UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

P element transposases in Drosophila melanogaster and other Eukaryotes

Permalink

https://escholarship.org/uc/item/3sb9581x

Author

Ghanim, George E

Publication Date

2019

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

P element transposases in *Drosophila melanogaster* and other Eukaryotes

By

George E. Ghanim

A dissertation submitted in the partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Donald C. Rio, Chair
Professor Kathleen Collins
Assistant Professor Dirk Hockemeyer
Professor Rasmus Neilson

Summer 2019

ABSTRACT

P element transposases in *Drosophila melanogaster* and other Eukaryotes

By

George E. Ghanim

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Donald C. Rio, Chair

Transposable elements are mobile genetic sequences that are found in the genomes of nearly all organisms. DNA transposons constitute a major class of transposable elements and can move throughout the host genome by a cut-and-paste mechanism catalyzed by an encoded transposase protein. Although transposase activity can be detrimental to the host, numerous examples of host benefit have been documented. Over evolutionary time transposon-related sequences and proteins have been adapted to serve a wide range of cellular functions, a process termed transposon domestication.

The *Drosophila* P element is one well-studied example of a eukaryotic DNA transposable element. Although the encoded P element transposase protein has been biochemically characterized, it exhibits several features that distinguish it from the other characterized DNA transposases. Namely, P element transposase requires a guanosine triphosphate (GTP) cofactor and generates unusually long 17 nucleotide staggered DNA breaks at the transposon ends during transposition. To gain insight into the molecular basis of these distinguishing features we determined the cryo-EM structure of the *Drosophila* P element transposase strand transfer complex (STC) to 3.6 Å - a nucleoprotein complex in which the transposase protein is bound to P element donor DNAs covalently joined to a target DNA. Our structure reveals that the STC is dimeric, the P element donor DNAs adopt a highly unusual DNA geometry and further reveals a function for GTP in positioning the P element ends into the transposase active site for catalysis. This structure provides the first view of the P element superfamily of eukaryotic DNA transposases, offers new insights in P element transposition and implies a transposition pathway that is mechanistically distinct from other cut-and-paste DNA transposases.

Furthermore, bioinformatic and biochemical analysis have identified C_2CH DNA binding domain termed the THAP domain. This novel and evolutionarily conserved domain is found across a wide range of animal genomes, including vertebrates, invertebrates, *Drosophila* P element transposase, in primates and in 12 human genes. Of the 12 THAP domain containing genes in

humans, THAP9 is homologous to the entirety of *Drosophila* P element transposase, still has DNA transposase activity, but lacks the hallmarks of an active DNA transposable element. maintained. The evidence implies that THAP9 has likely been domesticated/adapted by the cell in early chordates from an ancient THAP9-like P element transposon, such as those found in *Ciona*. However, a cellular function for THAP9 has not been identified. In an attempt to elucidate a cellular function for THAP9, we carried out genome-editing in human embryonic stem cells (hESCs) to either knockout or epitope tag the endogenous THAP9 gene. Disruption of THAP9 did not produce overt phenotypic changes in hESCs and did not affect differentiation into fibroblast-like cells, indicating that THAP9 is likely not required for the hESC maintenance. However, endogenously epitope tagged THAP9 is translated, can be immunoprecipitated and localizes to the nucleus in hESCs. To determine potential THAP9 human genome cleavage and binding sites, we raised an antibody to purified, recombinant human THAP9 protein, performed direct *in situ* breaks labeling, enrichment on streptavidin and next-generation sequencing, or BLESS, to detect potential DNA cleavage site, a method used successfully to find Cas9 off-target genomic cleavage sites and ChIP-Nexus experiment, a chromatin immunoprecipitation method similar to ChIP-Exo. The ongoing analysis and comparison of both the BLESS and ChIP-Nexus sequencing data should identify genomic binding sites, potential genomic DNA cleavage sites, motifs associated with human THAP9 DNA binding and cleavage and should uncover a cellular function for the human THAP9 gene.

While these projects are essentially independent of one another, they all relate to P element DNA transposases. Together, they hopefully contribute to a deeper understanding of the mechanisms of P element transposition and the expanding roles that transposase-related proteins play in the context of cellular function in human cells.

For Jamal.
From our family's humble beginnings
as olive tree farmers, plumbers, and carpet installers.
We can now add nucleoprotein biochemist to the list.
Thanks Jim.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I have truly stood upon the shoulder of giants. This journey was made easier by the insights, advice and support of many great individuals along the way. I couldn't have done any of this without all of you.

I would like to start off by thanking the Rio lab, particularly Yeon and Qingqing. It wasn't always easy being the only graduate student in the lab but it would have been much harder without your support, encouragement, and advice over these last few years. You have been phenomenal lab mates and I hope that I have reciprocated the help you have offered me.

Thanks to the Collins and Hockemeyer labs, particularly Chiba, Franzi, Kelly and Heather, for being such great comrades. I have always been able to count on you all for reagents, scientific advice and more importantly friendship. Chiba, I could always count on you to get a drink at 11PM, while an experiment was running or a gel was drying. Thank you, also, to my thesis committee for the constructive feedback and support. You have kept me on track with this difficult project and have always had my best interests in mind. Thank you Dirk and Kathy, for your infectious enthusiasm for science, insights and advice over the years. Dirk, you have always known what to say when it was needed the most. In the end, your clear perspective has often made the difference for me. Kathy, it was always a pleasure to share "beautiful" gels with you and get your feedback. There was always a breakthrough shortly after our meetings – not a coincidence.

A big thank you to Don. I cannot put into words my appreciation and reverence for your guidance and mentorship, both inside and outside of science. You have been an excellent mentor by providing me with criticism, feedback and insights on experiments and data, yet allowed me the freedom, flexibility and support to explore ideas that you "liked the least." You were never too busy to come in on the weekends to practice a talk or go over a manuscript. I will always look back fondly at our bench side chats and your anecdotes of Mizuuchi and Mu transposase. They kept my spirits high when I was ready to give up on the P element transposase project. Again, thank you Don.

To my family, thank you. Mom and Dad, thank you for encouraging me to go into science, rather than art because, in hindsight, I think I make a much better biochemist. You were always understanding, unconditionally supportive, and only phone call away - it made all the difference. Thank you to my three younger brothers. You three were always ready to help me in any way that you could. I will always appreciate your interest and excitement for the scientific projects I worked on even when it was too complicated to understand. I wouldn't be where I am without all of you.

And finally, a thank you to Mana. You were there for me during the highs and during the lows. I could write another dissertation on the ways you have supported me, but I won't. You've once again managed to end up last on the list – but I wouldn't have it any other way. Thank you.

CHAPTER ONE

Transposable Elements, Drosophila P Element Transposase and Human THAP9

Introduction

Transposable elements are repetitive genetic sequences that can be mobilized to increase their copy number within a host genome or through horizonal gene transfer. Due to modern genome sequencing efforts we now know that these mobile elements comprise a substantial fraction of many eukaryotic genomes and are present in nearly all extant organisms, with a few rare exceptions (*i.e. Plasmodium falciparum* (Rebollo et al., 2012)). The successful dissemination of transposable elements can be attributed to their self-contained, autonomous nature. However, these elements generally depend on host cell DNA repair pathways. Although sometimes regarded as parasitic or selfish-elements, transposon mobilization can have variable (either deleterious or beneficial) consequences for the host genome. In humans, deleterious transposable element insertion events have been shown to cause diseases, such as hemophilia A (Kazazian et al., 1988), and β-thalassemia (Thein, 2013), while a related set of transposable elements provide essential functions for telomere maintenance in *Drosophila* (Villasante et al., 2008). Over evolutionary timescales, transposable element-derived genes and other sequence elements have even been domesticated to perform essential cellular functions, such as telomerase in telomere maintenance (Nakamura and Cech, 1998), the V(D)J recombinase a key component of the adaptive immune system of jawed vertebrates (Huang et al., 2016; Kapitonov and Jurka, 2005) and the generation of introns at a genomic scale (Huff et al., 2016).

Classification of Transposable Elements

Historically, the classification of transposable elements has been based on either the presence of an RNA intermediate during transposition, leading to a “copy-and-paste” (retrotransposons) or the lack of, leading to a “cut-and-paste” (DNA transposon) mechanism. However, both bacterial and eukaryotic transposable elements have been discovered that violate this mechanistic distinction, challenging that classification system (Wicker et al., 2007). It is now known that there are a number of elements that “copy-and-paste” without an RNA intermediate (Wicker et al., 2007). Here, I will follow the classification system presented by Wicker *et al.*, to describe the major groups of transposable elements, their distinctions, modes of transposition and provide some examples.

Retrotransposons

Transposable elements are characterized into two major classes depending on the mechanism of mobilization, Class I or Class II. Class I transposable elements, or retrotransposons, mobilize by going through an RNA intermediate that is copied to DNA using a reverse transcriptase and are further organized into five orders based on the mechanism of retrotransposition; long-terminal repeat (LTR) elements, *Dictyostelium* intermediate repeat sequence-like elements (DIRS), Penelope-like elements (PLE), *long-interspersed nuclear elements* (LINE) and *short-interspersed nuclear elements* (SINE). Like the closely related retroviruses, LTR elements are flanked by direct long-terminal repeats and encode a capsid protein (GAG), protease (PRO), reverse transcriptase (RT) and integrase (IN), and sometimes an envelope protein (ENV)(Figure 1.1, top). The RT and IN are both are required for retrotransposition. After transcription of the LTR element, the element encoded RT generates a double-stranded cDNA copy of the element, which is then integrated back into the host genome by the action of IN (Beauregard et al., 2008). Examples of LTR elements include the *copia* and *gypsy* elements in *Drosophila* and the endogenous retroviral (ERV) elements, such as HERV-H or HERV-K in humans. *Dictyostelium* intermediate repeat sequence, or DIRS-like elements are similar to the LTR elements, however encode a tyrosine recombinase rather than an IN protein (Goodwin and Poulter, 2004).

Unlike the LTR class, LINEs, SINEs and PLE-like elements do not employ an integrase or recombinase for integration, but mobilize through a distinct mechanism termed target-primed reverse transcription (TPRT)(Figure 1.1, top). In TPRT the encoded RT engages an RNA copy of the transposable element and directly primes reverse transcription from an exposed 3'-OH group on the target DNA. LINEs and *Penelope* use an element-encoded endonuclease (EN) to expose the target 3'-OH. However, *Penelope*-like elements are flanked by LTRs. SINE elements are non-autonomous elements and thus do not possess the factors necessary for retrotransposition. SINEs instead rely on the RT and EN proteins encoded by LINE elements to achieve retrotransposition. The subclasses of SINE elements are ancestrally derived from tRNA, 7SL and 5S RNA sequences, and include the well-known primate-specific *Alu* elements (Wicker et al., 2007)(Figure 1.1, top).

DNA Transposons

Unlike Class I retrotransposons, Class II elements or DNA transposons, do not mobilize through an RNA intermediate. DNA transposons are found across a wide range of organisms and are usually present at low to moderate copy numbers. Class II elements that are flanked by terminal inverted repeats (TIR) can be thought of as the classic “cut-and-paste” elements (Figure 1.1, bottom). These terminal inverted repeat elements, or TIR elements, use an element-encoded transposase protein to recognize, excise and integrate transposon DNA from one genomic location to another, generating short, direct duplications of the target site flanking the new transposon insertion. TIR DNA transposons are further classified into superfamilies distinguished by terminal inverted repeat sequences and lengths, target site duplication (TSD) lengths and sequence/structural similarities of the element-encoded transposase proteins (*e.g. hAT*, *Transib*, *PiggyBac* and *P* elements). The TIR DNA transposon class includes predominantly eukaryotic elements. However, some prokaryotic DNA transposons and transposases (such as Tn5 and bacteriophage Mu) share similarities in transposon structure, transposase active sites and mechanisms and will therefore be included in the following description of TIR DNA transposons. Collectively, the TIR elements use a variety of mechanisms to cleave the DNA strands at the transposon ends and will be discussed in a later section.

The "cut-and-paste" DNA transposons characterized to date all use a transposase protein with a catalytic RNase H-like domain with an acidic DDD or DDE triad at the active site (Yuan and Wessler, 2011). However, modern genome sequencing efforts have revealed two novel groups of DNA transposons that do not use a DDD/DDE transposase and instead mobilize through a "copy-and-paste" mechanism, the so-called *Helitron* and *Maverik* elements. *Helitron* elements encode a large multidomain transposase and are thought to mobilize through a "rolling-circle" like mechanism (Grabundzija et al., 2016). The transposase contains an HUH nuclease and helicase domain which recognizes hairpin structures on the transposon ends. Transposition is initiated by strand nicking and DNA synthesis to generate a copy of the transposon. The HUH active site tyrosine residues form a covalent linkage with the DNA and integrate the transposon elsewhere. *Maverik* elements are unusually long (10-20 kb), are flanked by long terminal inverted repeats and encode a large number of proteins, including a DNA polymerase and an integrase. These elements are thought to undergo replicative transposition through a single-stranded DNA intermediate (Kapitonov and Jurka, 2006).

Effects of Transposition on the Host Genome

The act of transposon mobilization can burden and compromise the survival of the host organism due to effects on genome integrity. During cut-and-paste DNA transposition, the excision of the transposon leaves behind a double-stranded DNA break that must be repaired by cellular factors. Integration of the transposon into new sites can also be detrimental to the host (*e.g.* by disrupting host genes or gene regulation or generating DNA damage) and is often a source of mutations or illegitimate chromosomal recombination events. One of the best-studied examples of transposition events with a detrimental impact to the host is a genetic syndrome in *Drosophila melanogaster*, termed hybrid dysgenesis. In the early 1970s, a syndrome of genetic traits and abnormalities was characterized that occurred between crosses of established *Drosophila* laboratory strains (pre-1930s) and strains more recently collected from wild populations (Kidwell et al., 1977). Crosses between *Drosophila* males from wild strains and females from lab strains resulted in progeny with a number of abnormal traits, including elevated mutation rates, temperature-dependent sterility, male recombination and chromosomal aberrations. Further analysis demonstrated that hybrid dysgenesis phenotypes were explained by the recent DNA transposon invasion into *Drosophila melanogaster* (called P elements) and their mobilization within the developing germline (Bingham et al., 1982). For instance, sterility is caused by DNA damage-induced death of germ cells, while elevated mutation rates are attributed to P element insertions and improper excision events into or near genes (Engels, 1996).

The detrimental effects of transposition is not restricted to Class II elements or the *Drosophilid* genus, but can be caused by any mobile element and is rather widespread across organisms. For instance, it is known that LINE-1 and their non-autonomous counterparts (*e.g. Alu* and SVA elements) are the only active retroelements within the human genome (Kazazian and Moran, 2017; Sassaman et al., 1997). The rate of *de novo* LINE-1 and *Alu*/SVA insertion events is estimated to be 1 in 95 and 1 in 21 births, respectively (Bourque et al., 2018), with more than 120 insertion events associated with human disease (Kazazian and Moran, 2017).In fact, the first reported transposition event associated with a human disease was *de novo* LINE-1 insertions into exon 14 of the factor VIII gene in two unrelated patients with haemophilia A (Kazazian et al., 1988).

While there are numerous examples of genomic abnormalities caused by transposition, mobilization is not strictly required for transposable elements to impact the host. The mere presence of transposons in the genome can impose *cis*- and *trans*-acting effects on gene expression patterns. These effects include the generation of antisense RNAs, the recruitment of transcription factors and the methylation or heterochromatization of the transposon-containing genomic loci (Chuong et al., 2017). These effects were observed before the discovery of transposable elements by Barbra McClintock in her seminal work on “controlling elements” in maize (McClintock, 1984). She had noted variegated pigment color patterns of maize kernels in response to environmental challenges and later determined that mobile genetic factors could change, or control, gene expression patterns, hence the name “controlling elements”.

Conversely, there are numerous instances in which transposon insertions can provide an advantage to the host. Under the appropriate selective pressures, these beneficial transposon-containing alleles can spread throughout global populations. A study using *Drosophila melanogaster* as a model to dissect the genetic basis of insecticide resistance identified a transposon insertion at the 5' end of the *Cyp6g1* gene (Daborn et al., 2002). The transposon insertion allele increases *Cyp6g1* gene expression conferring resistance to the insecticide DDT and has spread across global *Drosophila* populations under this strong selective pressure. Other examples of beneficial transposon insertion alleles include a transposon insertion that enhances *cortex* gene expression underlying the industrial melanism of the peppered moth (Hof et al., 2016), and a retrotransposon insertion that functions as a temperature-inducible enhancer of the *Ruby* gene in blood oranges (Butelli et al., 2012).

Mobile Element Adaptation and Domestication

Over evolutionary timescales some transposon-derived sequences have been co-opted by the host to carryout various cellular functions, in a process termed domestication. Through this domestication process, transposable elements have provided numerous protein coding regions, non-coding sequences, protein domains and entire gene sequences that are involved in gene expression networks or that carry out other essential cellular processes. For instance, the interferon response is a signaling pathway that activates the transcription of innate immunity genes and comprises a major branch of innate immune pathway. LTR promoter regions from endogenous retroviruses (ERVs) are enriched within interferon-induced transcription factor binding sites. These ERV sequences have been independently dispersed across numerous mammalian genomes and act as INF-inducible enhancers shaping the transcriptional landscape of the immune response (Chuong et al., 2016). Similarly, transcribed LINE1 RNAs act as a nuclear scaffold to regulate the gene expression landscape essential for maintaining embryonic stem cell identity in the mouse (Percharde et al., 2018).

Transposable elements can also provide genes or protein domains to the cell for numerous and essential cellular functions. Well-known examples include telomerase derived from non-LTR retrotransposons (Nakamura and Cech, 1998), the spliceosome derived from mobile group II introns (Rodríguez-Trelles et al., 2006) and RAG1/RAG2 V(D)J recombinase derived from a DNA-based transposon (Huang et al., 2016; Kapitonov and Jurka, 2005). It is generally accepted that the recombinase signal sequences (RSSs) and recombinase-activating genes (RAG1/RAG2) of the V(D)J recombinase originated from an ancient DNA transposon belonging to the *Transib* superfamily (Kapitonov and Jurka, 2005). In a mechanism analogous to DNA transposon excision, the RAG1/RAG2 complex binds the RSSs and initiates recombination between the variable(D), diversity (D) and joining (J) gene segments at the immunoglobulin and T cell receptor genes of jawed vertebrates (Huang et al., 2016). This combinatorial assembly process, termed V(D)J recombination, allows for the massively diverse protein coding potential at the immunoglobulin and T cell receptor loci, generating a large repertoire of antibodies that form the basis of adaptive immunity. More recently, an ancestral *ProtoRAG* transposon was discovered in the Lancelet (Amphioxus) (Huang et al., 2016). The *ProtoRAG* transposon exists in a distant chordate lineage and is thought to be the mobile element from which RAG1/RAG2 and the RSSs of the jawed-vertebrate immune system originated from. Structural and biochemical characterization of the *ProtoRAG* transposase identified RAG1/RAG2-specific adaptations that suppress transposition activity, providing insights into the evolutionary pathway guiding transposon domestication (Zhang et al., 2019).

Overview of DNA Transposons, Transposases and Transposition

There are variations in the structure of DNA transposons and the corresponding mechanisms of transposition, yet several features can be generalized to describe the fundamental anatomy of a transposon and the transposition mechanism.

Anatomy of a Terminal Inverted Repeat DNA Transposon

As their name implies, a key feature of DNA terminal inverted repeat transposons is the presence of inverted repeat sequences at the transposon ends (TIRs)**.** These identical or nearly identical sequences are found in opposite orientations at the left and right ends of the transposon. Usually the DNA transposase will bind at or near the TIRs to initiate transposition. The TIRs of different transposon superfamilies vary in both size and structure. The terminal inverted repeats can either be relativity simple and short, such as the 19 bp outer and inner transposon ends of the Tn5 composite bacterial transposon (Reznikoff, 2008), or they can be long and complex, occurring as multiple copies and in alternating orientations. More complex transposon ends include the six Mu-A binding sites (designated L1, L2, L3 and R1, R2, R3) of bacteriophage Mu (Mizuuchi and Craigie, 1986) and the ~200 bp termini of eukaryotic *hAT* elements, which contain TIRs and numerous short repeats scattered throughout each end (Hickman et al., 2014). In addition to the flanking TIRs, inverted repeat sequences may be found inside the transposon termini and are usually called internal inverted repeats or IIRs. These repeat sequences are thought to facilitate recruitment of the transposase protein and can be required for the proper assembly of the transposase-transposon nucleoprotein complex before cleavage. As such, while minimal TIRs are often sufficient *in vitro*, the full transposon end complete with TIRs and IIRs is often necessary for efficient transposition *in vivo* (Coupland et al., 1989; Hickman et al., 2014; Li et al., 2005; Mullins et al., 1989). Furthermore, it is often the case that the left and right transposon ends are functionally distinct and not interchangeable (Hickman et al., 2014; Mullins et al., 1989).

DNA TIR transposons are also flanked by short direct repeats derived from host genomic DNA, termed target site duplications (TSDs). The target site duplication is a hallmark of transposition and results from the staggered integration of the transposon ends into the top and bottom strands of the target DNA followed by gap repair by host factors. The length of the target site duplication is characteristic to each DNA transposon superfamily and may be short (2 bp for Tc1/*mariner* elements) or longer (8 bp for P and *hAT* elements, 9 bp for Tn5) (Feschotte and Pritham, 2007).

Transposon-Encoded Transposase Proteins

DNA transposons additionally encode a transposase protein, the enzyme responsible for catalyzing the excision and integration of the DNA transposon from one genomic locus to another. The specific domain organization and architectures of DNA transposase proteins varies greatly among the characterized transposon superfamilies. However, several salient features of DNA transposases are common. DNA transposases contain one or more DNA-binding domains, a catalytic RNase H-like domain and oligomerization interfaces/domains that provide the potential to oligomerize.

Transposase DNA-binding domains

A variety of DNA-binding domains have been observed in the structures of TIR DNA transposases that can be sequence-specific, such as the P element THAP DNA-binding domain (Lee et al., 1996; Roussigne et al., 2003b) or sequence non-specific, such as the BED-finger domain of *Hermes* (Aravind, 2000; Hickman et al., 2005) and IIβ and IIIα domains in Mu transposase. The encoded transposase protein can possess one or more DNA-binding domains that adopt various structures, including the winged helix-turn-helix domain (Iα domain of Mu) (Clubb et al., 1994), helix-turn-helix domains (HTH1 and HTH2 of *Mos1*) (Richardson et al., 2006), AT-hooks-like motifs (*mariner*/*Mos1*) (Richardson et al., 2006), C_2CH zinc-coordinating THAP domain(P element) (Roussigne et al., 2003b; Sabogal et al., 2010) or the predicted C_2HC zinc-coordinating BED finger domain (*hobo* and *Hermes* elements) (Aravind, 2000). Interactions between the DNA-binding domains and one or more binding sites on the transposon end facilitate the proper assembly of the transposon-transposase nucleoprotein complex and ensures the fidelity of the transposition reaction.

Transposase Oligomerization

All TIR DNA transposases characterized to date are understood to from multimeric complexes within the nucleoprotein transposome. The multimeric architectures allow for "communication" between transposase subunits or domains, ensuring correct engagement and assembly of both transposon ends before catalysis. Thus, TIR DNA transposases often support *trans*-catalysis mechanisms, in which a transposase subunit binds a particular transposon end, but catalytically processes the other end. Taken together, oligomerization is thought to act as an assembly checkpoint, ensuring the proper engagement of both transposon ends, before continuing transposition. Although oligomerization is a common feature of DNA transposases, structural characterization of several major transposase superfamilies reveals that the mode of oligomerization is divergent (Hickman et al., 2010). Multimerization can be achieved through a dedicated domain as in the case of the leucine zipper domain of *Drosophila* P element transposase (Lee et al., 1996; Rio, 1990) or through the association of several domains as is the case for RAG1/RAG2 (Kim et al., 2015). In a RAG1 protomer, three separate domains/segments intertwine with the corresponding regions of another RAG1 protomer to dimerize, in addition to fulfilling other roles, such as site-specific DNA-binding (Yin et al., 2009).

The transposase oligomeric state may also be dependent on DNA binding. For instance, Tn5 and Mu are both monomers in solution and dimerize or tetramerize, respectively, upon DNA binding. In fact, the oligomeric state of Tn5 acts to directly regulate transposase activity. In the absence of DNA Tn5 adopts a monomeric, auto-inhibited state, in which the N terminus (that is required for DNA binding) engages and sequesters the C terminus (required for dimerization) (la Cruz et al., 1993)) (Reznikoff, 2008). Other transposases are multimeric on their own, such as the *Hermes* transposase which adopts an octameric ring shaped assembly, with or without DNA (Hickman et al., 2014).

The RNase H-like Catalytic Domain

The catalytic domain of DNA transposases folds to organize three acidic amino acid residues. These acidic residues, DDE or sometimes DDD, coordinate two divalent metals and are responsible for catalyzing the nucleophilic cleavage and joining of DNA phosphodiester bonds. Although there is little primary protein sequence similarity across DNA transposase superfamilies, the cores of the catalytic domains adopt a similar topological arrangement, termed the RNase H-like fold for its similarity to the catalytic fold first identified in RNase H (Yang et al., 1990). The RNase H fold at the core of the catalytic domain consists of mixed α-helixes and β-strands, β1-β2-β3-α1-β4-α2/3-β5-α4-α5 (Hickman et al., 2010). The central 5-stranded β-sheet adopts a characteristic 3-2-1-4-5 strand order and is buttressed above and below by α-helixes (Hickman et al., 2010; Yang and Steitz, 1995). The three catalytic residues are found in nearly identical topological positions: the first carboxylate on β1, the second on or after β4, and the third on or before α4 (Hickman et al., 2010).

The loops connecting these secondary structure elements of the RNase H-like fold can vary considerably in size among transposases. Of particular interest is the region connecting β5 and α4. In some transposases (Rice and Mizuuchi, 1995) and related retroviral integrases (Dyda et al., 1994) the β5 and α4 structural elements are connected by a short, often disordered, loop. In contrast, other transposases can have an entire domain inserted between β5 and α4 of the RNase H-like fold. Domains found at this position are called “insertion domains” and can play significant roles during the transposition reaction. For instance, in Tn5, a mostly β-stranded 96 amino acid insertion domain acts to stabilize a hairpin on the transposon DNA end, an intermediate formed during the transposition reaction of Tn5 (Davies et al., 2000). The entirely α-helical insertion domain found in *Hermes* transposase is considerably large (residues 265-552) and plays critical roles during transposition by facilitating hairpin formation and contributing to transposase oligomerization (Hickman et al., 2005). Insertion domains are also found in the members of the *hAT*, *Mutator*, *Transib*, *CACTA*, *piggyBac* and *P element* eukaryotic DNA transposase superfamilies (Hickman et al., 2010; Yuan and Wessler, 2011).

Transposon-Encoded Proteins

In addition to encoding the catalytic transposase protein, DNA transposons can also encode their own inhibitors through a variety of mechanisms. The Tn5 transposon encodes both the Tn5 transposase and an N-terminally truncated version of Tn5 through an alternate translation start codon (Johnson et al., 1982). This naturally-occurring truncated version of Tn5, called Inh, inhibits transposase activity by forming nonfunctional oligomers with Tn5 (la Cruz et al., 1993). Another well-characterized example of a naturally encoded inhibitor protein occurs within the *Drosophila* P element. Alternative splicing of the P element pre-mRNA third intron (IVS3) introduces 15 unique amino acids followed by a premature stop codon and encodes a smaller 66kD C-terminally truncated protein rather than the full-length transposase (Rio, 1990). Genetic experiments demonstrated that the 66kD protein-expressing P elements cause a marked reduction in P element transposase activity (Misra and Rio, 1990; Rio, 1990). While the exact mechanism of repression is not known, it is thought to occur through competitive binding to the P element transposon ends (Lee et al., 1998) or through protein-protein interactions by the production of nonfunctional oligomers between full-length transposase and the 66kD repressor protein. These element-encoded transposase repressors/regulators are important components to long-term transposon survival because rampant transposition would surely kill the host cell.

Some transposons, particularly prokaryotic mobile elements, encode a variety of other genes which can confer antibiotic resistance, function as transposase activators, or aid in transposase target site selection and delivery. The prokaryotic transposon Tn5 encodes three antibiotic resistance genes conferring kanamycin resistance, bleomycin resistance, and streptomycin resistance (Reznikoff, 2008). Transposon Tn10 has several open reading frames, some of which encode resistance to tetracycline (*tetA*, *tetR*, *tetD*, and *tetC*) and others with unclear functions (*jemA*, *jemB* and *jemC*). Like Tn5 and Tn10, bacterial Tn*7* encodes a number of genes, some of which confer antibiotic resistance. However, Tn7 is unusual in that functions of the "transposase" is encoded across five genes (*TnsA*, *TnsB*, *TnsC*, *TnsD*, and *TnsE*) (Craig, 1991). The transposase is formed by *TnsA*+*TnsB*, whereas *TnsC* acts to activate and recruit *TnsA*+*TnsB* to a target site selected and captured by *TnsD* or *TnsE* (Craig, 1991). Some Tn7-like transposons have even recruited CRISPR-Cas associated proteins (Peters et al., 2017), which have been recently shown to function as RNA-guided transposases (Klompe et al., 2019; Strecker et al., 2019).

Autonomous and Non-Autonomous Transposons

There is an additional distinction within a given DNA transposon family that is based on the ability to mobilize independently, described as autonomous or non-autonomous transposons. Autonomous transposons, like those described in the preceding sections, encode all the factors that are required for DNA transposition. Conversely, non-autonomous transposons cannot mobilize themselves and require the transposase protein encoded from an autonomous transposon for mobilization. Non-autonomous are often internally-deleted transposon copies, that arise from illicit transposition, or incomplete homologous recombination events (Engels et al., 1990). Numerous examples of non-autonomous elements exist across DNA transposon superfamilies and include the maize *Dissociation* (*Ds*) element (which requires the autonomous *Activator* (*Ac*) element for mobility) and the KP element, an internally-deleted derivative of the *Drosophila* P element. Interestingly, the KP element encodes a severely truncated copy of P element transposase. While the KP element cannot mobilize independently, the encoded KP protein acts as a transpositional repressor by mechanisms similar to those described for the naturally occurring 66kD repressor protein (Lee et al., 1998).

Host-Encoded Factors

While autonomous DNA transposons encode all the factors required for mobilization, they can also make use of cellularly-encoded host proteins. The assembly of a transposase with the transposon ends is highly regulated and can require or induce severe DNA distortions. As such, transposition is often stimulated by cellular DNA-bending proteins. For instance, it is known that the prokaryotic HU and IHF DNA-bending proteins are required for phage Mu transposition (Harshey, 2014). Similarly, the eukaryotic DNA-bending proteins, HMGB1 or HMGB2, are required for proper function of V(D)J recombinase and are thought to facilitate binding to and cleavage of the DNA recombinase signal sequences by stabilizing DNA distortions (Schatz and Swanson, 2011). Indeed, HMGB1 is observed to promote a nearly 90° bend within the 23 bp spacer recombinase signal sequence based on the recent structures of the RAG1/RAG2 nucleoprotein complexes (Kim et al., 2018; Ru et al., 2015). A similar stimulation of transposition by HMGB1 is observed with the lancelet *ProtoRAG* transposase, the transposable element from which V(D)J recombinase is believed to have originated (Huang et al., 2016).

Furthermore, both the excision and integration stages of “cut-and-paste” DNA transposition can generate DNA nicks, gaps or breaks that must be repaired by host cell factors. In *Drosophila*, a basic leucine zipper Xrp1/IRBP18 heterodimeric complex (IRBP complex), binds site-specifically to the P element transposon terminal inverted repeats and promotes repair of the resulting double-strand break at the donor site after P element excision (Francis et al., 2016). P element mobilization in IRBP18-null flies results in increased larval and pupal lethality, presumably from DNA break repair defects, further implicating the IRBP complex in genome stability and repair after P element excision (Francis et al., 2016). Another example in which host factors facilitate transposition occurs during the replicative pathway of bacteriophage Mu. Replicative transposition of Mu generates a transposition product that resembles two stalled DNA replication forks (Harshey, 2014). The post-transposition Mu-DNA transposome complex actively recruits cellular factors that are required to disassemble the transposome, degrade the transposase and initiate stalled replication fork restart (Burton and Baker, 2003; Harshey, 2014).

Mechanisms of DNA Transposition: Transposase DNA Binding and Synaptic Complex Formation

DNA transposition often proceeds through a sophisticated series of assembly reactions between the DNA transposon and the transposase protein. The finer details of transposition can be specific to each DNA transposon family, however, the process can be described by seven fundamental steps: transposase-transposon DNA binding, pairing of the transposon ends or synaptic complex formation, donor DNA cleavage, target DNA capture, strand transfer or integration, disassembly and DNA repair.

Transposition is initiated when the transposase protein recognizes the transposon DNA ends. This generally occurs through site-specific DNA-binding domains and DNA binding sites at or near the transposon end. The transposase will then proceed to synaptic complex formation in which the transposon DNA ends are paired and organized to be catalytically engaged within the transposase. Synaptic complex formation can have many requirements in addition to the transposon terminal inverted repeats and is often highly regulated. For instance, under physiological conditions Mu transposome assembly requires the left and right transposon ends in a supercoiled configuration, an internal enhancer sequence, and the *E. coli* DNA bending proteins, HU and IHF (Harshey, 2014). This level of regulation and complexity ensures proper Mu DNA engagement and transposome assembly before the cleavage reactions begin.

Mechanisms of DNA Transposition: Transposon Cleavage

After proper synaptic complex formation the transposase protein will catalyze excision of the transposon DNA from the surrounding host genomic DNA. While different "cut-and-paste" DNA transposases have adopted different excision pathways, excision generally involves successive nucleophilic S_N2 in-line attacks of the phosphodiester backbone to generate a double-strand break (Hickman and Dyda, 2016)(Figure 1.2). The chemistry of the DNA cleavage reaction is well understood within the context of two-metal ion catalysis. Three acidic residues in the RNase H-like active site coordinate two divalent metal ions which orient and activate a water molecule for nucleophilic attack on the phosphorous atom of the scissile phosphate (Hickman and Dyda, 2016; Yang et al., 2006). The resulting hydrolysis of the DNA strand generates a nicked DNA duplex, with a 3' – OH and a 5' – phosphate at the cleavage site (Hickman and Dyda, 2016). This water mediated "nicking" reaction is the first step and is common to all DNA transposases (Figure 1.2).

Some transposases, like Mu, will nick only one strand in each DNA duplex and proceed directly to integration (see footnote 1). Whereas other transposase systems, such as *Hermes*, *piggyBac* or Tn5, will use the newly generated 3' – OH group as a nucleophile to cleave the opposing DNA strand, in a transesterification reaction that generates a hairpin intermediate. In the cleavage reactions, the order and orientation of strand cleavage dictates whether the hairpin is generated on the host DNA, or at the transposon end and if the hairpin is opened by a third S_N2

nucleophilic attack. An important distinction is made between the two DNA strands at the transposon end. If the DNA strand becomes covalently joined to the target DNA during the next stage of transposition, it is called the "transferred strand," and if it is not, it is called the "non-transferred strand." If the transferred strand is "nicked" in the first step, the second step will generate a hairpin on the transposon end. Alternatively, if the non-transferred strand is nicked first, a hairpin is instead generated on the flanking host DNA. Hairpins at the transposon end must be resolved (opened) for transposition to proceed and is achieved through another transposase catalyzed nucleophilic attack by an activated water molecule. Hairpins on the flanking host DNA are not opened by the transposase but are instead repaired by cellular factors. Nonetheless, it is always the 3'-OH group at the transposon end, generated by donor DNA cleavage, that attacks the target DNA during strand transfer.

Some transposases, such as phage Mu, Tn7 or *Mos1*, do not generate a hairpin intermediate, but instead proceeded directly to strand transfer after cleaving the transferred strand (Mizuuchi, 1992; Mizuuchi and Craigie, 1986), cleave the second DNA strand with a second active site (Craig, 1991) or are thought undergo large structural changes to reposition the active site over the uncleaved second (non-transferred) strand (Richardson et al., 2006).

The position at which the DNA strands are cleaved varies and is dependent on the exact transposase superfamily. While the transferred strand is always precisely cleaved at the transposon end, the non-transferred strand can be cleaved inside of, at the end of, or outside of the transposon sequence. For instance, *piggyBac* cleavage leaves a 4 nucleotide TTAA 5' overhang on the non-transferred strand (Mitra et al., 2008). Conversely, *P* element transposase cleaves the non-transferred strand well into the transposon end generating a 17 nt 3' overhang at the transposon ends (Beall and Rio, 1997).

Mechanisms of DNA Transposition: Target Capture and Strand Transfer

Regardless of the pathway, cleavage liberates the transposon DNA from the surrounding DNA[1]. The transposon-transposase nucleoprotein complex will then capture an appropriate target DNA substrate and initiate strand transfer[2]. While some DNA transposase display no to low target site specificity (Goryshin et al., 1998), others can display a much higher preference for a target sequence or can even be exquisitely site-specific (Craig, 1997). For instance, *Drosophila* P elements preferentially integrate into a 14 bp palindromic target sequence motif (Linheiro and Bergman, 2008). Likewise, *piggyBac* specifically targets TTAA sequences (Ding et al., 2005; Li et

[1] Replicative DNA transposases, such as prokaryotic Tn3, do not generate double strand breaks at the transposon end. They instead cleave only the transferred strand and use the resulting 3'-OH directly in the strand transfer reaction. Replicative transposition generates a complex "*theta*" shaped DNA intermediate, (sometimes called a "Shapiro intermediate" (Shapiro, 1979), that must be resolved by the cellular replication machinery. Similarly, bacteriophage Mu uses replicative transposition during the phage lytic phase (Nakai et al., 2001). The replicative transposition pathway appears to be restricted to prokaryotes (Hickman and Dyda, 2016).

[2] This is not explicitly true, as Tn7 initiates transposon excision only after capturing an appropriate target DNA site (Craig, 1997).

al., 2005), in fact, flanking TTAA sequences play a key role in transposon excision (Mitra et al., 2008). Bacterial Tn7 also displays high target site specificity, transposing into a single location upstream of the *GlmS* gene, termed *attTn7*, within the *E. coli* genome (Craig, 1997). Tn7 uses a transposon-encoded protein, *TnsD*, to direct transposition into *attTn7*. As mentioned above, it is now known that some Tn7-like transposon have recruited CRISPR-Cas like proteins to select a specific target site by an RNA-guided mechanism (Klompe et al., 2019; Peters et al., 2017; Strecker et al., 2019).

There is emerging evidence that target DNA flexibility plays an important role in transposon target site selection (Fuller and Rice, 2017). Target DNA bending is thought to optimize protein-DNA contacts and facilitate positioning of the scissile phosphate into the transposase active site (Fuller and Rice, 2017; Wright et al., 2017). The contribution of target DNA flexibility to target site selection can be significant. For example, Mu transposase will integrate into a mismatched target even in the presence of a 300,000 fold excess on non-mismatched sites (Yanagihara and Mizuuchi, 2002). Similarly, *Drosophila Mos1* exhibits a higher affinity for and is stimulated by nicked target sites, presumably due to increased target DNA flexibility (Pflieger et al., 2014). Consistent with target DNA flexibility, severe target DNA distortions have been observed within the structures of both transposases and the related retroviral integrases (Maertens et al., 2010; Montaño et al., 2012; Morris et al., 2016). Distorted DNA at the target site may facilitate strand transfer and prevent reversal of the reaction.

The next step in transposition involves the catalytic joining of the transposon DNA ends to the captured target DNA in an integration reaction called “strand transfer.” Strand transfer proceeds through the nucleophilic attack on the target DNA scissile phosphates by the 3'-OH groups on the transposon transferred DNA strands. Each transferred strand becomes covalently joined to opposite strands of the target DNA in a staggered fashion. Under physiological conditions strand transfer is effectively irreversible, and the resulting nucleoprotein strand transfer complex (STC) is typically very stable. For instance, the Mu transposase STC is resistant to challenge by denaturants (6M Urea), high temperatures (75°C) and high salt (2 M NaCl) (Surette et al., 1987). After strand transfer, the transposon DNA-target DNA phosphodiester is often ejected out of the transposase active site. This conformational change is thought to prevent reversal of the strand transfer reaction and has been observed in the *Mos1* STC (Morris et al., 2016) and the related retroviral integrase STC structures (Maertens et al., 2010).

Mechanisms of DNA Transposition: Disassembly and Repair

The generation of new phosphodiester bonds between the target and transposon DNA strands is concomitant with DNA strand breaks at the target site. Transposition is completed after transposase disassembly and DNA repair by cellular host factors. The unusually high stability of the product strand transfer complex suggests that the complex does not simply fall apart after transposition but rather proceeds through a disassembly pathway, likely involving host factors. The disassembly and repair pathways have been extensively studied for Mu transposase, which actively recruits the molecular chaperone/protease ClpXP after transposition (Harshey, 2014). ClpXP belongs to a family of ATPases known to remodel and degrade multisubunit complexes and initiates disassembly and repair by selectively destabilizing and degrading the product Mu STC (Harshey, 2014).

Cellular factors, such as DNA polymerases and ligases, are often involved in repairing the DNA gaps and nicks left at the sites of transposition. DNA synthesis across the staggered integration event gives rise to direct duplications of host DNA flanking each transposon end, a feature that is characteristic of DNA transposon mobility. The length of the target site duplication (TSD) is specific to each transposon family and can range from 2 to 12 base pairs (Hickman and Dyda, 2016). Some DNA transposases use a unique excision mechanism and bypass DNA synthesis to generate a target site duplication (Mitra et al., 2008).

Description of Prokaryotic and Eukaryotic DNA Transposases

The details of many prokaryotic and eukaryotic DNA transposases have been used as illustrative examples in the preceding sections. Therefore, only salient features will be summarized in the following section. A special focus will be given to *Drosophila* P element transposition.

Transposon Tn5

Tn5 is one of the simpler prokaryotic DNA transposons belonging to the “cut-and-paste” class of mobile elements. The Tn5 transposon is a composite transposon, in which two inverted insertion sequence elements (IS50) flank a central region that often encodes antibiotic resistance genes (*e.g.* kanamycin resistance). The IS50 sequences encode the Tn5 transposase, which can transpose the IS50 sequences independently, or the entire composite Tn5 transposon as a unit (Reznikoff, 2008). Tn5 is an auto-inhibited monomer in solution but dimerizes upon DNA binding to initiate transposon cleavage (Reznikoff, 2008). Cleavage occurs in *trans,* generating a blunt end through a hairpin intermediate on the transposon end (Bhasin et al., 1999). Strand transfer into an appropriate target sequence generates a 9bp target site duplication. The generally low target sequence specificity and development of a hyperactive system with mosaic OE/IE ends has led to the use of Tn5 transposase many high-throughput next generation sequencing applications (Adey et al., 2010; Buenrostro et al., 2015; Kaya-Okur et al., 2019).

Bacteriophage Mu

DNA transposition plays a key role in the life cycle of bacteriophage Mu. Upon infection, the phage employs two encoded proteins, MuA and MuB, to randomly integrate its ~36 kb genome (Morgan et al., 2002) into the host DNA by a mechanism akin to replicative DNA transposition. The transposase MuA pairs the left and right Mu DNA ends in a sophisticated assembly pathway that requires the Mu left and right DNA ends in a supercoiled configuration, multiple terminal MuA binding sites, an internal enhancer site and host DNA bending proteins. MuA monomers assemble on the Mu DNA ends to form an active MuA tetramer that will nick at the Mu DNA ends, generating 3'-OH groups and transpose these ends into an appropriate target site delivered by the MuB protein. MuB is a AAA+ ATPase that binds DNA non-specifically to capture and deliver a target DNA to MuA. MuB is also aids MuA assembly, stimulates MuA activity and prevents transposition into or near Mu DNA, a process termed Mu genome immunity or *cis*-immunity, respectively. Flanking non-Mu DNA is acquired during packaging of the Mu genome from the previous host and is degraded after transposition into a target DNA in the new host genome. Repair of the staggered transposition products generates a 5 bp target site duplication.

During the lytic phase of the Mu life cycle replicative DNA transposition is used to amplify its genome at least 100-fold. Transposition of the prophage genome proceeds in a similar fashion, however, the DNA repair pathways differ. During the lytic phase Mu actively recruits replication restart factors and invokes the host replication machinery to synthesize across the Mu genome and thereby increase the Mu DNA copy number.

The structure of the Mu transposome represents the first transposase nucleoprotein product complex structure (Mu transposase in complex with bacteriophage DNA ends and target DNA) (Montaño et al., 2012). The transposome structure shows a tetrameric protein assembly highly intertwined with the two Mu DNA ends and the target DNA. Two MuA subunits are catalytically engaged with the target DNA and the R1 sites of the Mu donor DNAs, and the two other subunits are engaged with the R2 Mu DNA sites and stabilize the overall complex through protein-protein interactions. The domains of each MuA subunit assume different roles within the transposome structure. Domain Iβ binds either the R1 or R2 site in a subunit-dependent manner and domain IIIα either stabilizes the bent target DNA or stabilizes the complex by wrapping around the catalytic subunits, also in a subunit-dependent manner. Overall the Mu transposome structure emphasizes the complex and sophisticated transposition assembly pathway and the complexity of DNA-protein contacts and subunit architecture in synaptic complex formation.

Mos1 (Tc1/mariner)

Tc1/*mariner* elements are found across a diverse range of taxa and are likely one of the most widely distributed transposable elements found in nature (Munoz-Lopez and García-Pérez, 2010). *Mos1* is one of the few active Tc1/*mariner* elements and the first to be identified, isolated from *Drosophila mauritiana* (Hartl, 2001). The *Mos1* transposon is ~1.3 kb in length, possesses 28 bp imperfect terminal inverted repeats and is flanked by a TA dinucleotide target site duplication. *Mos1* transposase is thought to exist as an extended dimer in solution (Cuypers et al., 2013) and consists of two helix-turn-helix DNA-binding domains connected by an interdomain linker and an RNase H-like catalytic domain at the C-terminus (Tellier et al., 2015). Cleavage of the non-transferred strand occurs first and is recessed ~3 nucleotides into the transposon end (Dawson and Finnegan, 2003). The transferred strand is then cleaved precisely at the transposon end but does not occur through a hairpin intermediate (Dawson and Finnegan, 2003). Several models have been proposed for the mechanism of *Mos1* transposition. Although recent studies have illuminated the assembly pathway (Cuypers et al., 2013) and the order of strand cleavage, the exact mechanism by which a single RNase H-like active site cleaves opposite strands without a hairpin intermediate is unclear (Bouuaert et al., 2014). The proposed models invoke dimeric, subunit exchange or tetrameric states and involve reorganization of the transposome, reorientation of the transposase active site or DNA conformational changes (Bouuaert et al., 2014; Richardson et al., 2009).

After donor DNA cleavage *Mos1* specifically inserts into TA dinucleotide target sites with a 2 bp stagger (Tellier et al., 2015). Recent crystal structures of transposon-transposase complexes revealed a sharply bent target DNA in which both adenosines of the TA dinucleotide are flipped out into extrahelical positions and recognized by base-specific protein-DNA interactions (Morris et al., 2016). A staggered integration into the target DNA TA dinucleotide generates the 2 bp TA target site duplication flanking the new transposon insertion. The flanking TA dinucleotide TSDs also appear to play a role during the earlier cleavage reactions. Structures of a *pre*-second strand cleavage intermediate show that the flanking TA dinucleotides are recognized by base-specific amino acid interactions, partially through the interdomain linker and are thought to correctly position the second strand for cleavage (Dornan et al., 2015).

Sleeping Beauty, an ancient and previously inactive member of the Tc1*/mariner* family, was reconstructed from salmonids using accumulated phylogenetic data (Ivics et al., 1997) and has garnered much attention as a gene transfer tool in a wide range of organisms (Narayanavari et al., 2017).

Hermes (hAT elements)

Hermes belongs to a broader family of mobile elements, termed *hAT* elements, named for the *h̲obo* element in *Drosophilids*, the *A̲ctivator* elements of *Zea mays*, and the *T̲am3* element from *Antirrhinum majus*. An endogenous *hobo*-like transposase activity led to the discovery of *Hermes* in the common house fly, *Musca domestica* (Warren et al., 1994). The *Hermes* element is ~2.7 kb long, possesses short imperfect 17 bp terminal inverted repeats and like other *hAT* elements generates 8 bp target site duplications upon insertion. The element contains a single ORF encoding a 612 amino acid protein homologous to the *Drosophila hobo* transposase (Warren et al., 1994).

In vitro characterization of *Hermes* transposase revealed that transposition proceeds through a hairpin intermediate generated on the host DNA ends (Zhou et al., 2004). *Hermes* transposase is a four domain protein, consisting of an N-terminal predicted zinc-binding BED domain (nonspecific DNA-binding), an intertwined dimerization/DNA-binding domain, an RNase H-like catalytic domain, and an α-helical insertion domain. Structural characterization showed that the transposase is an octamer, with the overall assembly described as a tetramer of dimers arranged in a ring (Hickman et al., 2014). Dimerization is achieved through an intertwined, domain swapped N-terminal domain while tetramerization occurs through interactions between adjacent insertion domains (Hickman et al., 2005; 2014). Octamerization of *Hermes* is required for activity *in vivo* and it is thought that the six noncatalytic subunits of *Hermes* facilitate binding to several subterminal inverted repeats found at each transposon DNA end (Hickman et al., 2014).

RAG1/RAG2 (V(D)J recombinase)

While not strictly a transposase, the recombinase-activating genes (RAG1 and RAG2) and the recombinase signal sequences(RSSs) of jawed vertebrates appear to have been domesticated from a transposon/transposase from the *Transib* family (Kapitonov and Jurka, 2005). The RAG1/RAG2 V(D)J recombinase orchestrates DNA elimination/inversion and joining events at the T cell receptor and immunoglobulin loci, which gives rise to the diversity and plasticity of the vertebrate adaptive immune system. The RAG1/RAG2 complex cleaves DNA precisely at the RSSs flanking the variable (V), diversity (D), and joining (J) gene segments. The RSSs have conserved heptamer and nonamer sequences separated by a non-conserved 12 bp or 23 bp spacer (Fugmann et al., 2000; Gellert, 2002; Ru et al., 2018b).

V(D)J recombination will only occur between one 12-RSS and one 23-RSS, termed the 12/23 rule (Fugmann et al., 2000; Gellert, 2002). The strict requirement of differently spaced RSSs is reminiscent of the P element left 5' and right 3' transposon ends (Beall and Rio, 1997). An induced asymmetry in a flexible DNA-binding domain (nonamer binding domain) is the molecular basis for the 12/23 rule; binding to either RSS induces an asymmetry, such that only a differently spaced RSS can be accommodated (Kim et al., 2015; Lapkouski et al., 2015). Assembly of the RAG1/RAG2 complex upon the 12/23 RSSs induces severe DNA bending (nearly 90°), which is facilitated by HMGB1/2 DNA bending proteins (van Gent et al., 1997).

Recent structural characterization has illuminated the molecular mechanisms at each stage of V(D)J recombination in unprecedented detail (Kim et al., 2015; 2018; Ru et al., 2015; 2018a). Like *Hermes* transposase, RAG1/RAG2 cleaves at the RSSs through a transesterification mechanism, that generates a hairpin intermediate at the coding flank segment (analogous to the flanking host DNA) (van Gent et al., 1996; Zhou et al., 2004). Cleavage at the RSSs involves significant conformational rearrangements in both the protein and the DNA (Ru et al., 2018b). For instance, a nearly 180° rotation accompanied by DNA melting and interstitial base staking occurs at the coding flank to position the first scissile phosphate into the RAG1 active site (Kim et al., 2018; Ru et al., 2018b). These conformational changes further underscore the complexity and plasticity of transposases and transposase-related proteins.

Although RAG1/RAG2 can mediate transposition *in vitro* (Agrawal et al., 1998; Hiom et al., 1998), transposition is severely limited *in vivo* (Chatterji et al., 2006; Zhang et al., 2019).

P element transposase

The P element is thought to have entered the *D. melanogaster* genome sometime in the early 20th century (Kidwell et al., 1977). Despite the negative consequences of mobilization, this element rapidly spread throughout all wild *D. melanogaster* populations on every continent within 30-40 years (Engels, 1992; Kidwell, 1992). This relatively recent invasion by the P element has afforded researchers a rare opportunity to study horizontal transfer in related species (*D. simulans*) (Kofler et al., 2015) as well as the mechanisms that drive transposon adaptation (Khurana et al., 2011).

In the early 1980s, 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, spontaneous chromosomal recombination, malformed gonads and sterility (Bingham et al., 1982; Kidwell, 1992; Kidwell et al., 1977). Crosses between males from newly collected wild strains and females from long established laboratory stocks produced offspring that manifest the hybrid dysgenesis phenotype (Kidwell et al., 1977). After their initial discovery as mobile elements, P elements were engineered as a critically important tool for *Drosophila* molecular genetics and germline transformation (Majumdar and Rio, 2015). P elements also served as a model system for understanding DNA repair mechanisms (Sekelsky, 2017), the role of PIWI-interacting small RNA pathways that drive transposon adaption and limit transposon mobility (Khurana et al., 2011; Teixeira et al., 2017), and for identifying RNA binding proteins as regulators of tissue-specific alternative splicing (Laski et al., 1986; Siebel et al., 1992).

The 2.9 kb full-length P element possesses 31 bp perfect terminal inverted repeats, 10 bp internal transposase binding sites and internal 11 bp subterminal inverted repeats. The left 5' and right 3' ends differ in the spacing between the terminal inverted repeat and the 10 bp transposase binding site, 9 bp and 21 bp, respectively. This spacing is reminiscent of the 12 and 23 RSSs of V(D)J recombinase (Beall and Rio, 1997). The element-encoded transposase gene is punctuated by three introns and undergoes tissue-specific alternative splicing to produce two protein isoforms (Laski et al., 1986). Alternative splicing of the third intron restricts full-length P element transposase production to germline cells (Laski et al., 1986; Rio et al., 1986). Retention of the third intron in somatic cells (and also to a large extent in the germline) produces an mRNA that encodes for a 66kDa transpositional repressor protein (Rio et al., 1986).

Much effort has gone into the biochemical characterization of P element transposase and into uncovering the mechanism of transposition (Figure 1.3). Purification and characterization of the transposase protein from *Drosophila* tissue culture nuclear extracts showed that P element transposase binds to internal 10 bp sites found at each end of the transposon (Kaufman et al., 1989) and that a guanosine triphosphate co-factor (GTP) was required for *in vitro* transposition activity (Kaufman and Rio, 1992). Following the initial recognition of a single transposon end, P element transposase captures and pairs the second end in a GTP-dependent manner (Tang et al., 2005). Uncoupled cleavage at each P element end liberates the transposon from the flanking host genome (Tang et al., 2005; 2007). Like other transposable elements, the 3' cleavage site occurs at the end of the P element DNA, but top strand 5' 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 (Beall and Rio, 1997). The order and mechanism of strand cleavage is not currently known, however atomic force microscopy volume measurements indicate that a tetrameric form of transposase may be involved in the initial assembly steps (Tang et al., 2007).

After donor DNA cleavage the transposon-transposase nucleoprotein complex will capture and integrate into an appropriate target site. Transposition preferentially occurs into nearby target sites on the same chromosome (~50 -150 kb away) in a phenomenon termed “local hopping” (Tower et al., 1993). The sites of transposition are separated by 8 bp, which gives rise to the 8 bp target site duplications (TSDs), after transposome disassembly and DNA repair. Although P element transposition is not site-specific, a target sequence consensus motif was derived from over 10,000 accurately mapped P element insertions from the *Drosophila* genome project (Linheiro and Bergman, 2008). Although the disassembly and DNA repair mechanisms at the target site have not been investigated, it is understood that the double-strand break at the donor site can be repaired through both homologous recombination-dependent (HR) or non-homologous end joining (NHEJ) pathways involving IRBP18/Xrp1, Ku70/80 and the *Drosophila* Bloom's syndrome helicase homolog (Francis et al., 2016; Min et al., 2004; Sekelsky, 2017; Weinert et al., 2005).

The difficulty in expressing and purifying active P element transposase has previously precluded detailed structural analysis. Thus, many of the structural details of P element transposase, such as the domain organization or the location of catalytic residues, were not known. Several details emerged from the characterization of the KP repressor protein, a truncated version of P element transposase that is readily purified and refolded from *E. coli* inclusion bodies (Lee et al., 1996). These studies localized the DNA binding domain to an 98 amino acid N-terminal C_2HC zinc-coordinating domain called a THAP domain and confirmed that an adjacent leucine zipper domain was responsible for dimerization. It was thought that P element transposase would possess a noncanonical GTP binding domain (Mul and Rio, 1997) and an RNase H-like catalytic fold, however the locations were not known.

Several mechanistic features distinguish P element transposition from the other characterized "cut-and-paste" DNA transposases. Namely, the requirement of GTP for the pairing, donor cleavage and strand transfer reactions (Beall and Rio, 1998; Kaufman and Rio, 1992; Tang et al., 2005), and the unusually long 17 nt staggered cleavage at each P element end (Beall and Rio, 1997). To understand the mechanisms underlying the unique features of the P element transposase superfamily, we prepared and characterized protein-DNA transposition complexes and used cryo-electron microscopy (cryo-EM) to determine the structure of the P element transposase strand transfer complex (STC) at 3.6 Å resolution. This post-transposition nucleoprotein complex contains transposase and cleaved P element ends covalently joined to a target DNA in a dimeric assembly where four identifiable domains are closely intertwined with the transposon DNAs. Most unusually, the terminal single-stranded transposon DNA adopts unusual A-form and distorted B-form helical regions that are stabilized by extensive protein-DNA interactions with 4 major protein domains. Additionally, the bound GTP cofactor interacts via hydrogen bonding 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 superfamily of eukaryotic transposases, offers new insights into P element transposition and implies a transposition pathway mechanistically and fundamentally distinct from other cut-and-paste DNA transposases. These results are discussed in detail in Chapter 2.

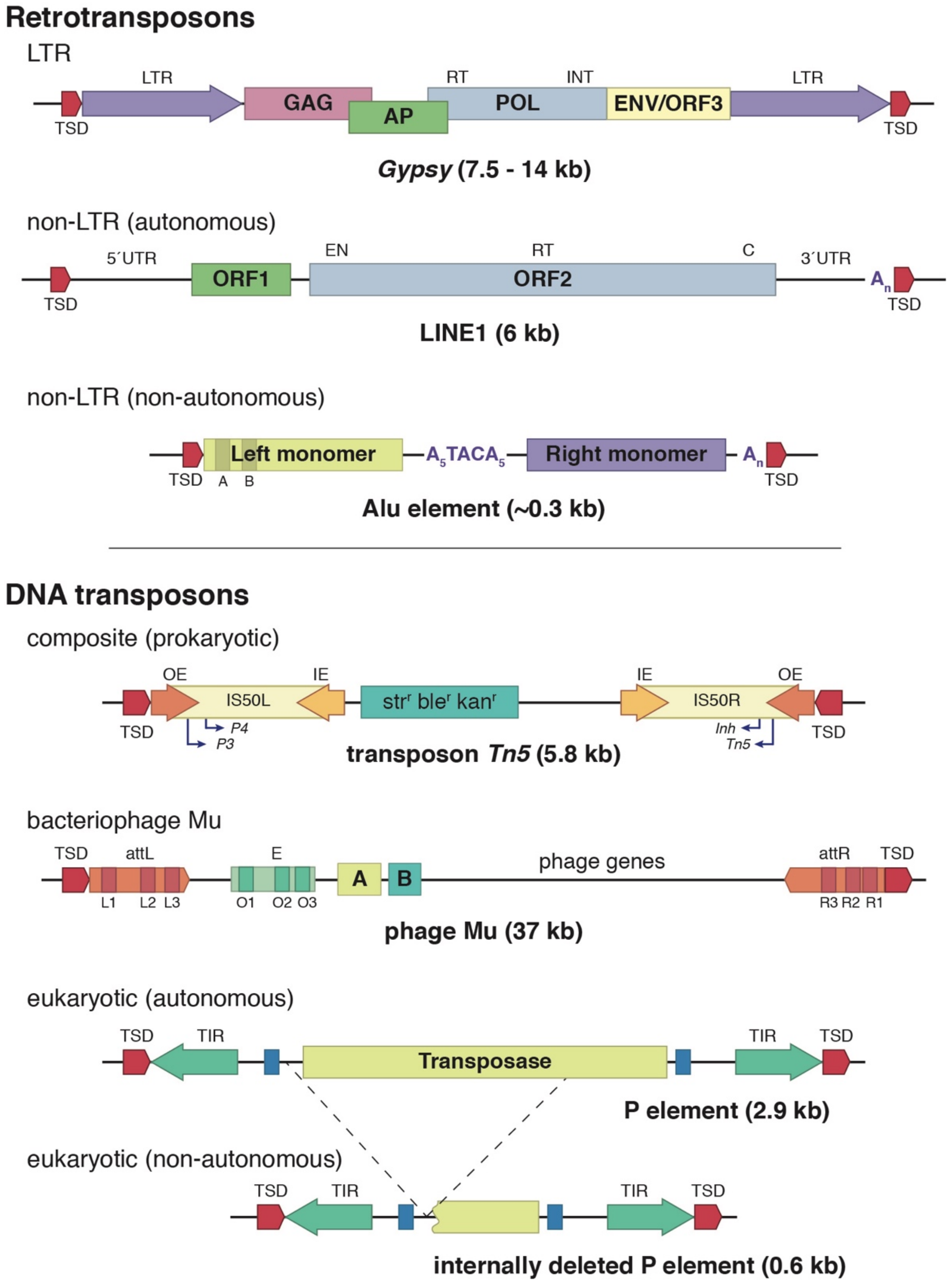
THAP9, a Domesticated P element Transposase

Like the adaptation of an ancient *Transib* element into V(D)J recombinase, a similar domestication process is thought to have occurred with P element transposase (Quesneville et al., 2005). The Thanatos-associated protein, or THAP, domain is a well-conserved and common C_2CH zinc-coordinating DNA-binding domain found at the N-terminus of *Drosophila* P element DNA transposase (Roussigne et al., 2003b; Sabogal et al., 2010). Surprisingly, the THAP DNA-binding domain is not only found in P element transposase, but in many cellular proteins across a wide range of taxa, such as *C. elegans, Drosophila* and vertebrates (Roussigne et al., 2003b). For instance, 12 THAP domain-containing proteins (THAP0 - THAP11) have been identified in the human genome and have been shown to play roles in diverse cellular functions including apoptosis (Roussigne et al., 2003a), histone deacetylation (Macfarlan et al., 2005) and maintenance of mouse embryonic stem cell pluripotency (Dejosez et al., 2008). The THAP9 family of proteins, in particular, exhibits a high degree of homology along the entirety of *Drosophila* P element transposase (~25% identity and 40% similarity), yet the THAP9 gene generally lacks the characteristics of a mobile element. THAP9-like P element transposons have been identified in *Ciona intestinalis,* an species of the most basal chordate lineage (Kimbacher et al., 2009) and *Danio rerio* (Hagemann and Hammer, 2006; Hammer, 2005), and display the hallmarks of mobility. This suggests that the cellular THAP9 gene may have been domesticated in early chordates from this DNA transposable element.

Furthermore, human THAP9 has retained DNA transposase catalytic activity, because it was shown to mobilize a genetically marked *Drosophila* P element in both *Drosophila* and human cell lines (Majumdar et al., 2013). DNA transposition activity is unusual among human cellular proteins and thus far only V(D)J recombinase, PGBD5, THAP9 have been identified as functional DNA transposases (Agrawal et al., 1998; Henssen et al., 2015; Majumdar et al., 2013). However, unlike V(D)J recombinase, a cellular function and/or true DNA recombination sites for PGBD5 and THAP9 have not been identified.

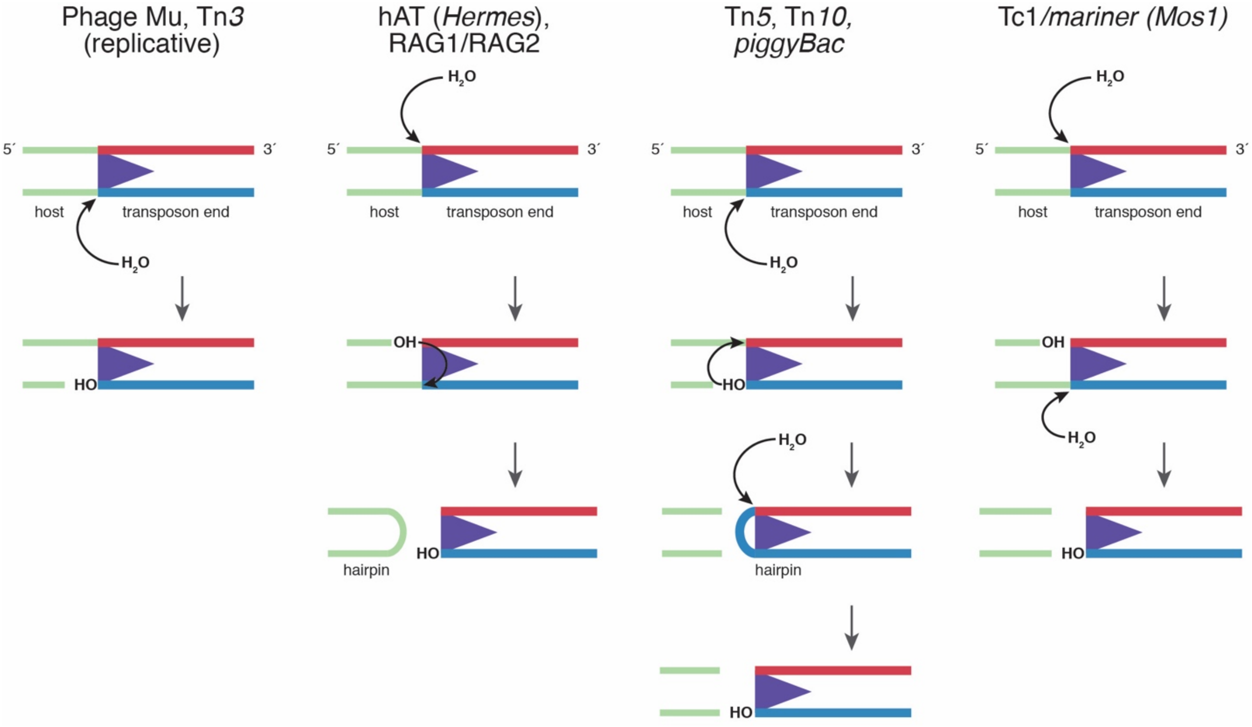
To elucidate the genomic DNA binding sites and cellular function of the human THAP9 gene, we generated a homozygous knockout human embryonic stem cell line, and carried out BLESS (direct in situ breaks labeling, enrichment on streptavidin and next-generation sequencing) to identify THAP9-induced genomic break sites. We also generated antibodies to human THAP9 and carried out Chromatin Immunoprecipitation and next-generation sequencing (ChIP-seq) to identify THAP9 DNA binding sites in the human genome. These results are discussed in detail in Chapter 3.

Figure 1.1



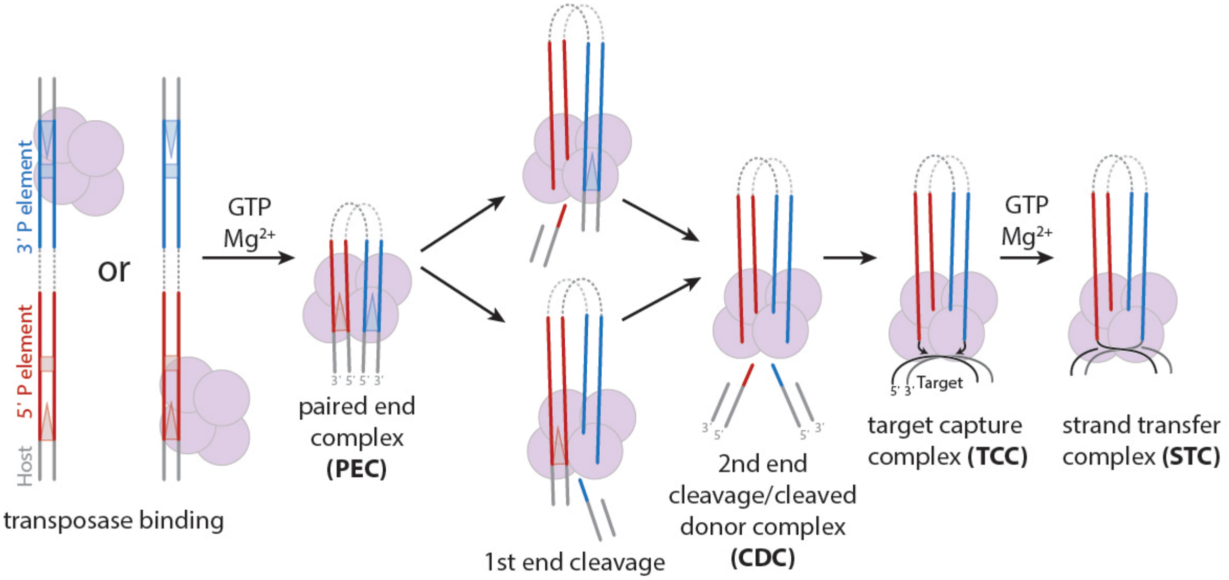
Organization of representative transposable elements.
Schematic examples of representative mobile genetic elements. Target site duplications (TSD) are colored red, long terminal repeats (LTR) are colored purple, and terminal inverted repeats (TIR) are colored green. Elements are not drawn to scale.

Figure 1.2


Phage Mu, Tn*3* (replicative)
hAT (*Hermes*), RAG1/RAG2
Tn*5*, Tn*10*, *piggyBac*
Tc1/*mariner* (*Mos1*)
5´
3´
host
transposon end
H_2O
OH
HO
hairpin

First and second strand cleavage mechanisms of DDE-Transposases.
The pathways focus on a single transposon end and generalize the features of each cleavage mechanism. The flanking host DNA is depicted in light green, the transposon transferred strand in blue and the transposon non-transferred strand in red. Arrows indicate a nucleophilic attack. The transferred strand is cleaved at the transposon end, while the non-transferred strand can be cleaved inside, at, or outside of the transposon end. The transferred strand 3'-OH group is used in the subsequent strand transfer reaction. Adapted from ((Curcio and Derbyshire, 2003)).

Figure 1.3


3' P element
5' P element
Host
or
transposase binding
GTP
Mg^{2+}
3' 5' 5' 3'
paired end complex **(PEC)**
1st end cleavage
3' 5'
5' 3'
2nd end cleavage/cleaved donor complex **(CDC)**
5' 3' Target
target capture complex **(TCC)**
GTP
Mg^{2+}
strand transfer complex **(STC)**

Model of P element mobilization.
The transposase protein (purple) first assembles on a single transposon (transposase binding), and brings both the left and right P element ends together upon binding GTP (PEC). After uncoordinated excision at the P element ends (CDC), the active transposome complex captures (TCC) and catalyzes transposition (STC) into a target DNA. The transposase oligomeric state at the early reaction stages and the order and mechanism of strand cleavage are not currently known.

CHAPTER TWO

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

Based on Ghanim et al., *NSMB*, 2019

Abstract

P element transposase (TNP) catalyzes the mobility of P transposable elements within the *Drosophila* genome. Compared to other eukaryotic DNA transposases, TNP exhibits several unique properties, including the requirement for guanosine triphosphate (GTP) as a cofactor and the generation of unusually long 17 nt staggered DNA breaks during donor DNA excision. To gain mechanistic insights into these features, we determined the atomic structure of the *Drosophila* P element transposase strand transfer complex (STC) using cryo-EM (resolution of 3.6 Å). This post-transposition nucleoprotein complex contains transposase and cleaved P element ends covalently joined to a target DNA in a dimeric assembly where four identifiable domains are closely intertwined with the transposon DNAs. Most unusually, the terminal single-stranded transposon DNA adopts unusual A-form and distorted B-form helical regions that are stabilized by extensive protein-DNA interactions with 4 major protein domains. Additionally, we infer the bound GTP cofactor interacts via hydrogen bonding 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 superfamily of eukaryotic transposases, offers new insights into P element transposition and implies a transposition pathway mechanistically and fundamentally distinct from other cut-and-paste DNA transposases.

Introduction

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. While transposon mobilization can be deleterious to the host, transposable elements can function in genome evolution by generating mutations, genetic polymorphisms, driving genome rearrangements, dispersing *cis*-regulatory sequences that modify gene expression networks or by supplying coding/non-coding RNAs that can be adapted to essential cellular functions (Bourque et al., 2018; Huang et al., 2012; 2015). 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 20th century (Kidwell et al., 1977; Majumdar and Rio, 2015). 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 (Engels, 1996). After their initial discovery as mobile elements, P elements were engineered as a critically important tool for *Drosophila* molecular genetics and germline transformation (Majumdar and Rio, 2015). P elements also served as a model system for understanding DNA repair mechanisms (Sekelsky, 2017), the role of PIWI-interacting small RNA pathways that drive transposon adaption and limit transposon mobility (Khurana et al., 2011; Teixeira et al., 2017), and for identifying RNA binding proteins as regulators of tissue-specific alternative splicing (Laski et al., 1986; Siebel et al., 1992).

Since the initial discovery of P elements, and through widespread genome sequencing efforts, it is now appreciated that the N-terminal site-specific DNA binding domain of P element transposase, termed a Thanatos-associated protein or THAP domain is a very common C_2CH zinc-binding, DNA binding domain that is restricted to animal genomes (Roussigne et al., 2003b; Sabogal et al., 2010). For instance, in the human genome there are 12 THAP domain-containing genes (Roussigne et al., 2003b). While most members of the THAP family share homology with the N-terminal DNA binding and dimerization domains of *Drosophila* P element transposase, THAP9, in particular, displays extensive homology along the entire length of P element transposase. More importantly, the human THAP9 gene exhibits transposase activity that can mobilize *Drosophila* P elements, but the THAP9 locus in the human genome lacks the hallmarks of a mobile genetic element (Majumdar et al., 2013). That is, unlike active DNA transposons, the human THAP9 gene is present as a single copy, lacks terminal inverted repeats and flanking target site duplication sequences, and is present in syntenic genomic locations in divergent genomes (Hammer, 2005; Quesneville et al., 2005). The cellular function of THAP9 has yet to be identified.

The 2.9 kbp full-length P element transposon possesses 31 base pair (bp) terminal inverted repeats (TIRs), internal THAP domain binding sites, internal 11 bp inverted repeats (IIRs), and an encoded transposase gene (Kaufman et al., 1989; Mullins et al., 1989; O'Hare and Rubin, 1983). The 5' and 3' P element transposon ends differ in the spacing between the THAP domain DNA-binding sites and the terminal inverted repeats (Fig. 2.1a). 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”. Formation of a paired end complex requires a guanosine triphosphate (GTP) cofactor (Kaufman and Rio, 1992; Tang et al., 2005; 2007). Formation of this higher-order nucleoprotein complex is required for the subsequent DNA cleavage (excision) reaction in which the P element transposon is excised from flanking host DNA. Like other transposable elements, the 3' cleavage site occurs at the end of the P element DNA, but top strand 5' 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 (Beall and Rio, 1997). These staggered transposon ends are the substrate that the protein uses to integrate P element DNA into a target site.

After excision of the P element from a donor site in the host genome, the resulting nucleoprotein complex, termed the cleaved donor complex (CDC), locates, captures and integrates the transposon DNA into a target site elsewhere in the genome, followed by host cell DNA repair to complete the transposition process. Large-scale analysis of P element insertion sites from the *Drosophila* genome project 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 (Linheiro and Bergman, 2008). Integration into the central portion of the TSM, followed by disassembly and DNA repair, gives rise to the characteristic 8 bp direct target site duplication (TSD).

Among the characterized DNA transposases, P element transposase is mechanistically distinct in the requirement of a GTP cofactor and the unusually long staggered cleavage of the transposon termini (Beall and Rio, 1997; Kaufman and Rio, 1992; Tang et al., 2005). To understand the mechanisms underlying the unique features of the P element transposase superfamily, we prepared and characterized protein-DNA transposition complexes and used cryo-electron microscopy (cryo-EM) to determine the structure of the P element transposase 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 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. There are two unusual duplex regions in the P element inverted repeats, one region in which part of the 17 nt single-stranded DNA base-pairs with a melted portion of the non-transferred strand of the inverted repeats adopting an A-form like DNA geometry and a second distorted B-form helical region comprising the remainder of the 31 bp inverted repeat. To our knowledge, this unusual arrangement of both A-form and distorted B-form helical DNAs has not been observed in other nucleoprotein structures. Our findings suggest that several structural transitions and rearrangements at the P element transposon ends must occur to generate the DNA organization observed in the STC structure. In addition, we observe direct interactions between the GTP guanine base and the terminal guanosine residue of the transposon DNA. This novel interaction likely acts to position the reactive transposon DNA end into the active site, providing a rationale for the requirement of GTP in the strand transfer reaction at this stage in the reaction pathway. Our structure also revealed severe DNA bending of the target DNA at the sites of transposition. We find that the preference for the TSM target sequence is dictated largely by DNA deformability rather than extensive base specific protein-DNA contacts. Finally, we observe flexibility and asymmetry regarding the transposase N-terminal THAP DNA binding and dimerization domains, suggesting a mechanism for pairing the differently spaced 5' and 3' P element ends during synaptic complex formation. Together, these results provide insight into the unique features of the P element transposition reaction and more generally how complex the interplay of the transposase/integrase enzymes with their DNA substrates can be.

Results

Reconstituted STC represents the active form of P element transposase

To achieve high level expression and purification of P element transposase (TNP) for structural determination, we generated complete *Drosophila* codon-optimized baculovirus expression constructs. Full-length TNP with N-terminal tandem maltose-binding protein (MBP) and SUMO* solubility tags was expressed in Sf9 cells and purified using a three-step chromatography strategy (see Methods). We reasoned that the strand transfer complex (STC) would likely be most stable and therefore amenable to structural determination. To assemble the STC, we first prepared the cleaved donor complex (CDC) by incubating TNP with a minimal pre-cleaved 3' P element end donor DNA end and SUMOstar protease to remove the solubility tags (Fig. 2.S1a), in the absence of Mg^{2+} and GTP. The STC was then prepared by incubating the CDC overnight at 30°C with GTP, Mg^{2+}, and an optimized target DNA derived from the *Drosophila singed* locus, a hotspot for P element transposition (Hawley et al., 1988; Linheiro and Bergman, 2008; Roiha et al., 1988) (Fig. 2.1b and Fig. 2.S1b). Fractionation by size exclusion chromatography (SEC) of either the CDC or STC sample produced higher-order species with elution profiles distinct from the donor DNA, target DNA or the liberated solubility tags (Fig. 2.1c and Fig. 2.S1d). Analysis of the DNA from deproteinized SEC fractions revealed that the CDC fraction contained donor DNA, while the STC fraction contained a slower-mobility species, resulting from strand transfer of the donor DNA into the target DNA generating the strand transfer product DNA (stDNA) (Fig. 2.1d). 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' donor DNA covalently attached to the target (Fig. 2.1b, stDNA, and Fig. 2.S1c), a strategy used for retroviral intrasomes (Ballandras-Colas et al., 2017; Passos et al., 2017; Yin et al., 2012; 2016). Particles in negative-stained electron micrographs of STC complexes assembled on stDNA were indistinguishable from authentically generated STC (Fig. 2.S1e). To assess the biological relevance of STC samples prepared this way, we exploited a property of transposases and retroviral integrases termed “disintegration” (Au et al., 2004; Beall and Rio, 1998; Chow et al., 1992; Jonsson et al., 1993; Melek and Gellert, 2000; Polard et al., 1996). In the presence of Mn^{2+}, transposase will reverse the transesterification reactions of strand transfer, liberating the donor DNA and rejoining the target DNA strands to give products that resemble an unintegrated donor DNA and a duplex target DNA (Beall and Rio, 1998). In the presence of transposase, disintegration of the stDNA to donor DNA (dDNA) and target DNA (tDNA) was observed in the presence of Mn^{2+}, but not in the presence of Mg^{2+} (Fig. 2.1e). Minor faster migrating bands were also observed (Fig. 2.1e, asterisks), and may arise from an alternate reversal foldback pathway that had been observed for Mu transposase (Au et al., 2004) and retroviral integrases (Donzella et al., 1996). 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, however this produced mixed 3'-3', 3'-5' and 5'-5' samples decreasing homogeneity.

The STC structure is dimeric and reveals four domains in each monomer

Cryo-EM data collection of the STC on an Arctica microscope equipped with a K2 detector resulted in data set containing 253,209 particles. Collecting images with a 40°-tilted stage overcame the effects of preferential orientation (Tan et al., 2017) (Fig. 2.S2a, b, d). *Ab initio* model generation using cryoSPARC (Punjani et al., 2017) resulted in a 6 Å reconstruction. Imposing C2-symmetry improved the reconstruction to 4 Å. Subsequent RELION-3.0 (Zivanov et al., 2018) refinement further improved the cryo-EM reconstruction to 3.6 Å and 3.9 Å for the symmetrized and unsymmetrized reconstructions, respectively (Figs. 2.S2c, f) (see Methods). As would be expected at this resolution, we can see density for large side chains (Fig. 2.S2e). Ions, such as magnesium, and GTP are also identifiable (Figs. 2.2e, f). Local resolution throughout the structure is fairly uniform, ranging between 3.5 and 4 Å, except for the flexible DNA ends, which are at lower resolution (>8 Å) (Fig. 2.S2g).

Our structure reveals that the STC adopts a dimeric assembly arranged with two-fold symmetry around the stDNA (Figs. 2.2c, d and Fig. 2.S3a). 26 bp of the 40 bp target DNA are well-resolved, while the first 23 bp of each donor DNA are not well-resolved in the symmetrized reconstruction. Each monomer closely interacts with the pre-cleaved P element 31 bp terminal inverted repeat donor DNAs. The two donor DNAs adopt a 55° angle relative to each central duplex axis (Figs. 2.2c, d) and insert into the target DNA, separated by 8 bp, which gives rise to the characteristic target site duplication (O'Hare and Rubin, 1983). The target DNA is distorted and bent, as observed in other transposase and retroviral integrase structures (Maertens et al., 2010; Montaño et al., 2012; Morris et al., 2016; Passos et al., 2017; Yin et al., 2016) (Fig. 2.2d).

Transposase can be divided into six structural domains (Fig. 2.2a, b), four of which could be *de novo* modeled (see Supp. Methods). The N-terminal THAP DNA-binding domain and a majority of the following dimerization domain (Lee et al., 1996; 1998; Roussigne et al., 2003a; Sabogal et al., 2010) are not resolved in our cryo-EM reconstruction due to flexibility. Thus, our model begins with the N-terminal DNA-binding helix-turn-helix domain (HTH; blue-green), followed by a split catalytic RNase H domain (RNase H; orange) that is interrupted by a GTP-binding insertion domain (GBD; blue), and a carboxy-terminal domain (CTD; red) (Fig. 2.2a-d). The linker between the RNase H and the C-terminal domain (residues 570 to 616) is not visible in the density map (Fig. 2.2c, left, white asterisks), consistent with the high probability for disorder in this region (Dunker et al., 2001) (Fig. 2.S3a). However, the orientation of the sparse density at the beginning and end of this linker suggests that the depicted RNase H and C-terminal domain are connected to constitute a monomer (Fig. 2.2c, left, white asterisks, and Fig. 2.S3b, c).

The RNase H domain conforms to the canonical RNase H fold and includes an active site similar to that in other DDE transposases and the related retroviral integrases (Hickman et al., 2010) (Fig. 2.2d, left). This similarity allowed the identification of the three catalytic acidic residues D230 located on β1, D303 after β4, and E531 on α4 (Fig. 2.2f), in agreement with previous computational predictions (Yuan and Wessler, 2011). Indeed, alanine substitution of any one of these acidic residues eliminates P element transposase excision activity *in viv*o (Beall and Rio, 1996), confirming their essential role for transposase catalytic activity (Fig. 2.S4a, b). The RNase H domains are located near the donor-target DNA junctions, with the catalytic D230, D303, E531 residues coordinating a Mg^{2+} ion (Fig. 2.2f). However, the scissile phosphate of the target DNA at the donor-target junction is rotated out of the active site (Fig. 2.2g, cyan phosphate). Since this is a product complex, this rotation may occur to prevent reversal of the integration reactions. A similar configuration of a donor-target DNA junction was observed with the PFV retroviral integrase strand transfer complex (Maertens et al., 2010).

The three additional domains of P element transposase, a previously unrecognized helix-turn-helix domain, the GTP-binding domain, and the C-terminal domain, all participate in protein-DNA interactions. The helix-turn-helix domain directly contacts the donor DNA (Fig. 2.3b, see below). 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 (Fig. 2.S4c-d). The GTP-binding domain packs against the RNase H fold and extends a loop to contact the central region of the target DNA (see below). The C-terminal domain contains many acidic residues and two predicted disordered regions (Dunker et al., 2001) (Fig. 2.S3a, b, d). The remaining 17 residues of the unmodeled C-terminus contain multiple basic residues and is ideally positioned to electrostatically interact with the target DNA (Fig. 2.S3d). As with the GTP-binding domain we also observe protein DNA contacts between the C-terminal domain and the stDNAs (see below).

The donor DNAs adopt an unusual, highly 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 nucleotide (nt) 3' single-stranded DNA (ssDNA) overhangs (Beall and Rio, 1997). We were able to place 12 of the 17 nts into our cryo-EM reconstruction. One unanticipated observation is the unusual configuration of the DNA at the P element end. 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. 2.3a and Fig. 2.S5a, b). The transferred strand is displaced from the non-transferred strand at nucleotide C_{-22} to accommodate the A-form duplex (Fig. 2.3a, schematic). This displaced transferred strand is stabilized by numerous contacts from the C-terminal and GTP-binding insertion domain, including aromatic base-stacking interactions from Y721, F722, F384, Y629, and Y519 (Figs. 2.3c and 2.4).

To investigate the importance of base pairing between distant strands in the donor DNA, we performed *in vitro* strand transfer assays with mutated donor DNAs substrates (Fig. 2.S5c).

Mismatches introduced into the transferred strand that disrupt base pairing at the A-form duplex region decreased or eliminated strand transfer activity at nearly all positions (Fig. 2.S5c, 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 (Fig. 2.S5c, 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 helix-turn-helix domain, which engages the donor DNA at the 31 bp terminal inverted repeats through a loop in the minor groove and an α-helix inserted into the major groove (Fig. 2.3a). Numerous backbone and base contacts are made by R154, S188, R189, T190, T191, R194, and W195 (Fig. 2.3b and Fig. 2.4). Of these, the positioning of R154, R189 and T190 lead us to infer that these sidechains form base-specific hydrogen bonds with either T_{12}, G_6 and T_7, and G_{-25} or G_{-26}, respectively.

Overall, we observe extensive protein-DNA contacts of a single subunit with both of the P element donor DNAs. The depicted protein subunit in Fig. 2.S6 (left, P element donor DNAs shown in red and blue) 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 (right in Fig. 2.S6) shows that the same subunit contacts the other P element end (blue) through the helix-turn-helix domain, a long loop in the RNase H domain, and the C-terminal domain. Overall the observed architecture supports a *trans*-catalysis mechanism, in which transposase binds to and holds one P element end, but catalyzes the strand transfer of the other end. This interlocking architecture likely acts as a checkpoint to ensure proper assembly of the nucleoprotein complex prior to catalysis of DNA integration.

The GTP cofactor interacts with the donor DNA

Among DNA transposases, P element transposase is unique in its requirement of GTP as a cofactor for assembly of the paired end complex and the strand transfer reaction. We were able to identify densities that correspond to GTP and a coordinated magnesium ion (Fig. 2.2e). 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 (Fig. 2.S7). Interestingly, residues that mediate GTP binding (D528, K385, K400, V401, S409, F443, D444, and N447) are conserved within members of the P element superfamily (Yuan and Wessler, 2011) (Fig. 2.3a, inset). We observe that GTP makes base-stacking interactions with the transferred strand (T_{-9}) and is most likely hydrogen bonding with G_{-1}, the terminal P element donor DNA nucleotide. The interaction between GTP and the donor DNA may act to position the attacking 3'OH in the active site and would explain why GTP is required for strand transfer.

To investigate the interactions of GTP in the strand transfer complex, we performed strand transfer assays with radiolabeled donor DNAs and different purine nucleoside triphosphate analogs (Fig. 2.3d). Nucleotides that lacked a C6 carbonyl group did not support strand transfer activity (2-aminopurine, ATP, 2-amino-ATP, Fig. 2.3d, lanes 3, 5, 6). Conversely, ITP and to lesser extent XTP, both of which carry the C6 carbonyl group, did support strand transfer activity, but not to the same level as GTP (Fig. 2.3d, lanes 2, 4, 7). This is likely due to differences of 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 likely facilitate nucleotide binding. These results and the structure support a model in which interactions with GTP act to position the donor DNA for strand transfer and explain the specificity of GTP. Thus, GTP is the only nucleotide that can fully support the observed interactions at this stage of transposition.

Altered target DNA structure stimulates transposition

Target DNA bending is a common feature among DDE transposases (Montaño et al., 2012; Morris et al., 2016) and the related retroviral integrases (Maertens et al., 2010; Passos et al., 2017; Yin et al., 2016). Recent studies indicate that DNA flexibility and deformability play a critical role in target site selection, where regions of flexibility optimize protein-DNA contacts and facilitate positioning of the scissile phosphate into the protein active site (Fuller and Rice, 2017; Wright et al., 2017). Consistent with these findings, we observe substantial distortion of the target DNA within the P element STC (Fig. 2.5a). At each strand transfer site, the target DNA duplex exhibits a sharp ~55° bend away from the central axis (Fig. 2.5b). This distortion is accommodated over the AT-rich flanking sequences, which display a widened minor groove (Fig. 2.5, green, Fig. 2.S8a). The central 8 bp GC-rich TSD duplex remains approximately B-form (Fig. 2.5, red).

The target DNA binds along a basic channel formed by the RNase H and GTP-binding insertion domain of each monomer (Fig. 2.S8b). Numerous residues from both the RNase H domain (K310, R538 and H546) and the GTP-binding insertion domain (H350, R394, Q399, and K487) are positioned to contact the phosphate backbone, likely stabilizing the observed target DNA conformation (Fig. 2.5c,d and Fig. 2.S8c). A loop from the GTP-binding insertion domain 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. 2.5c). RNase H domain residues T306 and Y253 are positioned within the minor groove of the flanking AT-rich regions (Fig. 2.5d). T306 contacts T_{11} at the extremity of the TSM. While Y253 is also positioned within 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 target DNA bending and thereby help position the scissile phosphate within the transposase active site.

Although P element transposition is not site-specific, integration preferentially occurs into the TSM or TSM like-sequences (Linheiro and Bergman, 2008). In our structure base-specific interactions between TNP and the target DNA are sparse, suggesting that the preference for TSM or TSM like-sequences is not achieved through direct target DNA sequence readout alone. Given these observations, the preference for the P element TSM is likely driven by a pattern of target DNA flexibility and further enforced by amino acid side chain-base interactions.

To further investigate the effects of target DNA flexibility on transposase activity we performed *in vitro* strand transfer assays with nicked or mismatched target DNA substrates. G-mismatches or nicks were included along the bottom strand to introduce deformability/flexibility into specific regions of the target DNA duplex (Fig. 2.5e). Mismatches did not appreciably stimulate activity, but rather decreased activity in specific instances (Fig. 2.5e, lanes 4, 5, and 9). Mismatches at positions G_6 and T_{11} coincide with observed TNP-target DNA base interactions, and likely decrease target DNA binding affinity by disrupting these contacts or altering crucial duplex geometries. Notably, nicks along the central GC-rich region increased strand transfer into the top strand of the target DNA, with the greatest stimulation observed at the site of bottom strand transfer, between nucleotides 8/9 (Fig. 2.5e, lane 14). This is the same region that accommodates the highest level of distortion within the target DNA duplex. Taken together, this supports a model in which target DNA flexibility is a contributing factor in transposition activity and in conjunction with the observed limited base-specific contacts likely dictates target site preference.

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 terminal inverted repeat (Beall and Rio, 1997) (TIR, Fig. 2.6a). The 5' and 3' ends are both required for the proper initiation of transposition. The 5' end cannot function as the 3' end during the initial stages of synaptic complex assembly before DNA cleavage (Beall and Rio, 1997; Mullins et al., 1989). 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. 2.6b and Fig. 2.S2g), while the THAP DNA-binding domain remains unresolved, likely due to flexibility. The additional 12 residues of the leucine zipper dimerization domain are oriented towards one of the 3' P element donor DNA 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. 2.6c), reminiscent of the flexible nonamer binding domain in the RAG1/RAG2-12/23 RSS complex, which enforces the 12/23 rule in V(D)J recombination (Beall and Rio, 1997; Rodgers, 2017). 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 (Kim et al., 2015; Lapkouski et al., 2015; Ru et al., 2015). 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 (Beall and Rio, 1997). 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 donor DNAs 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 into fundamental cellular processes, such as the regulation of tissue-specific alternative splicing and DNA repair pathways. Among all previously characterized DNA transposases, P element transposase 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. The structure presented here provides the first three-dimensional view of the P element superfamily of eukaryotic DNA transposases, illuminating many of the important mechanistic features distinct to this family of proteins.

Our structural data has revealed a complex nucleoprotein architecture and allowed the unambiguous identification of the domain organization of P element transposase. Previously, only the N-terminal THAP DNA binding domain had been structurally characterized (Sabogal et al., 2010). The new structural information presented here visualizes four additional protein domains: a helix-turn-helix domain, a catalytic RNase H domain, a GTP-binding domain, and a highly charged C-terminal domain. The GTP-binding domain is inserted into the RNase H catalytic domain. The location of this insertion domain is similar to other insertion domains found in bacterial Tn5, housefly Hermes and the jawed vertebrate V(D)J RAG1 enzymes (Fig. 2.S4c). This observation suggests that the RNase H fold readily tolerates an insertion at this position or that these transposases possibly share a common ancestor and diverged after the insertion of a primitive domain at this position of the RNase H fold. In fact, some of the insertion domains share structural similarity (Fig. 2.S4d).

P element transposase is unique in using GTP as a non-hydrolyzed cofactor for both the cleavage and integration steps of transposition. Earlier work confirmed the use of GTP by mutating a single amino acid that changed the nucleotide specificity from GTP to XTP, *in vivo* and *in vitro* (Mul and Rio, 1997). Our structure reveals that the guanine base of GTP makes hydrogen bond contacts the terminal G base at the transposon end, 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, and thereby providing a rationale for the GTP requirement for strand transfer (Beall and Rio, 1997).

Previous biochemical and atomic force microscopy studies with full-length P element ends indicated that a transposase tetramer acts at the early stages of transposition in forming synaptic paired end complexes (PEC) and cleaved donor complexes (CDC) (Tang et al., 2005; 2007). However, we observed that the strand transfer complex is dimeric. Assembly of the STC used minimal oligonucleotide donor DNA substrates, rather than the two full-length ~150 bp P element ends. The longer P element ends include the 11 bp internal inverted repeats, which act as transpositional enhancers *in vivo* (Mullins et al., 1989). 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 donor DNA 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, that captures a target DNA and performs strand transfer. Contributions to DNA binding by non-catalytic subunits has been observed in both the bacteriophage Mu transposome (Montaño et al., 2012) and the retroviral integrase structures (Passos et al., 2017; Yin et al., 2016) and is thought to occur in the octameric Hermes transposome (Hickman et al., 2014).

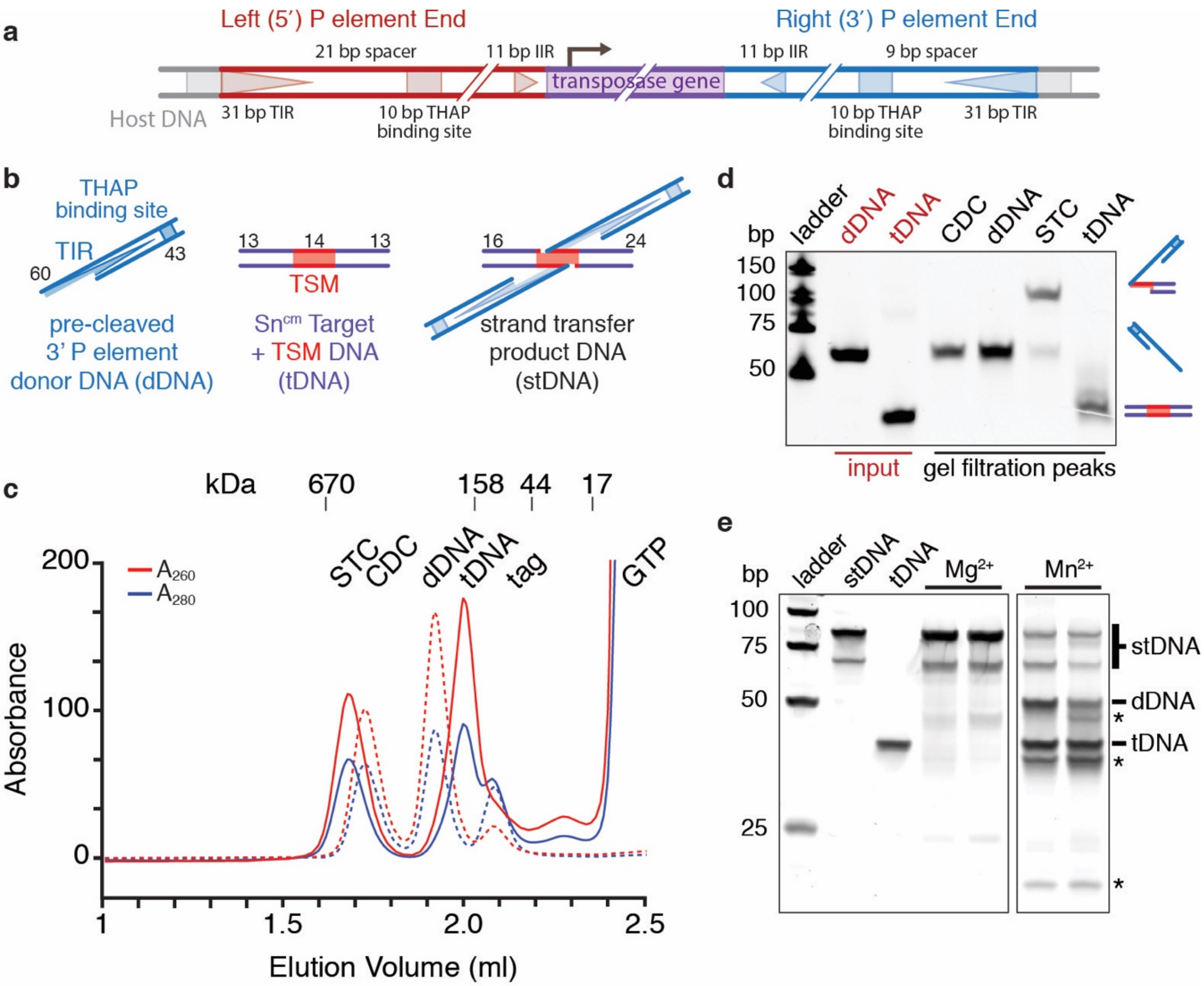
Overall, our structure suggests that during the early stages of transposition, when the THAP domains engage with the internal 10 bp transposase binding sites, that P element transposase acts to pair the two different P element ends in a manner reminiscent of the 12/23 rule imposed by the RAG1/RAG2 V(D)J recombinase (Kim et al., 2015; Lapkouski et al., 2015; Ru et al., 2015). The atypically long staggered cleavage and the arrangement of the donor DNAs 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 at must occur the P element transposon ends to generate the distorted terminal DNA conformations observed in the STC structure. Furthermore, GTP is required for pairing of the two P element ends prior to the DNA cleavage (Tang et al., 2005; 2007) indicating that GTP plays an additional role(s) at the early stages of transposition. While 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 (Chuong et al., 2017). For instance, the human THAP9 gene encodes a functional P element transposase homolog that can mobilize *Drosophila* P element DNA in both *Drosophila* and human cells (Majumdar and Rio, 2015). However, the natural DNA substrates and cellular functions of these P element transposase homologs are currently unknown. Our data provides a structural framework for understanding all future biochemical studies, not only of *Drosophila* P element transposase, but also of the related vertebrate P element transposase THAP9 homologs with as yet unidentified cellular functions.

Acknowledgments

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, Y. Li critical reading of the manuscript. Work in the Rio lab was supported by NIH grant R35GM118121. EHK was supported by NIH grant K99GM124463. EN is an Investigator of the Howard Hughes Medical Institute.

Figure 2.1



Reconstituted STC represents the active form of P element transposase.

a, Diagram of the full-length P element transposon depicting the differently spaced 5' and 3' ends. The 31 bp terminal inverted repeats (TIR, triangles), 10 bp THAP domain binding site (squares), the 11 bp internal inverted repeats (IIR, triangles), and the P element transposase gene (purple) are indicated. The 5' and 3' P element ends are colored red and blue, respectively. Not drawn to scale.

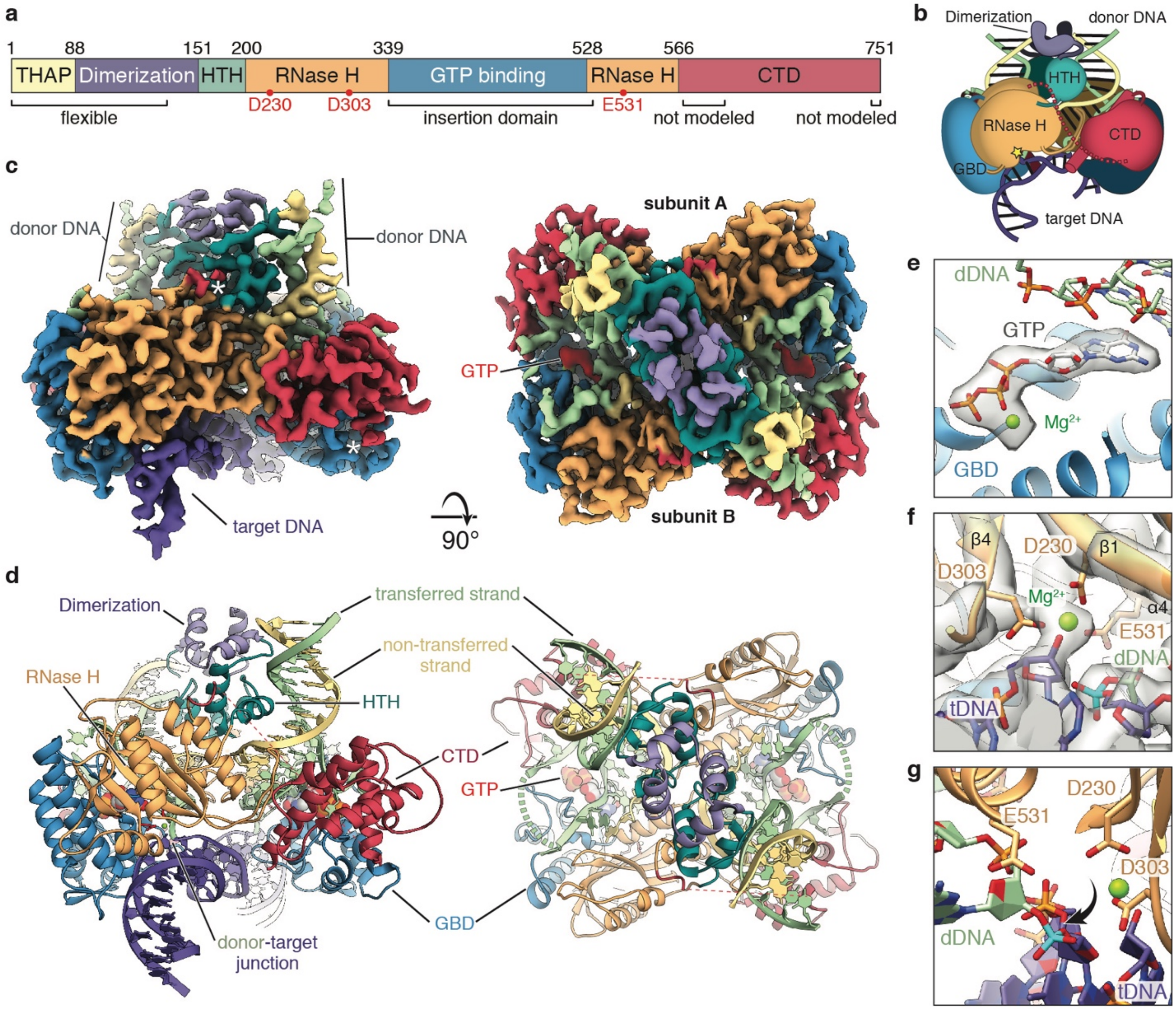
b, Schematic of DNA substrates used. The nucleotide length of each strand is indicated (TIR, terminal inverted repeat; TSM, target sequence motif; dDNA, donor DNA; tDNA, target DNA; stDNA, strand transfer product DNA). Not drawn to scale.

c, Cleaved donor complex (CDC) and strand transfer complex (STC) gel filtration elution profiles (CDC, dotted lines; STC, solid lines). A_{260} and A_{280} absorbance are indicated in red and blue, respectively. Elution positions of mass standards (in kilodaltons) are shown above.

d, SYBR Gold stained denaturing PAGE of dDNA input, tDNA input and peak fractions from fig. 2.1**b**. Schematic of DNAs are shown to the right. Input DNA standards and deproteinized gel filtration fractions are colored red, respectively. bp, base pairs of markers.

e, SYBR Gold stained native PAGE gel of disintegration assay with strand transfer product DNA. The expected mobility of the dDNA and tDNA products are indicated to the right. Unidentified bands are indicated with asterisks.

Figure 2.2



Structure of the *Drosophila* P element STC.

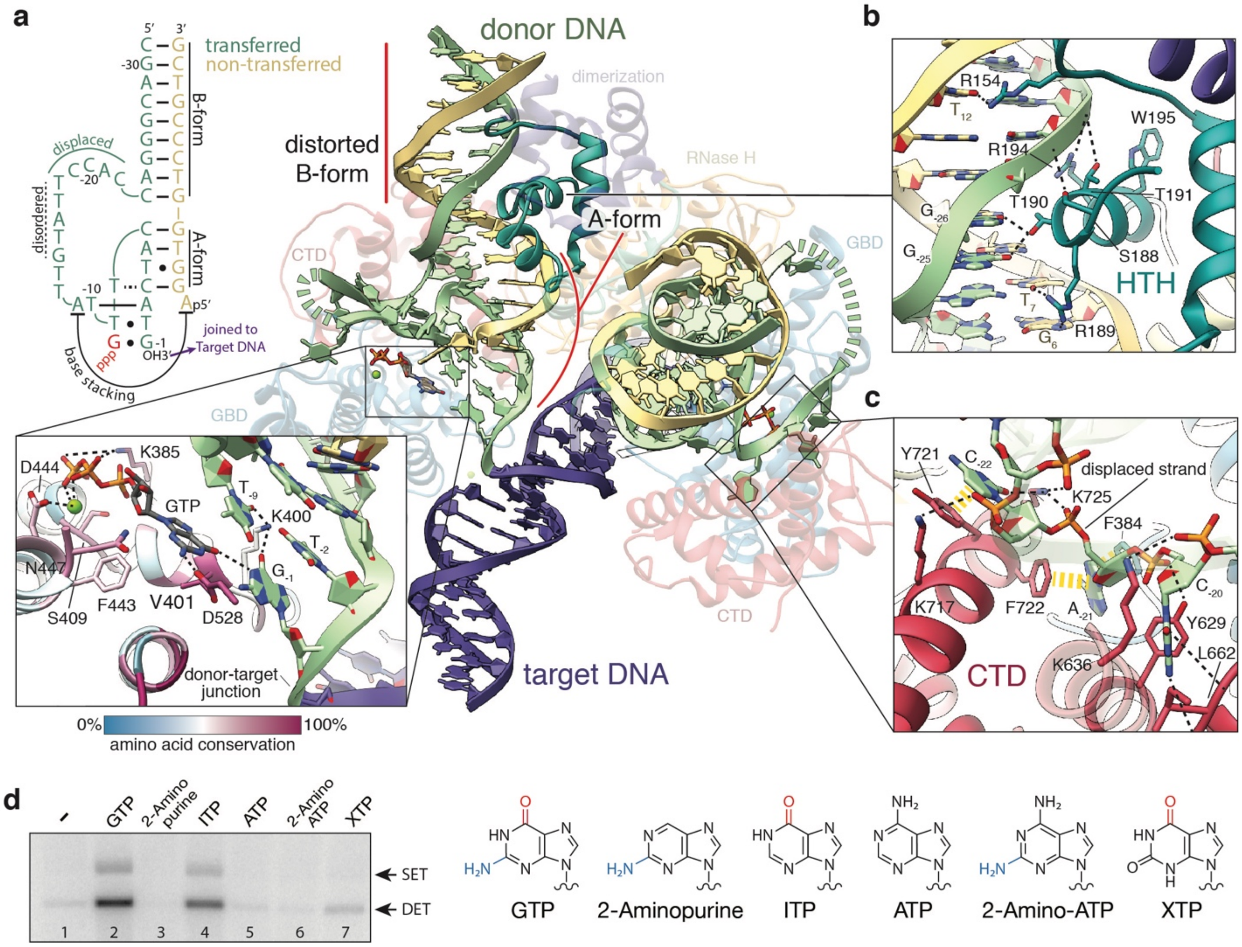
a, Domain architecture of the *Drosophila* P element transposase with the domain boundaries indicated by amino acid residue numbers. The RNase H-like catalytic domain (RNase H, orange) is interrupted by a GTP binding insertion domain (GBD, blue). The RNase H catalytic residues are indicated as red dots. THAP, THAP DNA-binding domain (yellow); Dimerization, leucine zipper dimerization domain (purple); HTH, helix-turn-helix domain (dark cyan); CTD, C-terminal domain (red).

b, Cartoon of the P element transposase strand transfer complex. The catalytic site is indicated with a yellow star and domains are colored as in fig. 2.2**a**. Domains of the other subunit are darkened.

c, Side (left) and top (right) views of the cryo-EM reconstruction at 3.6 Å. Domains are colored as in fig. 2.2**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 P element transposase STC model. Colored as in fig. 2.2**c**, with domains indicated. Catalytic residues are colored red, and unmodeled connections are shown as dashed lines (dashed green, dashed red). Target DNA is shown in purple, donor transferred strand in light green and donor, non-transferred strand in yellow. 26 bp of the 40 bp target DNA are well-resolved, while the first 23 bp of each donor DNA are not well-resolved in the symmetrized reconstruction.
e, Close up view of the modeled GTP density. Only the density corresponding to GTP is shown for clarity (dDNA, donor DNA).
f, Close up view of the RNase H catalytic residues (tDNA, target DNA). Density as in **b**, 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. Similar view as in fig. 2.2**f,** but rotated 90°. Density was omitted for clarity. The scissile phosphate is shown in cyan.

Figure 2.3



Donor DNA adopts a noncanonical geometry within the STC.

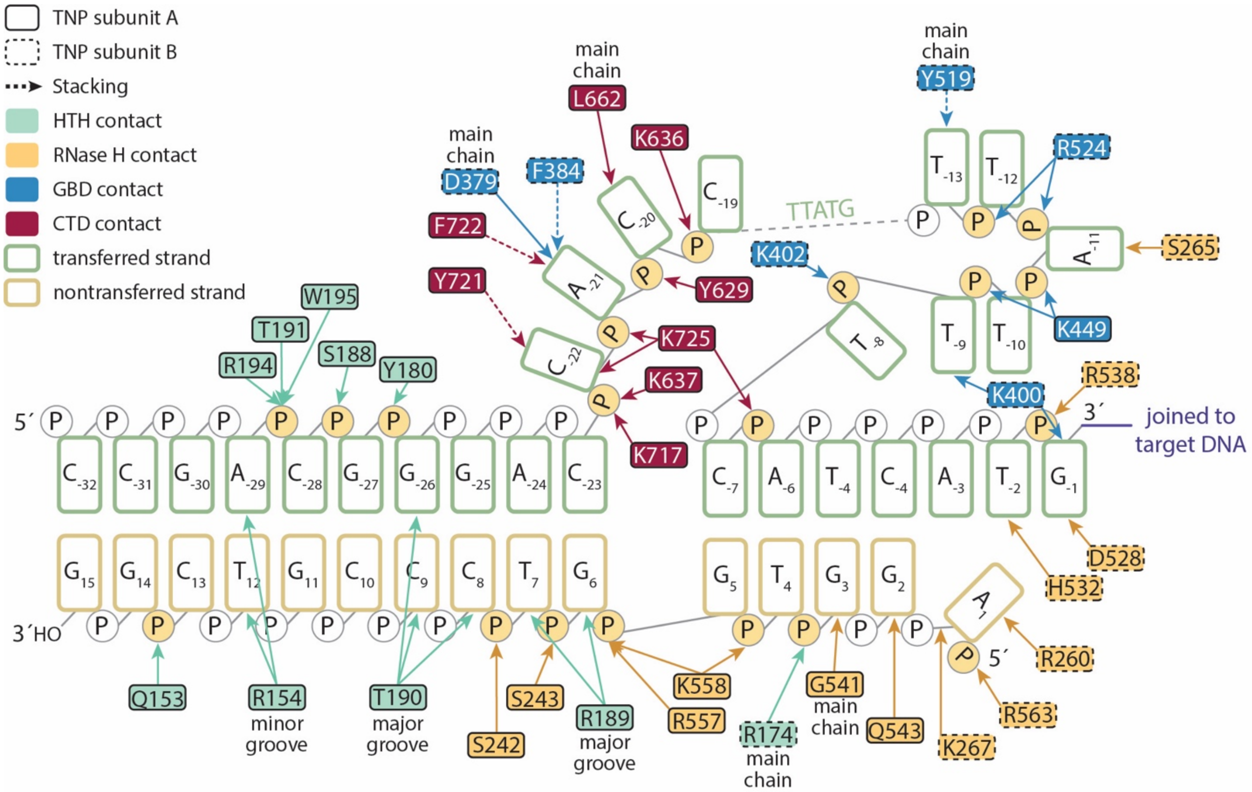
a, Overview of donor DNA structure within the strand transfer complex. Distorted B-form and A-form regions of the donor DNA are indicated. The transposase protein is faded out for clarity with relevant domains labeled (dimerization, leucine zipper dimerization domain; RNase H, RNase H-like catalytic domain; GBD, GTP binding insertion domain; CTD, C-terminal domain). 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 donor DNA terminal inverted repeat **(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 interaction between GTP, the GTP-binding insertion domain, and donor DNA (bottom, inset). 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 helix-turn-helix domain (HTH) and donor DNA contacts. Nucleotides are numbered as in fig. 2.3**a**.

c, Close-up view of C-terminal domain (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 triphosphates analogs. Agarose gel of a strand transfer assay (left). 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 (right). C6 carbonyl groups and C2 amino groups are colored red and blue, respectively.

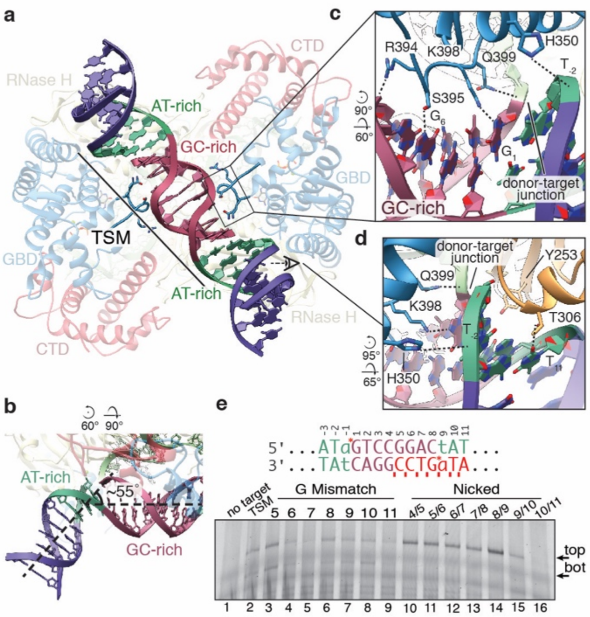
Figure 2.4



Each subunit makes extensive contacts with a single donor DNA.

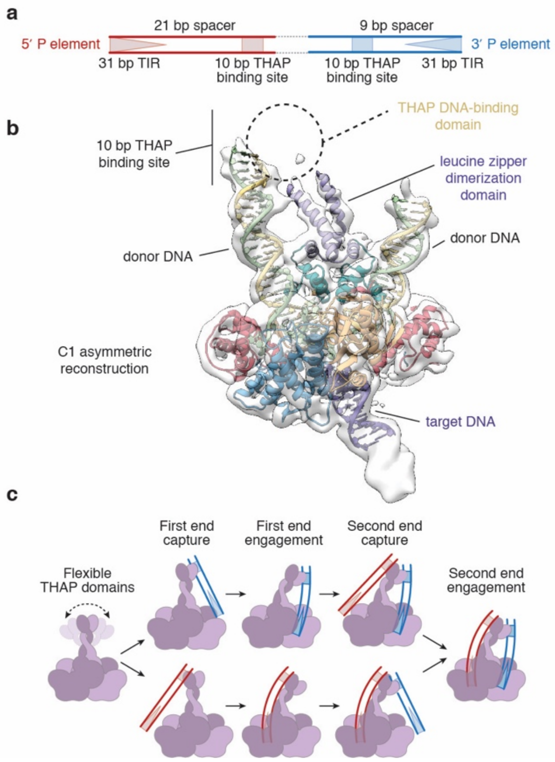
Schematic representation of the inferred base-specific and backbone contacts between transposase and the donor DNA. Nucleotides of the transferred strand (green fill) are numbered -1 to -32, starting at the 3' terminal guanosine. Nucleotides of the non-transferred strand (gold fill) are numbered 1 to 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; GDB, 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.

Figure 2.5



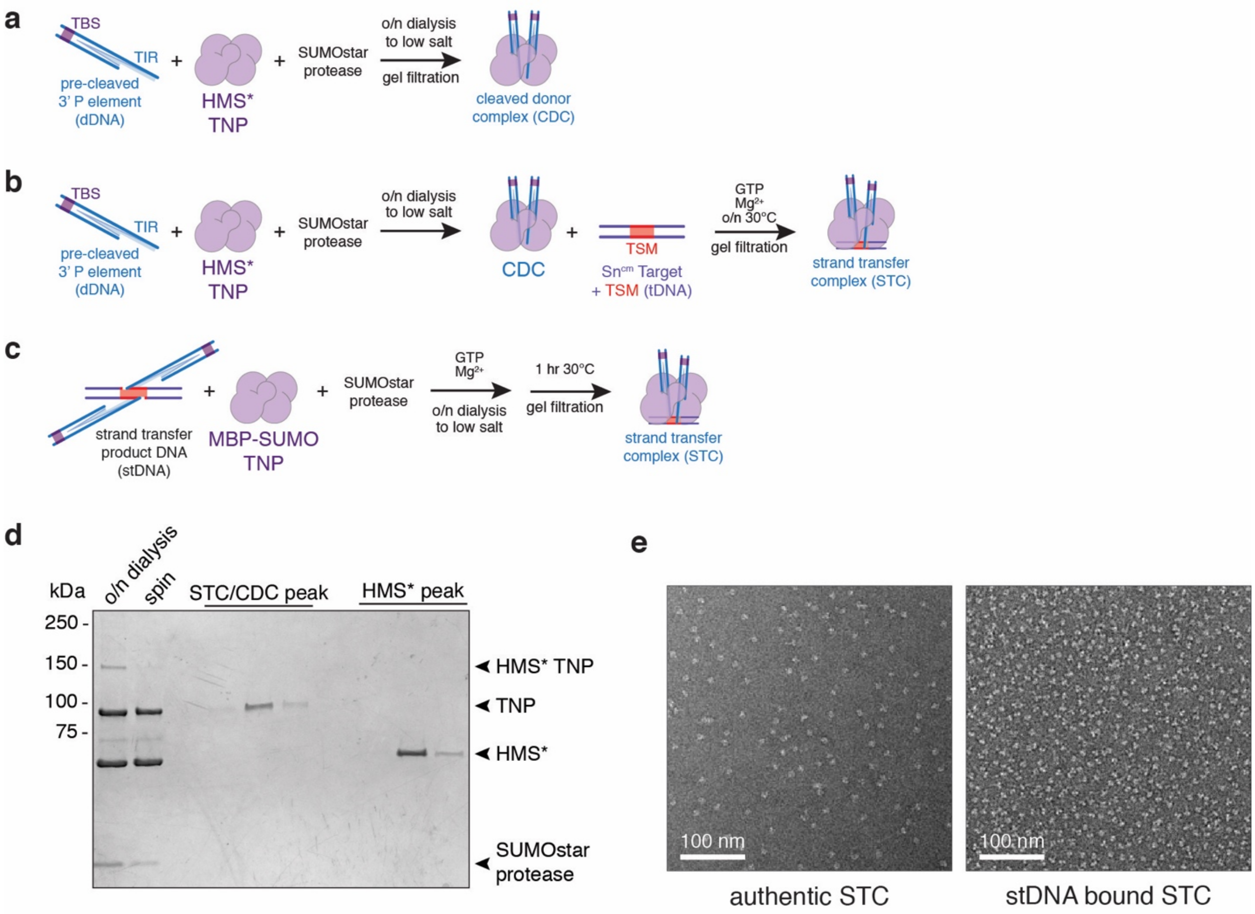
The target DNA is severely bent at AT-rich sites.
a, Bottom view of the strand transfer complex, highlighting the bent target DNA. AT-rich (green) and GC-rich (red) regions of the target DNA are indicated. The GBD loop that interacts with the target DNA is shown. The transposase protein is faded out for clarity with relevant domains labeled. (GBD, GTP-binding insertion domain; CTD, C-terminal domain; TSM, target sequence motif). All subsequent panel rotations are depicted with respect to fig. 2.5**a**.
b, Bend at flanking AT-rich sites. Bend is highlighted and dashed lines indicate the central axis of the DNA. The target DNA is colored as in fig. 2.5**a**.
c, Close-up view of the target DNA-GDB-loop interaction inferred from the atomic model. Site-specific interactions are indicated (S395:G1,K398:G6). Nucleotides are numbered as in fig. 2.5**e**.
d, Close-up view of target DNA-RNase H domain interaction inferred from the atomic model. Site-specific interactions are indicated (T306:T11). A region of target DNA backbone was made transparent for clarity.
e, Denaturing PAGE gel of a transposition assay using mismatched, or nicked target DNA substrates. The sequence of the target sequence motif (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. 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 target DNA are indicated to the right of the gel. The transferred strand of the donor DNA was fluorescently labeled at the 5' position with a TAMRA dye.

Figure 2.6



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 terminal inverted repeats (TIR, 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 donor DNA and the leucine zipper dimerization domain were modeled into the density. Expected position of the THAP domain, and THAP domain binding site are indicated.
c, Model for pairing of the 5' and 3' P element ends. P element transposase protein (purple and light purple), 3' P element transposon end (blue) and the 5' P element transposon end (red) are represented as cartoons.

Figure 2.S1



Assembly of the CDC, authentic STC, and STC bound to stDNA.

a, Diagram of the cleaved donor complex (CDC) assembly pathway. (TIR, terminal inverted repeat).

b, Diagram of the authentic strand transfer complex (STC) assembly pathway. CDCs were assembled as in fig. 2.S1**a**, then provided with an idealized hotspot target DNA from the *Drosophila* singed locus (tDNA). (GTP, guanosine triphosphate; TSM, target sequence motif).

c, Diagram of the strand transfer complexes assembled on strand transfer product DNA (stDNA).

d, Representative Coomassie-stained SDS-PAGE of the overnight dialysis and gel filtration fractions. ‘o/n dialysis’ and ‘spin’ lanes were diluted 1:10 before loading. (HMS*, 6xHis-Maltose binding protein-SUMO* tandem solubility tag).

e, Negative stain electron micrographs of authentic STC and stDNA bound STC.

Figure 2.S2

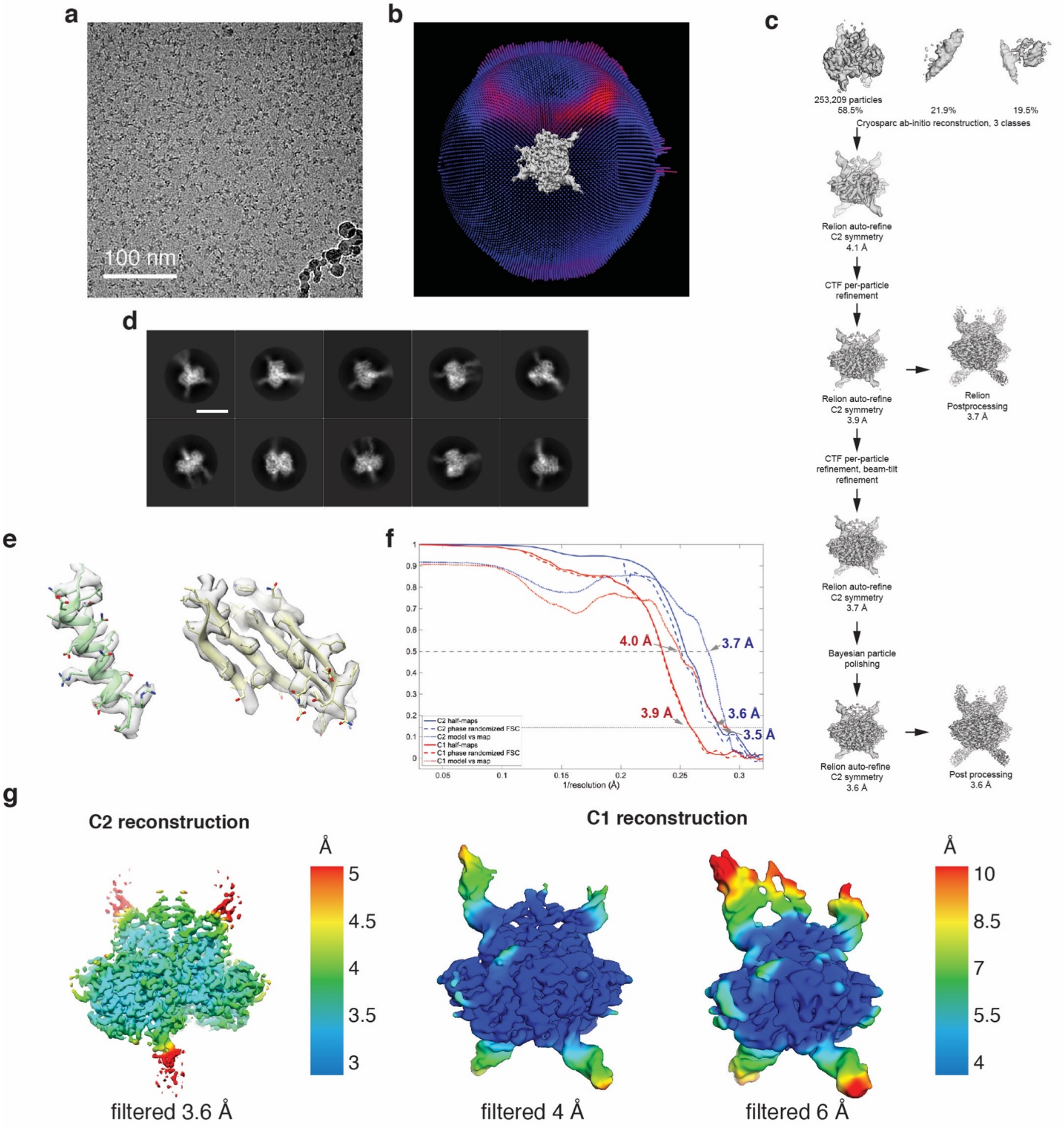


Image processing of tilted dataset leading to a 3.6 Å resolution cryo-EM reconstruction.

a, Representative cryo-EM image collected with a 40° tilt showing well-defined, monodispersed particles (scale bar represents 100 nm).

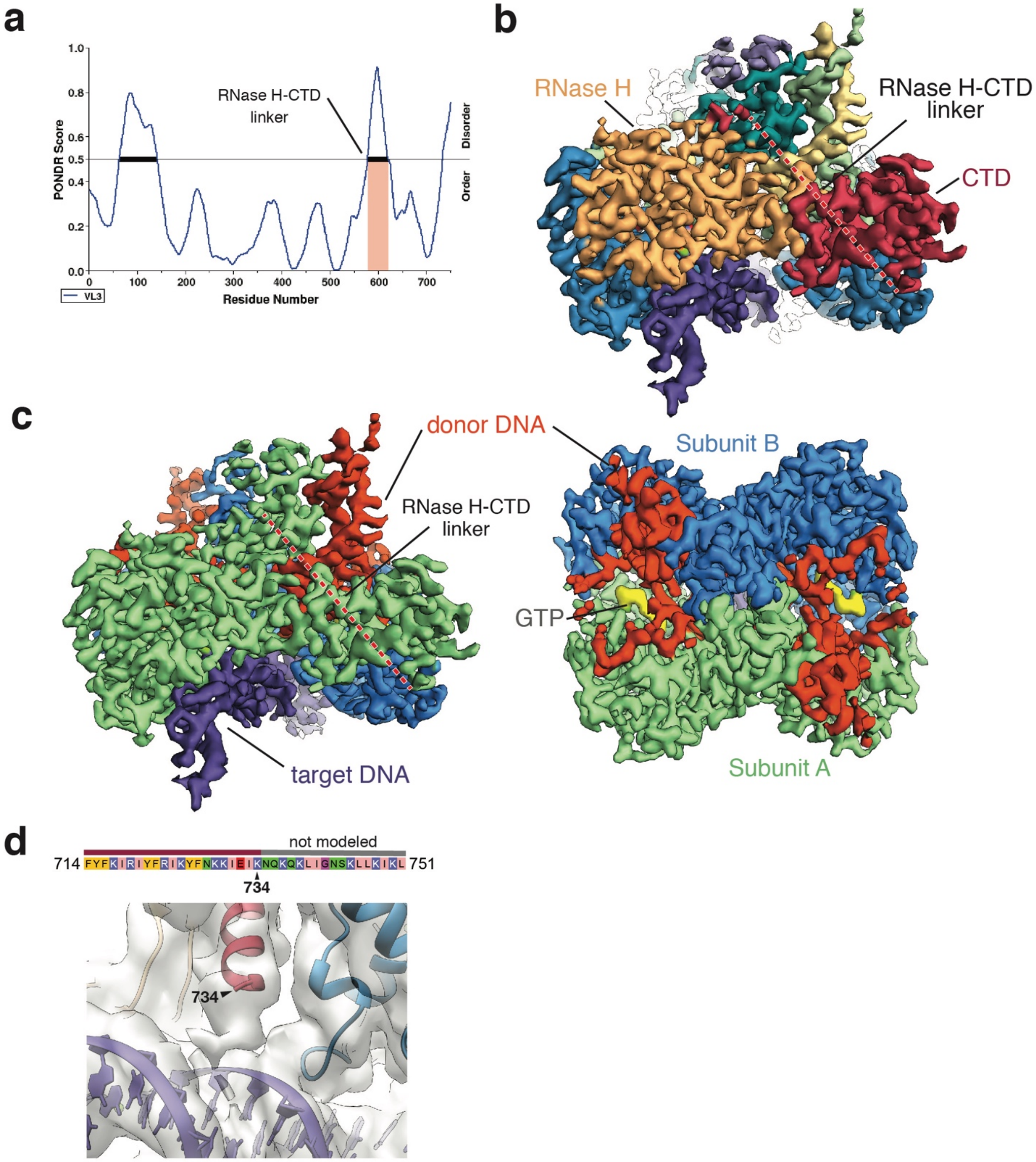
b. Angular distribution of particles from the tilted dataset is cone-like, corresponding to a majority of top-views.

c. A single, well-defined reconstruction was produced using cryoSPARC and subsequently refined to high resolution using RELION-3.0 (see methods for details).

d. Reference-free 2D classes of the tilted data reveal secondary structure features.

e. The secondary structure features are consistent with the estimated resolution of the map, with well-defined secondary structure and distinctive densities for large side-chain.
f, The overall resolution (based on the Fourier shell correlation (FSC) 0.143 criterion) for the symmetrized reconstruction is 3.5 Å (3.6 Å if using randomized phases), and 3.9 Å for the unsymmetrized reconstruction. The map versus model resolution is 3.7 and 4 Å, respectively, for the symmetrized and unsymmetrized maps.
g. 3D map for the C2 (symmetrized, left) and C1 (unsymmetrized, right) reconstructions colored by local resolution showing the core of the structure to be around 3.5 Å. To show some of the most disordered regions, the C1 map is shown low-pass filtered to both 4 Å and 6 Å.

Figure 2.S3

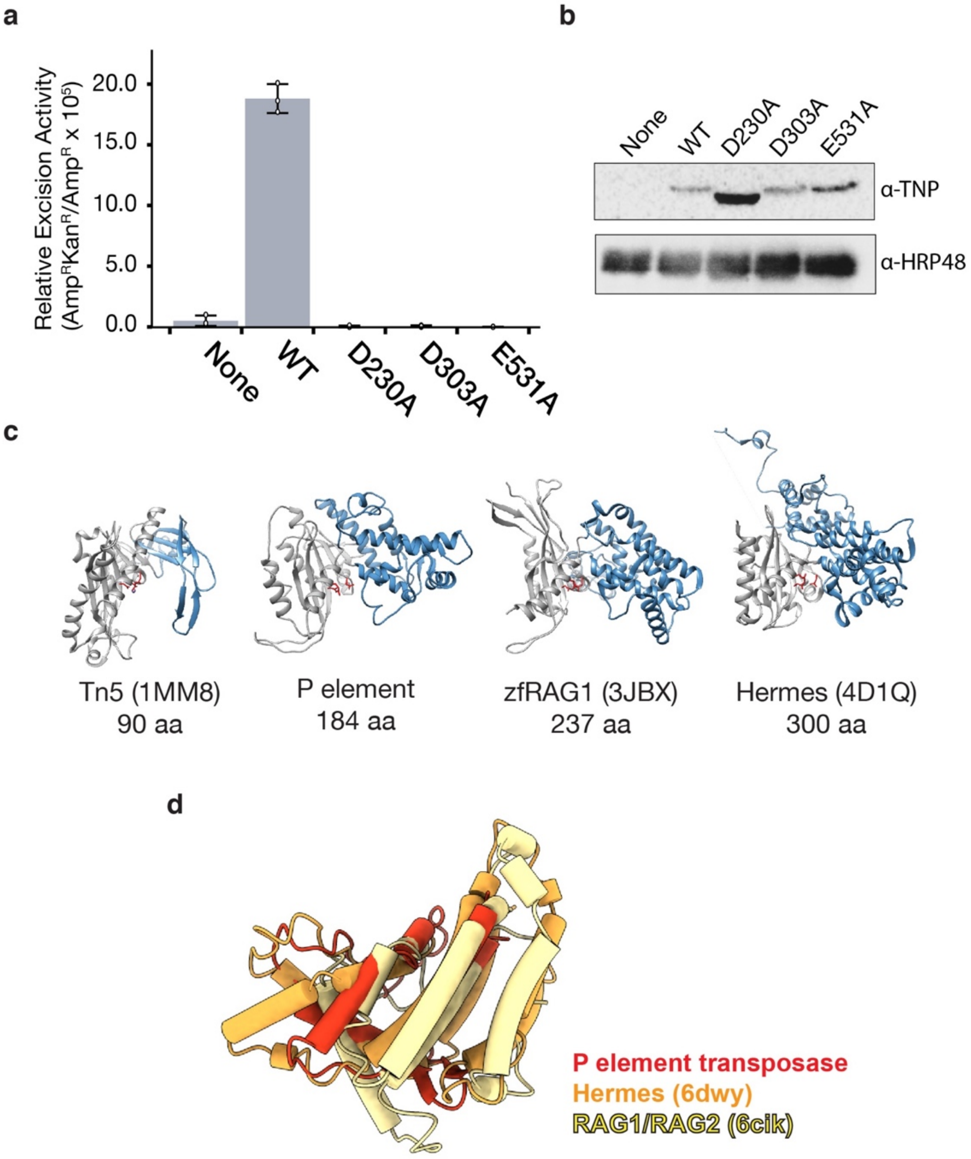


The STC is dimeric and contains disordered regions.

a, PONDR scores of predicted disordered regions. Disordered regions predicted with high confidence (indicated by black bars), are within the leucine zipper dimerization domain and the RNase H – CTD linker region. Contrary to the prediction results, we observe the dimerization

domain to be largely ordered in the C1 reconstruction, in spite of the prediction suggesting that this region may undergo a disorder-to-order transition upon dimerization (Dunker et al., 2001).
b, The disordered linker spanning RNase H and CTD domains is represented by the dashed line, with density for ordered regions colored by domain as in fig. 2.2**b**.
c, STC subunit organization. Densities are shown as in fig. 2.2**b** but are colored by subunit (blue and green). Donor DNAs are colored in red, and target DNA in purple. The density corresponding to GTP is indicated in yellow.
d, Unmodeled density at the C-terminus. The C-terminus terminates with a highly basic stretch of amino acids oriented towards the target DNA. The displayed map is low-pass filtered to 4 Å in order to show more clearly the presence of additional, poorly-ordered density at the C-terminus. While we could not confidently build into the density beyond position 734, the highly basic nature and positioning of the observed weak density near DNA suggests that this region likely plays an important role during transposition by binding the DNA. Consistent with this, C-terminal tags on transposase decrease the overall excision and strand transfer activity (unpublished results).

Figure 2.S4



P element transposase RNase H domain catalytic mutants are inactive and a comparison of RNase H insertion domains among different transposases.

a, Bar graph of relative *in vivo* P element excision activity of alanine-substituted catalytic mutants (D230, D303 and E531). Cell-based excision assays were performed as previously described (Beall and Rio, 1996; Rio et al., 1986). Single alanine mutants were generated by site-directed mutagenesis of pPBSKS (+)pAc-TNP and verified by sequencing over the entire coding sequence. The assay was conducted in triplicate (n = 3). Error bars indicate standard deviations. (WT, wild type).

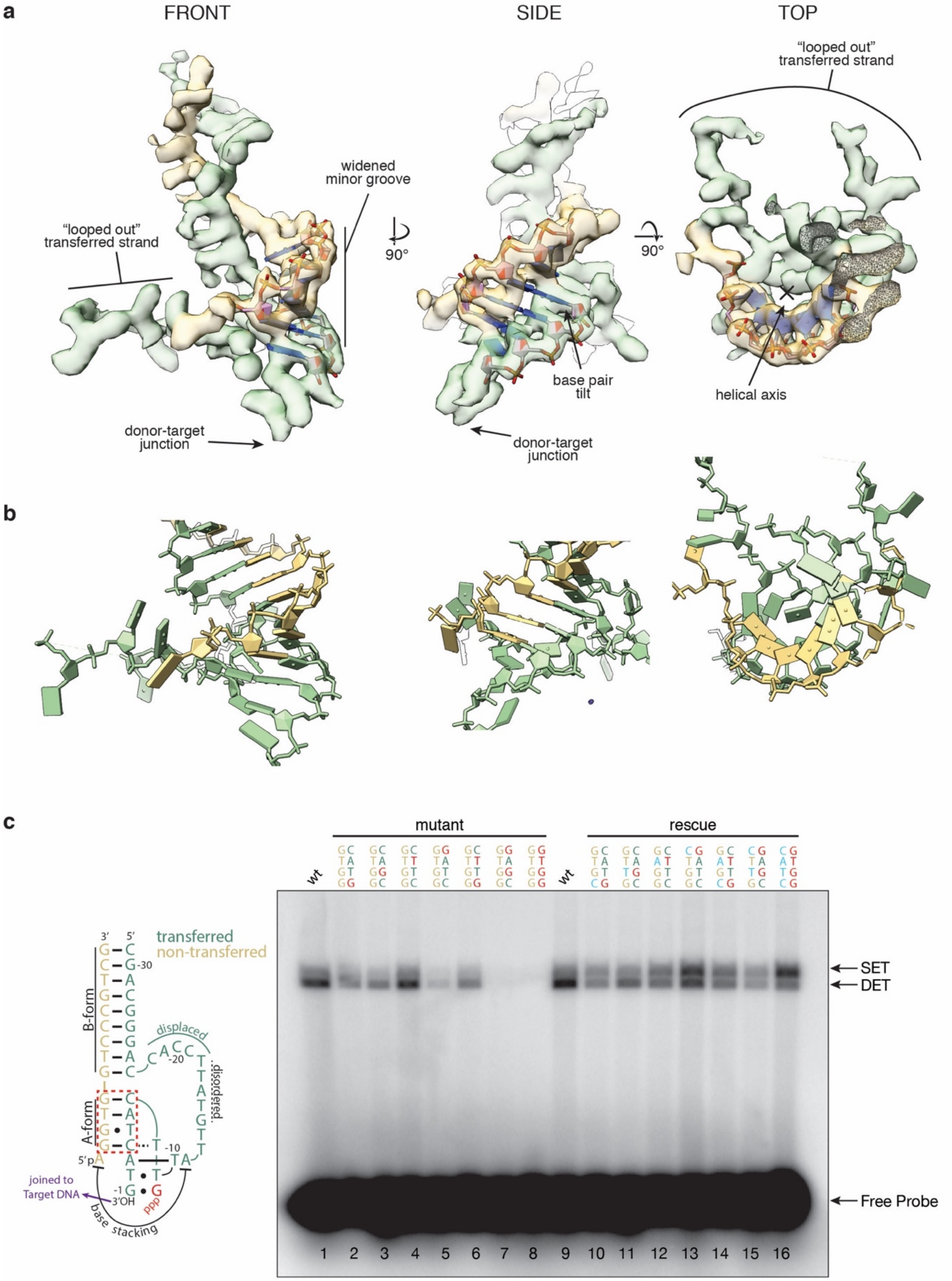
b, Representative immunoblot of wild type transposase and catalytic mutant protein expression levels. Cells were harvested 24hr after transfection and lysates were normalized to cell number.

Membrane was cut and then immunoblotted with anti-transposase antibodies (α-TNP) or a loading control (α-HRP48).

c, Architectures of insertion domains found in other DNA transposases. The RNase H domains (grey) of other structurally characterized DNA transposases (or the transposase-related RAG1 protein) were aligned by their respective catalytic residues (indicated in red) and ordered by increasing insertion domain size (blue). Insertion domain sizes (indicated below) were determined by approximate start and end insertion positions. The PDB numbers from which these structures were derived are in parentheses.

d. Structural alignment of the P element transposase insertion domain, the Hermes insertion domain (1dwy) and the RAG1/RAG2 insertion domain (6cik) reveals structural similarities at the fold level.

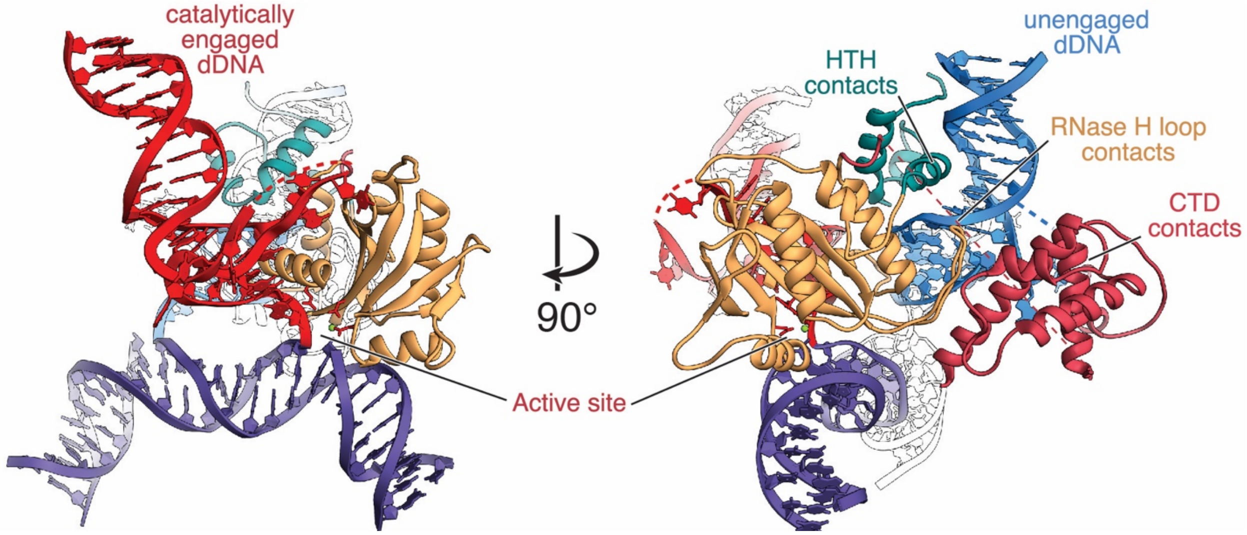
Figure 2.S5



Characteristics of A-form DNA are well resolved and base pairing between distant donor DNA regions is required for strand transfer activity.

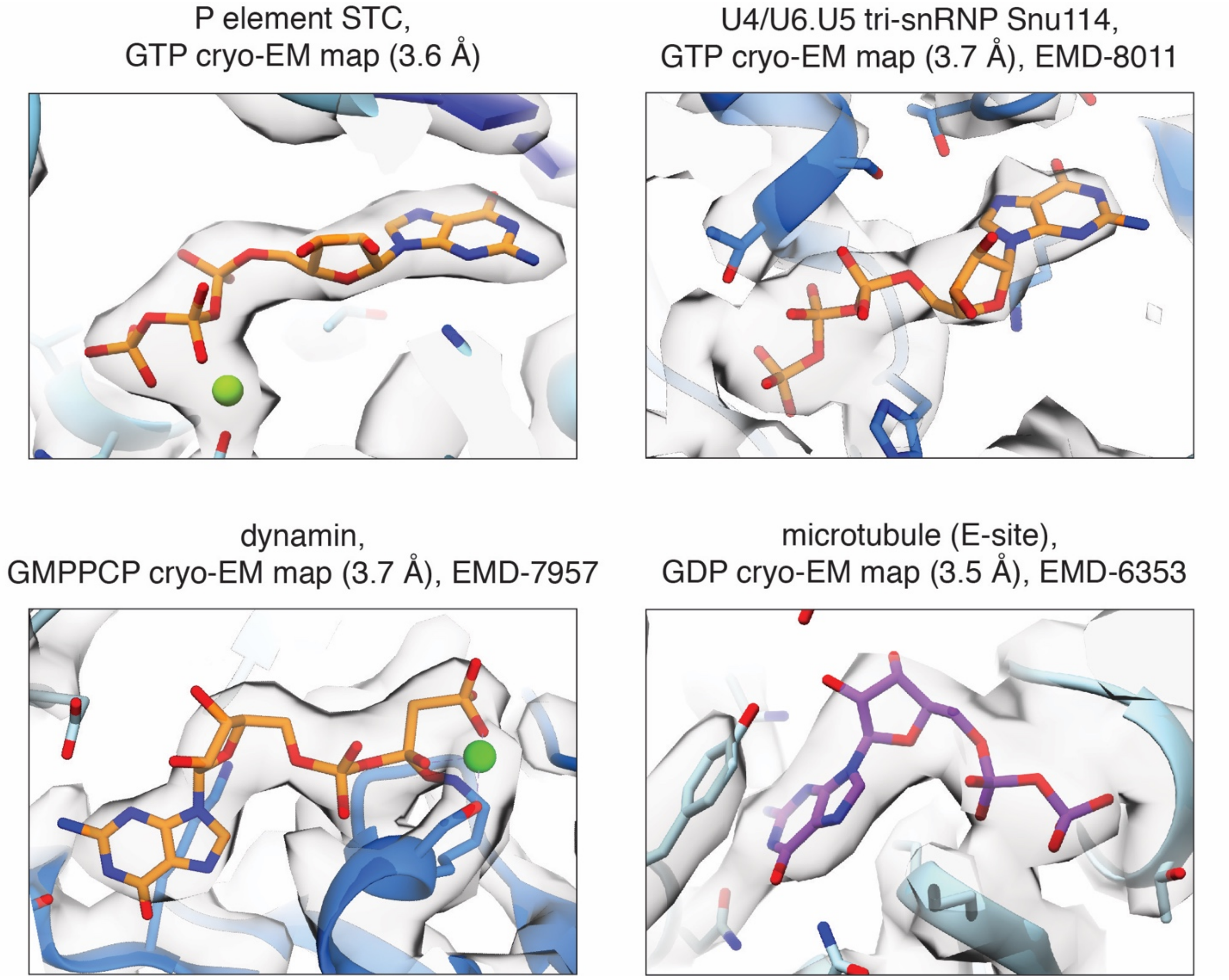
a, Ideal A-form DNA fitted into donor DNA reconstruction, depicting widened minor groove, base pair tilt, axial rise and helical axis dislocation relative to base pairs. A single donor DNA is depicted for clarity. The reconstruction is colored green and yellow, for transferred strand and nontransferred strand, respectively. Relevant regions of DNA are indicated.
b, Atomic model of donor DNA depicting A-form DNA characteristics. Views are as in **a**, except only relevant regions the donor DNA atomic model are depicted.
c, Schematic of the secondary structure of a donor DNA terminal inverted repeat (left). Watson-Crick base pairing is indicated by solid lines. Non-canonical base pairing is indicated by dots, or dotted lines. Nucleotides of the transferred strand are numbered -1 to -31, starting at the 3' terminal guanosine. Distant noncanonical A-form helical base pairing between the transferred and non-transferred strand is highlighted (dashed red box). Agarose gel of a strand transfer assay with 5'-radiolabeled mutant and/or rescue donor DNAs (right). Assays were largely performed as previously described (Beall and Rio, 1998). The base pairs are shown above each lane, with the substituted bases highlighted in red (mutant, lanes 2 - 8). Compensatory substitutions in the non-transferred strand are shown above each lane, with substitutions to restore base pairing highlighted in blue (rescue, lanes 10 - 16). The expected positions of single-ended integration (SET) and double-ended integration (DET), as well as free donor DNA, are indicated. wt, wild type donor DNA.

Figure 2.S6



A single transposase subunit engages both P element donor DNAs.
Left, the RNase H domain of one subunit of transposase is catalytically engaged with one P element donor DNA (red DNA). The domains are colored as in Fig. 2.2**b**. For clarity, the dimerization domain, the GBD, and the other transposase subunit are not shown. Catalytic residues are depicted in red. Right, a 90° rotated view shows the same subunit contacting the other P element donor DNA (blue DNA), through the HTH domain, a long loop in the RNase H domain, and through the CTD. This mode of engagement likely acts as a regulatory step to ensure proper assembly with both P element ends before proceeding to catalysis.

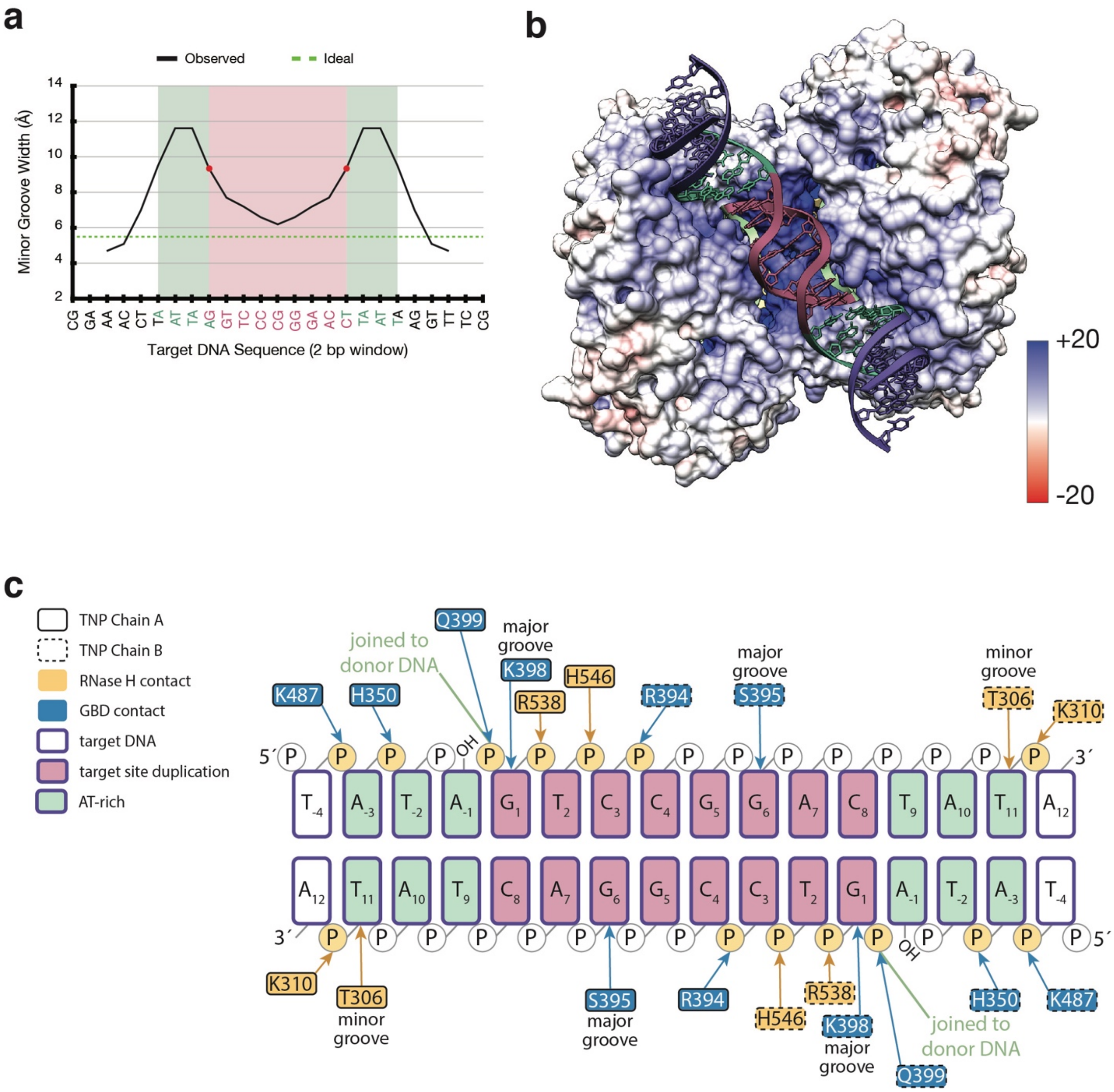
Figure 2.S7


P element STC,
GTP cryo-EM map (3.6 Å)
U4/U6.U5 tri-snRNP Snu114,
GTP cryo-EM map (3.7 Å), EMD-8011
dynamin,
GMPPCP cryo-EM map (3.7 Å), EMD-7957
microtubule (E-site),
GDP cryo-EM map (3.5 Å), EMD-6353

TNP nucleotide density is consistent with GTP and distinguishable from GDP within the reported resolution regime.

The GTP density in our cryo-EM map (top-left) is consistent with the GTP density observed the cryo-EM reconstruction of the U4/U6.U5 tri-snRNP Snu114 (3.7Å, top-right) and the non-hydrolyzable GTP analog (GMPPCP) of dynamin (3.7Å, bottom-left) and inconsistent with GDP in the β-tubulin subunit (3.5 Å, bottom-right), for which GTP is hydrolyzed during microtubule assembly.

Figure 2.S8


a
Observed
Ideal
14
12
10
8
6
4
2
Minor Groove Width (Å)
CG GA AA AC CT TA AT TA AG GT TC CC CG GG GA AC CT TA AT TA AG GT TT TC CG
Target DNA Sequence (2 bp window)
b
+20
-20
c
TNP Chain A
TNP Chain B
RNase H contact
GBD contact
target DNA
target site duplication
AT-rich
joined to donor DNA
Q399
major groove
K398
H546
R538
K487
H350
R394
major groove
S395
minor groove
T306
K310
5′ P P P P OH P P P P P P P P P P P P 3′
T_{-4} A_{-3} T_{-2} A_{-1} G_1 T_2 C_3 C_4 G_5 G_6 A_7 C_8 T_9 A_{10} T_{11} A_{12}
A_{12} T_{11} A_{10} T_9 C_8 A_7 G_6 G_5 C_4 C_3 T_2 G_1 A_{-1} T_{-2} A_{-3} T_{-4}
3′ P P P P P P P P P P P P P OH P P P P 5′
K310
T306
minor groove
S395
major groove
R394
H546
R538
K398
major groove
Q399
joined to donor DNA
H350
K487

Target DNA binds in a positively charged groove.

a, Plot of target DNA minor groove width. The minor groove width was calculated from the target DNA model using the 3DNA webserver (Li et al., 2019), with a 2 bp sliding window, accounting for phosphate van der Waals radii. The target DNA sequence is depicted on the x-axis and colored as in Fig. 2.5**a**. Red dots indicate transposition sites, on either the top or bottom strand of the target DNA.

b, Electrostatic surface potential of the STC as viewed from below the target DNA binding site. Calculations were performed in UCSF Chimera (Pettersen et al., 2004). Blue denotes a positive charge and red denotes a negative charge. Target DNA is shown as in Fig. 2.5**a**.

c, Schematic representation of observed base-specific and backbone contacts between transposase and the target DNA. Target DNA (purple border) is numbered as in Fig. 2.5**e** (target site duplication, pink fill; AT-rich flanks, green fill). 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 (RNase H, orange; GDB, blue). 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 yellow.

Table 2.S1

ID #	Description	Sequence
	Figure 1a-c	
399	3' P element pre-cleaved strand transfer oligo, extends 10 bp past thap binding site	CAAGCATACGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTCATCATG
401	5' phosphorylated nontransfered strand 3' P element strand transfer oligo, extends 9 bp past thap binding site	/5PHOS/AGGTGGTCCCGTCGGCAAGAGACATCCACTTAACGTATGCTT
409	Blunt Singed locus replaced with TSM TOP	CAACGGGTTTCATATAGTCCGGACTATAGTTCGTGAGCGG
410	Blunt Singed locus replaced with TSM BOTTOM	CCGCTCACGAACTATAGTCCGGACTATATGAAACCCGTTG
	Figure 1a,d & Figure 2	
623	3' strand transfer (55) half site integration into Sn60 target, extends 13 bp of "right side" of Sn60	ATACGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTCATCATGGTCCGGACTATAGTTCGTGAGCGG
624	bottom strand of Sn60 target for half site	CCGCTCACGAACTATA
625	5' phosphorylated nontransfered 3' strand transfer (55), extends 5 bp past thap binding site	/5PHOS/AGGTGGTCCCGTCGGCAAGAGACATCCACTTAACGTAT
	Figure 4e	
	5' TAMRA labeled 3' P element pre-cleaved strand transfer oligo, extends 2 bp past TBS	/TAMRA/CGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTCATCATG
441	Blunt Targets for high resolution STA, singed locus w/ TSM	CGCTCGCAACGGGTTTCATATAGTCCGGACTATAGTTCGTGAGCGGTCGTTCTCTCCTCT
442	Blunt Targets for high resolution STA, singed locus w/ TSM	AGAGGAGAGAACGACCGCTCACGAACTATAGTCCGGACTATATGAAACCCGTTGCGAGCG
476	Targets for oligonucleotide strand transfer assay, singed locus w/ TSM, GGMM at 4 (2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTATAGTCGGGACTATATGAAACCCGTTGCGAGCG
477	Targets for high resolution STA, singed locus w/ TSM, GGMM at 5 (2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTATAGTGCGGACTATATGAAACCCGTTGCGAGCG
478	Targets for high resolution STA, singed locus w/ TSM, GGMM at 6 (2009 NAR)	CGCTCGCAACGGGTTTCATATAGTCCGGGCTATAGTTCGTGAGCGGTCGTTCTCTCCTCT
479	Targets for high resolution STA, singed locus w/ TSM, GGMM at 6 (2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTATAGGCCGGACTATATGAAACCCGTTGCGAGCG
480	Targets for high resolution STA, singed locus w/ TSM, GGMM at 7 (2009 NAR)	CGCTCGCAACGGGTTTCATATAGTCCGGAGTATAGTTCGTGAGCGGTCGTTCTCTCCTCT
481	Targets for high resolution STA, singed locus w/ TSM, GGMM at 7 (2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTATAGTCCGGACTATATGAAACCCGTTGCGAGCG
482	Targets for high resolution STA, singed locus w/ TSM, GGMM at 8 (2009 NAR)	CGCTCGCAACGGGTTTCATATAGTCCGGACGATAGTTCGTGAGCGGTCGTTCTCTCCTCT
483	Targets for high resolution STA, singed locus w/ TSM, GGMM at 8 (2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTATGGTCCGGACTATATGAAACCCGTTGCGAGCG
484	Targets for high resolution STA, singed locus w/ TSM, GGMM at 9 (2009 NAR)	CGCTCGCAACGGGTTTCATATAGTCCGGACTGTAGTTCGTGAGCGGTCGTTCTCTCCTCT
485	Targets for high resolution STA, singed locus w/ TSM, GGMM at 9 (2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTAGAGTCCGGACTATATGAAACCCGTTGCGAGCG
486	Targets for high resolution STA, singed locus w/ TSM, GGMM at 10 (2009 NAR)	CGCTCGCAACGGGTTTCATATAGTCCGGACTAGAGTTCGTGAGCGGTCGTTCTCTCCTCT
487	Targets for high resolution STA, singed locus w/ TSM, GGMM at 10 (2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTGTAGTCCGGACTATATGAAACCCGTTGCGAGCG
488	Targets for high resolution STA, singed locus w/ TSM, Nick between 3/4(2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTATAGTCC
489	Targets for high resolution STA, singed locus w/ TSM, Nick between 3/4(2009 NAR)	GGACTATATGAAACCCGTTGCGAGCG
490	Targets for high resolution STA, singed locus w/ TSM, Nick between 4/5(2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTATAGTC
491	Targets for high resolution STA, singed locus w/ TSM, Nick between 4/5(2009 NAR)	CGGACTATATGAAACCCGTTGCGAGCG
492	Targets for high resolution STA, singed locus w/ TSM, Nick between 5/6(2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTATAGT
493	Targets for high resolution STA, singed locus w/ TSM, Nick between 5/6(2009 NAR)	CCGGACTATATGAAACCCGTTGCGAGCG
494	Targets for high resolution STA, singed locus w/ TSM, Nick between 6/7(2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTATAG
495	Targets for high resolution STA, singed locus w/ TSM, Nick between 6/7(2009 NAR)	TCCGGACTATATGAAACCCGTTGCGAGCG
496	Targets for high resolution STA, singed locus w/ TSM, Nick between 7/8(2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTATA
497	Targets for high resolution STA, singed locus w/ TSM, Nick between 7/8(2009 NAR)	GTCCGGACTATATGAAACCCGTTGCGAGCG
497	Targets for high resolution STA, singed locus w/ TSM, Nick between 8/9(2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTAT
499	Targets for high resolution STA, singed locus w/ TSM, Nick between 8/9(2009 NAR)	AGTCCGGACTATATGAAACCCGTTGCGAGCG
500	Targets for high resolution STA, singed locus w/ TSM, Nick between 9/10(2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACTA
501	Targets for high resolution STA, singed locus w/ TSM, Nick between 9/10(2009 NAR)	TAGTCCGGACTATATGAAACCCGTTGCGAGCG
502	Targets for high resolution STA, singed locus w/ TSM, Nick between 10/11(2009 NAR)	AGAGGAGAGAACGACCGCTCACGAACT
503	Targets for high resolution STA, singed locus w/ TSM, Nick between 10/11(2009 NAR)	ATAGTCCGGACTATATGAAACCCGTTGCGAGCG
504	Targets for high resolution STA, singed locus w/ TSM, Nick between 11/12(2009 NAR)	AGAGGAGAGAACGACCGCTCACGAAC
505	Targets for high resolution STA, singed locus w/ TSM, Nick between 11/12(2009 NAR)	TATAGTCCGGACTATATGAAACCCGTTGCGAGCG
506	Targets for high resolution STA, singed locus w/ TSM, Nick between 12/13(2009 NAR)	AGAGGAGAGAACGACCGCTCACGAA
507	Targets for high resolution STA, singed locus w/ TSM, Nick between 12/13(2009 NAR)	CTATAGTCCGGACTATATGAAACCCGTTGCGAGCG
	Extended Data Figure 5b	
727	3' P element pre-cleaved strand transfer oligo, wild type transferred strand	ATACGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTCATCATG
728	3' P element pre-cleaved strand transfer oligo, double swap at nt 26 and 28 of 31 bp TIR	ATACGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTCTTGATG
729	3' P element pre-cleaved strand transfer oligo, double swap at nt 25 and 27 of 31 bp TIR	ATACGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTGAGCATG
730	3' P element pre-cleaved strand transfer oligo, complete swap at nt 25-28 of 31 bp TIR	ATACGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTGTGGATG
731	3' P element pre-cleaved strand transfer oligo, wild type non-transferred strand	AGGTGGTCCCGTCGGCAAGAGACATCCACTTAACGTAT
732	3' P element pre-cleaved non-transferred strand oligo, compensatory swap for **#728**	ACGAGGTCCCGTCGGCAAGAGACATCCACTTAACGTAT
733	3' P element pre-cleaved non-transferred strand oligo, compensatory swap for **#729**	AGTTCGTCCCGTCGGCAAGAGACATCCACTTAACGTAT
734	3' P element pre-cleaved non-transferred strand oligo, compensatory swap for **#730**	ACTACGTCCCGTCGGCAAGAGACATCCACTTAACGTAT
736	3' P element pre-cleaved strand transfer oligo, single swap at nt 28 of 31 bp TIR	ATACGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTCATGATG
737	3' P element pre-cleaved non-transferred strand oligo, compensatory swap for **#736**	ACGTGGTCCCGTCGGCAAGAGACATCCACTTAACGTAT
738	3' P element pre-cleaved strand transfer oligo, single swap at nt 27 of 31 bp TIR	ATACGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTCAGCATG
739	3' P element pre-cleaved non-transferred strand oligo, compensatory swap for **#738**	AGTTGGTCCCGTCGGCAAGAGACATCCACTTAACGTAT
740	3' P element pre-cleaved strand transfer oligo, single swap at nt 26 of 31 bp TIR	ATACGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTCTTCATG
741	3' P element pre-cleaved non-transferred strand oligo, compensatory swap for **#740**	AGGAGGTCCCGTCGGCAAGAGACATCCACTTAACGTAT
742	3' P element pre-cleaved strand transfer oligo, single swap at nt 25 of 31 bp TIR	ATACGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTGATCATG
743	3' P element pre-cleaved non-transferred strand oligo, compensatory swap for **#742**	AGGTCGTCCCGTCGGCAAGAGACATCCACTTAACGTAT

DNA substrates used in this study.

Table 2.S2

Data collection	
Microscope	Arctica
Voltage (kV)	200
VPP	No
Camera	K2
Defocus range (μm)	-1.0 to -3.0
Stage tilt (degrees)	40
Dose rate (e^-/pixel/s)	8
Total exposure time (s)	10
Frame rate (ms/fr)	250
Frames	39
Total dose (e^-/Å^2)	60
Nominal pixel size (Å)	1.16
Number of micrographs	1,857

Refinement	
Starting number of particles	547,929
Number of particles (final map)	252,574
Overall resolution (Å)	3.6
Map sharpening B-factor (Å^2)	100

Modeling statistics	
Favored rotamers	100%
Allowed rotamers	0%
Poor rotamers	0%
Favored Ramachandran	98%
Ramachandran Outliers	0%
Molprobity score	1.22
Cβ deviation	0%
Bad bonds	0%
Band angles	0%
Cis-prolines	0%
model-map FSC (0.5)	3.7 Å
cross-correlation	0.8

Refinement and modeling statistics.

Materials and Methods

Protein expression

Drosophila codon optimized His$_8$-MBP-TEV protease cleavage site-TNP was kindly provided by Arzeda Inc. *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 HMS* TNP. The 5' untranslated region was replaced with a lobster tropomyosin cDNA leader sequence (Sano et al., 2002) 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 (Trowitzsch et al., 2010). 10 ml 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. Cells were cultured in paddle flasks in TNM-FH/10% fetal bovine serum/1X Penicillin/Streptomycin (Gibco). Infected cells were incubated for 72 hr (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 phenylmethylsulfonyl fluoride (PMSF), 1X 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 160,000xG 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 hr. The resin was washed three times with 10 column volumes (CV) 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 1 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 1000 mM $(NH_4)_2SO_4$ over 5 CV. 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/RNA synthesizer (Applied Biosystems, Inc.), and were purified using denaturing polyacrylamide gel electrophoresis (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.

Strand transfer complex assembly

For Strand transfer complex assembly, 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 TCEP) at 4 °C overnight. After dialysis, a white precipitate was observed that could not be solubilized by the addition of salt (Ballandras-Colas et al., 2016; Yin et al., 2016). The mixture was centrifuged to remove precipitates. Soluble TNP DNA complexes were incubated at 25 - 30°C for 1 hr 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

~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 20X STOP buffer (85 mM EDTA, 5% SDS), then incubated at 37°C for 2 hr with 0.1 mg/ml proteinase K. 2 µl 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 (Beall and Rio, 1998). Briefly, 250 ng HMS* TNP (1.8 pmol monomer) was preincubated with 0.4 pmol of radiolabeled minimal pre-cleaved 3' donor DNA 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 bovine serum albumin). The reaction was initiated by the addition of 14 µl of 0.35x 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 hr. Reactions were terminated by the addition of 1.5 µl of 20X STOP buffer, then incubated at 37°C for 30 min. with 0.1 mg/ml proteinase K. Reaction products were analyzed by electrophoresis on 0.7% agarose gel, dried and visualized by phosphorimaging. Strand transfer assays in Fig. 2.3d, were performed as described but with 5 µM of either GTP, ATP, inosine triphosphate (ITP, Jena Bioscience), xanthosine triphosphate (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' donor DNA 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 bovine serum albumin, 10 mM Mg $(OAc)_2$, 2 mM GTP). Reactions were initiated by the addition of 5 pmol of target DNAs, then incubated at 30°C for 2 hr. Reactions were terminated by the addition of 1.5 µl of 20X STOP buffer (85 mM EDTA, 5% SDS), then incubated at 37°C for 30 min. with 0.1 mg/ml proteinase K. 22 µl of deionized Formamide and 2 µl 100 mM NaOH was added, boiled for 5 min., then 6 µl to 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

In vivo excision assays were performed in triplicate essentially as previously described (Beall and Rio, 1996; Rio et al., 1986). 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). 24 hr after transfection, cells were washed with phosphate buffered saline (PBS), then harvested for immunoblot analysis and plasmid DNA recovery. Plasmid DNA was recovered as previously described (Rio et al., 1986), resuspended in 10 µl TE buffer (10 mM Tris-HCL pH 7.5, 1 mM EDTA). 1 µl was used to transform $RecA^-$ *E. coli* strain AG1574 (Kaufman and Rio, 1992) with a BioRad Gene Pulser as described by the manufacturer. Cells were grown for 1.5 hr at 37°C with shaking, then plated onto Luria broth plates containing either 100 µg/ml of ampicillin (1 µl of a 1:1000 dilution) or 100 µg/ml of ampicillin and 50 µg/ml of kanamycin (50 µl undiluted cells). Colonies were allowed to develop for 16 hr at 37°C, then counted.

Cryo-EM sample vitrification and data collection

Samples were vitrified using a Mark IV vitrobot (FEI). 4 µl of concentrated STC complex was applied to a Quantifoil 1.2/1.3 UltraAuFoil grid after being plasma cleaned (Solarus) for 10 sec. in air. After a 30 sec. incubation, the sample was blotted using a blot force of 8 pN and a blot time of 6 sec. Images were collected on an Arctica scope (Thermo Fisher) using a K2 detector (Gatan) using SerialEM (Mastronarde, 2005). During data collection, the stage was tilted by 40° to circumvent preferential orientation. 1857 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 sec. per movie. Please see Supplementary Table 2 for additional details.

Image processing

After motion correction with MotionCor2 (Zheng et al., 2017), particle-picking using Gautomatch, an initial per-micrograph CTF estimation and a subsequently per-particle CTF estimation were carried out using GCtf (Zhang, 2016) was completed. *Ab initio* model generation using cryoSPARC (Punjani et al., 2017) 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 (Zivanov et al., 2018) and an initial refinement in a ~4 Å reconstruction. Subsequent rounds of automatic refinement, followed by per-particle CTF refinement and Bayesian polishing, were iterated until convergence (Fig. 2.S2**c**) 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 (Fig. 2.S2**g**). The alignment parameters from this final C2 reconstruction was then refined without imposing symmetry (C1) resulting in an overall 3.9 Å structure (masked half-map), which matches the phase-randomized FSC estimate (Fig. 2.S2**f**).

De novo model building

An initial Cα trace and the initial sequence register were built manually using COOT (Emsley and Cowtan, 2004). Subsequent rounds of refinement using RosettaES (Frenz et al., 2017) filled in loops and rebuilt regions that were incorrect. The model for the nucleic acid was generated using COOT and refined with PHENIX (Adams et al., 2011). 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 rmsd difference between input and minimized atomic models is ~0.1 Å rmsd. The calculated final model-map FSC (0.5 cutoff) was 3.7 Å.

Map and model visualization

Maps were visualized in Chimera (Pettersen et al., 2004) and all model illustrations were prepared using either Chimera or ChimeraX (Goddard et al., 2018).

Data and Software Availability

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 (C

CHAPTER THREE

Elucidating the cellular function of human THAP9, an endogenous Drosophila P element transposase homolog present in the human genome

Abstract

Bioinformatic and biochemical analyses have identified a novel and evolutionarily conserved DNA binding domain, termed the THAP domain. This C_2CH zinc-binding DNA domain is found in animal genomes, vertebrates, invertebrates, *Drosophila* P element transposase, primates and in 12 human genes. Of the 12 *THAP* genes in humans, *THAP9*, in particular, is homologous to *Drosophila* P element transposase along its entire length and was shown to encode a functional DNA transposase that can mobilize *Drosophila* and zebrafish P elements in both insect and human cells. While the THAP9 protein possesses DNA transposase activity, the gene is expressed in human embryonic stem cells and orthologs are found across many animal species, including primates but not mouse or rat, a cellular function for THAP9 has not been identified. In an attempt to elucidate a cellular function for THAP9, we carried out genome-editing in human embryonic stem cells (hESCs) to either knockout or epitope tag the endogenous *THAP9* gene. Disruption of *THAP9* did not produce overt phenotypic changes in hESCs and did not affect differentiation into fibroblast-like cells, indicating that *THAP9* is likely not required for the hESC maintenance. However, endogenously epitope tagged THAP9 is translated, can be immunoprecipitated and localizes to the nucleus in hESCs. To determine potential THAP9 human genome cleavage and binding sites, we raised an antibody to purified, recombinant human THAP9 protein, performed BLESS to detect potential DNA cleavage site, a method used successfully to find Cas9 off-target genomic cleavage sites and ChIP-Nexus experiment, a chromatin immunoprecipitation method. The ongoing analysis and comparison of both the BLESS and ChIP-Nexus sequencing data should identify genomic binding sites, potential genomic DNA cleavage sites, motifs associated with human THAP9 DNA binding and cleavage and should uncover a cellular function for the human THAP9 gene.

Chapter 3

Introduction

Transposable elements are repetitive defined genetic sequences that can be mobilized throughout a genome by the action of an element-encoded transposase protein. Through this ability to self-propagate, transposons are ubiquitous in the genomes of all organisms, with few exceptions. For instance, in humans it is known that nearly 50% of the genome is derived from mobile elements (Lander et al., 2001). While transposon mobilization can be detrimental to the host, transposable elements are considered drivers of genome evolution by generating mutations and genetic polymorphisms, driving genome rearrangements, dispersing *cis*-regulatory sequences that modify gene expression networks or by supplying coding or non-coding sequences that can be adapted to carry out essential cellular functions (Bourque et al., 2018; Huang et al., 2012; 2015).

Over evolutionary timescales transposon-derived sequences have often been co-opted by the host to carryout various cellular functions in a process termed domestication. Through this domestication process, transposable elements have provided numerous protein coding regions, non-coding sequences, protein domains and even entire gene sequences that are involved in gene expression networks or that carry out other essential cellular processes. Long terminal repeat (LTR) promoter regions from endogenous retroviruses (ERVs) are enriched within interferon (IFN)-induced transcription factor binding sites and act as INF-inducible enhancers shaping the transcriptional landscape of the innate immune response (Chuong et al., 2016). Similarly, transcribed LINE1 RNAs act as a nuclear scaffold to regulate the gene expression landscape essential for maintaining embryonic stem cell identity in the mouse (Percharde et al., 2018).

Transposable elements can also provide genes or protein domains to the cellular genome for numerous and essential functions. Well-known examples of this domestication process include telomerase derived from non-LTR retrotransposons (Nakamura and Cech, 1998), the spliceosome derived from mobile group II introns (Rodríguez-Trelles et al., 2006) and RAG1/RAG2 V(D)J recombinase derived from a DNA-based transposon found in the lancelet Amphioxus (Huang et al., 2016; Kapitonov and Jurka, 2005). It is generally accepted that the recombinase signal sequences (RSSs) and recombinase-activating genes (RAG1/RAG2) of the V(D)J recombinase originated from an ancient DNA transposon belonging to the *Transib* superfamily of mobile eukaryotic elements (Kapitonov and Jurka, 2005). In a mechanism analogous to DNA transposon excision, the RAG1/RAG2 complex binds the RSSs and initiates an essential DNA recombination reaction at the immunoglobulin and T cell receptor genes of jawed vertebrates (Huang et al., 2016). This combinatorial assembly process, termed V(D)J recombination, allows for the massively diverse protein coding potential at the immunoglobulin and T cell receptor loci, generating a large repertoire of antibodies that form the basis of adaptive immunity in vertebrates (Zhang et al., 2019).

Like the adaptation of an ancient *Transib* element into V(D)J recombinase, a similar domestication process is thought to have occurred with the *Drosophila* P element (Quesneville et al., 2005). P elements are thought to have invaded *Drosophila* by horizontal transfer from a parasitic mite (Houck et al., 1991). The P element is a well-characterized DNA transposon flanked by 31 bp terminal inverted repeats that mobilizes via a “cut-and-paste” type mechanism. P element transposition is catalyzed by an element-encoded transposase protein (Majumdar and Rio, 2015). The Thanatos-associated protein, or THAP, domain is a common and well-conserved C_2CH zinc-coordinating DNA-binding domain found at the N-terminus of the *Drosophila* P element transposase (Roussigne et al., 2003b; Sabogal et al., 2010). Surprisingly, the THAP DNA binding domain is not only found in P element transposase, but in many cellular proteins across a wide range of animal taxa, such as *Ciona*, *C. elegans, Drosophila* and vertebrates (Roussigne et al., 2003b). For instance, 12 THAP domain-containing proteins (THAP0 - THAP11) have been identified in the human genome and have been shown to play roles in diverse cellular processes, including apoptosis (Roussigne et al., 2003a), histone deacetylation (Macfarlan et al., 2005) and maintenance of mouse embryonic stem cell pluripotency (Dejosez et al., 2008).

The human THAP9 gene exhibits a high degree of amino acid sequence homology along the entirety of *Drosophila* P element transposase (~25% identity and 40% similarity). More importantly, human THAP9 has retained DNA transposase catalytic activity, because it was shown to mobilize a genetically marked *Drosophila* P element in both *Drosophila* and human cell lines (Majumdar et al., 2013). Yet, the human THAP9 gene lacks the hallmarks of a mobile genetic element (Majumdar et al., 2013). That is, unlike active DNA transposons, the human THAP9 gene is present as a single copy, lacks both terminal inverted repeats and flanking target site duplication sequences, and is present in syntenic genomic locations in divergent genomes (Hammer, 2005; Quesneville et al., 2005). However, THAP9-like P element transposons have been identified in the sea squirt, *Ciona intestinalis,* a species from the most basal chordate lineage (Kimbacher et al., 2009) and the zebrafish, *Danio rerio* (Hagemann and Hammer, 2006; Hammer, 2005). The P-like elements in these two species display all the hallmarks of mobility, including terminal inverted repeats and 8 bp target site duplications. This suggests that the cellular THAP9 gene found in the human and other primate genomes may have been domesticated in early chordates from a P element-related DNA transposable element.

The activity of DNA transposase-related proteins is very uncommon among human cellular proteins and thus far only V(D)J recombinase, PGBD5 and THAP9 have been identified as functional DNA transposases found in the human genome (Agrawal et al., 1998; Henssen et al., 2015; Majumdar et al., 2013). However, unlike V(D)J recombinase, cellular function(s) and/or true DNA recombination sites for PGBD5 and THAP9 have not been identified.

To study the role of THAP9 in hESCs we initially carried out endogenous epitope tagging and gene knockout strategies and showed that the THAP9 protein is expressed at low levels in this cell type, is nuclear localized and is not essential for maintenance of the pluripotent state. To determine the genomic DNA cleavage and binding sites for THAP9, we raised and affinity-purified a rabbit polyclonal antibody against the entire THAP9 protein, generated inducible THAP9 HEK293 cell lines and carried out *in vivo* BLESS DNA cleavage and ChIP-Nexus DNA binding experiments.

Results

THAP9 disruption in hESCs

Publicly available bulk and single-cell RNA sequencing data suggest that human *THAP9* is expressed during embryogenesis and in hESCs (Davis et al., 2018; Dunham et al., 2012; Yan et al., 2013). Given this expression pattern (Fig. 3.1a), we aimed to understand the role of THAP9 in this cell type. To address this question, we employed two CRISPR/Cas9-mediated genome-editing approaches to disrupt the THAP9 gene and derive human pluripotent stem cells (WIBR#3) that lack THAP9 expression. In the first approach, we designed a single guide RNA (sgRNA) targeted downstream from the THAP9 ATG start codon and directed repair with a puromycin selection cassette driven from the phosphoglycerol kinase (PGK) promoter donor construct (Fig. 3.1b). Introduction of the PGK puromycin resistance (PURO) cassette at this position should disrupt THAP9 gene expression and generate THAP9-minus hESCs. After editing and puromycin selection we were able to derive several clonal hESC lines with insertion of the selection cassette at either one or both alleles (THAP9$^{-/-}$ #14) as confirmed by genomic DNA Southern blot hybridization analysis (Fig. 3.1c, clones 14 and 15).

In the second approach, we deleted exons of the THAP9 gene using two sgRNAs that cut at either positions -28 and +83 for exon 1 or +5413 and +5837 for exon 3 relative to the ATG start codon (Fig. 3.1d). Deletion of exon 1 should eliminate expression of full-length protein and disrupt exon 1 splicing whereas deletion of exon 3 should introduce a frameshift and premature stop codon within the THAP9 coding sequence and again generate THAP9-minus hESC lines. Clonal hESC lines derived from single-cells were validated for exon deletions by PCR and sequencing (Fig. 3.1d, e). Using this approach we were able to derive two clonal homozygous THAP9 exon 1 deleted lines (THAP9$^{E1\Delta/E1\Delta}$ #26, THAP9$^{E1\Delta/E1\Delta}$ #31) and a single heterozygous exon 3 deleted line (THAP9$^{E3\Delta/+}$ #40).

To evaluate the phenotypic impact of THAP9 disruption, both the THAP9$^{-/-}$ and THAP9$^{E1\Delta/E1\Delta}$ hESC lines were maintained in culture and observed for several weeks. Over this time the cell lines did not display any growth defects or phenotypic changes when compared to wild type hESCs (Fig. 3.2, left panels). To assess if the THAP9$^{-/-}$ or THAP9$^{E1\Delta/E1\Delta}$ mutant cells maintained pluripotency, we differentiated the mutant and wild type hESCs lines into embryoid bodies (EBs) and eventually into fibroblast-like cells (Fig. 3.2). Over the 15 day differentiation period we did not observe any overt phenotypic differences between THAP9$^{-/-}$, THAP9$^{E1\Delta/E1\Delta}$ and wild type hESCs.

Endogenously epitope-tagged THAP9 is expressed and localizes to the nucleus in hESC

The THAP9 gene contains five exons and can produce seven alternatively spliced isoforms that are predicted to undergo nonsense mediated RNA decay (NMD) or predicted to produce an N-terminally truncated protein variant (Hunt et al., 2018). To determine if the full-length THAP9 protein is produced in hESCs and to determine the sub-cellular localization, we endogenously tagged the THAP9 gene at the C-terminus with 2XStrep-Tag II and 3XFLAG epitopes (Jaeger et al., 2011)(Fig. 3.3). We employed CRISPR/Cas9-mediated genome editing using an sgRNA targeted near the stop codon and directed repair with a 2xStrep-Tag II 3xFLAG epitope and LoxP (floxed) PGK PURO cassette donor construct (Fig. 3.3a). An SV40 polyadenylation signal was included downstream of the epitope tags to allow for proper THAP9 transcript termination during puromycin selection (Fig. 3.3a). We were able to derive several clonal THAP9 2XStrep-Tag II 3XFLAG epitope-tagged hESC lines (THAP9-SF) at either one (Fig. 3.3b, clone #64) or both alleles (Fig. 3.3b, clone #67) as confirmed by Southern blot analysis. Transient transfection with Cre recombinase encoding mRNA was then performed to remove the PGK PURO selection cassette and restore the endogenous 3'UTR and polyadenylation sequence. Removal of the PGK PURO cassette was validated by PCR and susceptibility to puromycin selection on several hESC clonal lines derived from single-cells. Together, the editing strategy leaves behind 68 bp (containing a single loxP site and 18 bp of the THAP9 ORF) resulting in nearly scarless epitope-tagging of the endogenous THAP9 ORF. Edited cell lines were maintained in culture for several weeks and did not show any overt phenotypic changes when compared to unedited wild type hESCs.

To determine if the endogenously-tagged THAP9 protein is produced in hESCs, we performed immunoprecipitation with wild type and THAP9-SF hESCs whole cell lysates (Fig. 3.3c). Immunoprecipitation with α-FLAG antibody resin and immunoblotting with α-Strep-Tag II antibody revealed two protein bands, of ~ 110 kDa and 80 kDa, in the tagged cell lines that were absent in the wild type control samples. The slower mobility species is consistent with the mobility of full-length human THAP9-SF cDNA transiently expressed in HEK293 cells. The higher mobility species has been observed in some HEK293 THAP9-SF transfections and may correspond to a degradation product or the usage of a predicted alternative internal translation start codon (Hunt et al., 2018). We then performed indirect immunofluorescence (IF) staining for the 3XFLAG epitope on THAP9-SF and wild type cells to determine the subcellular localization of THAP9-SF in hESCs. As shown in representative cell images (Fig. 3.3d), we detected signal localized to the nucleus with a clear punctate pattern visible throughout the nucleoplasm in THAP9-SF hESCs (Fig. 3.3d, right), but not in wild type cells (Fig. 3.3d, left). In wild type cells, the signal was weak and diffuse over the entire cell, with some weak punctate staining visible the nucleus (Fig. 3.3d, DAPI), likely corresponding to background signal. Taken together, these data indicate that the tagged THAP9-SF protein is produced in hESCs and is localized to the nucleus, as would be expected for a DNA interacting protein.

Mapping THAP9 genomic cleavage and binding sites

Given THAP9-SF protein localization in hESCs and previously reported DNA transposase activity (Majumdar et al., 2013), we wanted to determine where THAP9 could cleave and bind within the human genome. Towards this aim, we raised polyclonal anti-THAP9 antibodies and engineered tetracycline-inducible THAP9 or THAP9-SF HEK293 cell lines. Rabbit polyclonal antibodies were raised against a recombinantly expressed and purified full-length human THAP9 protein, then subjected to THAP9 affinity purification (Fig. 3.4a). Both anti-THAP9 serum and antigen affinity purified anti-THAP9 antibodies detected protein species corresponding to THAP9 protein, in THAP9 or THAP9-SF transiently-transfected HEK293 cell lysates (Fig. 3.4a, TH9 and TH9-SF lanes). This THAP9 signal was altogether absent in the transiently transfected eGFP lysate control (Fig. 3.4a, eGFP lane), indicating the specificity of the affinity purified anti-THAP9 antibody.

To generate tetracycline-inducible THAP9 cell lines, we employed the Flp-In T-REx 293 system, from Invitrogen. THAP9-SF was detectable within 24 hr of induction with 1 µg/ml of tetracycline, and undetectable in uninduced cells (Fig. 3.4b). THAP9-SF expression did not appear to change HEK293 cell growth or viability. However, we did observe a slight increase in γ-H2AX histone signal by immunoblotting, consistent with a potential DNA double strand breaks (Fig. 3.4b). The histone variant, H2AX, becomes rapidly phosphorylated (γ-H2AX) in response to DNA damage (Kuo and Yang, 2008), and associates at the sites of V(D)J recombination–induced double-strand DNA breaks (Chen et al., 2000).

To identify the double-strand breaks potentially generated by induced THAP9 expression, we used direct *in situ* breaks labeling, enrichment on streptavidin and next-generation sequencing, or BLESS (Crosetto et al., 2013). The BLESS methodology labels and identifies double-strand DNA breaks with biotinylated adapters and next-generation sequencing (Fig. 3.4c). BLESS has been used to capture off-target sites for RNA-guided nucleases (Ran et al., 2015; Slaymaker et al., 2016; Zischewski et al., 2017). Transiently transfected THAP9, or tetracycline-induced THAP9-SF cells were subjected to BLESS, alongside control pUC18 or eGFP transiently transfected and uninduced THAP9-SF cells (Fig. 3.4d). Initial analysis of the BLESS data identified many break sites, several of which were not enriched or found in control samples, as depicted in a representative genome browser track (Fig. 3.4e). The BLESS data are being analyzed and correlated with the DNA binding data below to find sites of THAP9 binding near sites of DNA breaks.

To identify THAP9 binding sites and facilitate the analysis of the BLESS data, ChIP-Nexus (He et al., 2015) was performed on THAP9 and a catalytically-dead mutant THAP9 inducible cell lines. ChIP-Nexus is an improved version of ChIP-Exo (Rhee and Pugh, 2011), in which immunoprecipitated chromatin DNA is digested to the protein crosslinking site thereby providing near nucleotide resolution of protein-DNA interactions. The THAP9 wild type and THAP9(E613A) mutant inducible HEK293 cells were generated as before. Homology modeling of human THAP9 against the Cryo-EM structure of *Drosophila* P element transposase described in chapter 2 allowed for the identification of potential catalytic residues (D374 and E613). The ongoing analysis and comparison of both the BLESS and ChIP-Nexus sequencing data should identify THAP9 genomic binding and potential cleavage sites. Any identified binding/cleavage sequences will be biochemically validated with purified full-length THAP9 protein or the THAP9 DNA binding THAP domain, as well as tested in a cellular DNA cleavage assay as previously used for *Drosophila* P element transposase (Beall and Rio, 1996).

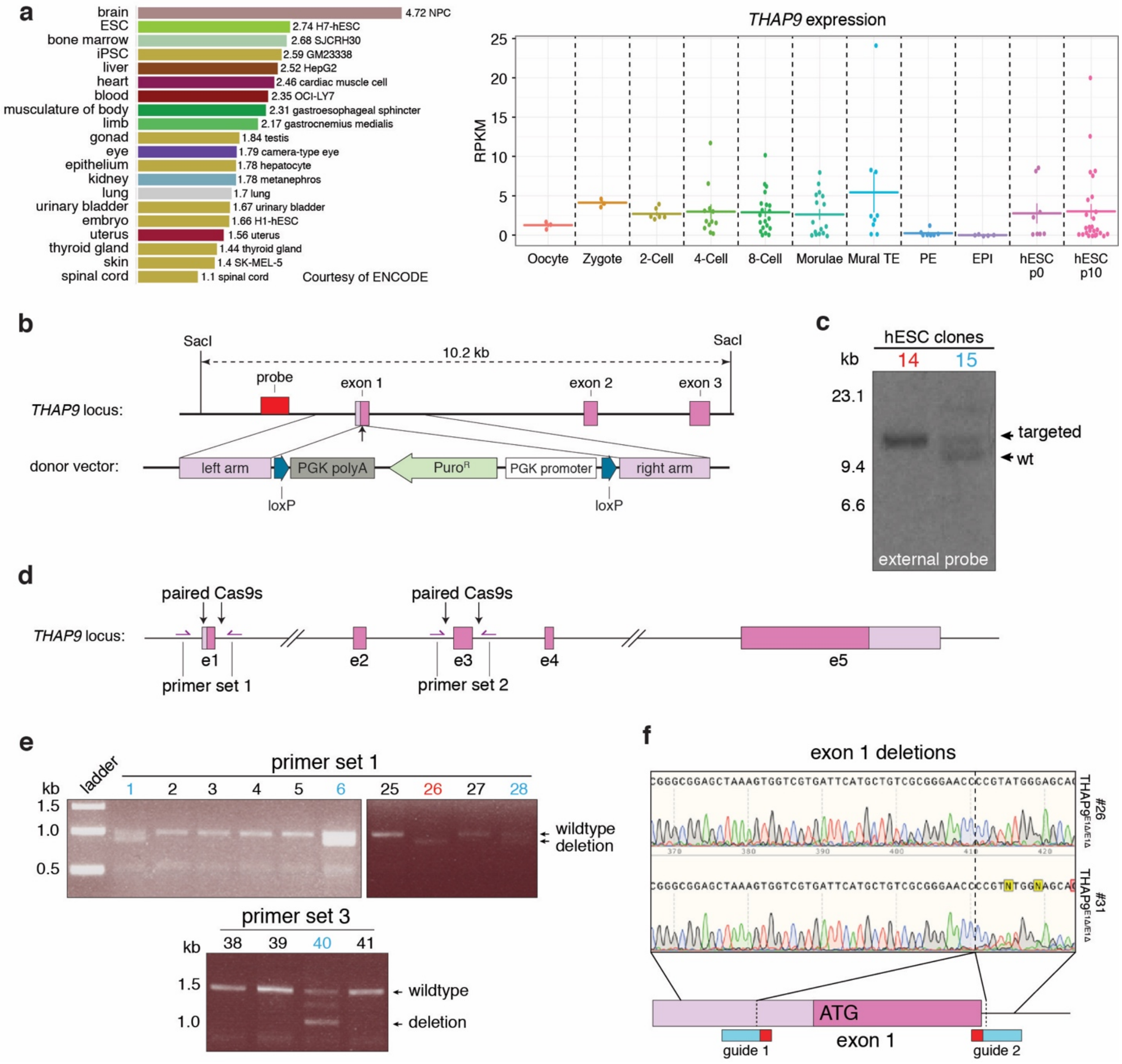
Discussion

THAP9 is one of 12 THAP domain-containing proteins identified in humans, and unlike the other THAP proteins, THAP9 is homologous to *Drosophila* P element transposase across its entire length. It was recently demonstrated that THAP9 possess DNA transposase catalytic activity and can mobilize Drosophila and zebrafish P elements in human and insect cells (Majumdar et al., 2013). DNA transposase activity is somewhat unusual for human genes and has only been observed for V(D)J recombinase (in rare circumstances; (Agrawal and Schatz, 1997; Messier et al., 2003)), THAP9, and more recently for PGBD5 (Henssen et al., 2015). With the exception of V(D)J recombinase, cellular functions have not been identified for PGBD5 and THAP9. Here, we aimed to identify the function of THAP9 by generating a knockout line in hESCs, a cell type in which the gene is expressed. However, the knockout hESC lines did not reveal an obvious phenotype, indicating that THAP9 is not required for the maintenance and proliferation of the stem cell state. Yet, endogenously epitope-tagged THAP9 can be immunoprecipitated as two distinct species and localizes to the nucleus by immunofluorescence analysis.

THAP9 encodes seven alternatively spliced mRNA isoforms that would generate truncated THAP9 protein and are predicted to undergo nonsense-mediated mRNA decay (NMD) (Hunt et al., 2018). In particular, transcript variant 5 (NM_001317776) has a longer 5'UTR and introduces a premature in-frame stop codon through exon inclusion splicing. Although this transcript is a candidate for NMD, leaky scanning could allow translation initiation at a downstream start codon to encode a shorter protein isoform, completely lacking the THAP DNA-binding domain. The size of this isoform is consistent with the size of the smaller species observed in immunoblots (data not shown) and immunoprecipitation of endogenously-tagged THAP9 from hESCs. Truncated forms of P element transposase are known to act as transpositional repressors or inhibitors (Lee et al., 1998; Misra and Rio, 1990) and this isoform may act to inhibit THAP9 function. However, we have not excluded the possibility that the smaller species is a simply a proteolytic degradation product.

Furthermore, ribosome profiling data in hESCs, neuronal precursor cells, and mature neurons indicate the is ribosome occupancy on an upstream open reading frame (uORF) (Blair et al., 2017). uORFs are prevalent across eukaryotic transcripts and generally act to transcriptionally regulate the downstream ORF (Somers et al., 2013). Taken together, these observations suggests that there is a high level of regulation on full-length THAP9 protein production, through nonfunctional mRNA isoforms or possibly through a uORF translational control mechanism. We note that we were unable to successfully target a THAP9 overexpression construct to the AAVS1 locus, consistent with the idea that THAP9 production is highly regulated and overexpression of the full-length protein may be harmful or toxic to hESCs.
A thorough analysis of THAP9 genomic binding sites and potential DNA cleavage sites could illuminate a cellular function for the protein. Toward this goal, we raised and purified a polyclonal THAP9 antibody, generated inducible THAP9 HEK293 cell lines, and carried out BLESS and ChIP-Nexus experiments. Ongoing bioinformatic and biochemical analyses should identify and validate motifs and reveal more details about potential THAP9 cleavage sites. Comparisons of both the BLESS and ChIP-Nexus sequencing data should identify THAP9 genomic binding and potential cleavage sites. Any identified binding/cleavage sequences will be biochemically validated with purified full-length THAP9 protein or the THAP9 DNA binding THAP domain, as well as tested in a cellular DNA cleavage assay as used for *Drosophila* P element transposase (Beall and Rio, 1996).

Figure 3.1

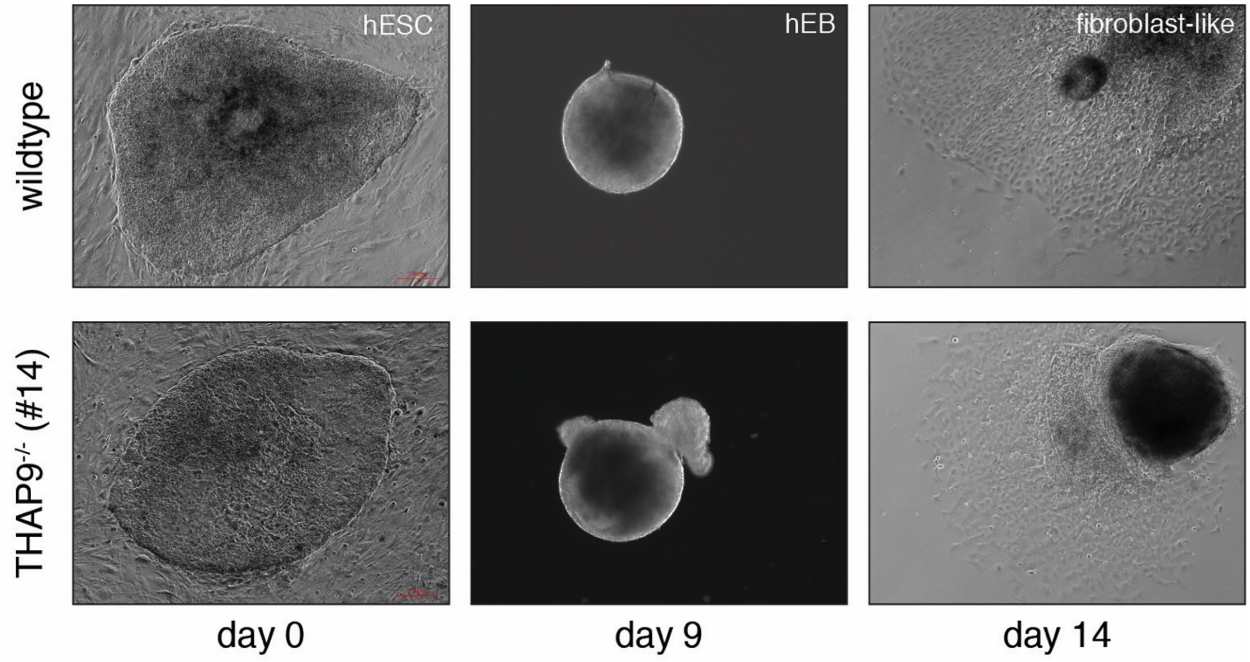


Targeted gene disruption of the *THAP9* gene in hESCs.

a, Publicly available RNA-sequencing data for expression of the THAP9 gene across various cell types. Left, THAP9 protein-coding transcript (ENST00000302236.5) expression levels from various cell types from the human ENCODE RNA expression database (Davis et al., 2018; Dunham et al., 2012). Log2 scale transcripts per kilobase million values are indicated before each cell type. Right, THAP9 expression profiles in single-cell human preimplantation embryos and embryonic stem cells from Yan et al., 2013 (Yan et al., 2013). Cell types are listed below. RPKM, reads per kilobase million; TE, trophectoderm; PE, primitive endoderm; EPI, epiblast; hESC, human embryonic stem cells; p0, passage 0; p10, passage 10.

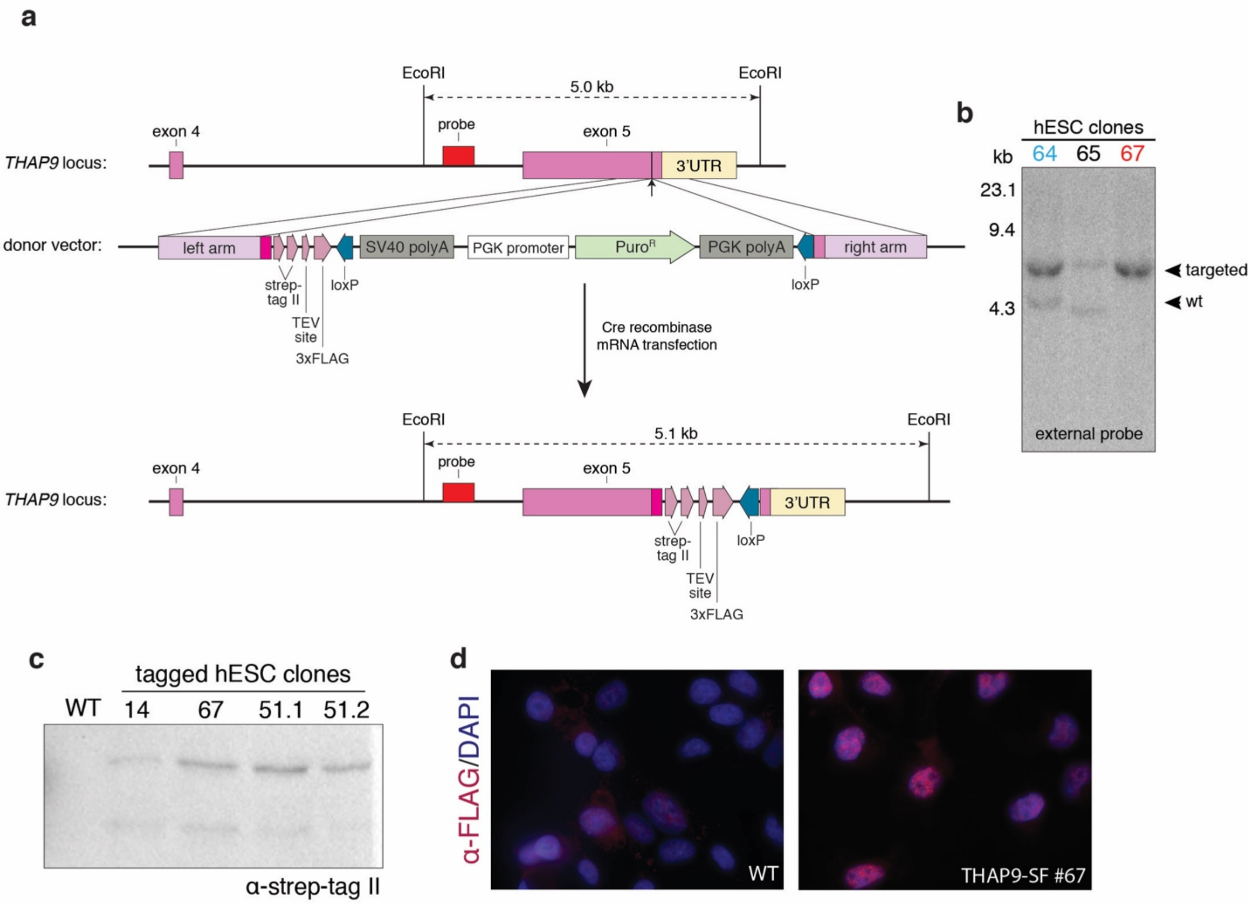
b, Schematic overview depicting the disruption strategy for the *THAP9* gene. Donor vector used to target the *THAP9* locus is depicted below. Red box, external probe used for Southern blot validation; pink boxes, first 3 exons of *THAP9*; arrow, Cas9 genomic cleavage site; loxP, loxP sites; polyA, polyadenylation sequence; Puro^{R}, puromycin resistance gene; PGK, phosphoglycerol kinase promoter. Not drawn to scale.
c, Representative southern blot DNA hybridization analysis of monoallelic or biallelic *THAP9*-targeted hESCs cells. Genomic DNA was digested with SacI and hybridized with the external probe. A correctly targeted monoallelic clone is indicated in cyan and biallelic targeted clone is indicated in red. Fragment sizes: wt, 10.2 kb, targeted, 12.1 kb.
d, Schematic overview depicting the exon deletion strategy for the *THAP9* gene. Pink boxes, exons of *THAP9*; purple arrows, location of PCR primers for validation; black arrows, paired Cas9 genomic cleavage sites. Not drawn to scale.
e, Agarose gel of PCR validation of exon 1 or exon 3 deletion. Genomic amplificon sizes for primer set 1: wild type, 1002 bp; deleted, 891 bp. Primer set 3: wild type, 1513 bp; deletion, 1089 bp. Monoallelic exon deletion lines are indicated in cyan and biallelic-targeted clone is indicated in red. PCR primers are listed in table 3.1.
f, Sanger sequencing profiles of PCR fragments from clones derived from the paired Cas9 deletion strategy confirms homozygous deletion of *THAP9* exon 1. Schematic overview of exon 1 deletion is depicted below. Pink box, *THAP9* exon 1; red/blue boxes, sgRNA locations; dotted lines, Cas9 cut sites. Not drawn to scale.

Figure 3.2



THAP9 $^{-/-}$ **hESCs differentiate into fibroblast-like cells.**
Representative phase-contrast images of *THAP9*$^{-/-}$ and wild type hESCs differentiation into embryoid bodies and fibroblast-like cells. hESC, human embryonic stem cells on day 0; hEB, human embryoid body after 9 days; fibroblast-like cells, after 14 day differentiation protocol.

Figure 3.3



Endogenously-tagged THAP9 protein localizes to the nucleus in hESCs.

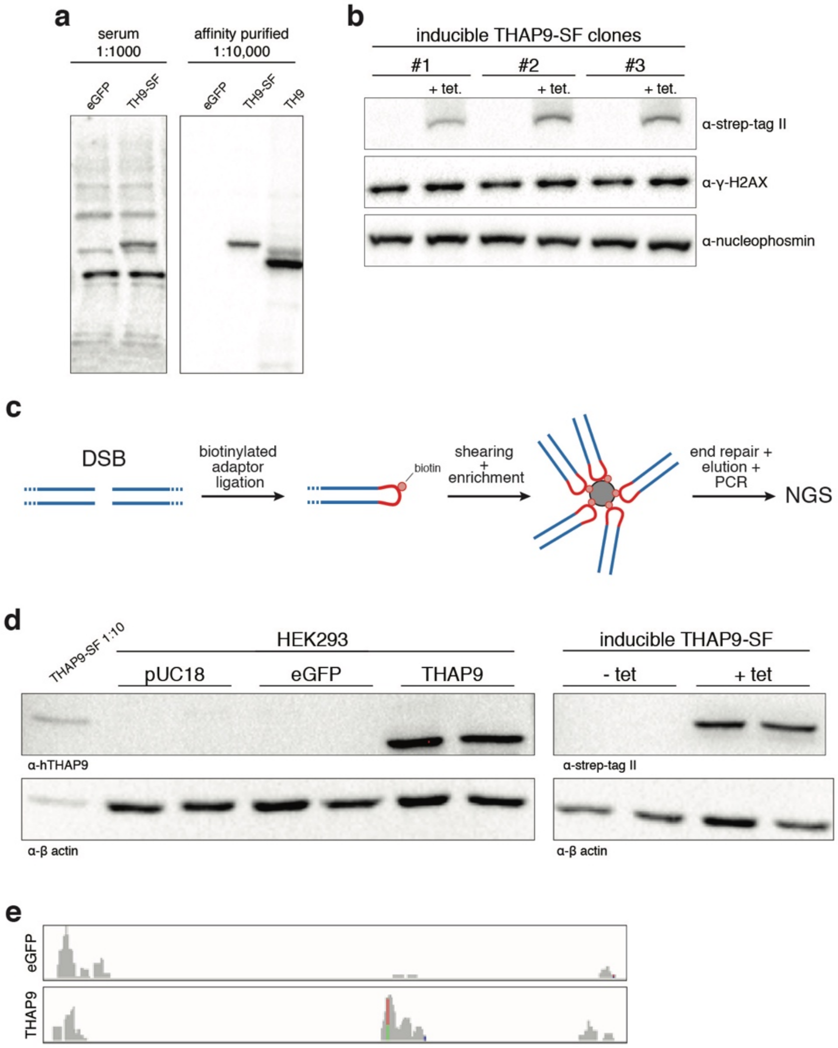
a, Schematic overview depicting the epitope tagging strategy for the *THAP9* gene (THAP9-SF). Donor targeting vector used to target the *THAP9* locus is depicted below. Red box, external DNA probe used for Southern blot validation; pink boxes, ORF of *THAP9*; beige box, 3' untranslated region (UTR); magenta box, altered codons for six terminal amino acids; strep-tag II, Strep-tag II epitope; TEV site, Tobacco Etch Virus (TEV) protease cleavage site; 3XFLAG, 3XFLAG epitope; Cas9 genomic cleavage site; loxP, loxP sites; SV40, polyomavirus simian virus 40; polyA, polyadenylation sequence; Puro^R, puromycin resistance gene; PGK, phosphoglycerol kinase promoter. Schematic overview of *THAP9* locus after transient Cre recombinase mRNA transfection is depicted below. Not drawn to scale.

b, Representative southern blot hybridization analysis of monoallelic or biallelic *THAP9* epitope tagged hESCs cells. Genomic DNA was digested with EcoRI and hybridized with the external probe. A correctly targeted monoallelic clone is indicated in cyan and biallelic targeted clone is indicated in red. Fragment sizes: wt, 5.0 kb, targeted, 7.3 kb.

c, Anti-FLAG M2 immunoprecipitation of endogenously tagged THAP9 and wild type (WT) hESC lines. α-Strep-Tag II was used for immunoblotting. Tagged THAP9 clonal lines are indicated above each lane.

d, Immunofluorescence images of wild type and THAP9-SF hESCs. Left panel shows α-FLAG (magenta) and DAPI (blue) staining of wild type cells. Right panel shows immune staining of endogenously-tagged THAP9 cells. WT, wildtype; THAP9-SF, THAP9-StrepTagII-3xFLAG.

Figure 3.4



Strategies for identifying THAP9 genomic cleavage and binding sites.

a, Immunoblot validation of THAP9 rabbit anti-serum and affinity purified anti-THAP9 antibodies. THAP9 and control eGFP proteins were transiently expressed in HEK293 cells. Dilutions are indicated above the lanes. eGFP, enhanced green fluorescent protein; TH9-SF, CMV promoter driven THAP9-StrepTagII-3xFLAG; TH9, Chicken beta actin gamma globulin synthetic promoter driven untagged THAP9. Note that untagged THAP9 has a faster mobility than SF tagged THAP9.

b, Immunoblot analysis of tetracycline inducible THAP9-SF cell lines. Clonal line numbers are indicated above. anti-nucleophosmin was used as a loading control.

c, Schematic overview of BLESS methodology. Adapted from (Crosetto et al., 2013).

d, Immunoblot analysis of transiently transfected and THAP9 inducible cells used in BLESS. Anti-β actin was used as a loading control.

e, Representative BLESS clusters for THAP9- or eGFP-transfected control cells.

Table 3.1

#	Name	Sequence	Description
172	EP hTHAP9 ORF1.1.F	GCAATCTTGTTAGGCCTGGA	Forward primer for external probe, upstream of hTHAP9 ORF1
174	EP hTHAP9 ORF1.1.R	ATGTGATACCGGAGGAGCAG	Reverse primer for external probe, upstream of hTHAP9 ORF1
161	hTHAP9 Exon1Deletion.2.F	CCCGATATCCTCCAGTTTCA	Forward primer for exon 1 deletion genotyping strategy
162	hTHAP9 Exon1Deletion.2.R	ATCAAATCCAGCCAGAATCG	Reverse primer for exon 1 deletion genotyping strategy
159	hTHAP9 Exon3Deletion.1.F	CCAAGTCCCAAGAGCTTCCT	Forward primer for exon 3 deletion genotyping strategy
160	hTHAP9 Exon3Deletion.1.R	GGTAGCCTTTCCATGGGTTT	Reverse primer for exon 3 deletion genotyping strategy
240	EP hTHAP9 ORF5.2.F	GCCTCTGTTGCCTGAAACTT	Forward primer for external probe, upstream of THAP9 stop codon
241	EP hTHAP9 ORF5.2.R	CAAAGCGCCAAGTCTTTCCT	Reverse primer for external probe, upstream of THAP9 stop codon

Number, name and description of primers used in this study.

Chapter 3

Materials and Methods

Cell culture

hESC experiments were performed in WIBR#3 hESCs (Lengner et al., 2010), NIH stem cell registry # 0079. hESCs culturing was carried out as previously described (Soldner et al., 2009). Briefly, all hESC lines were maintained on a layer of inactivated mouse embryonic fibroblasts (MEFs) in hESC medium (DMEM/F12 (Lifetech) supplemented with 15% fetal bovine serum (FBS, Lifetech), 5% KnockOutTM Serum Replacement (Lifetech), 1 mM glutamine (Lifetech), 1% non-essential amino acids (Lifetech), 0.1 mM β-mercaptoethanol (Sigma), 1000 U/ml penicillin/streptomycin (Lifetech), and 4 ng/ml FGF2 (Lifetech)). Cultures were enzymatically passaged every 5-7 days with collagenase type IV (Lifetech) (1.5 mg/ml) and gravitational sedimentation by washing three times in wash media (DMEM/F12 with 5% fetal bovine serum, and 1000 U/ml penicillin/streptomycin).

HEK 293 (UC Berkeley Cell Culture Facility) and Flp-In T-REx 293 (Invitrogen) cell lines were cultured in HEK media (DMEM (Gibco) supplemented with 10% heat inactivated FBS (Gibco), 1 mM glutamine, and 1000 U/ml penicillin/streptomycin). Cells were enzymatically passaged using 0.25% Trypsin-EDTA (Gibco) every 5 days. Tetracycline-inducible hTHAP9, hTHAP9(E613A) or hTHAP9-SF Flp-In T-REx 293 cell lines were generated according to the manufacturer's instructions. Where required, expression was induced by the addition of 1 μg/ml tetracycline to the culture media and incubated for 24 hr before harvesting.

Gene editing in hESCs

Cas9 and all sgRNAs were expressed using the pX330 plasmid (Cong et al., 2013). All targeting experiments were performed as previously described (Chiba and Hockemeyer, 2014). Briefly, hESCs were cultured in rho kinase (ROCK) inhibitor (Calbiochem; Y-27632) 24 hr before electroporation. To generate the hTHAP9 disruption and tagged cell lines, ~2.0 x 10^7 hESC were co-electroporated with 10 μg of pX330 plasmids targeting either the hTHAP9 first exon (CCCGAAGTTGCTCCGCAGTGGGC), or near the hTHAP9 stop codon (AGGCATTTGCTAAGTAACGATGG) and 40 μg of corresponding repair template plasmid. Cells were subsequently plated on MEF feeder layers (DR4 MEFs for puromycin selection) in hESC medium supplemented with ROCK inhibitor for the first 24 hr. Individual colonies were picked and expanded after puromycin selection (0.5 μg/ml) 10-14 days after electroporation. Gene editing was confirmed by Southern blot hybridization using probes amplified from hESC genomic DNA. Primers used to generate the Southern blot probes are described in table 3.1.

Endogenously tagged hTHAP9 cell lines were transfected with Cre recombinase mRNA to remove the puromycin selection cassette and restore the endogenous 3' UTR. Briefly, cells were plated on Matrigel (Corning) in a 12-well plate in mTeSR1 medium (Stem Cell Technologies). The next day a single well was transfected with 100 ng nGFP mRNA (Stemgent) and 400 ng Cre mRNA (Miltenyi Biotec) using Stemfect RNA transfection reagent (Stemgent). Cells were sorted for GFP fluorescence 72 hr after transfection. Single-cell derived hESC colonies were isolated and removal of the selection was validated by PCR and susceptibility to puromycin selection.

To generate the hTHAP9 exon deletion lines, ~2.0 x 10^7 hESC were co-electroporated with 7.5 µg of a GFP-expression plasmid and 15 µg of two pX330 plasmids targeting upstream (exon 1: TGCTGTCGCGGGAACCCCGAAGG, exon 3: CCTAACTAACTCTCCACAGCAAC) and downstream (exon 1: CCAGTGCGTATGGGAGCAGCCTC, exon 3: CCCCCTAGTAACCTGTAGTATTT) of either exon 1 or exon 3, respectively. GFP fluorescence positive cells were sorted 72 hr after electroporation. Single-cell derived hESC colonies were isolated and editing was confirmed by PCR followed by sequencing. Primers used to confirm exon deletions are described in Table 3.1**.**

Differentiation to fibroblast-like cells

For the formation of EBs hESC colonies were grown on petri dishes in fibroblast medium (DMEM/F12) supplemented with 15% FBS, 1 mM glutamine, 1% non-essential amino acids, and penicillin/streptomycin. After 9 days EBs were transferred to tissue culture dishes to attach. Fibroblast-like cells were passaged with 0.25% Trypsin EDTA (Gibco), triturated into a single-cell suspension and plated on tissue culture dishes. Cultures were maintained in fibroblast media and passaged every 6 days.

Immunoprecipitation

For analysis by immunoprecipitation, one well of hESCs from a six-well plate were mechanically harvested, washed with PBS, then lysed in 500 µl RIPA buffer (50 mM TRIS pH 8.0, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, 150 ml NaCl, 2mM EDTA, 50 mM NaF) supplemented with 1 mM PMSF (phenylmethylsulfonyl fluoride) and protease inhibitor cocktail. Lysates were briefly sonicated to shear DNA then centrifuged to remove insoluble material. The soluble fraction was incubated with 10 µl of anti-FLAG M2 affinity resin (Sigma), for 1 hr with rotation at 4°C. Resin was collected, washed three times in 1 ml RIPA buffer with rotation, then bound proteins were eluted by boiling in 40 µl SDS sample buffer. Eluates were subjected to electrophoresis on a 7.5% SDS-PAGE, transferred to nitrocellulose membranes, then immunoblotted with an anti-Strep-tag II antibody (Abcam, ab76949).

Antibodies

Anti-Strep-tag II (mouse, ab76949), anti-gamma-H2AX (rabbit, ab11174), anti-nucleophosmin (mouse, ab10530), anti-β actin (rabbit, ab8227) antibodies used for general immunoblotting were purchased from Abcam. Rabbit anti-hTHAP9 antibodies were raised in rabbits using purified recombinant hTHAP9 protein and antigen-affinity-purified in house.

Immunofluorescence

Indirect immunofluorescence analysis was carried out as follows: cells were briefly rinsed with PBS and fixed with 4% formaldehyde in PBS. Cells were then blocked with PBS supplemented with 0.3% Triton X-100 and 5% horse serum. Fixed cells were incubated with anti-FLAG M2 antibodies (mouse, F3165 Sigma 1:1000) overnight in PBSTB buffer (PBS supplemented with 0.3% Triton X-100 and 1% BSA). The next day the cells were washed with PBS and then stained with secondary antibodies (Alexa Fluor 546 goat anti-mouse,(Lifetech); 1:500), for 1 hr in PBSTB buffer. Cells were washed with PBS and stained with 1 ng/µl DAPI (Sigma) in PBS.

BLESS

BLESS was performed in transiently transfected HEK 293 cells or hTHAP9-SF inducible Flp-In T-REx 293 cell lines essentially as described (Crosetto et al., 2013; Slaymaker et al., 2015). Briefly, a total of 10 million cells were harvested 24 hr post-transfection or post-induction with 1 µg/ml tetracycline. Cells were fixed, nuclei were isolated and permeabilized then treated with proteinase K (Thermo) for 4 min at 37°C before inactivation with PMSF. Nuclei double strand breaks were repaired and then labeled with 200 mM of annealed proximal linkers overnight at 16°C. The next day, nuclei were washed and sheared by sonication to approximately 300 bp (BioRuptor). 20 µg were captured on streptavidin magnetic beads (Dynabeads MyOne Streptavidin C1), washed, then ligated to 200 mM of distal linker. DNA was then eluted by I-SceI (NEB) digestion for 4 hr at 37°C, then PCR-enriched for 18 cycles before proceeding to library preparation with an NEB Ultra II Kit (NEB). For the negative controls, cells were either uninduced or transfected with pUC18 or eGFP plasmids and were processed alongside. Lipofectamine 2000 was used according to the manufacturer’s instructions for all DNA transfections.

ChIP-Nexus

ChIP-Nexus was performed essentially as previously described (He et al., 2015). Briefly, 50 million THAP9 or THAP9(E613A) induced cells were fixed in 1% formaldehyde, quenched, and resuspended in lysis buffer. Chromatin from the lysates was sonicated to approximately 500 bp (BioRuptor). DNA was immunoprecipitated from 300 µl chromatin extract with 100 µl protein A Dynabeads (ThermoFisher) bound to anti-THAP9 rabbit antibody. DNA repair, adapter ligation, lambda exonuclease (NEB) treatment, DNA elution and circularization with CircLigase II (Lucigen) were carried out as described (He et al., 2015) before proceeding to library preparation with an NEB Ultra II Kit. THAP9 or THAP9(E613A) expression was induced with 1 µg/ml tetracycline 24 hr before fixation.

REFERENCES

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Echols, N., Headd, J.J., Hung, L.-W., Jain, S., Kapral, G.J., Grosse Kunstleve, R.W., et al. (2011). The Phenix software for automated determination of macromolecular structures. Methods *55*, 94–106.

Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X., et al. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. *11*, R119.

Agrawal, A., and Schatz, D.G. (1997). RAG1 and RAG2 form a stable postcleavage synaptic complex with DNA containing signal ends in V(D)J recombination. Cell *89*, 43–53.

Agrawal, A., Eastman, Q.M., and Schatz, D.G. (1998). Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature *394*, 744–751.

Aravind, L. (2000). The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. Trends Biochem. Sci. *25*, 421–423.

Au, T.K., Pathania, S., and Harshey, R.M. (2004). True reversal of Mu integration. Embo J. *23*, 3408–3420.

Ballandras-Colas, A., Brown, M., Cook, N.J., Dewdney, T.G., Demeler, B., Cherepanov, P., Lyumkis, D., and Engelman, A.N. (2016). Cryo-EM reveals a novel octameric integrase structure for betaretroviral intasome function. Nature *530*, 358–361.

Ballandras-Colas, A., Maskell, D.P., Serrao, E., Locke, J., Swuec, P., Jonsson, S.R., Kotecha, A., Cook, N.J., Pye, V.E., Taylor, I.A., et al. (2017). A supramolecular assembly mediates lentiviral DNA integration. Science *355*, 93–95.

Beall, E.L., and Rio, D.C. (1996). 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.

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

Beall, E.L., and Rio, D.C. (1998). Transposase makes critical contacts with, and is stimulated by, single-stranded DNA at the P element termini in vitro. *17*, 2122–2136.

Beauregard, A., Curcio, M.J., and Belfort, M. (2008). The Take and Give Between Retrotransposable Elements and their Hosts. Annu. Rev. Genet. *42*, 587–617.

References

Bhasin, A., Goryshin, I.Y., and Reznikoff, W.S. (1999). Hairpin formation in Tn5 transposition. Journal of Biological Chemistry *274*, 37021–37029.

Bingham, P.M., Kidwell, M.G., and Rubin, G.M. (1982). The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. Cell *29*, 995–1004.

Blair, J.D., Hockemeyer, D., Doudna, J.A., Bateup, H.S., and Floor, S.N. (2017). Widespread Translational Remodeling during Human Neuronal Differentiation. CellReports *21*, 2005–2016.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about transposable elements. Genome Biol. *19*.

Bouuaert, C.C., Walker, N., Liu, D., and Chalmers, R. (2014). Crosstalk between transposase subunits during cleavage of the mariner transposon. Nucleic Acids Research *42*, 5799–5808.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current Protocols in Molecular Biology *109*, 21.29.1–.29.9.

Burton, B.M., and Baker, T.A. (2003). Mu transpososome architecture ensures that unfolding by ClpX or proteolysis by ClpXP remodels but does not destroy the complex. Chemistry & Biology *10*, 463–472.

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato-Recupero, G., and Martin, C. (2012). Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges. Plant Cell *24*, 1242–1255.

Chatterji, M., Tsai, C.L., and Schatz, D.G. (2006). Mobilization of RAG-generated signal ends by transposition and insertion in vivo. Mol. Cell. Biol. *26*, 1558–1568.

Chen, H.T., Bhandoola, A., Difilippantonio, M.J., Zhu, J., Brown, M.J., Tai, X.G., Rogakou, E.P., Brotz, T.M., Bonner, W.M., Ried, T., et al. (2000). Response to RAG-mediated V(D)J cleavage by NBS1 and gamma-H2AX. Science *290*, 1962–1964.

Chiba, K., and Hockemeyer, D. (2014). Genome Editing in Human Pluripotent Stem Cells Using Site-Specific Nucleases. In Chromosomal Mutagenesis, (New York, NY: Springer New York), pp. 267–280.

Chow, S.A., Vincent, K.A., Ellison, V., and Brown, P.O. (1992). Reversal of Integration and Dna Splicing Mediated by Integrase of Human-Immunodeficiency-Virus. Science *255*, 723–726.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087.

References

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics 2013 14:12 *18*, 71–86.

Clubb, R.T., Omichinski, J.G., Savilahti, H., Mizuuchi, K., Gronenborn, A.M., and Clore, G.M. (1994). A Novel Class of Winged Helix-Turn-Helix Protein - the Dna-Binding Domain of Mu Transposase. Structure/Folding and Design *2*, 1041–1048.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. Science *339*, 819–823.

Coupland, G., Plum, C., Chatterjee, S., Post, A., and Starlinger, P. (1989). Sequences Near the Termini Are Required for Transposition of the Maize Transposon Ac in Transgenic Tobacco Plants. Proc Natl Acad Sci USA *86*, 9385–9388.

Craig, N.L. (1991). Tn7 - a Target Site-Specific Transposon. Mol. Microbiol. *5*, 2569–2573.

Craig, N.L. (1997). Target site selection in transposition. Annu. Rev. Biochem. *66*, 437–474.

Crosetto, N., Mitra, A., Silva, M.J., Bienko, M., Dojer, N., Wang, Q., Karaca, E., Chiarle, R., Skrzypczak, M., Ginalski, K., et al. (2013). Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat Meth *10*, 361–365.

Curcio, M.J., and Derbyshire, K.M. (2003). The outs and ins of transposition: From mu to kangaroo. Nat. Rev. Mol. Cell Biol. *4*, 865–877.

Cuypers, M.G., Trubitsyna, M., Callow, P., Forsyth, V.T., and Richardson, J.M. (2013). Solution conformations of early intermediates in Mos1 transposition. Nucleic Acids Research *41*, 2020–2033.

Daborn, P.J., Yen, J.L., Bogwitz, M.R., Le Goff, G., Feil, E., Jeffers, S., Tijet, N., Perry, T., Heckel, D., Batterham, P., et al. (2002). A single P450 allele associated with insecticide resistance in Drosophila. Science *297*, 2253–2256.

Davies, D.R., Goryshin, I.Y., Reznikoff, W.S., and Rayment, I. (2000). Three-Dimensional Structure of the Tn5 Synaptic Complex Transposition Intermediate. Science *289*, 77–85.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Research *46*, D794–D801.

Dawson, A., and Finnegan, D.J. (2003). Excision of the Drosophila Mariner Transposon Mos1: Comparison with Bacterial Transposition and V(D)J Recombination. Mol. Cell *11*, 225–235.

References

Dejosez, M., Krumenacker, J.S., Zitur, L.J., Passeri, M., Chu, L.-F., Songyang, Z., Thomson, J.A., and Zwaka, T.P. (2008). Ronin Is Essential for Embryogenesis and the Pluripotency of Mouse Embryonic Stem Cells. Cell *133*, 1162–1174.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005). Efficient Transposition of the piggyBac (PB) Transposon in Mammalian Cells and Mice. Cell *122*, 473–483.

Donzella, G.A., Jonsson, C.B., and Roth, M.J. (1996). Coordinated disintegration reactions mediated by Moloney murine leukemia virus integrase. J. Virol. *70*, 3909–3921.

Dornan, J., Grey, H., and Richardson, J.M. (2015). Structural role of the flanking DNA in mariner transposon excision. Nucleic Acids Research *43*, 2424–2432.

Dunham, I., Consortium, T.E.P., Davis, C., Doyle, F., Epstein, C.B., Harrow, J., Khatun, J., Lee, B.-K., Pauli, F., Sanyal, A., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.R., Hipps, K.W., et al. (2001). Intrinsically disordered protein. J. Mol. Graph. Model. *19*, 26–59.

Dyda, F., Hickman, JENKINS, T.M., Engelman, A., Craigie, R., and DAVIES, D.R. (1994). Crystal-Structure of the Catalytic Domain of Hiv-1 Integrase - Similarity to Other Polynucleotidyl Transferases. Science *266*, 1981–1986.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr Sect D Biol Crystallogr *60*, 2126–2132.

Engels, W.R. (1992). The Origin of P Elements in Drosophila-Melanogaster. Bioessays *14*, 681–686.

Engels, W.R. (1996). P elements in Drosophila. Current Topics in Microbiology and Immunology *204*, 103–123.

Engels, W.R., Johnsonschlitz, D.M., Eggleston, W.B., and Sved, J. (1990). High-Frequency P-Element Loss in Drosophila Is Homolog Dependent. Cell *62*, 515–525.

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

Francis, M.J., Roche, S., Cho, M.J., Beall, E., Min, B., Panganiban, R.P., and Rio, D.C. (2016). DrosophilaIRBP bZIP heterodimer binds P-element DNA and affects hybrid dysgenesis. Proc Natl Acad Sci USA *113*, 13003–13008.

Frenz, B., Walls, A.C., Egelman, E.H., Veesler, D., and DiMaio, F. (2017). RosettaES: a sampling strategy enabling automated interpretation of difficult cryo-EM maps. Nature Publishing Group *14*, 797–800.

Fugmann, S.D., Lee, A.I., Shockett, P.E., Villey, I.J., and Schatz, D.G. (2000). The rag proteins and V(D)J recombination: Complexes, ends, and transposition. Annu. Rev. Immunol. *18*, 495–527.

Fuller, J.R., and Rice, P.A. (2017). Target DNA bending by the Mu transpososome promotes careful transposition and prevents its reversal. eLife *6*, 257.

Gellert, M. (2002). V(D)J recombination: RAG proteins, repair factors, and regulation. Annu. Rev. Biochem. *71*, 101–132.

Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H., and Ferrin, T.E. (2018). UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci. *27*, 14–25.

Goodwin, T., and Poulter, R. (2004). A new group of tyrosine recombinase-encoding retrotransposons. Molecular Biology and Evolution *21*, 746–759.

Goryshin, I.Y., Miller, J.A., Kil, Y.V., Lanzov, V.A., and Reznikoff, W.S. (1998). Tn5/IS50 target recognition. Proc Natl Acad Sci USA *95*, 10716–10721.

Grabundzija, I., Messing, S.A., Thomas, J., Cosby, R.L., Bilic, I., Miskey, C., Gogol-Döring, A., Kapitonov, V., Diem, T., Dalda, A., et al. (2016). A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat Comms *7*, 10716.

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

Hammer, S.E. (2005). Homologs of Drosophila P transposons were mobile in zebrafish but have been domesticated in a common ancestor of chicken and human. Molecular Biology and Evolution *22*, 833–844.

Harshey, R.M. (2014). Transposable Phage Mu. Microbiol Spectr *2*.

Hartl, D.L. (2001). Discovery of the transposable element mariner. Genetics *157*, 471–476.

Hawley, R.S., Steuber, R.A., Marcus, C.H., Sohn, R., Baronas, D.M., Cameron, M.L., Zitron, A.E., and Chase, J.W. (1988). 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.

He, Q., Johnston, J., and Zeitlinger, J. (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat Biotechnol *33*, 395–401.

Henssen, A.G., Henaff, E., Jiang, E., Eisenberg, A.R., Carson, J.R., Villasante, C.M., Ray, M., Still, E., Burns, M., Gandara, J., et al. (2015). Genomic DNA transposition induced by human PGBD5. eLife *4*.

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

Hickman, A.B., Ewis, H.E., Li, X., Knapp, J.A., Laver, T., Doss, A.-L., Tolun, G., Steven, A.C., Grishaev, A., Bax, A., et al. (2014). Structural Basis of hAT Transposon End Recognition by Hermes, an Octameric DNA Transposase from Musca domestica. Cell *158*, 353–367.

Hickman, A.B., Perez, Z.N., Zhou, L., Musingarimi, P., Ghirlando, R., Hinshaw, J.E., Craig, N.L., and Dyda, F. (2005). Molecular architecture of a eukaryotic DNA transposase. Nat Struct Mol Biol *12*, 715–721.

Hickman, A.B., Chandler, M., and Dyda, F. (2010). Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. Critical Reviews in Biochemistry and Molecular Biology *45*, 50–69.

Hiom, K., Melek, M., and Gellert, M. (1998). DNA transposition by the RAG1 and RAG2 proteins: A possible source of oncogenic translocations. Cell *94*, 463–470.

Hof, A.E.V., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., Hall, N., Darby, A.C., and Saccheri, I.J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. Nature *534*, 102–105.

Houck, M.A., Clark, J.B., Peterson, K.R., and Kidwell, M.G. (1991). Possible Horizontal Transfer of Drosophila Genes by the Mite Proctolaelaps-Regalis. Science *253*, 1125–1129.

Huang, C.R.L., Burns, K.H., and Boeke, J.D. (2012). Active transposition in genomes. Annu. Rev. Genet. *46*, 651–675.

Huang, S., Tao, X., Yuan, S., Zhang, Y., Li, P., Beilinson, H.A., Zhang, Y., Yu, W., Pontarotti, P., Escriva, H., et al. (2016). Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. Cell *166*, 102–114.

Huff, J.T., Zilberman, D., and Roy, S.W. (2016). Mechanism for DNA transposons to generate introns on genomic scales. Nature *538*, 533–.

Hunt, S.E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., Parton, A., Armean, I.M., Trevanion, S.J., Flicek, P., et al. (2018). Ensembl variation resources. Database (Oxford) *2018*.

Ivics, Z., Hackett, P.B., Plasterk, R.H., and Izsvák, Z. (1997). Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell *91*, 501–510.

References

Jaeger, S., Gulbahce, N., Cimermancic, P., Kane, J., He, N., Chou, S., D'Orso, I., Fernandes, J., Jang, G., Frankel, A.D., et al. (2011). Purification and characterization of HIV-human protein complexes. Methods *53*, 13–19.

Johnson, R.C., Yin, J., and Reznikoff, W.S. (1982). Control of Tn5 Transposition in Escherichia-Coli Is Mediated by Protein From the Right Repeat. Cell *30*, 873–882.

Jonsson, C.B., Donzella, G.A., and Roth, M.J. (1993). Characterization of the forward and reverse integration reactions of the Moloney murine leukemia virus integrase protein purified from Escherichia coli. Journal of Biological Chemistry *268*, 1462–1469.

Kapitonov, V.V., and Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. Proc Natl Acad Sci USA *103*, 4540–4545.

Kapitonov, V.V., and Jurka, J. (2005). RAG1 Core and V(D)J Recombination Signal Sequences Were Derived from Transib Transposons. Plos Biol *3*, e181–14.

Kaufman, P.D., and Rio, D.C. (1992). P element transposition in vitro proceeds by a cut-and-paste mechanism and uses GTP as a cofactor. Cell *69*, 27–39.

Kaufman, P.D., Doll, R.F., and Rio, D.C. (1989). Drosophila P element transposase recognizes internal P element DNA sequences. Cell *59*, 359–371.

Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Comms 1–10.

Kazazian, H.H., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., and Antonarakis, S.E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature *332*, 164–166.

Kazazian, H.H.J., and Moran, J.V. (2017). Mobile DNA in Health and Disease. New England Journal of Medicine *377*, 361–370.

Khurana, J.S., Wang, J., Xu, J., Koppetsch, B.S., Thomson, T.C., Nowosielska, A., Li, C., Zamore, P.D., Weng, Z., and Theurkauf, W.E. (2011). Adaptation to P Element Transposon Invasion in Drosophila melanogaster. Cell *147*, 1551–1563.

Kidwell, M.G. (1992). Horizontal Transfer of P-Elements and Other Short Inverted Repeat Transposons. Genetica *86*, 275–286.

Kidwell, M.G., Kidwell, J.F., and Sved, J.A. (1977). Hybrid dysgenesis in Drosophila melanogaster: a syndrome of aberrant traits including mutation, sterility and male recombination. Genetics *86*, 813–833.

Kim, M.-S., Chuenchor, W., Chen, X., Cui, Y., Zhang, X., Zhou, Z.H., Gellert, M., and Yang, W. (2018). Cracking the DNA Code for V(D)J Recombination. Mol. Cell *70*, 358–370.e4.

Kim, M.-S., Lapkouski, M., Yang, W., and Gellert, M. (2015). Crystal structure of the V(D)J recombinase RAG1-RAG2. Nature *518*, 507–511.

Kimbacher, S., Gerstl, I., Velimirov, B., and Hagemann, S. (2009). Drosophila P transposons of the urochordata Ciona intestinalis. Mol Genet Genomics *282*, 165–172.

Klompe, S.E., Vo, P.L.H., Halpin-Healy, T.S., and Sternberg, S.H. (2019). Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. Nature 1.

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

Kuo, L.J., and Yang, L.-X. (2008). gamma-H2AX - A novel biomarker for DNA double-strand breaks. In Vivo *22*, 305–309.

la Cruz, de, N.B., Weinreich, M.D., Wiegand, T.W., Krebs, M.P., and Reznikoff, W.S. (1993). Characterization of the Tn5 transposase and inhibitor proteins: a model for the inhibition of transposition. Journal of Bacteriology *175*, 6932–6938.

Lander, E.S., Consortium, I.H.G.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lapkouski, M., Chuenchor, W., Kim, M.-S., Gellert, M., and Yang, W. (2015). Assembly Pathway and Characterization of the RAG1/2-DNA Paired and Signal-end Complexes. Journal of Biological Chemistry *290*, 14618–14625.

Laski, F.A., Rio, D.C., and Rubin, G.M. (1986). Tissue specificity of Drosophila P element transposition is regulated at the level of mRNA splicing. Cell *44*, 7–19.

Lee, C.C., Beall, E.L., and Rio, D.C. (1998). DNA binding by the KP repressor protein inhibits P-element transposase activity in vitro. Embo J. *17*, 4166–4174.

Lee, C.C., Mul, Y.M., and Rio, D.C. (1996). The Drosophila P-element KP repressor protein dimerizes and interacts with multiple sites on P-element DNA. Mol. Cell. Biol. *16*, 5616–5622.

Lengner, C.J., Gimelbrant, A.A., Erwin, J.A., Cheng, A.W., Guenther, M.G., Welstead, G.G., Alagappan, R., Frampton, G.M., Xu, P., Muffat, J., et al. (2010). Derivation of Pre-X Inactivation Human Embryonic Stem Cells under Physiological Oxygen Concentrations. Cell *141*, 872–883.

Li, S., Olson, W.K., and Lu, X.-J. (2019). Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. Nucleic Acids Research *47*, W26–W34.

Li, X., Harrell, R.A., Handler, A.M., Beam, T., Hennessy, K., and Fraser, M.J. (2005). piggyBac internal sequences are necessary for efficient transformation of target genomes. Insect Mol Biol *14*, 17–30.

Linheiro, R.S., and Bergman, C.M. (2008). Testing the palindromic target site model for DNA transposon insertion using the Drosophila melanogaster P-element. Nucleic Acids Research *36*, 6199–6208.

Macfarlan, T., Kutney, S., Altman, B., Montross, R., Yu, J.J., and Chakravarti, D. (2005). Human THAP7 is a chromatin-associated, histone tail-binding protein that represses transcription via recruitment of HDAC3 and nuclear hormone receptor corepressor. Journal of Biological Chemistry *280*, 7346–7358.

Maertens, G.N., Hare, S., and Cherepanov, P. (2010). The mechanism of retroviral integration from X-ray structures of its key intermediates. Nature *468*, 326–329.

Majumdar, S., and Rio, D.C. (2015). P Transposable Elements in Drosophila and other Eukaryotic Organisms. Microbiol Spectr *3*, MDNA3–0004–2014.

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

Mastronarde, D.N. (2005). Automated electron microscope tomography using robust prediction of specimen movements. Journal of Structural Biology *152*, 36–51.

McClintock, B. (1984). The Significance of Responses of the Genome to Challenge. Science *226*, 792–801.

Melek, M., and Gellert, M. (2000). RAG1/2-mediated resolution of transposition intermediates: Two pathways and possible consequences. Cell *101*, 625–633.

Messier, T.L., O'Neill, J.P., Hou, S.M., Nicklas, J.A., and Finette, B.A. (2003). In vivo transposition mediated by V(D)J recombinase in human T lymphocytes. Embo J. *22*, 1381–1388.

Min, B., Weinert, B.T., and Rio, D.C. (2004). Interplay between Drosophila Bloom's syndrome helicase and Ku autoantigen during nonhomologous end joining repair of P element-induced DNA breaks. Proc Natl Acad Sci USA *101*, 8906–8911.

Misra, S., and Rio, D.C. (1990). Cytotype Control of Drosophila-P Element Transposition - the 66 Kd Protein Is a Repressor of Transposase Activity. Cell *62*, 269–284.

Mitra, R., Fain-Thornton, J., and Craig, N.L. (2008). piggyBac can bypass DNA synthesis during cut and paste transposition. Embo J. *27*, 1097–1109.

Mizuuchi, K. (1992). Transpositional Recombination - Mechanistic Insights From Studies of Mu and Other Elements. Annu. Rev. Biochem. *61*, 1011–1051.

Mizuuchi, K., and Craigie, R. (1986). Mechanism of Bacteriophage-Mu-Transposition. Annu. Rev. Genet. *20*, 385–429.

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

Morgan, G.J., Hatfull, G.F., Casjens, S., and Hendrix, R.W. (2002). Bacteriophage Mu genome sequence: analysis and comparison with Mu-like prophages in Haemophilus, Neisseria and Deinococcus. J. Mol. Biol. *317*, 337–359.

Morris, E.R., Grey, H., McKenzie, G., Jones, A.C., and Richardson, J.M. (2016). A bend, flip and trap mechanism for transposon integration. eLife *5*.

Mul, Y.M., and Rio, D.C. (1997). Reprogramming the purine nucleotide cofactor requirement of Drosophila P element transposase in vivo. Embo J. *16*, 4441–4447.

Mullins, M.C., Rio, D.C., and Rubin, G.M. (1989). cis-acting DNA sequence requirements for P-element transposition. Genes Dev. *3*, 729–738.

Munoz-Lopez, M., and García-Pérez, J.L. (2010). DNA Transposons: Nature and Applications in Genomics. Curr. Genomics *11*, 115–128.

Nakai, H., Doseeva, V., and Jones, J.M. (2001). Handoff from recombinase to replisome: Insights from transposition. Proc Natl Acad Sci USA *98*, 8247–8254.

Nakamura, T.M., and Cech, T.R. (1998). Reversing time: origin of telomerase. Cell *92*, 587–590.

Narayanavari, S.A., Chilkunda, S.S., Ivics, Z., and Izsvák, Z. (2017). Sleeping Beauty transposition: from biology to applications. Critical Reviews in Biochemistry and Molecular Biology *52*, 18–44.

O'Hare, K., and Rubin, G.M. (1983). Structures of P transposable elements and their sites of insertion and excision in the Drosophila melanogaster genome. Cell *34*, 25–35.

Passos, D.O., Li, M., Yang, R., Rebensburg, S.V., Ghirlando, R., Jeon, Y., Shkriabai, N., Kvaratskhelia, M., Craigie, R., and Lyumkis, D. (2017). Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. Science *355*, 89–92.

Percharde, M., Lin, C.-J., Yin, Y., Guan, J., Peixoto, G.A., Bulut-Karslioglu, A., Biechele, S., Huang, B., Shen, X., and Ramalho-Santos, M. (2018). A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. Cell 1–35.

Peters, J.E., Makarova, K.S., Shmakov, S., and Koonin, E.V. (2017). Recruitment of CRISPR-Cas systems by Tn7-like transposons. Proc Natl Acad Sci USA *114*, E7358–E7366.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem *25*, 1605–1612.

Pflieger, A., Jaillet, J., Petit, A., Auge-Gouillou, C., and Renault, S. (2014). Target Capture during Mos1 Transposition. Journal of Biological Chemistry *289*, 100–111.

Polard, P., TonHoang, B., Haren, L., Betermier, M., Walczak, R., and Chandler, M. (1996). IS911-mediated transpositional recombination in vitro. J. Mol. Biol. *264*, 68–81.

Punjani, A., Rubinstein, J.L., Fleet, D.J., and Brubaker, M.A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat Meth *14*, 290–296.

Quesneville, H., Nouaud, D., and Anxolabehere, D. (2005). Recurrent recruitment of the THAP DNA-binding domain and molecular domestication of the P-transposable element. Molecular Biology and Evolution *22*, 741–746.

Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., et al. (2015). In vivo genome editing using Staphylococcus aureus Cas9. Nature *520*, 186–191.

Rebollo, R., Romanish, M.T., and Mager, D.L. (2012). Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annu. Rev. Genet. *46*, 21–42.

Reznikoff, W.S. (2008). Transposon Tn5. Annu. Rev. Genet. *42*, 269–286.

Rhee, H.S., and Pugh, B.F. (2011). Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell *147*, 1408–1419.

Rice, P., and Mizuuchi, K. (1995). Structure of the Bacteriophage-Mu Transposase Core - a Common Structural Motif for Dna Transposition and Retroviral Integration. Cell *82*, 209–220.

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

Richardson, J.M., Dawson, A., O'Hagan, N., Taylor, P., Finnegan, D.J., and Walkinshaw, M.D. (2006). Mechanism of Mos1 transposition: insights from structural analysis. Embo J. *25*, 1324–1334.

Rio, D.C. (1990). Molecular mechanisms regulating Drosophila P element transposition. Annu. Rev. Genet.

Rio, D.C., Laski, F.A., and Rubin, G.M. (1986). Identification and immunochemical analysis of biologically active Drosophila P element transposase. Cell *44*, 21–32.

Rodgers, K.K. (2017). Riches in RAGs: Revealing the V(D)J Recombinase through High-Resolution Structures. Trends Biochem. Sci. *42*, 72–84.

Rodríguez-Trelles, F., Tarrío, R., and Ayala, F.J. (2006). Origins and evolution of spliceosomal introns. Annu. Rev. Genet. *40*, 47–76.

Roiha, H., Rubin, G.M., and O'Hare, K. (1988). P element insertions and rearrangements at the singed locus of Drosophila melanogaster. Genetics *119*, 75–83.

Roussigne, M., Cayrol, C., Clouaire, T., Amalric, F., and Girard, J.-P. (2003a). THAP1 is a nuclear proapoptotic factor that links prostate-apoptosis-response-4 (Par-4) to PML nuclear bodies. Oncogene *22*, 2432–2442.

Roussigne, M., Kossida, S., Lavigne, A.-C., Clouaire, T., Ecochard, V., Glories, A., Amalric, F., and Girard, J.-P. (2003b). The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. Trends Biochem. Sci. *28*, 66–69.

Ru, H., Chambers, M.G., Fu, T.-M., Tong, A.B., Liao, M., and Wu, H. (2015). Molecular Mechanism of V(D)J Recombination from Synaptic RAG1-RAG2 Complex Structures. Cell *163*, 1138–1152.

Ru, H., Mi, W., Zhang, P., Alt, F.W., Schatz, D.G., Liao, M., and Wu, H. (2018a). DNA melting initiates the RAG catalytic pathway. Nat Struct Mol Biol 1–15.

Ru, H., Zhang, P., and Wu, H. (2018b). Structural gymnastics of RAG-mediated DNA cleavage in V(D)J recombination. Current Opinion in Structural Biology *53*, 178–186.

Sabogal, A., Lyubimov, A.Y., Corn, J.E., Berger, J.M., and Rio, D.C. (2010). THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves. Nat Struct Mol Biol *17*, 117–U145.

Sano, K.-I., Maeda, K., Oki, M., and Maéda, Y. (2002). Enhancement of protein expression in insect cells by a lobster tropomyosin cDNA leader sequence. FEBS Letters *532*, 143–146.

Sassaman, D.M., Dombroski, B.A., Moran, J.V., Kimberland, M.L., Naas, T.P., DeBerardinis, R.J., Gabriel, A., Swergold, G.D., and Kazazian, H.H. (1997). Many human L1 elements are capable of retrotransposition. Nat Genet *16*, 37–43.

Schatz, D.G., and Swanson, P.C. (2011). V(D)J Recombination: Mechanisms of Initiation. Annu. Rev. Genet. *45*, 167–202.

Sekelsky, J. (2017). DNA Repair in Drosophila: Mutagens, Models, and Missing Genes. Genetics *205*, 471–490.

Shapiro, J.A. (1979). Molecular-Model for the Transposition and Replication of Bacteriophage Mu and Other Transposable Elements. Proc Natl Acad Sci USA *76*, 1933–1937.

Siebel, C.W., Fresco, L.D., and Rio, D.C. (1992). 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.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2015). Rationally engineered Cas9 nucleases with improved specificity. Science 1–7.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. Science *351*, 84–88.

Soldner, F., Hockemeyer, D., Beard, C., Gao, Q., Bell, G.W., Cook, E.G., Hargus, G., Blak, A., Cooper, O., Mitalipova, M., et al. (2009). Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. Cell *136*, 964–977.

Somers, J., Pöyry, T., and Willis, A.E. (2013). A perspective on mammalian upstream open reading frame function. Int. J. Biochem. Cell Biol. *45*, 1690–1700.

Strecker, J., Ladha, A., Gardner, Z., Schmid-Burgk, J.L., Makarova, K.S., Koonin, E.V., and Zhang, F. (2019). RNA-guided DNA insertion with CRISPR-associated transposases. Science.

Surette, M.G., Buch, S.J., and Chaconas, G. (1987). Transpososomes: stable protein-DNA complexes involved in the in vitro transposition of bacteriophage Mu DNA. Cell *49*, 253–262.

Tan, Y.Z., Baldwin, P.R., Davis, J.H., Williamson, J.R., Potter, C.S., Carragher, B., and Lyumkis, D. (2017). Addressing preferred specimen orientation in single-particle cryo-EM through tilting. Nat Meth *14*, 793–796.

Tang, M., Cecconi, C., Bustamante, C., and Rio, D.C. (2007). Analysis of P element transposase protein-DNA interactions during the early stages of transposition. Journal of Biological Chemistry *282*, 29002–29012.

Tang, M., Cecconi, C., Kim, H., Bustamante, C., and Rio, D.C. (2005). Guanosine triphosphate acts as a cofactor to promote assembly of initial P-element transposase-DNA synaptic complexes. Genes Dev. *19*, 1422–1425.

Teixeira, F.K., Okuniewska, M., Malone, C.D., Coux, R.-X., Rio, D.C., and Lehmann, R. (2017). piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. Nature Publishing Group *552*, 268–272.

Tellier, M., Bouuaert, C.C., and Chalmers, R. (2015). Mariner and the ITm Superfamily of Transposons. Microbiol Spectr *3*, MDNA3–0033–2014.

Thein, S.L. (2013). The Molecular Basis of beta-Thalassemia. Cold Spring Harb Perspect Med *3*.

Tower, J., Karpen, G.H., Craig, N., and Spradling, A.C. (1993). Preferential transposition of Drosophila P elements to nearby chromosomal sites. Genetics *133*, 347–359.

Trowitzsch, S., Bieniossek, C., Nie, Y., Garzoni, F., and Berger, I. (2010). New baculovirus expression tools for recombinant protein complex production. Journal of Structural Biology *172*, 45–54.

van Gent, D.C., Hiom, K., Paull, T.T., and Gellert, M. (1997). Stimulation of V(D)J cleavage by high mobility group proteins. Embo J. *16*, 2665–2670.

van Gent, D.C., Mizuuchi, K., and Gellert, M. (1996). Similarities between initiation of V(D)J recombination and retroviral integration. Science *271*, 1592–1594.

Villasante, A., de Pablos, B., Méndez-Lago, M., and Abad, J.P. (2008). Telomere maintenance in Drosophila - Rapid transposon evolution at chromosome ends. Cell Cycle *7*, 2134–2138.

Warren, W.D., Atkinson, P.W., and Obrochta, D.A. (1994). The Hermes Transposable Element From the House-Fly, Musca-Domestica, Is a Short Inverted Repeat-Type Element of the Hobo, Ac, and Tam3 (Hat) Element Family. Genet. Res. *64*, 87–97.

Weinert, B.T., Min, B., and Rio, D.C. (2005). P element excision and repair by non-homologous end joining occurs in both G1 and G2 of the cell cycle. DNA Repair *4*, 171–181.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 2013 14:12 *8*, 973–982.

Wright, A.V., Liu, J.-J., Knott, G.J., Doxzen, K.W., Nogales, E., and Doudna, J.A. (2017). Structures of the CRISPR genome integration complex. Science *357*, 1113–1118.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol *20*, 1131–1139.

Yanagihara, K., and Mizuuchi, K. (2002). Mismatch-targeted transposition of Mu: A new strategy to map genetic polymorphism. Proc Natl Acad Sci USA *99*, 11317–11321.

Yang, W., and Steitz, T.A. (1995). Recombining the Structures of Hiv Integrase, Ruvc and Rnase-H. Structure/Folding and Design *3*, 131–134.

Yang, W., Hendrickson, W.A., Crouch, R.J., and Satow, Y. (1990). Structure of Ribonuclease-H Phased at 2-a Resolution by Mad Analysis of the Selenomethionyl Protein. Science *249*, 1398–1405.

Yang, W., Lee, J.Y., and Nowotny, M. (2006). Making and breaking nucleic acids: Two-Mg2+-ion catalysis and substrate specificity. Mol. Cell *22*, 5–13.

Yin, F.F., Bailey, S., Innis, C.A., Ciubotaru, M., Kamtekar, S., Steitz, T.A., and Schatz, D.G. (2009). Structure of the RAG1 nonamer binding domain with DNA reveals a dimer that mediates DNA synapsis. Nat Struct Mol Biol *16*, 499–508.

Yin, Z., Lapkouski, M., Yang, W., and Craigie, R. (2012). Assembly of prototype foamy virus strand transfer complexes on product DNA bypassing catalysis of integration. Protein Sci. *21*, 1849–1857.

Yin, Z., Shi, K., Banerjee, S., Pandey, K.K., Bera, S., Grandgenett, D.P., and Aihara, H. (2016). Crystal structure of the Rous sarcoma virus intasome. Nature *530*, 362–366.

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

Zhang, K. (2016). Gctf: Real-time CTF determination and correction. Journal of Structural Biology *193*, 1–12.

Zhang, Y., Cheng, T.C., Huang, G., Lu, Q., Surleac, M.D., Mandell, J.D., Pontarotti, P., Petrescu, A.J., Xu, A., Xiong, Y., et al. (2019). Transposon molecular domestication and the evolution of the RAG recombinase. Nature 1–23.

Zheng, S.Q., Palovcak, E., Armache, J.-P., Verba, K.A., Cheng, Y., and Agard, D.A. (2017). MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nature Publishing Group *14*, 331–332.

Zhou, L.Q., Mitra, R., Atkinson, P.W., Hickman, Dyda, F., and Craig, N.L. (2004). Transposition of hAT elements links transposable elements and V(D)J recombination. Nature *432*, 995–1001.

Zischewski, J., Fischer, R., and Bortesi, L. (2017). Detection of on-target and off-target mutations generated by CRISPR/Cas9 and other sequence-specific nucleases. Biotechnol. Adv. *35*, 95–104.

Zivanov, J., Nakane, T., Forsberg, B.O., Kimanius, D., Hagen, W.J., Lindahl, E., and Scheres, S.H. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION-3. eLife *7*, 163.

Craig, N. L. and Chandler, M. and Gellert, M. and Lambowitz, A. M. and Rice, P. A. and Sandmeyer, S. B.(ed). (2015). Mobile DNA III (American Society of Microbiology).