB variants such as RE63/RE63 can be excised but not integrated, indicating that excision is more permissive than integration.

The PB from T. ni is but one member of a larger superfamily of PB-like elements. The superfamily includes piggyBat, one of the very few DNA transposons natively active in vertebrates, domesticated elements such as PiggyMac which play a pivotal role in ciliate nuclear development, and the larger group of piggyBac-derived (PGBD) proteins. In humans, PGBD proteins have been implicated in premature ovarian failure and childhood tumor development, and the structures here will provide insight into their action.

A popular version of PB in genomic applications is hyperactive PB, hyPBase, that has seven amino acid substitutions relative to PB and the structures suggests how they may stimulate activity. As two (I30V, S103P) are in the disordered N-terminus, step of mobilization, PB does not require that its flanking sequence be TTAA; non-TTAA flanked versions are integrated with comparable efficiency relative to WT, but the TTAA targeting specificity remains. Also, as shown here (Fig. 6a), PB variants such as RE63/RE63 can be excised but not integrated, indicating that excision is more permissive than integration.
this region is clearly important but most likely not for chemical catalysis; it will be interesting to determine if this region is involved in PB–PB interactions or interactions with cellular factors that contribute to transposition. G165S, located close to the donor DNA, presumably forms an additional interaction with it; M282V is in the omega-loop which may allow better stacking of Y283 with A35 of the target sequence. S509G is at the kink of an α-helix (a12) that interacts with both LE DNAs, and the increased flexibility afforded by the G mutation may strengthen DNA binding. While N538K is the last residue visible in the DDBD before the disordered loop that connects to the CRD, the loop itself is very negatively charged (E539, D545, E549, E550) and the K mutation might provide a stabilizing effect. N571S appears well-positioned to interact favorably with a phosphate at the CRD-binding site. Finally, the phenotype of the R372A/K375A mutation might provide a stabilizing effect. N571S appears itself is very negatively charged (E539, D545, E549, E550) and the K mutation might provide a stabilizing effect. N571S appears well-positioned to interact favorably with a phosphate at the CRD-binding site. Finally, the phenotype of the R372A/K375A mutation might provide a stabilizing effect. With hyPBase, there is some detectable activity with LE/LE TIRs, suggesting that a catalytic PB dimer alone can function if the system is pushed. One fundamental difference between transposition experiments in vitro and in cells is the vast amount of DNA in the nucleus of a cell. It is possible that a higher order assembly is required to provide enough binding strength so that PB LEs and REs present in cells can be brought into each other’s physical proximity to form an effective synapse in the background of a large amount of other DNA. This raised the intriguing possibility that dB transposition in cells might be limited at the synapse step. To test this, we designed a modified LE TIR (LE tandem) by adding additional pB-binding sites. Indeed, LEtandem in combination with the RE TIR showed increased transposition activity in cells (Fig. 6d), mostly likely by directing the assembly of an even higher order transpososome. This tandem TIR also suggests a strategy for designing improved pB vectors for gene delivery. A similar approach has been tried with SfiI with success. However, as the SfiI transpososome structure is not known, the binding sites were simply duplicated and a mutation to prevent cleavage had to be introduced. The requirement for large multimeric assemblies of transposases and integrases for in cell activity while only a dimeric subset is used for cleavage and strand transfer steps is a common feature of many of the systems studied.

### Methods

**Protein expression and purification.** The gene for full-length Trichoplusia ni (T. ni) PB transposase was codon-optimized for mammalian expression (IDT) and cloned into the pD2610 expression vector (a gift from Dr. Alex Kentsis from Memorial Sloan Kettering Cancer Center) between BamHI and KpnI restriction sites, downstream of an N-terminal maltose-binding protein (MBP) tag followed by a TEV protease cleavage site. The plasmid pD2610-MBP-PB was transfected into 500 ml EPI293F cells (Thermo Fisher Scientific) for transient protein expression using a standard PEI transfection protocol. The transfected cells were supplied with 500 ml Epi293 expression medium after 24 h. Cells were harvested 3 days after transfection at 300 × g and stored at −80°C. Cells expressing MBP-tagged PB were resuspended in lysis buffer containing 25 mM Tris–Cl, pH 7.5, 500 mM NaCl, 1 mM TCEP, and protease inhibitor.

*Fig. 7 Model of piggyBac transposition.* Upon liberating the transposon from the donor site, PB then catalyzes hairpin DNA opening and integration into a 4-bp TTAA target site. Target DNA is accessed through the minor groove and undergoes a severe bend and distortion. Hairpin opening and target DNA capture may or may not be concerted at the active site. The observed metal ion is shown as a red sphere and a second metal ion as a black sphere. Catalytic residues are shown as green sticks. DNA molecules are colored as in Fig. 1d. NTS non-transferred strand, TS transferred strand.

The SNHP and STC structures here, which have captured two snapshots along the pB transposition pathway (Fig. 7), reveal the remarkable way that pB achieves seamless excision and its strict TTAA target site specificity. They serve as a foundation for future development of improved genetic tools.
cocktail (Roche). The cells were lysed by three cycles of sonication. Cell lysates were centrifuged at ~95,000 g for 30 min at 4 °C (Beckman Coulter Optima L-100 XP ultracentrifuge). The supernatant was filtered, mixed with 10 ml amylase resin (New England Biolabs) equilibrated with lysis buffer. After one hour of centrifugal rotation, the mixture was loaded onto a gravity flow column and washed with 100 ml lysis buffer. The protein was eluted with 50 ml elution buffer (25 mM Tris–Cl, pH 7.5, 500 mM NaCl, 10 mM maltose, 1 mM TCEP, and protease inhibitor cocktail). The eluate was incubated with TEV protease and TCEP for 16–20 h at 4 °C. The cleaved MBP tag and TEV protease were separated from PB using a 5 ml HiTrap Heparin HP column (GE Healthcare) with a linear-gradient elution from 0 to 1 M NaCl, and further purified on a Superdex 200 HiLoad 16/60 column in 50 mM Tris–Cl, pH 7.5, 500 mM NaCl, and 0.5 mM TCEP. Purified PB was concentrated to about 30 mg/ml and stored at ~80 °C. The yield is about 3 mg/l cell culture.

SEC–MALS analysis of full length PB protein. For determination of molecular mass by SEC coupled with multi-angle light scattering (SEC–MALS) analysis, purified PB at 4 and 2 mg/ml was applied to a Superdex 200 Increase 10/300 (GE Healthcare) column in 20 mM HEPES, pH 7.5, 100 mM NaCl, and 1 mM TCEP. The column eluent was passed through a 660 nm photometer and a refractometer with a dynamic light scattering attachment (Wyatt Technology). Analysis was performed using ASTRA version 7.1.2.5 software and yielded a single molar mass of 135,700 g/mol, consistent with a dimer. The SEC–MALS was performed in the Biophysics Core Facility at NHLBI/NIH.

SNHP and STC complex assembly. The SNHP was prepared by mixing equi-molar amounts of PB protein and LE hairpin DNA, and incubating for 1 h at room temperature (25 °C), followed by SEC purification using a Superdex 200 Increase 10/300 column (GE Healthcare) in 50 mM Tris–Cl, pH 8.0, 100 mM NaCl, 5 mM CaCl2, and 1 mM TCEP. The LE hairpin DNA was prepared by heating LE35hrpn (5′-CAGGATCCCTGGAATGTTAAGCTGTACCAATGTCATGCGTCAATTTTACGC-3′) and LE STC DNA in a 1:2 molar ratio in assembly buffer (50 mM Tris–Cl, 10 mM NaCl, 5 mM MgCl2, and 1 mM TCEP) and dialyzing to a final concentration of 130,000×, corresponding to a calibrated super resolution pixel size of 0.53 Å per pixel on the specimen. The total exposure time for each movie was 10 s with a total exposure dose of 73.7 e-/Å2. The defocus values ranged from ~2.0 μm to a calibrated super resolution pixel size of 0.53 Å per pixel on the specimen. The overall resolution was estimated based on the Fourier Shell Correlation (FSC) threshold at 0.143 between the two half-maps, after a soft mask was applied to mask out the solvent region. The local resolutions were estimated from two unfiltered half maps, using LocalRes within RELION 3.0β8. The sequence was assigned based on visible side chains densities. The bound DNA was built manually. The NMR model of the CRD structure (PDB 5LME) was docked into the density as a rigid body, followed by manual adjustment in COOT. The linker between the CRD and catalytic domain had no interpretable density and was built as a coil using COOT. The model was rebuilt with Rosetta’s rebuilding tools77 using Rosetta’s Cartesian Sampler generating 200 decoys. Execution was parallelized by GPU parallel85. Out of these, the five lowest Rosetta energy decoys were disabilities in O47 and a consensus model that best fit the density was generated manually. This model was refined by the Relax protocol in Rosetta with tightly restrained B factor optimization.

For the STC, the initial model was built from the SNHP complex and fitted into the density with Coot. Real-space refinement against the cryo-EM density map was carried out by the Rosetta Relax protocol including tightly restrained B factor refinement. The validation check for model stereochemistry of both complexes was performed in MolProbity86 (Supplementary Table 1).

In vitro transposition assay using TIR DNA. Purified PB (200 nM) was incubated with SC 100 ng pUC19 plasmid and 100 nM TIR DNA in reaction buffer containing 25 mM Tris–Cl pH 8.0, 25 mM NaCl, 5 mM MgCl2, 2 mM DTT, 1X BSA (NEB) (final volume 100 μl). The mixture was incubated at 30 °C for 1 h unless otherwise indicated. The reaction was stopped by adding 5 μl 0.5 M EDTA, 2 μl for the reaction followed incubation at 37 °C for 30 min. This linearized plasmid DNA was extracted for analysis by ethanol precipitation, dissolved in 10 μl 10 mM Tris–Cl pH 8.0 solution followed by electrophoresis on 1.5% agarose gel running in 1X TAE buffer at 100 V for 70 min. Gels were stained with ethidium bromide for 20 min and imaged using a Chemilumines 5500 gel imaging system (Alpha Innotech). For the STC, the final model was constructed using COOT and MolProbity. The model was further refined and improved by a second round of CTF refinement and Bayesian polishing. A final model was improved by a second round of CTF refinement and Bayesian polishing. A final 3D refinement generated a 3.47 Å map (Supplementary Fig. 4b).

All refinements followed the gold-standard refinement procedure, in which two half-maps were refined independently. The overall resolution was estimated based on the Fourier Shell Correlation (FSC) threshold at 0.143 between the two half-maps, after a soft mask was applied to mask out the solvent region. The local resolutions were estimated from two unfiltered half maps, using LocalRes within RELION 3.0β8. Map visualization was carried out using UCSF Chimera87. All figures were prepared using either PyMol (http://www.pymol.org) or Chimera88.

Ab initio model building and refinement. Except the CRD, no homology model was available to be fitted into the density maps. Both SNHP and STC cryo-EM density maps were of sufficient quality for ab initio atomic model building since distinct structural features (backbones of protein and DNA, secondary structure elements, and large sidechains) appeared in the maps, although the CRD density was not as well resolved as other regions of the complex. We generated a sharpened cryo-EM map in Phenix45 using the SNHP map as input. The SNHP map was prepared from the model by map_to_model also within Phenix. The output model was examined in COOT and the sequence was assigned based on visible side chain densities. The bound DNA was built manually. The NMR model of the CRD structure (PDB 5LME) was docked into the density as a rigid body, followed by manual adjustment in COOT. The linker between the CRD and catalytic domain had no interpretable density and was built as a coil using COOT. The model was rebuilt with Rosetta’s rebuilding tools using Rosetta’s Cartesian Sampler generating 200 decoys. Execution was parallelized by GPU parallel. Out of these, the five lowest Rosetta energy decoys were disabilities in O and a consensus model that best fit the density was generated manually. This model was refined by the Relax protocol in Rosetta with tightly restrained B factor optimization.

To generate linearized pUC19 as target DNA, SC pUC19 was treated by Plasmid Maxi Prep Kit. The unextracted plasmid DNA was purified using a Qiagen Mini Prep Kit. The fraction was incubated at 30 °C for 1 h unless otherwise indicated. The reaction was stopped by adding 5 μl 0.5 M EDTA, 2 μl for the reaction followed incubation at 37 °C for 30 min. This linearized plasmid DNA was extracted for analysis by ethanol precipitation, dissolved in 10 μl 10 mM Tris–Cl pH 8.0 solution followed by electrophoresis on 1.5% agarose gel running in 1X TAE buffer at 100 V for 70 min. Gels were stained with ethidium bromide for 20 min and imaged using a Chemilumines 5500 gel imaging system (Alpha Innotech).

To generate linearized pUC19 as target DNA, SC pUC19 was treated by XmaI (a single cutter) and re-purified. Purified PB (200 nM) was incubated with 100 ng linearized pUC19 plasmid and 100 nM RE TIR DNA variants. The reactions were incubated overnight at 30 °C. Other steps are as described above.

In vitro transposition assays using a mini-transposon. To generate a mini-transposon, we first cloned the PB LE35 and RE65 into pHLL257 using the Quick Change method, and then used the plasmid (called pBLE35RE65) as a template to generate the mini-transposon by PCR. After purification and extraction from an agarose gel, we incubated the 2.9 kb mini-transposon in which the Kan’
gene was flanked by L35 and RE63 with purified PB, pUC19 in the above buffer overnight at 30 °C. The ethanol-preserved DNA was transformed into TOP10 cells and colonies were selected by Amp–Kan– agar plates. The plasmids were extracted and sequenced using primers (5′-CTGATGTCGACGAGCAGCAGC-3′ and 5′-CCGGCGGGGGGGCGAACGGAGCAGC-3′) for sequences inside the mini-transposon to detect the TSD sites generated by transposition89.

In vitro DNA cleavage and hairpin DNA formation assay. Purified PB (200 nM) was incubated with FAM-labeled DNA substrates (5 nM) in reaction buffer (25 mM Tris–Cl pH 8.0, 25 mM NaCl, 5 mM MgCl2, or CaCl2, 1x BSA, 5% glycerol and 2 mM DTT, 50 µl final reaction volume) at 30 °C for 1 h. The reaction was stopped by adding 12.5 µl 98% formamide. Realizations were loaded on a denaturing 15% TBE-urea gel, followed by electrophoresis at 250 V for 25 min at room temperature in 1X TBE buffer. The gel was imaged using a Typhoon FLA 9500 (GE Healthcare).

In cell excision and transposition assays, pCMV-HA-PB90 encodes a hemagglutinin (HA)-tagged PB transposase. All transposon plasmids were derived from pTP88 using standard molecular biology techniques. All plasmids were confirmed by DNA sequencing. For excision assays, 3 million HT-1080 cells were seeded into a 100 mm dish. The next day, cells were transfected with 10 µg of the transposon and 5 µg transposase plasmids using lipofectamine (ThermoFisher). One day later, cells were trypsinized and excision analysis PCR was performed using the following primers (forward: 5′-ATGCGCATCGAGCAGATT-3′ and reverse: 5′-TGTGTGGAATTGTAGGGGA-3′). For colony counts, 0.4 million HT-1080 cells were seeded into each well of a six-well plate. The next day, cells were transfected with 1 µg of transposon and 0.5 µg of transposase plasmids using lipofectamine. One day later, cells were trypsinized and diluted to 100 nm dishes followed by selection with 600 µg/ml of the antibiotic G418 for 8–10 days. Colonies of cells were fixed in 10% formaldehyde/phosphate-buffered saline (PBS), stained with 1% methylene blue in PBS, and counted91.

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

**Data availability**

The data that support this study are available from the corresponding author upon reasonable request. The atomic coordinates of SNHP and STC have been deposited in the protein data bank (PDB) with the accession codes 6X67 for the STC, and 6X68 for the SNHP complex. The EM maps have been deposited in the Electron Microscopy Data Bank (EMDB) with accession codes: EMD-22072 for the STC and EMD-22073 for the SNHP complex. The source data underlying Fig. 6 and Supplementary Figs. 1, 2, and 3 are provided as a Source Data file.

Received: 27 February 2020; Accepted: 2 June 2020;
Published online: 10 July 2020

**References**

1. Azziz, R. K., Breitbart, M. & Edwards, R. A. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38**, 4207–4217 (2010).
2. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of euarkytic genomes. *Annu. Rev. Genet.* **41**, 331–368 (2007).
3. Smit, A. F. Interspersed repeats and other mementos of transposable elements in eukaryotic genomes. *Eukaryot. Cell* **6**, 1220–1229 (2007).
4. Bourque, G. et al. Ten things you should know about transposable elements. *Trends Genet.* **24**, 9–16 (2008).
5. Park, M. A., Jung, H. S. & Slukvin, I. Genetic engineering of human induced pluripotent stem cells by the piggyBac transposon. *Curr. Protoc. Stem Cell Biol.* **47**, e63 (2018).
6. Woltjen, K., Hamalainen, R., Kibschull, M., Mileikovsky, M. & Nagy, A. Preclinical and transition to clinical trials. *Cell* **171**, 1391–1393 (2017).
7. Tipanee, J., Chai, Y. C., VandenDriessche, T. & Chuah, M. K. Preclinical and transition to clinical trials. *Cell* **171**, 1391–1393 (2017).
8. Hancks, D. C. & Kazazian, H. H. Jr. Roles for retrotransposon insertions in human disease. *Nat. Rev. Genet.* **16**, 19–29 (2016).
9. Ivics, Z. & Izsvak, Z. Sleeping beauty transposition. *Annu. Rev. Genet.* **43**, 995–1001 (2009).
10. Hencken, C. G., Li, X. & Craig, N. L. Functional characterization of an active Rag-like transposase. *Nat. Struct. Mol. Biol.* **19**, 834–836 (2012).
11. Liu, C., Yang, Y. & Schatz, D. G. Structures of a Rag-like transposase during cut-and-paste transposition. *Nature* **575**, 540–544 (2019).
12. Reznikoff, W. S. Transposon Tn5. *Annu. Rev. Genet.* **42**, 269–286 (2008).
13. Jin, Y. et al. DNA-PK facilitates piggyBac transposition by promoting paired-end complex formation. *Proc. Natl Acad. Sci. USA* **114**, 7408–7413 (2017).
43. Morelli, N. et al. Sequence-specific DNA binding activity of the cross-brace zinc finger motif of the piggyBac transposase. *Nucleic Acids Res.* 46, 2660–2677 (2018).

44. Li, M. A. et al. The piggyBac transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites. *Mol. Cell. Biol.* 33, 1317–1330 (2013).

45. Terwilliger, T. C., Adams, P. D. Automated model building by maximization of detail and connectivity. *Acta Crystallogr. D* 74, 545–559 (2018).

46. Wang, R. Y. et al. Automated structure refinement of macromolecular assemblies by cryo-EM maps using Rosetta Maps 5, e17219 (2016).

47. Wang, R. Y. et al. Automated structure reprocessing reaction. *Nucleic Acids Res.* 36, 2595–2605 (2008).

48. Wang, R. Y. et al. Automated structure reprocessing reaction. *Nucleic Acids Res.* 36, 2595–2605 (2008).

49. Wang, R. Y. et al. Automated structure reprocessing reaction. *Nucleic Acids Res.* 36, 2595–2605 (2008).

50. Wang, R. Y. et al. Automated structure reprocessing reaction. *Nucleic Acids Res.* 36, 2595–2605 (2008).

51. Wang, R. Y. et al. Automated structure reprocessing reaction. *Nucleic Acids Res.* 36, 2595–2605 (2008).

52. Wang, R. Y. et al. Automated structure reprocessing reaction. *Nucleic Acids Res.* 36, 2595–2605 (2008).

53. Wang, R. Y. et al. Automated structure reprocessing reaction. *Nucleic Acids Res.* 36, 2595–2605 (2008).

54. Wang, R. Y. et al. Automated structure reprocessing reaction. *Nucleic Acids Res.* 36, 2595–2605 (2008).

55. Richardson, J. M., Colloms, S. D., Finnegan, D. J. & Walkinshaw, M. D. Mutation of Tn5 transposase beta-loop residues affects all steps of Tn5 transposition: the role of conformational changes in three-dimension. *Biochem. Biophys. Acta* 1804, 966–1010 (2010).

56. Ghanim, G. E., Kellogg, E. H., Nogales, E. & Rio, D. C. Structure of a P element transposase–DNA complex reveals unusual DNA structures and GTP–DNA complex. *Nat. Struct. Mol. Biol.* 26, 1013–1022 (2019).

57. Holm, L. Benchmarking fold detection by Dalili v5. *Bioinformatics* 35, 5326–5327 (2019).

58. Pacheco, S. et al. Structural basis for the 3′-5′ exonuclease activity of *Escherichia coli* DNA polymerase I: a two metal ion mechanism. *EMBO J.* 10, 2536–2543 (1991).

59. Wang, Y., Lee, J. Y. & Nowotny, M. Making and breaking nucleic acids: two Mg2+–ion catalysis and substrate specificity. *Mol. Cell.* 22, 5–13 (2006).

60. Stefl, R., Wu, H., Ravindranathan, S., Sklenar, V. & Feigon, J. DNA-A tract bending in three dimensions: solving the DA4T4 vs. T4A4 conundrum. *Proc. Natl. Acad. Sci. USA* 110, 1717–1182 (2014).

61. Davies, D. R., Guo, L. Y., Reznikoff, W. S. & Rayment, I. Three-dimensional structure of the Tn5 synapt complex transposition intermediate. *Science* 289, 77–80 (2000).

62. 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).

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

64. Steiniger, M., Metzler, J. & Reznikoff, W. S. Mutation of Tn5 transposase beta-loop residues affects all steps of Tn5 transposition: the role of conformational changes in three-dimension. *Biochemistry* 45, 15552–15562 (2006).

65. Ason, B. & Reznikoff, W. S. Mutation analysis of the base flipping event found in Tn5 transposition. *J. Biol. Chem.* 277, 11284–11291 (2002).

66. Bischouni, J. & Chalmers, R. Base-flipping dynamics in a DNA hairpin processing reaction. *Nucleic Acids Res.* 35, 2584–2599 (2007).

67. Hare, S., Maertens, G. N. & Cherepanov, P. 3′-processing and strand transfer catalysed by retroviral integrase in crystals. *EMBO J.* 31, 3020–3028 (2012).

68. Crooks, G. E. et al. Crystal structure of the Rous sarcoma virus intasome. *Nature* 530, 362–366 (2016).

69. Ballandras-Colas, A. et al. A supramolecular assembly mediates lentiviral DNA integration. *Science* 355, 93–95 (2017).

70. Passos, D. O. et al. Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. *Science* 355, 89–92 (2017).

71. Morris, B. R., Gray, H., McKenzie, G., Jones, A. C. & Richardson, J. M. A bend, flip and trap mechanism for transposon integration. *elife* 5, e15357 (2016).

72. Wilson, M. H., Coates, C. J. & George, A. L. Jr. PiggyBac transposon-mediated gene transfer in human cells. *Mol. Ther.* 15, 139–145 (2007).

73. Burmigh, E. R. et al. A hyperactive transposase promotes persistent gene transfer of a piggyBac DNA transposon. *Mol. Ther. Nucleic Acids* 1, e1502 (2012).

74. Chen, X. et al. How mouse RAG recombinase avoids DNA transposition. *Nat. Struct. Mol. Biol.* 27, 127–133 (2020).

75. Boulougue, M., Rouault, J. D., Hua-Van, A., Makni, M. & Capy, P. Molecular evolution of piggyBac superfamily: from selfishness to domestication. *Genome Biol.* 9, 332–339 (2017).

76. Mitra, R. et al. Functional characterization of piggyBac from the bat *Myotis lucifugus* unveils an active mammalian DNA transposon. *Proc. Natl. Acad. Sci. USA* 110, 234–239 (2013).

77. Ray, D. A. et al. Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. *Genome Res.* 18, 717–728 (2008).

78. Baudry, C. et al. PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramaecium tetraurelia*. *Genes Dev.* 23, 2478–2483 (2009).

79. Qin, Y. et al. CSB-PGBD3 mutations cause premature ovarian failure. *PLoS Genet.* 11, e1005419 (2015).

80. Hennses, A. G. et al. PGBD3 promotes site-specific oncogenic mutations in human tumors. *Nat. Genet.* 49, 1005–1014 (2017).
