Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

presented by

M.Sc., Irma Querques

born in Lucera, Italy

Oral examination: 7th June 2018

MECHANISMS AND DESIGN OF *TC1/MARINER* TRANSPOSONS FOR GENOME ENGINEERING

Referees: Dr. Kiran Raosaheb Patil

Prof. Dr. Irmgard Sinning

SUMMARY

Transposons are DNA segments that autonomously move within and between genomes across the tree of life. *Tc1/mariners* in particular have frequently crossed species boundaries in nature and provide powerful broad-host-range genetic vectors. Among them, the *Sleeping Beauty* (*SB*) transposon inserts DNA in vertebrate genomes with extraordinarily high efficiency, making it a prime genetic tool with applications expanding to gene therapy clinical trials. Nevertheless, the molecular principles of *SB*'s distinctive activity remain elusive, greatly hampering its further development.

In the first part of this thesis, I investigated the molecular mechanisms of the *SB* transposon in comparison to *Human mariner 1* (*Hsmar1*), a representative transposon of the same superfamily. Using biochemical and biophysical techniques together with fluorescence-based assays, I have characterized the initial steps of *SB* and *Hsmar1* transposition and shown that the two transposons assemble their molecular machineries (or transpososomes) differently. By combining crystallographic data and SAXS-based modelling, I visualized the structural basis of these differences and explained how transpososome assembly is coupled to catalysis in the *Hsmar1* transposon. Moreover, the data demonstrated that the unique assembly pathway of *SB* largely contributes to its exceptional efficiency and that it can be chemically modulated to control insertion rates in living cells. I have further reconstituted *in vitro* the ordered series of events comprising *SB* transposition, including transposon end binding, cleavage, and integration, and dissected previously unrevealed molecular features of the process.
In the second part of my work, building on these mechanistic insights, I developed a novel SB transposase variant (hsSB) by employing a structure-based protein design approach. Using hsSB allowed for establishing a new genome engineering method based on the direct delivery of recombinant SB protein to cells. We showed that this new method, named SBprotAct, provides safer and more controlled genome modification of several cell types (including stem cells and human T cells), as compared to the state-of-art technology.

This work sheds first light on the molecular determinants of *SB* transposition and its hyper-activity, providing a unique resource for the rational design of improved genome engineering platforms for research and medicine.

ZUSAMMENFASSUNG

Transposons sind mobile genetische Elemente, die sich selbständig innerhalb des Genoms, sowie zwischen den Genomen verschiedener Spezies bewegen. Insbesondere *Tc1/mariners* haben häufig Speziesgrenzen in der Natur überschritten und stellen deshalb leistungsfähige genetische Vektoren für einen breiten Wirtsbereich dar. *Sleeping Beauty* (*SB*) Transposon übertragt DNS in die Genome von Vertebraten mit außergewöhnlich hoher Effizienz. Diese Eigenschaft macht *SB* immer häufiger zu einem genetischen Werkzeug in klinischen Studien zur Gentherapie. Die molekularen Prinzipien von *SB*'s ausgeprägter Aktivität sind weitgehend unbekannt, was eine weiter Entwicklung stark beeinträchtigt.

Im ersten Teil dieser Arbeit habe ich die molekularen Mechanismen des *SB* Transposons mit *Human mariner 1* (*Hsmar1*), einem repräsentativen Transposon aus der selben Superfamilie, verglichen. Ich habe die anfänglichen Schritte der Transposition von *SB* und *Hsmar1* mittels biochemischer und biophysikalischer Techniken, sowie fluoreszenzbasierten Untersuchungen, charakterisiert. Dies zeigte dass die beiden Transposons ihre molekularen Maschinen (Transpososomes) auf unterschiedliche Weise zusammenfügen. Durch die Verbindung von kristallografischen Daten und SAXS-basiertem Modellierungen zeige ich die strukturelle Grundlage dieser Unterschiede und erkläre wie das Zusammenfügen des Transpososomes von *Hsmar1* mit der Katalyse gekoppelt ist. Des Weiteren zeigen diese Daten, dass der einzigartige Assemblierungsprozess von *SB* zu dessen außergewöhnlich hoher Effizienz beiträgt. Diese Erkenntnis wiederum eröffnet die Möglichkeit Insertionsraten in lebenden Zellen chemisch zu kontrollieren. Außerdem habe ich die geordnete Reihenfolge der Ereignisse der *SB* Transposition, einschließlich Bindung, Ausschneiden und Integration des Transposons *in vitro* rekonstruiert und untersuchte bislang unbekannte molekulare Eigenschaften des Prozesses.
Im zweiten Teil meiner Arbeit, die auf diesen mechanistischen Erkenntnissen aufbaut, habe ich mittels strukturbasiertem Protein Design eine neuartige SB Transposase Variante (hsSB) entwickelt. Die Verwendung von hsSB erlaubte die Entwicklung einer neuen Methode zur Genomveränderung, basierend auf dem direkten Einschleusen von rekombinantem SB Protein in Zellen. Verglichen mit anderen modernsten Technologien erlaubt diese neue Methode, die wir SBprotAct getauft haben, eine sicherere und kontrolliertere Modifikation von Genomen verschiedenster Zelltypen, einschließlich Stammzellen und humaner T-Zellen.

Diese Arbeit gibt zum ersten mal Aufschluss über die molekularen Faktoren der *SB* Transposition und seiner Hyperaktivität und bietet eine einzigartige Quelle für das rationale Design von verbesserten Genom-Modifikationsplattformen für Wissenschaft und Medizin.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

aa	amino acid
Abs	Absorbance
Ac	*Activator*
Ala/A	Alanine
ALL	Acute lymphoblastic leukemia
Amp	Ampicillin
AMPD	2-amino-2-methyl-1,3-propanediol
APS	Ammonium persulphate
Arg/R	Arginine
Asn/N	Asparagine
ASO	Assembly site occlusion
Asp/D	Aspartic acid
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
ATCC	American Type Culture Collection
bp	Base pair
BSA	Bovine serum albumin
C-terminus	Carboxy-terminus
CAR	Chimeric antigen receptor
CAT	Catalytic domain
CD	Circular dichroism
χ^2	Chi-square
CHO cell	Chinese hamster ovary cell
Cm	Chloramphenicol
CPP	Cell-penetrating peptide
Cys/C	Cysteine
Da	Dalton
DAPI	4',6-diamidino-2-phenylindole
dATP	Deoxyadenosine triphosphate
DBD	DNA binding domain
dCTP	Deoxycytidine triphosphate
DEJ	Double end joining
dGTP	Deoxyguanosine triphosphate
Dmax	Maximum dimension
DMEM	Dulbecco's Modified Eagle Medium
DMF	Dimethylformamide
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNAse	Deoxyribonuclease
DR	Directed repeat
DRi	Inner directed repeat
DRo	Outer directed repeat
Ds	*Dissociation*
DSB	Double strand break

DSS	Dissucimidyl suberate
DTT	Dithiothreitol
dTTP	Dideoxythymidine triphosphate
EDTA	Ethylenediaminetetraacetic acid
EMBL	European Molecular Biology Laboratory
ESRF	European Synchrotron Radiation Facility
FBS	Fetal bovine serum
FDA	Food and drug administration
FKBP12	FK506 Binding Protein 12
for	forward
FRB	FKBP Rapamycin Binding domain
FT	Fourier transform
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
Gln/Q	Glutamine
Glu/E	Glutamic acid
Gly/G	Glycine
GST	Glutathione transferase
HA	Human influenza hemagglutinin
HSPC	Hematopoietic stem and progenitor cell
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
His/H	Histidine
HMGB1	High mobility group protein 1
Hsmar1	*Human mariner 1* (transposon)
Hsmar1	Human mariner 1 (transposase)
Hsmar1-Ra	Reconstituted ancestral transposase Human mariner 1
hsSB	high solubility Sleeping Beauty (transposase)
HTH	Helix-Turn-Helix
I(0)	Intensity at zero angle
IDT	Integrated DNA Technologies
Ile/I	Isoleucine
iPSC	Induced pluripotent stem cell
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IR	Inverted repeat
kbp	Kilobase pair
K_d	Dissociation constant
kDa	Kilodalton
keV	Kiloelectron volt
Km	Kanamycin
LB	Lysogeny Broth
LE	Left end
Leu/L	Leucine
Li	Left inner
LIF	Leukemia inhibitory factor
LLG	Log-likelihood gain
Lo	Left outer
Lys/K	Lysine
MAD	Multi-wavelength anomalous diffraction

MALDI-TOF MS	Matrix-assisted laser desorption/ionization time of flight mass spectrometry
MDACC	MD Anderson Cancer Center
mESC	Mouse embryonic stem cell
Met/M	Methionine
MIR	Multiple isomorphous replacement
Mw	Molecular weight
MR	Molecular replacement
M.Sc.	Master of Science
N-terminus	Amino-terminus
NC	Not characterized
NCBI	The National Center for Biotechnology Information
NEB	New England Biolabs
neo	Neomycin
NHEJ	Non-homologous end joining
NLS	Nuclear localization signal
NMR	Nuclear magnetic resonance
nt	Nucleotide
NT	Not transfected
NTS	Non-transferred strand
nvAMD	Neovascular age-related macular degeneration
OD_{600}	Optical density at 600 nm
OPI	Overproduction inhibition
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PFA	Paraformaldehyde
PB	*piggyBac*
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PEC	Paired-end complex
PEDF	Pigment epithelium-derived factor
PEG	Polyethylene glycol
PFA	Paraformaldehyde
PFV	Prototype foamy virus
PGBD5	*piggyBac transposable element derived 5*
Phe/F	Phenylalanine
PI	Propidium iodide
pI	Isoelectric point
PMSF	Phenylmethanesulphonylfluoride
Pro/P	Proline
q	Momentum transfer
RAG	Recombination activating genes
RE	Right end
rev	reverse
RF	Restriction free (cloning)
RFP	Red fluorescent protein
Rg	Radius of gyration

Ri	Right inner
RNA	Ribonucleic acid
RNase	Ribonuclease
Ro	Right outer
RPE	Retinal pigment epithelial
rpm	Round per minute
SAXS	Small-angle X-ray scattering
SB	*Sleeping Beauty* (transposon)
SB	Sleeping Beauty (transposase)
SB	Sleeping Beauty transposon system (where indicated)
SDS	Sodium dodecyl sulphate
SEC	Size exclusion chromatography
SEC1	Single-end complex 1
SEC2	Single-end complex 2
SEJ	Single-end joining
SeMet	Selenomethionine
Ser/S	Serine
SETMAR	Suppressor of variegation 3-9, Enhancer-of-zeste, Trithorax domain and Mariner transposase
SOB	Super Optimal Broth
TAE	Tris-acetate-EDTA
TALEN	Transcription activator-like effector nuclease
TAT	Transactivator of transcription
TBE	Tris-borate-EDTA
Tc1	*Transposon Caenorhabditis 1*
Tc3	*Transposon Caenorhabditis 3*
TCA	Trichloracetic acid
TCC	Target capture complex
TCEP	Tris(2-carboxyethyl)phosphine
tDNA	target DNA
TE	Transposable element
TE buffer	Tris-EDTA buffer
TEMED	Tetramethylethylenediamine
Tet	Tetracycline
Thr/T	Threonine
Tris	Tris(hydroxymethyl)aminomethane
Trp/W	Tryptophan
TRX	Thioredoxin
TS	Transferred strand
TSD	Target site duplication
Tyr/Y	Tyrosine
UV	Ultraviolet
Val/V	Valine

1 GENERAL INTRODUCTION

'They thought I was crazy, absolutely mad.'

B. McClintock, 1944

This chapter introduces transposable elements, with focus on DNA transposons, and their impacts on evolution and genetic engineering. This introduction highlights the motivation of the research carried out in my PhD project.

Genomes of all organisms have the capacity to undergo DNA rearrangements, which can result in modification of their genetic pool and modulation of their gene expression. Among the factors responsible for these recombination events, transposable elements (TEs, also known as transposons) have been found in all genomes sequenced to date and generally occupy a large fraction of them (Craig, 2002). Notably, they make up about half of the genome content in humans and represent even up to the ~85% of sequences in some plant genomes (Lander *et al.*, 2001; Lee *et al.*, 2014).

1.1 "Jumping" genes: an overview

TEs are defined as discrete DNA segments with the distinctive ability to autonomously "jump" from one location to another in a host genome without any requirement for sequence homology.

They were discovered in the late 1940s (even before the molecular structure of the DNA was revealed) by Barbara McClintock (McClintock, 1950), who was studying mutable genetic loci in maize. The mutations she observed were the consequences of movement of DNA transposons known as *Ac* (for "*Activator*") and *Ds* (for "*Dissociation*"), *Ac* being the autonomous element regulating *Ds*. McClintock's work was revolutionary since it suggested that an organism's genome is not a stationary entity, but rather is subject to dynamic alterations (or mobile). This principle was met with heavy criticism from the scientific community until McClintock was finally awarded the Nobel Prize in 1983 in recognition of her contributions to the field of genetics and epigenetics.

Nowadays, scientists are well aware of the extraordinary number of transposons existing in nature. As new transposons were discovered and the related molecular mechanisms were elucidated, several classification systems were proposed (Jurka *et al.*, 2005; Piegu *et al.*, 2015; Wicker *et al.*, 2007). One of the most common and still valid classifications is based on the TEs' general mechanism of transfer, according to which they are categorized into two classes (Figure 1-1):

- Class I elements, generally referred to as retroelements, retrotransposons or RNA transposons, which move by reverse transcription of an RNA intermediate;
- Class II elements, or DNA transposons, which move only via DNA-mediated mechanisms (Craig, 2002).

Retrotransposons follow a so-called copy-and-paste mechanism for their mobilization, while most DNA transposons move through a cut-and-paste pathway.

This work focuses on TEs belonging to Class II, or DNA transposons. DNA transposons are further subclassified into different families based on the general transposon architecture and on the nature of their associated proteins.

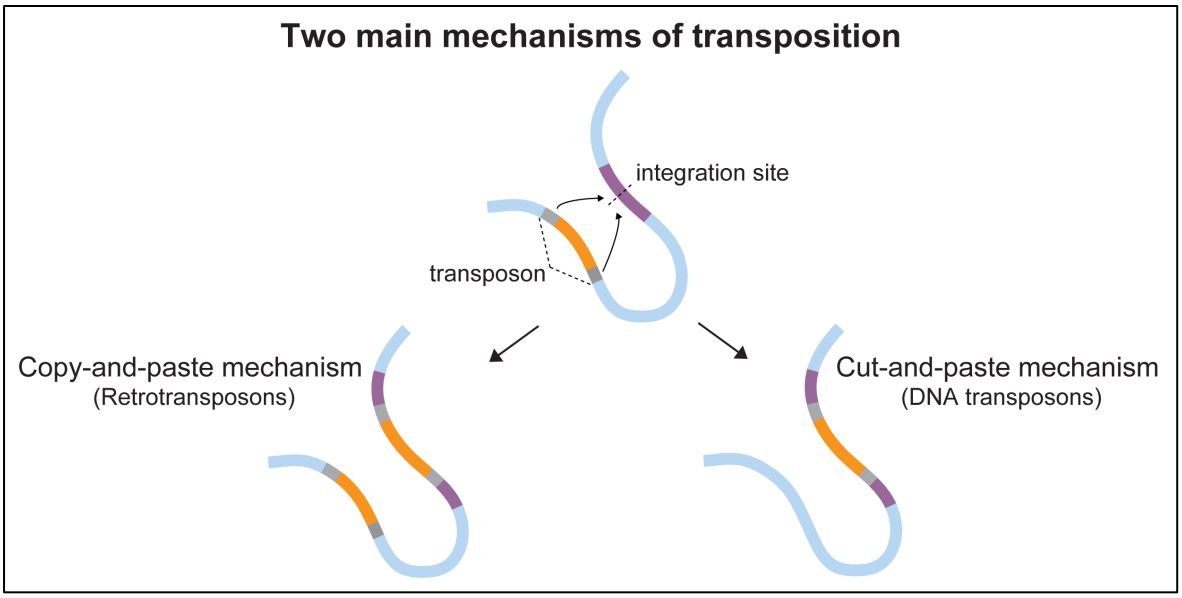


Figure 1-1: Transposons move within genomes mainly by two distinct mechanisms: 1. the copy-and-paste pathway, typical of RNA transposons, which involves single-stranded RNA intermediates and 2. the cut-and-paste mechanism, followed by most DNA transposons, which exclusively relies on DNA intermediates. If the integration site is located in a gene (represented in purple), insertion of the transposon (in orange, flanked by transposon's termini in grey) results in interruption of the gene's DNA sequence.

All complete or "autonomous" DNA transposons consist of at least one open reading frame (ORF) encoding the transposase protein, an enzyme needed for autonomous mobility (transposition), flanked by two inverted repeat (IR) sequences at the transposon termini (Figure 1-2). These components are an indispensable requirement for transposition: the transposase binds to specific DNA sequences within the IRs and catalyses the complete excision of the transposon from the donor locus and its integration to a different target locus in the host genome. Of note, "non-autonomous" transposons also exist, which accumulated inactivating mutations in their transposase-encoding ORF and thus "borrow" the transposition machinery from their active mates for their mobility (as the *Ds* element found by McClintock).

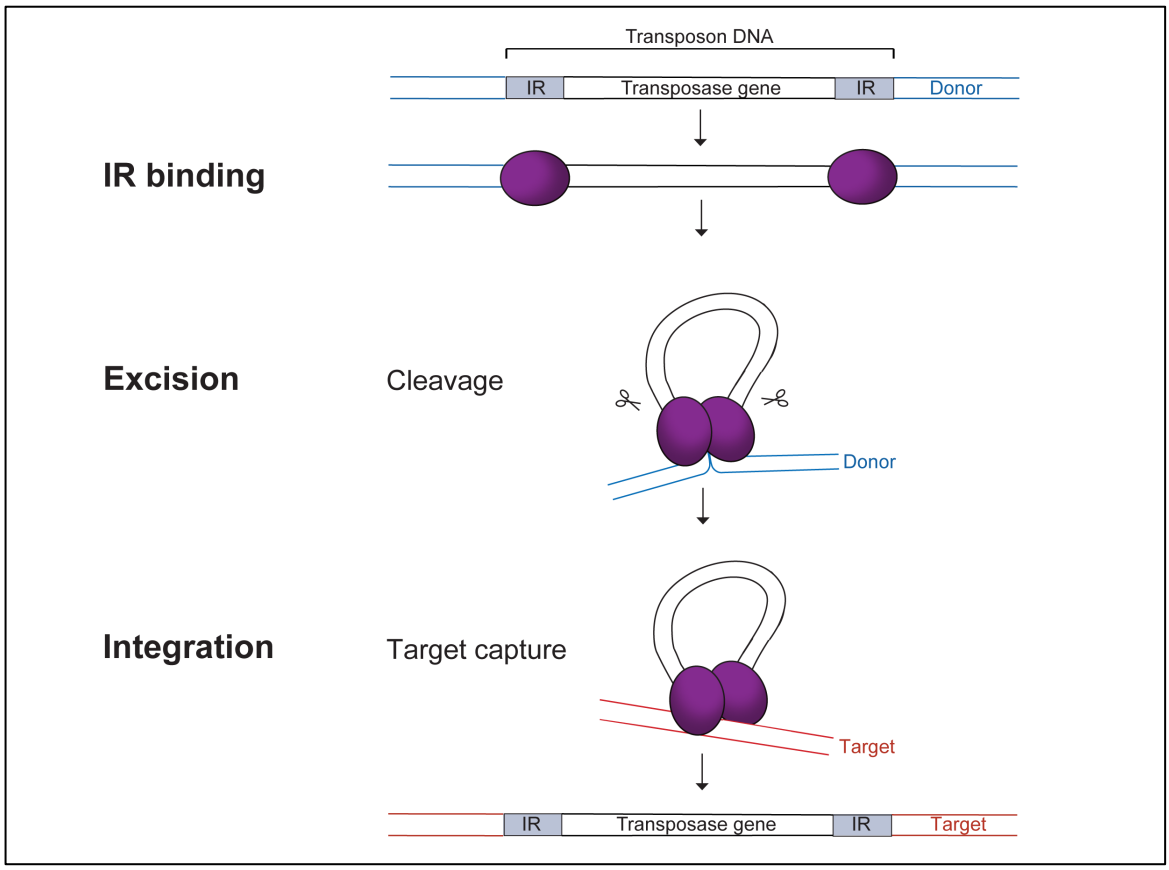


Figure 1-2: Schematic representation of the general cut-and-paste DNA transposition mechanism. Individual transposase molecules are represented as purple ovals. They initially bind to specific inverted repeat (IR) sequences at the transposon ends, they then perform cleavage at these sites to liberate the transposon from its donor locus (excision) and finally promote transposon insertion into a new locus (integration) upon capture of a target DNA. Donor DNA is shown in blue; target DNA is in red.

Transposases represent the most abundant genes known in nature (Aziz *et al.*, 2010) and contain a high variety of protein sequences. This generates an extraordinary diversity of transposases' structures and transposition mechanisms, which are mostly unexplored (Dyda *et al.*, 2012; Hickman *et al.*, 2016). Therefore, biochemical and structural investigation of transposons and their associated transposases represents a thrilling and fast-evolving research field.

1.2 Biological impacts of DNA transposons

Due to their ability to carry and modify genes (Figure 1-1), TEs are prominent evolutionary forces that act in genome remodelling and antibiotic resistance spreading in prokaryotes (Alekshun *et al.*, 2007) and contribute to the emergence of new biological

functions (S. Huang *et al.*, 2016; Sinzelle *et al.*, 2009) and gene-regulatory networks (Chuong *et al.*, 2017; Imbeault *et al.*, 2017) in eukaryotes.

Eukaryotic DNA transposons are active in plants and lower-order animals (C. R. Huang *et al.*, 2012). Interestingly, they have established a special relationship with some unicellular eukaryotic organisms, ciliates, where transposon-derived proteins have been domesticated to carry out programmed genome rearrangements required for sexual reproduction (Baudry *et al.*, 2009; Cheng *et al.*, 2010; Nowacki *et al.*, 2009). In higher organisms (except bats) activity of DNA transposons is often severely down-regulated or disabled through mutations. In mammals, for example, the vast majority of DNA transposons are non-autonomous and in some genomes, such as the human genome, almost none of the ca. 300,000 DNA transposons have a functional transposase gene (Feschotte *et al.*, 2007). Thus, higher-order organisms genomes are riddled by remnants of once active elements and, for this reason, transposons have been for a long time dismissed as "junk" DNA. However, an overwhelming body of evidence highlights that DNA transposons can introduce genomic changes that have beneficial or detrimental functional consequences in their hosts, including in higher-order animals (Chenais *et al.*, 2012; Feschotte *et al.*, 2007).

The beneficial roles of DNA transposons are obvious for exaptation, meaning the repurposing of transposon-derived DNA sequences or proteins to execute a different task other than transposition. In fact DNA transposons have been proven to provide *cis*-regulatory elements and thus contribute to the evolution of regulatory networks (Imbeault *et al.*, 2017). For instance, in the peppered moth a polymorphic *carbonaria* TE insertion within an intron of the *cortex* gene enhances its expression levels, underlying the adaptive cryptic coloration of moths emerged during the industrial revolution (Van't Hof *et al.*, 2016). Regarding transposon-derived proteins, the RAG1 and RAG2 (recombination activating genes) proteins of the vertebrate V(D)J recombination system have arisen from an ancient RAG transposon (S. Huang *et al.*, 2016). Moreover, the once functional transposase protein of the *Human mariner 1* (*Hsmar1*) transposon makes up the C-terminal domain of the SETMAR [Suppressor of variegation 3-9, Enhancer-of-zeste, Trithorax (SET) domain and Mariner transposase] fusion protein, a DNA repair factor involved in the non-homologous end joining (NHEJ) pathway in humans (Goodwin *et al.*, 2010). Notably, transposases have been also an important source of DNA binding and chromatin-associated domains (Feschotte, 2008).

Apart from beneficial outcomes, the DNA breaking activity and mobility of DNA transposons present also a danger for their hosts as these can scramble and dramatically disrupt the host genomes. For these reasons, the activity and retention of DNA transposons are finely modulated by a combination of intrinsic (host- or self-mediated) and extrinsic (environmental, ecological) factors over evolutionary time (Feschotte *et al.*, 2007) to minimize detrimental effects. Perhaps, it is more correct to say that an arms race is established between host organisms and transposons, which can eventually escape existing repression mechanisms. Recently, a *piggyBac transposable element derived 5* (*PGBD5*) gene was found in humans, which encodes an active transposase expressed in the majority of childhood solid tumours, including lethal rhabdoid tumours (Henssen *et al.*, 2015). This discovery provides the first example of a DNA transposon with autonomous (even if conditional) activity in the human genome and which can exhibit oncogenic consequences.

1.3 Technological impacts of DNA transposons

DNA transposons are natural gene delivery vehicles capable of efficient genomic insertion. They exhibit important features that make them particularly attractive as genetic vectors. Due to specificity of donor sites, DNA transposons are mobilized very precisely and only single transposon copies are cut-and-pasted at a time. This ensures that single copies of desired transgenes are inserted with absolute fidelity. Differently to viruses that often have highly immunogenic protein coats, transposons are mostly DNA-based vectors and hence avoid immune defence mechanisms evolved by the cells against viral proteins. DNA transposons are generally not restricted by the size of the transgene to be delivered (Zayed *et al.*, 2004), in contrast to viral vectors in which DNA packaging capacity is fairly limiting.

Due to these intrinsic properties, the mobile nature of DNA transposons has been successfully exploited to provide efficient insertional mutagens or sensors in functional genomics (de la Rosa *et al.*, 2017; Dupuy *et al.*, 2006; Ruf *et al.*, 2011) as well as artificial gene carriers in transgenesis (Miskey *et al.*, 2005; Skipper *et al.*, 2013) and even in gene therapy applications (Kebriaei *et al.*, 2016). One recent and exciting application is the high resolution mapping of nucleosome positions using transposon integrations into accessible chromatin followed by sequencing [called ATAC-seq method; (Buenrostro *et al.*, 2013; Chen *et al.*, 2016)].

In their applications, the mobility of DNA transposons is experimentally controlled by separating the two functional components of the transposon: the IR sequences are placed to flank a gene of interest, and the transposase protein is conditionally supplied in *trans* to drive the transposition reaction. In principle, any sequence between the IRs can be mobilized by the transposase and can be stably integrated into the genome in a regulated and highly efficient manner (Ammar *et al.*, 2012).

The most widely applied transposons are *piggyBac* (*PB*), and *Tc1/mariners*. *PB* was discovered as an active mobile element in the cabbage moth *Trichoplusia ni* (Cary *et al.*, 1989) and is used as a versatile tool in various genetic modification experiments (Woodard *et al.*, 2015). Recent reports have shown that the *PB* transposon vector can be mobilized by the human transposase-derived PGBD5 protein (Henssen *et al.*, 2015; Henssen *et al.*, 2017; Ivics, 2016), raising major concerns about the use of *PB* in human applications. *Tc1/mariner* elements are preferred systems for the development of transposon-based genetic tools, because of their unique evolutionary, mechanistic and structural characteristics. These TEs have frequently crossed species boundaries in nature and provide very powerful gene delivery vehicles in a broad range of hosts (Grabundzija *et al.*, 2010; Plasterk *et al.*, 1999). Their transposition is not-host restricted, probably because it requires minimal transposon- and host-encoded machineries, if any. *Tc1/mariners* are generally not limited by the size of the transgene to be delivered (Zayed *et al.*, 2004) and have no preference for integration into genes or regulatory units (Sultana *et al.*, 2017). In addition, no human proteins have been identified that could remobilize genomically integrated *Tc1/mariner* transposon vectors to date.

Among *Tc1/mariners*, the *Sleeping Beauty* (*SB*) transposon (Ivics *et al.*, 1997) is exceptionally efficient in inserting into vertebrate genomes (Dupuy *et al.*, 2006), making it a prime genetic tool with applications spanning from transgenesis (Ammar *et al.*, 2012), forward mutagenesis screening (de la Rosa *et al.*, 2017; Dupuy *et al.*, 2006; Ruf *et al.*, 2011) to even cancer immunotherapy trials in humans (Kebriaei *et al.*, 2016; Singh *et al.*, 2015). However, the mechanistic principles underlying *SB*'s transposition and distinctive performance have remained elusive, greatly hampering the rational design and further improvement of this transposon as genetic tool in research and medicine.

1.4 Aims of this study

Nowadays, mobility of DNA transposons can be investigated using integrated genetic, biochemical, and structural approaches. Owing to the technological impact of *Tc1/mariners*, novel insights into the mechanism and regulation of these transposons will support the development of attractive tools for basic research and human therapy. Therefore, the overarching goal of my PhD project is to investigate the molecular mechanisms of *Tc1/mariner* transposition to design new improved transposon systems for genome engineering.

In the first part of my work (chapter 2), I have focused on studying the structural and biochemical principles underlying transposition of two prominent *Tc1/mariners*: *Hsmar1*, the best biochemically characterized member (section 2.2), and *SB*, a prime genetic tool applied both in research and medicine (sections 2.3 and 2.4). In particular, I analysed the biologically relevant conformations of their associated transposase proteins and nucleoprotein complexes by a combination of low- and high-resolution structural biology methods, classical biochemistry, fluorescence-based, and biophysical techniques. Moreover, by establishing novel biochemical assays, I dissected the discrete DNA intermediates involved in *SB* transposition. Finally, these comparative insights gained *in vitro* have been explored in the cellular context using cell biology techniques, shedding first light on the molecular determinants of *SB*'s hyperactivity.

In the second section of my work (chapter 3), I carried out the most ambitious part of my project, which is the rational design of novel SB transposon systems. To this aim, using available structural data as a resource, I generated an improved transposase variant by protein design and characterized its biophysical and functional properties (section 3.2), taking advantage of specific assays established in the previous section (chapter 2). Based on the use of this rationally designed protein variant, a new genetic engineering method was developed, utilizing cell biology methodologies such as *in vivo* transposition assays and flow cytometry (section 3.3).

In brief, by learning more about the working mechanism of *Tc1/mariners,* and of *SB* in particular, we may open up new avenues for the development of tailor-made genome engineering tools for research and clinical use.

2 MECHANISMS OF *TC1/MARINER* TRANSPOSITION

The part of my PhD work presented in this chapter focuses on elucidating the molecular mechanisms of *Tc1/mariner* transposition. Section 2.1 introduces *Tc1/mariner* transposons and their movement, and highlights the motivation of the research conducted in this part of my thesis. This is followed by a section presenting the structural investigation of *Hsmar1* transposition (2.2) and two sections devoted to the *SB* transposon, focusing on the biochemistry of its transposition (section 2.3) and its self-regulation at the molecular level (section 2.4). In the last section (2.5), a model for *Hsmar1* and *SB* transposition, based on the obtained results, is proposed and described. Finally, future perspectives in the investigation of the studied transposons (and of other family members) at the structural and biochemical level are discussed.

Note: A number of the figures in this chapter are based on and modified from the following manuscript that is in preparation for submission. I created all figures that can be found in the manuscript and in this thesis.

- Querques *et al.*, (2018); "Distinct assembly pathways of Tc1/mariner transpososomes dictate different self-regulatory mechanisms."

2.1 Introduction – *Tc1/mariner* transposons

Tc1/mariners constitute an extraordinary group of DNA transposons characterized by an exceptionally widespread distribution in the tree of life, a distinctive mechanism of transposition and a tremendous potential for biotechnological applications (Plasterk, 1996). This introduction to *Tc1/mariner* transposons describes the main distinctive features that make these elements unique among DNA transposons as well as fascinating biological systems for biochemical and structural investigation.

2.1.1 Evolution, life cycle, inactivation and molecular resurrection

All identified eukaryotic DNA transpososons have been classified in ~18 superfamilies, according to similarities in the sequences of their associated transposases and transposon ends (Jurka *et al.*, 2005; Piegu *et al.*, 2015; Wicker *et al.*, 2007). Among these, the *Tc1/mariner* superfamily is ubiquitously present in eukaryotes, and is named after two of the first identified family members: the *Tc1* (*Transposon Caenorhabditis 1*) transposable element from *Caenorhabditis elegans* (Emmons *et al.*, 1983) and the *mariner* element *Mos1* from *Drosophila mauritiana* (Jacobson *et al.*, 1986). The *Tc1/mariner* superfamily appears to constitute a monophyletic group of transposons and can be further grouped in the *Tc1*, *mariner* and *pogo* clades (Figure 2-1).

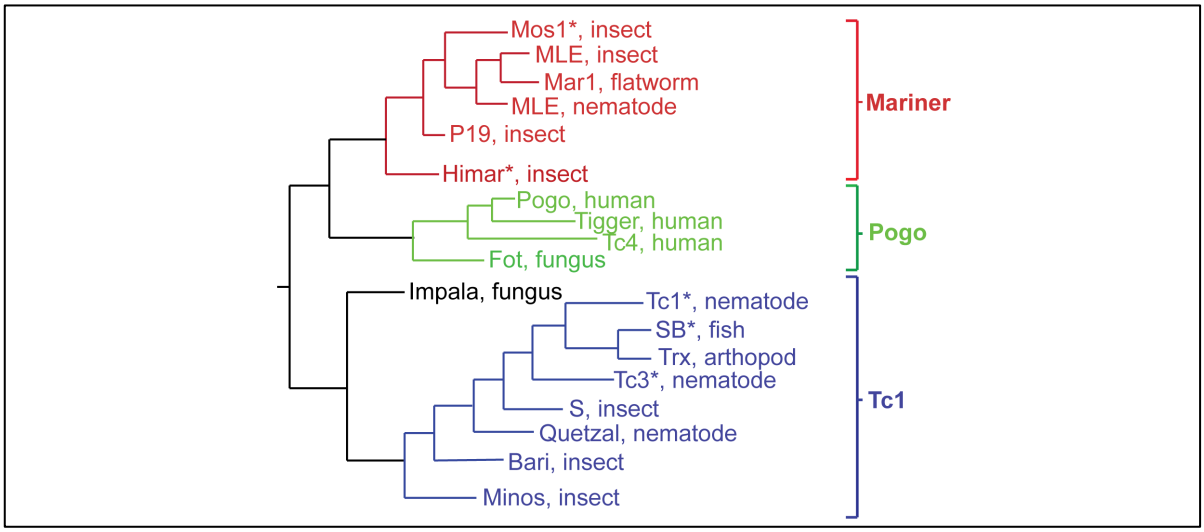


Figure 2-1: Phylogeny of the *Tc1/mariner* superfamily (Plasterk *et al.*, 1999). The *mariner* (red), *pogo* (green) and *Tc1* (blue) transposon families are monophyletic. The phylogenetic tree is based on similarities between Tc1/mariner transposases and represents only a subset of all known elements. Members of the family mentioned in this study are marked with asterisks. Next to each element, the host organism is indicated.

Tc1/mariners are the most widespread DNA transposons existing in nature. They are represented in all eukaryotes, spanning from rotifers, fungi, plants, and ciliates to fish and mammals, including humans (Plasterk, 1996). Remarkably, phylogenetic relationships of *Tc1/mariner* elements are often incongruent with those of their host species (Ivics *et al.*, 1996; Robertson *et al.*, 1993) (Figure 2-1). For example, close relatives of *mariner* transposons in humans are found in insects and worms (Lampe *et al.*, 2001). This suggests that *Tc1/mariners* have frequently crossed species boundaries by a so-called "horizontal transfer" process during evolution (Lohe *et al.*, 1995).

How this occurs exactly is still subject of investigation. Since transposons are not infectious *per se*, it is likely that additional genetic vehicles, like viruses and parasites, are needed as vectors for their transfer between organisms (Houck *et al.*, 1991; Kidwell, 1992). Once in the new host, *Tc1/mariners* can move in the newly invaded genome without any host-specific restrictions because the transposon-encoded transposase protein is the only factor strictly required for their transposition (Clayes Bouuaert *et al.*, 2010; Lampe *et al.*, 1996; Vos *et al.*, 1996).

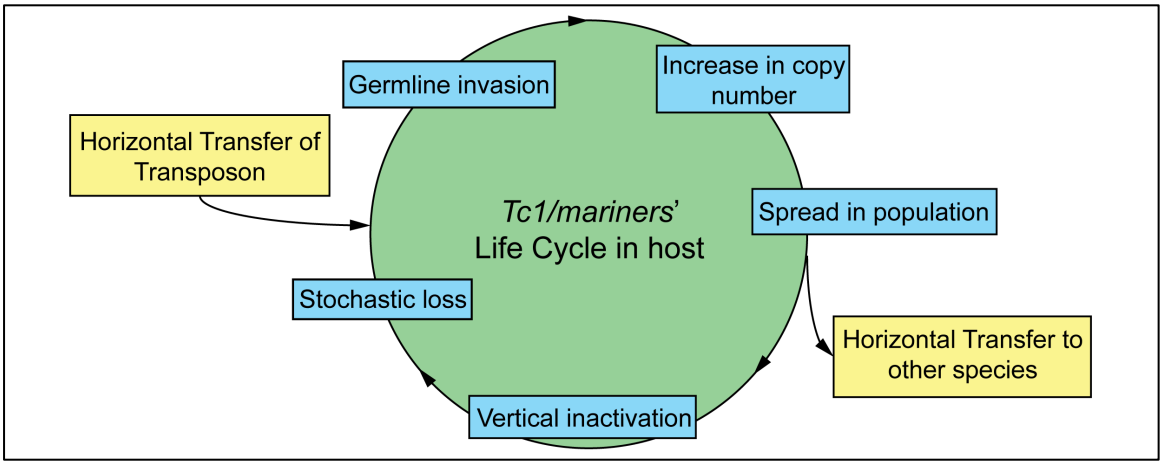


Figure 2-2: Proposed life cycle of *Tc1/mariner* elements. The figure has been adapted from (Miskey *et al.*, 2005).

Once transmitted to a new host, *Tc1/mariners* go through a peculiar life cycle: first, they must introduce themselves in the germline genome; then, they experience an exponential amplification of their activity and copy number; finally, more and more transposon copies gradually undergo vertical inactivation, ultimately leading to complete suppression of their activity. To circumvent complete extinction, *Tc1/mariners* must therefore transfer

themselves to new species repeatedly, even to rather distantly related ones (Miskey *et al.*, 2005) (Figure 2-2).

As a consequence of repeated transposon invasions and inactivation by mutation, many genomes are scattered with relics of once active transposons. Out of the hundreds identified sequences, only ten naturally occurring *Tc1/mariner* elements have proven to be active [i.e. *Tc1* and *Tc3* from *C. elegans* (Emmons *et al.*, 1983), *Minos* from *Drosophila hydei* (Franz *et al.*, 1991), *Mos1* from *D. mauritiana* (Hartl, 2001; Medhora *et al.*, 1991), *Famar1* from *Forficula auricularia* (Barry *et al.*, 2004), *Osmar5* from *Oryza sativa* (G. Yang *et al.*, 2006), *Fot1* and *Impala* from the fungus *Fusarium oxysporum* (Daboussi *et al.*, 1992; Langin *et al.*, 1995), ISY100 isolated in bacteria (Feng *et al.*, 2007), and *Mboumar-9* from the ant *Messor bouvieri* (Munoz-Lopez *et al.*, 2008)]. Notably, although almost 3% of the human genome is built of *Tc1/mariner* transposons, all the estimated 20,000 elements are inactive due to multiple mutations within the coding regions of their transposase genes (Munoz-Lopez *et al.*, 2010).

In some cases transposases enzymes have escaped mutational inactivation through domestication for a different task in the host cell. The Hsmar1 transposase provides a remarkable case of domestication; it has been incorporated in the SETMAR fusion protein upon insertion of the Hsmar1-encoding ORF downstream of the SET gene. SETMAR acts as a DNA repair factor in the NHEJ pathway, where the transposase-derived moiety of the protein provides DNA binding and processing activities (Goodwin *et al.*, 2010; Liu *et al.*, 2007).

In the past decades, molecular reconstruction of artificial active *Tc1/mariner* transposons has been carried out by combination of comparative sequence analysis and mutagenesis screens. By eliminating inactivating mutations in their transposases and transposon end sequences, an increasing number of *Tc1/mariners* have been brought back to life. Among these, prominent examples are the *SB* transposon from salmonid-type fish (Ivics *et al.*, 1997) and the above-mentioned *Hsmar1* element from *H. sapiens* (Miskey *et al.*, 2007), which constitute the main focus of this study. The awakening of these extinct elements opened the door to the exploration of *Tc1/mariner* transposons in basic research and biotechnological applications.

2.1.2 Transposon ends and transposases

Tc1/mariners are about 1.3 to 2.4 kbp long elements. Their boundaries are marked by the two transposon ends that are conventionally referred to as Left End (LE) and Right End (RE). Specific inverted repeats (IRs) are present at each end and constitute binding sites for the transposase. They are 30-40 bp long perfect IRs in the *mariners* and long bipartite IRs, each containing two 30-35 bp direct repeats (DRs) separated by DNA segments of variable sequence, in some *Tc1-like* elements (Plasterk *et al.*, 1999). The functional role of the varying and complex arrangement of DRs within the family is still uncharacterized.

Between the two ends, *Tc1/mariner* transposons harbor a single ORF encoding the transposase. Although Tc1/mariner transposases share as little as 15% amino acid sequence identity (Robertson *et al.*, 1995), they feature a conserved structural organization: the N-terminal DNA binding domain (DBD), which specifically recognizes the IR sequences at the transposon ends, comprises two Helix-Turn-Helix motifs (HTH) and is connected by a flexible linker to the C-terminal RNase H-type catalytic domain (CAT) (Figure 2-3).

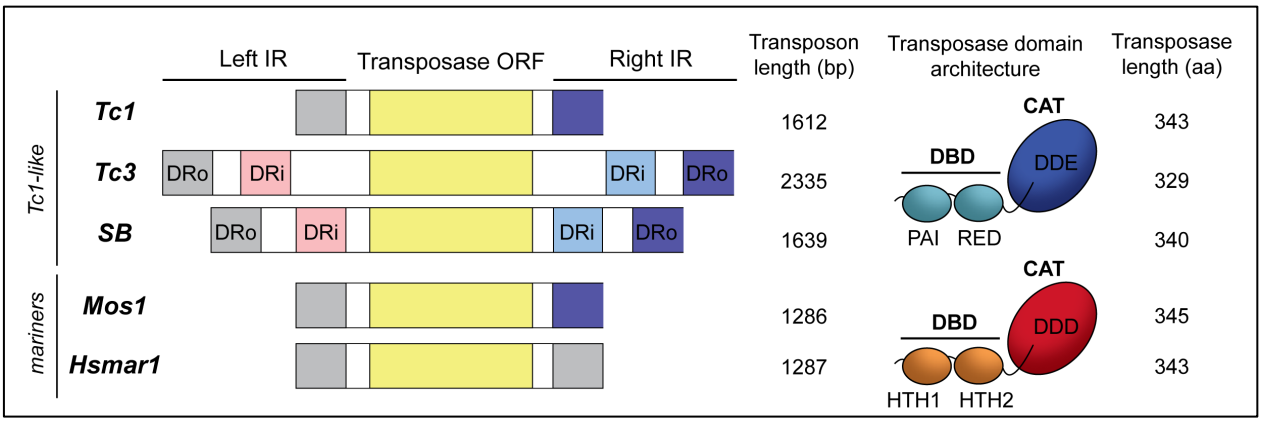


Figure 2-3: General architecture of *Tc1/mariner* transposon ends and transposases. The transposase open reading frame (ORF) is represented as a yellow rectangle and is flanked by inverted repeats (IRs). *Mariners* and the *Tc1* element contain only single left (grey square) and right (purple square) repeats. In the *Hsmar1* transposon, the two IRs have identical sequences (both shown as grey squares). Most *Tc1-like* elements feature bipartite IRs (so called IR-DR), where each IR contains outer (DRo) and inner (DRi) direct repeats. Representative Tc1/mariner transposases (right panel; SB for Tc1-like in blue and Hsmar1 for mariners in red) are depicted with their bipartite DNA binding domains (DBD; small ovals) and their catalytic domains (CAT; big ovals). The same colour code (blue for SB; red for Hsmar1) is used to represent these transposases in the rest of this thesis. DDE and DDD indicate the different catalytic triads. HTH: Helix-Turn-Helix motif.

In Tc1-like transposases, the DBD is similar to the paired domain of some transcription factors and thus its subdomains are also called PAI and RED (referring to HTH1 and HTH2 respectively) (Ivics *et al.*, 1996; Vos *et al.*, 1994). The CAT of all Tc1/mariner

transposases assembles into a typical RNase H-like fold and contains a conserved triad of acidic residues - three aspartates (DDD) in the *mariners* and two aspartates and one glutamate (DDE) in *Tc1-like* elements - that are responsible for catalysing DNA rearrangements in a two-metal-ion-dependent manner (Montano *et al.*, 2011; Plasterk *et al.*, 1999; W. Yang *et al.*, 2006). Tc1/mariner transposases feature also a nuclear localization signal (NLS), located at varying positions in their amino acid sequence. The presence of the NLS indicates that these transposons can take advantage of receptor-mediated transport of their host cells for the nuclear uptake of their transposases (Ivics *et al.*, 1996).

2.1.3 Overall transposition pathway

Transposition of *Tc1/mariner* elements occurs by a cut-and-paste process. Specific binding of the transposase to IR sequences triggers coordinated cleavage of both transposon ends, excising the transposon from its original locus. Both strands of each transposon end are directly cut by the transposase that is therefore able to perform four distinct cleavage reactions for a complete transposon excision (Dawson *et al.*, 2003). Mg^{+2}-mediated hydrolysis generates 3'-hydroxyls at the transposon ends, which then attack one strand of the target DNA at staggered positions for transposon integration. Integration always occurs precisely at the 5' of an invariant TA dinucleotide in the target DNA (Craig, 2002) (Figure 2-4) and generates a DNA structure in which the transposon is flanked by short single-stranded gaps. These gaps are then repaired by host machineries, creating TA target site duplications (TSDs) at the TE flanks characteristic of transposon integration (Craig, 2002) (Figure 2-4).

A fundamental requirement for *Tc1/mariner* transposition is the coordination of the DNA cleavage reactions at each transposon end. This allows to avoid aberrant single-end cleavage events that are non-productive for the transposon and potentially genotoxic for the host. To ensure this, transposition occurs in the context of a specific protein-DNA machinery, called transpososome, which contains the two transposon ends and an even number of transposase subunits. This specific stoichiometry ensures that the two ends are held together (synapsed) by at least two protein protomers and processed in a coordinated fashion (Hickman *et al.*, 2016). As transposition proceeds, the transpososome undergoes multiple conformational rearrangements to ensure correct and unidirectional progression of

the reactions (Claeys Bouuaert *et al.*, 2010; Claeys Bouuaert, Lipkow, *et al.*, 2013; Claeys Bouuaert *et al.*, 2014; Cuypers *et al.*, 2013; Richardson *et al.*, 2009).

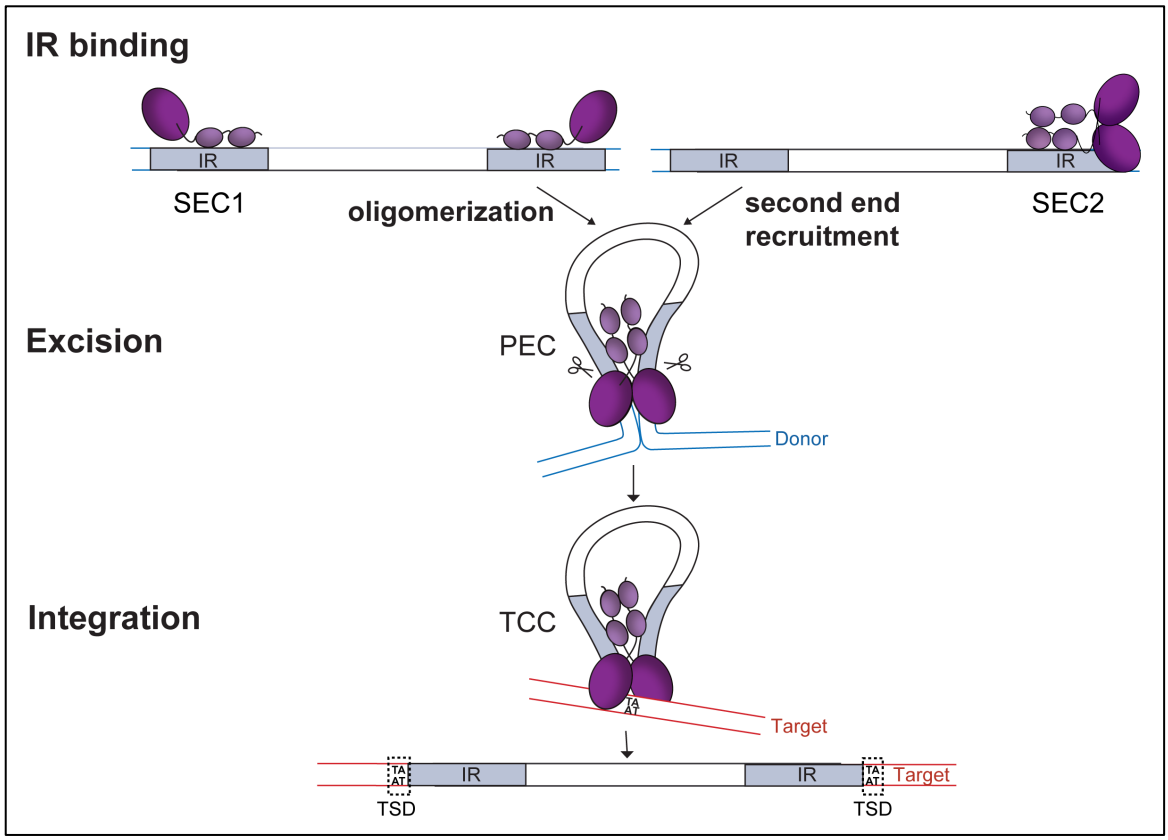


Figure 2-4 Transposition mechanism proposed for *Tc1/mariners*. Specific binding of the transposon-encoded transposase to inverted repeats (IRs) at each transposon end is followed by paired-end complex (PEC) assembly, coordinated cleavage of both transposon termini (excision) to liberate the transposon from the donor locus, subsequent target recognition in the target capture complex (TCC), and finally transposon integration. IR binding can be carried out by monomers (forming a single-end complex 1; SEC1) or by dimers (forming a single-end complex 2; SEC2). Transposon integration and repair of the target locus lead to characteristic TA target site duplications (TSDs). The different domains of the transposases are shown as ovals. Large oval: catalytic domain, lighter small ovals: DNA binding domain.

In the following paragraphs, the main steps and intermediates of the *Tc1/mariner* transposition reaction are described in the light of the biochemical and structural information available.

2.1.4 Transposon end recognition and synapsis

The first step of transposition requires sequence-specific recognition of the two transposon ends by the transposase, leading to transposon end synapsis in the so-called paired-end

complex (PEC) (Figure 2-4). Early events of this assembly process remain poorly understood at the molecular level.

Tc1/mariner transposases have been reported to exist in various oligomeric states prior to transposon end binding, from monomers to tetramers (Brillet *et al.*, 2007). This diversity might reflect differences in their transposition mechanisms, altered oligomerization equilibrium, or simply suboptimal conditions *in vitro*. Depending on the oligomeric state of the transposase, two main models for transposon end recognition and subsequent PEC assembly have been formulated (Claeys Bouuaert, Lipkow*, et al.*, 2013) (Figure 2-4):

1. individual transposase subunits initially bind to separate transposon ends (forming a so-called single-end complex 1, SEC1) and subsequentially undergo oligomerization for transposon end pairing in the PEC;

or

2. a preformed transposase dimer (or higher oligomer) binds a single transposon end (single-end complex 2, SEC2) and then recruits the second DNA end to form the PEC.

The first model was proposed based on biochemical evidences obtained with *Tc1* (Vos *et al.*, 1994) and *Tc3* (Fischer *et al.*, 1999). However, these two transposons differ in the number of DRs in their IRs [two for *Tc1* and four for *Tc3* (Figure 2-3)] and how the varying architecture of DRs influences transposon ends binding and synapsis is still unknown. Notably, the crystal structure of the Tc3 DBD in complex with DRs visualizes four subunits held together by two distinct dimer interfaces (Watkins *et al.*, 2004). As the physiological relevance of these interfaces remains to be verified, it is still unclear if the Tc3 transposase forms dimers or tetramers in the PEC.

The second model, implying transposon end recruitment by preformed dimers, has been attributed to Mos1, Mboumar-9 and Hsmar1 transposases, mainly based on their fairly stable dimeric state *in vitro* (Auge-Gouillou *et al.*, 2005; Claeys Bouuaert *et al.*, 2017; Trubitsyna *et al.*, 2014). Based on structural studies performed by small angle X-ray and neutron scattering, the Mos1 transposase is an elongated dimer that binds a single transposon end, forming a SEC2 (Cuypers *et al.*, 2013). However, the exact architecture of such single-end complexes has not been determined yet for any Tc1/mariner transposase.

To sum up, due to differences in the transposon end architecture within *Tc1/mariners* and to limited structural and biochemical information on their transposases, it is still unknown which assembly model correctly applies to each element. Nevertheless, both proposed

models agree with the generally recognized principle of DNA transposition, according to which at least two transposase subunits are required to completely process the two transposon ends. This rule is also reflected in the architecture of the Mos1 transpososome, the only transpososome structure revealed for eukaryotic transposases to date (Dornan *et al.*, 2015; Morris *et al.*, 2016; Richardson *et al.*, 2009) (Figure 2-5).

All available structures of the Mos1 transpososome (Dornan *et al.*, 2015; Morris *et al.*, 2016; Richardson *et al.*, 2009) contain a transposase homodimer bound to the two transposon ends in a *trans*-arrangement: each IR sequence is recognized by the N-terminal DBD of one transposase subunit and by the C-terminal CAT of the other subunit. Thus, both subunits act together to carry out cleavage. Each transposase protomer establishes protein-protein contacts in two distinct regions: the so-called clamp loop in the CAT of one protomer interacts with the linker region of the other protomer, and the two HTH1 motifs interact with each other (Figure 2-5). Notably, Tc1-like transposases feature a different linker region (with a conserved KKPL/F motif) than mariners (with a WVPHEL motif) and a distinct essential glycine-rich strip at the base of the clamp loop, suggesting that their PEC architecture might be significantly different from the mariners (Voigt *et al.*, 2016).

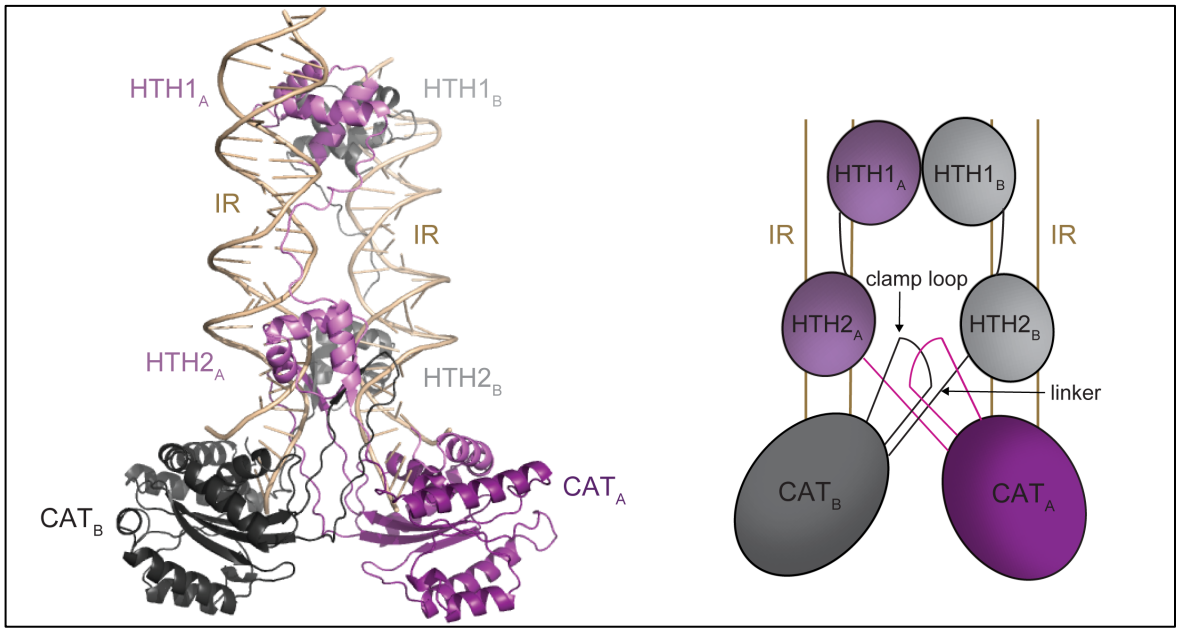


Figure 2-5: Molecular architecture of the Mos1 paired-end complex (PEC). Left: cartoon representation of the Mos1 PEC crystal structure [PDB ID: 3HOT; (Richardson *et al.*, 2009)]. Subunit A is coloured purple and subunit B grey. Each inverted repeat (IR) is represented as light brown dsDNA. Right: schematic representation of the structure and the intermolecular interactions. Large oval: catalytic domain (CAT), lighter small oval: Helix-Turn-Helix (HTH) subdomains. Based on (Richardson *et al.*, 2009).

Perhaps the most puzzling question about the early stages of *Tc1/mariner* transposition regards how the transposase enzyme restrains itself from processing single transposon ends before synapsis. A model of synapsis by oligomerization implies that catalysis is not possible in monomers (SEC1), and thus occurs only when both transposon ends are brought together in an oligomeric PEC. However, in the alternative pathway of "synapsis by recruitment" transposases feature a dimeric state even at the first step of single end recognition (SEC2), and thus inhibitory mechanisms must prevent catalysis on single ends (Claeys Bouuaert *et al.*, 2014; D. Liu *et al.*, 2014) (Figure 2-4). Nevertheless, the structural principles underlying this regulation are still missing.

2.1.5 Transposon excision

Although different DD(E/D) tranposases adopt different strategies to liberate their transposon ends from a donor site, excision of DNA transposons always begins with the nucleophilic attack of an activated water molecule at or close to the transposon end, and follows a two-metal-ion mechanism of phosphodiester bond hydrolysis (W. Yang *et al.*, 2006) (Figure 2-6).

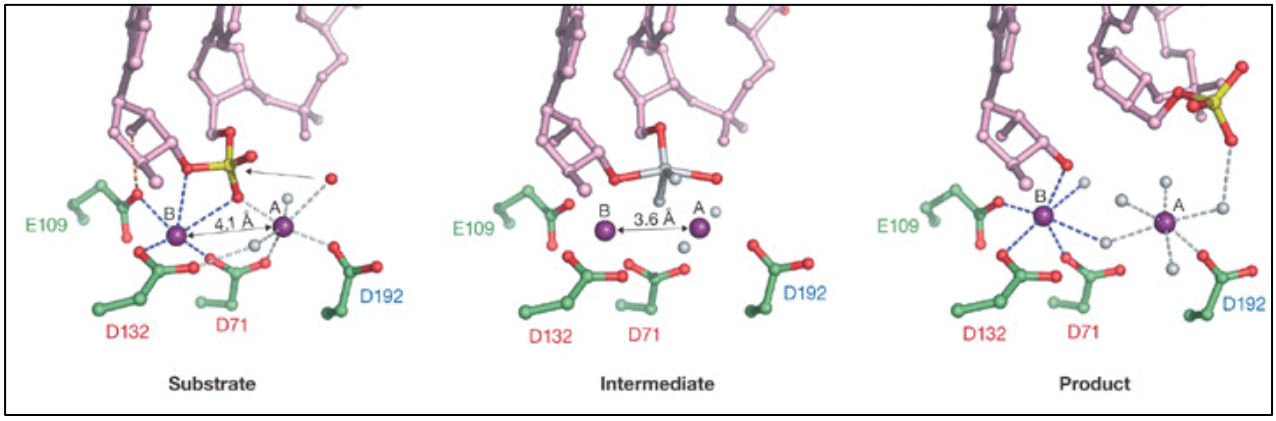


Figure 2-6: Two-metal-ion mechanism of phosphodiester bond hydrolysis as catalysed by RNase H1 enzymes. The nucleic acid is shown in pink ball-and-stick representation with the scissile phosphate shown in red and yellow. The attacking nucleophile is shown as a small red sphere and the two Mg^{2+} ions as purple spheres. Active-site carboxylates are shown in green ball-and-stick representation. Active site residues labelled in red and blue (D71, D132, D192) are conserved in DDD/E transposases. The additional glutamate (E109, green) is specific to RNase H1, RNA H2 and viral reverse transcriptases. The representations are based on the structures of *Bacillus halodurans* RNase H1 enzyme in complex with RNA/DNA hybrids (PDB ID: 1ZBL, 2G8H and 2G8V). The figure has been reprinted from (Nowotny, 2009). Copyright © 2009 European Molecular Biology Organization, with permission from John Wiley and Sons (License Number 4275220590271).

Protein residues in the enzyme active site coordinate two divalent metal ions (Mg^{2+}, physiologically) and orient a nucleophile water molecule for S_N2 in-line attack on the phosphorus atom of the scissile phosphate of the bound DNA substrate. The reaction proceeds through a pentacovalent intermediate and leads to inversion of the phosphorus configuration. The phosphodiester bond is broken such that the products are two free DNA ends, one with a 3'-OH group and one with a 5'-phosphate. Two conserved catalytic aspartates (the first two D-s of the DD(E/D) motif) directly coordinate the catalytic metal ions (labelled red in Figure 2-6). The third catalytic residue also coordinates one of the metal ions (labelled blue in Figure 2-6), but it is relatively tolerant to mutation (Nowotny, 2009).

Apart sharing chemistry with other transposases, Tc1/mariner transposases use a distinctive mechanism to generate double strand breaks (DSBs) at each transposon end during excision (Figure 2-7). Most cut-and-paste transposases cut just one DNA strand at each transposon end by hydrolysis, and the generated free 3′-OH then attacks the phosphodiester bond on the second strand at the same transposon end, forming a DNA hairpin intermediate. The hairpin may then need to be resolved by hydrolytic cleavage to free the 3′-OH required for the further integration reactions (Craig, 2002).

No DNA hairpin intermediate was detected in *Tc1/mariner* transposition (Dawson *et al.*, 2003). Thus, excision of *Tc1/mariners* includes additional reactions performed by the transposase, which must cleave two DNA strands of opposite polarity, conventionally referred to as transferred (TS) and non-transferred (NTS) strands at each transposon end. The TS is the strand whose free 3'-OH attacks and joins the target DNA in the subsequent integration reactions, whereas the NTS is not directly integrated in the target DNA. Throughout excision, the NTS is nicked first at 2-3 nucleotides recessed position within the transposon end and then the TS is cleaved exactly at the transposon boundary (Claeys Bouuaert *et al.*, 2014; Clayes Bouuaert *et al.*, 2010). The DNA sequence that flanks the transposon, in particular the invariable TA dinucleotide, which derives from the target site of the last integration, seems to be crucial for excision (Claeys Bouuaert *et al.*, 2010; Cui *et al.*, 2002; Dornan *et al.*, 2015).

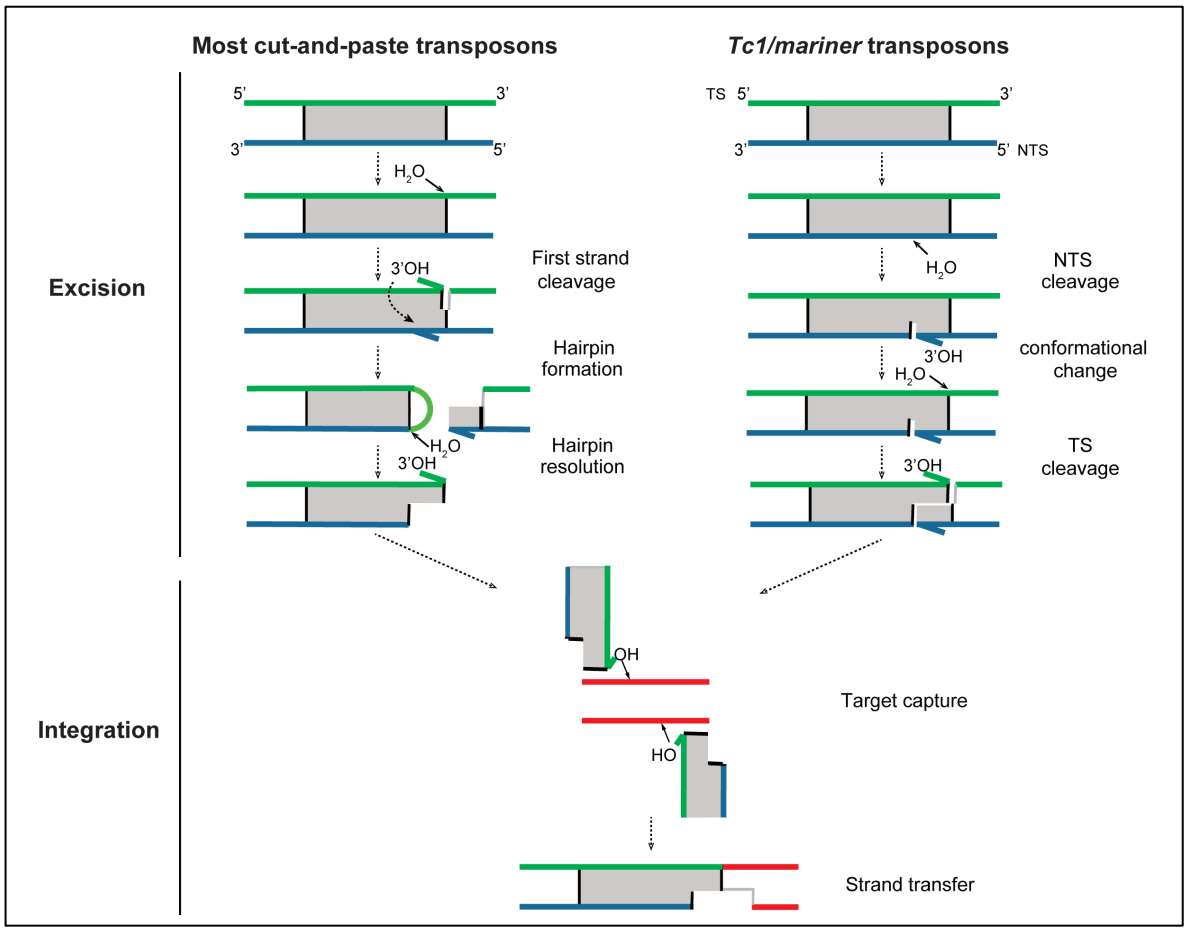


Figure 2-7: Different mechanisms for DNA cleavage by cut-and-paste transposases. A single inverted repeat is represented as a grey rectangle and the green and blue lines indicate the two DNA strands. DNA strands of the target DNA are depicted as red lines. In most DNA transposons double strand breaks at each transposon end are generated via a hairpin intermediate. In the figure a 3'-OH belonging to the transposon end is depicted to attack the other strand for hairpin formation. In an alternative pathway, first cleavage generates a 3′-OH at the flanking DNA that is then used as a nucleophile, producing a hairpin on the donor flank. Members of the *Tc1/mariner* superfamily, perform excision via two hydrolysis reactions at each transposon end, first on the non-transferred strand (NTS) and successively on the transferred strand (TS). The transposition reaction proceeds with target capture and strand transfer, as for all DNA transposons.

Due to the difficulty in isolating and analysing transposition intermediates *in vitro* and to the consequent lack of informative structures, the mechanism of transposon excision in *Tc1/mariners* is still poorly understood. Open questions concern the exact steps involved in hydrolysis and the structural assemblies in which reactions occur. Recent reports showed that each active site of the Hsmar1 transposase dimer is responsible for two hydrolysis and one transesterification reactions at the same transposon end (Claeys Bouuaert *et al.*, 2017). However, it is not clear whether some of the hydrolysis reactions (i.e. initial NTS cleavage) could already occur prior to transposon end-pairing, in SEC1 or SEC2 complexes, or in yet unknown transpososome conformations.

Once *Tc1/mariner* transposons are excised, they leave DSBs behind, which can be repaired using a homologous chromosome or a sister chromatid as a template. The repair machinery can therefore recreate a new copy of the original transposon in its old site (Craig, 2002). This results in exponential amplification of *Tc1/mariners* after invasion. Alternatively, the donor site can be repaired via the NHEJ pathway, potentially resulting in major genome modifications or deletions at the donor site. Consistent with NHEJ involvement, the Ku70/Ku80 protein complex has been implicated to interact with the *SB* transposon, helping to ensure its efficient transposition (Izsvak *et al.*, 2004; Yant *et al.*, 2003).

2.1.6 Transposon integration

Once transposon ends have been released from their donor site, the 3'-OH group at each end serves as the nucleophile for the following transesterification reactions, allowing transposon integration into a new site (Figure 2-7). *Tc1/mariners* always integrate at a TA dinucleotide (Figure 2-4). Recent structural data on the Mos1 post-integration complex (Morris *et al.*, 2016) and the catalytic domain of the SB transposase (Morris *et al.*, 2016; Voigt *et al.*, 2016) shed first light on the structural principles of this target specificity. However, no structure of a eukaryotic transpososome contacting target DNA before integration (the so-called target capture complex, TCC) has been yet determined (Figure 2-4). Based on available structures of the related prototype foamy virus (PFV) integrase enzyme, catalysis of strand transfer likely uses the same active sites as for the initial step of DNA cleavage at the transposon ends. Only, now the 3'-OH groups at each excised TS end serve as activated nucleophiles attacking the two opposite strands of a target DNA at 2 nt staggered positions (Hare *et al.*, 2010; Hare *et al.*, 2012; Maertens *et al.*, 2010).

Concomitantly with formation of new phosphodiester bonds upon strand transfer, nicks are generated in the target DNA strands that must be repaired by the host machinery to fill in the 4-5 nt long gaps introduced by integration (Figure 2-7). Repair of such gaps generates TA TSDs on each side of the integrated transposon (Figure 2-4). TA TSDs constitute a specific signature of *Tc1/mariner* transposition and can be monitored to trace their movements in genomes.

2.1.7 Positive regulation of transposition

Faithful progression of the transposition pathway is important to ensure successful DNA processing and avoid deleterious events such as DSBs formation at an isolated end, imprecise excision, or failure to reintegrate. Such aberrant events may also result in undesired genome rearrangements, and hence present a danger to the host as well. Thus, careful regulation of an efficient and precise cut-and-paste mechanism can be viewed as "structural checkpoints" in the positive regulation of the transposition pathway.

A first prominent checkpoint regards the early events of transposon end recognition and PEC assembly, which ensure coordinated and complete excision. Small angle scattering data indicated that the Mos1 transposase binds one transposon end as an elongated dimer, assembling a SEC2 complex (Cuypers *et al.*, 2013). Yet, how the second transposon end is recruited remains unclear. Based on biochemical studies, the binding of the first transposon end is believed to affect the affinity for the second end (Claeys Bouuaert and Chalmers, 2013; Claeys Bouuaert, Lipkow*, et al.*, 2013; D. Liu *et al.*, 2014), but the mechanistic details are unknown. Moreover, how catalysis on single ends is mechanistically prevented is one of the most intriguing questions concerning transposition autoregulation.

Another structural checkpoint acts concomitantly to the synapsis of the two transposon ends. Major conformational changes are proposed to occur in the Mos1 transposase (180 degrees rotation of one transposase subunit relatively to the other one) for transition from an elongated SEC2 to a compact PEC (Cuypers *et al.*, 2013). Available structures of the Mos1 PEC revealed that the *trans*-architecture of the complex ensures coordination of transposon end recognition and processing (Dornan *et al.*, 2015; Morris *et al.*, 2016; Richardson *et al.*, 2009) (Figure 2-5). However, these structures fail to explain how the two ends are actively paired and correctly positioned in the complex for initial cleavage.

Yet, another key checkpoint concerns the conformational changes occurring within the transpososome between first and second strand cleavage (Figure 2-7). These rearrangments are required to remove the first cut strand from an active site of the transposase and correctly position the second strand for cleavage and successive nucleophilic attack on the target DNA. This structural switch ensures that both DNA strands have been processed before integration can proceed.

2.1.8 Negative regulation of transposition

Unregulated transposition would result in an exponential amplification of the number of transposon copies in the host genome, inevitably leading to genomic meltdown. To circumvent this, several mechanisms have evolved to down-regulate transposition. In addition to host defence machineries that preserve genome integrity [i.e. DNA methylation, heterochromatin formation and the piRNA pathway (Slotkin *et al.*, 2007)], evidences exist that transposons negatively regulate their own dissemination via taming the activity of their transposases (Claeys Bouuaert, Lipkow*, et al.*, 2013).

First experimental evidence for autoregulation of *Tc1/mariners*, and eukaryotic transposons in general, has been provided by observation of a phenomenon called overproduction inhibition (OPI) for the *mariner* element *Mos1* (Lohe *et al.*, 1996). OPI manifests in a reduction in the frequency of excision above a certain transposase concentration and has been described to affect several *Tc1/mariner* transposons both *in vitro* and *in vivo* (Bire *et al.*, 2013; Clark *et al.*, 2009; Clayes Bouuaert *et al.*, 2010; Lampe *et al.*, 1998; Lohe *et al.*, 1996).

Transpososome assembly has been implicated to play a critical role in establishing OPI. Claeys Bouuaert and colleagues have proposed that mariner transpososomes assemble by initial binding of transposase dimers to one transposon end and subsequent recruitment of the second naked end (Figure 2-8). As the transposase concentration rises, occupation of both ends by transposase dimers impairs productive synapsis, and consequently causes OPI (Claeys Bouuaert, Lipkow*, et al.*, 2013). Whether this so-called “assembly site occlusion” (ASO) model of autoregulation is conserved in other members of the *Tc1/mariner* superfamily has remained unclear. To note, *SB* exhibits sustained transposition even at high transposase concentrations (Kolacsek *et al.*, 2014), suggesting that this transposon regulates its activity differently.

Alternatively, OPI might arise from an alternative oligomeric state of the transposase, which might form inactive oligomeric species at high transposase concentrations (G. Carpentier *et al.*, 2011) (Figure 2-8). In agreement, structural evidences for the Mos1 transposase revealed that the dimer interface involved in PEC formation is different from the one observed in the crystal structure of the catalytic domain in the absence of DNA

(Richardson *et al.*, 2006). This suggests that an alternative dimer interface in the catalytic domain could be involved in down-regulation of transposition by OPI.

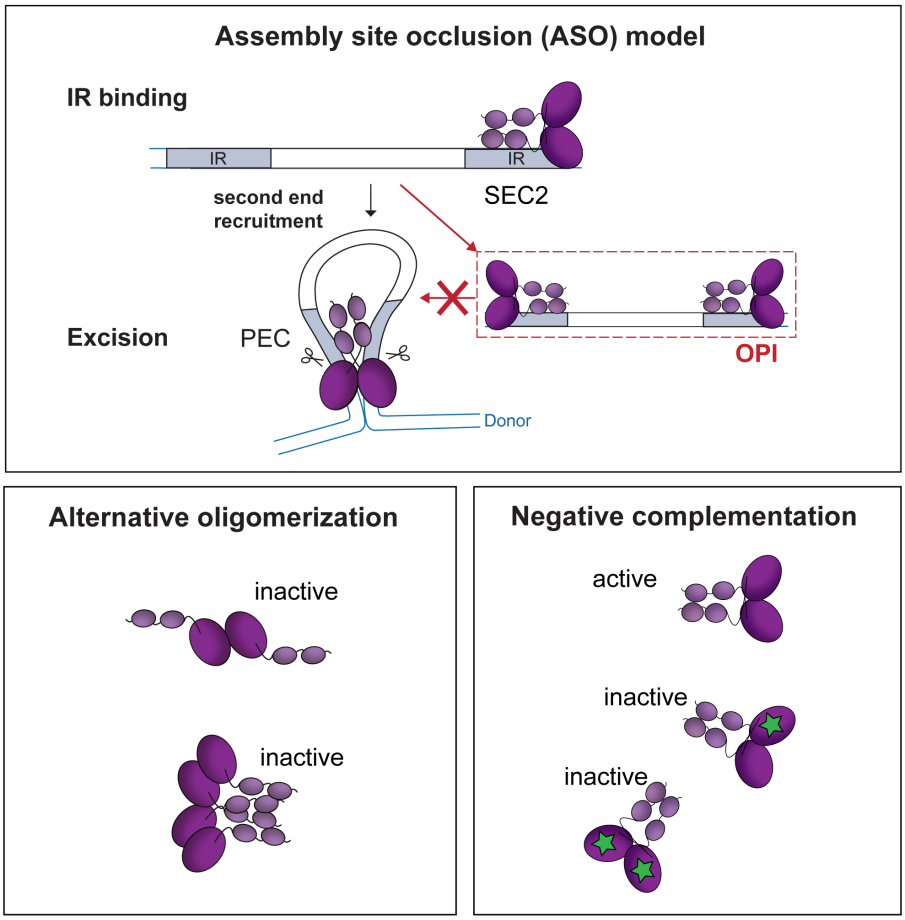


Figure 2-8: Autoinhibitory mechanisms proposed for *Tc1/mariners*. Assembly site occlusion (ASO) model: consistent with the "assembly by second end recruitment" pathway, occupation of both transposon ends by transposase dimers impairs productive synapsis, causing overproduction inhibition (OPI). Alternative oligomerization: several transposase conformations can co-exist in solution and some are inactive in transposition. Negative complementation: interaction of wild type transposase subunits with mutated ones (marked with green stars) result in formation of inactive oligomers. The figure follows the same schematic code as Figure 2-4. SEC2: single-end complex 2; PEC: paired-end complex.

Another mechanism for transposition down-regulation might act at the stage when transposon copies start to accumulate mutations in their transposase sequence. Combination of wild type and mutated subunits of the transposases to create oligomers would result in impaired activity of the hetero-oligomers, leading to transposition inhibition (G. Carpentier *et al.*, 2011) (Figure 2-8). Such dominant negative complementation is consistent with the intrinsic oligomerization properties of the transposases; however, no experimental evidences support a physiological role of this mechanism in transposon autoinhibition to date.

2.1.9 *Human mariner 1*: a representative *mariner* transposon

Human mariner 1 (*Hsmar1*) was the first *mariner* transposon discovered in the human genome and was probably active about 50 million years ago (Cordaux *et al.*, 2006; Liu *et al.*, 2007; Miskey *et al.*, 2007; Robertson *et al.*, 1997). Now, all 200 copies of *Hsmar1* present in the human genome bear mutations that inactivate the transposase. Using an *in silico* phylogenetic approach, the ancestral sequence of the transposase gene was successfully reconstructed, providing the so-called "reconstituted ancestral transposase" Hsmar1-Ra (later simply referred to as the Hsmar1 transposase, Hsmar1). The reconstructed transposase gene is capable of catalysing precise cut-and-paste transposition of its specific IRs in its natural (human) host and in the zebrafish embryo, making up the first functional vertebrate *mariner* transposon system (Miskey *et al.*, 2007).

The complete *Hsmar1* transposon consensus sequence is 1287 bp long. The transposon architecture is one of the simplest within the *Tc1/mariner* family. The transposon is bordered by two 30 bp perfect, identical IRs, each containing a single transposase binding site, and contains a single ORF encoding for a 343 amino acids long transposase (Robertson *et al.*, 1997) (Figure 2-3). The reconstructed Hsmar1 transposase shares 94% sequence identity with the transposase version found in the SETMAR fusion protein and is 37% identical to the related Mos1 transposase.

Most of the mechanistic information available for *mariner* transposons has been derived from *Hsmar1* and *Mos1*. *In vitro* analysis revealed that *Hsmar1* transposition follows a typical cut-and-paste pathway. The transposase generates DSBs at each transposon ends by nicking first the NTS 3 nt inside the transposon end and later the TS, always exactly at the transposon termini (Claeys Bouuaert *et al.*, 2010). Based on recent biochemical data, double strand cleavage and integration are performed by a single transposase subunit within the stable homodimer at each transposon end (Claeys Bouuaert *et al.*, 2017). However, detailed characterization of the Hsmar1 transpososome assembly has remained challenging.

Initial biochemical analysis of the Mos1 transpososome has also lead to several alternative models for the arrangement of subunits in the active complex. On the other hand, structural analysis of the Mos1 PEC revealed a *trans* dimeric assembly and recent evidences support a dimeric state for the Hsmar1 transposase as well [reported by (Claeys Bouuaert *et al.*,

2017)]. Consistently, in my M.Sc. work in the Barabas Laboratory, we found that the Hsmar1 transposase is a stable dimer both before and after binding to transposon ends (Querques *et al.*, manuscript in preparation). Hsmar1 dimerization is mainly driven by its DBD, is not required for transposon end binding, but it is strictly needed for catalysis of DNA cleavage *in vitro* and productive transposition in mammalian cells. Transposase dimerization thus controls both transposon end synapsis and cleavage activity in *Hsmar1*, which is consistent with the finding that catalysis in *Hsmar1* transposition is strictly dependent on prior synapsis of the ends (Claeys Bouuaert *et al.*, 2011).

Regarding the kinetics of the transposition reaction, the Hsmar1 transposase binds rapidly to one transposon end; however, binding to the second end is much slower. This delay in synapsis might result from (i) an allosteric conformational change induced by single end binding, which prevents the second end to bind at comparable rates (Claeys Bouuaert and Chalmers, 2013; D. Liu *et al.*, 2014) and/or (ii) electrostatic repulsion between two charged dsDNA molecules. Following synapsis, catalysis also occurs at different rates on the two DNA strands, where the NTS is nicked rapidly, whereas TS cleavage is much slower. This suggests that a second conformational change is required between these events. Target capture and integration are also relatively slow and are believed to be tightly coupled to the previous excision reactions (Claeys Bouuaert *et al.*, 2010).

The only structural information available for full length Tc1/mariner transposases derives from the structures of the Mos1 transpososome in its pre-TS cleavage, post-excision and post-integration conformations (Dornan *et al.*, 2015; Morris *et al.*, 2016; Richardson *et al.*, 2009). Based on sequence identity, we can hypothesize that Hsmar1 forms a similar PEC to Mos1. However, due to lack of structural information on other transposition intermediates, how the DNA breaking and ligation reactions involved in *Hsmar1* transposition occur at the atomic level has remained unclear.

2.1.10 The *Sleeping Beauty* transposon and its new artificial life

Sleeping Beauty (*SB*) is an artificial transposon that was resurrected from a 20 million-years long "evolutionary sleep" by determining a consensus sequence for 12 mutated Tc1-family transposons found in eight salmonid fish species. Its reconstruction in 1997 was the first demonstration that ancient transposons can be brought back to life and that they can actively move in vertebrates. The 340 amino acid transposase was reconstituted, in parallel

to the circa 230 bp IRs, to give rise to the so-called "SB10" transposase, which was proved to be active in fish, mouse and human cell lines (Ivics *et al.*, 1997).

The overall domain organization of the SB transposase is similar to other *Tc1/mariners,* although sequence identity is quite low (15% with the Mos1 transposase) (Figure 2-3). As other Tc1-like transposases, its DBD is formed by both PAI and RED subdomains (Ivics *et al.*, 1996; Vos *et al.*, 1994). The CAT of SB, in conformity with Tc1-like transposases, features a conserved DDE catalytic triad, a glycine rich-strip region and a conserved KKPL motif, differently to mariner transposases. Structures are available of the separate domains of the SB transposase, namely its PAI [NMR structure, (C. E. Carpentier *et al.*, 2014)], RED [NMR structure, (Konnova *et al.*, 2017)] and CAT domain [crystal structure, (Voigt *et al.*, 2016)], which visualize the fold and arrangement of specific structural regions, but fail to explain how transposition is coordinated in the context of the SB transpososome.

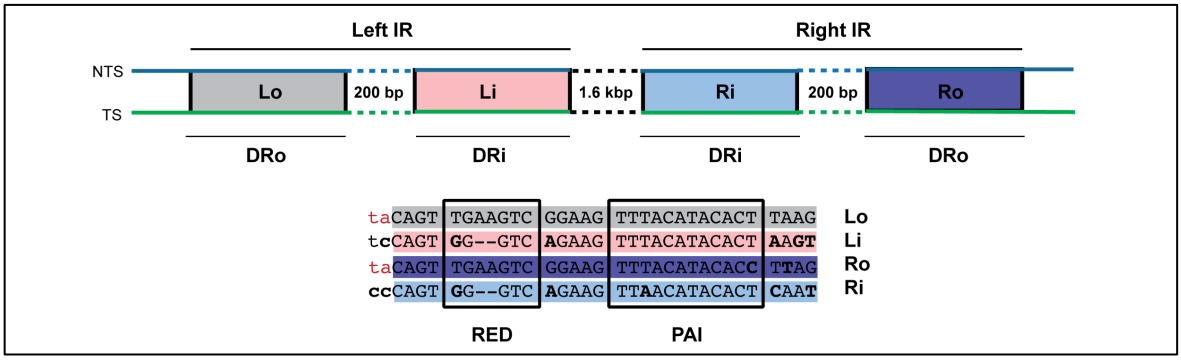


Figure 2-9: *SB* transposon ends architecture and sequences. Top: schematic representation of *SB* transposon ends. Transposase binding sites (outer directed repeats-DRo and inner directed repeats-DRi) are shown as rectangles and are also named as left outer (Lo), left inner (Li), right inner (Ri) and right outer (Ro) repeats. Blue line: non-transferred strand (NTS); Green line: transferred strand (TS). Bottom: the binding sites are not identical even though they share a consensus sequence. DNA regions contacted by the PAI and RED subdomains of the SB transposase are indicated (Izsvak *et al.*, 2002). Transposon end DNA sequence is shown in capital letters (bold letters indicate nucleotide differences between binding sites). Flanking DNA is indicated in lower case letters. TA dinucleotides (red) in the flanks of the outer repeats (Lo and Ro) represent *Tc1/mariners*-specific target site duplications.

SB features a complex transposon end architecture. In contrast to *mariner* transposons, considerably longer IRs (ca. 230 bp) are required for transposition (Figure 2-3 and Figure 2-9). Each IR contain two ~32 bp transposase binding sites, one at the terminus of the transposon (outer DR) and an interior site located ~165 bp from the outer DRs (inner DR). Accordingly, the binding sites are also named left outer (Lo), left inner (Li), right inner

(Ri) and right outer (Ro) DRs. Outer and inner DRs have similar, but not identical sequences (Figure 2-9) and are both necessary and not interchangeable for efficient transposition *in vivo*. Moreover, the inner DRs are thought to be bound more tightly than the outer DRs by the transposase, suggesting distinct roles for these binding sites in transposition (Cui *et al.*, 2002).

SB moves through a cut-and-paste mechanism; however, the exact products and DNA intermediates of *SB* transposition as well as the minimal protein requirements of these reactions have not been yet characterized. The only information concerning the nature of the SB cleavage reactions is in fact derived from sequencing of repaired transposon excision sites in mammalian cells (Luo *et al.*, 1998): footprint analysis in several cell lines indicate that breakages at the transposon ends are staggered and consistent with cuts on the NTS 2-3 bp within the transposon. Transposition assays in HeLa cells indicate that the TA dinucleotide at the transposon flanks, originating from the insertion sites, also affect transposon excision (Cui *et al.*, 2002).

As other *Tc1/mariners*, *SB* inserts preferentially into TA sites, however, in addition to the TA, a palindromic AT-repeat consensus sequence in AT-rich tracts was found to constitute preferred target locations (Vigdal *et al.*, 2002). Once remobilized from a chromosomal site, *SB* tends to re-integrate into *cis*-linked sites in the vicinity of the donor locus, according to a phenomenon known as “local hopping”. Nevertheless, *SB* integration can be considered fairly random at the genome level and does not show any preference for integration into transcription units or regulatory regions of genes (Moldt *et al.*, 2011). This constitutes a distinct advantage for its applications in random mutagenesis screens as well as transgenesis.

To optimize *SB*’s “jumping” activity, both the original SB10 tranposase and IRs have undergone a number of improvements (Ammar *et al.*, 2012; Cui *et al.*, 2002; Mates *et al.*, 2009; Voigt *et al.*, 2016; Yant *et al.*, 2004; Zayed *et al.*, 2004). The most substantial increase of transposase activity (100-fold) has been achieved by close-to-random mutagenesis of the SB10 to originate the SB100X version (Mates *et al.*, 2009). This hyperactive derivative, indicated simply as SB in the rest of the thesis, has been used, unless otherwise indicated, in all experiments reported in sections 2.3 and 2.4.

Since these ground-breaking studies, *SB* has proven to be exceptionally efficient in inserting into vertebrate genomes (up to 95% efficiency) (Dupuy *et al.*, 2006) and has become a prime genetic tool applied in transgenesis of higher organisms (Ammar *et al.*, 2012), cancer gene discovery (Dupuy *et al.*, 2006), and forward mutagenesis screening (de la Rosa *et al.*, 2017; Dupuy *et al.*, 2006; Ruf *et al.*, 2011). It also provided the first non-viral gene delivery vector used in ongoing clinical trials to *ex vivo* modify human T cells for cancer immunotherapy (Kebriaei *et al.*, 2016; Singh *et al.*, 2015) (as extensively reported in section 3.1). Other *Tc1/mariner* transposons such as *Mos1* and *Hsmar1* are far less efficient. However, the mechanistic principles of *SB*'s extraordinary success as genetic tool have remained uncharacterized.

Two decades after *SB*'s resurrection, the crystal structure of the SB catalytic domain provided proof-of-concept for the rational design of new hyperactive transposase variants (Voigt *et al.*, 2016). However, the current lack of knowledge on the transposition mechanism and regulation of *SB*, and of *Tc1/mariners* in general, greatly impairs further development of this transposon for basic research and clinical applications. This part of my work aims to fill up some of these gaps.

2.2 Results – Structural principles of *Hsmar1* transposition

In order to elucidate the mechanism of *Tc1/mariner* transposition, I used the *Hsmar1* transposon as a model for the superfamily. The *Hsmar1* transposon is the best-characterized member with respect to the biochemistry of its transposition pathway. Previous reports (Claeys Bouuaert *et al.*, 2010; Claeys Bouuaert and Chalmers, 2013; Claeys Bouuaert *et al.*, 2017) also indicated that the Hsmar1 transposase has particularly high solubility and stability appropriate for *in vitro* work, as compared to other related proteins.

Despite the available biochemical information, it is still poorly understood how a single Hsmar1 transposase enzyme orchestrates all transposition steps, and which structural assemblies and conformations are involved. In particular, open question concern:
1. Structural basis for inhibition of single end cleavage. The Hsmar1 transposase forms stable dimers in solution and in the PEC. So, what prevents the preformed Hsmar1 dimer to cleave single transposon ends before PEC assembly?
2. Mechanism of transposon excision. How do two consecutive hydrolysis reactions (on the NTS and TS) occur in a single transposase active site and which transpososome arrangements are required for excision?

I aimed to address the above questions by structural characterization of *Hsmar1* transposition. For this, I analysed the Hsmar1 transposase in its DNA-free and DNA-bound states by biophysical methods, crystallography, small angle X-ray scattering (SAXS), and structural modelling.

2.2.1 Architecture of the Hsmar1 transposase

Prior to my PhD work, I was involved in determining a crystal structure of the Hsmar1 transposase together with a PhD Fellow, Franka Voigt (Barabas group, EMBL Heidelberg, manuscript in preparation). The structure contains the catalytic domain (CAT) of the transposase (aa 118-343) including most of the flexible inter-domain linker (spanning aa 112-124) and provided preliminary insights into the architecture of the full-length Hsmar1 transposase that I investigated further in this thesis.

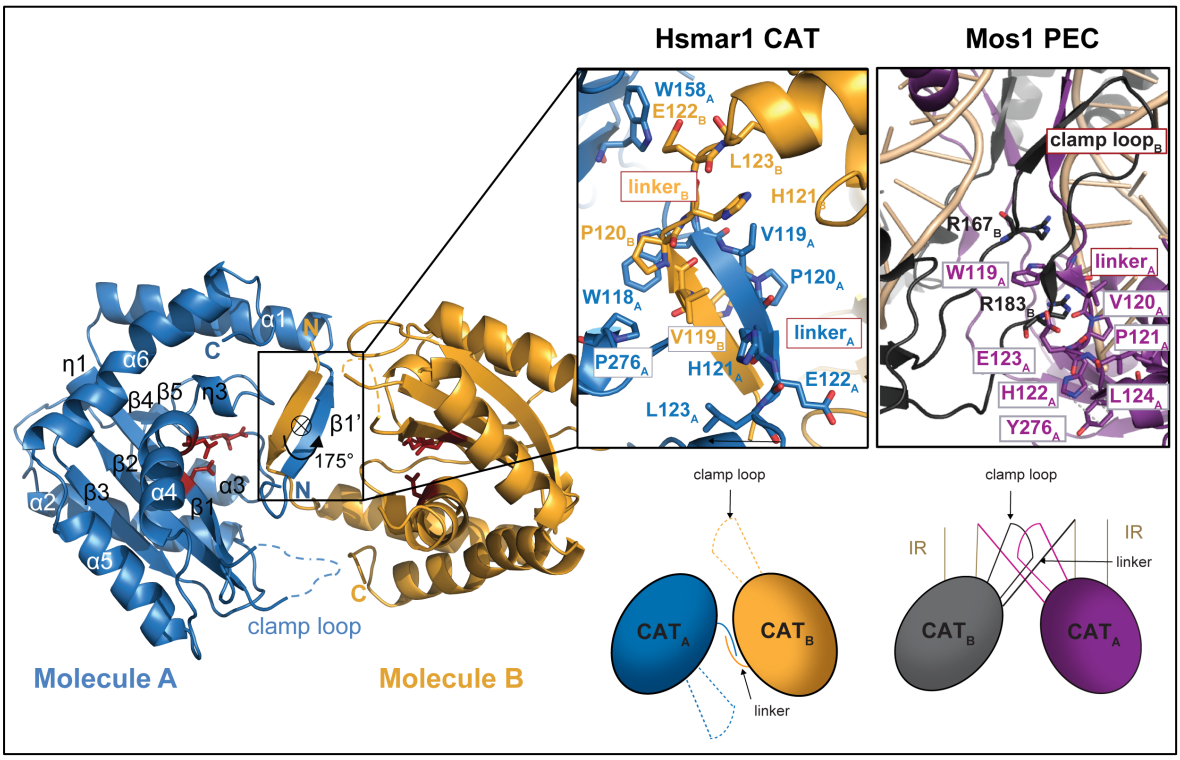


Figure 2-10: Crystal structure of the Hsmar1 transposase catalytic domain (CAT) reveals an autoinhibited dimer *in trans*. The two molecules (Molecule A, blue and Molecule B, orange) are asymmetric and superpose by 175° rotation. α-helices (α), β-strands (β) and turns (η) are indicated. The clamp loops that are not visible in the structure are indicated as dashed lines. The three catalytic residues (D155, D247 and D282; red sticks) are assembled in each active site. Left insert: specific amino acid interactions involving the conserved WVPHEL motif (aa 118-123) in the linker region create an autoinhibited conformation. Right insert: interactions involving the inter-domain linkers and the clamp loops in the Mos1 PEC [PDB ID: 3HOT; (Richardson *et al.*, 2009)]. Mos1 subunits are shown in purple and black. IR: inverted repeat. In both inserts, residues belonging to the WVPHEL motif and their interacting partners are shown as sticks. Below each inserts, a schematic representation of the interactions is shown.

In the structure, two molecules of the Hsmar1 CAT form a unique dimer assembly that is similar, but remarkably more compact, than the one observed in the Mos1 PEC structures (Dornan *et al.*, 2015; Morris *et al.*, 2016; Richardson *et al.*, 2009) (Figure 2-10). Moreover, the linker regions of the two Hsmar1 protomers interact with each other (Figure 2-10, left insert), whereas in the Mos1 PEC each linker contacts the clamp loop of the partner molecule (Figure 2-10, right insert; see also Figure 2-5 in section 2.1.4). The inter-linker interactions are established mainly by the conserved *mariner*-specific WVPHEL motif (spanning aa 118-123) (Figure 2-10, left insert). By these interactions, the linker of one molecule is tethered to the catalytic core of the other molecule in *trans*, covering the active sites and locking the catalytic domains in a closed catalytically incompetent conformation (Figure 2-10).

Intriguingly, mutations of the WVPHEL motif have been shown to result in hyperactivity of the corresponding transposase variants (D. Liu *et al.*, 2014). For example, substitution of W118 with proline (W118P) results in 4-fold increase of transposition activity. In light of our structure, this data implies that disruption of the linker-linker interactions by mutagenesis allows higher accessibility of the active sites, leading to a hyperactive phenotype.

Thus, the structure of the Hsmar1 CAT and previous mutagenesis data together suggested that the Hsmar1 transposase assumes a closed, inactive conformation in absence of DNA. In order to test this hypothesis, I first investigated the full length wild type Hsmar1 and hyperactive Hsmar1-W118P transposases by size exclusion chromatography (SEC). For this, I overexpressed both proteins in *Escherichia coli* (*E. coli*), purified them to homogeneity and finally analysed them on a SEC column (Figure 2-11).

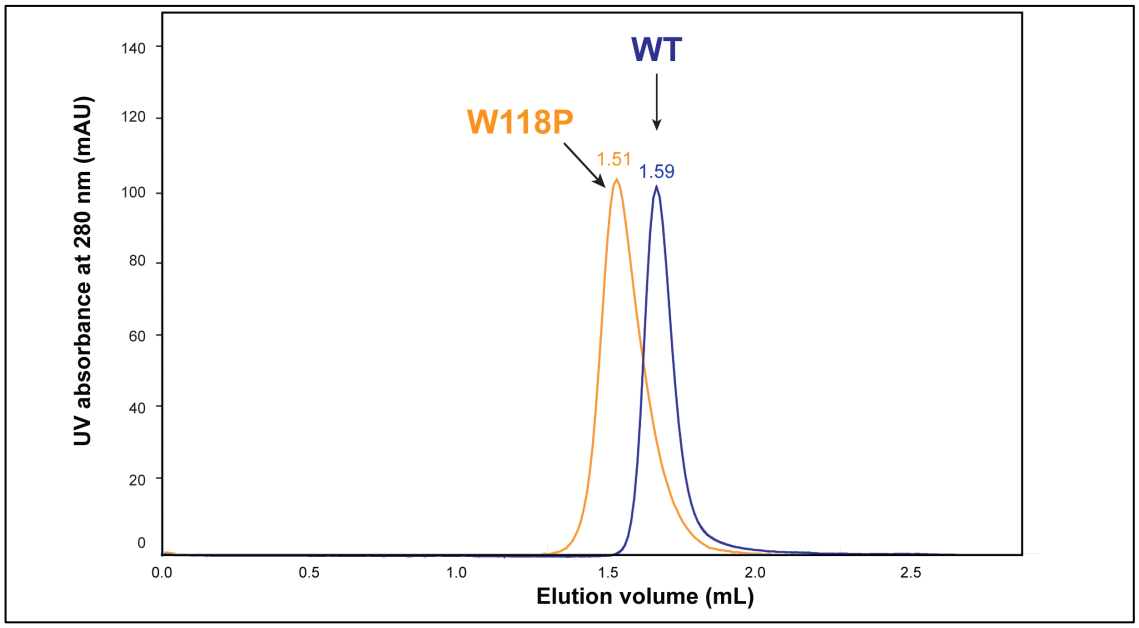


Figure 2-11: Analytical size exclusion chromatogram showing that wild type Hsmar1 (WT, blue curve) elutes later than Hsmar1-W118P (W118P, orange curve) from the column (Superdex 200, 3.2/30). Based on their elution volumes, both proteins form dimers.

In SEC, proteins are separated on the basis of their Stokes radius, which depends on their molecular size. The volume at which a protein elutes from the column (elution volume) is a direct reflection of its size and molecular shape, where larger or more elongated entities elute earlier, and smaller more compact proteins elute later (Nenortas, 1995). Interestingly, although both proteins elute at a volume consistent with homodimers, the Hsmar1-W118P

mutant shows a lower elution volume than the wild type Hsmar1, indicating a more expanded conformation for the mutant (Figure 2-11). This is consistent with the idea that disruption of the W118-P120 interaction by the W118P substitution causes "opening-up" of the Hsmar1-W118P transposase conformation.

To further confirm these findings and obtain structural insights into the different conformations of the Hsmar1 transposase, I then analysed the two protein variants in their apo state by SAXS (Figure 2-12). Based on the SAXS data, the calculated molecular weight (Mw) of the wild type transposase correspond to 80.59 kDa and is therefore consistent with a dimer in solution (theoretical Mw of the monomer is 40.5 kDa) (Figure 2-12 A). I then performed Guinier analysis of the obtained SAXS curves (Figure 2-12 B) to obtain the average radius of gyration (Rg): 35.5±0.9 Å for Hsmar1 wild type and 44.6±0.5 Å for Hsmar1-W118P (Figure 2-12 A). According to the paired distance distribution P(r) function, the maximum dimension (Dmax) of wild type Hsmar1 is 116.2 Å, whereas this parameter is significantly larger for Hsmar1-W118P (150 Å) (Figure 2-12 A and Figure 2-12 C). Thus, analogously to SEC, SAXS analysis indicates a compact shape for the wild type transposase and an expanded arrangement for the mutated variant. In agreement, *ab initio* shape reconstruction of the Hsmar1 dimer from the experimental SAXS data also revealed a compact molecular envelope shape (Figure 2-12 D).

The finding that the Hsmar1 transposase is quite compact in solution is in contrast with SAXS experiments performed with the Mos1 transposase, which showed that Mos1 forms an elongated dimer that is held together only by its N-terminal DBD (Cuypers *et al.*, 2013). The Rg and Dmax values reported for the Mos1 transposase (apo state) in solution were in fact significantly larger (Rg=49.2 Å, Dmax=185 Å) than the ones I obtained for the Hsmar1 transposase (apo) (Figure 2-12 A). The experimental Dmax of the apo Hsmar1 dimer (116.2 Å) is instead comparable to the calculated maximum length of the Mos1 transposase in the PEC (110 Å) (Figure 2-12 A), indicating an overall similar structural arrangement.

A

	I(0)	Mw (kDa)	Rg (Å) Guinier	Rg (Å) Gnom	Dmax (Å)	Porod volume
Hsmar1-WT apo	65.60±0.095	80.59	35.5±0.9	35.7±0.1	116.2	120.88
Hsmar1-W118P apo	90.96±0.640		44.6±0.5	43.8±1.1	150	131.48
Mos1 apo			49.2±0.1	49.5±0.2	185	
Mos1 PEC				38.2	110	

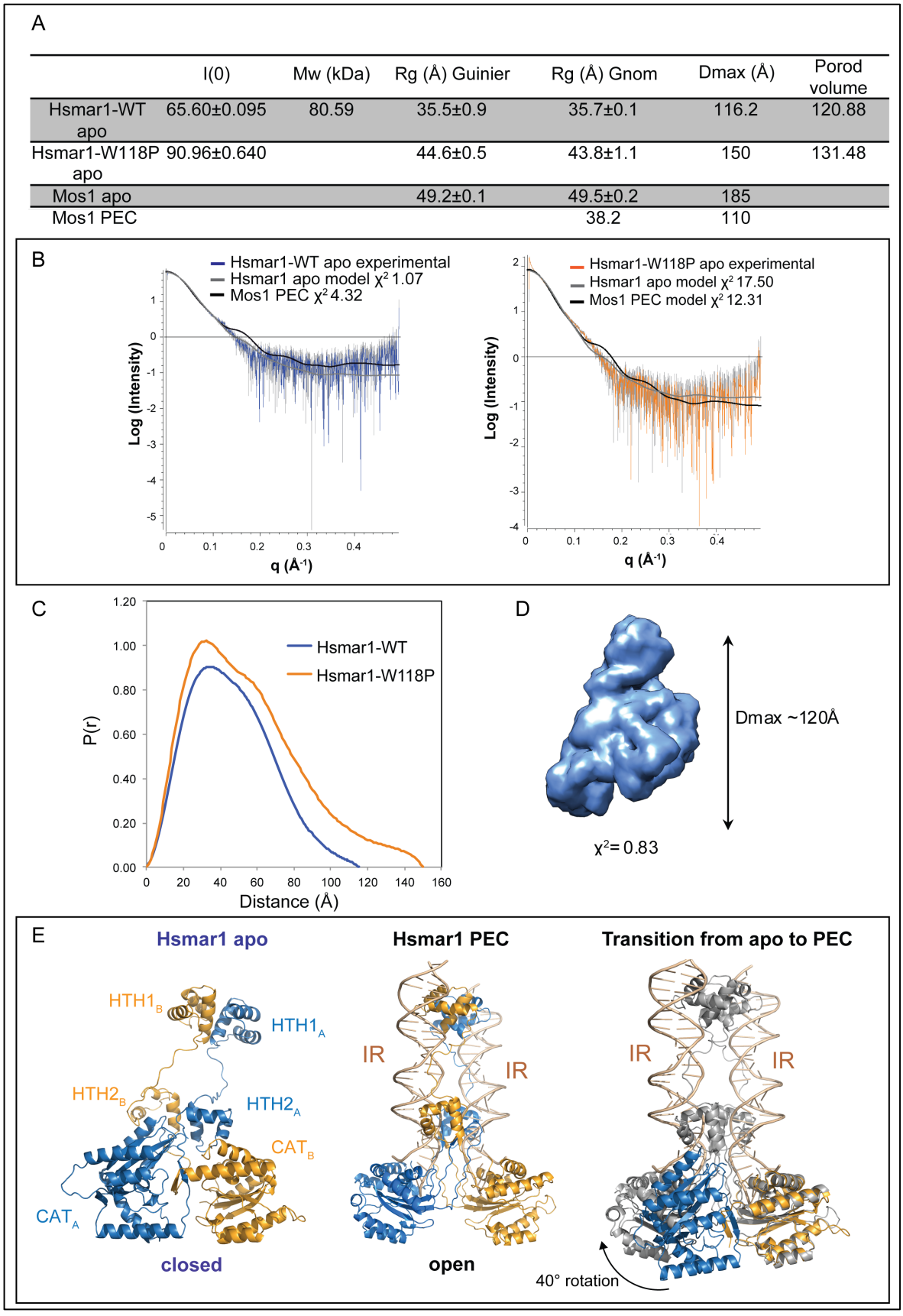


Figure 2-12: SAXS analysis of the wild type Hsmar1 (Hsmar1-WT) and W118P Hsmar1 (Hsmar1-W118P) transposases. (A) Comparison of structural parameters of the Hsmar1 variants and of the Mos1 transposase [Mos1 apo - parameters extracted from experimental SAXS data (Cuypers *et al.*, 2013), and Mos1 PEC -

parameters calculated from the atomic coordinates of the structure - PDB ID: 3HOT; (Richardson *et al.*, 2009)]. (B) Comparison of the experimental SAXS curves (for Hsmar1-WT, blue line, left; and for Hsmar1-W118P, orange line, right) with calculated scattering curves for the model of apo Hsmar1 (grey line) and for the Mos1 PEC (black line) (see models in E below). The quality of fit to the experimental data is evaluated by the chi square (χ^2) values, as indicated on the graphs. (C) Pair distribution function P(r) calculated from the SAXS data of the Hsmar1-WT (blue line) and the Hsmar1-W118P (orange line) transposases. (D) *Ab initio* shape reconstruction of the wild type Hsmar1 dimer calculated from the SAXS data using GASBOR (D. I. Svergun *et al.*, 2001). χ^2 value and maximum dimension (Dmax) are shown next to the model. (E) Structural models of the DNA-free Hsmar1 homodimer (Hsmar1 apo) predicted using the SAXS data (left), and of the Hsmar1 PEC (middle). Right: Superposition of the Hsmar1 catalytic domain structure (blue and orange molecules) with the Hsmar1 PEC model (grey dimer). Arrow illustrates the transformation (40° rotation) that would bring the blue subunit into the PEC conformation. IR: inverted repeat.

To visualize how the individual domains might be arranged in the full-length DNA-free Hsmar1 transposase dimer, I generated a structural model based on the SAXS data (Figure 2-12 E, Hsmar1 apo). As source of high-resolution information, I used the crystal structure of the catalytic domain dimer and homology models of HTH1 and HTH2 [generated using the Mos1 structure as template, PDB ID: 3HOT; (Richardson *et al.*, 2009)]. An ensemble of the Hsmar1 domains has been iteratively constructed and minimized against the low-resolution SAXS data. The resulting *in silico* model (Figure 2-12 E, Hsmar1 apo) approximates the experimental scattering data well [chi-square (χ^2) value of 1.07] (Figure 2-12 B, left panel) and visualizes the Hsmar1 transposase in a compact conformation, where both N-terminal and C-terminal domains are involved in protein dimerization, differently than in the Mos1 transposase.

In turn, the conformation assumed by the Mos1 transposase in the PEC poorly agrees with the experimental SAXS data of the Hsmar1 transposase (χ^2 value of 4.32, Figure 2-12 B). Given the significant sequence similarity and largely similar biochemistry of the two proteins it is reasonable to assume that Hsmar1 assembles a highly similar active PEC as Mos1. Consistently, I was able to confidently model the Hsmar1 PEC based on the Mos1 PEC structure (Figure 2-12 E, middle panel). Notably, the experimental SAXS curve of Hsmar1-W118P is not well represented by either the Hsmar1 apo model or the PEC (Figure 2-12 B, right panel). This suggests that Hsmar1-W118P assumes an open conformation, probably due to disruption of specific inter-linker interactions by mutagenesis, and its assembly is distinct from the ones assumed by the wild type protein.

In conclusion, SAXS data and molecular modelling suggests that the Hsmar1 transposase forms "closed" inhibited dimers in its DNA-free state (Figure 2-12 E, Hsmar1 apo) and "open" active dimers in its PEC form (Figure 2-12 E, Hsmar1 PEC). Structural transition from one conformation to the other would mostly involve discrete movements of the catalytic domains (circa 40 degree rotation between the two states, Figure 2-12 E, right panel) and restructuring of the WVPHEL motifs in particular. We speculate that the Hsmar1 transposase exists as a closed, inactive dimer at early stage of the transposition pathway to prevent catalysis on the DNA prior to transposon end synapsis (as discussed in section 2.5.3).

2.2.2 Biochemistry of *Hsmar1* excision

In order to elucidate the mechanism of *Hsmar1* excision, I first focused on identifying the order of the DNA hydrolysis reactions performed by the Hsmar1 transposase on transposon ends. To this aim, I performed *in vitro* cleavage assays using radioactively labelled substrates (Figure 2-13 A). To test efficiency and specificity of cleavage on each strand, different substrates imitating a single intact (IR) and nicked (N_1 and N_2) *Hsmar1* inverted repeats (Figure 2-13 A, top) were incubated with recombinant Hsmar1 transposase. Transposase-mediated cleavage products were analysed via denaturing PAGE (Figure 2-13 A, bottom). In order to detect cleavage on a specific strand, the 5' end of the NTS or the TS of the dsDNA substrate was labelled with ^{32}P (as indicated above the gel lines).

Consistently with previous reports (Miskey *et al.*, 2007), I found that the Hsmar1 transposase can specifically cleave both strands of an intact transposon end (IR). Remarkably, the protein can still perform cleavage on the TS of the N_1 substrate, containing a nick in the NTS strand. On the contrary, the NTS could not be cut when the TS was nicked (N_2 substrate; Figure 2-13 A, bottom). These findings indicate that NTS cleavage strictly precedes TS cleavage (Figure 2-13 B). None of these cleavage events occurs in the presence of Ca^{+2} (Figure 2-13 A, bottom), as reported for other RNase H-like nucleases (W. Yang *et al.*, 2006).
Several additional reaction products were detected in the *in vitro* cleavage reactions (Figure 2-13 A, bottom), which likely result from integration of the cleaved transposon end substrates at TA dinucleotides present in their own DNA sequence. This indicates that the Hsmar1 transposase is able to perform also the integration reactions *in vitro*.

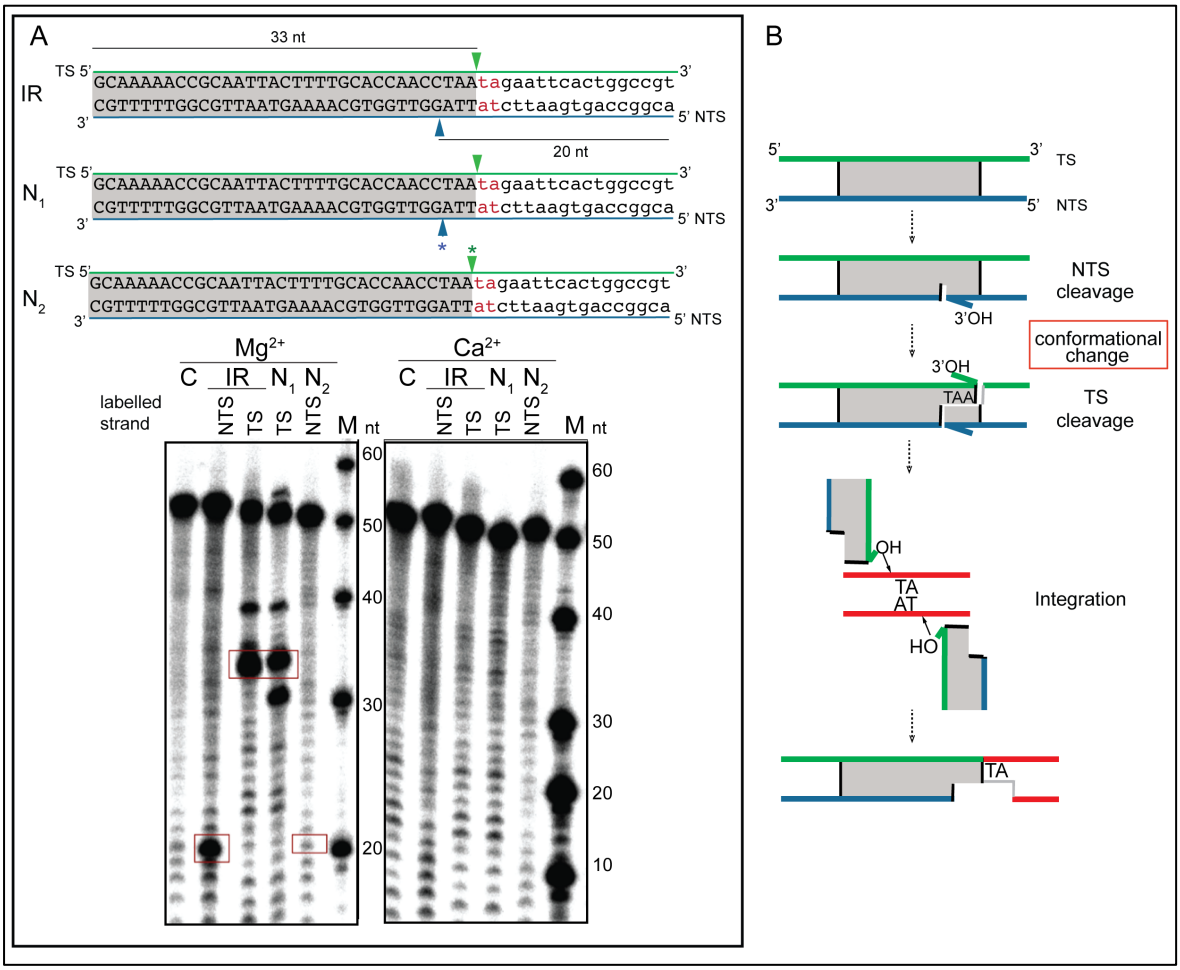


Figure 2-13: Characterization of Hsmar1 cleavage activity *in vitro*. (A) *Hsmar1 in vitro* cleavage assay. Top: dsDNA substrates used in the assay. The inverted repeat is shaded grey and its DNA sequence is shown in capital letters; flanking DNA is indicated in lower case letters. The invariable flanking TA dinucleotide is in red. Green and blue arrowheads indicate the position of cleavage on the transferred strand (TS) and non-transferred strand (NTS), respectively. Arrowheads with asterisks indicate the position of a nick in the N_1 and N_2 substrates. Bottom: 12% Urea PAGE resolving ^{32}P-labelled DNA substrates (IR, N_1 or N_2, labelled on the NTS or TS, as indicated above each gel lane) and products after incubation with Hsmar1 transposase in a 10:1 DNA to protein molar ratio. Bands corresponding to specific cleavage products are framed by red rectangles on the gel. The control samples, labelled C, did not contain any protein. Marker (M) sizes are indicated next to the gel. (B) Proposed mechanism of transposon end cleavage and strand transfer in the *Hsmar1* transposon [adapted from (Robertson *et al.*, 1997)]. Schematic code equivalent to A. DNA strands of the target DNA are depicted as red lines.

Taken together, my biochemical data revealed that the two transposon DNA strands are cleaved in a specific order, with NTS preceding TS cleavage. This implies that a conformational change in the transposase-DNA complex occurs between these breakage events. For energetic reasons, it is likely that the last step, namely TS cleavage, occurs in a complex conformation immediately competent for successive transposon integration (Figure 2-13 B).

2.2.3 *In vitro* reconstitution and crystallization of pre-excision Hsmar1 protein-transposon end DNA complexes

In order to identify the structural basis of *Hsmar1* excision, I aimed to determine the structure of pre-excision Hsmar1 protein-DNA complexes by X-ray crystallography. To this aim, a preliminary requirement was the reconstitution of such complexes *in vitro*. For this, I designed complexes for structural studies based on my biochemical insights.

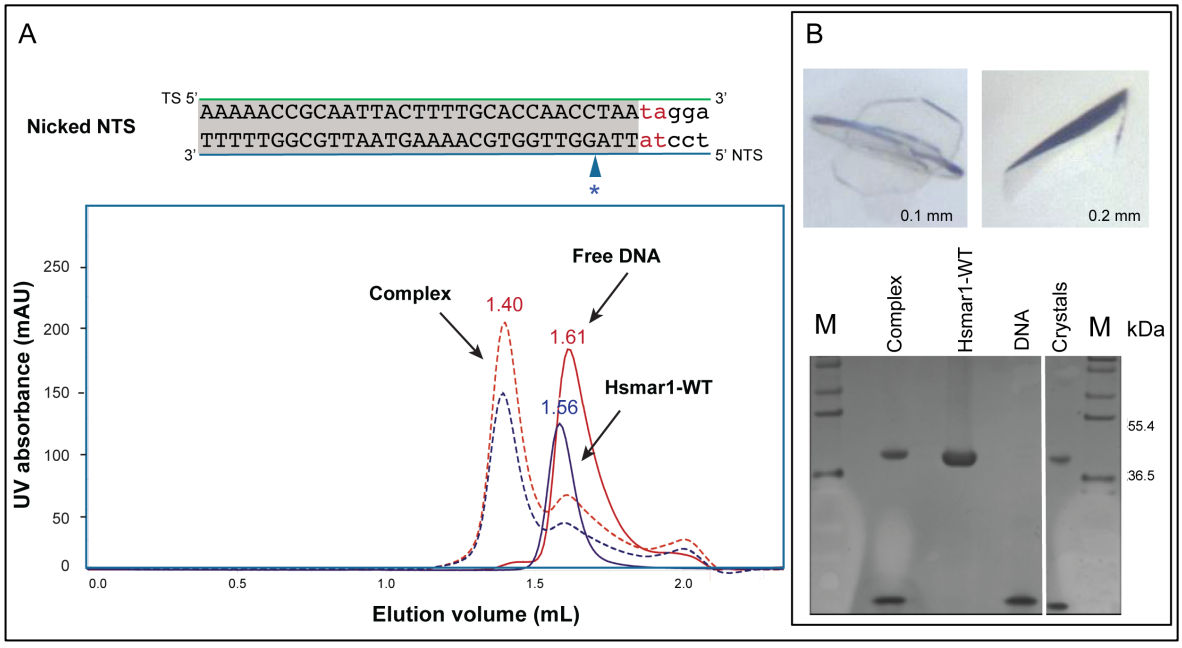


Figure 2-14: *In vitro* reconstitution and crystallization of Hsmar1 protein-transposon end DNA complexes. (A) Top: dsDNA probe used for nucleoprotein complex formation. Schematic code equivalent to Figure 2-13. The asterisk indicates the position of a synthetic nick on the NTS of the dsDNA probe. Bottom: analytical size exclusion chromatogram (column Superdex 200, 3.2/30) showing that the Hsmar1 transposase (Hsmar1-WT, blue curve, absorption at 280 nm) binds to the Nicked NTS probe (Free DNA, red curve, absorption at 260 nm), forming a stable, homogeneous nucleoprotein complex (Complex, dashed blue and dashed red lines representing absorption at 280 and 260 nm, respectively). Elution volume of the complex shifts with respect to individual complex components. (B) Top: crystals of the reconstituted Hsmar1-Nicked NTS complexes. Maximum crystal sizes are stated on the images. Bottom: SDS-PAGE analysis of the crystals. Silver-stained 4-12% Bis-Tris NuPAGE gel showing both the Hsmar1 transposase and the Nicked NTS DNA present in the crystals. The relevant molecular weights of the protein marker (M) are marked next to the gel.

I incubated the wild type Hsmar1 transposase with a dsDNA probe representing a transposon end after first strand nicking in non-catalytic conditions (Ca^{+2}). The DNA contained a single IR flanked by some genomic DNA sequences and was nicked at the cleavage site on the NTS strand (Nicked NTS, Figure 2-14 A, top). Successively, the samples of protein only, DNA only, and protein-DNA mixture were run separately on an

analytical SEC column to assess complex formation. In SEC-based binding assays, if a complex between protein and the DNA is formed, the size of the complex will be larger than that of the single components, resulting in a shift of the complex's elution volume in comparison to the elution volumes of protein and DNA alone (Nenortas, 1995). The elution profile of the Hsmar1 protein-Nicked NTS DNA mixture in analytical SEC showed the formation of a stable and homogenous complex (Figure 2-14 A, bottom). This indicates that the reconstituted complex likely assumes a discrete pre-excision conformation, since in the presence of Ca^{+2} even NTS cleavage is blocked [(Figure 2-13 A and (Claeys Bouuaert *et al.*, 2010)].

Using this complex, I performed crystallization trials and obtained crystals that grew in two months in 20% (w/v) PEG3350 and 0.2 M sodium thiocyanate, pH 6.9 in sitting drops using the vapour diffusion method (Figure 2-14 B, top). Crystals dissolved in water, run and visualized on a silver-stained gel show to contain both the Hsmar1 transposase and the DNA probe (Figure 2-14 B, bottom).

2.2.4 Crystal optimization and X-ray data collection

As initial crystals grew as multiple crossing plates and diffracted X-rays poorly, I made use of several strategies, as listed in Table 2-1, to optimize crystal quality.

DNA SUBSTRATES	Crystallization conditions
Pre-cleaved, full, gapped, nicked, mutated transposon end DNA	- All available sparse matrix and grid screens - 0.2 M sodium thiocyanate, 20 % (w/v) PEG 3350 - 0.1 M TRIS HCl pH 8.5, 15 % (w/v) PEG 20000 - 0.2 M sodium acetate, 20 %(w/v) PEG 3350 - different iodide, fluoride, tartate, formate salts
PROTEIN VARIANTS	**Cryoprotection**
- wild type - catalytic mutants - cysteine mutants - thioredoxin-tagged - TS cleavage-deficient mutant - SeMet derivative	Cryoprotectants: Glycerol, PEG3350, PEG400, MPD, Ethylene, glycol-NDSB-glucose, 2,3-butenediol - Flash-cooling in the cryo-stream
COMPLEX FORMATION	
Buffers: HEPES, BisTris propane, pH Ions: $CaCl_2$, $MgCl_2$, EDTA Protein:DNA ratios (1:1.5, 1:1.7, 1:1.6)	
CRYSTAL SCREENS	**(Post-) crystallization treatments**
- 24- and 96–wells plates - Sitting drop, hanging drop - Additives - Complex solution:crystallization solution ratios (1:1,1:2, 2:1)	- Annealing - Dehydration - Seeding - CrystalDirect™ method

Table 2-1: Summary of constructs and conditions tested in order to improve the diffraction limit of the crystals. The approaches resulting in diffraction improvements and/or described in details are in red.

Firstly, I designed and tested diverse DNA substrates aiming to improve complex homogeneity and stability and/or the crystal packing. New DNA substrates varying in length, sequence, and containing nicks and/or gaps and sticky and/or blunt ends have been designed and tested for crystallization. Moreover, I introduced mutations of the transposon end in positions surrounding the cleavage site to create "uncleavable" DNA substrates, as probed by *in vitro* cleavage assays (Figure 2-15 A), to help locking the complex in the pre-cleavage state.

Secondly, I tried to crystallize the protein-DNA complex using different constructs of the Hsmar1 transposase: catalytic mutant (where aspartate 155 in the catalytic triad was substituted with alanine); cysteine mutants (to prevent adventitious protein oxidation); Thioredoxin fusion protein (so as to form additional protein interactions in the crystal supporting the packing); and a TS-cleavage deficient mutant. In the latter variant, I substituted the residue Val 199 with Gly in the transposase, as this mutation was shown to block TS cleavage [(Claeys Bouuaert *et al.*, 2014) and Figure 2-15 B], supposedly trapping a pre-excision complex. However, no or poorly diffracting crystals grew in these conditions.

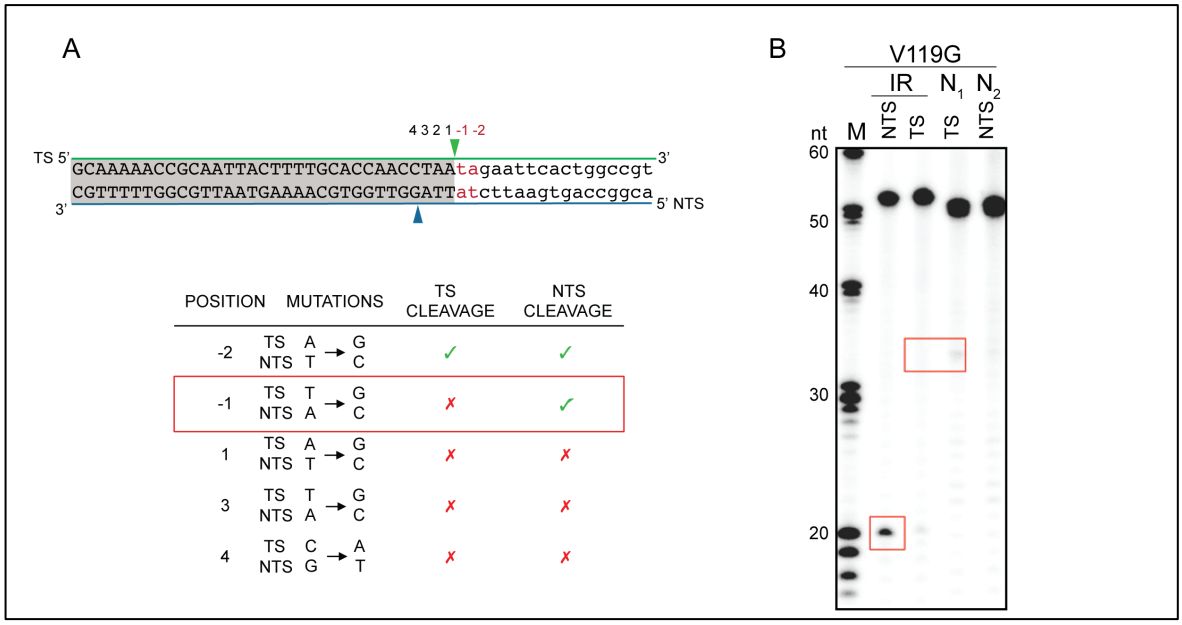


Figure 2-15: Effects of mutating *Hsmar1* transposon ends or transposase residues on NTS and TS cleavage. (A) Effect of *Hsmar1* transposon end mutations on NTS and TS cleavage. Arrowheads mark the specific positions of DNA cleavage on transposon ends by the Hsmar1 transposase. The table summarizes the results of *Hsmar1 in vitro* cleavage assays monitoring Hsmar1-mediated cleavage of mutated transposon ends [on the non-transferred strand (NTS) or on the transferred strand (TS)]. (B) *Hsmar1 in vitro* cleavage assays showing that the Hsmar1-V119G mutant is deficient in TS cleavage (see *in vitro* cleavage assay with wild type protein in Figure 2-13 for comparison). 12% Urea PAGE resolving ^{32}P-labelled dsDNA substrates (IR, N_1 or N_2, as indicated) and products after incubation with Hsmar1-V119G transposase in a 10:1 DNA to

protein molar ratio. Positions corresponding to expected cleavage products are framed by red rectangles on the gel. M: marker.

Finally, I focused on refining the crystallization conditions by exhaustive screening around the initial crystallogenic condition [20% (w/v) PEG3350 and 0.2 M sodium thiocyanate, pH 6.9] in larger drops. By this method, I obtained remarkable improvements in crystal size (Figure 2-16 A). Then, I applied different treatments, such as annealing, dehydration (Russo Krauss *et al.*, 2012), and seeding to the obtained crystals. Since my crystals were highly sensitive to handling, cooling, and cryoprotection procedures, I also screened different cryoprotectants, I tested direct freezing of the crystals in the 100 K nitrogen gas stream, and I used the CrystalDirect™ technology (Cipriani *et al.*, 2012) for automated crystal harvesting and cryo-cooling (together with V. Mariaule and J. Márquez, Márquez Laboratory, EMBL Grenoble, France).

Out of these attempts, the crystal quality improved notably by applying the manual hanging drop setup in combination with streak seeding methods (Figure 2-16 A). In addition, the use of 2,3-butanediol as cryoprotectant and flash-cooling in the cryo stream resulted in improved diffraction quality. By these methods, I could obtain diffraction and collected datasets with resolution in the 4-5 Å range (Figure 2-16 A) at synchrotron facilities ESRF in Grenoble and PetraIII in Hamburg. The best crystals diffracted to 4.25 Å (Figure 2-16 B and C). Data processing by XDS and XSCALE (Kabsch, 2010) permitted to calculate the space group (C2) and the unit cell parameters (212.97 187.17 461.6 90 102.51 90). Due to high diffraction anisotropy, data processing has been challenging and available datasets, are currently being processed by the STARANISO server alone and in combinations to improve data quality (Tickle *et al.*, 2017).

Using the available data, initial attempts to solve the structure of the Hsmar1 complexes were performed by using the molecular replacement (MR) method in Phaser-MR (McCoy *et al.*, 2007). For this, the structure of the Mos1 PEC [PDB ID: 3HOT (Richardson *et al.*, 2009), 37% transposase sequence similarity with Hsmar1] has been used as search model. Molecular replacement solutions with high log-likelihood gain (LGG) scores were obtained. Inspection of the calculated electron density maps showed continuity of the function along all three unit cell directions, indicating a meaningful solution with compact packing of the molecules in the crystal lattice. However, due to the large number of molecules in the asymmetric unit and the limited diffraction data quality, refinement of

these MR solutions was challenging. Thus, the correctness and completeness of the solutions could not be unambiguously assessed.

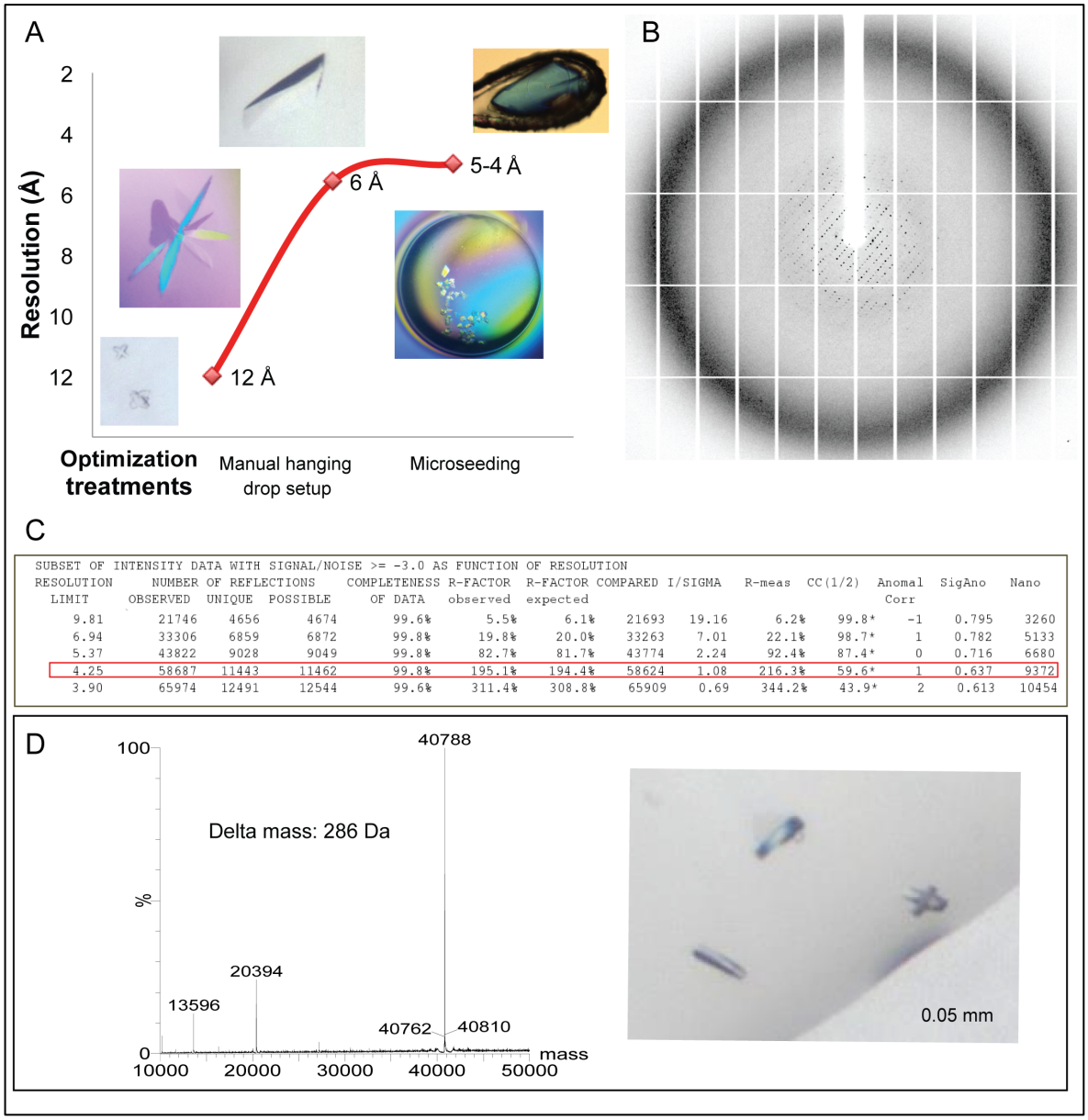


Figure 2-16: Optimization, X-ray data collection and derivatization of the Hsmar1-Nicked NTS crystals. (A) Hsmar1-Nicked NTS crystals before and after optimization by the indicated techniques. (B) Diffraction image of optimized Hsmar1-Nicked NTS crystals. (C) X-ray diffraction data statistics obtained from data processing using XDS and XSCALE (Kabsch, 2010). (D) Left: mass-spectrometry showing the calculated mass of the Hsmar1 selenomethionine derivative (40,788 Da) and its mass shift in comparison with the native protein (286 Da). Right: selenomethionine derivative crystals of the Hsmar1 protein-Nicked NTS DNA complex. Maximum crystal size is indicated on the image.

To improve the confidence and quality of phases, I next focused on combining MR with experimental phasing methods, namely multi-wavelength anomalous (MAD) and multiple isomorphous replacement (MIR) phasing approaches. For applying MAD, I first produced recombinant selenomethionine (SeMet) derivatives of the Hsmar1 transposase in *E. coli*

grown in a minimal medium supplemented with L-SeMet. Complete derivatization of the protein with SeMet (incorporation of 6 SeMet residues) was confirmed by matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) analysis (Figure 2-16 D, left). Crystals of the complexes formed with the SeMet protein derivative grew in the same conditions as the native crystals (Figure 2-16 D, right) and were diffracted at synchrotron facilities ESRF in Grenoble and PetraIII in Hamburg. For MIR experiments, I soaked native crystals with heavy atom (e.g. cadmium and platinum) compounds and collected datasets at the absorption edge of these elements. At PetraIII P13 beamline, I also performed data collection on native crystals close to the absorption edge of calcium, since this ion is included in my complexes and likely present in my crystals.
Derivative datasets are currently being processed in the XDS suite. Once completed, the use of MAD and MIR approaches will be tested to determine the correctness and improve the quality of the previous MR solutions. To overcome the challenges, data analysis and structure solution is being carried out in collaboration with A. McCarthy, EMBL Grenoble, France.

2.3 Results – Biochemistry of *SB* transposition

Despite its broad applications as a genetic tool, very little is known about the transposition mechanism of the *SB* transposon. This is mainly due to limited biochemical and structural information available on the SB transposase and its active complexes. To date, recombinant production of the full-length SB transposase in sufficient quantity and quality for *in vitro* characterization has been challenging. However, in the Barabas Laboratory, purification of the SB transposase and preliminary activity analysis have been successfully established prior to my thesis work (Voigt *et al.*, 2016). Building on this, I focused on elucidating the biochemical principles governing *SB* transposition and on providing structural insights. To this aim, here I reconstituted and analysed *in vitro* the molecular species involved in each step of the transposition reaction, namely transposon end recognition, transposon excision and transposon integration.

2.3.1 *SB* transposon end recognition

SB transposon ends contain four transposase binding sites in total: left outer (Lo), left inner (Li), right inner (Ri) and right outer (Ro) direct repeats (DRs). Outer and inner DRs have similar, but not identical sequences (Figure 2-9 in section 2.1.10) and are both necessary

and not interchangeable for efficient transposition *in vivo* (Cui *et al.*, 2002). Previous *in vitro* binding assays reported for *SB* have shown that the transposase DBD (residues 1-123) binds the inner DRs more tightly than the outer DRs, suggesting distinct roles for these binding sites in transposition (Cui *et al.*, 2002). Thus, in order to elucidate the molecular features of transposon end recognition, I analysed the binding activity of the full length SB to the different DRs *in vitro* by SEC.

2.3.1.1 Visualization of SB protein-transposon end DNA complexes

I first overexpressed the full length SB transposase in *E. coli* and purified it to homogeneity. I then tested the purified SB protein for its ability to bind specific dsDNA probes comprising the distinct sequences of single DRs and some flanking DNA (Lo′, Li′, Ri′ and Ro′, Figure 2-17, above corresponding chromatogram) using SEC. For this, SB was mixed with the different DNA probes in non-catalytic conditions [with Ca^{2+} substituting the native cofactor Mg^{2+} in the buffer (Dawson *et al.*, 2003)]. Then, the samples of protein only, DNA only, and protein-DNA mixtures were run on a SEC column (Figure 2-17).

Incubation of SB with the outer DRs (Lo′ and Ro′, respectively) resulted only in a partial shift in elution volumes compared to DNA alone and protein alone controls. The peaks corresponding to the SB-Lo′ and SB-Ro′ samples are quite broad, reflecting a heterogeneous population of species, including protein only, DNA only and low-affinity nucleoprotein complexes. In turn, more stable nucleoprotein complexes have formed when SB was incubated with the Li′ or Ri′ probes, respectively (SB-Li′ and SB-Ri′ samples). This indicates that the SB transposase binds more strongly the inner DRs than the outer DRs. Based on the elution volumes, the estimated size of these complexes corresponds to a monomer of SB bound to a single DR. This is in agreement with previous observations in our laboratory that the SB transposase is a stable monomer and binds to transposon end DNA as a monomer (Franka Voigt, unpublished).

Of note, the SB-Li′ complexes show higher UV absorbance than the SB-Ri′ complexes despite mixing equal protein and DNA amounts (Figure 2-17, Li′ and Ri′), This indicates that some precipitation occurred upon mixing SB and Ri′. I observed the same pattern with the outer DRs, where the SB-Lo′ mixture contains more protein and DNA than the SB-Ro′ mixture, as judged by the height of the corresponding SEC peaks (Figure 2-17, Lo′ and

Ro′). Based on these observations, SB seems to have higher affinity to the left DRs than to the right DRs.

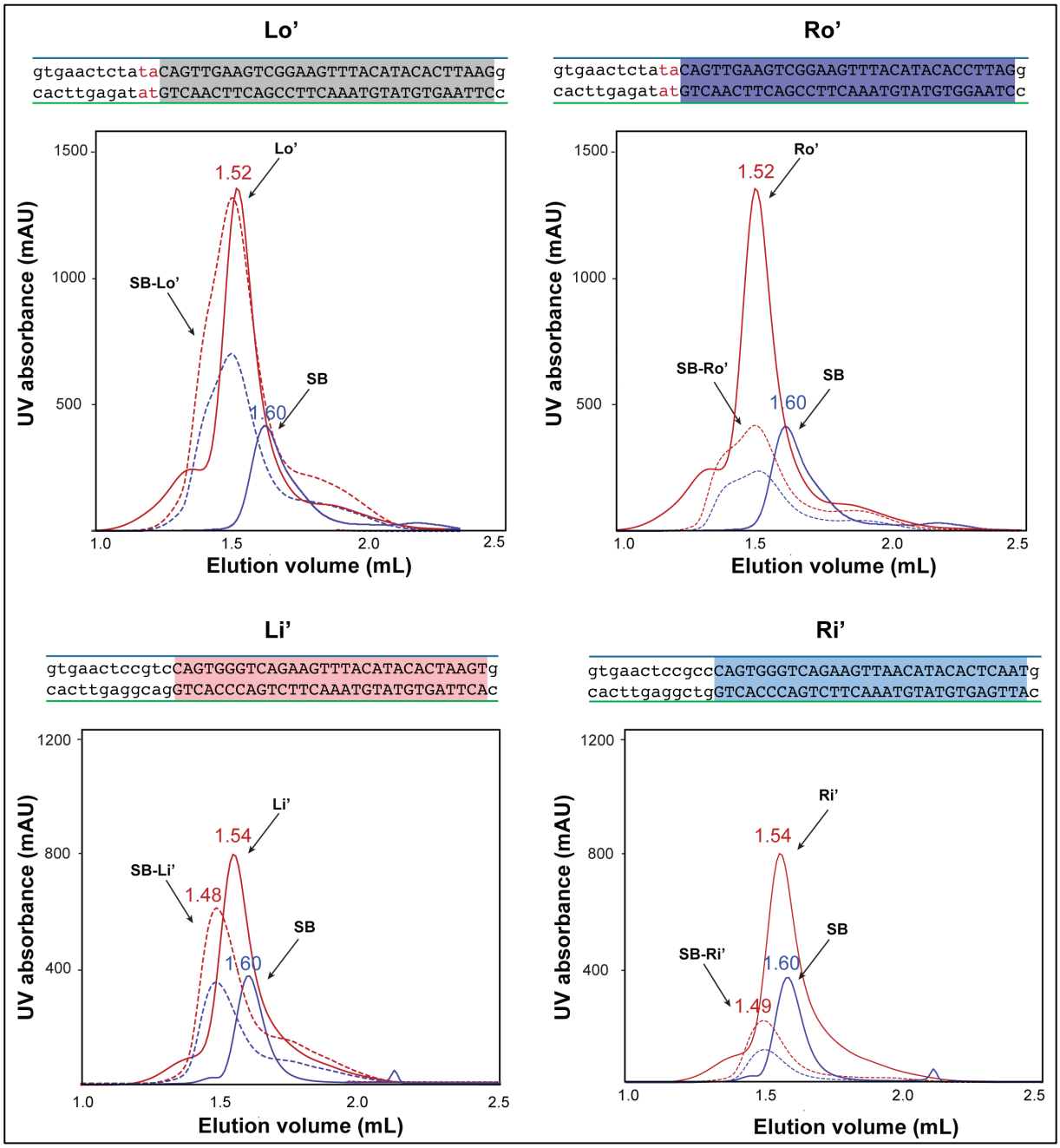


Figure 2-17: Characterization of SB binding to transposon end DNA *in vitro*. Analytical SEC (column Superdex S200 3.2/30) profile of the SB transposase (SB; blue curve, absorption at 280 nm), dsDNA probes (Lo′, Ro′, Li′, Ri′; red curves, absorption at 260 nm) and SB protein-DNA mixtures (dashed blue and dashed red lines representing absorption at 280 and 260 nm, respectively). Sequence of dsDNA probes used in the binding studies are shown above the corresponding chromatograms. Schematic code equivalent to Figure 2-9 in section 2.1.10.

2.3.1.2 Identification of high affinity SB binding sites at the outer repeats

After transposon end binding, the SB transposase performs DNA cleavage and ligation reactions exclusively at the outer DRs (Lo and Ro). However, my initial attempts to reconstitute functional SB transposase-outer DRs complexes *in vitro* have shown limited success, probably due to the moderate affinity of the SB transposase to these sites prior to catalysis (Figure 2-17, Lo′ and Ro′). Lack of stable, functional, and homogeneous SB-outer DRs complexes greatly hampers both biochemical and structural characterization of *SB* transposition.

To overcome this, I set out to identify high affinity outer repeat substrates. To this end, I mixed the SB transposase with dsDNA probes representing processed outer DRs at different stages of the transposition reaction and tested complex formation by SEC. In particular, I used dsDNA probes mimicking Lo after (i) first strand nicking (Nicked Lo); (ii) double-strand DNA cleavage (Cleaved Lo); (iii) ligation to target DNA (Gapped Lo 1-2-3 and Gapped Lo-Li) (Figure 2-18 and Figure 2-19). I also tested different buffer conditions, several DNA oligonucleotides of varying length and sequence, and various protein-DNA mixing ratios for optimal complex formation. In these attempts, I could reconstitute stable nucleoprotein complexes with various substrates, as shown in Figure 2-18 and Figure 2-19 with the specific DNA probes indicated above each chromatogram.

Interestingly, compared to pre-cleaved substrates (Cleaved Lo), the use of gapped or nicked substrates increased the stability and homogeneity of the corresponding complexes significantly, as revealed by the single, narrow peaks observed in SEC (Figure 2-18). This indicates that the presence of nicks or gaps at the cleavage site promotes binding to the SB transposase. This might be explained by specific recognition of nicks and gaps by the SB transposase, as a signature of specific transposition steps. Alternatively, they might confer higher flexibility to the DNA, facilitating its correct accommodation in the catalytic pocket of the transposase. Additionally, short stretches of flanking genomic DNA also increased transposase binding, probably due to direct interaction of specific protein surfaces with flanking DNA (in particular, the TA dinucleotide immediately at the transposon flanks) during transposon binding and target recognition (Dornan *et al.*, 2015).

Moreover, the SEC data shows that the length and position of the gaps on the NTS greatly affect the affinity of the SB transposase for Lo. In particular, DNA substrates with a

specific 5 nt gap spanning from the cleavage site to some flanking DNA (see Gapped Lo2, Lo3 and Lo-Li in Figure 2-18 and Figure 2-19), formed more stable complexes than those where the gap covers a longer region (7 nt, data not shown) or the cleavage site only (Gapped Lo1, Figure 2-18).

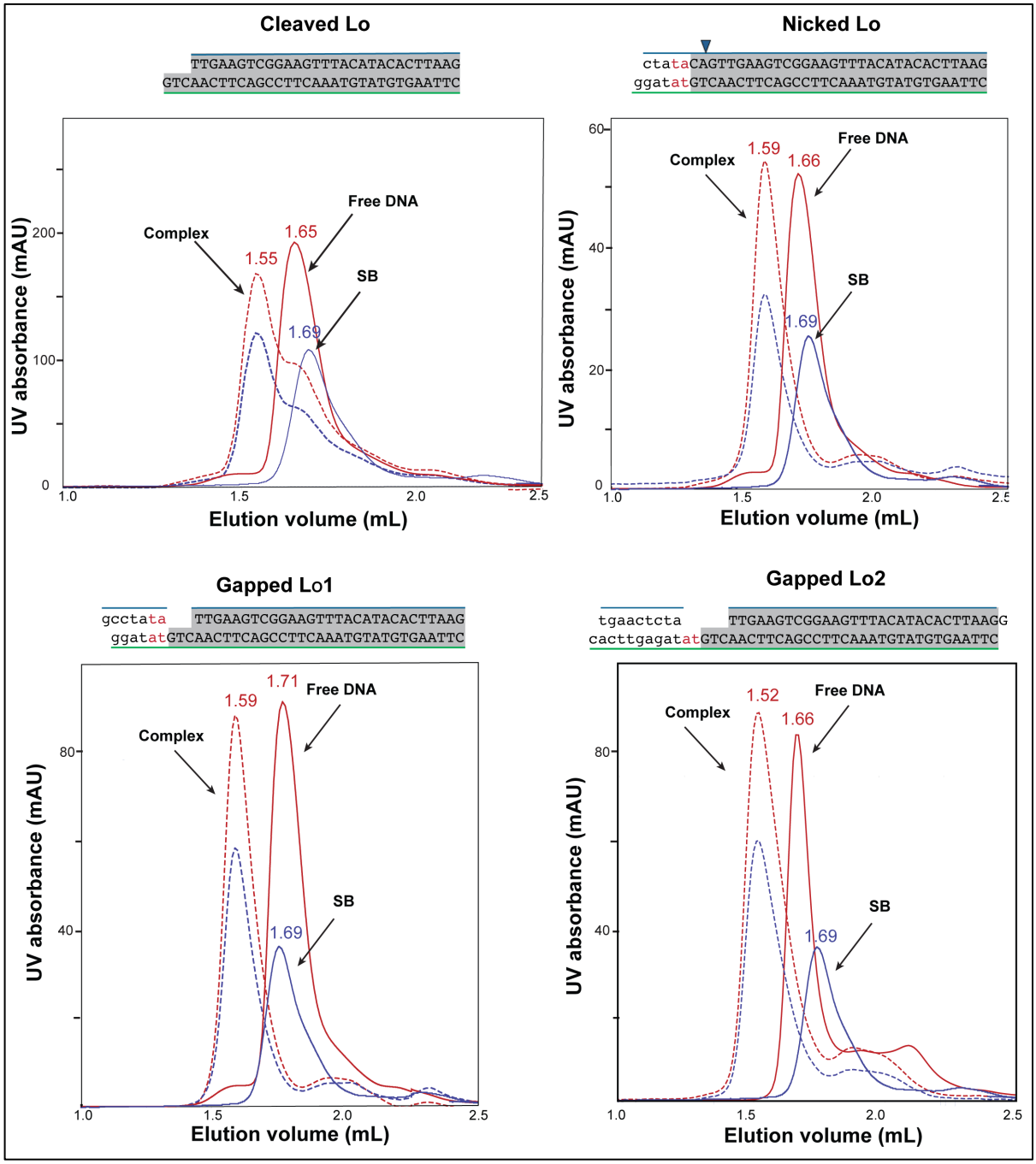


Figure 2-18: Characterization of SB binding to left outer repeats (Lo) *in vitro*. Analytical SEC (column Superdex S200 3.2/30) profiles of the SB transposase (SB; blue curve, absorption at 280 nm), dsDNA probes (Cleaved Lo, Nicked Lo, Gapped Lo1 and Gapped Lo2; red curves, absorption at 260 nm) and SB protein-DNA mixtures (dashed blue and dashed red lines representing absorption at 280 and 260 nm, respectively).

Sequence of the dsDNA probes used are shown above the corresponding chromatograms. Schematic code equivalent to Figure 2-9 in section 2.1.10. Blue arrowhead indicates the position of the nick on the non-transferred strand (NTS) in Nicked Lo.

Despite the identification of high affinity Lo DNA substrates, I detected significant precipitation during complex formation in all cases. To overcome these problems, I identified specific amino acid substitutions in the SB protein (SB100X) sequence that conferred high solubility and stability to the corresponding mutant. In particular, residues C176 and I212 were mutated into serines to obtain the $SB_{C176S-I212S}$ transposase variant (also named "high solubility SB"- hsSB). The design and characterization of this transposase variant is described in more detail in section 3.2. By using this specific transposase mutant, I successfully formed stable nucleoprotein complexes with a gapped DNA probe (Gapped Lo3) without significant protein precipitation (Figure 2-19, left).

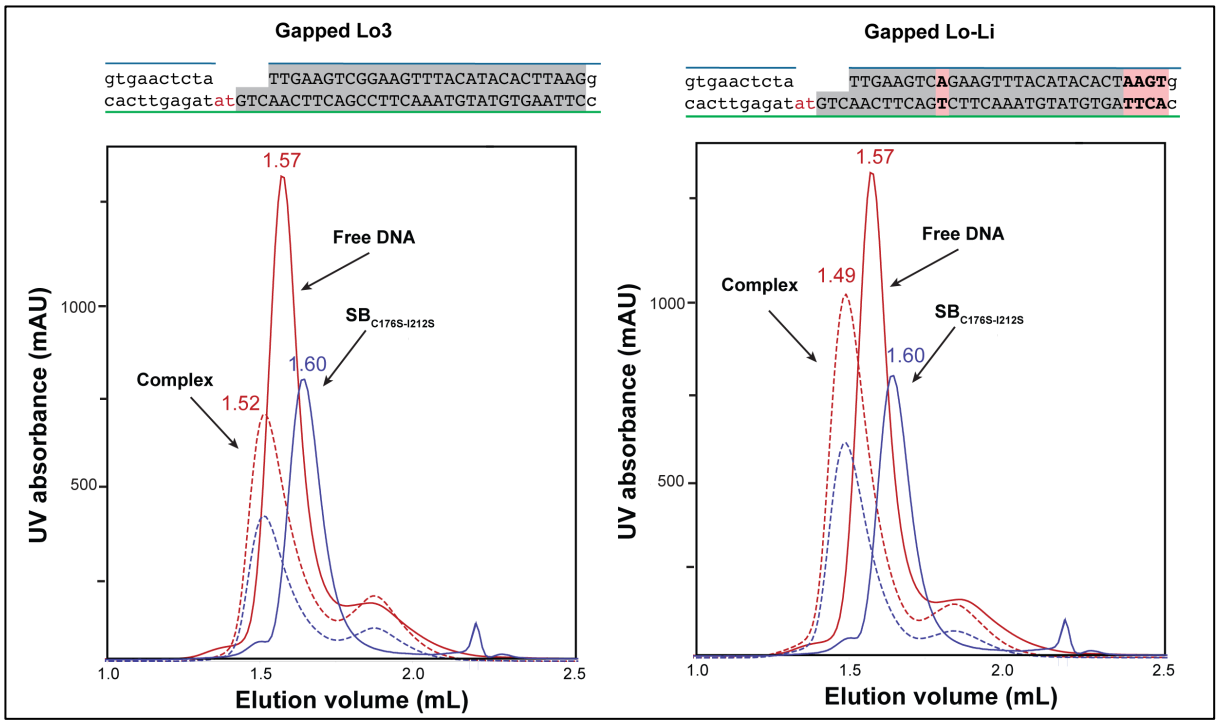


Figure 2-19: Reconstitution of high affinity SB protein-transposon end DNA complexes *in vitro*. Analytical SEC (column Superdex S200 3.2/30) chromatograms showing that the $SB_{C176S-I212S}$ transposase ($SB_{C176S-I212S}$; blue curve, absorption at 280 nm) stably binds to dsDNA probes (Gapped Lo3 and Gapped Lo-Li; red curves, absorption at 260 nm), forming homogeneous complexes (dashed blue and dashed red lines representing absorption at 280 and 260 nm, respectively). Sequence of dsDNA probes used for complex formation are shown above the corresponding chromatograms. Schematic code equivalent to Figure 2-9 in section 2.1.10.

In addition, I designed a new gapped substrate containing sequences of both the outer (Lo) and inner (Li) repeats, named Gapped Lo-Li. In particular, specific positions (shown in pink in Figure 2-19) in Lo were mutated to their counterparts in Li. Use of this new substrate for complex formation increased the stability and homogeneity of the SB transposase-DNA complexes significantly, forming a single, narrow peak in analytical SEC (Figure 2-19, right). The obtained complexes did not show protein precipitation and could be concentrated up to concentration ranges suitable for crystallization. These results allowed me to reconstitute SB nucleoprotein complexes that are currently being screened for crystallization.

2.3.1.3 Identification of specific SB protein-transposon end DNA contacts

In order to gain structural insights into transposon end DNA recognition, I aimed to identify specific protein residues involved in DNA binding by site-directed protein-DNA disulfide crosslinking. For these experiments, cysteine residues are introduced by site-directed mutagenesis at defined positions in the DNA-binding surface of the protein and an alkanethiol tether is incorporated at specific positions in the DNA backbone. If the cysteine-substituted residues are in close proximity to the modified DNA backbone, they will form disulfide bonds with the alkanethiol tethers, leading to the formation of covalently trapped protein-DNA complexes (Verdine *et al.*, 2003) (Figure 2-20 A).

By *in silico* analysis of a homology model of the SB DBD [modelled based on the Mos1 structure, PDB ID: 3HOT, (Richardson *et al.*, 2009)], I first identified candidate residues both in the PAI and in the RED domains that are likely in contact with the transposon end DNA (Figure 2-20 B). Subsequently, I performed a comparative analysis of transposon end DNA sequences of several *Tc1/mariner* transposons (*Tc3*, *Tc1*, *Mos1*, *Hsmar1*) to predict specific nucleotide positions recognized by the candidate SB residues. From this analysis, I identified eight different amino acids that likely interact with the transposon ends in a single position or two possible positions in the DNA sequence (Figure 2-21 A).
I then expressed and purified eight single point mutants, with each of the previously selected candidate residues substituted to Cys. Then, I assessed the DNA binding properties of the soluble SB mutants (7 out of 8) by analytical SEC (Figure 2-21 A). The three mutants that retained DNA binding (indicated in red in Figure 2-20 and Figure 2-21) to a pre-cleaved left transposon end (LE) have been selected for crosslinking experiments (Figure 2-21).

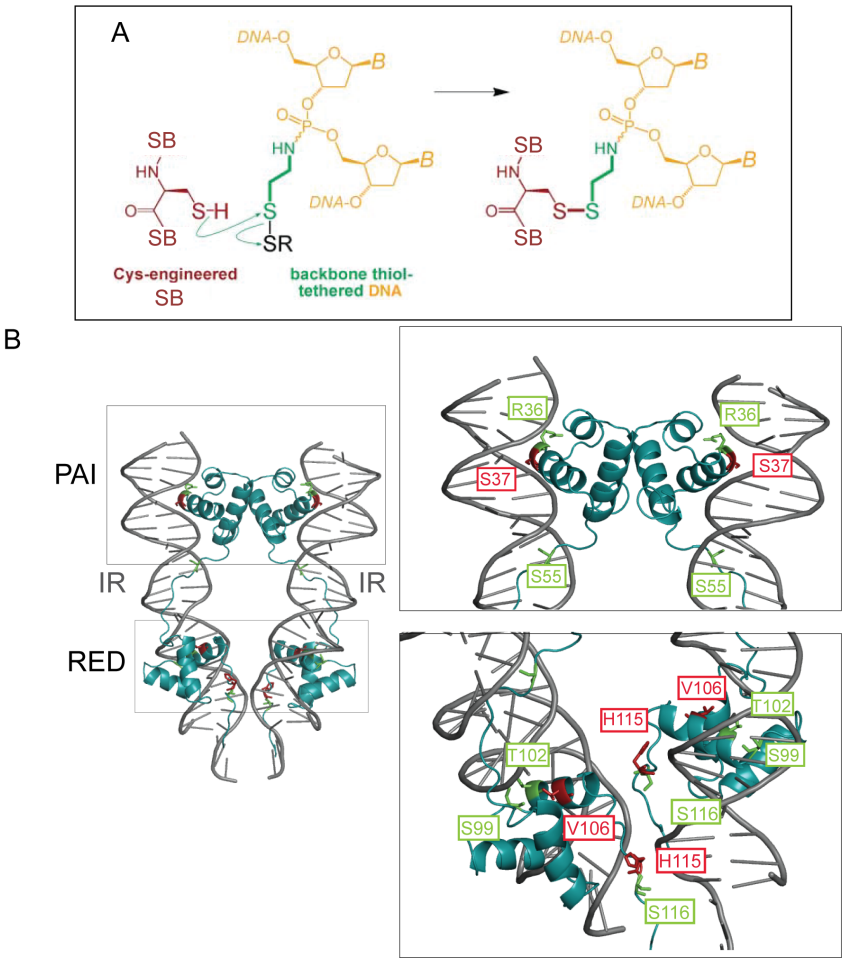


Figure 2-20: Design of protein-DNA crosslinking to map SB-Lo interactions. (A) Schematic representation of the crosslinking strategy. Specific cysteine mutations are introduced in the SB transposase (red) and specific backbone phosphates are substituted with a N-thiolalkyl phosphoramidate moiety (thiol-containing tether in green, DNA backbone in yellow). Upon disulfide crosslinking reaction, the protein gets covalently attached to DNA. Adapted from (Banerjee *et al.*, 2006). (B) Structural model of SB DNA binding domain (DBD) bound to two inverted repeats (IR, grey). The PAI and RED HTH motifs are indicated. The side chains of predicted DNA binding residues are shown in red (for residues selected for further crosslinking experiments) and green.

I performed preliminary crosslinking experiments using the SB_{S37C}, SB_{V106C} and SB_{H115C} mutants in combination with thiol-tethered Lo DNA substrates, (LE3, LE4, LE1 and LE2; containing the thiol modification at different positions in SB's left outer direct repeat) (Figure 2-21 B). The crosslinking protocol [adjusted from the method described in (Gorecka *et al.*, 2013)] is based on incubation of the cysteine-substituted SB with the modified DNA substrate in non-reducing conditions. The products of the reactions were analysed on non-reducing SDS-PAGE gels, stained for protein (or DNA where indicated; Figure 2-21 B). Bands having retarded mobility, corresponding to a molecular weight of approximately 49 kDa, appeared upon DNA addition, indicating crosslink formation

between the mutants (39 kDa) and the modified strand of the DNA substrates (9.9 kDa). The relatively low amount of the observed crosslinked complex in these experiments was probably due to low efficiency of the crosslinking reaction in the tested condition. Unspecific bands were also detected in the reactions and likely indicate adventitious oxidation of the protein in non-reducing conditions.

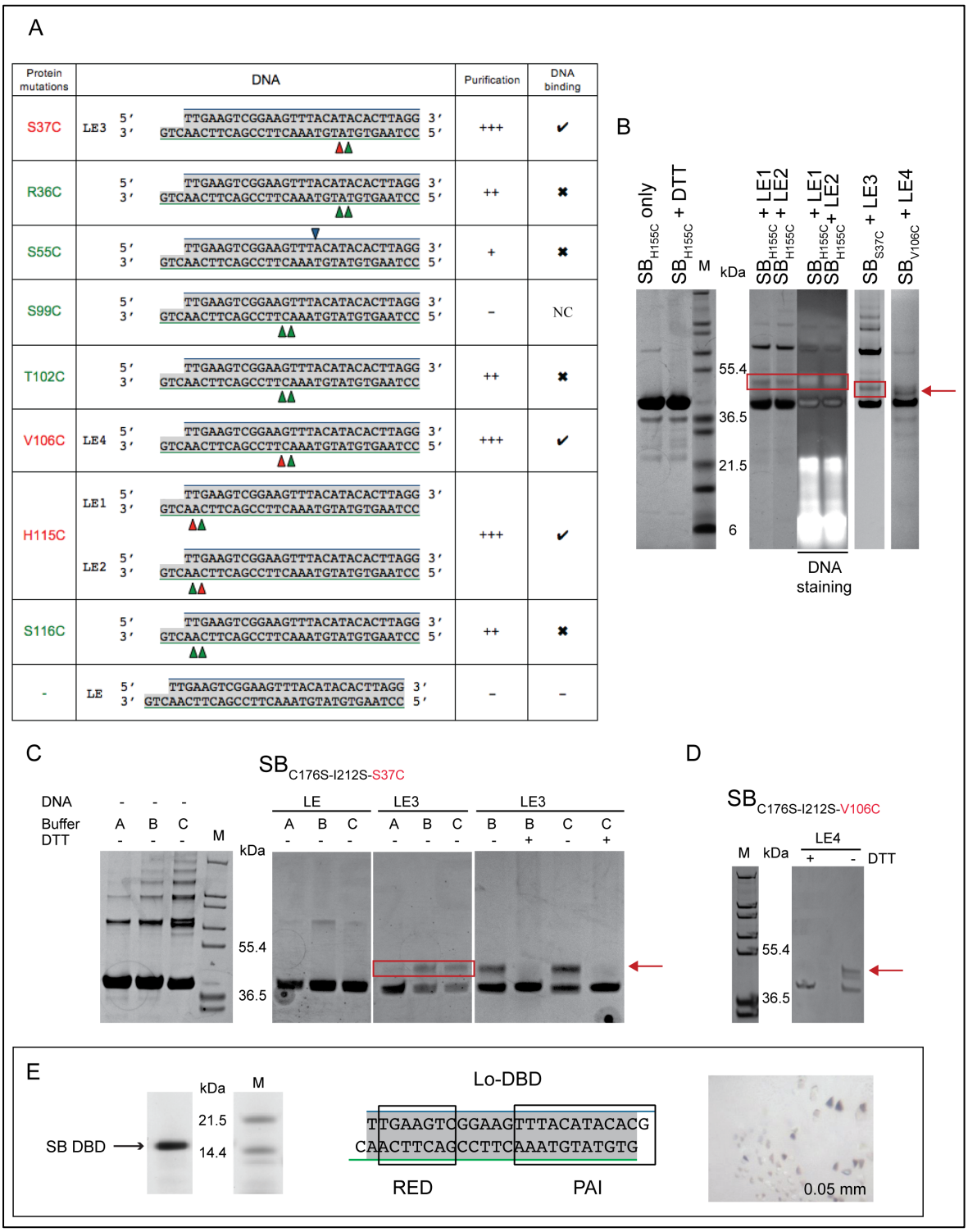


Figure 2-21: Identification of specific SB protein-transposon end DNA contacts. (A) Candidate DNA binding residues (Cys-substituted) and predicted corresponding protein-binding positions on transposon end DNA

(marked with green and red arrowheads; red arrowheads mark the positions of P-cystamine modification in LE1, LE2, LE3 and LE4 DNA substrates used in crosslinking experiments). Mutants shown in red have been used in crosslinking experiments. Results of protein purification ('+': positive; '-': negative) and DNA binding assays ('✔': positive; '✖': negative; 'NC': not characterized) are indicated. (B-D) Site-directed protein-DNA disulfide crosslinking. 4-12% Bis-Tris NuPAGE gels separating crosslinking reactions with the indicated SB transposase variants and the modified DNA substrates. The resulting protein-DNA covalent complexes are framed by a red rectangle or indicated by a red arrow. Gels are stained for protein, unless otherwise indicated. Reaction buffer A, B and C contained no, 0.2 mM and 0.5 mM DTT, respectively. After the crosslinking procedure, samples were resuspended in loading buffer containing 300 mM DTT (DTT +) or no DTT (DTT -). M: marker. (E) Formation and crystallization of SB DNA binding domain-transposon end DNA complexes. Left panel: Left: SDS-PAGE analysis of the purified N-terminal DNA binding domain (aa 1-117) of the SB transposase (SB DBD); M: marker. Middle: dsDNA probe (Lo-DBD) used for complex formation and crystallization. It contains complementary overhangs at its termini. DNA regions recognized by the PAI and RED domains are indicated. Right: crystals formed of SB DBD protein- Lo-DBD DNA complexes.

To overcome these technical issues, I optimized the crosslinking protocol and I introduced the C176S and I212S mutations mentioned in the previous section, to increase solubility and stability of the cysteine-substituted mutants. These conditions enhanced the specificity of the crosslinking reaction and prevented undesired protein-protein oxidation. With the $SB_{C176S\text{-}I212S\text{-}S37C}$ (Figure 2-21 C) and $SB_{C176S\text{-}I212S\text{-}V106C}$ (Figure 2-21 D) variants, specific bands appeared upon addition of modified LE DNA, corresponding to crosslinked nucleoprotein complexes. This indicates that S37 in the PAI and V106 in the RED domain are DNA-binding residues and contact transposon end DNA at the predicted specific positions.

In addition to the crosslinking experiments, in order to gain high-resolution insights into transposon end DNA recognition, I worked on structure determination of SB DBD-transposon end DNA complexes (Figure 2-21 E). For this, I overexpressed and purified the DBD of the SB transposase (residues 1-117) to homogeneity (Figure 2-21 E, left panel) and reconstituted *in vitro* nucleoprotein complexes with different length transposon end DNA substrates (e.g. Lo-DBD in Figure 2-21 E, middle panel). From several crystallization attempts, I obtained crystals of the reconstituted complexes (Figure 2-21 E, right panel) that have been tested for diffraction. However, the diffraction limit remained moderate (5 Å) and the obtained datasets did not allow me to determine the structure by molecular replacement; the presence of protein in the crystals could also not be

unambiguously assed. Crystal optimization attempts resulted in thin clusters of needles, so far unsuitable for data collection.

2.3.2 *SB* transposon excision

The mechanism of excision of the *SB* transposon is still poorly characterized. To date, studies performed in living cells (Luo *et al.*, 1998) have provided no direct evidence on the exact positions of cleavage at the transposon ends or on the minimal protein and DNA requirements of these reactions. To elucidate these aspects, I reconstituted and analysed the specific cleavage activity of the SB transposase on transposon end DNA sequences *in vitro* and mapped the positions of the DNA breaks with single nucleotide precision.

2.3.2.1 Mapping the specific cleavage sites of SB

Previous reports showed that the SB transposase is able to cleave dsDNA substrates imitating isolated Lo sites *in vitro*, but multiple cleavage positions were detected (Voigt *et al.*, 2016). In order to increase specificity of cleavage, I first optimized the design of the Lo dsDNA substrate used for the cleavage assay. I thus incubated this optimized dsDNA substrate, referred to as Lo (Figure 2-22, top), with the purified SB transposase and I analysed the resulting cleavage products via denaturing PAGE (Figure 2-22, left and middle panels). To detect cleavage on a specific strand, the NTS or the TS of the dsDNA substrate was 5' end labelled with ^{32}P (as indicated by asterisks in Figure 2-22).

The results of the assay showed that the Lo substrate is cleaved at specific positions on both DNA strands in the presence of the SB protein (Figure 2-22, left and middle panels). This indicates that the SB transposase is able to specifically cleave the outer repeats in the transposon ends *in vitro* without any additional protein factors or DNA requirements. Remarkably, I observed that the SB transposase cleaves the TS exactly at the transposon end boundary and the NTS at a 2 nt recessed position within the transposon end, generating 2 nt staggered dsDNA breaks (Figure 2-22, top).

2.3.2.2 Dissecting the role of the inner repeats in SB cleavage activity

The outer and inner repeats of the *SB* transposon differ only in few nucleotides but likely have distinct roles in transposition (Cui *et al.*, 2002). The purified SB transposase binds dsDNA probes imitating the left inner repeat, Li (Figure 2-17). However, I found that SB

does not cleave Li sites *in vitro* (Figure 2-22, right panel), consistently with previous *in vivo* reports (Cui *et al.*, 2002). Therefore, recognition of specific nucleotides in the sequence of the outer repeat, not present in the inner repeats, is necessary for transposase-mediated specific cleavage.

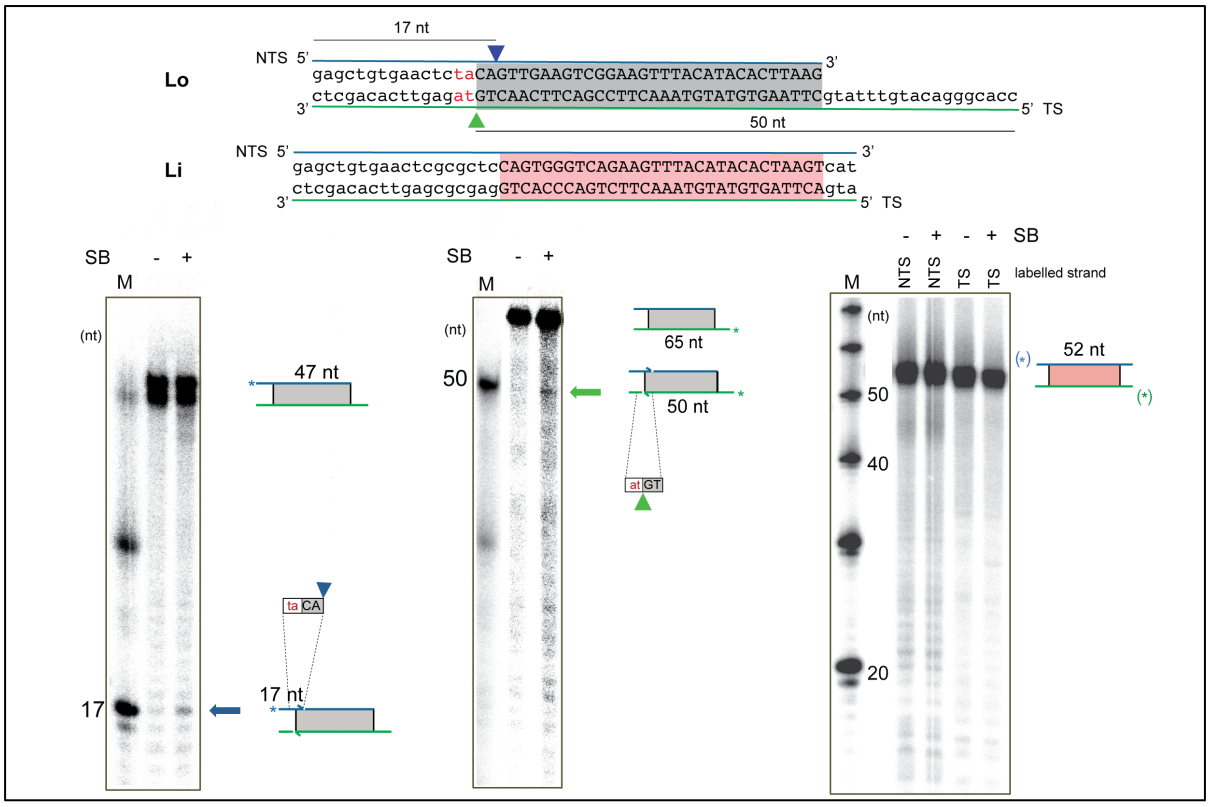


Figure 2-22: Characterization of the specific cleavage activity of SB *in vitro*. Top: dsDNA substrates used in the *SB* cleavage assay. A single left outer (in Lo) or inner (in Li) inverted repeat is shaded grey or pink and its DNA sequence is shown in capital letters; flanking DNA is indicated in lower case letters. The invariable flanking TA dinucleotide is in red. Green and blue arrowheads indicate the position of cleavage on the transferred strand (TS) and non-transferred strand (NTS) of Lo, respectively. Bottom: *SB* cleavage assay. 12% Urea PAGE resolving ^{32}P-labelled DNA substrates (Lo, left and middle panels; Li, right panel, as indicated by cartoons next to the gels) and products after incubation with SB transposase (SB) in a 50:1 protein to DNA molar ratio. The specific cleavage products are indicated by arrows. Marker (M) sizes are indicated next to the gel.

Successively, I tested whether the presence of the inner repeats can affect cleavage on the outer repeat by the SB transposase *in vitro*. Addition of Li substrates in the Lo cleavage reactions (Figure 2-23, left and middle panels) or the use of a Lo-Li substrate, where the outer and inner sites are linked by a flexible single strand DNA linker, (Figure 2-23, right panel) did not affect either the specificity or the efficiency of Lo cleavage by the SB transposase. Therefore, the recognition of the inner repeats by the SB transposase does not seem to facilitate transpososome assembly and thus transposon end cleavage *in vitro*.

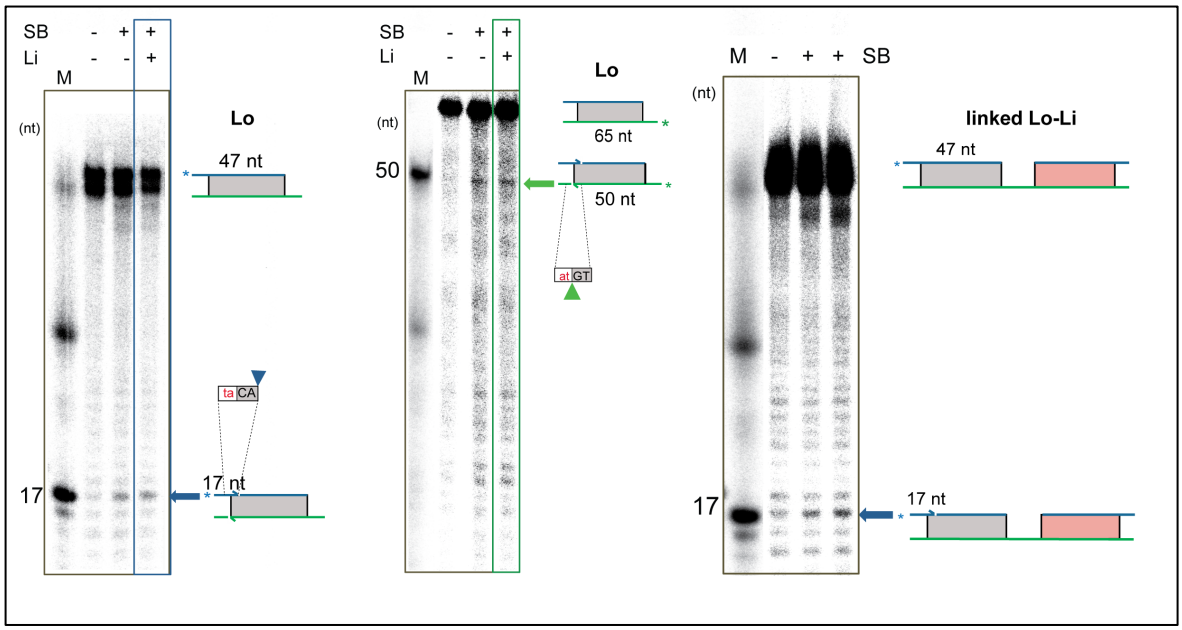


Figure 2-23: Testing the effect of the inner repeats on SB cleavage activity. 12% Urea PAGE resolving ^{32}P-labelled DNA substrates (Lo, with or without unlabelled Li, left and middle panels; linked Lo-Li, right panel, as indicated by cartoons next to the gels) and products after incubation with SB transposase (SB) in a 50:1 protein to DNA molar ratio. The specific cleavage products are indicated by arrows next to the gel. Marker (M) sizes are indicated next to the gel. Schematic code as Figure 2-22.

2.3.2.3 Identifying the role of flanking DNA in SB cleavage activity

The outer repeats of the SB transposon are always flanked by TA dinucleotides sequences derived from the insertion sites (Figure 2-24, 'ta' in red). Transposition assays in HeLa cells indicated that the loss of this TA motif on both flanks of the transposon abolishes transposition and the addition of a TATA motif in the flanking sequences increases the overall transposition rates up to 195% (Cui *et al.*, 2002). This suggests that these flank sequences also affect transposon excision. Therefore, I analysed the effect of this additional TA dinucleotide in *in vitro* cleavage assays.

The results of this assay (Figure 2-24) showed that if a double TA is present in the flanking DNA (as in the Lo-TATA dsDNA substrate, Figure 2-24, top), the SB transposase cleaves the NTS of Lo at two positions: at the canonical cleavage site (marked as 1) and, additionally, 2 nucleotides away from the 3' of the additional TA (marked as 2) (Figure 2-24, bottom). This indicates that the recognition of this flanking TA by the SB transposase might be essential for selecting the position of cleavage on the outer repeat 2 nucleotides 3' of a TA dinucleotide.

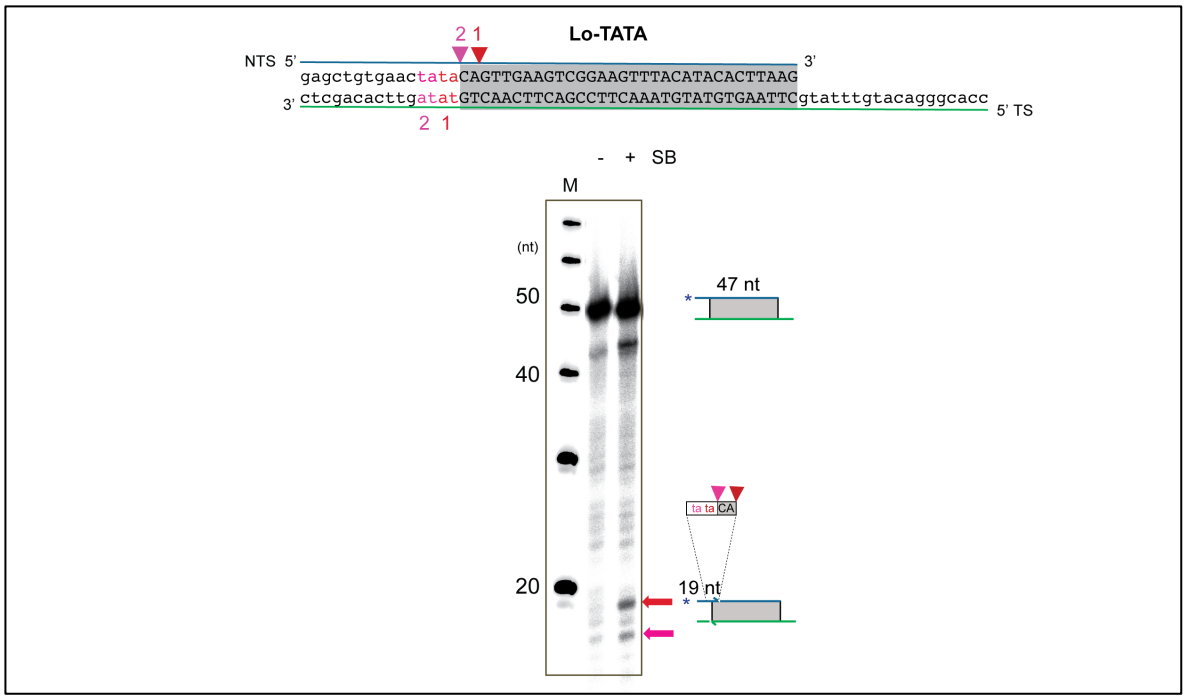


Figure 2-24: Identifying the role of flanking TA dinucleotides in SB cleavage activity. Top: dsDNA substrate, named Lo-TATA used in *SB in vitro* cleavage assays. The invariant TA dinucleotide (number 1) in the flanks is shown in red; the additional TA (number 2) is in magenta. Bottom: 12% Urea PAGE resolving ^{32}P-labelled Lo-TATA DNA substrate and products, as indicated by cartoons next to the gel, after incubation with SB transposase (SB) in a 50:1 protein to DNA molar ratio. The specific cleavage products are indicated by arrows. Marker (M) sizes are indicated next to the gel.

2.3.3 *SB* transposon integration

The biochemical principles and the molecular factors involved in *SB* integration are yet uncharacterized. This is mainly due to technical challenges in reconstituting the integration activity of the SB transposase *in vitro*. To overcome this lack of knowledge, I therefore endeavoured to establish biochemical assays to monitor *SB* integration.

2.3.3.1 *In vitro* reconstitution and analysis of *SB* integration activity

I first focused on identifying permissive conditions in which the recombinant SB transposase can perform integration (i.e. strand transfer reactions) *in vitro* in the presence of transposon end DNA (Figure 2-25).

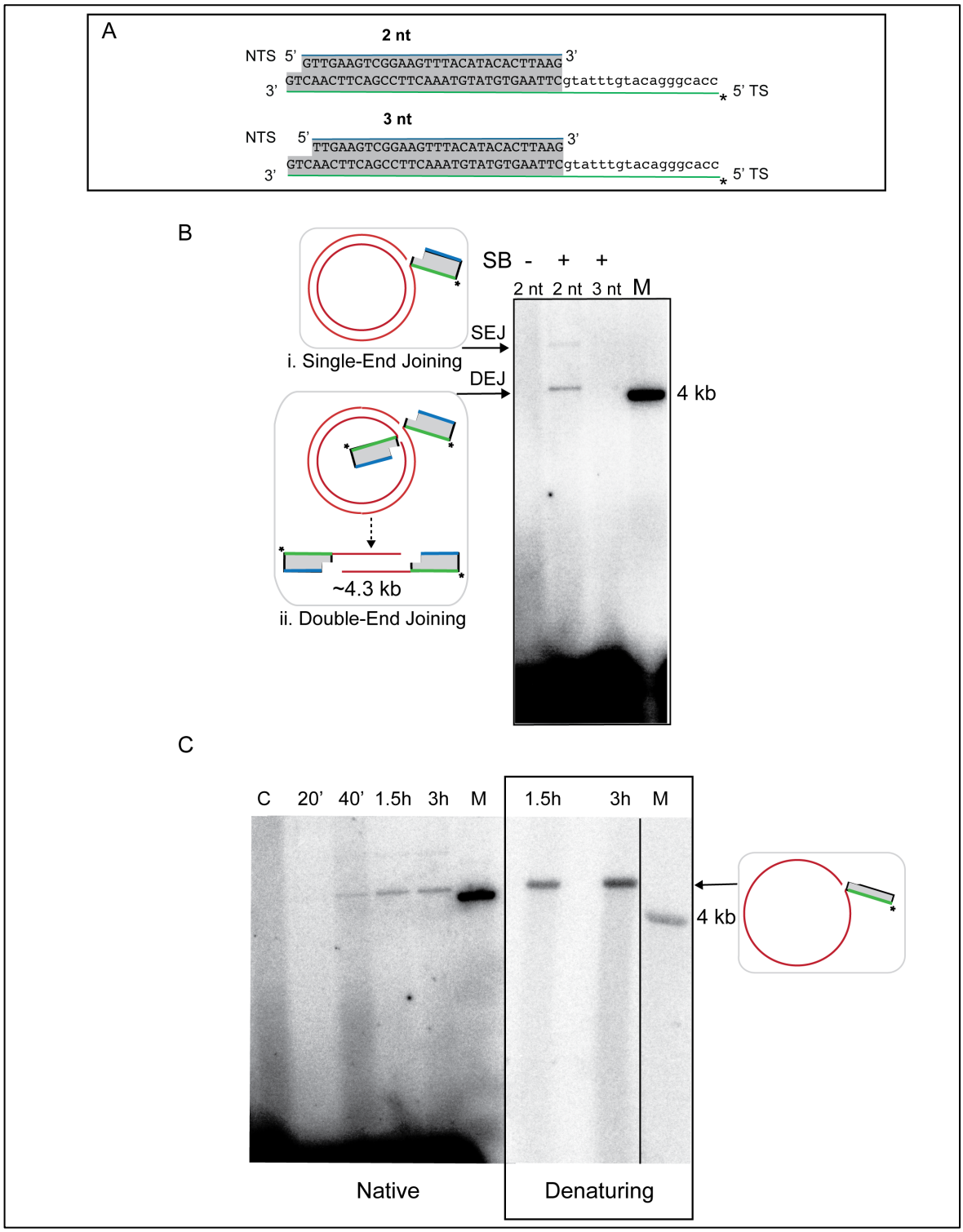


Figure 2-25: Analysis of SB integration activity *in vitro*. (A) dsDNA substrates used in *SB in vitro* integration assay. Schematic code as Figure 2-22. (B) *SB in vitro* integration assays, showing that the SB transposase (SB) joins specifically 2 nt staggered pre-cleaved Lo DNA (2 nt) to target plasmid DNA (each strand shown as a red line). Strand transfer reactions generate two different products, as displayed on a native agarose gel: i. Single-End Joining (SEJ) products and ii. Double-End Joining (DEJ) products. (C) Time-course *SB in vitro* integration assay, showing the kinetics of SEJ and DEJ product formation. Incubation times are shown above the corresponding lines. Control samples (C) did not contain any SB protein.

Reactions are run on a native or on a denaturing gel, as indicated above each gel. In (A), (B) and (C), asterisks indicate the positions of ^{32}P-labelling. M: marker.

For this, I incubated the SB transposase with target plasmid DNA and dsDNA substrates imitating *SB* Lo sites in different ratios and reaction conditions. Out of these attempts, I identified specific experimental conditions in which a pre-cleaved Lo dsDNA substrate (identified in my previous *in vitro* cleavage assay in section 2.3.2.1; referred as "2 nt" in Figure 2-25) is ligated to the target plasmid DNA by SB. Of note, the use of an intact, "uncleaved" dsDNA substrate did not support integration (data not shown). For the *in vitro* integration assays, the 5' end of "2 nt" was ^{32}P-labelled (specifically on the TS, as indicated by asterisks in Figure 2-25). Therefore, ligation of two "2 nt" dsDNA substrates to separate strands of the same circular plasmid (Double-End Joining-DEJ) generated a radiolabelled linear dsDNA product, which could be detected on a native agarose gel (Figure 2-25 B, see schematics).

SB is able to generate such DEJ products *in vitro*, which accumulate rapidly over time and reach saturation within 3 hours (Figure 2-25 C, left native gel). "Single-End Joining" (SEJ) is less frequent and generates a nicked circular plasmid, in which the strand attached by a single "2nt" becomes radioactively labelled (Figure 2-25 B, see schematics next to the gel). Consistently, when I visualized the same reactions on a denaturing agarose gel, DEJ and SEJ products could not be differentiated from each other because only one band was present, corresponding to the radiolabelled strand of "2 nt" joined to a single target strand (Figure 2-25 C, right panel, denaturing gel).

2.3.3.2 Confirmation of specific requirements for *SB* integration activity

Next, I tested whether the 2 nt staggered overhangs of the pre-cleaved transposon ends are required for the integration step. For this, I performed the previously described integration assay using a Lo substrate with 3 nt overhangs at the 3' end of its TS (named "3 nt" in Figure 2-25 A and B). This substrate was not able to support either DEJ or SEJ formation. This indicates that the 2 nt overhangs, introduced at the outer repeats during transposon excision, are specifically required for the successive step of transposon integration.

In conclusion, in this biochemical analysis of *SB* transposition, I showed that the recombinant SB transposase is completely proficient in transposon end binding, cleavage, and integration *in vitro*. Moreover, I identified specific molecular features of the process that were previously unrevealed for the *SB* transposon and will be discussed in details in section 2.5.6.

2.4 Results - Role of transposase oligomerization in *SB* transposition

In the previous section, the main steps of *SB* transposition have been reconstituted *in vitro*, but how these events are orchestrated in the context of the SB transpososome has remained uncharacterized. The recombinant SB transposase is prevalently monomeric on its own (Franka Voigt, manuscript in preparation; see also section 2.3.1.1). However, the active forms of DNA transposases are protein-DNA assemblies that are at least dimeric, since two subunits of the transposase are always required to act on two transposon ends (Hickman *et al.*, 2016). For this reason, in this part of my thesis, I investigated oligomerization of the SB transposase and its regulatory role in transposition.

2.4.1 *In vitro* analysis of SB transposase oligomerization

The first question I asked is whether SB monomers can form oligomers and what conditions influence oligomer assembly. Thus, I analysed SB oligomerization by employing different oligomerization assays, based on fluorescence (section 2.4.1.1) as well as on crosslinking techniques (section 2.4.1.2).

2.4.1.1 *In vitro* fluorescence-based oligomerization assay

Prior to this thesis work, to study SB oligomerization in living cells, an assay has been designed that relies on red fluorescent protein (RFP) derived biosensors, named A1 and B1 (Alford *et al.*, 2012) (Franka Voigt, manuscript in preparation). A1 and B1 are dimerization-dependent fluorescent proteins that interact with very low affinity (K_d = 33 μM) and thus exist mainly in their dissociated state at the concentrations used in the experiments. Due to the low affinity of their interaction, they only light up when they are brought into close proximity upon association of proteins that are genetically fused to them

(Figure 2-26 A). These properties make them well suited to study the potentially low-affinity or transient interactions between SB monomers.

For establishing *in vivo* oligomerization assay, fusion proteins were generated between the SB transposase and the A1 and B1 proteins. The resulting constructs (A1-SB and B1-SB) were then overexpressed in HeLa cells. Notably, a strong fluorescent signal could be detected in the nuclei of cells expressing both A1-SB and B1-SB fusion proteins, indicating that the SB transposase is capable of oligomerization *in vivo* (Franka Voigt and Cecilia Zuliani-Lab Manager, Barabas Laboratory, EMBL Heidelberg, manuscript in preparation).

In order to corroborate these findings and identify permissive conditions, I then probed and analysed oligomerization of the SB transposase *in vitro,* together with Cecilia Zuliani. For this, we established an *in vitro* oligomerization fluorescence-based assay, conceptually similar to the one used *in vivo* (Figure 2-26 A). We first overexpressed in *E. coli* and purified recombinant SB fusion proteins, A1-SB and B1-SB, as well as A1 and B1 proteins, used as negative controls. We then mixed the purified proteins in pairs (A1 with B1; A1-SB with B1-SB) in presence (+ DNA) or absence of transposon end DNA (Cleaved Lo, see Figure 2-19, section 2.3.1.2). Finally, the fluorescence signal emitted from the mixtures was measured in a plate reader at different time points (Figure 2-26 B, left).

Consistently with previous reports (Alford *et al.*, 2012), the A1 and B1 pair exhibited close-to-background fluorescence, even 72 hours after protein mixing. Interestingly, this was also true for the A1-SB/B1-SB pair. In fact no significant increase of fluorescence signal was detected when A1-SB and B1-SB fusion proteins were mixed together, indicating that SB oligomerization is quite weak in these conditions. However, the A1-SB/B1-SB pair exhibited significantly high fluorescence signal upon addition of transposon end DNA (+ DNA), as compared to the condition without DNA. As expected, fluorescence of the A1/B1 controls was not affected by the presence of DNA. A strong fluorescence signal of the A1-SB/B1-SB pair, together with transposon end DNA, was already detectable after 5 minutes of incubation and increased at 24 hours of incubation (Figure 2-26 B, left).

These findings indicate that the SB transposase does form oligomers *in vitro* and that its oligomerization depends on the presence of transposon end DNA. Moreover, SB oligomers are progressively assembled in time, but, once formed, result quite stable in the conditions tested.

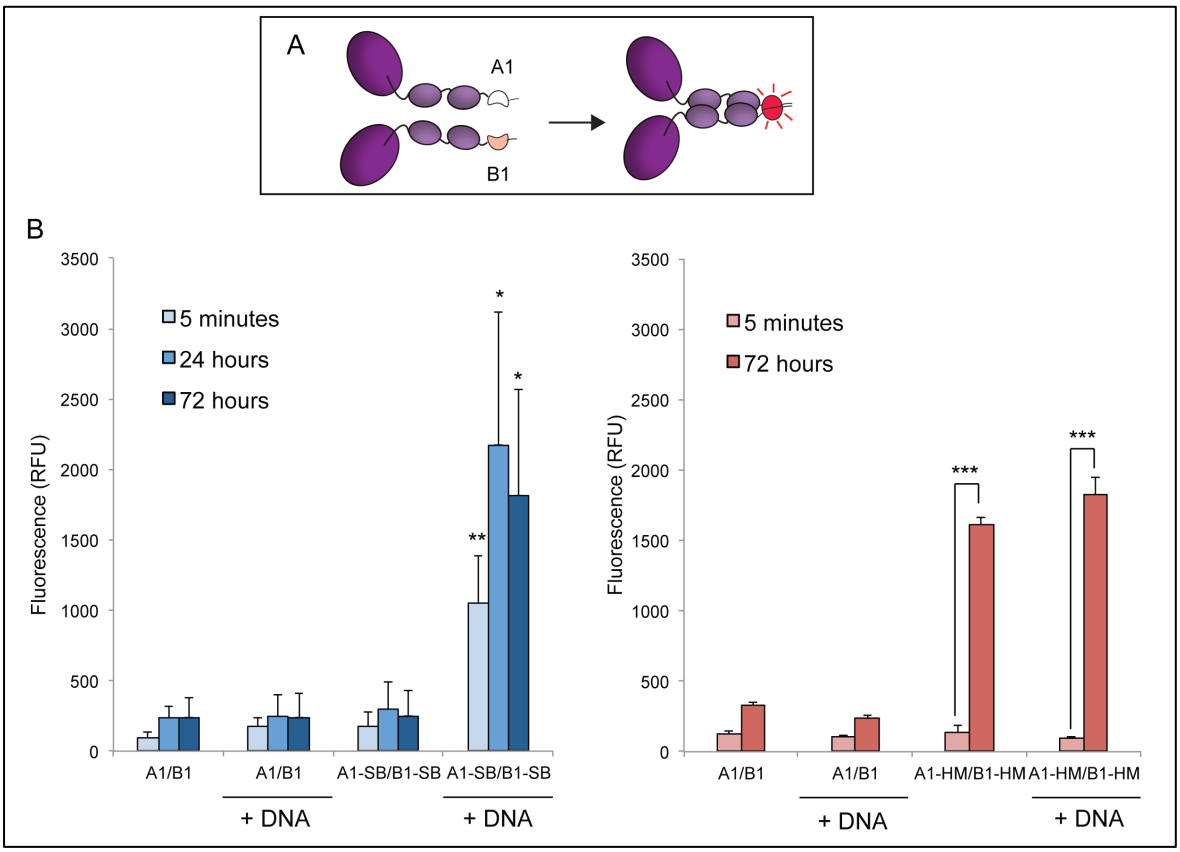


Figure 2-26: *In vitro* fluorescence-based oligomerization assay. (A) Schematic representation of the assay [adapted from (Alford *et al.*, 2012)]. A1 and B1 proteins come in close proximity and emit fluorescence only upon dimerization of the transposase (SB or Hsmar1), genetically fused to them. The transposase is depicted with its bipartite DNA binding domain (small light purple ovals) and its catalytic domain (big dark purple ovals). (B) Left: *in vitro* SB oligomerization-dependent fluorescence measurements of fusion proteins in absence and presence (+ DNA) of *SB* transposon end DNA. For each condition, mean values of three experiments are shown and error bars represent the standard deviation. Statistical analysis was performed by a two-sample *t*-test comparing fluorescence of A1-SB/B1-SB mixtures in presence and in absence of DNA. A single asterisk indicates *P* values ≤0.05; two asterisks indicate *P* values ≤0.01. Right: *in vitro* Hsmar1 oligomerization-dependent fluorescence measurements of fusion proteins in absence and presence (+ DNA) of *Hsmar1* transposon end DNA. For each condition, mean values of three experiments are shown and error bars represent the standard deviation. Statistical analysis was performed by a two-sample *t*-test comparing fluorescence of A1-HM/B1-HM mixtures at two different time points. Three asterisks indicate *P* values ≤0.001.

As a positive control, we then performed similar assays using recombinant A1 and B1 proteins fused to the Hsmar1 transposase, which is a stable dimer by itself (Figure 2-26 B, right). We did not detect any significant increase of fluorescence signal 5 minutes after mixing the A1-Hsmar1 and B1-Hsmar1 proteins in comparison to the A1/B1 controls. However, the A1-Hsmar1/B1-Hsmar1 pair emitted very strong fluorescence 72 hours after mixing, and the fluorescence of the protein pair was not significantly affected by the presence of *Hsmar1* transposon end DNA (Figure 2-26 B, right). The results of this experiment can be interpreted as follows. The A1-Hsmar1 and B1-Hsmar1 transposases are purified as stable homodimers, as the wild type Hsmar1. Thus, no heterodimer formation between A1-Hsmar1 and B1-Hsmar1 monomers, and consequently fluorescence emission, likely occurs immediately after mixing (5 minutes). Consistently with this hypothesis, long incubation times (72 hours) would allow exchange of Hsmar1 monomers in solution, resulting in heterodimerization and consequent detection of a strong fluorescence signal.

Taken together, the data reported in this section indicate that the SB transposase, differently to the Hsmar1 transposase, exists both in a monomeric and oligomeric form. Binding to the transposon end DNA is likely the switch regulating the transition between these states throughout the transposition pathway (as discussed in section 2.5).

2.4.1.2 Disuccinimidyl suberate (DSS)-mediated crosslinking of SB

In order to further confirm oligomerization of the SB transposase *in vitro*, I tested the formation of covalent bonds between subunits of the purified SB transposase using chemical crosslinking. For this, I incubated the SB protein with or without transposon end DNA and then treated the samples with disuccinimidyl suberate (DSS), a chemical crosslinker for primary amines present in the side chains of lysine residues and the N-terminus of the proteins (Wong, 1993). By resolving the reactions by SDS-PAGE, I could detect DSS-mediated crosslinking of two SB subunits both in the presence and in the absence of transposon end DNA (Figure 2-27 A). Moreover, higher molecular weight species, likely corresponding to tetramers, could also be detected in the sample containing DNA (Figure 2-27 A). This shows that oligomerization of SB occurs *in vitro* under these experimental conditions and it is compatible with complex formation.

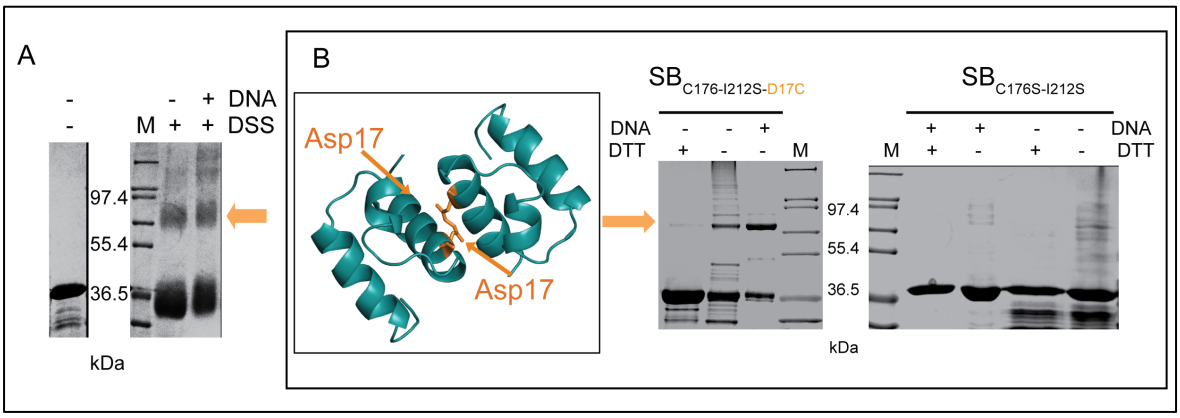


Figure 2-27: *In vitro* analysis of SB transposase oligomerization by crosslinking. (A) Chemical crosslinking of the SB transposase by disuccinimidyl suberate (DSS). SDS-PAGE analysis of the SB transposase treated with DSS in absence or presence of transposon end DNA (DNA). Orange arrow indicates a band with retarded mobility, corresponding to the molecular weight of a SB dimer. (B) Site-directed cysteine-based crosslinking of the SB transposase. Left: structural model of the dimer interface in the SB PAI, showing residue Asp17 in sticks. Right: SDS-PAGE analysis showing covalent SB dimers (indicated by the yellow arrow) obtained by disulfide protein-protein crosslinking of the $SB_{C176S-I212S-D17C}$ mutant with or without transposon end DNA (DNA). SDS-PAGE analysis of the $SB_{C176S-I212S}$ protein in the same conditions is shown for comparison.

2.4.1.3 Specific residues in the PAI mediates SB oligomerization

In order to dissect specific amino acids involved in SB oligomerization, I identified candidate residues of the SB transposase and tested their location on the transposase dimer interface by site-specific cysteine-based crosslinking.

The work done on the Hsmar1 transposase during my M.Sc. thesis let me identified several residues in the HTH1 motif that are involved in its dimerization. I therefore analysed an *in silico* model of the SB PAI [generated based on the Mos1 structure, PDB ID: 3HOT, (Richardson *et al.*, 2009)] (Figure 2-27 B, left), and identified candidate amino acids in this specific region of the SB transposase that could be involved in its dimerization, likely forming salt bridges between the two protein protomers (Asp 10, Arg 14, Asp 17, Lys 20 and Arg 31). I then mutated these candidate residues one by one into cysteine in the SB protein sequence. Additionally, these cysteine-substituted mutants contained the mutations C176S and I212S, described in sections 2.3.1.2 and 2.3.1.3, to increase transposase stability and decrease adventitious oxidation. Successively, I overexpressed in *E. coli* and purified two of these cysteine-substituted mutants ($SB_{C176S-I212S-D17C}$ and $SB_{C176S-I212S-R14C}$) as recombinant proteins and I incubated them with transposon end DNA in reducing conditions (0.2 mM tris(2-carboxyethyl)phosphine-TCEP). Once stable nucleoprotein

complexes have been formed, I removed completely the reducing agent by dialysis and incubated the samples for two days at 25° C to favour the formation of disulfide bonds between cysteines.

I then analysed the reactions on non-reducing SDS-PAGE gels. A band having retarded mobility corresponding to a molecular weight of approximately 80 kDa appeared after the described treatment of the $SB_{C176S-I212S-D17C}$ both as protein alone and in complex with DNA (Figure 2-27 B) (less prominent for the $SB_{C176S-I212S-R14C}$ mutant; data not shown). This is indicative of disulfide bond formation between two SB protomers and is specific to the incorporation of the cysteine in position 17 in the PAI of the SB transposase (compare with $SB_{C176S-I212S}$ transposase under same conditions; Figure 2-27 B). The formation of a disulfide bond between specific cysteines belonging to different SB subunits is indicative of the high proximity of the PAIs of the two protomers (the S-S bond length is about 2.05 Å). This suggests that the specific residue Asp 17 in the PAI of the SB protein is involved in protein dimerization (Figure 2-27, B).

2.4.2 Oligomerization-based strategies for *in vitro* reconstitution of SB protein-DNA complexes

Experimental data presented in the previous sections indicate that the recombinant SB transposase is predominantly monomeric by itself, but it transiently forms oligomers upon binding to transposon end DNA. Lack of stable oligomerization might greatly impair formation of functional SB complexes *in vitro* and, consequently, crystallization. To overcome this, I explored different strategies to stabilize SB oligomerization for functional complex formation, towards crystallization and structure determination. Higher symmetry of oligomeric complexes might also increase *per se* the chances of crystallization.

I therefore aimed to generate stable dimeric SB derivatives. For this, I first substituted the PAI of the SB transposase with the HTH1 motif of the homologous Tc3 and Mos1 transposases, creating transposase chimeras. Differently to SB, Tc3 [PDB ID: 1U78; (Watkins *et al.*, 2004)] and Mos1 [PDB ID: 3HOT; (Richardson *et al.*, 2009)] form in fact fairly stable dimers by their HTH1 motifs. After successful protein overexpression in *E. coli*, the Tc3-based chimeric constructs showed very low solubility, whereas one constructs of the Mos1-SB transposase (comprising residues 1-56 of Mos1 and residues 53-340 of

SB) could be purified at high protein yields and eluted as a dimer in SEC experiments (Figure 2-28 A).

However, I could not identify any transposon end DNA substrate that would support correct complex formation (tested by SEC) or activity (probed by *in vitro* cleavage and integration assays, described in sections 2.3.2 and 2.3.3) by the obtained dimeric Mos1-SB chimera. To note, the design of such DNA substrates (represented in Figure 2-28 A, bottom) is particularly challenging because they should contain mostly *SB* transposon ends sequences, apart from the predicted SB PAI-binding site, which has to be substituted with the HTH1-binding site of Mos1. However, no structural information is available on the exact contacts established by the SB transposase with transposon end DNA.

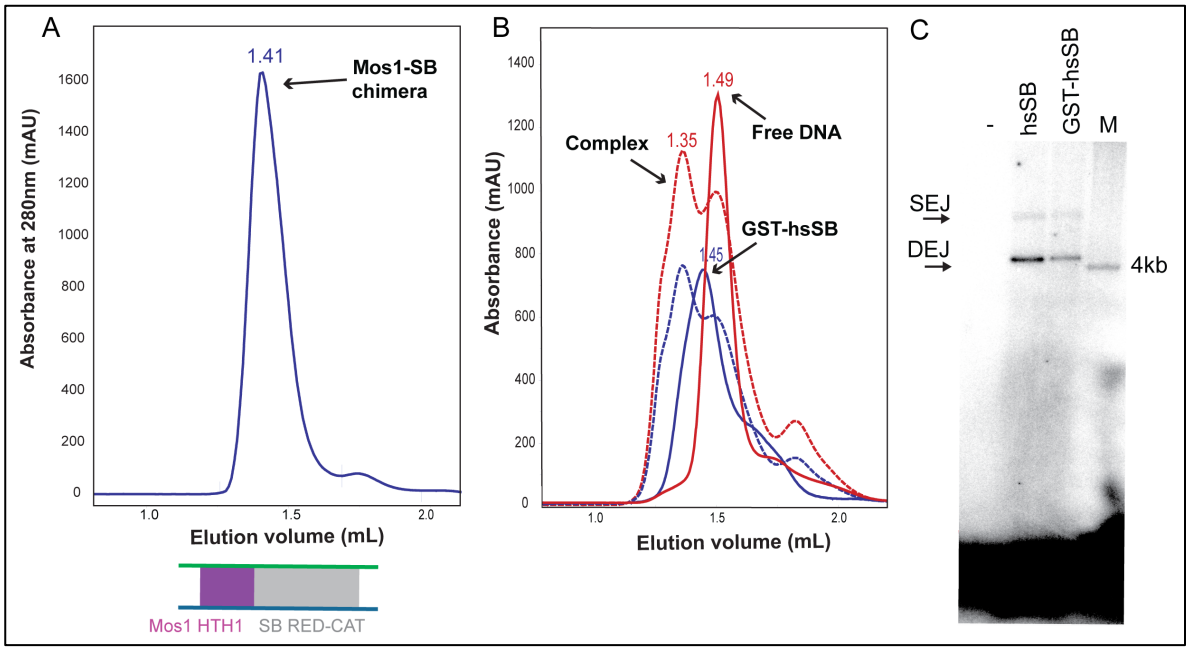


Figure 2-28: Dimeric derivates of the SB transposase. (A) Analytical SEC (column Superdex 200 3.2/3.0) chromatogram showing that the Mos1-SB chimera elutes as a dimer. Below the chromatogram, a chimeric dsDNA substrate is shown containing the Mos1 HTH1-binding site (purple) and the SB RED and catalytic domain-binding sites (grey). (B) Formation of GST-hsSB transposase-DNA complexes. Analytical SEC (column Superdex 200 3.2/3.0) chromatogram showing that the GST-hsSB transposase binds to Gapped Lo-Li dsDNA substrates. Elution volumes of complex peaks (dashed red and dashed blue lines, absorbance at 260 nm and 280 nm respectively) shift with respect to individual complex components (GST-hsSB protein, blue line, absorbance at 280 nn; Free DNA, red line, absorbance at 260 nm). GST-hsSB protein elutes as a dimer and mantains its dimeric state upon binding to DNA. (C) *In vitro* integration assays, showing that the GST-hsSB transposase (GST-hsSB), similarly to the hsSB transposase (hsSB), joins pre-cleaved Lo DNA to target plasmid DNA generating SEJ (Single-End Joining) and DEJ (Double-End Joining) products, as displayed on a native agarose gel. M, marker.

I therefore employed another strategy to form dimeric SB derivatives, based on the use of the glutathione transferase (GST) tag. GST is commonly used as an affinity tag for protein purification, and is dimeric by itself. Moreover, as previously mentioned, I could identify a new SB transposase variant ($SB_{C176S-I212S}$ or hsSB, extensively described in section 3.2), which exhibits higher solubility and stability than the SB transposase. Therefore, I expressed and purified the hsSB transposase as a GST-tagged fusion protein and assessed its oligomeric state by SEC. The GST-tagged hsSB transposase (GST-hsSB) elutes as a stable, homogeneous dimer from the analytical SEC column (Figure 2-28 B, blue curve). Upon addition of the Gapped Lo-Li dsDNA substrate (see substrate sequence in Figure 2-19) to the protein, I detected a shift in the elution volume of the mixture in SEC experiments, indicating formation of a dimeric nucleoprotein complex (Figure 2-28 B). Functional characterization of GST-hsSB showed that this derivative is active both in cleavage and integration assays *in vitro*, similarly to the untagged protein (integration assays shown in Figure 2-28 C). In summary, I was able to generate a functional, stable GST-tagged-SB dimeric variant and I used this to form dimeric nucleoprotein complexes *in vitro* that are currently being screened for crystallization.

2.4.3 Interfering with SB transposase oligomerization affects transposition efficiency in living cells

The data presented in this thesis indicate that the SB transposase is predominantly a stable monomer, which binds transposon end DNA and sequentially undergoes oligomerization. SB would therefore behave differently from mariner transposases (as Hsmar1 and Mos1), which form fairly stable dimer by themselves. Thus, in this part of my work, I asked which is the functional role of SB's unusual oligomerization in transposition. Experiments described in this section were performed together with Cecilia Zuliani (Lab Manager, Barabas lab, EMBL Heidelberg).

To answer the above question, we tested the effects of inducing stable dimerization of the SB transposase on transposition efficiency in HeLa cells (Figure 2-29). To manipulate SB oligomerization *in vivo*, we took advantage of the chemically-inducible dimerization system constituted by the FRB and FKBP12 (FKBP Rapamycin Binding domain and FK506 Binding Protein 12) proteins, which form stable heterodimers solely upon addition of the chemical dimerizer rapamycin (Figure 2-29) (Putyrski *et al.*, 2012). We expected

that monomers of the SB transposase genetically fused to FRB and FKBP12, respectively, (FRB-SB and FKBP-SB fusion proteins) would be brought together in the presence of rapamycin, thus resulting in very stable SB heterodimers and leading to alterations in SB transposition efficiency.

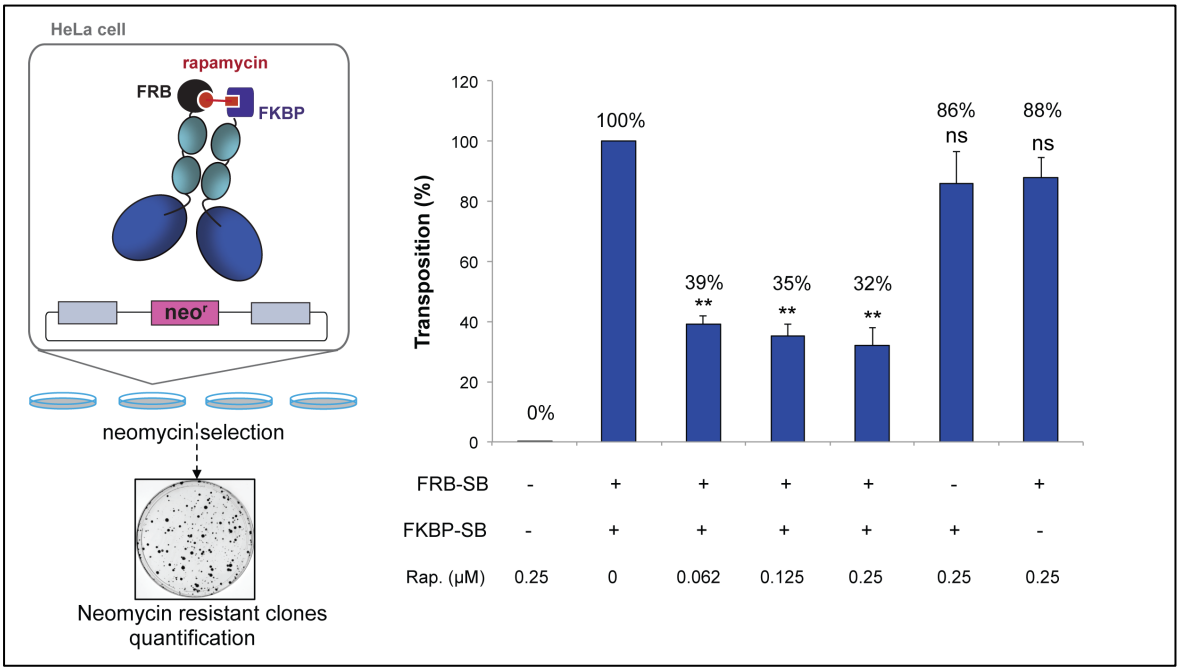


Figure 2-29: Rapamycin-based *SB* transposition assay in HeLa cells. Left: schematic representation of the assay. Donor plasmid is shown with neomycin resistance cassette (neo^r, magenta rectangle), flanked by *SB* transposon ends (grey rectangles). The SB transposase is depicted as a fusion protein with its bipartite DNA binding domain and its catalytic domain. The FRB-SB and FKBP-SB fusion proteins are expressed from plasmids. Right: transposition activity of FRB-SB and FKBP-SB proteins in HeLa cells, showing that stable rapamycin-induced oligomerization of the SB transposase negatively affects transposition rates. Mean values of three independent experiments are shown and error bars represent the standard error of the mean. Statistical analysis was performed by a one-sample *t*-test. Two asterisks indicate *P* values ≤0.01; ns - not statistically significant difference from 100% (*P* values >0.05). Rapamycin (Rap.).

The FRB-SB and FKBP-SB proteins were co-expressed from plasmids in HeLa cells treated with increasing concentration of rapamycin and their activity was assessed using *in vivo* transposition assay [based on assays established by (Ivics *et al.*, 1997)]. The assay monitors the transposition of a neomycin resistance gene (neo^r) flanked by the *SB* transposon ends from a donor plasmid. Upon transposition, the neomycin cassette undergoes genomic integration and provides long-lasting drug resistance to the cells. Following neomycin selection, transposition efficiency is calculated from the numbers of neomycin-resistant colonies in the presence versus absence of the transposase (Figure 2-29). As expected, no significant alteration of transposition activity was detected in HeLa

cells expressing only the FRB-SB or the FKBP-SB protein in the presence of rapamycin, showing that rapamycin *per se* does not influence transposition rates. On the contrary, transposition efficiency was drastically reduced in the cells co-expressing the FRB-SB and FKBP-SB proteins together in the presence of rapamycin (Figure 2-29). This indicates that rapamycin-mediated induction of stable SB oligomers negatively affects *SB* transposition.

We next performed similar assays using the FRB and FKBP12 proteins fused to the Hsmar1 transposase (Figure 2-30). We did not detect any significant alteration of the transposition activity of the FRB-Hsmar1 and FKBP-Hsmar1 proteins, expressed alone or together in HeLa cells, upon rapamycin treatment (Figure 2-30, right). This is consistent with the Hsmar1 transposase being a preformed stable functional dimer, whose activity is not perturbed by rapamycin-induced dimerization.

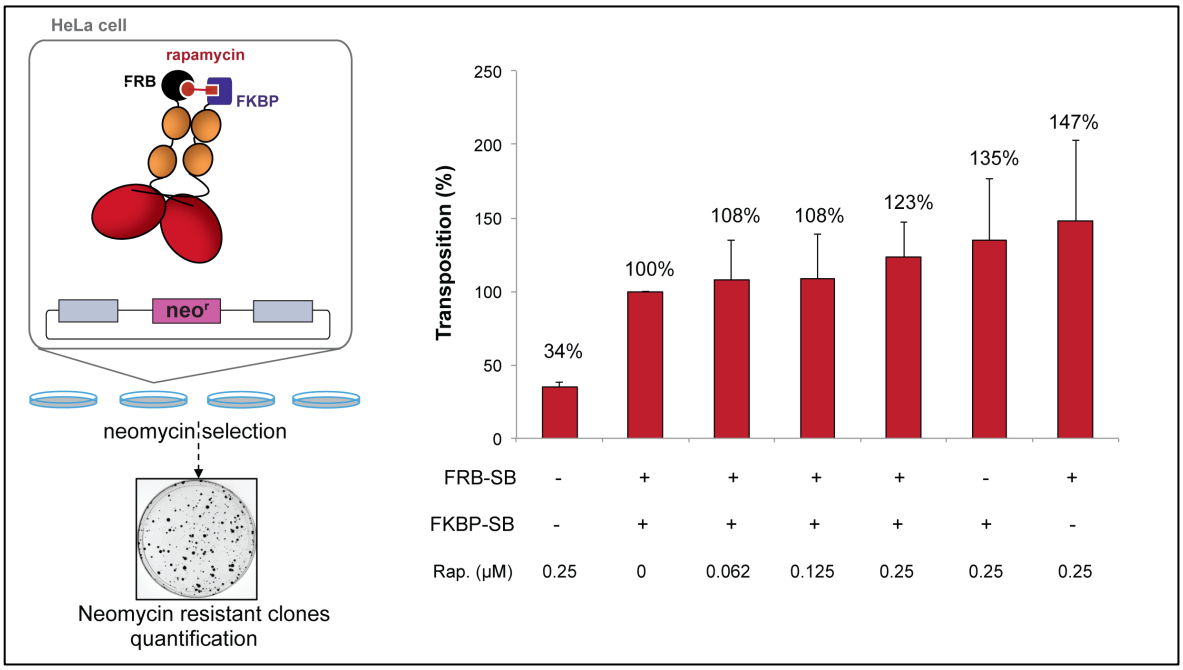


Figure 2-30: Rapamycin-based *Hsmar1* transposition assay in HeLa cells. Left: schematic representation of the assay. Donor plasmid is shown with neomycin resistance cassette (neo^r, magenta rectangle), flanked by *Hsmar1* transposon ends (grey rectangles). The Hsmar1 transposase is depicted as a fusion protein with its bipartite DNA binding domain and its catalytic domain. The FRB-Hsmar1 and FKBP-Hsmar1 fusion proteins are expressed from plasmids. Right: transposition activity of FRB-Hsmar1 and FKBP-Hsmar1 proteins in HeLa cells, showing that rapamycin treatment does not affect transposition rates. Mean values of three independent experiments are shown and error bars represent the standard error of the mean. Statistical analysis was performed by a one-sample *t*-test. No samples (apart from the negative control containing no proteins) show statistically significant difference from 100% (*P* values >0.05). Rapamycin (Rap.).

Taken together, these findings indicate that the distinctive oligomerization of the SB transposase plays an essential role in correct transpososome assembly and remarkably regulates the efficiency of *SB* transposition in the cellular context. The functional consequences of these insights are discussed in details in sections 2.5.7 and 2.5.9.

2.5 Discussion – *Tc1/mariner* transposition

Among DNA transposons, *Tc1/mariners* are unique because they ubiquitously invaded eukaryotic genomes during evolutionary time, they move through a distinctive and yet incompletely characterized mechanism, and have been resurrected for genome manipulation in research and medicine (Hudecek *et al.*, 2017; Plasterk, 1996). Comprehensive structural and biochemical understanding of these transposons will provide invaluable insights into their transposition mechanisms and promote their further development in biotechnological and therapeutic applications.

In this part of my work, I have investigated the molecular mechanism and regulation of two prominent transposons of the *Tc1/mariner* superfamily: *Hsmar1* (Miskey *et al.*, 2007), a representative transposon of the *mariner* clade (in section 2.2), and the *Tc1-like* element *SB* (Ivics *et al.*, 1997), a popular genetic tool applied in research and in clinical settings (in sections 2.3 and 2.4).

2.5.1 Summary of experimental findings

With support from colleagues in the Barabas Laboratory, I have employed a multidisciplinary approach, including bioinformatics, structural biology, biochemistry, and cell biology to investigate the mechanism of *Hsmar1* and *SB* transposition.

i. *Hsmar1* transposition. Structural analysis of *Hsmar1* transposition revealed that the Hsmar1 transposase exists as a compact, autoinhibited dimer in its apo state. This dimer conformation differs from the PEC assembly and is not compatible with DNA binding and processing. Consistently, mutagenesis of the WVPHEL motif, which holds the catalytically incompetent state, results in an elongated conformation (as shown by SAXS and SEC analysis of the W118P mutant) and hyperactivity of the resulting transposases [as reported by (D. Liu *et al.*, 2014)]. Based on these findings, we hypothesize that the WVPHEL motif

acts as a structural "zipper", which prevents DNA catalysis in the transposase active site prior to transposon end synapsis in the PEC.

By biochemical analysis of *Hsmar1* excision, I have established that a specific order exists in the DNA hydrolysis reactions at the transposon ends, where NTS nicking strictly precedes TS cleavage. This supports the hypothesis that conformational changes are required in the transpososome between these two cleavage events. To gain high-resolution insights into these structural rearrangements, I have reconstituted *in vitro* and crystallized Hsmar1 transposase-transposon end DNA complexes in pre-excision conformations. Following successful X-ray data collection, structural determination is currently in progress. High-resolution structures of Hsmar1 pre-excision complexes, and their comparison with available post-excision structures (Richardson *et al.*, 2009), will reveal unprecedented insights into the mechanism of *Tc1/mariner* transposon excision.

ii. *SB* transposition. I have reconstituted *in vitro* the entire set of *SB* transposition events, from transposon end binding to integration, and this allowed to reveal unique molecular features of the process. For transposon recognition, the SB transposase binds with higher affinity to the inner binding sites than the outer sites at the ends. Then it performs two DNA hydrolysis reactions exclusively on the outer sites to liberate the transposon from its donor locus, generating 2 nt staggered transposon termini. TA dinucleotides in the transposon flanks are likely involved in establishing cleavage specificity. The 2 nt overhangs at the excised transposon ends represent specific molecular signatures, which are strictly required for the following step of transposon integration into a target DNA.

By combining biochemical, fluorescence, and functional *in vivo* assays, we provided evidence that the SB transposase assumes different oligomeric states throughout the transposition pathway. SB exists prevalently as a monomer by itself and can recognize transposon ends as a monomer; however, it forms oligomers following transposon end binding. We propose that this allows synapsis and coordinated cleavage of two transposon ends in the active transpososome. Remarkably, we also found that the unusual assembly mode of the *SB* transposon largely contributes to its exceptional efficiency. In fact, interfering with this specific pathway reduces transposition rates *in vivo*.

2.5.2 Proposed model of *Hsmar1* and *SB* transposition

Building on the presented comparative analysis, the following mechanistic model can be proposed for *Hsmar1* and *SB* transposition and autoregulation (Figure 2-31).

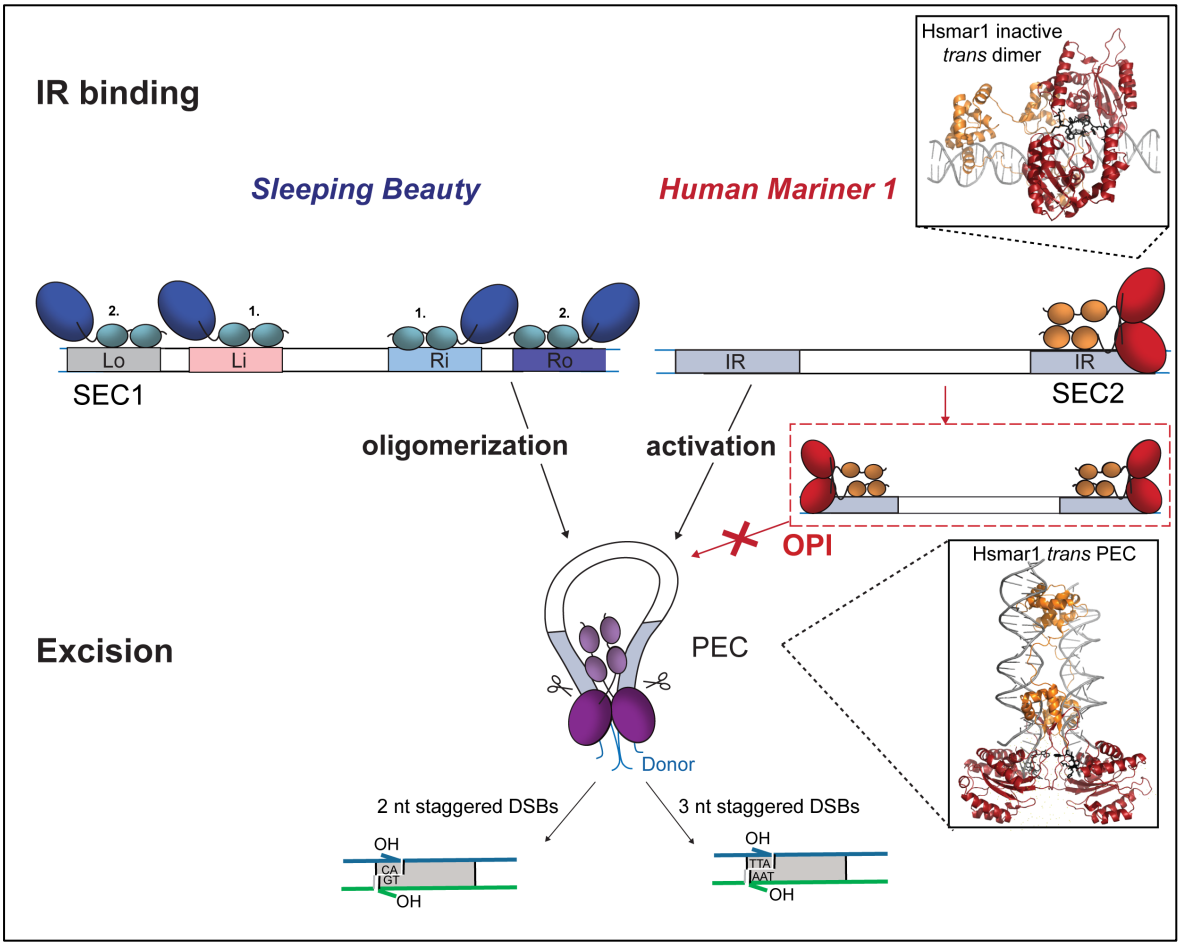


Figure 2-31: Proposed model for *SB* and *Hsmar1* tranposition. IR binding is performed by monomers (coloured blue; forming single-end complexes 1-SEC1) for *SB*; by dimers (coloured red; forming a single-end complex 2-SEC2) for *Hsmar1* and is followed by paired end complex (PEC) assembly (transposase subunits shown in purple). If two Hsmar1 transposase dimers contemporarily occupy the two transposon ends, productive synapsis is impaired and overproduction inhibition (OPI) arises. In the PEC, excision of the *SB* transposon generates 2 nt staggered DSBs, while *Hsmar1* is excised by producing 3 nt staggered breakages. Structural models of the Hsmar1 inactive dimer and PEC are shown. Transposase domains are depicted as ovals. Large oval: catalytic domain, lighter small ovals: DNA binding domain. Transposase binding sites are shown as rectangles, and are named left outer (Lo), left inner (Li), right inner (Ri), right outer (Ro) and simple inverted repeats (IRs). Blue line: non-transferred strand (NTS); green line: transferred strand (TS).

i. *Hsmar1* transposition. The Hsmar1 dimer (shown in red in Figure 2-31) initially assumes an autoinhibited *trans* conformation, mediated by interactions between the WVPHEL motifs in the interdomain linker (Figure 2-31, structural model of the inactive *trans* dimer). *Hsmar1* transposition starts with such a dimer binding to a single transposon end (SEC2). At this stage, the Hsmar1 dimer maintains its catalytically incompetent state, thereby preventing aberrant single end cleavage. Then, recruitment of the second end will resolve the inhibited transposase conformation, allowing coordinated cleavage of both transposon ends in the PEC (modeled in Figure 2-31). However, at high transposase concentrations, two transposase dimers can simultaneously occupy both transposon ends. In these conditions, PEC assembly is prohibited and, as a consequence, transposition is inhibited (overproduction inhibition-OPI) (Claeys Bouuaert, Lipkow*, et al.*, 2013). According to our structural data and models, only second end binding will trigger the unzipping of the WVPHEL motifs, which is required for catalytic activation of the transposase. This is consistent with the finding that the Hsmar1 dimer can occupy, but not cleave single transposon ends, both in transposition initiation and in OPI conditions. Following PEC assembly, catalysis occurs on transposon end DNA in a specific order, where NTS strictly precedes TS cleavage. Coordination and ordered progression of the excision reactions requires major conformational changes in the transpososome. After NTS nicking, TS cleavage likely occurs in a complex conformation immediately competent for successive target binding and transposon integration.

ii. *SB* transposition. *SB* follows an alternative mechanism of trasposososome assembly, as compared to *Hsmar1*. Monomers of the SB transposase (shown in blue in Figure 2-31) bind the direct repeats at the two transposon ends separately. The inner repeats are likely bound first, the outer sites later. Transposon end binding then triggers oligomerization of the SB transposase to assemble a synaptic complex that is immediately active for catalysis. In comparison to *Hsmar1*, this constitutes a different mechanism for ensuring that cleavage activity is tightly coupled to transposon end synapsis, which relies on DNA-mediated assembly rather than autoinhibition. By this specific assembly pathway, SB can escape OPI and optimize its efficiency even at high transposase concentrations. Transposition proceeds with transposon end cleavage, resulting in 2 nt staggered processed ends, which are finally ligated to target DNA during transposon integration.

Specific aspects of this proposed model are discussed in the following sections.

2.5.3 Structural principles of *Hsmar1* transposition: inhibition of single end cleavage

One of the most unexpected findings is the autoinhibitory conformation assumed by the Hsmar1 transposase in its apo state. In the crystal structure of the Hsmar1 CAT, the WVPHEL motifs of the transposase subunits are tangled, each covering the active site of the other subunit in a *trans* arrangement mediating autoinhibition of the Hsmar1 transposase. Consistently, specific mutations in the WVPHEL motif generate hyperactive variants of both Hsmar1 (D. Liu *et al.*, 2014) and the homologous Himar (Lampe, 2010) transposases.

According to the proposed structural model, single transposon ends are first contacted only by the DBD and not by the catalytic domain of the transposase (see structural model of the inactive *trans* dimer in Figure 2-31). This is consistent with the work that I performed during my M.Sc. thesis showing that the DBD of Hsmar1 is capable of transposon end binding alone. Single end binding appears not to be energetically sufficient to impair stability of the inhibitory interactions that link the two WVPHEL motifs. Only the subsequent second end recruitment will trigger the untangling of the two intertwined WVPHEL motifs and catalysis in a *trans* PEC (see structural model of the *trans* PEC in Figure 2-31). This is consistent with previous biochemical findings that second end binding is the rate-limiting step of synapsis and transposition (Clayes Bouuaert *et al.*, 2010), allocating major conformational changes to this reaction step.

In fact, in the PEC conformation the WVPHEL motif is involved in an alternative set of interactions: each WVPHEL contacts the conserved YSPDL motif of the partner subunit and the transposon end DNA that is bound to the DBD of the other transposase molecule in *trans* (based on homology with Mos1; see Figure 2-5 and Figure 2-10). Our model strongly agrees with previous data showing that Hsmar1 transposases mutated in the WVPHEL motif synapse and cleave transposon ends faster, but less accurately, than the wild type protein (D. Liu *et al.*, 2014). We propose that these mutations destroy interactions of the WVPHEL motifs, increasing the rate of PEC assembly and synapsis, while loosening inhibition of aberrant cleavage reactions.

The WVPHEL motif is conserved in mariner transposases and has been proposed to execute several functions in their transposition (Bouuaert *et al.*, 2014). Our findings strongly indicate a structural role of the WVPHEL motif in transposition autoinhibition and regulation. This motif appears to play a central role in transposon regulation, establishing mutually exclusive interactions with itself, with transposon end DNA intermediates and with other specific protein regions throughout the transposition process, orchestrating the ordered progression and fidelity of the reactions.

High-resolution structures of the full length Hsmar1 dimer on its own and/or bound to a single transposon end DNA would provide final experimental proof of our model for *Hsmar1*'s autoinhibition. Such structures would also reveal for the first time how a eukaryotic transposon initiates and regulate its movement at the structural level.

2.5.4 Structural principles of *Hsmar1* transposition: OPI

In the previous section, the structural principles preventing aberrant single end cleavage in *Hsmar1* transposition are proposed. These molecular constrains ensure efficient and precise cut-and-paste transposition and at the same time reduce potential genotoxicity of transposons in their hosts. Another inhibitory mechanism that is encoded in the transposon itself is OPI (see section 2.1.8).

Two main models have been proposed to explain OPI:

1. Formation, of transposase oligomers that are inactive in transposition at high transposase concentration (G. Carpentier *et al.*, 2011).
2. Assembly site occlusion (ASO) model, where OPI arises from the multimeric state of the transposase and is mediated by competition for the two transposon ends (Claeys Bouuaert, Lipkow*, et al.*, 2013).

The presented structural and functional data for *Hsmar1* now provide evidence supporting the second model, where OPI is connected to the specific mechanism of transpososome assembly. The autoinhibited conformation of the Hsmar1 transposase explains how, at high transposase concentrations, two dimers could block both transposon ends without cleavage, preventing synapsis.

Nevertheless, the two models are not mutually exclusive and could both contribute to establishment of the OPI phenomenon. Similarly to the previous Mos1 crystal structures (Richardson et al., 2006), the structure of the Hsmar1 catalytic domain revealed several

dimer interfaces in the crystal, which are different from the autoinhibitory interface and from the one observed in the Mos1 PEC. These alternative interfaces might be involved in higher oligomerization and/or formation of inactive assemblies, as postulated in Model 1. Consistently, I observed formation of Hsmar1 transposase aggregates at high protein concentration in SAXS experiments (data not shown). Whether alternative inactive oligomers exist as functional intermediates in *Hsmar1* transposition remains to be investigated.

2.5.5 Structural principles of *Hsmar1* transposition: transposon end cleavage

My biochemical data [together with (Claeys Bouuaert *et al.*, 2014)] indicate a specific order of the DNA hydrolysis reactions involved in transposon end cleavage, and imply that discrete conformational changes in the transpososome are required to execute and order these events. The specific cleavage order, NTS followed by TS ensures that after transposon excision the TS strand remains in the transposase active site and can directly integrate into a target DNA [as suggested by (Robertson *et al.*, 1997)]. Due to lack of informative structures, how these reactions are coordinated in the context of the transpososome have so far remained unknown. Structure determination of the crystallized pre-excision complexes described in this thesis will allow to elucidate these aspects at high-resolution. However, preliminary hypotheses can be drawn based on available biochemical data.

In my M.Sc. thesis work, I showed that monomers of Hsmar1 are catalytically incompetent even in the first step of NTS nicking. This indicates that the Hsmar1 cannot cleave DNA in *cis* and all reactions occur only in a *trans* synaptic complex. Of note, our structural model supports the notion that the Hsmar1 transposase assumes a *trans* conformation both in its autoinhibited and in its PEC state (see structural models in Figure 2-31). Thus, it is unlikely that Hsmar1 would act in *cis* on the NTS after DNA binding, and then reassume a *trans* assembly to cleave the TS and integrate. We therefore propose that monomers of the transposase are always in a *trans* arrangement and specific structural changes in distinct regions (as in the WVPHEL) guide the transposition reactions forward to completion.

Moreover, it is worth noting that the Hsmar1 transposase can cleave the TS when the NTS is already cut (e.g. contains a synthetic nick). This suggests that it is the recognition of a cleaved NTS, rather than the actual cleavage by the transposase that is required to enable progression towards TS cleavage. Thus, DNA is probably the main factor that, by changing its structural and biochemical properties upon processing, dictates ordered remodelling of the transpososome throughout the transposition reaction.

2.5.6 Biochemistry of *SB* transposition

Concerning the *SB* transposon, my biochemical analysis showed that the SB transposase mediates transposon end binding, excision and integration *in vitro* and these events are characterized with unprecedented, distinctive features.

Regarding transposon end binding, SB binds DRs as a monomer and exhibits higher affinity to the inner repeats than to the outer repeats. This might dictate a specific order in recognition and assembly of active complexes, as will be discussed in detail in section 2.5.7. Moreover, left DRs seem to bind more tightly than right DRs. These differences probably reflect a distinct role of the left and right transposon ends in transposition and provide an explanation for previous *in vivo* transposition assays. In these studies, a modified symmetric *SB* transposon containing left end sequences at both ends (LoLi-LiLo) was nearly as active as the standard transposon, while a construct made up of right sites only (RoRi-RiRo) retained only 50% of *SB*'s original activity (Cui *et al.*, 2002). Taken together, these findings suggest that the left binding sites are strictly required, while the right ones are dispensable for successful *SB* transposition.

By reconstituting the excision and integration activity of SB *in vitro,* I demonstrated that synaptic assembly and catalysis only requires the transposase enzyme and no additional host-derived protein factors [in contrast to bacterial transposases as Tn7, Tn10 and the bacteriophage Mu transposase (Nesmelova *et al.*, 2010)]. This is also true for all biochemically characterized *Tc1/mariners* to date (Clayes Bouuaert *et al.*, 2010; Lampe *et al.*, 1996; Vos *et al.*, 1996) and agrees with the hypothesis that *Tc1/mariners*' widespread distribution in the tree of life reflects their independence from host-encoded factors. Accessory proteins, such as DNA bending proteins (e.g. the high mobility group protein HMGB1) or repair factors (e.g. the Ku heterodimer), might still facilitate and/or cooperate

with *SB* transposition in the cellular context (Izsvak *et al.*, 2004; Zayed *et al.*, 2003), but are not critical for movement *per se*.

Using cleavage assays, I showed that SB specifically cuts single outer repeats, but not inner repeats *in vitro* for transposon excision. Thus, recognition of specific DNA nucleotides at the outer repeats dictates specificity of cleavage, rather than the position of the repeats in the full transposon end architecture. Perhaps the most apparent difference between DRs is the presence of an invariable TA dinucleotide exclusively at the outer repeat flanks. Mutational analysis showed that altering the dinucleotide flanking the repeats completely abolished transposition *in vivo* (Cui *et al.*, 2002), demonstrating the critical role of the flanking TA in the reaction. Thus, the absence of TA dinucleotide does not seem the only reason why inner repeats are "uncleavable", at least *in vivo*. Consistently, I observed that the SB transposase cleaves outer repeats always 2 nucleotides 3' of the TA dinucleotide. With two consecutive TAs (TATA) in the flank, two cleavage positions were observed. Notably, higher transposition rates were obtained *in vivo* with the TATA motif, perhaps due to an increased flexibility of NTS cleavage (Cui *et al.*, 2002). Thus, recognition of the flanking TA dinucleotide by the transposase seems to be strictly required for cleavage. Consistently, the Mos1 transposase makes specific contacts with the flanking TA dinucleotide through its WVPHEL motif in the crystal structure, positioning the transposon ends precisely for cleavage (Dornan *et al.*, 2015). Since the WVPHEL motif is not conserved in SB, the structural principles of flank recognition remain to be explored for this transposon.

I demonstrated that the SB transposase generates 2 nt staggered dsDNA breaks at the transposon ends, which are essential for subsequent transposon integration (as highlighted in Figure 2-31). This is particularly interesting because the excision of other *Tc1-like* transposons, as *Tc1* and *Tc3*, also leads to the formation of 2 nt staggers on their transposon ends (van Luenen *et al.*, 1994; Vos *et al.*, 1996), whereas members of the *mariner* clade, as *Hsmar1* and *Mos1*, excise by producing 3 nt staggered DSBs (Claeys Bouuaert *et al.*, 2010; Dawson *et al.*, 2003). These differences might reflect distinctive structural and mechanistic features of the two transposon subfamilies within the *Tc1/mariner* superfamily. Regarding the structural determinants of these differences, it can be speculated that the DDE catalytic triad in the Tc1-like transposases allows

accommodating only 2 nucleotides of the TS in the catalytic pocket, whereas the mariner-specific DDD motif fits 3 nucleotides better.

Concerning the final step of transposon integration, I noticed that SB integrates the transposon end DNA only into target DNA plasmids *in vitro*. In turn, Hsmar1 (this thesis work) and Mos1 (Morris *et al.*, 2016) transposases can use linear DNA as targets. This indicates that integration by SB greatly depends on DNA topology, where more bendable structures are favoured [in agreement with computational analysis of unique *SB* insertions in HeLa cell genomes (Vigdal *et al.*, 2002)]. Consistently, it has been shown that DNA flexibility has a strong positive effect on transposition of many transposons. Structural studies on several DDE/D transposases revealed that these proteins strongly bend their target DNA prior to integration (Montano *et al.*, 2012; Morris *et al.*, 2016). Such target bending may be required to avoid immediate re-excision of the integrated transposon: a bent target would snap away from the transposase active site immediately after strand transfer, thereby rendering the reaction irreversible (Montano *et al.*, 2012).

Finally, I observed that coordinated ligation of both transposon ends in a target DNA occurs preferentially over aberrant single end integration *in vitro*. This is quite striking because other integrases, such as the HIV integrase (Li *et al.*, 2006), and transposases, like piggyBac (Mitra *et al.*, 2008), Hermes (Zhou *et al.*, 2004), and Mu (Haapa *et al.*, 1999), produce both integration products at similar rates in *in vitro* assays. This suggests that SB regulates its transposition more tightly than other integrases. Such increased coordination of integration might result from SB's unique transpososome assembly pathway (see next section, 2.5.7), as integration of a single end by SB monomers or transient dimers is likely very limited.

2.5.7 The unique assembly mode of the SB transpososome

In this thesis work, the most surprising difference found between *Hsmar1* and *SB* concerns their mechanisms of transpososome assembly. While Hsmar1 (and other mariner transposases) achieves synapsis by sequential transposon end recruitment (Claeys Bouuaert, Lipkow*, et al.*, 2013), the SB transposase brings the two ends together by protein oligomerization (Figure 2-31). Both mechanisms are consistent with the generally accepted principle that coordinated, complete processing of the two transposon ends requires at least two transposase subunits in the transpososome (Hickman *et al.*, 2016). Transpososome

assembly by oligomerization is a very effective mechanism to directly coordinate catalysis with synapsis and has been reported to be used extensively by prokaryotic transposases (as Tn7, Tn10 and Mu). For example, the Mu transposase is also purified in a monomeric form but forms tetramers *in vitro* upon DNA recognition for transpososome assembly (Baker *et al.*, 1994). Eukaryotic transposases have been generally thought to employ sequential end recruitment, which provides an opportunity for sensitive autoregulation. This work now shows that eukaryotic elements can also conform with the prokaryotic pathway, allowing them to escape OPI.

Several structural explanations for the difference between *Hsmar1* and *SB* can be proposed. The crystal structure of the SB catalytic domain (Voigt *et al.*, 2016) has pointed out that SB complex assembly likely involves Tc1-like specific structural regions that are not present in mariner transposases. These include the distinct glycine rich strip at the base of the clamp loop and the conserved KKPL linker region, which differs from the *mariner*-specific WVPHEL. I discussed above how the WVPHEL motif contributes to CAT dimerization (and autoinhibition) of the Hsmar1 transposase; thus, its absence in SB might contribute to its primarily monomeric state.

I could also show that oligomerization of the SB transposase is likely mediated by hydrophilic interactions in its PAI domain (especially by residue Asp17). This differs from the largely hydrophobic character of the HTH1 dimerization interfaces of Hsmar1 and Mos1 [PDB ID: 3HOT, (Richardson *et al.*, 2009)]. Hydrophilicity of SB's PAI domain might destabilizes initial homodimers, while allowing to alter the oligomeric state upon DNA binding.

Of note, the data presented establish that binding to outer repeats triggers oligomerization of the transposase *in vitro*; however, how this occurs in the context of the full transposon end remains uncharacterized. In particular, it is still unknown whether the inner repeats can trigger transposase oligomerization and how cooperative binding to the four repeats is established. In this respect, two possible mechanisms of transposition initiation can be proposed (based also on binding studies in section 2.3.1.1):

1. Two SB monomers bind with high affinity to inner repeats and form oligomers, bringing the outer repeats into close proximity. Outer repeats could be occupied by SB monomers before or after these events. Oligomerization at the inner repeats

facilitates oligomerization at the outer repeats, triggering assembly of the excision-competent complex.

2. Two SB monomers bind the inner repeats first and sequentially recruit additional monomers at the outer sites by specific protein-protein interactions. Binding of the second pair of monomers at the outer repeats then destroys the previous protein interactions and favours transposase oligomerization in the PEC.

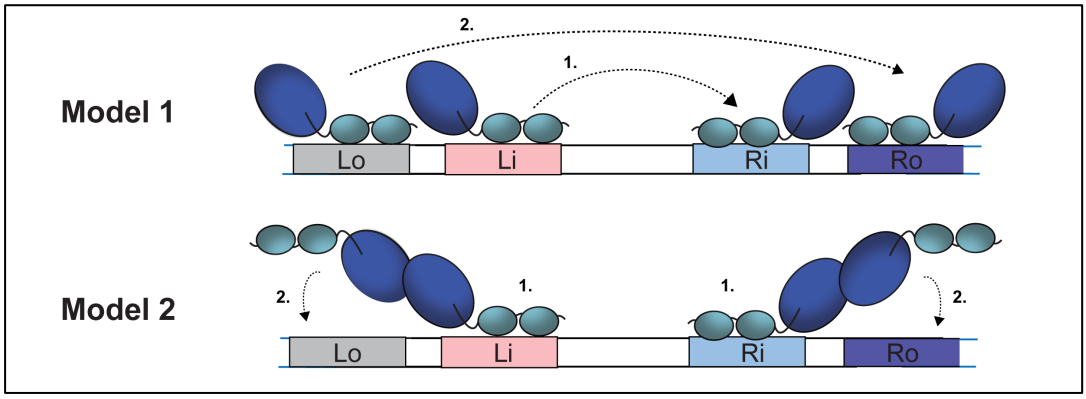


Figure 2-32: Proposed model for the role of the inner repeats in *SB* transposon end recognition and assembly. Model 1: SB transposase monomers bound to inner repeats oligomerize first, triggering oligomerization of protomers also at the outer sites. Model 2: SB transposase monomers bound to inner repeats recruite protomers to the outer sites by protein-protein interactions. Colours and shapes as in Figure 2-31.

In this work I specifically identified the PAI domain to be involved in transposase oligomerization. This is consistent with previous biochemical data based on isolated SB DNA binding domains (Izsvak *et al.*, 2002). Thus, similarly to Mos1 and Hsmar1, SB protomers interact through their N-terminal domain upon binding to outer repeats in the PEC. However, several alternative dimer interfaces might be involved in *SB* transposition. A discrete, inactive dimer interface was identified in the structure of the SB catalytic domain and has been proposed to link two monomers during the putative recruitment step described in Model 2 (Voigt *et al.*, 2016). Structural studies of complete transposon ends in complex with the SB protein (for example by Cryo-electron microscopy or crosslinking mass spectrometry techniques) are required to characterize the exact stoichiometry of the transpososome and elucidate these aspects at the molecular level.

Furthermore, it is still unclear whether any DNA cleavage and/or ligation reactions can occur in SB monomers. In the light of what I have discussed for the Hsmar1 transposase

(see section 2.5.5), it is expected that even the first strand cleavage occurs in assembled transpososome. In addition, insertion of single transposon ends barely occurs in the established *SB* integration assays even *in vitro*, despite the abundance of monomeric species in this setting, indicating that oligomers are strictly required for correct transposon integration.

2.5.8 Conservation across *Tc1/mariners*

This work reveals a remarkable mechanistic diversity between the *Hsmar1* and *SB* transposons. This raises the question if characteristic mechanistic principles hold true for other *Tc1/mariners.*

Mariner transposases (like Mos1, Mboumar-9, Himar1 and Hsmar1) share circa 40% sequence similarity, suggesting they might follow similar transposition pathway as Hsmar1. Consistently, the Mos1 and Mbourmar-9 transposases have been described as stable dimers that first cleave the NTS and then the TS generating 3 nt staggered transposon ends (Trubitsyna *et al.*, 2014), analogously to Hsmar1. Conservation of the *mariner*-specific WVPHEL motif also points to the idea that this structural region has a role in dimer formation and autoinhibition of other mariner transposases. Nevertheless, the Mos1 transposase has been described to form an elongated dimer prior to DNA binding that is held together only by its DBD. Thus, second end recruitment and synaptic complex formation would require rotation of one transposase subunit around its DBD (Cuypers *et al.*, 2013). In contrast, our structural data indicate that minor conformation rearrangements, mostly in the WVPHEL motif, trigger the activation of the Hsmar1 transposase. For these reasons, the inter-WVPHEL interactions in the Hsmar1 transposase might represent a specific and unique layer of autoregulation among the *Tc1/mariners*. This means that, despite its conservation, the WVPHEL motif might play different roles in distinct *mariner* elements, resulting in different levels of control and autoregulation. Of note, the Hsmar1 transposase is considered to be the most specific and tightly controlled Tc1/mariner transposase studied so far (Tellier *et al.*, 2015), perhaps related to the extensive inter-WVPHEL interactions observed in this thesis.

About *Tc1-like* transposons, no experimental evidences are available to support a shared mechanism of synapsis by protein oligomerization among the clade. SB and Tc3 have been proposed to form tetramers, based mainly on their transposon ends architecture (Brillet *et*

al., 2007) and available data on isolated DNA binding domains (Izsvak *et al.*, 2002; Watkins *et al.*, 2004). The PAI domain is also only weakly conserved in Tc1-like transposases, such as SB, Tc1, Tc3 and Frog Prince. Therefore, the interesting idea of a correlation between hydrophilicity of the dimer interface and DNA-promoted oligomerization would need further investigation. It is worth noting that SB is an artificially resurrected transposase; therefore, its specific sequence and structural features might be due to the molecular reconstruction process and may thus differ from the original native protein and from other Tc1-like transposases.

On the other hand, SB contains the canonical Tc1-like specific catalytic triad (DDE) and shares the same cleavage specificity (2 nt staggered DSB) at transposon ends with other biochemically characterized Tc1-like transposases, like Tc1 (Vos *et al.*, 1996) and Tc3 (van Luenen *et al.*, 1994). Whether this cleavage specificity extends to all Tc1-like transposases, it is still unknown. In addition, the monomeric state of SB and the conserved absence of the WVPHEL motif in Tc1-like transposases suggest a different structural basis for excision regulation compared to the *mariners*, which may be shared in Tc1-like transposases.

In conclusion, due to low sequence conservation and sparseness of experimental data, it is difficult to evaluate conservation across *Tc1-like* elements. Thus, many open questions remain about the working mechanism of these transposons. To shed light on these aspects, comparative biochemical and structural characterization of SB and other Tc1-like transposases and transpososomes is needed. This thesis provides methods, assays and strategies (including for crystallization) that could be specifically employed for this purpose.

2.5.9 Regulation of *SB* transposition

One of the most puzzling questions about *SB* transposition regards the mechanistic principles behind its extraordinary efficiency. In this work, first experimental evidences is provided that the specific pathway of transpososome assembly directly dictates the high transposition rate.

Previous reports indicated that *SB* is susceptible to very limited, if any, OPI (Kolacsek *et al.*, 2014). In this work, we show that forcing SB dimerization, and thus interfering with its

synapsis by oligomerization, strongly reduces transposition efficiency in living cells. This data indicate that *SB* employs a specific assembly mechanism, which allows it to escape OPI, largely contributing to its extraordinary high transposition efficiency. In agreement with these findings, computer simulations also predicted that transposons using the synapsis by oligomerization mechanism would sustain long-lasting high transposition levels in eukaryotes (Claeys Bouuaert, Lipkow*, et al.*, 2013) and can only be inactivated by mutations.

SB's unique assembly pathway also provides a distinctive advantage for its technological use, since the transposase is generally overexpressed in these applications. Thus, in this work, we shed first light on the molecular principles behind *SB*'s extraordinary success as a genetic tool. Furthermore, the presented mechanistic findings provide a unique opportunity to control *SB* transposition in living cells. For example, we were able to artificially reduce insertion rates by chemically modulating transposase oligomerization in HeLa cells. By creating dimerization defective SB mutants that retain DNA binding properties and are fused to heterologous inducible dimerization domains, one could trigger transposase activation at specific desired time points. Such dimerization-dependent strategies could provide the basis for regulating time, efficiency and conditions of *SB* transposition on demand in its genetic engineering applications.

3 DEVELOPMENT OF NOVEL *SB*-BASED GENETIC TOOLS

The part of my PhD work presented in this chapter focuses on the development of novel genetic tools based on the SB transposon system. The first section (section 3.1) introduces the use of *SB* as a prominent genetic tool both in research and in clinical trials and highlights advantages and current limitations of its applications. Based on this introduction, the motivation and aims of my research are stated. Section 3.2 presents the design and functional characterization of a novel SB transposase variant, named hsSB. This provided the basis for the development of a new genome engineering method based on direct cellular delivery of the hsSB protein, named SBprotAct (described in section 3.3). In the last section (section 3.4), the advances and impacts offered by SBprotAct are discussed in comparison to the state-of-start technologies and future directions in the further development of the method are proposed.

Note: A number of the figures in this chapter are based on and modified from the following manuscript and patent application.

- Querques, Mades, Zuliani *et al.*, (2018); "Genome engineering by direct delivery of engineered Sleeping Beauty transposase protein.", submitted
- Barabas, Querques and Zuliani (filed August 2017); "Improved Transposase Polypeptide and Uses Thereof", European Patent Application 17187128.8

3.1 Introduction – SB transposon system as a genetic tool

Transposons can be viewed as natural DNA transfer vehicles that are capable of efficient genomic insertion. In this respect, *SB* is the element that has advanced the fastest and furthest in research and the clinics, with applications spanning from cancer gene discovery to even gene therapy in humans. This introduction describes the "story" of the *SB* transposon as a genetic tool, including successes and pitfalls, highlighting the importance of further development of the technology, especially for clinical use.

3.1.1 Overview of *SB*'s applications

The potential of the *SB* transposon for genome manipulation in vertebrates has been recognized concomitantly to its resurrection from fish (Ivics *et al.*, 1997). The possibility to externally control transposition by separating the two obligatory components: the IR sequences and the SB transposase is critically important. This allows virtually any DNA of interest (a fluorescent marker, a mutagenic gene trap, or a therapeutic gene) placed between the IR sequences in the transposon to be stably inserted in a target genome by the transposase conditionally provided in *trans* (Figure 3-1).

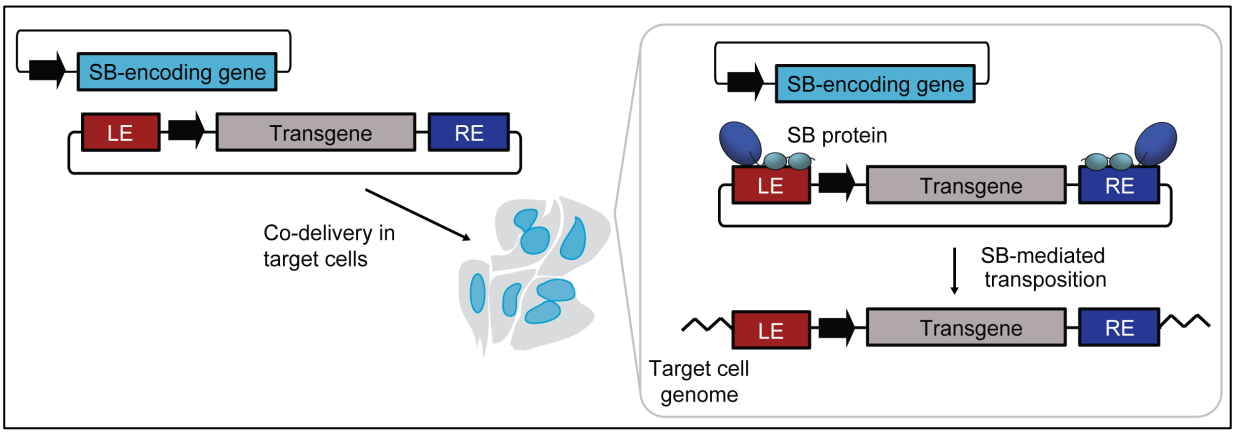


Figure 3-1: Schematic representation of genetic engineering with the SB transposon system. The transposon-carrying and the transposase-encoding plasmids are co-delivered into target cells, where overexpression of the SB transposase leads to transgene insertion into the target cell genome by transposition. LE and RE indicate transposon end sequences. LE: left end; RE: right end; black arrow: promoter.

This engineered two-component entity, referred to as the SB transposon system, can be placed on various DNA carriers, such as plasmids or minicircles (Monjezi *et al.*, 2017) (non-viral vectors) or can be combined with various viruses (hybrid transposon/virus

vectors) (Grabundzija *et al.*, 2013). The typical set-up for delivery into cells is by supplying both components as conventional plasmids (Ammar *et al.*, 2012; Ivics *et al.*, 2009). However, the transposase can be also provided in the form of a messenger RNA (mRNA) synthetized *in vitro* (Monjezi *et al.*, 2017; Wiehe *et al.*, 2007; Wilber *et al.*, 2006; Wilber *et al.*, 2007).

Since its discovery, SB has proven to support highly efficient transgene integration in various cell types, including somatic and germ cells, differentiated and stem cells. It is very efficient in cells of all vertebrate species (Izsvak *et al.*, 2000) as well as in preclinical animal models (Hackett *et al.*, 2010) and livestock. Consequently, it has been extensively applied for creating transgenic cell lines and gene knockdowns, also in combination with other genetic engineering tools (Grabundzija *et al.*, 2013; Petrakis *et al.*, 2012) and various delivery techniques (viral or non-viral). Of particular biotechnological interest is its use for generating commercial cell lines, such as Chinese hamster ovary (CHO) cells, for large-scale production of recombinant proteins (Balasubramanian *et al.*, 2016).

One of SB's most prominent applications is its use as insertional mutagen in cancer research. It was used to generate cancer-causing mutations, where analysis of its insertions allowed identification of cancer genes (Collier *et al.*, 2005; Dupuy *et al.*, 2005) and important cell signalling pathways involved in different cancer types (Bermejo-Rodriguez *et al.*, 2015; Moriarity *et al.*, 2015; Takeda *et al.*, 2016; Wu *et al.*, 2016). In addition to functional oncogenomics, incorporating a regulatory sensor within the transposon allowed mapping the regulatory architecture of chromosomes in mice (Ruf *et al.*, 2011).

Additionally, SB has been successfully used to introduce reprogramming factor-encoding cassettes for generating induced pluripotent stem cells (iPSCs) in various organisms and models (Davis *et al.*, 2013) (Talluri *et al.*, 2015). Once reprogrammed, engineered iPSCs can be expanded and differentiated into multiple cell types for use in research and gene therapy applications. Moreover, SB has been also extensively probed and utilized to generate transgenic animals [including fish, frog, rat, mouse, rabbit, pig, cow and others, reviewed in (Narayanavari *et al.*, 2017)].

Importantly, several preclinical gene therapy trials have been carried out using the SB transposon system, aimed at treating hematologic disorders, lysosomal storage diseases,

pulmonary, dermatological, metabolic and neurologic disorders, and cancer (Aronovich *et al.*, 2007; Escobar *et al.*, 2016; Eyjolfsdottir *et al.*, 2016; X. Huang *et al.*, 2008; Latella *et al.*, 2016; Singh *et al.*, 2008; Turunen *et al.*, 2016; Yant *et al.*, 2000). These studies have paved the way for SB's breaking on through to the clinical side, which is discussed in the following paragraph.

3.1.2 Clinical applications of the SB transposon system

Among SB-based applications, the ones that trigger perhaps most excitement at the moment relate to in human gene therapy.

In the last years, important steps have been made toward the application of SB for treatment of neovascular age-related macular degeneration (AMD). AMD refers to the degeneration of retinal pigment epithelial (RPE) cells in the macula area of the eye and is a major cause of acquired irreversible blindness in adults worldwide. One form of this disorder, the neovascular AMD (nvAMD), can be successfully treated by replacing degenerated RPE cells with genetically modified ones that overexpress the pigment epithelium-derived factor (PEDF). In recent preclinical trials, the hyperactive SB100X was used to efficiently deliver a human PEDF transgene cassette to cultured RPE cells. These modified cells were subsequently transplanted, and significantly suppressed the characteristic neovascularization associated with nvAMD both in rabbit (Kuerten *et al.*, 2015) and rat (Johnen *et al.*, 2015) models. Following these pioneering studies and recent developments of optimized protocols (Thumann *et al.*, 2017), clinical trials have been approved (in Switzerland) and will soon be initiated. In these trials, pigment epithelial cells isolated from iris biopsies of nvAMD patients will be transfected with a PEDF-carrying SB vector and transplanted back by a simple surgical session (Hudecek *et al.*, 2017).

Even more advanced is the use of SB-mediated gene transfer for engineering human T cells for anticancer immunotherapy. Adoptive immunotherapy is based on genetic modification of human T cells (derived from patients in autologous therapies or healthy donors in allogenic ones) *ex vivo* by incorporating the genetic information for a chimeric antigen receptor (CAR) against malignancy-specific antigens. CARs are artificial receptors that are used to provide T cells with new specificities and trigger effector functions after antigen encounter (Fesnak *et al.*, 2016; Sadelain *et al.*, 2017). The engineered CAR T cells are then re-infused in the cancer patient to seek and kill cancer cells. In particular, CAR T cell therapy targeting CD19, a biomarker overexpressed on the surface of malignant B

cells, has shown unprecedented high response rates (70%-90%) in the treatment of acute lymphoblastic leukemia (ALL) (Turtle *et al.*, 2016) and B cell lymphomas (Locke *et al.*, 2017). Although CD19 is expressed also in healthy B cells, the global destruction of the B cells can be compensated with a globulin treatment.

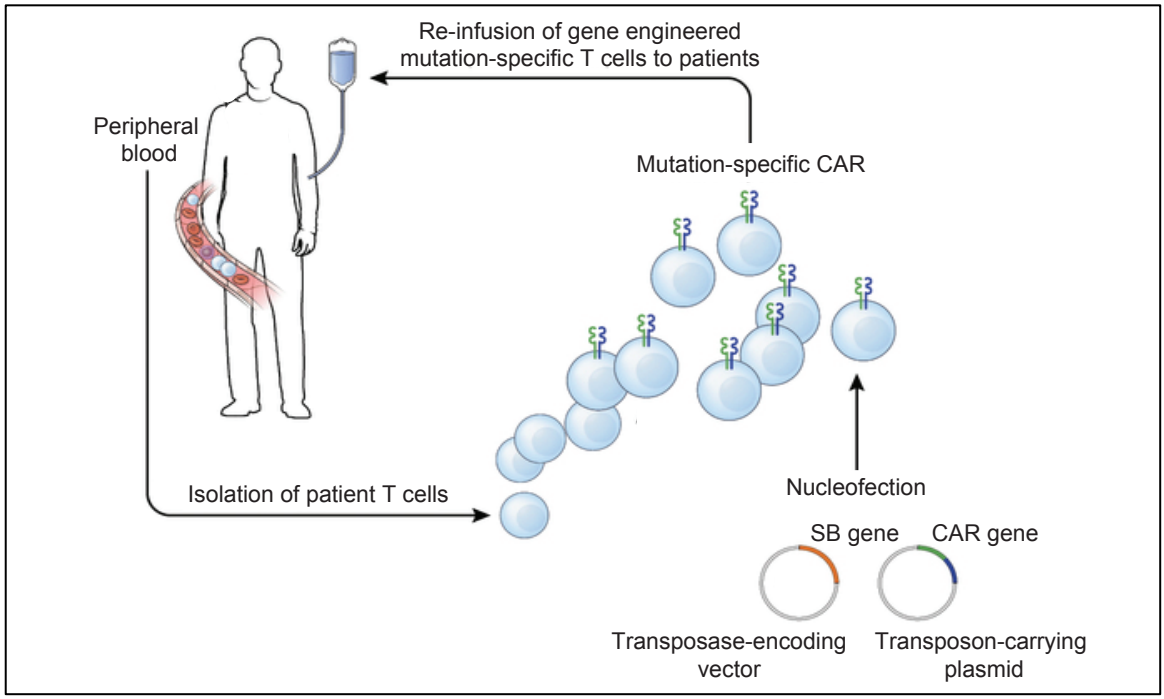


Figure 3-2: General procedure for autologous chimeric antigen receptor (CAR) gene therapies targeting malignant antigens using the SB system. The specific CAR gene that confers reactivity against the cancer-specific antigen is cloned into the transposon. T cell subsets isolated from the peripheral blood of the patient are modified by the dual SB transposon/transposase system upon nucleofection of the two vectors, expanded *in vitro*, and re-infused into the patient. The figure has been adapted from (Klebanoff *et al.*, 2016).

SB has been applied as a non-viral gene delivery vector in several CAR T clinical trials at the MD Anderson Cancer Center (MDACC) (Clinical Trials.gov NCT00968760, NCT01497184, NCT02529813 -still ongoing-; and NCT01362452, NCT01653717 -concluded), using SB to introduce a CAR against the CD19 antigen into (both autologous and allogenic) T cells. First published results [related to Clinical Trials.gov NCT00968760 and NCT01497184; (Kebriaei *et al.*, 2016)] for patients with relapsed B cell ALL or lymphoma showed a promising outcome, establishing proof-of-concept for the feasibility and safety of SB-mediated CAR T engineering, and initiating a follow-up trial (Clinical Trials.gov NCT02529813) with further optimized protocols.

On the 30th August 2017, the first CAR T therapy treatment (named Tisagenlecleucel, using viral vectors for CAR delivery) has obtained FDA (Food and drug administration) approval for the treatment of young patients with B cell ALL. This approval represents a milestone in the history of anticancer therapy, but it has important implications for health-care systems owing to its considerable logistical, toxicological, and financial challenges ("CAR T cells - what have we learnt?," 2017). In the approved approach, T cell modification is achieved via the use of viral vectors that are produced via an extremely cumbersome and expensive manufacturing process. Due to the elaborate treatment procedures, very few specialized cancer centers in the world can offer the therapy. This results in the enormously high cost of Tisagenlecleucel (US$ 475,000 for a single infusion) and clearly impairs access to the technology for a large number of patients. In addition, adverse effects associated with CAR T cell therapy, for example the cytokine-release syndrome, are largely uncharacterized and poorly predictable to date. Even more concerning is the fact that ~50% of patients who responded to Tisagenlecleucel in the first pediatric global trial (ELIANA Clinical Trials.gov NCT02435849) relapsed within 1 year.

Therefore, to enable CAR T therapies to enter routine clinical practice, there is a pressing need to improve manufacturing feasibility, efficacy, and safety, which are critical requirements of gene therapy in general. The use of SB for gene delivery helps to overcome several limitations in these aspects, as described in the following section.

3.1.3 Advantages of the SB system for clinical use

SB transposon-based gene delivery provides distinctive advantages for clinical use in comparison to viral vectors and to other non-viral delivery vehicles (e.g. naked DNA molecules).

SB is a simple synthetic DNA-based binary system, which is cheaper, easier, and faster to produce and implement than viral vectors. This provides a particular advantage especially for single-use and personalized applications. In addition, it offers high scalability of production as well as easy quality control for clinical applications. The use of the non-viral SB vector also reduces the risks of undesired immune response activation in patients (Yant *et al.*, 2000), which constitutes a major safety concern related to the use of viral proteins and nucleotides in viral delivery approaches.

Due to the transposition mechanism, SB belongs to the class of integrating vectors, which directly insert the transgene in the target cell's genome. Among non-viral delivery vehicles, SB is the only one that offers sufficient insertion efficiency to allow stable transgene incorporation suitable for medical use. Even some viral vehicles such as adenoviral and adeno-associated viral vectors remain episomal, requiring several doses for sustained transgene acquisition, which in turn can lead to uncontrolled immune responses (Basner-Tschakarjan *et al.*, 2014).

Moreover, standard gamma-retroviral and lentiviral vectors preferentially insert into actively transcribed or regulatory regions creating a risk for insertional mutagenesis and genotoxicity (Cesana *et al.*, 2012; Hargrove *et al.*, 2008). In contrast, SB presents a close-to-random integration pattern, greatly reducing the risk of genomic aberrations. It has been also reported that retroviral stock can accumulate mutations at significant frequencies as a consequence of reverse transcription (Menendez-Arias, 2009). As transposition involves exclusively DNA intermediates, the risk of mutagenesis of the SB vector is very limited. Furthermore, since SB's cargo capacity is large (Zayed *et al.*, 2004) and not subjected to physical constrains (e.g. by DNA packaging in capsids as for viruses), it provides a distinct advantage over viral vectors for delivery of large transgenes or multigene cassettes (e.g. CARs).

Compared with integrating viral vectors, the SB system has generally lower gene-transfer efficiency. However, in the last years, novel strategies, as improved design and delivery methods for its components as well as selective propagation of CAR positive cells, have increased the success of SB-mediated T cell engineering (Hudecek *et al.*, 2017). For example, the use of minicircle vectors as carriers of the SB transposon system components has resulted in an improved efficiency (up to fivefold) in genome modification, as well as an increase (up to threefold) of T cell survival rates due to decreased cytotoxicity upon nucleofection (Monjezi *et al.*, 2017).

Finally, differently to other genome editing nucleases as zinc-finger nucleases, TALENs (Transcription activator-like effector nucleases) and Cas9, the SB transposase directly and precisely integrates its cargo into chromatin. Designer nucleases generate DSBs at the target locus and co-opt the cellular repair machinery to integrate exogenous genetic cargos. This mechanism limits integration efficiency, induces excessive DNA damage responses

(Ihry *et al.*, 2017) and can also cause undesired genome modifications, limiting the utility of nuclease-based systems in the clinical setting to date. In contrast, SB's insertion rates do not depend on the efficiency of the repair machinery in the target cells and its transposition does not generate genomic DSBs, making it a more efficient and safer alternative. These advantages make SB the only non-viral vehicle currently used for CAR T cell engineering and soon also in other clinical trials.

3.1.4 Limitations and desired improvements

Despite the advantageous features of the SB system, important limitations remain and specific technological improvements are highly desirable, especially for advancement in the clinic:

1. Improved efficiency. Several current and past efforts have been directed towards the further enhancement of SB's insertion efficiency, and lead to substantial successes. However, transgenesis rate is still limiting in several medically relevant primary cell types and has remained below the ones reported in clinical trials using viral vectors.
2. Targeted integration. As discussed above, SB integrates its gene cargo randomly across the genome. If SB-based insertions could be directed to specific genomic sites (such as 'safe harbour sites'), that would further increase its safety profile in the clinical setting. A targeted (and ideally programmable) integrase that combines the advantages of genome editing nucleases (programmable site-specific modification of the genome) with those of the current SB system (high integration rate and safety) would be of at most interest for research and medicine.
3. Tight copy number control. Especially in therapeutic applications, it is extremely desirable to limit and tightly control the number of transgene copies inserted into the genome of the target cell. This is not possible with the current SB system (or with viral vectors), due to long-term exposure to the gene delivery system.
4. Direct delivery of the SB transposase protein. Current SB-based applications exclusively depend on administration of the system components in the form of nucleic acids. Direct delivery of the transposase would greatly improve control over the genome engineering procedure and increase the safety profile of the SB system. However, recombinant production of the SB transposase for cellular delivery could not been established before.

Regarding the last point, the form in which the transposase is supplied has a major impact on the outcome and safety of the SB-based engineering procedure. Long-term transposase expression from a plasmid vector in the target cells can result in uncontrolled ongoing transposition, potentially leading to transgene remobilization, undesired insertion events, genome instability, and cytotoxicity (Cai *et al.*, 2014; Galla *et al.*, 2011; Hackett *et al.*, 2013; X. Huang *et al.*, 2010; Q. Liang *et al.*, 2009). Moreover, insertion of the SB transposase gene from the expression vector (e.g. by homologous recombination) would result in infinite transposase production and activity. Acquisition of the transposase promoter might also cause activation of oncogenes or disrupt gene regulatory networks in the target cells.

To circumvent the concerns associated with the use of transposase-coding plasmids, the use of mRNA as transposase source has been explored (Galla *et al.*, 2011; X. Huang *et al.*, 2006; Jin *et al.*, 2011; Wilber *et al.*, 2006; Wilber *et al.*, 2007). However, these mRNA-based delivery approaches resulted in heterogeneous and drastically reduced transposition efficiency (Galla *et al.*, 2011; Jin *et al.*, 2011), probably due to the intrinsic instability of mRNA and difficult delivery to cells. The requirement of cumbersome *in vitro* synthesis procedures and the lack of quality control methods for mRNA further limit its use in a commercial or clinical setting. Moreover, with mRNA-based delivery, *SB* transposition is strictly dependent on cellular translation efficiency that varies in different cell types and conditions. Thus, the use of mRNA does not permit direct control of the engineering procedure.

An alternative to the use of nucleic acid as source of transposase is to directly provide the cells with the SB protein. Delivery of the SB transposase is highly desirable to achieve tighter efficiency/temporal control of transposition-based cell engineering and improve safety, in particular for gene therapy applications. However, the intrinsic instability of the SB protein *in vitro* has limited its large-scale recombinant production, and thus prohibited cellular delivery to date. The specific objective of this part of my thesis was to overcome the difficulties related to SB transposase production, and to develop a novel SB system based on the application of the recombinant protein.

3.2 Results –The high solubility SB (hsSB) variant

In this part of my work (sections 3.2 and 3.3), I set out to develop a SB-based genetic engineering strategy based on direct delivery of the transposase protein to target cells. To this end, I first aimed to produce the SB transposase in suitable quality and quantity to enable efficient protein delivery (section 3.2).

3.2.1 Identification of the hsSB variant by structure-based design

The successful application of currently available protein delivery strategies [e.g. based on cell-penetrating peptides, lipidic and polycationic carriers, or electroporation (Fu *et al.*, 2014)] depends greatly on specific features of the protein sample. The protein has to be provided in high amounts and must be stable in the low volumes and specific buffers used for the delivery procedure. In addition, high stability of the protein at physiological temperature is strictly required, so that the delivered molecule is fully folded and functional in the target cells.

In this respect, the SB100X transposase (referred to simply as "SB" in this thesis) is not a promising candidate for protein delivery. It provides only modest yields after purification and it is unstable in low salt containing buffers (that are required for example for electroporation), impairing protein delivery attempts. To overcome these challenges, I sought to generate a novel SB variant with improved protein solubility and stability. To this aim, I employed a rational design approach based on the crystal structure of the SB catalytic domain and on a homology model of its TCC (Voigt *et al.*, 2016).

I identified and introduced specific amino acid substitutions into SB to increase the hydrophilic character of the protein surface (Figure 3-3). First, I substituted isoleucine 212 in the target DNA binding surface with a serine. This mutation was previously associated with hyperactive phenotype (Voigt *et al.*, 2016) (Figure 3-3 A) and also increased protein solubility *in vitro*. Then, I proceeded with mutagenesis of all cysteines (C176, C197, C304 and C316) into serines one by one. This approach can help reduce protein aggregation as shown for other systems (Avramopoulou *et al.*, 2004; Cozzolino *et al.*, 2008; Slusarczyk *et al.*, 2000; Xie *et al.*, 2009).

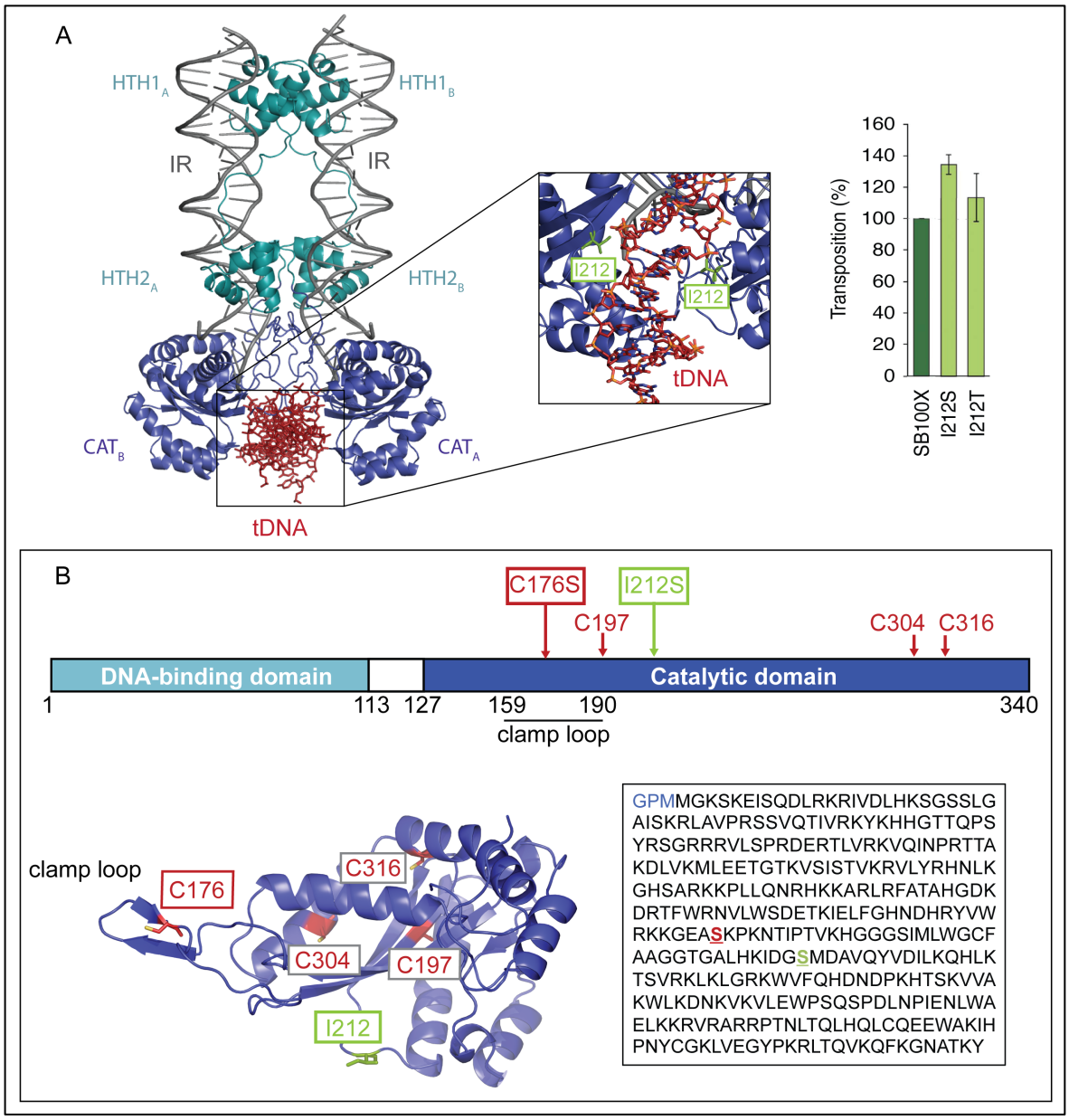


Figure 3-3: Rational mutagenesis of the SB transposase. (A) Hyperactive I212 mutants. Left: model of the SB TCC [target capture complex (Voigt *et al.*, 2016)], where two SB molecules (blue) contact the transposon end sequences (IR- inverted repeats, grey) and target DNA (tDNA, red). Insert shows that residue I212 is located on the target-binding interface. Right. I212 mutants are hyperactive in *in vivo* transposition assays (Voigt *et al.*, 2016). (B) Solubilizing mutations in SB. Schematic representation of SB's domain composition (top) and crystal structure of the catalytic domain (Voigt *et al.*, 2016) (bottom left) with the analyzed amino acid substitutions marked. Bottom right: amino acid sequence of the full length hsSB transposase variant. Bold underlined characters in red and green indicate serines substituting C176 and I212, respectively. Residues colored in blue have been introduced at the N-terminus for cloning.

I found that the C197S, C304S and C316S mutations largely compromised protein solubility, consistent with their role in assembling the core fold of the catalytic domain (Figure 3-3 B). In turn, mutation of C176 that is located in the surface-exposed flexible

clamp loop, in combination with the I212S substitution, generated a remarkably highly soluble SB variant (hsSB), as described in detail in the following paragraphs.

3.2.2 Functional characterization of the hsSB variant

I overexpressed the hsSB variant in *E. coli* fused to N-terminal purification and solubility tags and purified it by Ni-affinity chromatography, followed by tag removal and SEC (Figure 3-4 A and B). SDS-PAGE analysis of purified recombinant hsSB showed that the protein variant is highly pure (Figure 3-4 A). SEC analysis also revealed a single symmetric peak, showing that the recombinant hsSB is highly homogenous chemically and conformationally. hsSB was produced at significantly higher yields compared to the SB transposase (Figure 3-4 B), yielding up to 5 mg of 99.9% pure transposase protein from 1liter *E. coli* culture.

Following purification, I concentrated the purified hsSB protein using ultrafiltration devices. Remarkably, the hsSB protein could be concentrated up to 20 mg/ml final concentration, whereas SB precipitates at concentrations higher than 7 mg/ml (Figure 3-4 C). Moreover, hsSB is highly soluble in low salt containing buffers, including the R buffer used for electroporation with the Neon® transfection system (Figure 3-4 C). While SB heavily precipitates in these conditions, the vast majority of hsSB remains soluble upon concentration in R buffer.

I then tested the stability of hsSB at physiological temperature in which protein delivery is performed. SDS-PAGE analysis of purified SB proteins upon incubation at 37 °C revealed that hsSB is less prone to degradation than SB (Figure 3-4 D, left). SB exhibits degradation even upon short incubation at physiological temperature, whereas hsSB does not show significant degradation even after 18 hours. The improved solubility of hsSB does not seem related to decreased protein oxidation, as hsSB and SB were similarly affected by oxidation when incubated at 37 °C (Figure 3-4 D, right).

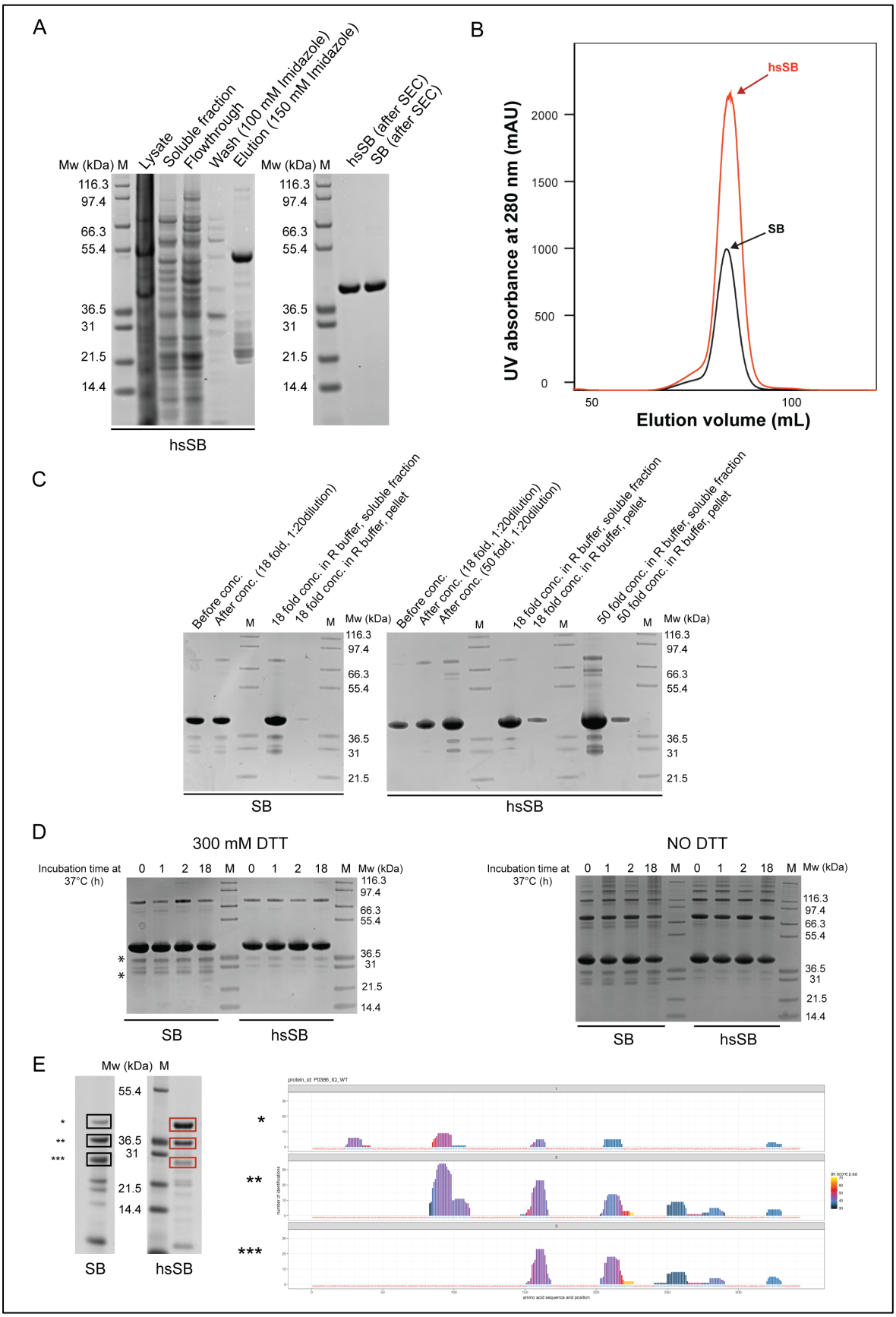


Figure 3-4: Purification and characterization of the hsSB protein variant. (A) SDS-PAGE analysis of hsSB purification. hsSB is highly pure after tag removal and size exclusion chromatography (SEC). (B) Size exclusion chromatogram (column Superdex 200, 16/600) showing that recombinant hsSB (red curve) is

homogenous and yields at significantly higher level compared to SB (black curve). (C) SDS-PAGE analysis of purified SB proteins after concentration. hsSB can be concentrated up to 50 fold (corresponding to 20 mg/ml), whereas SB undergoes precipitation at concentrations higher than 7 mg/ml. hsSB is highly soluble in the low salt R buffer, even at high protein concentration. (D) SDS-PAGE analysis of purified SB protein samples upon incubation at 37 °C with (left, 300 mM DTT) or without DTT (right, NO DTT). SB exhibits degradation (major degradation products marked with asterisks) even upon short incubation at physiological temperature, while hsSB does not. Oxidation patterns are similar for both proteins regardless of DTT concentration. (E) Stability of SB variants upon freezing. Left: SDS-PAGE analysis of purified SB proteins following long-term storage at -80 °C. SB undergoes significant degradation, while hsSB does not. Right: mass spectrometry analysis of the product bands (black boxes and asterisks on the left) confirms degradation.

Remarkably, I was able to show that hsSB is more stable than SB also during storage at low temperatures. SDS-PAGE analysis of SB proteins subjected to long-term storage (several months) at -80 °C revealed that hsSB is well preserved after freezing and thawing, whereas SB undergoes significant degradation (Figure 3-4 E, left) as confirmed by mass spectrometry analysis (Figure 3-4 E, right). Taken together, my data shows that hsSB is more resistant to degradation at different temperatures compared to SB, indicating increased conformational integrity and stability.

To further investigate hsSB's reduced susceptibility to degradation, I performed limited proteolysis of the SB proteins. Limited proteolysis is a simple biochemical method that can be used to analyse protein structure and conformational flexibility. Upon incubation of the target protein with a limited amount of protease, this cleaves at exposed flexible regions. Detected cleavage products allow to map stable subdomains and flexible linker regions. Among the different enzymes used for limited proteolysis, trypsin is widely applied, and breaks the peptide bond at C-terminus of basic amino acids such as arginine and lysine (Fontana *et al.*, 2004). With this assay, I observed similar trypsin-mediated degradation patterns for hsSB and SB, indicating that the overall protein fold is conserved between the two variants (Figure 3-5 A).

To confirm this hypothesis, I investigated the folding and secondary structure of the SB proteins by circular dichroism (CD) spectroscopy. The method is based on the difference in the absorption of left-handed and right-handed circularly polarised light by chiral chromophores, such as the amides of the polypeptide backbone in proteins. Secondary structure also imparts a distinct CD feature to proteins, so that all proteins have CD spectral signatures that are representative of their individual structures (Greenfield, 2006).

CD measurements of SB and hsSB proteins in close to physiological (1x PBS, 200 mM NaCl, pH 7.5) buffer conditions confirmed that the folding of hsSB is indistinguishable from the folding of SB (Figure 3-5 B, left). Thus, mutation of residues 176 and 212 does not affect the overall folding of the protein and hsSB's improved solubility and stability are not due to major changes in the protein structure. However, collecting the CD signal of the two proteins as a function of temperature, revealed that hsSB is partially folded even at 95 °C and thus is significantly more thermostable than SB (Figure 3-5 B, right). Considering that electroporation heats the sample, hsSB provides a distinct advantage over SB for direct protein delivery by this method.

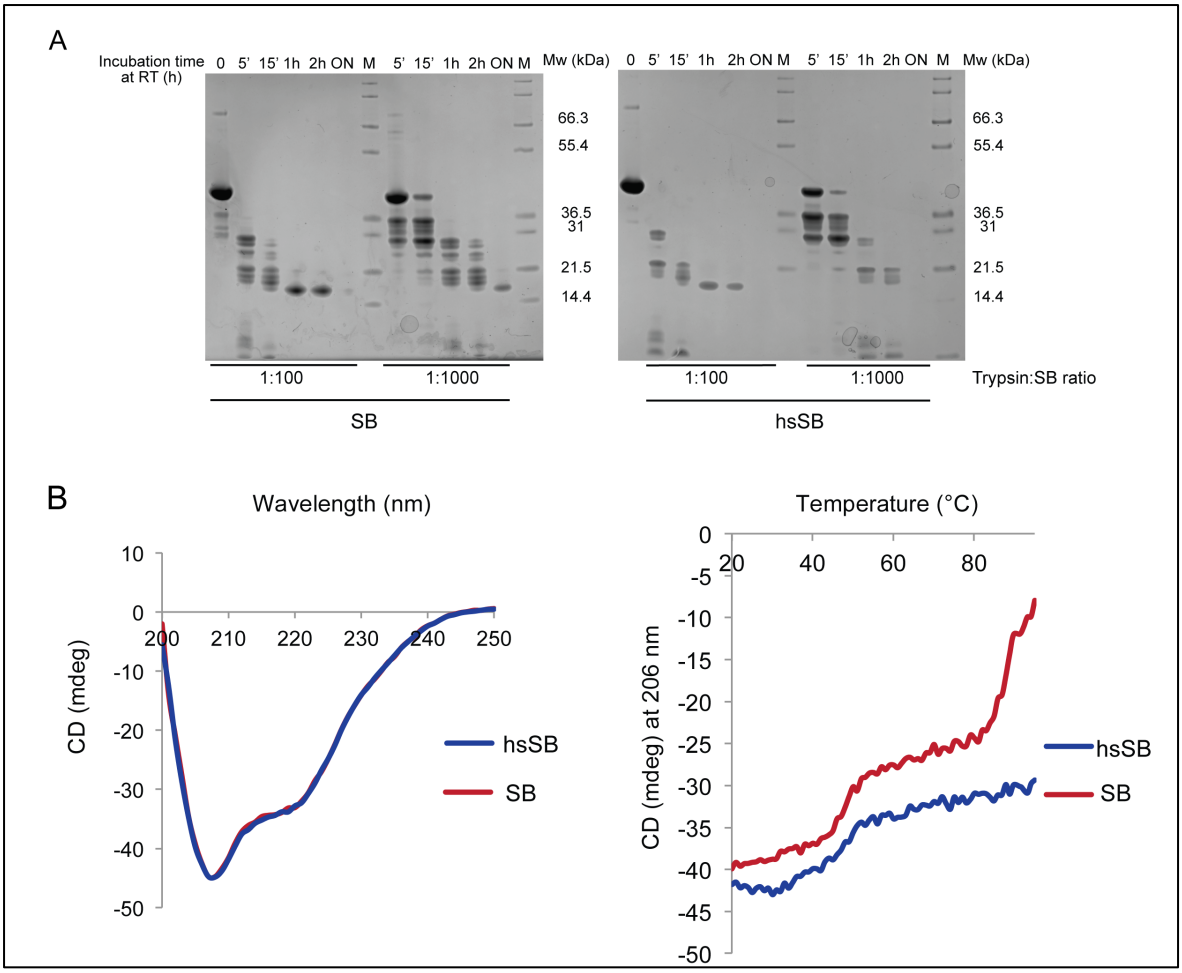


Figure 3-5: Folding and thermostability of SB proteins. (A) SDS-PAGE analysis of SB proteins upon treatment with trypsin for indicated times at room temperature (RT). ON: overnight. (B) Circular dichroism measurements of SB proteins in close to physiological (1x PBS, 200 mM NaCl, pH 7.5) buffer condition. CD spectra show that hsSB has the same fold as SB (left panel). However, hsSB is significantly more thermostable, as it still does not completely unfold at 95 °C (CD values at 206 nm plotted against temperature, right panel).

After establishing the improved solubility and stability of the hsSB variant, I tested its transposition activity *in vitro*. Using the cleavage and integration assays described in chapters 2.3.2 and 2.3.3, respectively, I confirmed that the hsSB transposase cleaves (Figure 3-6 A) and integrates transposon end DNA (Figure 3-6 B) *in vitro* with similar efficiencies as SB. hsSB was also able to mediate transposition in living HeLa cells [when provided from an expression plasmid, assays performed by Cecilia Zuliani, Lab Manager, Barabas Laboratory, EMBL Heidelberg] with identical efficiency and specificity as SB (Figure 3-6 C). This shows that the amino acid substitutions introduced in the hsSB variant did not affect transposition activity.

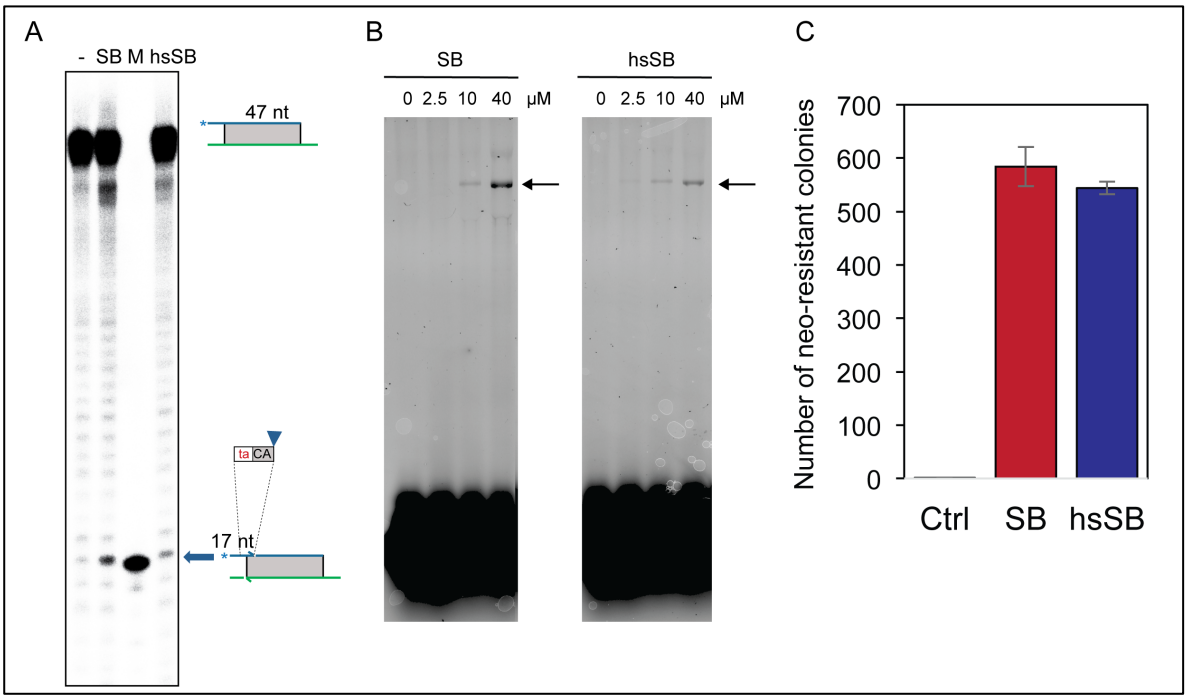


Figure 3-6: Characterization of hsSB's transposition activity. (A) *In vitro* cleavage assays showing that hsSB cleaves Lo sequences specifically. 12% Urea PAGE resolving ^{32}P-labelled Lo DNA substrates (see assays in Figure 2-22) and products after incubation with SB or hsSB (in a 50:1 protein to DNA molar ratio) or no protein (-). Blue arrow indicates the specific cleavage product (17 nt on NTS). Marker (M) is 17 nt long. (B) *In vitro* integration assay showing that hsSB can integrate 2 nt staggered pre-cleaved Lo DNA into a target plasmid (see assays in Figure 2-25). Products are detected on a native agarose gel. Bands indicated by arrows correspond to concerted integration products (DEJ, Double-End Joining products). (C) Transposition assays (Ivics *et al.*, 1997) in HeLa cells demonstrate identical activity of SB and hsSB (provided on an expression plasmid). Mean values of three independent experiments are shown and error bars represent standard error of the mean. Difference between SB and hsSB samples is not statistically significant as extimated by a two-sample *t*-test (*P* value >0.05).

3.3 Results – Direct transfection of the hsSB protein for mammalian cell engineering

The studies described in the previous paragraphs led to the identification of the hsSB transposase variant (section 3.2). hsSB is characterized with extraordinary solubility, stability and purification yields and is fully active in all steps of transposition, both *in vitro* and in HeLa cells. Following up on this, we aimed to explore the possibility to directly deliver this variant into mammalian cells for transposition-mediated genome engineering. The experiments described below in this section (3.3) have been designed and performed together with Cecilia Zuliani (Lab Manager, Barabas Laboratory, EMBL Heidelberg).

3.3.1 Genome engineering of HeLa cells by hsSB delivery

We first focused on delivering the purified hsSB transposase into HeLa cells. Several methods have been described in literature for protein delivery into mammalian cells (Fu *et al.*, 2014). Electroporation currently provides highest success, and has been successfully used for direct delivery of genome editing nucleases [as TALENs (J. Liu *et al.*, 2014; Ru *et al.*, 2013) and Cas9 (Kim *et al.*, 2014; Ramakrishna *et al.*, 2014) proteins]. Thus, we tested the delivery of the hsSB protein via electroporation using the Neon® system.

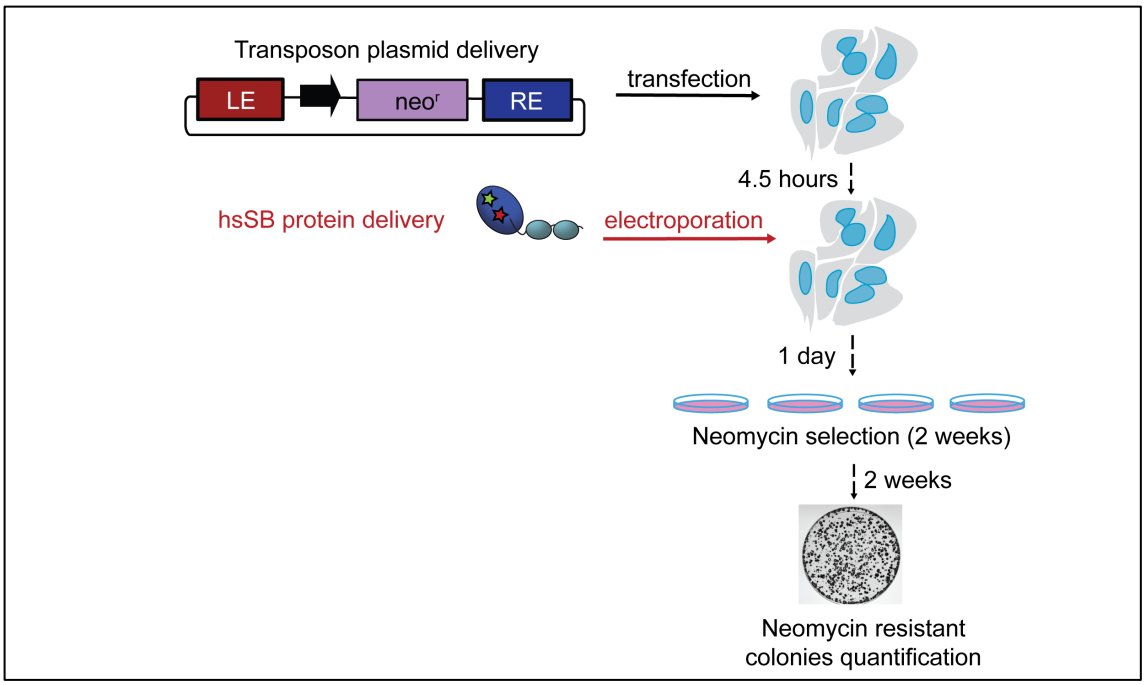


Figure 3-7: Schematic representation of the hsSB delivery and testing procedure using the neomycin-based reporter system. LE and RE indicate transposon end sequences. LE: left end; RE: right end; black arrow: promoter; neo^r: neomycin-resistance cassette. hsSB is depicted with its domain representation. Green and red stars indicate the C176S and I212S mutations, respectively.

In order to probe successful protein delivery into HeLa cells, we made use of a neomycin-based reporter system [protocols adapted from (Ivics *et al.*, 1997)]. For this, we transfected HeLa cells with a neomycin-resistance (neo^r) cassette-carrying transposon plasmid. Upon successful transposase delivery, the neo^r-transposon can undergo genomic integration and provides long-lasting drug resistance in the modified cells. Following drug selection, the number of neomycin-resistant colonies provides a measure of transposition efficiency (Figure 3-7).

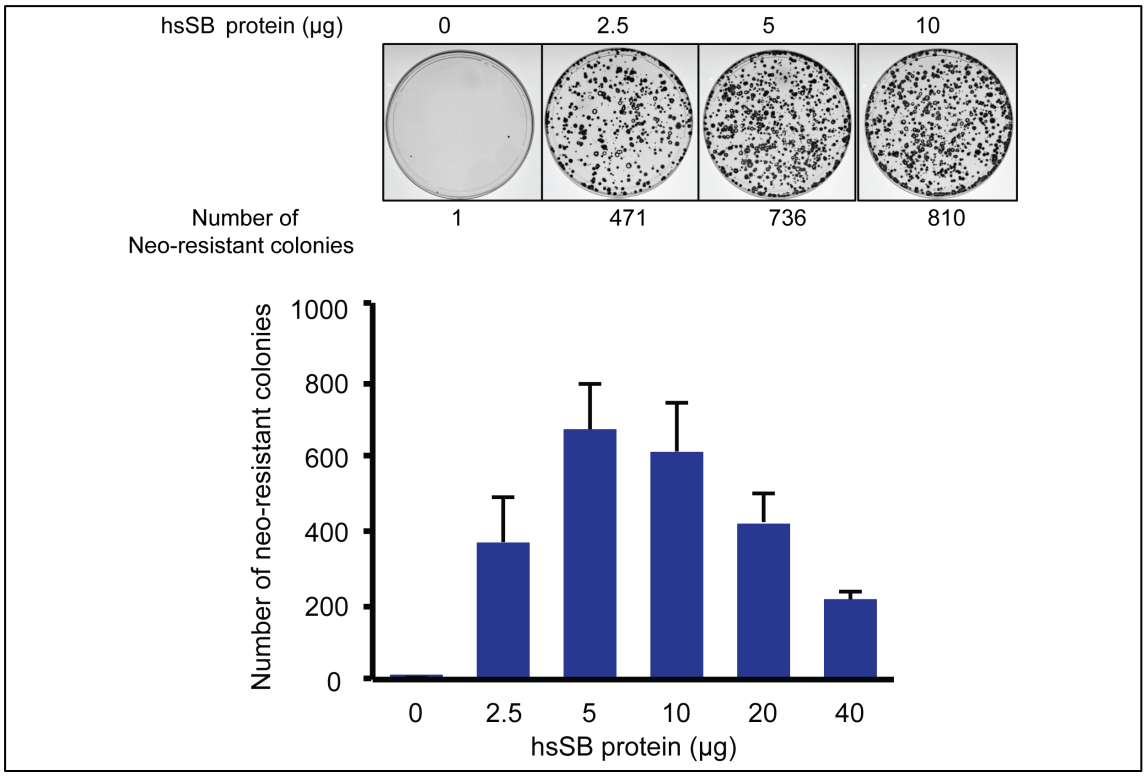


Figure 3-8: Neomycin-based transposition assay in HeLa cells using direct delivery of hsSB protein. Top: pictures of neomycin (neo) resistant colonies obtained in a single experiment. Bottom: graph showing the number of neomycin-resistant colonies obtained in the assay. Mean values of three independent experiments are shown and error bars represent the standard error.

No transposition was detected when the transposon donor plasmid was delivered alone. Delivery of hsSB protein together with transposon donor plasmid by electroporation (in a single electroporation step) also did not result in detectable transposition (data not shown). In turn, we observed efficient transposition using a sequential delivery protocol, where the donor plasmid was first transfected using a cationic transfection reagent (JetPei®), followed by hsSB protein delivery via electroporation (Figure 3-7 and Figure 3-8). Using this strategy, we obtained robust and dose-dependent transposition, as demonstrated by the

gradually increasing number of neomycin resistant HeLa cell colonies with increasing transposase concentration. Maximum activity was achieved using 5-10 µg protein (Figure 3-8).

To better quantify hsSB-mediated transposition, we then took advantage of a fluorescent reporter system (Izsvak *et al.*, 2009) (Figure 3-9). We first transfected HeLa cells with a Venus gene-carrying transposon plasmid (using JetPei®) and subsequently delivered the hsSB protein by electroporation. Green fluorescent cells (expressing the fluorescent Venus protein) were selected and sorted by flow cytometry two days post-transfection to obtain a pure population of cells containing the transposon plasmid. These cells are able to undergo cargo gene insertion into the genome in the presence of the hsSB protein. Three weeks later, we quantified by flow cytometric analysis the number of green fluorescent cells that stably expressed Venus as a consequence of genomic insertion by hsSB (Figure 3-9). By this method, we could precisely determine the efficiency of hsSB's delivery and transposition, independently from the efficiency of transposon plasmid transfection.

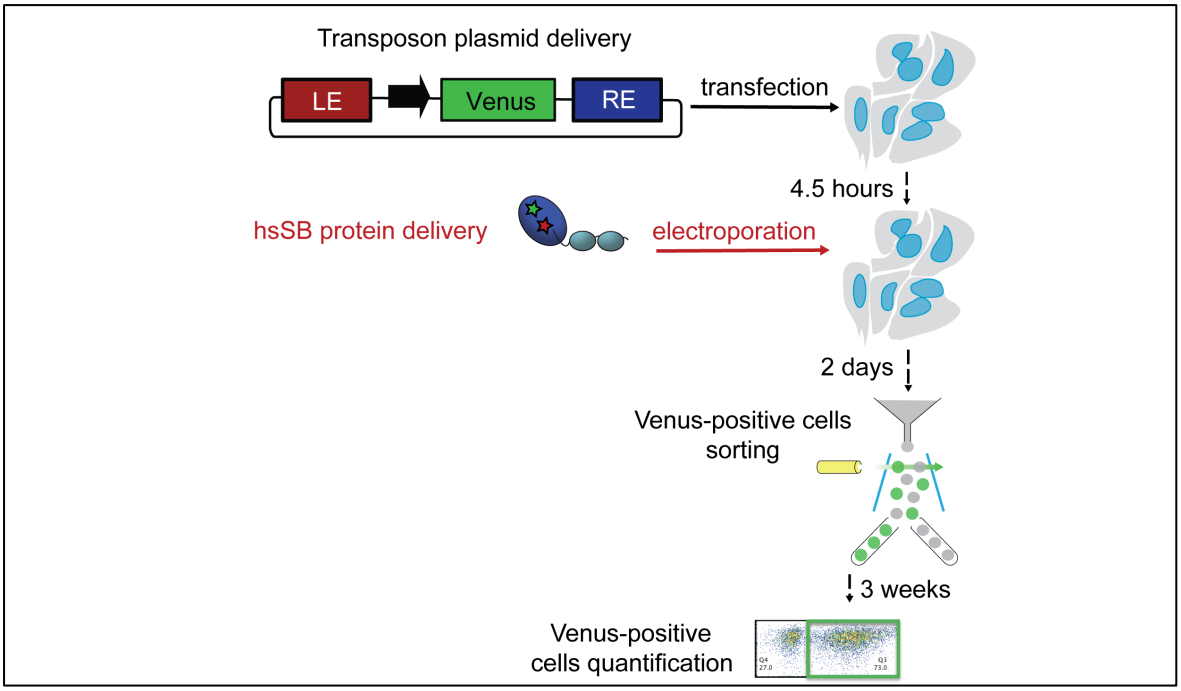


Figure 3-9: Schematic representation of hsSB delivery and quantification using the fluorescence-based reporter system. LE and RE indicate transposon end sequences. LE: left end; RE: right end; black arrow: promoter; Venus: Venus fluorescent protein-encoding gene. hsSB is depicted with its domain representation. Green and red stars indicate the C176S and I212S mutations, respectively.

Two days post-transfection, we observed similar amounts of fluorescent cells among different treated cell populations, irrespective of the amount of hsSB protein used (Figure

3-10, top row). This shows that donor plasmid transfection occurred at roughly equal levels in all successfully transfected cells, resulting in similar numbers of Venus expressing cells. Venus positive cells were then cultured for three weeks and probed for successful transposition. After three weeks, the plasmid-carried Venus gene is lost from the cells, unless it is integrated in the HeLa cell genome by hsSB. Consistently, we barely detected fluorescent cells (0.19%) among those that were electroporated without hsSB. In turn, we detected a clear dose-dependent increase in the percentage of fluorescent cells with electroporation of the hsSB transposase, reflecting efficient transposition. Maximum efficiency (42%) was achieved with 10-20 µg of protein (Figure 3-10, bottom row).

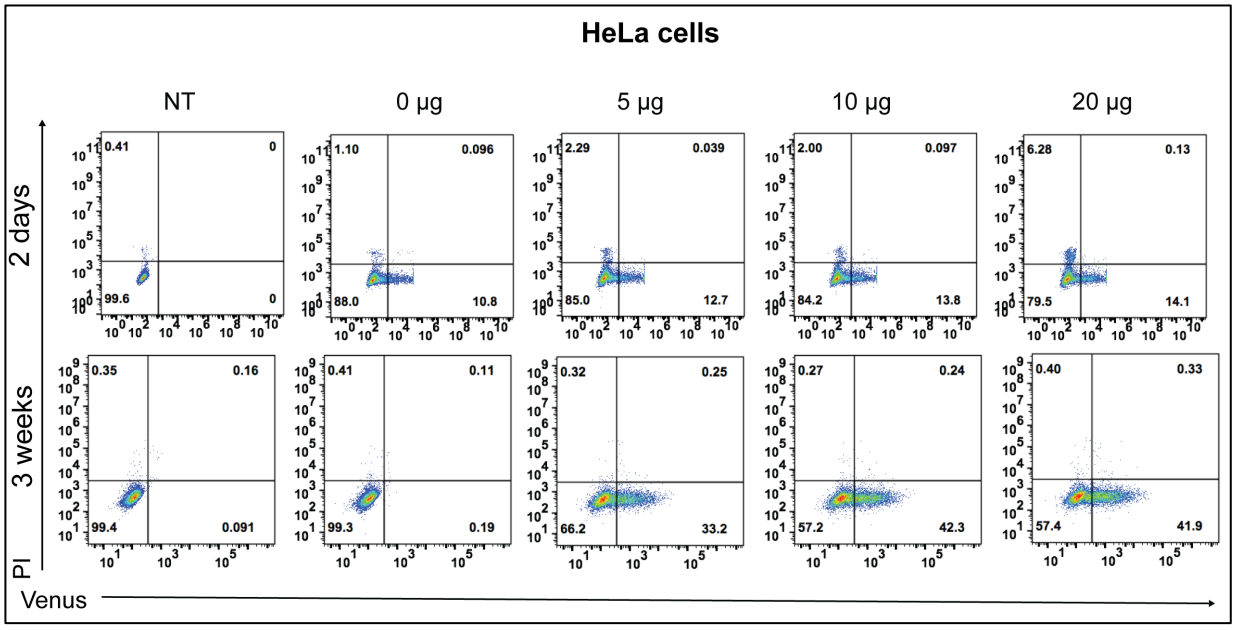


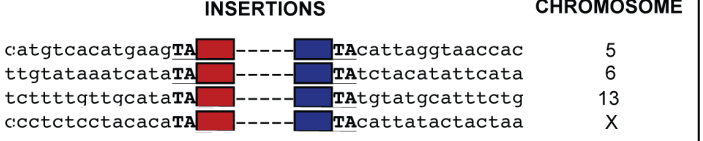
Figure 3-10: Representative graphs of flow cytometric analysis of HeLa cells transfected with Venus-carrying transposon plasmid and electroporated with increasing amounts of hsSB (hsSB concetration indicated above each graph). Venus-positive cells are sorted 2 days post-transfection (top row), so as to select the population bearing the transposon plasmid, and 3 weeks post-transfection (bottow row), so as to quantify the percentage of cells that underwent transposition. Y axis: propidium iodide (PI) staining to exclude dead cells. X-axis: green fluorescence from Venus expression. NT: non-transfected with transposon plasmid.

3.3.2 Characterization of the engineered HeLa cells

Once we achieved efficient and stable engineering of HeLa cells by hsSB delivery, we focused on characterizing the cell product obtained by our novel method.

We first identified the positions where the neomycin-resistance gene was inserted in the HeLa cell chromosomes. For this, we performed sequence analysis of isolated neo^r-positive cells with the help of the EMBL Genomics Core facility (T. Rausch and V. Benes,

EMBL Heidelberg). This analysis showed that insertions occurred at TA dinucleotides, following the typical integration pattern of SB transposition (Figure 3-11). The identified insertions mapped to different positions in the HeLa cell genome (Figure 3-11), indicating a close-to-random integration pattern at the genome level, as previously described for the SB transposon in general (Turchiano et al., 2014).



Single insertions sequence	Chromosome
tgtatatatatata TA CAGTTGAAGTC	1
gacacatacataca TA CAGTTGAAGTC	9
ctgttggatgcctc TA CAGTTGAAGTC	7
gatatatacatatg TA CAGTTGAAGTC	17
gtactgagtgtatg TA CAGTTGAAGTC	9
ctttcaggaacaaa TA CAGTTGAAGTC	17
tcaacttcagaaatg TA CAGTTGAAGTC	1
ggacacatacataca TA CAGTTGAAGTC	9
gtactgagtgtatg TA CAGTTGAAGTC	9
tttcaggaacaaa TA CAGTTGAAGTC	17
actctcctatgata TA CAGTTGAAGTC	2
actctcctatgata TA CAGTTGAAGTC	2
aataatgctagtta TA CAGTTGAAGTC	19
atggcgagttaaca TA CAGTTGAAGTC	12
tattccatggcata TA CAGTTGAAGTC	6
tatagctaacaata TA CAGTTGAAGTC	1
gcaagtcctgtca TA CAGTTGAAGTC	12
ttaaatggaataat TA CAGTTGAAGTC	12
gacacatacataca TA CAGTTGAAGTC	9
aaagcaatagcaca TA CAGTTGAAGTC	17
ttgtataaatcata TA CAGTTGAAGTC	6
tatagctaacaata TA CAGTTGAAGTC	1
cctaatcatctact TA CAGTTGAAGTC	1
ttgtataaatcata TA CAGTTGAAGTC	6
catgtcacatgaag TA CAGTTGAAGTC	5
agtgaggtttaaca TA CAGTTGAAGTC	22
ctattttggaaaca TA CAGTTGAAGTC	1

Figure 3-11: Insertion sites identified by sequence analysis of isolated neomycin resistant HeLa cell clones. Insertions of both *SB* transposon ends correctly occur at TA dinucleotides (underlined). Red: left end. Blue: right end.

Next, we investigated the time frame in which the hsSB transposase is present in the HeLa cells, when it is (a) provided as protein or (b) expressed from plasmid DNA. By Western blot analysis, we monitored at different time points the presence of hsSB protein in HeLa cells transfected with transposon donor plasmid together with (a) 10 µg of hsSB protein or (b) 500 ng of hsSB-encoding plasmid.

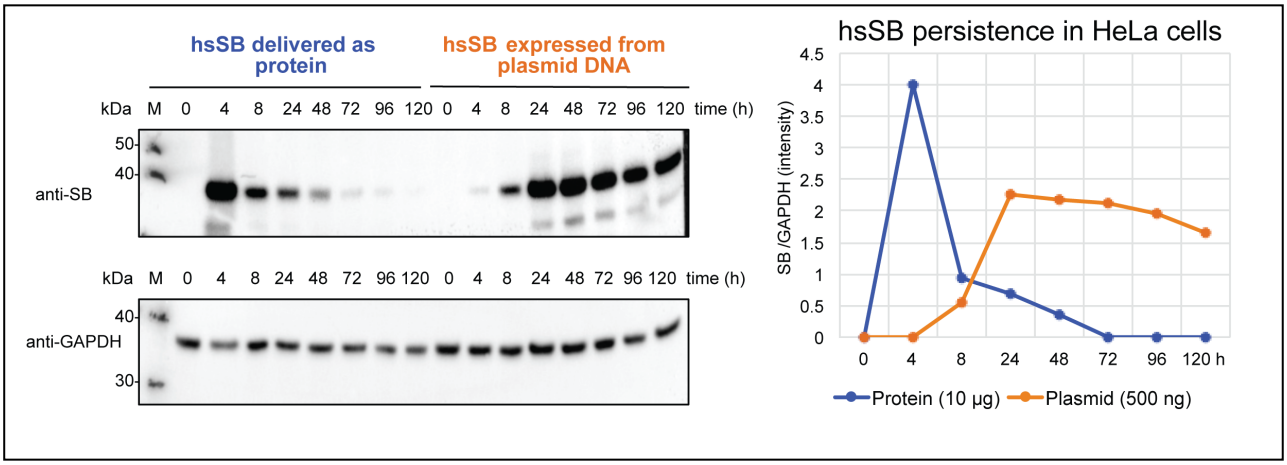


Figure 3-12: Retention of hsSB delivered into HeLa cells as protein or expressed from plasmid DNA. Western blot analysis shows almost complete loss of directly delivered hsSB protein 48 hours after electroporation, whereas cells transfected with hsSB expression plasmids produce high level of protein continuously from 24 hours to over 5 days after transfection. Western blot was performed on lysate from HeLa cells that were transfected with 500 ng of neomycin-carrying transposon plasmid and electroporated with 10 µg hsSB protein or transfected with 500 ng hsSB expression plasmid. Samples were taken at the indicated time points, and 20 µg of the total cell lysate was separated by electrophoresis and transferred to a nitrocellulose membrane. hsSB was detected with anti-SB antibody. The internal loading control was glyceraldehyde 3-phosphate dehydrogenase (GAPDH) detected with anti-GAPDH antibody. Right: plot showing the intensity of the Western blot bands at the analyzed time points.

We observed substantial loss of directly delivered hsSB protein 24 hours after it was electroporated to the cells. In turn, cells transfected with hsSB expression plasmids produced high level of protein continuously from 24 hours to over 5 days post transfection (Figure 3-12). This shows that direct delivery of hsSB protein drastically reduces the retention of the SB transposase in the cells and, ultimately, the time span of active transposition.

Finally, in collaboration with C. Miskey (Ivics Laboratory, Paul-Ehrlich-Institut, Langen, Germany), we analyzed the number of insertions in the HeLa cell genome when hsSB is provided (a) as protein or (b) as expression plasmid. For this, we first performed

transposition experiments using different amounts of hsSB or hsSB-encoding plasmid and selected positive cells using the neomycin-based assay as described in section 3.3.1. For each condition, we extracted genomic DNA from a pool of neomycin-resistant HeLa cells, and subjected it to insertion copy number analysis in the Ivics Laboratory. We found that the number of insertions gradually increased by electroporation of increasing amounts of hsSB protein into the HeLa cells (Figure 3-13 A, courtesy of C. Miskey). Such direct correlation was not observed when providing the protein from an expression plasmid. Moreover, direct protein delivery generally generated fewer insertions (on average) per cell and a reduced variation of integrated copy numbers compared to the plasmid-based strategies.

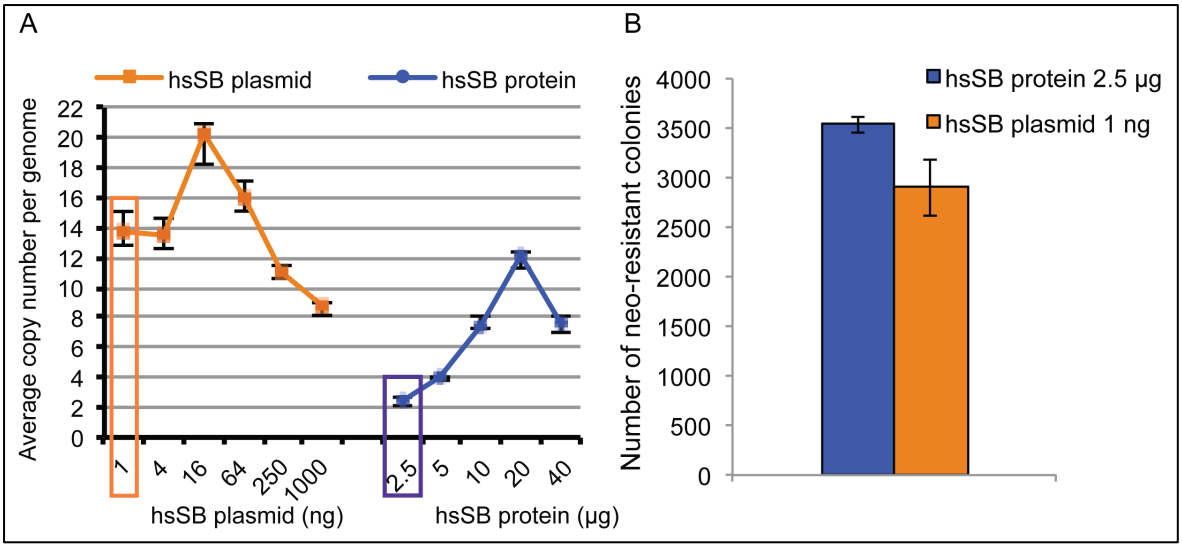


Figure 3-13: Comparative analysis of the number of SB-mediated insertions. (A) Average number of transposon insertions in HeLa cells, resulting from hsSB delivered as protein or expressed from plasmids. Copy number is measured by digital droplet PCR of HeLa cell genomic DNA (C. Miskey, Ivics Laboratory, Paul-Ehrlich-Institut, Langen, Germany). (B) Transposition assay in HeLa cells showing comparable transgenesis rates in cells where hsSB was directly delivered as protein (2.5 µg) or expressed from plasmids (1 ng). Mean values of three independent experiments are shown and error bars represent the standard deviation. Difference between hsSB protein and plasmid samples is not statistically significant as extimated by a two-sample *t*-test (*P* value >0.05).

In order to properly compare insertion copy numbers generated by the two methods, we identified the specific amounts of (a) the hsSB protein and (b) the hsSB-encoding plasmid (2.5 µg and 1 ng respectively) that produce similar transposition rates (i.e. similar number of engineered cells as quantified by neomycin-based transposition assays in HeLa cells) (Figure 3-13 B). We observed that at identical transposition efficiencies, hsSB protein delivery results in 7-fold less insertions than plasmid delivery (Figure 3-13). This shows

that the protein-based strategy creates fewer modifications and hence less perturbation in the HeLa cell genome.

3.3.3 Genome engineering of CHO cells by hsSB delivery

To explore the range of applicability of our technique, we then used hsSB for genome engineering of CHO cells. CHO cells are widely used in biological and medical research as well as for commercial production of therapeutic proteins. CHO cells are in fact the most commonly used mammalian hosts for industrial production of recombinant protein therapeutics (Omasa *et al.*, 2010). Thus, strategies for efficient and stable genome engineering of CHO cells are in high demand.

CHO cell lines used in this study are intrinsically resistant to neomycin. Therefore, in order to test hsSB delivery and transposition, we exclusively used the fluorescence-based reporter assay (previously established for HeLa cells, see section 3.3.1). Briefly, we applied a sequential delivery protocol, where CHO cells were first transfected with the Venus-carrying transposon plasmid using the non-liposomal transfection reagent FuGENE® and consequentially electroporated with the hsSB protein.

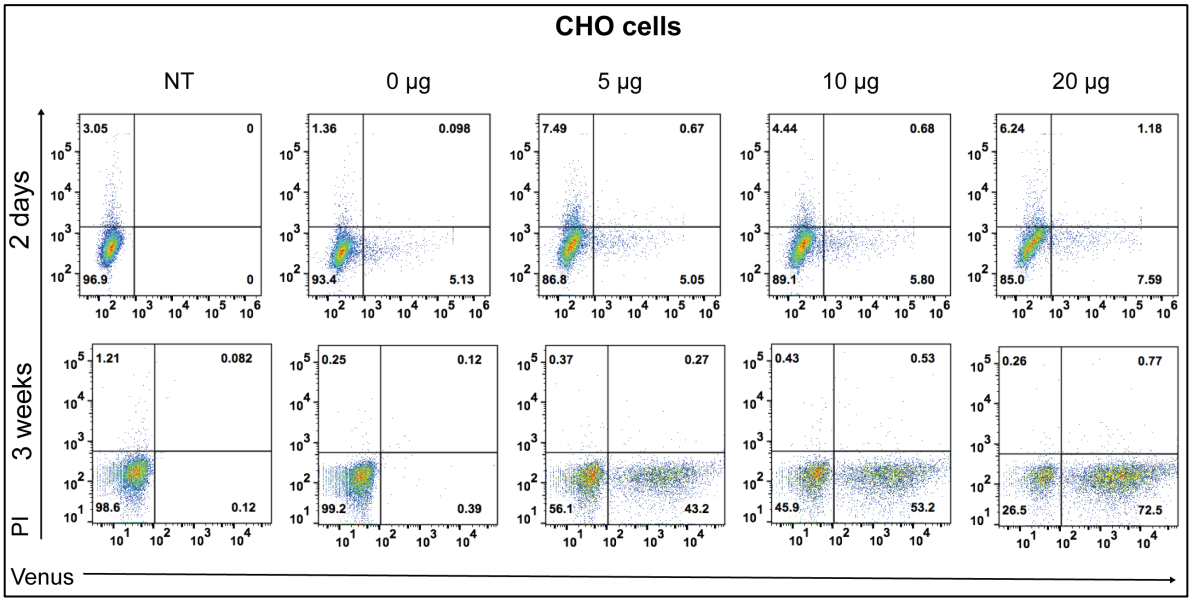


Figure 3-14: Representative flow cytometric graphs of Chinese Hamster Ovary (CHO) cells that were transfected with Venus-carrying transposon plasmid and electroporated with hsSB (amount indicated above each graph). Venus-positive cells are sorted 2 days post-transfection (top row), so as to select the population bearing the transposon plasmid, and 3 weeks post-transfection (bottow row), so as to quantify the percentage of cells that underwent transposition. Y axis: propidium iodide (PI) staining to exclude dead cells. X-axis: green fluorescence from Venus. NT: non-transfected with transposon plasmid.

Cytofluorimetric analysis performed after two days revealed relatively low efficiency of transposon plasmid transfection (Figure 3-14, top row). Nonetheless, three weeks post-transfection we observed a strong dose-dependent increase in green fluorescent CHO cells to a maximum of the 73% with 20 µg of hsSB protein (Figure 3-14, bottom row), revealing a remarkably high efficiency of the delivered hsSB in CHO cell engineering.

3.3.4 Genome engineering of mESCs and human primary cells by hsSB delivery

Next, we tested the efficacy of our hsSB-based engineering procedure for the genetic modification of stem cells. Stem cells are extensively used in research to study processes like differentiation (e.g. haematopoiesis) and development and as *in vitro* systems for toxicology testing (Trounson *et al.*, 2016). The efficient engineering of stem cells, and in particular of embryonic stem cells, is also critical to establish animal model lines for laboratory usage. In addition, due to their ability to differentiate into any cell type, stem cells and progenitor cells exhibit high potential for regerative medicine.

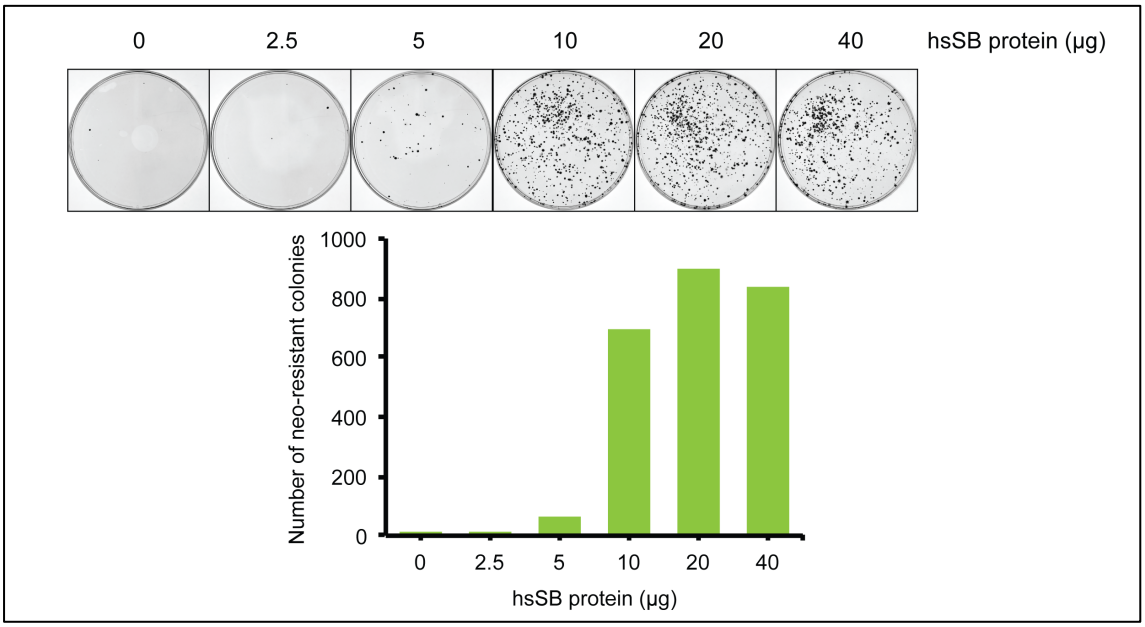


Figure 3-15: Representative neomycin-based transposition assay in mouse embryonic stem cells (mESCs), performed by direct delivery of hsSB protein. Top: pictures of mESCs neomycin-resistant colonies. Bottom: graphic showing the number of neomycin-resistant mESCs colonies obtained in the experiment.

In order to provide proof-of-concept for the use of hsSB delivery for stem cell engineering, we applied our protocol to mouse embryonic stem cells (mESCs). We first tested hsSB in these cells using the neomycin-based transposition assay (see section 3.3.1). Here, mESCs were transfected with the neomycin resistance-carrying transposon plasmid using the

FuGENE® reagent and then electroporated with the hsSB protein. Upon neomycin selection, we observed robust and dose-dependent transposition, as estimated by the number of neomycin resistant mESC colonies, with maximum activity achieved using 20 µg protein (Figure 3-15). To quantify transposition efficiency in a more accurate manner, we then applied the fluorescence-based reporter system (Figure 3-9, see also section 3.3.1 for the assay design).

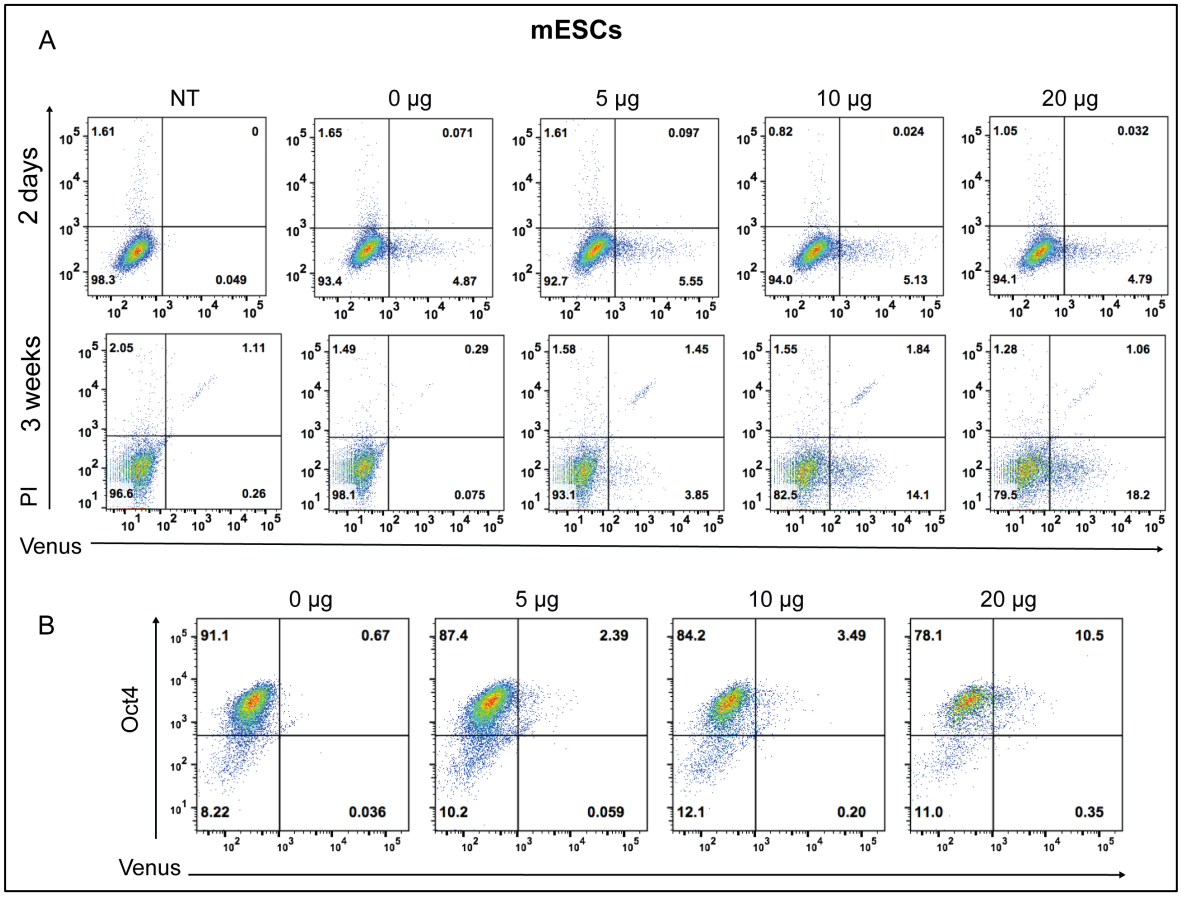


Figure 3-16: Genetic engineering of mESCs using the fluorescence-based reporter system and characterization of the modified cells. (A) Representative flow cytometric graphs of mESCs that were transfected with Venus-carrying transposon plasmid and electroporated with different amounts of hsSB (indicated on top of each graph). Venus-positive cells are sorted 2 days post-transfection (top row), so as to select the population bearing the transposon plasmid, and 3 weeks post-transfection (bottow row), so as to quantify the percentage of cells that underwent transposition. Y axis: propidium iodide (PI) staining to exclude dead cells. X-axis: green fluorescence from Venus. NT: non-transfected with transposon plasmid. (B) Flow cytometric analysis of engineered mESCs; Oct4 staining confirms that the large majority of engineered mESCs retain pluripotency.

The same delivery procedure was used as in the neomycin-based assay (transposon plasmid transfection by FuGENE®, followed by hsSB protein electroporation), except that

the transposon plasmid contained the Venus-encoding reporter gene. Two days after transfection, we observed comparable transposon plasmid transfection efficiencies as for the CHO cells (circa 5%) (Figure 3-16 A, top row). Three weeks post-transfection, we then identified up to 18% of genetically modified mESCs (green fluorescent cells) using 20 µg hsSB (Figure 3-16 A, bottom row). The engineered cells fully maintained pluripotency after treatment, as confirmed by immunofluorescent detection of the Oct4 self-renewal marker (Zeineddine *et al.*, 2014) (Figure 3-16 B).

Collectively, our data show that the hsSB transposase is effectively transfected into a variety of cell lines from different organisms - including HeLa cells, CHO cells, and mESCs -, and allows for stable and efficient genetic engineering of these cell types (as summarized in Figure 3-17). Our findings provide proof-of-concept for the use of hsSB for stem cell engineering (Figure 3-15, Figure 3-16 and Figure 3-17) that is quite remarkable since stem cells are generally hard-to-transfect and are considered refractory to genetic modification by nature.

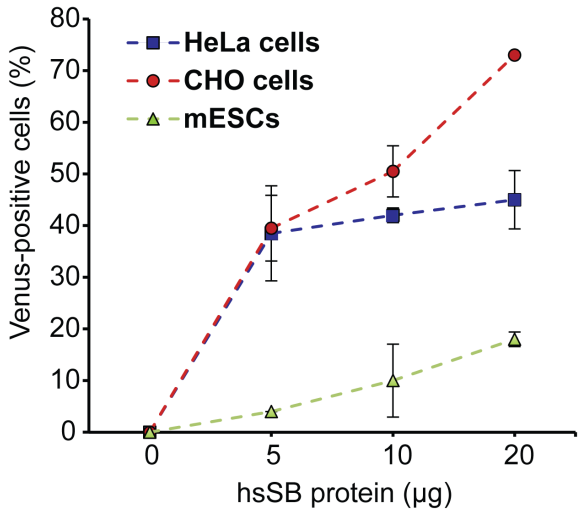


Figure 3-17: Plot showing the efficiency of hsSB-mediated genome engineering in different cell types quantified by flow cytometric analysis. Mean values of two independent experiments are shown. Error bars represent the standard deviation.

Using our purified hsSB variant and adapting our protocol, our collaborators (Ivics Laboratory, Paul-Ehrlich-Institut, Langen, Germany and Hudecek Laboratory, Uniklinikum Würzburg, Würzburg, Germany) were recently also able to demonstrate the functionality of the hsSB-based procedure for engineering of human primary cells. In the Ivics Laboratory, they achieved successful introduction of a reporter gene in primary

human Hematopoietic Stem and Progenitor Cells (HSPCs, $CD34^+$) (E. Grueso, Ivics Laboratory, unpublished). In the Hudecek Laboratory, they were able to generate genetically modified CD19-targeted human CAR T cells (A. Mades, Hudecek Laboratory, unpublished), which exerted powerful anti-tumor activity in murine xenograft models of human lymphoma.

In conclusion, we developed a new SB variant suitable for large-scale recombinant protein production and transfection, which allows us to successfully engineer a range of mammalian cell lines and even human primary cells. We named our engineering strategy SBprotAct, since it is based on the direct delivery and action of the active (Act) SB protein (SBprot). The name also showcases the increased safety associated with the use of this method, because SBprotAct provides a novel approach to alleviate safety issues and to enable maximal control of the SB system in clinical applications (as discussed in the following section section 3.4).

Note: A patent application ("Improved Transposase Polypeptide and Uses Thereof", Europe Patent Application 17187128.8) related to the content of sections 3.2 and 3.3 has been filed in August 2017 to the European Patent Office.

3.4 Discussion – The SBprotAct system

The SB system is the most clinically advanced non-viral genetic engineering approach used in gene therapy; however, important issues remain related to its use. For example, long-term expression of SB transposase and/or its genomic integration from an expression vector can result in uncontrolled transposition and genotoxicity. To circumvent the risks associated with the use of transposase-coding plasmids, direct delivery of the SB transposase protein is highly desired. However, the production of active recombinant SB transposase in a quantity and quality required for cellular delivery has been challenging to date.

In this part of my work, I have designed and characterized a new SB transposase variant (hsSB) (see section 3.2), which allowed efficient genetic engineering of various mammalian cell types and primary cells by direct SB protein delivery for the first time (see section 3.3). The specific advantages of this technology and the opportunities it offers for research and medical application will be discussed in the following paragraphs.

3.4.1 Summary of experimental findings

Using structural analysis and modelling, I first selected and engineered specific amino acids in the SB transposase to generate a new transposase variant, hsSB. While preserving full functionality in transposition, these mutations conferred extremely valuable properties to hsSB for protein delivery. These include easy recombinant production, increased solubility, reduced susceptibility to degradation, resistance to freezing and long-term storage, and outstanding thermostability. Following strategic optimization of conditions and nucleofection protocols, hsSB was then successfully delivered to a number of mammalian cell types, allowing genetic engineering of HeLa, CHO, mESC cells, and primary cells. This established a new cell engineering platform, named SBprotAct, that is based on the direct delivery of the hsSB transpososase.

SBprotAct avoids the use of the SB gene for transgene delivery, thereby prohibiting genomic integration of the transposase cassette in the engineered cells. Remarkably, we found that SBprotAct reduces exposure of the engineered cells to the transposase protein down to 2 days, and thereby allows to sensitively modulate the rates of transgenesis and to

adjust the integrated gene copy number to few copies per genome. We further demonstrated that SBprotAct provides a reliable tool to engineer a broad range of mammalian cell lines, spanning from biotechnology-relevant cells (CHO) to stem cells and human primary cells.

3.4.2 Advances of SBprotAct

SBprotAct exhibits several valuable properties, which advance the state-of-the-art in the use of SB and transposons in general for biotechnological and medical applications. In particular, novel features and technological advances of SBprotAct concern:

Development of SB by rational protein design. Generation of the hsSB variant - as well as previously reported hyperactive derivatives (Voigt *et al.*, 2016) - demonstrates that structure-based protein design is a valuable approach for the development of custom-made SB transposases and improved SB systems. Therefore, availability of structural and biochemical data on SB is of key importance to support current efforts in improving SB' efficiency, specificity and/or safety for clinical applications.

Direct delivery of active transposases. Delivery of purified genome engineering nucleases is an effective strategy to avoid regulatory complications, safety concerns and financial limitations connected to the use and the manufacture of viral vectors. This approach has been employed for Cre recombinase (Pfeifer *et al.*, 2001), TALENs (J. Liu *et al.*, 2014; Ru *et al.*, 2013), zinc finger (Gaj *et al.*, 2012), and Cas9 (Kim *et al.*, 2014) proteins. Yet, direct delivery of transposases to mammalian cells has proven challenging. Recently, the mariner Mos1 and Mboumar-9 transposases could be directly delivered in HeLa and HEK cells, but the transposition activity of the transfected proteins was very low (Trubitsyna *et al.*, 2017). To date, high-efficiency mammalian cell engineering by transposase delivery has been only achieved for the piggyBac transposase, but required incorporation of the protein into lentiviral particles (Cai *et al.*, 2014). Thus, our work provides the first completely virus-free system for efficient delivery of a transposase protein.

No need for gene or mRNA as transposase sources. In current SB-based applications, expression of the SB transposase is achieved either from an expression plasmid (Izsvak *et al.*, 2000) or from protein-encoding mRNA (Galla *et al.*, 2011) delivered into target cells.

Thus, SBprotAct establishes a new generation of the SB transposon system, based on the use of purified transposase, which is unprecedented.

In comparison to transposase gene delivery, direct hsSB protein delivery provides no risks of transposase-gene or promoter insertion from the expression vector (e.g. by homologous recombination) into the genome of target cells. Unintended integration of the SB gene may result in infinite transposase production and uncontrolled transposition. Moreover, acquisition of the transposase promoter might cause undesired transcriptional activation (of e.g. oncogenes) or disruption of gene regulatory networks in the target cells. Thus, delivery of SB in form of protein, instead of DNA, ensures that the engineered cells are free of undesired exogenous DNA and proteins derived from the vector, reducing the risks of genome instability or phenotypic perturbations in the target cells. For therapeutic cells, exogenous DNA and proteins can also trigger undesired immunogenic responses in the patients. It is particularly important to circumvernt these aberrant effects for therapeutic cells, such as CAR T cells, which are reported to persist for years in the treated patients.

hsSB protein delivery offers several advantages also in comparison to the use of mRNA as source of transposase, including:

i. Independency from cellular translation efficiency and regulation;
ii. Maximal, direct and fast control of transposition efficiency, since the transposase works immediately after transfection;
iii. Possibility to probe protein activity prior to application *in vitro*, which is of particular relevance for quality control procedures in a commercial or clinical setting. Intrinsic instability of mRNA raises quality control issues that could hinder widespread use for therapy.

Finally, since the rates and time frame of transgene insertion do not depend on transposase expression (from plasmids) and translation (from plasmids or mRNAs) by the cellular machinery, SBprotAct expands the applicability of SB-mediated engineering to cells in which protein overexpression is difficult and/or compromises cell viability.

Short retention times of the protein in the target cells. As the hsSB protein is immediately available and rapidly clears from the cells (less than 2 days), SBprotAct enables fast and safe engineering. In fact, the hsSB protein acts in a hit-and-run fashion, minimizing the temporal window of transposition and thus the associated off-target activities such as

transgene remobilization, insertional mutagenesis, and cytotoxicity. Moreover, since the protein is rapidly degraded, we expect that the immunogenic potential of the SBprotAct system is essentially reduced to that of the genetic cargo alone.

Control of transgenesis rates and copy numbers. As a consequence of direct delivery and fast transposase turnover, we found that SBprotAct allows maximal control of the cell engineering procedure. By varying the hsSB protein dose, transposition efficiency as well as the discrete number of insertions per genome can be tightly controlled. Remarkably, SBprotAct allows adjusting the copy number of the integrated gene to potentially a single copy per genome, reducing variance between individual cells and minimizing the risk of insertional mutagenesis and of undesired recombination events between multiple genomic copies. This fine tuning is not possible to achieve using viral and transposon-based vectors, including the current SB plasmid system, that typically generate high and variable number of integrations per genome in the cell population. Controlling copy numbers and minimizing aberrant genomic events is of critical importance for therapeutic applications to prevent potential post-therapeutic oncogenic transformation of the treated cells in the patients.

Applicability in different cells. We showed that SBprotAct finds broad applicability for engineering of several mammalian cell lines and primary cells. Its potency in CHO cells is of high interest for the industrial production of protein therapeutics (or biologics, e.g. therapeutic antibodies). In particular, the ability of SBprotAct to tightly control the number of inserted transgene copies constitutes a special benefit, as optimal gene dosage is directly linked to the yield of the therapeutic product (Chusainow *et al.*, 2009). Furthermore, SBprotAct can be efficiently used in difficult-to-engineer embryonic and hematopoietic stem cells and thus may facilitate engineering of stable stem cells for both research and therapy.

3.4.3 Potential impacts of SBprotAct on CAR T cell therapy

Safety, efficacy, and accessibility are the aspects currently limiting the translation of CAR T cell therapy into time- and cost-effective treatments for a large number of patients (Fesnak *et al.*, 2016; Kebriaei *et al.*, 2017; Sadelain *et al.*, 2017). SBprotAct can help overcome specific technical issues in all these three areas.

Safety. SBprotAct reduces exposure of the engineered cells to the transposase protein down to 2 days. In turn, in the first clinical trials the SB transposase gene was detected in CAR T cells even 21 days after modification. Recent developments in the *ex vivo* cell manufacturing process will allow reinfusion of the engineered cells in upcoming clinical trials after less than 2 days (Kebriaei *et al.*, 2017). The use of SBprotAct would guarantee that the transposase protein is not present in the cells at the time of administration, prohibiting post-infusion genotoxic events and minimizing immunogenic reactions in the patient. Moreover, by enabling tight control of the discrete number of insertions, SBprotAct generates a better-defined therapeutic cell product with constant transgene expression levels and minimal copy number variation between cells. Together, these properties will increase safety, consistency, and reproducible efficacy of CAR T cells produced by SBprotAct.

Efficacy. Currently, long waiting times before CAR T cell administration greatly limit the fitness of the therapeutic cells and their treatment success in cancer patients both in early and advanced cancer progression stages (Kebriaei *et al.*, 2016). The SBprotAct strategy is independent of transposase expression by the cellular machinery, which reducing the time-frame of genetic modification and the burden on the fitness of the target cell. In addition, the possibility to reinfuse transposase-free CAR T cells 2 days after engineering (as discussed in the previous paragraph) will ensure shorter waiting times and higher efficacy of the therapy, without compromising safety.

Different responses observed among patients restrain broad application of CAR T therapy to date. Generation of a more homogeneous population of engineered CAR T cells by SBprotAct (with minimal transgene copy number variation) could increase robustness and reproducibility of therapies. Moreover, by generating well-defined cell products, SBprotAct could help to identify engineered cell pools with the highest fitness and anti-tumour activity to be used as ‘off-the-shelf’ universal CAR T cells (Poirot *et al.*, 2015) or in personalized treatments.

Accessibility. Treatment delays and huge costs of CAR T therapy are mainly associated to the fact that vector generation, T cells modification and expansion and quality control of the final cell product are costly, time-consuming and possible only in small number of highly-specialized centers. The use of the simple SB system minimizes the time and cost of

vector manufacture, and evades laborious testing for replication-competent virus in the infusion product. Additionally, SBprotAct would provide a higher quality and safer cell product (as discussed above), thus reducing cost per treatment and improving scalable/exportable vector production to serve large numbers of patients.

Moreover, hsSB is well amenable for large-scale, automated and cost-effective manufacture. Its production and handling is also simple and could be performed even in point-of-care facilities, like hospitals and clinics. Moreover, the possibility to store the hsSB protein without compromising the efficacy allows easy delivery to de-centralized medical facilities for personalized treatment procedures. Thus, SB's use is not restricted to specialized centers, and together with the reduction in T cell culturing times (see above), it can help facilitate the dispersal of CAR T technologies to regular hospitals and regular patients.

3.4.4 Future directions

This study has paved the way for the development of genetic engineering methods based on direct delivery of the SB protein.

A current limitation of the SBprotAct system lies in the application of two successive transfection steps. Although electroporation constitutes one of the most efficient method for DNA and/or protein delivery, especially in hard-to-transfect cells, it greatly compromises cell viability. Thus, optimization of protocols for direct delivery of pre-formed SB transposase-transposon DNA complexes by a single electroporation event is highly desirable, especially for engineering sensitive cell lines. In this regard, different DNA carriers, such as minimalistic transposon-carrying vectors [minicircles (Monjezi *et al.*, 2017)], might provide an advantage for concerted protein-DNA delivery in comparison to plasmids. In fact, DNA charge, size and shape are key determinant for efficient transfection and can greatly influence the success of co-delivery.

Furthermore, alternative delivery methods to electroporation could be explored. For example, delivery of Cas9 protein has been achieved using a plethora of protocols, with variable success rates in different target cells (Kim *et al.*, 2014; X. Liang *et al.*, 2015; Ramakrishna *et al.*, 2014). In addition to different delivery methods, many strategies to increase protein's uptake across the cell and nuclear membranes have been described in

literature. Most of them are based on the use of cell-penetrating peptides (CPPs) [e.g. the transactivator of transcription (TAT) peptide from human immunodeficiency virus] that can be either associated by non-convalent interactions or covalently linked to the protein of interest (Fu *et al.*, 2014). In the case of SB, these extra moieties could be genetically fused to the N-terminus of the protein, as modification of the C-terminus generally compromise protein activity.

Most of the advantageous features of SBprotAct derive from the rapid turnover of the hsSB protein in the target cells. In order to achieve an even tighter control of protein clearance and thus of transposition, the engineering of faster degrading protein variants could be tested. Moreover, we provided proof-of-concept that SBprotAct is a versatile method that can be applied to different cells. Further application to a larger portfolio of cell lines and primary cells will strongly support this idea. Notably, the current SB system is used to *ex vivo* modify retinal pigment epithelium (RPE) cells to treat patients affected by macular degeneration (Hudecek *et al.*, 2017). If SBprotAct could be successfully applied to RPE cells, this will further expand its potential in gene therapy applications.

In conclusion, while preserving all advantages of the current SB system, including simplicity, ease and low cost, SBprotAct enables maximal control and enhanced safety of gene delivery and we anticipate that its valuable features will promote the advance of SB-mediated genetic engineering applications both in research and in the clinics.

4 CONCLUSIONS

'DNA neither cares nor knows. DNA just is. And we dance to its music.'
R. Dawkins, 1995

This thesis describes the results obtained during my PhD work, where I investigated the molecular mechanisms of a prominent group of DNA transposons, the *Tc1/mariners*, aiming to develop new transposon systems for genome engineering. In conjunction with being natural wide-host-range gene carriers, *Tc1/mariners* are extraordinarily powerful artificial vectors for genome manipulation in vertebrates. Hence, mechanistic insights into *Tc1/mariner* transposition and regulation will guide the rational design of novel genetic tools for basic research and human therapy.

The first part of my work focused on the molecular principles underlying transposition of two representative *Tc1/mariners*: *Hsmar1*, the best biochemically characterized family member, and *SB*, a prime genetic tool currently applied in gene therapy clinical trials. A combination of structural biology methods, biochemistry, fluorescence-based, and biophysical techniques allowed to characterize the early steps of *Hsmar1* and *SB* transposition. Comparative *in vitro* analysis revealed that *SB* assembles its transposition machinery (or transpososome) differently to *Hsmar1*. Remarkably, the unexpected assembly pathway of *SB* seems to underlie its exceptionally high transposition efficiency, as demonstrated by cell biology assays. Additionally, I characterized the key steps involved in *SB* transposition, namely transposon end binding, excision and integration, *in vitro*, revealing previously unknown features of the process. Taken together, these results

allow to build a mechanistic model for *SB* and *Hsmar1* transposition and provide first insights into the molecular reasons behind *SB*'s extraordinary success as a genetic tool.

The second part of my work addresses the most ambitious objective of my research that is the development of novel *SB*-based systems for genome engineering. Taking advantage of assays established above and of available structural data, I generated a novel SB transposase variant (hsSB) by rational protein design. hsSB exhibited unmatched solubility and stability in comparison to previous versions laying the foundation for the development of a new genome engineering platform, named SBprotAct. SBprotAct is based on the direct delivery of recombinant SB protein to cells and allows for efficient genome modification of several cell types (including stem cells and human T cells) in a more controlled and secure manner than the state-of-art technology. Thus, SBprotAct constitutes a promising strategy to improve the safety and performance of *SB* in its genetic engineering applications.

In summary, this work exemplifies the general principle inspiring biomolecular engineering: knowledge on molecular systems is used to guide their rational design for applications. By learning more about the working mechanism of *Tc1/mariners*, and of *SB* in particular, via an integrated multi-disciplinary approach, we provided an invaluable resource for the development of tailor-made transposon-based tools. On the long term, we expect these novel genetic engineering technologies to have tremendous implications for life sciences and human health.

5 MATERIALS AND METHODS

This chapter describes the experimental methodologies employed in the thesis.

5.1 Materials

5.1.1 Chemicals and reagents

Unless otherwise indicated, all chemicals used in this study were purchased from Sigma-Aldrich or Merck Millipore. All restriction enzymes were supplied by New England Biolabs (NEB) apart from DpnI, which was purchased from Thermo Scientific. Enzymes and reagents used for molecular biology were supplied by NEB, Bioline, and Thermo Scientific as stated in the text. Antibiotics were purchased from Sigma-Aldrich and Carl Roth. All buffers and stocks were made using deionized, distilled water.

5.1.2 Bacterial growth media

All media for bacterial growth were prepared by EMBL Media Kitchen Facility. Liquid *E. coli* cultures were grown in Lysogeny Broth (LB) medium containing 10 g tryptone, 5 g yeast extract, and 5 g sodium chloride per litre (pH 7.2) or in Super Optimal Broth (SOB) medium containing 20 g tryptone, 5 g yeast extract, 0.585 g sodium chloride, and 0.186 g potassium chloride per litre (pH 7.2). Solid medium for *E. coli* growth consisted of 10 g tryptone, 5 g yeast extract, 10 g sodium chloride, and 15 g agar per litre (pH 7.2). When applicable, the following antibiotics were added to the growth media: kanamycin (kanamycin sulphate, 50 μg/ml), ampicillin (ampicillin sodium salt, 100 μg/ml) and chloramphenicol (33 μg/ml). Bacterial growth was monitored by measuring the optical density at the wavelength of 600 nm (OD_{600}).

5.1.3 Bacterial strains

E. coli strain Rosetta™ 2 (DE3) cells [Genotype: F^- *ompT hsdS*$_B$(r_B^- m_B^-) *gal dcm* pRARE2 Cm^R- Chloramphenicol resistant] were purchased from Novagen. *E. coli* strain XL10-Gold cells [Genotype: Tet^R *Δ(mcrA)183 Δ(mcrCB-hsdSMR-mrr)173 endA1 supE44 thi-1 recA1 gyrA96 relA1 lac Hte* [F′ *proAB lacIqZDM15 Tn10* (Tet^R- Tetracycline resistant) Amy Cm^R] were purchased from Stratagene.

5.1.4 Plasmids

Table 5.1 contains a list of plasmids used in this study together with the description, carried antibiotic resistance, origin of replication, and source of each plasmid.

Table 5-1: Plasmids used in this study

Plasmid	Description	Resistance*	Origin	Source
pETM-22	T7 expression vector with N-terminal TRX-6xHis tag	Km	pBR322	PEP-core EMBL Heidelberg
Hsmar1_fl_pETM-22	T7 expression vector encoding TRX-6xHis tag-Hsmar1_fl	Km	pBR322	I. Querques, M.Sc. thesis
Hsmar1_fl_W118P_pETM-22	T7 expression vector encoding TRX-6-His tag-Hsmar1_fl_W118P	Km	pBR322	This study
Hsmar1_fl_V119G_pETM-22	T7 expression vector encoding TRX-6xHis tag-Hsmar1_fl_V119G	Km	pBR322	This study
SB_fl_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl	Km	pBR322	Voigt et al., 2016
SB_fl_R36C_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_R36C	Km	pBR322	This study
SB_fl_S37C_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_S37C	Km	pBR322	This study
SB_fl_S55C_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_S55C	Km	pBR322	This study
SB_fl_S99C_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_S99C	Km	pBR322	This study
SB_fl_T102C_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_T102C	Km	pBR322	This study
SB_fl_V106C_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_V106C	Km	pBR322	This study
SB_fl_H115C_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_H115C	Km	pBR322	This study
SB_fl_S116C_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_S116C	Km	pBR322	This study
SB_fl_C176C-I212S_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_C176S-I212S	Km	pBR322	This study
SB_fl_S37C-C176C-I212S_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_S37C-C176S-I212S	Km	pBR322	This study
SB_fl_V106C-C176C-I212S_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_V106C-C176S-I212S	Km	pBR322	This study
SB_fl_R14C-C176C-I212S_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_C176S-I212S-R14C	Km	pBR322	This study
SB_fl_D17C-C176C-I212S_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_fl_C176S-I212S-D17C	Km	pBR322	This study
A1_SB_fl_pETM22	T7 expression vector encoding TRX-6xHis tag-A1-SB_fl	Km	pBR322	F. Voigt, PhD thesis
B1_SB_fl_pETM22	T7 expression vector encoding TRX-6xHis tag-B1-SB_fl	Km	pBR322	F. Voigt, PhD thesis
A1_pETM22	T7 expression vector encoding TRX-6xHis tag-A1	Km	pBR322	F. Voigt, PhD thesis
B1_pETM22	T7 expression vector encoding TRX-6xHis tag-B1	Km	pBR322	F. Voigt, PhD thesis
A1_Hsmar1_fl_pETM22	T7 expression vector encoding TRX-6xHis tag-A1-Hsmar1_fl	Km	pBR322	This study
B1_Hsmar1_fl_pETM22	T7 expression vector encoding TRX-6xHis tag-B1-Hsmar1_fl	Km	pBR322	This study
SB_1-117_pETM-22	T7 expression vector encoding TRX-6xHis tag-SB_1-117	Km	pBR322	F. Voigt, PhD thesis
Tc3_1-135_SB_110-340_pETM-22	T7 expression vector encoding TRX-6xHis tag-Tc3_1-135_SB_110-340	Km	pBR322	F. Voigt
pCFJ601 - Peft-3 Mos1 transposase	Template plasmid for amplification of Mos1 gene	Amp	pBR322	Addgene
Mos1_1-56_SB_53-340_pETM-22	T7 expression vector encoding TRX-6xHis tag-Mos1_1-56_SB_53-343	Km	pBR322	This study
GST_SB_fl_pETM33	T7 expression vector encoding 6xHis tag-GST-SB_fl_C176S-I212S	Km	ColE1	O. Barabas, NIH Bethesda, USA
TEV_GST_SB_fl_pETM33	T7 expression vector encoding 6xHis tag-TEV cleavage site-GST-SB_fl	Km	ColE1	This study
TEV_GST_SB_fl_C176S-I212S_pETM33	T7 expression vector encoding 6xHis tag-TEV cleavage site-GST-SB_fl_C176S-I212S	Km	ColE1	This study
pETMCN-YC	Target plamid used for *in vitro* integration assays	Cm	p15a	Romier et al., 2006; courtesy of C. Müller, EMBL Heidelberg
A1_SB_fl_EWS	Mammalian expression vector encoding 6xHis-HA-Strep-Tag_A1-SB_fl	Amp	ColE1	F. Voigt, PhD thesis
A1_Hsmar1_fl_EWS	Mammalian expression vector encoding 6xHis-HA-Strep-Tag_A1-Hsmar1_fl	Amp	ColE1	F. Voigt, PhD thesis
FRB_CMV	Template plasmid for amplification of FRB gene	Km	pUC	Varnai et al., 2006; courtesy of D. Yushchenko, EMBL Heidelberg
FKBP_CMV	Template plasmid for amplification of FKBP gene	Km	pUC	Varnai et al., 2006; courtesy of D. Yushchenko, EMBL Heidelberg
FRB_SB_fl_EWS	Mammalian expression vector encoding 6xHis-HA-Strep-Tag_FRB-SB_fl	Amp	ColE1	This study
FKBP_SB_fl_EWS	Mammalian expression vector encoding 6xHis-HA-Strep-Tag-FKBP-SB_fl	Amp	ColE1	This study
FRB_Hsmar1_fl_EWS	Mammalian expression vector encoding 6xHis-HA-Strep-Tag-FRB-Hsmar1_fl	Amp	ColE1	This study
FKBP_Hsmar1_fl_EWS	Mammalian expression vector encoding 6xHis-HA-Strep-Tag-FKBP-Hsmar1_fl	Amp	ColE1	This study
pUC19	Filler plamid used for *in vivo* transposition assays	Amp	pBR322	Addgene
pHsmar1-neo	Reporter plasmid with neomycin resistance gene flanked by the Hsmar1 transposon ends	Amp	pMB1	Miskey et al., 2007
SB_fl_pCMV(CAT)T7	Mammalian expression vector encoding 6xHis-SB_fl	Cm	pBR322	Mates et al., 2009
SB-C176S-I212S_pCMV(CAT)T7	Mammalian expression vector encoding 6xHis-SB_fl_C176S-I212S	Cm	pBR322	This study
SB_PGK-neo	Reporter plasmid with neomycin resistance gene flanked by the SB transposon ends	Amp	ColE1	Neuromics; courtesy of F. Spitz, EMBL Heidelberg
SB_T2/Venus	Reporter plasmid with Venus-encoding gene flanked by the SB transposon ends	Amp	ColE1	Mates et al., 2009

fl-full length
* Km-kanamycin; Cm-Chloramphenicol; Amp-ampicillin

5.1.5 Oligonucleotides

All oligonucleotides used in this study were purchased from Integrated DNA Technologies (IDT, Leuven, Belgium) and were resuspended in distilled water or Tris-EDTA (TE) buffer [10 mM tris(hydroxymethyl)aminomethane (Tris), pH 8, and 1 mM

ethylenediaminetetraacetic acid (EDTA)] to a final concentration of 1 mM or 100 µM. Oligonucleotides used as primers in polymerase chain reaction (PCR) are listed in Table 5-2. Melting temperature of the primers was estimated using the IDT OligoAnalyzer tool (http://eu.idtdna.com/calc/analyzer). Oligonucleotides used as DNA substrates in biochemical assays, crystallization and binding studies are shown in the method sections for individual experiments.

Table 5-2: Primers for PCR reactions used in this study.

No	Primer	Sequence**	Purpose
1	Hsmar1_W118P	[Phos]GCAAATTGGAAAGGTGAAAAAGCTCGATAAGCCGGTGCCTCATGAGCTGAGTG	Site-directed mutagenesis of Hsmar1_fl_W118P
2	Hsmar1_V119G	[Phos]GAAAGGTGAAAAAGCTCGATAAGTGGGGTCCTCATGAGCTGAGTGAAAATCAAAAAAATCG	Site-directed mutagenesis of Hsmar1_fl_V119G
3	SB_R36C	[Phos]GCCTGGCGGTACCATGCTCATCTGTACAAACAATAGTACGCAAA	Site-directed mutagenesis of SB_R36C
4	SB_S37C	[Phos]CTGGCGGTACCACGTTGCTCTGTACAAACAATAGTACGCAAGT	Site-directed mutagenesis of SB_S37C
5	SB_S55C	[Phos]GACCACGCAGCCGTGCTACCGCTCAGGAAGGAGAC	Site-directed mutagenesis of SB_S55C
6	SB_S99C	[Phos]TGGAGGAAACAGGTACAAAAGTATGCATATCCACAGTAAAACGAGTCCTATATC	Site-directed mutagenesis of SB_S99C
7	SB_T102C	[Phos]GGAAACAGGTACAAAAGTATCTATATCCTGCGTAAAACGAGTCCTATATCGACATAAC	Site-directed mutagenesis of SB_T102C
8	SB_V106C	[Phos]ACAAAAGTATCTATATCCACAGTAAAACGATGCCTATATCGACATAACCTGAAAGGC	Site-directed mutagenesis of SB_V106C
9	SB_H115C	[Phos]GTCCTATATCGACATAACCTGAAAGGCTGCTCAGCAAGGAAGAAGCCACT	Site-directed mutagenesis of SB_H115C
10	SB_S116C	[Phos]ATATCGACATAACCTGAAAGGCCACTGCGCAAGGAAGAAGCCACTGCT	Site-directed mutagenesis of SB_S116C
11	SB_C176S	[Phos]AGGAAGAAGGGGGAGGCTTCCAAGCCGAAGAACACCATCCC	Site-directed mutagenesis of SB_C176S
12	SB_I212S	[Phos]GGTGCACTTCACAAAATAGATGGCAGCATGGACGCGGTGCAGTAT	Site-directed mutagenesis of SB_I212S
13	SB_R14C	[Phos]GCCAAGACCTCAGAAAATGCATTGTAGACCTCCACAAG	Site-directed mutagenesis of SB_R14C
14	SB_D17C	[Phos]GACCTCAGAAAAAGAATTGTATGCCTCCACAAGTCTGGTTC	Site-directed mutagenesis of SB_D17C
15	Hsmar1_fl_for	ACAAGTTTGTACAAAAAAGCAGGCTGCATGGAAATGATGTTAGACAAAAAGC	Cloning of Hsmar1_fl in A1_SB_fl_pETM22 (to substitute SB)
16	Hsmar1_fl_rev	GCTCGAGTGCGGCCGCAAGCTTCTAATCAAAATAGGAACCATTACAA	Cloning of Hsmar1_fl in A1_SB_fl_pETM22 (to substitute SB)
17	Hsmar1_fl_for	ACAAGTTTGTACAAAAAAGCAGGCTGCATGGAAATGATGTTAGACAAAAAGC	Cloning of Hsmar1_fl in B1_SB_fl_pETM22 (to substitute SB)
18	Hsmar1_fl_rev	GCTCGAGTGCGGCCGCAAGCTTCTAATCAAAATAGGAACCATTACAA	Cloning of Hsmar1_fl in B1_SB_fl_pETM22 (to substitute SB)
19	Mos1_1-56_SB_53-340_for	AGTTCTGTTCCAGGGGCCCATGATGTCGAGTTTCGTGCCG	Cloning of Mos1_1-56 in SB_fl_pETM22
20	Mos1_1-56_SB_53-340_rev	AGCGGTATGACGGCTGCGTGGTACCACTTTTGAAGCGTTGAAA	Cloning of Mos1_1-56 in SB_fl_pETM22
21	TEV_GST_SB_fl_pETM33	[Phos]CCATCACCATCACAACACTAGTGAGAACCTGTACTTCCAAGGCATGTCCCCTATACTAGG	Cloning of TEV cleavage site in GST_SB_fl_pETM33
22	FKBP_SB_for	ACCATCACCATCACCTCGAGACCATGGGAGTGCAGGTGGAA	Cloning of FKBP in A1_SB_fl_EWS (to substitute A1)
23	FKBP_SB_rv	TGTACAAACTTGTGATATCGGCCGCTTCCAGTTTTAGAAGCTCCACA	Cloning of FKBP in A1_SB_fl_EWS (to substitute A1)
24	FRB_SB_for	ACCATCACCATCACCTCGAGACCTCTAGAATCCTCTGGCATGAGA	Cloning of FRB in A1_SB_fl_EWS (to substitute A1)
25	FRB_SB_rv	TGTACAAACTTGTGATATCGGCCGCACTAGTCTTTGAGATTCGTCGG	Cloning of FRB in A1_SB_fl_EWS (to substitute A1)
26	FKBP_Hsmar1_for	ACAAGTTTGTACAAAAAAGCAGGCTGCATGGAAATGATGTTAGACAAAAAG	Cloning of FKBP in A1_Hsmar1_fl_EWS (to substitute A1)
27	FKBP_Hsmar1_rv	ACCACTTTGTACAAGAAAGCTGGGTCCTAATCAAAATAGGAACCATTACAA	Cloning of FKBP in A1_Hsmar1_fl_EWS (to substitute A1)
28	FRB_Hsmar1_for	ACAAGTTTGTACAAAAAAGCAGGCTGCATGGAAATGATGTTAGACAAAAA	Cloning of FRB in A1_Hsmar1_fl_EWS (to substitute A1)
29	FRB_Hsmar1_rv	ACCACTTTGTACAAGAAAGCTGGGTCCTAATCAAAATAGGAACCATTACAA	Cloning of FRB in A1_Hsmar1_fl_EWS (to substitute A1)

** [Phos]- 5' Phosphorylation

5.2 Molecular biology methods

5.2.1 Constructs for protein overexpression

The open reading frame coding for amino acids 1-343 of the reconstituted full length Hsmar1 transposase as reported by Miskey and colleagues [(Miskey *et al.*, 2007); GenBank accession code EF517118] was kindly provided by Csaba Miskey and Zoltán Ivics. The *Sleeping Beauty 100X* (*SB*) gene [encoding amino acids 1-340 of the hyperactive SB100X; (Mates *et al.*, 2009)] was kindly provided by Zoltán also and Zsuzsanna Izsvák. The Hsmar1- and SB-encoding genes were cloned into vectors pETM-22 for

overexpression in *E. coli* prior to this thesis work, to give vectors Hsmar1_fl_pETM-22 (I. Querques, M.Sc. thesis) and SB_fl_pETM-22 (Voigt *et al.*, 2016) respectively.

All prokaryotic expression constructs contained in pETM-22 vectors were cloned in frame with N-terminal ThioredoxinA (TRX) and 6XHis tags that could be removed from the protein constructs via incubation with PreScission protease (3C; Protein Expression and Purification Core Facility, EMBL Heidelberg). The GST_SB_fl construct in pETM-33 vector contained a N-terminal 6XHis- GST tag that could be removed via incubation with PreScission protease (3C; Protein Expression and Purification Core Facility, EMBL Heidelberg). Additionally, a TEV protease cleavage site was inserted between the 6XHis tag and the GST tag in vector GST_SB_fl_pETM33 via a specific loop-in restriction-free cloning protocol (described in section 5.2.3) to generate vector TEV_GST_SB_fl_pETM33 (using primer 21). In this way, the 6XHis tag only could be removed from the GST-SB protein construct through incubation with TEV protease (Protein Expression and Purification Core Facility, EMBL Heidelberg).

All mutations of the Hsmar1 and SB protein sequences were introduced via site-directed mutagenesis (following protocols described in (Makarova *et al.*, 2000) and in section 5.2.3).

The genes for the dimerization-dependent fluorescent fusion proteins ddRFP-A1 (A1, GenBank accession code: JN381545) and ddRFP-B1 (B1, GenBank accession code: JN381546) (Alford *et al.*, 2012) were cloned in frame with the SB transposase gene into vector SB_fl_pETM-22 to obtain vectors A1_SB_fl_pETM-22 and B1_SB_fl_pETM-22 prior to this thesis work (F. Voigt, PhD thesis). The Hsmar1 transposase gene was substituted to the SB gene in vectors A1_SB_fl_pETM-22 (using primers 15/16) and B1_SB_fl_pETM-22 (using primers 17/18) via restriction-free cloning (as described in section 5.2.2) to generate vectors A1_Hsmar1_fl_pETM-22 and B1_Hsmar1_fl_pETM-22 respectively. A1-SB, B1-SB, A1-Hsmar1 and B1-Hsmar1 proteins contain a 13 amino acids long flexible linker (AADITSLYKKAGC) connecting the N-terminal fluorescent proteins with the respective transposases. A1 and B1 proteins were overexpressed from vectors A1_pETM-22 and B1_pETM-22, respectively (F. Voigt, PhD thesis).

The DNA sequence encoding for amino acids 1-56 of the Mos1 transposase (UniProtKB accession code: Q7JQ07.1) was cloned from vector pCFJ601 - Peft-3 Mos1 transposase (purchased from Addgene) to vector SB_pETM22 via restriction-free cloning (protocol in section 5.2.2) (using primers 19/20) to generate vector Mos1_1-56_SB_53-340_pETM-22.

The genes of FKBP12 and the FRB proteins (Varnai *et al.*, 2006) from vectors FKBP-CMV and FRB-CMV (courtesy of D. Yushchenko, EMBL Heidelberg, Germany) were substituted via restriction free cloning to the A1 genes into vectors A1_SB_fl_EWS and A1_Hsmar1_fl_EWS respectively, to generate vectors FKBP_SB_fl_EWS (primers 22/23), FRB_SB_fl_EWS (primers 24/25), FKBP_Hsmar1_fl_EWS (primers 26/27) and FRB_Hsmar1_fl_EWS (primers 28/29). The resulting vectors are used for mammalian cell expression of FKBP-SB, FRB-SB, FKBP-Hsmar1 and FRB-Hsmar1 fusion proteins, respectively. All constructs were cloned in frame with N-terminal 6XHis-, HA- (Human influenza hemagglutinin) and Strep-tags into EWS vectors. No cleavage site was present to remove the tags from the protein constructs.

5.2.2 Restriction-free (RF) cloning

Restriction-free (RF) cloning hybrid primers containing complementary sequences to both the desired insert and target plasmids were generated using the primer design tool at www.rf-cloning.org. All DNA concentrations were determined by UV spectroscopy at 260 nm using a NanoDrop instrument (Thermo Scientific).

The first PCR mix contained 1x Phusion Flash High-Fidelity PCR Master Mix (Thermo Scientific), 500 nM of each forward (for) and reverse (rev) primer, and 20 ng of template vector in a total volume of 50 µl. The thermocycling conditions are shown in Table 5-3. The resulting product, so-called Mega Primer, was purified using a GenElute PCR Clean-Up Kit (Sigma-Aldrich) according to the manufacturer's instructions and used in a second PCR reaction, with the target plasmid acting as template in a reaction with 1x Phusion Flash High-Fidelity PCR Master Mix, 350 ng of Mega Primer, and 50 ng of template DNA in a total volume of 50 µl. Thermocycling conditions of second PCR are also shown in Table 5-3. The reaction products were purified using a GenElute PCR Clean-Up Kit (Sigma-Aldrich) and eluted in 17 µl of distilled water. In order to degrade any remaining parental plasmid DNA, the sample was incubated with 2 µl of 10x FastDigest Buffer and 2 µl of FastDigest DpnI restriction enzyme (both from Thermo Scientific) at 37 °C for at least 4 hours. 10 µl of the sample was then transformed into *E. coli* XL10-Gold chemically competent cells (Stratagene).

Table 5-3: Thermocycling conditions in restriction-free cloning PCR.

Step	Temperature	Time	
PCR1			
1. Initial Denaturation	98 °C	20 seconds	
2. Denaturation	98 °C	1 second	repeated 35 times
3. Annealing	50-60 °C	5 seconds	
4. Extension	72 °C	60 seconds/kbp	
5. Final extension	72 °C	5 minutes	
6. Hold	4 °C	hold	
PCR2			
1. Initial Denaturation	98 °C	20 seconds	
2. Denaturation	98 °C	1 second	repeated 18 times
3. Annealing	60-65 °C	5 seconds	
4. Extension	72 °C	8 minutes	
5. Final extension	72 °C	10 minutes	
6. Hold	4 °C	hold	

The presence of the insert in single colonies of *E. coli* transformants was validated by a standard colony PCR protocol. The reaction mix consisted of 1x MangoMix™ (Bioline), 400 μM of each forward and reverse primers (specific primers for the insert of interest) and single transformant colony cells as template. Thermocycling conditions of the colony PCR are shown in Table 5-4. The PCR products were directly loaded on 1% agarose gels [1% (w/v) agarose powder dissolved in 1x Tris-acetate- EDTA (TAE) buffer (40 mM Tris, 20 mM acetic acid, and 1 mM EDTA)]. After applying a voltage of 7-10 V per cm for 30-60 minutes, the products of gel electrophoresis were analysed on a UV transilluminator (Alpha Innotech) using AlphaImager® HP software (Fisher Scientific). A HyperLadder™ 1kb, Bioline DNA marker was used.

Table 5-4: Thermocycling conditions in colony PCR

Step	Temperature	Time	
1. Initial Denaturation	98 °C	5 minutes	
2. Denaturation	96 °C	30 seconds	repeated 35 times
3. Annealing	60 °C	30 seconds	
4. Extension	72 °C	2 minutes	
5. Final extension	72 °C	4 minutes	
6. Hold	4 °C	hold	

Positive transformant colonies were inoculated into 5 ml LB medium containing appropriate antibiotics and grown overnight at 37 °C. After cell harvesting, the plasmid DNA was extracted using GenElute Plasmid DNA Miniprep Kit (Sigma-Aldrich) and validated for the presence of the correct insert by DNA sequencing performed by GATC Biotech.

5.2.3 Site-directed mutagenesis

Point mutations and short amino acid motifs (i.e. TEV cleavage site) were introduced through site-directed mutagenesis using a loop-in protocol (Makarova et al., 2000). A single 5' phosphorylated primer was designed for each mutation so that it contained mutation of interest at the center of the primer and each half-site flanking the mismatch has a melting temperature of 58 °C with the melting temperature of the full primer close to 68 °C. The PCR mutagenesis reaction was performed in a reaction volume of 50 µL that included 2.5 units PfuUltra High Fidelity DNA polymerase enzyme in 1x PfuUltra HF DNA polymerase reaction buffer (Agilent Technologies), 1 unit Taq DNA Ligase in 1x Taq DNA Ligase buffer (New England Biolabs), 200 nM dNTP mixture (Bioline), 500 nM of a single 5' phosphorylated primer, and 100 ng of template plasmid DNA. Thermocycling conditions are shown in Table 5-5. Since only one primer containing the desired mutation is used, PCR results in amplification of a single-stranded plasmid, ligated by Taq DNA ligase in the same reaction. After purification of the reaction products and DpnI digestion, performed as previously described, the sample was then transformed into *E. coli* XL10-Gold competent cells and the sequences of the cloned plasmids were confirmed at the end of the construction by nucleotide sequencing by GATC Biotech.

Table 5-5: Thermocycling conditions in site-directed mutagenesis

Step	Temperature	Time	
1. Initial Denaturation	95 °C	1 minute	
2. Denaturation	95 °C	30 seconds	repeated 30 times
3. Annealing	55 °C	30 seconds	
4. Extension	65 °C	60 seconds/kbp	
5. Hold	4 °C	hold	

5.2.4 Protein overexpression in *E. coli*

All protein constructs described in this thesis were overexpressed in *E. coli* strains Rosetta™ 2 (DE3) cells from pET vectors under a T7 promoter. The expression vectors were transformed into cells, and overnight cultures of the transformants were set up in LB medium with appropriate antibiotics. In large-scale expressions, 500 ml of LB medium containing appropriate antibiotics was inoculated with 5 ml of overnight cultures and grown at 37 °C and 200 rpm until OD_{600} of 0.6-0.8. Next, expression was induced by addition of 0.5 mM of isopropyl β-D-1- thiogalactopyranoside (IPTG). Cells were grown

at 16 °C and 200 rpm for 18 hours and then harvested by centrifugation at 5000 rpm for 15 minutes at 4 °C. The obtained pellet was then washed once with 25 ml LB and centrifuged at 3000 *g* for 45 minutes at 4 °C. After the final harvest, the pellet was flash-frozen in liquid nitrogen and stored at -80 °C until further use or immediately used for purification (see section 5.2.5).

The selenomethionine (SeMet) derivative of Hsmar1 was expressed in methionine-prototrophic strain Rosetta™ 2 (DE3) from plasmid Hsmar1_fl_pETM-22. 5 ml of the overnight culture was washed twice with M9 medium containing 1x M9 salts (Table 5-6), 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.4% (w/v) glucose, 1 µg/ml thiamine, and 1x trace elements solution (Table 5-6). 50 ml of M9 medium was then inoculated with the washed culture (5 ml) and grown overnight at 37 °C. The overnight culture was then added to 1 litre of fresh M9 medium and grown to OD_{600} of 0.6. Next, essential amino acids were added at the following concentrations: 100 mg/l of lysine, phenylalanine, and threonine and 50 mg/l of isoleucine, leucine, valine, and selenomethionine. The cultures were then incubated for 20 minutes at 4 °C without shaking, followed by induction of expression with 0.5 mM IPTG. After expression for 18 hours at 16 °C, the cultures were harvested and frozen as described for the other proteins.

Table 5-6: Composition of 10x M9 salts and 100x trace elements.

Components	Amount per litre
10x M9 salts	
$Na_2HPO_4 \bullet 7\ H_2O$	70 g
KH_2PO_4	30 g
NaCl	5 g
NH_4Cl	10 g
100x trace elements	
EDTA	5 g
$FeSO_4 \bullet 7\ H_2O$	0.85 g
$ZnSO_4 \bullet 7\ H_2O$	0.173 g
$CuCl_2 \bullet 2\ H_2O$	0.013 g
$CoCl_2 \bullet 6\ H_2O$	0.010 g
H_3BO_3	0.010 g
$MnCl_2 \bullet 6\ H_2O$	0.0016 g
$(NH_4)_6Mo_7O_{24} \bullet 4\ H_2O$	0.010 g

5.2.5 Protein purification

All protein constructs described in this study were overexpressed as fusions with an N-terminal 6xHis affinity tag and purified through a three step affinity purification scheme, including 1) first purification by nickel affinity chromatography, 2) tag cleavage and

removal by second nickel affinity chromatography, and 3) size exclusion chromatography (SEC) on a gel filtration column. All purification steps were performed using the ÄKTApurifier Protein Purification System (GE Healthcare) at 4 °C.

The harvested expression pellets were resuspended in 40 ml Lysis buffer (Table 5-7). The resuspended cells were then lysed by sonication using a Branson Sonifier 250 set to 60% duty cycle and 55% output control, in 6 cycles of 45 seconds sonication and 60 seconds rest on ice. Next, the lysed cells were ultracentrifuged at 18000 rpm for 40 minutes at 4 °C. The supernatant, corresponding to the soluble fraction of *E. coli* proteins, was loaded onto a 5 ml HisTrap HP column (GE Healthcare), previously equilibrated in Loading buffer (Table 5-7) for the first nickel affinity purification. After immobilization on the column, elution of the 6xHis-tagged protein was achieved using gradually increasing imidazole concentrations (applying a gradient from 5% to 50% Elution buffer, Table 5-7).

In order to remove the 6xHis- (in case of protein expressed from vector TEV_GST_SB_fl_C176S-I212S_pETM33) or TRX-6xHis tags (for all other protein constructs), 10 µg of 6xHis tagged TEV protease or 10 µg of 6xHis tagged 3C protease per mg of protein, respectively, were added and the sample was then dialysed against Gel filtration buffer (Table 5-7) overnight at 4 °C in order to remove imidazole and allow for protease cleavage.

Table 5-7: Composition of purification buffers.

Buffer	Composition
Purification of all Hsmar1 constructs	
Lysis buffer	1x PBS, 1 M NaCl, 10 mM imidazole, 0.2 mM TCEP, 1 tablet of cOmplete Protease Inhibitor Cocktail (Roche), 25 mM PMSF, 25 µg/ml RNaseA (Roche), 50 µg/ml DNaseI (Roche), pH 7.5
Loading buffer	1x PBS, 1 M NaCl, 10 mM imidazole, 0.2 mM TCEP, pH 7.5
Elution buffer	1x PBS, 1 M NaCl, 1 M imidazole, 0.2 mM TCEP, pH 7.5
Gel filtration buffer	1x PBS, 1 M NaCl, 0.2 mM TCEP, pH 7.5
Purification of all SB constructs and of A1 and B1 proteins	
Lysis buffer	1x PBS, 1 M NaCl, 20 mM imidazole, 0.2 mM TCEP, 1 tablet of cOmplete Protease Inhibitor Cocktail (Roche), 25 mM PMSF, 25 µg/ml RNaseA (Roche), 50 µg/ml DNaseI (Roche), pH 7.5
Loading buffer	1x PBS, 1 M NaCl, 20 mM imidazole, 0.2 mM TCEP, pH 7.5
Elution buffer	1x PBS, 1 M NaCl, 1 M imidazole, 0.2 mM TCEP, pH 7.5
Gel filtration buffer	1x PBS, 1 M NaCl, 0.2 mM TCEP, pH 7.5

1xPBS – 1x phosphate buffered saline (0.2 g KCl, 0.2 g KH_2PO4, 1.15 g Na_2HPO_4, and 8 g NaCl per litre, prepared by EMBL Media Kitchen Facility); TCEP - tris(2-carboxyethyl)phosphine; PMSF – phenylmethanesulphonylfluoride;

A second nickel affinity purification step was performed to remove the 6xHis tag and the uncleaved fusion protein from the solution. The cleaved protein was further purified by SEC using a Superdex 200 16/60 column (GE Healthcare). When applicable, the fractions

of interests were dialysed step-wise against appropriate buffers, with each dialysis step carried out for at least 4 hours at 4 °C. The dialysed proteins were concentrated to desired concentrations using Vivaspin® Turbo 15 or/and Turbo 2 concentrators (molecular weight cut-off of 10 kDa, Sartorius stedim biotech) and kept at 4 °C or frozen in liquid nitrogen and stored at -80 °C until further use.

5.2.6 Sodium dodecyl-sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

Protein samples were routinely analysed by electrophoresis on 12% Sodium dodecyl-sulphate polyacrylamide (SDS-PA) gels. Protein samples were mixed with 4x SDS loading buffer [200 mM Tris-Cl, pH 6.8, 24% glycerol, 8% SDS, 0.02% bromophenol blue, and 0.5 M dithiothreitol (DTT)], heated to 95° C and loaded on SDS-PA gels [12% resolving gel: 0.375 M Tris-Cl, pH 8.8, 0.1% SDS, 12% Acrylamide: Bisacrylamide (37.5:1), 0.1% (w/v) ammonium persulfate (APS), 0.16% tetramethylethylenediamine (TEMED); 4% stacking gel: 0.125 M Tris-Cl, pH 8.8, 0.1% SDS, 4% Acrylamide: Bisacrylamide (37.5:1), 0.1% APS, 0.16% TEMED]. Gels were run for 45 min at 175 V in 1x Laemmli buffer (prepared by the EMBL Media Kitchen Facility). When applicable, precast NuPAGE 4-12% Bis-Tris Gels (Invitrogen) were used for electrophoresis according to the manufacturer's instructions. Protein marker used was Mark12^{TM} (Life Technologies). After the run, gels were stained with Coomassie staining solution containing 0.075% (w/v) G250 Coomassie Brilliant Blue (Thermo Scientific) and 0.1% (v/v) HCl. The products of gel electrophoresis were analysed under white light on a transilluminator (Alpha Innotech) using AlphaImager® HP software (Fisher Scientific).

The molecular weight and theoretical extinction coefficient of the purified proteins was estimated using the ProtParam bioinformatic tool provided by ExPASy (http://www.expasy.org). Protein samples were quantified based on their UV absorption at 280 nm wavelength measured using a Nanodrop spectrophotometer (Thermo Fisher Scientific).

5.2.7 Mass spectrometry

The efficiency of selenomethionine incorporation in the Hsmar1 derivative (see section 2.2.4) and the identity of SB degradation products (see section 3.2.2) were investigated by

mass spectrometry performed by the EMBL Proteomics Core Facility. In brief, the purified Hsmar1 sample was concentrated to >1 mg/ml and submitted to the Core Facility, whereas the SB degradation products identification was performed from PA gels. Mass determination was performed on a Q-Tof2 iMass Spectrometer (Micromass/Waters). In the case of the Hsmar1 derivative, the raw data were deconvoluted using MaxEnt1 software (Micromass/Waters) and the number of methionine sites occupied by selenomethionine was estimated based on the single peak with assigned mass.

5.3 Biochemical methods

5.3.1 Annealing of DNA substrates

All DNA substrates used in this work were annealed by mixing desired amounts of the oligonucleotides (resuspended in TE buffer, unless otherwise indicated), followed by incubation at 98 °C for 5 minutes and slow cooling (2-3 hours) in a switched-off heating block until the temperature reached the room temperature.

5.3.2 Radioactive labelling of DNA substrates

All oligonucleotides used in labelling reactions were unphosphorylated at the 5'-end. The labelling reaction contained 1x T4 Polynucleotide Kinase Reaction Buffer, 10 units of T4 Polynucleotide Kinase (both from NEB), 20 μM oligonucleotide, and 92.5 MBq (2.5 mCi)/ml of [γ-^{32}P]-ATP (PerkinElmer) in 10 μl reaction mix. The reactions were incubated at 37 °C for 1 hour, after which 40 μl of TE buffer was added. The kinase was then heat-inactivated by incubation at 80 °C for 30 minutes. In order to remove all unincorporated [γ-^{32}P]-ATP, the samples were applied to Micro Bio-Spin® Chromatography Columns (Bio-Rad) and purified according to the manufacturer's instructions. The Oligo Length Standards 10/60 and 20/100 Ladder markers (IDT) were labelled as above. To 40 μl of the marker, 60 μl of distilled water and 100 μl of 2x Formamide loading buffer [1x TBE (100 mM Tris base, 100 mM boric acid, 2 mM EDTA) 90% formamide, 0.005% xylene cyanol, and 0.005% bromophenol blue] were added.

5.3.3 *Hsmar1 in vitro* cleavage assay

Hsmar1 in vitro cleavage assay was based on reaction conditions described in (Claeys Bouuaert *et al.*, 2010). The Hsmar1 wild type and V119G mutant proteins were

overexpressed in *E. coli* and purified as described in sections 5.2.4 and 5.2.5. The DNA substrates used (Table 5-8) were 5' labelled with ^{32}P either on the NTS or on the TS.

The Hsmar1 wild type protein or V119G mutant (in 20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 500 mM NaCl, 1 mM $MgCl_2$ and 0.2 mM TCEP, pH 7.5) were mixed with 5' ^{32}P- labelled DNA substrates (80 nM) (sequences provided in Table 5-8) at a 1:2 molar ratio in Activity buffer (20 mM HEPES, 150 mM NaCl, 20 µg/ml bovine serum albumin (BSA), 1 mM $MgCl_2$ or $CaCl_2$ and 0.2 mM TCEP, pH 7.5) in the final volume of 10 µl. The control samples contained no protein.

Table 5-8: Oligonucleotides used in *Hsmar1 in vitro* cleavage assay.

Oligonucleotide	Sequence
***In vitro* cleavage assays with Hsmar1 wild type and V199G**	
IR-NTS	ACGGCCAGTGAATTCTATTAGGTTGGTGCAAAAGTAATTGCGGTTTTTGC
IR-TS	GCAAAAACCGCAATTACTTTTGCACCAACCTAATAGAATTCACTGGCCGT
N_1-NTS_1	ACGGCCAGTGAATTCTATTA
N_1-NTS_2	GGTTGGTGCAAAAGTAATTGCGGTTTTTGC
N_1-TS	GCAAAAACCGCAATTACTTTTGCACCAACCTAATAGAATTCACTGGCCGT
N_2-NTS_1	ACGGCCAGTGAATTCTATTAGGTTGGTGCAAAAGTAATTGCGGTTTTTGC
N_2-TS_1	GCAAAAACCGCAATTACTTTTGCACCAACCTAA
N_2-TS_2	TAGAATTCACTGGCCGT
In vitro* cleavage assays with Hsmar1 wild type and mutated IR	
IR-NTS_(-2)	ACGGCCAGTGAATTCCATTAGGTTGGTGCAAAAGTAATTGCGGTTTTTGC
IR-TS_(-2)	GCAAAAACCGCAATTACTTTTGCACCAACCTAATGGAATTCACTGGCCGT
IR-NTS_(-1)	ACGGCCAGTGAATTCTCTTAGGTTGGTGCAAAAGTAATTGCGGTTTTTGC
IR-TS_(-1)	GCAAAAACCGCAATTACTTTTGCACCAACCTAAGAGAATTCACTGGCCGT
IR-NTS_(1)	ACGGCCAGTGAATTCTACTAGGTTGGTGCAAAAGTAATTGCGGTTTTTGC
IR-TS_(1)	GCAAAAACCGCAATTACTTTTGCACCAACCTAGTAGAATTCACTGGCCGT
IR-NTS_(3)	ACGGCCAGTGAATTCTATTCGGTTGGTGCAAAAGTAATTGCGGTTTTTGC
IR-TS_(3)	GCAAAAACCGCAATTACTTTTGCACCAACCGAATAGAATTCACTGGCCGT
IR-NTS_(4)	ACGGCCAGTGAATTCTATTATGTTGGTGCAAAAGTAATTGCGGTTTTTGC
IR-TS_(4)	GCAAAAACCGCAATTACTTTTGCACCAACATAATAGAATTCACTGGCCGT

* Mutations are shown in red

Reactions were incubated at 37 °C for 18 h and terminated by Proteinase K (NEB) treatment according to the manufacturer's instructions. DNA was purified by ethanol precipitation: the sample was mixed with 1/10 sample volume of 3 M sodium acetate (pH 5.2), 3 sample volumes of absolute ethanol, and 20 µg of glycogen (Thermo Scientific), mixed, and stored at –20 °C overnight. Next, the DNA was resuspended in 10 µl of distilled water, 1x Formamide loading buffer (1x TBE, 90% formamide, 0.005% xylene cyanol, and 0.005% bromophenol blue) was added, and the sample was analysed by PAGE on 12% urea-TBE sequencing gels. The gel [12% Acrylamide: Bisacrylamide (19:1) (BioRad), 8.3 M Urea, 1x TBE buffer, 0.17% (w/v) APS, 0.05% (v/v) TEMED] was

prepared in a Model SA Sequencing Gel Electrophoresis Apparatus (Biometra). The gel was pre-run in 1x TBE buffer at 40 W for 30 minutes before loading and then run for 2 h hours at 20 W. After the run was finished, the gel was removed from the glass plates and a BAS-SR 2040 Phosphor Screen (FujiFilm) was exposed for 1 hour in a HypercassetteTM (Amersham Life Science). The screen was imaged in a Typhoon FLA 7000 Phosphoimager (GE Healthcare Life Sciences).

5.3.4 *SB in vitro* cleavage assay

SB in vitro cleavage assays (using SB or hsSB proteins; assays in Figure 2-22, Figure 2-23, Figure 2-24 and Figure 3-6) were performed according to (Voigt *et al.*, 2016) with some modifications. SB proteins were overexpressed in *E. coli* and purified as described in sections 5.2.4 and 5.2.5. The DNA substrates used (Table 5-9) were 5' ^{32}P-labelled either on the NTS or on the TS. SB derivatives (in 20 mM HEPES, 1 M NaCl, 1 mM $MgCl_2$ and 1 mM DTT, pH 8.0) were mixed with 5' ^{32}P- labelled Lo, Li, linked Lo-Li DNA or Lo-TATA DNA substrate (20 nM) (sequences shown in Table 5-9) at a 50:1 molar ratio in activity buffer (20 mM HEPES, 100 mM NaCl, 20 µg/ml bovine serum albumin (BSA), 10 mM $MgCl_2$ and 1 mM DTT, pH 8.0). Reactions were incubated at 25 °C for 18 h, processed and analysed on the 12% urea-TBE sequencing gel as previously described in section 5.3.3. For assays shown in Figure 2-23, unlabelled Li DNA was added to the reactions containing the Lo DNA (20 nM) at 1:2, 1: 1 and 2:1 final molar ratios (Li:Lo).

Table 5-9: Oligonucleotides used in *SB in vitro* cleavage assay.

Oligonucleotide	Sequence
***In vitro* cleavage assays with SB or hsSB**	
Lo-NTS	GAGCTGTGAACTCTACAGTTGAAGTCGGAAGTTTACATACACTTAAG
Lo-TS	CCACGGGACATGTTTATGCTTAAGTGTATGTAAACTTCCGACTTCAACTGTAGAGTTCACAGCTC
Li-NTS	GAGCTGTGAACTCGCTATCCAGTGGGTCAGAAGTTTACATACACTAAGTCAT
Li-TS	ATGACTTAGTGTATGTAAACTTCTGACCCACTGGATAGCGAGTTCACAGCTC
linked Lo-Li-NTS_1	GAGCTGTGAACTCTACAGTTGAAGTCGGAAGTTTACATACACTTAAG
linked Lo-Li-NTS_2	GAGCTGTGAACTCGCTATCCAGTGGGTCAGAAGTTTACATACACTAAGTCAT
linked Lo-Li-TS	ATGACTTAGTGTATGTAAACTTCTGACCCACTGGATAGCGAGTTCACAGCTCCCACGGGACATGT TTATGCTTAAGTGTATGTAAACTTCCGACTTCAACTGTAGAGTTCACAGCTC
Lo-TATA-NTS	GAGCTGTGAACTATACAGTTGAAGTCGGAAGTTTACATACACTTAAG
Lo-TATA-TS	CCACGGGACATGTTTATGCTTAAGTGTATGTAAACTTCCGACTTCAACTGTATAGTTCACAGCTC

5.3.5 *SB in vitro* integration assay

SB integration assays were performed using either 5' ^{32}P-labelled (assays in Figure 2-25 and Figure 2-28) or fluorescently labelled 5' 6-FAM (Fluorescein)-labelled (purchased with modification from IDT) (assays in Figure 3-6) DNA substrates. SB proteins were overexpressed in *E. coli* and purified as described in sections 5.2.4 and 5.2.5. The DNA substrates used (Table 5-10) were labelled on the TS.

In radioactive-based assays, SB or hsSB or GST-hsSB (1 µM in 20 mM HEPES, 500 mM NaCl and 0.2 mM TCEP, pH 8.0) and 60 nM 5' ^{32}P-labelled 2nt DNA or 3nt DNA (sequences and position of labelling shown in Table 5-10) were preincubated at 25 °C for 1 h in activity buffer (20 mM HEPES, 100 mM NaCl, 20 µg/ml BSA, 10 mM $MgCl_2$ and 1 mM DTT, pH 8.0) in a 6 µl reaction volume. 0.6 µg of target plasmid DNA pETMCN-YC was then added to a final reaction volume of 10 µl. A 4000 bp DNA (NoLimits 4000bp, Thermo Scientific) was also 5' ^{32}P-labelled (as described in section 5.3.1) and used as marker.

Table 5-10: Oligonucleotides used in *SB in vitro* integration assay.

Oligonucleotide	Sequence
***In vitro* integration assays with SB or hsSBor GST-hsSB**	
2nt-NTS	GTTGAAGTCGGAAGTTTACATACACTTAAG
2nt-TS	(6-FAM-)CCACGGGACATGTTTATGCTTAAGTGTATGTAAACTTCCGACTTCAACTG
3nt-NTS	TTGAAGTCGGAAGTTTACATACACTTAAG
3nt-TS	(6-FAM-)CCACGGGACATGTTTATGCTTAAGTGTATGTAAACTTCCGACTTCAACTG

6-FAM- 5' Fluorescein

Reactions were then incubated at 37 °C for 2 h (unless otherwise indicated) and terminated by Proteinase K (NEB) treatment according to the manufacturer's instructions. DNA was purified by ethanol precipitation: the sample was mixed with 1/10 sample volume of 3 M sodium acetate (pH 5.2) and 3 sample volumes of absolute ethanol, mixed, and stored at -20 °C overnight. Next, the DNA was resuspended in 10 µl of distilled water, 1x GelPilot Loading Dye (QIAGEN) was added, and the sample was analysed by gel electrophoresis on 1% agarose gel [1% (w/v) agarose powder dissolved in 1x Tris-acetate- EDTA (TAE) buffer (40 mM Tris, 20 mM acetic acid, and 1 mM EDTA)] in 1x TAE buffer. The voltage of 120 V was applied for 120 minutes. The gel was dried 1 h at 60 °C on a gel drying system (AlphaMetrix). A BAS-SR 2040 Phosphor Screen (FujiFilm) was then exposed for

1 hour in a HypercassetteTM (Amersham Life Science). The screen was imaged in a Typhoon FLA 7000 Phosphoimager (GE Healthcare Life Sciences).

When applicable, reaction samples were also run on a 1% alkaline (denaturing) agarose gel [1% (w/v) agarose powder dissolved in 30 mM NaCl, 2 mM EDTA, pH 7.5]. Prior to loading, the gel was immersed for 1 hour into alkaline electrophoresis buffer (30 mM NaOH, 2 mM EDTA) and 5 volumes of sample and ladder were diluted with one volume of 6X alkaline electrophoresis loading buffer (180 mM NaOH, 6 mM EDTA, 18% Ficoll 400, 0.05% bromcresol green). Samples and ladder were heated at 70 °C for 5 minutes and then loaded onto the gel. The voltage of 30 V was applied for 17 hours. After electrophoresis, the gel was immersed for 30 minutes in 7% trichloroacetic acid (TCA) and then was dried and imaged as previously described.

In fluorescence-based assays, indicated amount of SB or hsSB (in 20 mM HEPES, 500 mM NaCl and 0.2 mM TCEP, pH 8.0) and 1.5 μM 5'-6 FAM-labelled 2 nt DNA were preincubated at 25 °C for 1 h in activity buffer (20 mM HEPES, 100 mM NaCl, 20 μg/ml BSA, 10 mM $MgCl_2$ and 1 mM DTT, pH 8.0) in a 6 μl reaction volume. 5 μg of target plasmid DNA was then added to a final reaction volume of 10 μl. Reactions were then incubated at 37 °C for 2 h and processed by Proteinase K as previously described. DNA was purified by ethanol precipitation and resuspended in with GelPilot Loading Dye (QIAGEN). The sample was analysed by gel electrophoresis on 1% agarose gel (gel was not dried after run) and visualized by fluorescence using a Typhoon FLA 9500 fluorescence scanner (GE Healthcare).

5.3.6 Analytical size exclusion chromatography (SEC)

The oligomeric state of Hsmar1 wild type and W118P proteins and the Mos1-SB chimera was determined by analytical size exclusion chromatography (SEC) on a Superdex 200 3.2/30 column (GE Healthcare) using a ÄKTAmicro liquid chromatography system (GE Healthcare) according to the manufacturer's instructions. Hsmar1 wild type and W118P proteins were analysed in 20 mM HEPES, 200 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5; the Mos1-SB chimera in 20 mM HEPES, 350 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5.

Formation of nucleoprotein complexes, upon mixing of Hsmar1 and SB derivates with a variety of DNA substrates (Table 5-11), was assessed by SEC on Superdex 200 3.2/30 gel filtration column (GE Healthcare) using ÄKTAettan and ÄKTAmicro liquid chromatography systems (GE Healthcare).

Hsmar1 protein-Nicked NTS DNA complexes (in Figure 2-14) were formed by mixing the Hsmar1 wild type (concentration: 90 μM) with Nicked NTS DNA (sequences provided in Table 5-11) at a 1:1.5 molar ratio in Gel filtration buffer (Table 5-7). The sample was then dialysed in two steps against dialysis buffer A (20 mM HEPES, 500 mM NaCl, 1 mM CaCl2 and 0.2 mM TCEP, pH 7.5) and then dialysis buffer B (20 mM HEPES, 200 mM NaCl, 1 mM CaCl2 and 0.2 mM TCEP, pH 7.5). Sample was diluted 4-fold in dialysis buffer B before being loaded into the analytic gel filtration column.

Table 5-11: Oligonucleotides used for reconstitution of nucleoprotein complexes of Hsmar1 and SB protein derivatives analysed by analytical SEC.

Oligonucleotide	Sequence
Nicked NTS-TS	AAAAACCGCAATTACTTTTGCACCAACCTAATAGGA
Nicked NTS-NTS_1	TCCTATTA
Nicked NTS-NTS_2	GGTTGGTGCAAAAGTAATTGCGGTTTTT
Lo′-NTS	GTGAACTCTATACAGTTGAAGTCGGAAGTTTACATACACTTAAGG
Lo′-TS	CCTTAAGTGTATGTAAACTTCCGACTTCAACTGTATAGAGTTCAC
Ro′-NTS	CCTTAAGTGTATGTAAACTTCCGACTTCAACTGTATAGAGTTCAC
Ro′-TS	GTGAACTCTATACAGTTGAAGTCGGAAGTTTACATACACTTAAGG
Li′-NTS	GTGAACTCCGTCCAGTGGGTCAGAAGTTTACATACACTAAGTG
Li′-TS	CACTTAGTGTATGTAAACTTCTGACCCACTGGACGGAGTTCAC
Ri′-NTS	GTGAACTCCGCCCAGTGGGTCAGAAGTTAACATACACTCAATG
Ri′-TS	CATTGAGTGTATGTTAACTTCTGACCCACTGGGCGGAGTTCAC
Cleaved Lo-NTS	TTGAAGTCGGAAGTTTACATACACTTAAG
Cleaved Lo-TS	CTTAAGTGTATGTAAACTTCCGACTTCAACTG
Nicked Lo-NTS_1	CTATACA
Nicked Lo-NTS_2	GTTGAAGTCGGAAGTTTACATACACTTAAG
Nicked Lo-TS	CTTAAGTGTATGTAAACTTCCGACTTCAACTGTATAGG
Gapped Lo1-NTS_1	GCCTATA
Gapped Lo1-NTS_2	TTGAAGTCGGAAGTTTACATACACTTAAG
Gapped Lo1-TS	CTTAAGTGTATGTAAACTTCCGACTTCAACTGTATAGG
Gapped Lo2-NTS_1	TGAACTCTA
Gapped Lo2-NTS_2	TTGAAGTCGGAAGTTTACATACACTTAAGG
Gapped Lo2-TS	CTTAAGTGTATGTAAACTTCCGACTTCAACTGTATAGAGTTCAC
Gapped Lo3-NTS_1	GTGAACTCTA
Gapped Lo3-NTS_2	TTGAAGTCGGAAGTTTACATACACTTAAGG
Gapped Lo3-TS	CCTTAAGTGTATGTAAACTTCCGACTTCAACTGTATAGAGTTCACC
Gapped Lo-Li-NTS_1	GTGAACTCTA
Gapped Lo-Li-NTS_2	TTGAAGTCAGAAGTTTACATACACTAAGTG
Gapped Lo-Li-TS	CACTTAGTGTATGTAAACTTCTGACTTCAACTGTATAGAGTTCACC
Lo-DBD-NTS	TTGAAGTCGGAAGTTTACATACACG
Lo-DBD-TS	GTGTATGTAAACTTCCGACTTCAAC

SB protein-DNA mixtures analysed in Figure 2-17 and Figure 2-18 were formed by mixing SB or hsSB proteins (concentration: 23 μM) with different DNA substrates (sequences provided in Table 5-11) in Gel filtration buffer (Table 5-7), at specific molar ratios and dialyzed step-wise against complex buffers, as indicated in Table 5-12.

GST-hsSB protein-DNA complexes (Figure 2-28) were formed by mixing GST-hsSB fusion protein (concentration: 23 μM) with the Gapped Lo-Li DNA substrate (sequence provided in Table 5-11) in Gel filtration buffer (Table 5-7), at a 1:1.7 molar ratio and dialyzed step-wise against complex buffers, as indicated in Table 5-12.

The SB DBD protein (concentration: 26.86 μM) was incubated with the Lo DBD DNA substrate (sequence provided in Table 5-11) in Gel filtration buffer (Table 5-7), at a 1:1.5 molar ratio and dialyzed step-wise against complex buffers (in order, to buffers A, then B, and when applicable, C), as indicated in Table 5-12, to form SB DBD nucleoprotein complexes (Figure 2-21).

Table 5-12: Composition and reconstitution of SB protein derivative-DNA complexes analysed by analytical SEC.

Protein	DNA	Protein:DNA molar ratio	Dialysis buffer
SB	Lo′	1:1.7	A) 20 mM HEPES, 500 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 B) 20 mM HEPES, 350 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.6
SB	Ro′	1:1.7	A) 20 mM HEPES, 500 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 B) 20 mM HEPES, 350 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5
SB	Li′	1:1.7	A) 20 mM HEPES, 500 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 B) 20 mM HEPES, 350 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5
SB	Ri′	1:1.7	A) 20 mM HEPES, 500 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 B) 20 mM HEPES, 350 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5
SB	Cleaved Lo	1:1.1	40 mM HEPES, 500 mM NaCl, 1 mM $MgCl_2$ and 0.2 mM TCEP, pH 7.5
SB	Nicked Lo	1:1.5	20 mM HEPES, 500 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 8.0
SB	Gapped Lo1	1:2	A) 20 mM HEPES, 500 mM NaCl, 10 mM $CaCl_2$ and 0.2 mM TCEP, pH 8.0 B) 20 mM HEPES, 300 mM NaCl, 10 mM $CaCl_2$ and 0.2 mM TCEP, pH 8.0
SB	Gapped Lo2	1:1.5	A) 20 mM HEPES, 500 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 B) 20 mM HEPES, 300 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5
hsSB	Gapped Lo3	1:1.7	A) 20 mM HEPES, 500 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 B) 20 mM HEPES, 350 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 C) 20 mM AMPD, 250 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 9
hsSB	Gapped Lo-Li	1:1.7	A) 20 mM HEPES, 500 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 B) 20 mM HEPES, 350 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 C) 20 mM AMPD, 250 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 9
SB DBD	Lo-DBD	1:1.7	A) 20 mM HEPES, 500 mM NaCl and 0.2 mM TCEP, pH 7.5 B) 20 mM HEPES, 250 mM NaCl and 0.2 mM TCEP, pH 7.5
GST-hsSB	Gapped Lo-Li	1:1.7	A) 20 mM HEPES, 500 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 B) 20 mM HEPES, 350 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 C) 20 mM AMPD, 250 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 9

AMPD: 2-Amino-2-methyl-1,3-propanediol

Together with the protein-DNA mixtures, samples of protein and DNA oligonucleotides were analysed by analytical SEC in same complex buffer and at the same concentrations used for complex formation. In each case, 30 µl of each sample was filtered using Centrifugal Filter Units (Millipore) and loaded onto the gel filtration column prewashed with the appropriate buffer. The eluting samples were detected by UV absorbance at 280 nm (protein) and 260 nm (nucleic acid). The performance of the column and of the chromatography system was tested using Gel Filtration Calibration LMW standards (Sigma).

5.3.7 Site-directed protein-DNA disulfide crosslinking

Specific cysteine-substituted protein mutants were overexpressed in *E. coli* and purified as described in sections 5.2.4 and 5.2.5. DNA oligomers (Table 5-13) were synthetized and purified by PAGE (purchased from FUTUREsynthesis, Poznań, Poland). Backbone modification to a phosphoramidate with a cystamine functionalization (Figure 5-1) at the non-bridging position was present at specific sites of the oligonucleotides (as indicated by red asterisks in Table 5-13). DNA was dissolved in TE buffer supplemented with 50 mM KCl and annealed at 1 mM final concentration.

Prior to the disulfide crosslinking experiments, the cysteine-substituted mutants were tested for their ability to bind the LE DNA substrate by analytical SEC. SB protein mutants (concentration: 12.64 µM) were mixed with LE DNA (sequence provided in Table 5-13) at a 1:2 molar ratio in Gel filtration buffer (Table 5-7). The mixtures were then dialysed in two steps against first dialysis buffer (20 mM HEPES, 500 mM NaCl, 10 mM $CaCl_2$ and 0.2 mM TCEP, pH 8) and second dialysis buffer (20 mM HEPES, 400 mM NaCl, 10 mM $CaCl_2$ and 0.2 mM TCEP, pH 8), and then run on a on a Superdex 200 3.2/30 column (GE Healthcare) as described in section 5.3.6.

All crosslinking reactions were performed at a concentration of 6.8 µM protein and 10.2 µM DNA. Complexes were formed by mixing protein and DNA in Gel filtration buffer (Table 5-7), followed by a first dialysis step (against buffer containing 20 mM HEPES, 500 mM NaCl and 0.5 mM DTT, pH 7.5) and then a second dialysis step alternatively in dialysis buffer A (20 mM HEPES and 350 mM NaCl, pH 7.5), B (20 mM HEPES, 350 mM NaCl and 0.2 mM DTT, pH 7.5) or C (20 mM HEPES, 350 mM NaCl and 0.5 mM DTT, pH 7.5) (as indicated in Figure 2-21 C; for complexes shown in Figure 2-21 B and D,

the second dialysis step was performed with buffer B only). Samples of the mutant proteins alone were subjected to the same dialysis procedure. The crosslinking reaction was carried out at 25 °C for 24 h in a 15 µl reaction volume. 4x SDS loading buffer (with or without 300 mM DTT, as indicated in Figure 2-21) was added to the reactions that were then analysed by SDS-PAGE on precast NuPAGE 4-12% Bis-Tris Gels (Invitrogen) (as described in section 5.2.6). Apart from Coomassie staining, when applicable, gels were also stained with 1x SYBR® Gold Nucleic Acid Gel Stain (Life technologies) and visualized in the UV transilluminator (Alpha Innotech) using AlphaImager® HP software (Fisher Scientific).

Table 5-13: Oligonucleotides used for SB protein-DNA crosslinking experiments.

Oligonucleotide	Sequence
LE-NTS	TTGAAGTCGGAAGTTTACATACACTTAGG
LE-TS	CCTAAGTGTATGTAAACTTCCGACTTCAACTG
LE1-NTS	TTGAAGTCGGAAGTTTACATACACTTAGG
LE1-TS	CCTAAGTGTATGTAAACTTCCGACTTCA*ACTG
LE2-NTS	TTGAAGTCGGAAGTTTACATACACTTAGG
LE2-TS	CCTAAGTGTATGTAAACTTCCGACTTC*AACTG
LE3-NTS	TTGAAGTCGGAAGTTTACATACACTTAGG
LE3-TS	CCTAAGTGTA*TGTAAACTTCCGACTTCAACTG
LE4-NTS	TTGAAGTCGGAAGTTTACATACACTTAGG
LE4-TS	CCTAAGTGTATGTAAAC*TTCCGACTTCAACTG

Red asterisks (*) mark the site of the P-cystamine modification on the backbone phosphate.

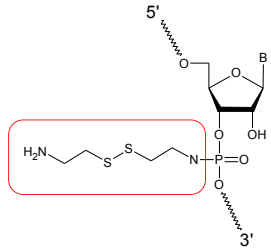


Figure 5-1: Structure of P-cystamine.

5.3.8 Dissucimidyl suberate (DSS)-mediated crosslinking of the SB protein

All samples (protein alone or nucleoprotein complexes) contained 23.76 µM SB protein and were dialyzed first against buffer containing 20 mM HEPES, 500 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 and then against a second dialysis buffer (20 mM HEPES, 350 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5). Nucleoprotein

complexes were formed by mixing SB protein and Gapped Lo-Li DNA (sequence in Table 5-11) in Gel filtration buffer (Table 5-7) and they were subjected to a third dialysis step against buffer containing 20 mM 2-amino-2-methyl-1,3-propanediol (AMPD), 250 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 9.0. All samples were incubated at 25 °C for 6 h in a 10 µl reaction volume. Then DSS [Thermo Fisher Scientific; dissolved in dimethylformamide (DMF)] at a final concentration of 2.5 µM was added to the samples for crosslinking. DMF only was added to control samples. All samples were incubated at 25 °C for 30 minutes. Crosslinking reactions were then stopped by adding 1 M TrisHCl pH 7.5 and by incubation at 25 °C for 15 minutes. Finally, samples were analysed by SDS-PAGE (as described in section 5.2.6) after addition of 4x SDS loading buffer (containing 100 mM DTT).

5.3.9 Site-specific cysteine-based crosslinking of the SB protein

Specific protein mutants ($SB_{R14C-C176S-I212S}$, $SB_{D17C-C176S-I212S}$ and $SB_{C176S-I212S}$) were overexpressed in *E. coli* and purified as described in sections 5.2.4 and 5.2.5. All samples (protein alone or nucleoprotein complexes) contained 22 µM SB derivatives and were dialyzed first against buffer containing 20 mM HEPES, 500 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5 and then against a second dialysis buffer (20 mM HEPES, 350 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 7.5). The protein alone samples were subjected to a third dialysis step against buffer containing 20 mM HEPES, 350 mM NaCl, 1 mM $CaCl_2$, pH 7.5. Nucleoprotein complexes were formed by mixing $SB_{D17C-C176S-I212S}$ protein and Gapped Lo3 DNA (sequence in Table 5-11) at 1:1.5 ratios in Gel filtration buffer (Table 5-7) and, after the above-mentioned two-step dialysis, they were subjected to a third dialysis step against buffer containing 20 mM AMPD, 250 mM NaCl, 1 mM $CaCl_2$ and 0.2 mM TCEP, pH 9.0. All samples were incubated at 25 °C for 48 h in a 10 µl reaction volume. Finally, samples were analysed by SDS-PAGE (as described in section 5.2.6) after addition of 4x SDS loading buffer (containing 300 mM DTT or no DTT, as indicated in Figure 2-27).

5.3.10 Limited proteolysis of SB proteins

Trypsin (Roche) was resuspended in 1% acetic acid in appropriate concentrations to obtain working aliquots that were snapped freeze in liquid nitrogen, and stored at -80 °C. Stock solution of protease were thawed and transferred to ice quickly before use. After

purification (see section 5.2.5), SB and hsSB proteins were dialysed against buffer containing 1x PBS and 300 mM NaCl, pH 7.5. Appropriate amounts of trypsin were mixed with 45 µg SB or hsSB protein to molar ratios of 1:100 or 1:1000 in a final reaction volume of 90 µl. All samples were incubated at 25 °C. After the desired incubation time, reactions were stopped by adding 4x SDS loading buffer and by incubating them at 70 °C for 10 minutes, followed by boiling at 98 °C for 5 minutes. Samples were then analysed by SDS-PAGE on precast NuPAGE 4-12% Bis-Tris Gels (Invitrogen) (as described in section 5.2.6).

5.3.11 Analysis of SB proteins' stability

Following purification (section 5.2.5), SB and hsSB proteins were dialysed against buffer containing 1x PBS and 300 mM NaCl, pH 7.5. 10 µg SB or hsSB protein were incubated at 37 °C in a final volume of 20 µl. After the desired incubation time, reactions were stopped by adding 4x SDS loading buffer (containing 300 mM DTT or no DTT, as indicated in Figure 3-4) and by boiling at 98 °C for 5 minutes. Samples were then analysed by SDS-PAGE on precast NuPAGE 4-12% Bis-Tris Gels (Invitrogen) (as described in section 5.2.6).

5.4 Biophysical methods

5.4.1 *In vitro* fluorescence-based oligomerization assay

A1, B1, A1-SB, B1-SB, A1-Hsmar1 and B1-Hsmar1 proteins were overexpressed in *E. coli* and purified as described in sections 5.2.4 and 5.2.5. Proteins [1 µM in 1x PBS, 1 M NaCl, 5% (v/v) glycerol, 1 mM $MgCl_2$ and 0.2 mM TCEP, pH 7.5] were measured directly or upon mixing with DNA (Cleaved Lo for SB and Cleaved IR for Hsmar1, see sequences in Table 5-14) at a 1:4 molar ratio in 40 mM HEPES, 200 mM NaCl, 5 mM $MgCl_2$ and 0.2 mM TCEP, pH 7.5. Samples were incubated for the indicated incubation times at 4℃ and 50 µl aliquots of each sample were measured in a 96- well plate reader (BioTeK Synergy4 from BioTeK Germany, Bad Friedrichshall, Germany) at 22℃ at excitation wavelength 535 nm. Emission was recorded at 600 nm. To evaluate statistical significance of the measurements, two-tailed two-sample *t*-test was performed.

Table 5-14: Oligonucleotides used for *in vitro* fluorescence-based oligomerization assay.

Oligonucleotide	Sequence
Cleaved Lo-NTS	TTGAAGTCGGAAGTTTACATACACTTAAG
Cleaved Lo-TS	CTTAAGTGTATGTAAACTTCCGACTTCAACTG
Cleaved IR-NTS	GGTTGGTGCAAAAGTAATTGCGGTTTTT
Cleaved IR-TS	AAAAACCGCAATTACTTTTGCACCAACCTAA

5.4.2 Circular dichroism spectroscopy

SB and hsSB proteins were overexpressed in *E. coli* and purified as described in sections 5.2.4 and 5.2.5. After purification, protein samples were dialyzed against buffer containing 1x PBS and 300 mM NaCl, pH 7.5. Circular dichroism spectra of the protein samples (200 µl at a concentration of 7.6 µM) were recorded on a Jasco J-715 spectropolarimeter using fused silica cuvettes (1 mm palmlenght) that are maintained at temperatures between 10-25° C (± 0.1 °C) using a Jasco PTC-348W peltier temperature control system. Thermal stability was assessed by monitoring the ellipticity at 206 nm while heating the sample at a rate of 1 °C/min. Data were analysed using the JASCO software program.

5.5 Small Angle X-ray Scattering

5.5.1 Principles of Small Angle X-ray Scattering

Small Angle X-ray Scattering (SAXS) has been employed in this work to obtain low-resolution structural information on the Hsmar1 transposase. Therefore, a brief introduction to the basic principles of SAXS is provided in this section [composed based on (D. Svergun *et al.*, 2003)]. SAXS is an experimental technique that provides structural analysis and physical information for a variety of 1–100 nm and beyond particle systems. In biological applications, SAXS is used to determine the structure of biological macromolecules in terms of average particle size and shape.

In comparison to other structural biology methods, as nuclear magnetic resonance (NMR) and X-ray crystallography, SAXS provides some technical advantages and thus, represents a powerful complementary tool for these techniques. For example, application of solution NMR is limited to protein size, whereas SAXS can be used for studying small molecules as well as for large multi-molecular assemblies. Structure determination by X-ray crystallography may take several weeks or even years and requires generation of well-diffracting macromolecular crystals, whereas SAXS measurements take days and the

macromolecule can be analysed in solution. The SAXS method is accurate, mostly non-destructive and usually requires minimum amounts of sample. However, with SAXS it is not possible to obtain structural information of macromolecules at atomic resolution.

Conceptually, SAXS experiments are simple: a solution of macromolecules is exposed to a highly collimated X-ray beam (with wavelength λ typically around 0.15 nm) that scatters elastically between 0 and 5 degrees ("small angles"). Biological SAXS is often performed at synchrotron radiation sources because biological molecules normally scatter weakly and the measured solutions are dilute. The scattered intensity I(q) is recorded on a detector and plotted as a function of momentum transfer q ($q= 4\pi\sin\theta/\lambda$, where 2θ is the angle between the incident and scattered radiation). The random positions and orientations of particles in solution result in an isotropic intensity distribution that, for monodisperse non-interacting particles, is proportional to the scattering from a single particle averaged over all orientations.
From the intensity of the macromolecule-containing solution the scattering from only the solvent is subtracted to obtain a final scattering curve. Generally, in a SAXS experiment, several solutions with varying concentrations of the macromolecule are measured to check that concentration effects do not affect the scattering curve. The SAXS curve contains information about the gross structural features -shape, quaternary and tertiary structure- of the macromolecule. At low resolution (2-3 nm resolution) the curves are rapidly decaying functions of q essentially determined by the particle shape and differ between proteins with different folds and molecular masses. At medium resolution (2 to 0.5 nm) the differences are already less pronounced and above 0.5 nm resolution all curves are very similar.

Data analysis of the extrapolated scattering curve begins with the inspection of the start of the scattering curve in the region around q=0. If the region follows the Guinier approximation, the sample is not aggregated. The lowest resolution portion of the SAXS data curve is dictated by a single size parameter, the radius of gyration Rg, which is the square root of the average distance of each scatterer from the particle center. For ideal monodisperse systems, the so-called Guinier plot [Log(I(q)) versus q^2] should be a linear function, whose slope yields the radius of gyration Rg and intercept gives the forward-scattered intensity at zero angle, I(0). Whereas the radius of gyration Rg characterizes the particle size, the parameter I(0) is useful to determine the molecular mass. Furthermore, by computing the Fourier transform of the scattering curve into real space, it is possible to

derive a pair-distribution function P(r), which provides information on the distances between electrons in the scattering sample and permits to calculate the maximum linear dimension (Dmax) of the macromolecule. Moreover, the Porod plot [q^4 I(q) versus q] provides other useful information such as an estimation of the particle volume (Porod volume).

From the SAXS curve is also possible to obtain an *ab initio* low-resolution model of the macromolecule. However, reconstruction of a three-dimensional model of an object from its one-dimensional scattering pattern is a challenging task. In fact, SAXS data does not imply a single solution and thus, reconstruction of the 3D structure might result in large number of different models. One of the available approaches is to use an ensemble of dummy residues (each amino acid is represented as one entity) and simulated annealing to build a locally "chain-compatible" dummy residue model inside a sphere of diameter Dmax, as implemented in the program GASBOR (D. I. Svergun *et al.*, 2001). Furthermore, the X-ray scattering data can be used to fit separate domains (X-ray or NMR structures) into the SAXS envelope to generate models of macromolecular complexes.

5.5.2 Experimental procedures

SAXS data were collected at 293 K at 0.992 Å on beamline BM-29 at the European Synchrotron Radiation Facility (ESRF, Grenoble, France). The sample to detector distance was 2.87 m, and scattering data were collected within the momentum transfer (q) range 0.001–0.50 Å. Data were collected in multiple 1 s frames, inspected and averaged in PRIMUS (Konarev *et al.*, 2003) normalized to the incident beam intensity and the scattering of the buffer subtracted. To check for concentration-dependent effects, scattering data were collected from Hsmar1 samples (Hsmar1 wild type and Hsmar1 W188P proteins) (40µl in 20 mM HEPES, 500 mM NaCl, 0.2 mM TCEP, pH 7.5) at 1.0 mg/ml, 2.0 mg/ml, 3.0 mg/ml and 4.0 mg/ml. The averaged curved shown in Figure 2-12 correspond to 1.0 mg/mL solution. SAXS data were calibrated against bovine serum albumin (BSA) at 4.3 mg/ml. Radius of gyration Rg and forward-scattered intensity at zero angle I(0) were determined by Guinier analysis. The SAXS data were transformed using GNOM (Petoukhov *et al.*, 2007) into a distribution of paired distances, P(r), from which the maximum dimension (Dmax) of the transposase and the Porod volume (Porod, 1982) were obtained. An estimate of the molecular mass (Mw) of the transposase was obtained from the Porod volume. Low resolution protein structure models were reconstructed *ab*

initio from the SAXS data by a chain-like ensemble of dummy residues using GASBOR (D. I. Svergun *et al.*, 2001). No symmetry was imposed for the calculations. Theoretical scattering curves for the Mos1 PEC and apo models were calculated from their crystallographic atomic coordinates using CRYSOL (D.I. Svergun *et al.*, 1995).

5.6 X-ray crystallography methods

5.6.1 Principles of biomolecular X-ray crystallography

X-ray crystallography has been employed in this work as a technique for macromolecular structure determination. Therefore, a brief introduction to the basic principles of biomolecular X-ray crystallography is provided in this section [composed based on (Rupp, 2010)]. X-ray crystallography is a powerful method that allows obtaining accurate and precise molecular structures of macromolecules (even of large molecular assemblies) at the atomic resolution level. About 90% of the structures deposited in Protein Data Bank (PDB) have been determined by X-ray crystallography.

In order to apply X-ray crystallography, well-ordered crystals of the macromolecules of interest need to be produced in the first place. Crystals are periodic assemblies of macromolecules, held together by weak inter-molecular interactions. In order to form crystals, the protein solution must be brought into a supersaturated, thermodynamically metastable state, in which crystal nucleation and growth can occur while the solution equilibrates. Supersaturation is achieved through the addition of crystallizing (or precipitating) agents such as salts or poly-alcohols, and by the manipulation of various parameters that include temperature, ionic strength and pH. This state should be reached slowly enough to allow the growth of single, well-ordered crystals. The most common technique used for crystal growth is the vapour diffusion method, where a drop containing the protein, buffer and precipitant is allowed to equilibrate with a larger reservoir containing similar buffers and precipitants in higher concentrations. Searching for successful crystallization conditions requires sampling of a vast combination of parameters. Moreover, several crystallization methods, in combination with post-crystallization treatments, may be applied to obtain well-ordered crystals that diffract to atomic resolution.

Once crystals are obtained, a single-crystal diffraction experiment is carried out. Briefly, a single crystal mounted on a goniostat is exposed to a finely collimated, intense X-ray beam in the 5-20 keV energy range (circa 2.3 to 0.6 Å wavelength). Individual diffraction images are recorded on an area detector during small rotation increments of the crystal and combined into a diffraction data set. Since crystals are highly sensitive to radiation damage by the intense ionizing X-ray radiation, cooling of crystals to cryogenic temperatures (100 K) is general practice. The diffraction images consist of spots, also known as reflections, which are created by scattered X-rays in specific directions of constructive interference. For each reflection, its position *hkl* (describing the direction of diffraction) and its intensity I_{hkl} can be observed directly from the diffraction pattern. The diffraction images are not direct pictures of the molecule, rather transforms of the molecular shape into diffraction space (called reciprocal space). Each reflection position in the diffraction image (or the so-called reciprocal lattice) corresponds to a specific real lattice plane in the crystal and contains information about the entire molecular structure.

The basic mathematical tool of back-transformation from reciprocal diffraction space into direct molecular space is the Fourier transform (FT). FT allows deconvolution of any function into the sum of a series of sine functions with their individual amplitudes, frequencies and phase shifts. Each diffracted X-ray beam that results in a reflection spot in the diffraction experiment corresponds to one term in the Fourier sum that describes the image (precisely the electron density) of the macromolecule of interest. This means that if the amplitude, frequency, and the phase of each individual collected reflection (which together constitute the structure factor function) is available, the electron density distribution of the diffracting object can be reconstituted by performing an inverse FT.

However, the parameters available for all reflections in the diffraction pattern comprise their frequency, the angle of diffraction (θ; described by position *hkl*) and the amplitude (described by intensity I_{hkl}), whereas the phase information is missing from the obtained diffraction pattern. For this reason, in order to complete the Fourier summation, phase information has to be estimated separately. For this, the phases can be derived experimentally (experimental phasing) or from a structurally-related macromolecular model (molecular replacement).

The molecular replacement (MR) method is a phasing technique that can be applied when a structurally similar search model (sequence identity of approximately 30%) to a molecule in the target crystal structure is available. Thus, the phase values from the structure factors of the model are used as initial phase estimates for the new macromolecular structure. MR requires a rotational and translational search to determine the proper orientation and location of the search model in the unknown crystal structure. The overlap between search model and target structure is generally assessed by comparing the observed and calculated structure factor amplitudes or Patterson functions (density functions calculated as the FT of the squared structure function amplitudes without phases) of the data and the search model.

Experimental phasing by isomorphous replacement involves collecting at least two diffraction datasets: one on a crystal of the native protein of interest and one on its heavy atom derivative. The heavy atom derivative crystals can be produced by soaking of already formed crystals in heavy atom-containing crystallization solutions, or by heavy atom derivatization of the macromolecule, for example by replacing methionines with selenomethionines in a protein. Heavy atoms contain large number of electrons and therefore scatter X-rays strongly, resulting in increased reflection intensities at specific positions. The differences between the spot intensities in the diffraction patterns of native and derivative datasets can be used to determine the heavy atom substructure. Therefore, the phases for the heavy atoms can be calculated, which provide the initial phase estimate for the entire protein structure. Depending on the number of different derivative crystals used, the method is referred to as single isomorphous replacement (SIR) or multiple isomorphous replacement (MIR).

Experimental phasing can be also applied by measuring the so-called anomalous scattering (or anomalous dispersion) that corresponds to a phase shift of scattered photons by heavy atoms. This phase shift can be used to locate the heavy atom positions with Patterson methods, and then to determine the phases of all reflections. Commonly used is the multi-wavelength anomalous dispersion (MAD) phasing that uses multiple datasets collected on the same crystal at wavelengths corresponding to the absorption peak, inflection point, and remote positions on the heavy atom's absorption spectrum. If crystals are particularly sensitive to radiation damage, only one dataset at the absorption peak is collected and thus single-wavelength anomalous diffraction (SAD) phasing is performed.

Once the phases are estimated by any of the phasing methods described, an initial electron density map can be calculated from the structure factors. Based on the map, a model of the macromolecule is built either automatically or manually. The obtained initial phases are only approximate, and thus need to be further refined in an iterative process that includes manual model building and real space and reciprocal space refinement. During model building and refinement, atomic coordinates are optimized to maximize the agreement between measured and calculated structure factors. The fit of the model to the experimental data is constantly monitored by specific statistical indicators (R-factors). The final model is also validated in terms of protein chemistry and geometry, and if satisfactory, it is deposited in the PDB for public use.

5.6.2 Preparation of protein-DNA complexes for crystallization

For crystallization experiments, Hsmar1 protein-Nicked NTS DNA complexes were prepared as for SEC experiment (section 5.3.6) with at least 95%-pure protein, as determined by SDS-PAGE analysis, and PAGE purified oligonucleotides (IDT). Moreover, prior to its use, the required amount of Nicked NTS DNA was lyophilized using a vacuum concentrator (Vacufuge plus, Eppendorf) and Hsmar1 wild type protein (in Gel filtration buffer) was added to lyophilized DNA at a 1:1.5 molar ratio. The mixture was then subjected to dialysis as described in section 5.3.6.

SB DBD protein-Lo DBD DNA complexes were prepared as for SEC experiment (section 5.3.6). The dialysed protein-DNA mixture was then concentrated to a protein concentration of 3.6 mg/ml using Vivaspin® Turbo 2 concentrators (molecular weight cut-off of 3 kDa, Sartorius stedim biotech). The concentration of the protein was estimated assuming that the protein amount remained the same before and after complex formation and concentration.

5.6.3 Crystallization of protein-DNA complexes and post-crystallization treatments

In order to form crystals of macromolecular complexes, the macromolecular solution has to be brought to a supersaturated state. For this, the vapour diffusion method was used, in which the supersaturation of the macromolecule is accomplished by increasing the concentration of the macromolecule and the precipitant simultaneously. In detail, a 0.1- 2 µl drop of the macromolecule solution is mixed with a 0.1-2 µl drop of the precipitant

solution and the mixture is placed in a sealed environment with a larger (0.5-1 ml) reservoir of the precipitant solution.

The initial crystallization screens were prepared by the EMBL Crystallization Facility and included commercially available crystallization screens The Classics Suite (QIAGEN) and PEGs II (QIAGEN), as well as the JCSG+ screen (Page *et al.*, 2003) prepared in-house. The screens were prepared in 96-well MRC sitting drop Crystallization PlatesTM (Molecular Dimensions) using a Mosquito Crystallization Robot (TTP Labtech). To a 0.1 µl drop of the reservoir precipitant solution, 0.1 µl of the protein-DNA complex solution was added and the plates were sealed. The screens were handled and stored for crystal growth at 20 °C.

Crystals likely containing SB DBD-Lo DBD DNA were obtained in 18% (w/v) PEG4000 and 0.1 M sodium acetate, pH 4.6. However, the presence of protein in the crystals could not be unambiguously assed and crystal optimization attempts resulted in thin clusters of needles, incompatible with data collection.

For Hsmar1 protein-DNA crystals, the initial condition that yielded small crystals [20% (w/v) PEG3350 and 0.2 M sodium thiocyanate, pH 6.9] was further optimized to yield larger crystals. For this, fine screening of varying PEG3350 concentrations against varying pH was manually set up in 24-well crystallization plates (Hampton Research), as shown in Figure 5-2. After 500 µl of specified solution was placed in the reservoir, a 1-2 µl drop of the mother liquor was set on the glass cover slip and 1-2 µl of Hsmar1 protein-DNA complex solution was added depending on the desired ratios. The cover slips were placed over the reservoir and sealed, allowing for vapour diffusion to occur. The plates were set up and stored at 20 °C. This setup allowed for fine-tuning of parameters such as precipitant amount, pH, protein-DNA complex concentration, and precipitant-to-complex ratio in large drops that supported growth of large crystals. Additionally, sitting drop 24-well plates were also set up but yielded no crystals.

Crystals obtained after optimization were washed several times in drops containing the crystallogenic condition and were finally dissolved in water, and analysed by SDS-PAGE using precast NuPAGE 4-12% Bis-Tris Gels (Invitrogen) stained with Pierce^{TM} Silver Stain Kit (Thermo-Fischer) according to the manufacturer's instructions.

In addition, crystallization experiments were carried out at the High Throughput Crystallization Laboratory of the EMBL Grenoble Outstation (in collaboration with J.A. Márquez Laboratory) using the sitting-drop vapour-diffusion method and CrystalDirect plates (MiTeGen, Ithaca, USA). Crystallization experiments were set up with 500 nl sample and 500 nl crystallization solution on the inner surface of the films within a CrystalDirect plate using a Cartesian PixSys robot (Cartesian Technologies). The plates were sealed on their upper side with CrystalClear film (Hampton Research) and the experiments were incubated at 20 °C in a RockImager system (Formulatrix Inc.).

SCREEN 1

HEPES pH		% (w/v) PEG3350 10	15	17	19	19.5	20
		1	2	3	4	5	6
6.0	A						
6.5	B						
6.8	C						
6.9	D						

SCREEN 3

HEPES pH		% (w/v) PEG3350 20.5	21	23	25	27	30
		1	2	3	4	5	6
6.0	A						
6.5	B						
6.8	C						
6.9	D						

SCREEN 2

HEPES pH		% (w/v) PEG3350 10	15	17	19	19.5	20
		1	2	3	4	5	6
7.0	A						
7.2	B						
7.5	C						
8.0	D						

SCREEN 4

HEPES pH		% (w/v) PEG3350 20.5	21	23	25	27	30
		1	2	3	4	5	6
7.0	A						
7.2	B						
7.5	C						
8.0	D						

Figure 5-2: Optimization screen of crystallization parameters (precipitant concentration and pH) in 24-well crystallization plates. All wells contained 0.1 M HEPES buffer at the indicated pH, PEG 3350 at the indicated concentrations and 0.2 or 0.25 M sodium thiocyanate.

In order to further improve the diffraction power of the Hsmar1 protein-DNA crystals, the streak seeding method was used. In brief, the biggest crystals obtained from SCREEN 2 and 3 (growing in 20-25% (w/v) PEG 3350, 0.2-0.25 M sodium thiocyanate and 100 mM HEPES, pH 7.5) were crushed into micro crystals with a probe and transferred to a Seed Bead tube (Hampton Reseach) and treated according to the manufacturer's instruction to obtain a seed stock. Different dilutions of this seed stock were used for the streak seeding experiments, in which a seeding probe was used to deposit the seeds in a recipient drop containing the sample and reagent, pre-equilibrated for two hours (Figure 5-3). Streak seeding was carried out by running the tip of the probe in a straight line across the middle

of the drop. This procedure was repeated sequentially in more drops to obtain dilutions of the original seed stock.

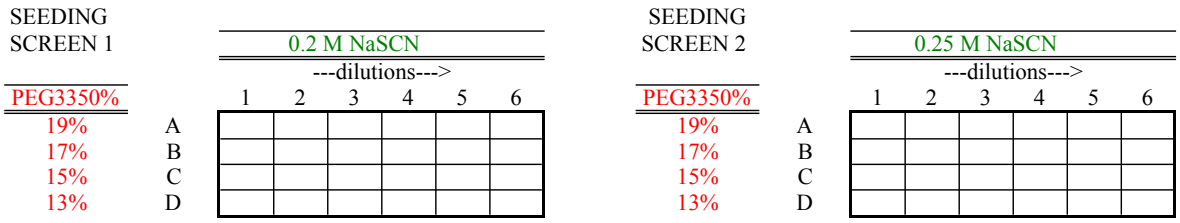


Figure 5-3: Seeding screen in 24-well crystallization plates. All wells contained 0.1 M HEPES buffer, pH 7.5, PEG 3350 at the indicated concentrations and 0.2 or 0.25 M sodium thiocyanate.

As post-crystallization treatments, dehydration protocols and annealing have been employed (Heras *et al.*, 2005). Crystal dehydration was performed by equilibrating the protein crystals over a reservoir with a higher percentage of precipitant (up to 25% PEG 3350). The hanging drop containing the crystals was allowed to dehydrate for 24 h. The annealing method was applied by transiently returning the flash-cooled crystal to ambient temperature. However, application of both dehydration and annealing protocols failed to improve the crystal diffraction quality.

5.6.4 Heavy atom derivative crystals

Heavy atom derivatization of Hsmar1 crystals was performed by crystal soaking. The platinum and cadmium derivative crystals were prepared by transferring the crystals to a 2 µl drop of the reservoir solution supplemented with 1 mM K_2PtCl_4, $K_2Pt(CN)_4$ or CdI_2 (Heavy Atom Screens^{TM}, Hampton Research) for 5-30 minutes, and then transferring the crystal back to a fresh drop of reservoir solution (back-soaking). The crystal were immediately fished and frozen in liquid nitrogen as described in section 5.6.5. Additionally, SeMet derivative crystals were obtained with purified SeMet-containing Hsmar1, following the same complex formation and crystallization protocols as for native complex crystallization (see sections 5.2.4 and 5.6.3).

5.6.5 Data collection

Crystals obtained from 24-well plates were fished using 0.05 – 0.4 mm CryoLoops (Hampton Research) and transferred immediately into liquid nitrogen-containing MicroTubes (Hampton Research). Alternatively, crystals were first flash-cooled to liquid

nitrogen temperatures by placing them in 100 K nitrogen gas cold stream and then transferred to liquid nitrogen-containing MicroTubes for storage. Crystals were transferred into solutions containing cryoprotecting agents [the optimal one being a solution of 2,3-butanediol at 20% (w/v)] before being frozen in liquid nitrogen. The crystals were transferred to the X-ray source in Cryogenic Dewar Flasks (TED PELLA) under cryogenic conditions.

X-ray diffraction data from native and derivative Hsmar1 protein-Nicked NTS DNA crystals were collected by the rotation method on tuneable beamlines ID-23 and ID-29 at the European Synchrotron Radiation Facility (ESRF) in Grenoble, France and on beamline P13 Petra-III in Hamburg, Germany. The samples were mounted onto the goniometer by an automatic sample changer and rotated as the data were collected. The best native dataset (1774 images) was collected on beamline P13 Petra-III at 12.7 keV using 0.15° oscillation and 0.037 sec exposure.

The best SeMet derivative datasets were collected at multiple wavelengths (peak and inflection point of the anomalous signal as well as at the high energy remote) on beamline P13 Petra-III using 0.10° oscillation and 0.097 sec exposure. The platinum and cadmium derivative crystals did not show any diffraction. At PetraIII P13 beamline, data collection was also performed on native crystals close to the absorption edge of calcium (6 keV) using 0.10° oscillation and 0.097 sec exposure.

5.6.6 Data processing

All datasets were initially processed in XDS, part of the XDS Program Package (Kabsch, 2010), to select the datasets with highest resolution and the best statistics. In addition, due to high diffraction anisotropy, available datasets, alone or in combination, are being currently processed by the STARANISO server (Tickle *et al.*, 2017). Together, the information from the data statistics table gave an indication for which datasets should be chosen for further data processing, and these datasets were internally scaled in XSCALE from the XDS Program Package, producing a final table of data statistics that are reported in the results sections (Figure 2-16). The reflection file was then converted to .mtz format using XDSCONV from the XDS Program Package, and the resulting reflection file was used for first structure solution attempts.

The initial attempts to solve the crystal structure of the crystallized complex used the molecular replacement method. For this, various homologous transposases structures and structure-derived models were used as search models in Phaser-MR (McCoy *et al.*, 2007), part of the Phenix suite. However, no correct solutions could be unambiguously identified.

5.7 Structural modelling

The N-terminal SB DNA binding domain, containing the dimer interface between two HTH domains, was modelled based on the Mos1 DBD [3HOT, (Richardson *et al.*, 2009)] using I-TASSER (Yang *et al.*, 2015). The full length Hsmar1 transposase apo dimer was modelled using the crystal structure of the Hsmar1 catalytic domain dimer and homology models of HTH1 and HTH2 [generated based on Mos1 by threading in Phyre2 (Karaca *et al.*, 2011; Kelley *et al.*, 2015)]. An ensemble of the Hsmar1 domains has been iteratively constructed and minimized against the low-resolution SAXS data [calculating the corresponding theoretical scattering curves with CRYSOL (D.I. Svergun *et al.*, 1995)]. Flexible parts have been manually built in Coot (Emsley *et al.*, 2004). The Hsmar1 PEC model was generated based on the Mos1 PEC [PDB: 3HOT, (Richardson *et al.*, 2009)] using I-TASSER (Yang *et al.*, 2015). Finally, all the assembled models were refined in HADDOCK (van Zundert *et al.*, 2016). The obtained models were visualized using the PyMOL Molecular Graphics System (Version 1.5.0.4, Schrödinger, LLC).

5.8 Cell biology methods

5.8.1 Cell culture

HeLa cells [American Type Culture Collection (ATCC); provided by Protein Expression and Purification Core Facility, EMBL Heidelberg] were maintained in Dulbecco's Modified Eagle Medium [DMEM low Glucose medium (Gibco)] supplemented with 10% (v/v) fetal bovine serum (FBS; Sigma) and 2 mM L-glutamine (Life Technologies). Chinese hamster ovary (CHO) cells (Invitrogen; R758-07) were maintained in Ham's F12 K (Kaighn's) medium (Gibco) containing 10% (v/v) FBS. Mouse ESCs (mESCs) [AB2.2 ES cell line (*Mus musculus*), ATCC, courtesy of S. Henkel, Neveu Laboratory, EMBL Heidelberg] were maintained without feeders in "serum plus LIF" medium composed of DMEM (high glucose, no glutamine, with sodium bicarbonate, Invitrogen) supplemented with 15% (v/v) ES-qualified fetal calf serum (EmbryoMax, Millipore), 10 ng/ml murine

Leukemia inhibitory factor (LIF; Protein Expression and Purification Core Facility, EMBL Heidelberg, Germany), 1× non-essential amino acids, 2 mM L-glutamine, 1 mM sodium pyruvate, 100 U/ml penicillin, 100 µg/ml streptomycin, and 0.1 mM β-mercaptoethanol (all Invitrogen) on culture dishes (Nunc) coated with 0.1% (v/v) gelatine (Sigma) solution.

The medium of all cell lines was changed daily and cells were passaged every other day with 0.05% Trypsin-EDTA (Invitrogen) at passaging ratios of 1/6–1/8. All cell lines were maintained at 37 °C in a humidified 5% CO_2 atmosphere in a Heracell™ 150i (Thermo Fisher Scientific).

5.8.2 *In vivo* transposition assay using SB coding plasmids

Transposition efficiency of SB and hsSB proteins was quantified in HeLa cells using a modified form of the protocol described by Ivics and colleagues (Ivics *et al.*, 1997). One day prior to transfection, cells were plated in a 6-well plate (Sigma) in 2 ml of complete media at a density of 3×10^5 cells per well. Transfection was performed using JetPei® DNA Transfection Reagent (Polyplus-transfection) at 60-80% of cell confluency following the manufacturer's instructions. 0.5 µg of SB_PGK-neo plasmid (containing the neomycin resistance gene flanked by the SB transposon ends) and 0.5 µg (unless otherwise indicated) of the SB_fl_pCMV(CAT)T7 (Mates *et al.*, 2009) or the SB-C176S-I212S_pCMV(CAT)T7 plasmid (encoding SB or hsSB protein, respectively) were mixed prior to transfection. All the transfection reactions were filled up to the same total DNA amount using a filler plasmid (pUC19). The recommended amount of DNA/JetPei® mixture was added in a drop-wise manner to the cells.

After a recovery period of 24 h, transfected cells were trypsinized in a total volume of 1 ml (in one well of a 6-well plate) and 10 or 100 µl aliquots of the cell suspension were re-plated (1:100 and 1:10, respectively) onto 10 cm dishes (Sigma). The medium was exchanged for selection medium containing 1 mg/ml G-418 (Geneticin, Life Technologies) 24 h later. Cells were selected for two weeks by replacing the selection medium twice a week. After selection, surviving colonies were fixed with 5% formaldehyde (Sigma), stained with Methylene blue (Sigma), imaged and counted using ImageJ software. In the assay in Figure 3-13, the final number of neomycin-resistant colonies shown in the graph has been corrected for the dilution used. Three independent experiments were performed

for each condition. To evaluate statistical significance, two-tailed one-sample *t*-test was performed.

5.8.3 Rapamycin-based *in vivo* transposition assay

Transposition activity of FRB- and FKBP- transposases (SB and Hsmar1) fusion proteins was quantified in HeLa cells using a modified version of the protocol described in (Ivics *et al.*, 1997) and in the previous section. One day prior to transfection with JetPei®, cells were plated in a 6-well plate at a density of 3 x 10^5 cells per well. 0.5 µg of donor plasmid SB_PGK-neo was co-trasfected with FRB_SB_fl_EWS and/or FKBP_SB_fl_EWS plasmids (0.5 µg each). 0.5 µg of donor plasmid pHsmar1-neo was co-trasfected with FRB_Hsmar1_fl_EWS and/or FKBP_Hsmar1_fl_EWS plasmids (0.03 µg each).

One hour after transfection, different concentrations of rapamycin (Sigma-Aldrich; provided by D. Yushchenko, EMBL Heidelberg) were added to the cells to induce dimerization. After a recovery period of 24 h, cells were trypsinized and a 1:100 dilution of the cell suspension was plated onto 10 cm Petri dishes. Selection, staining and quantification of the surviving colonies were performed as previously described in section 5.8.2. Three independent experiments were performed for each condition. The number of neomycin resistant colonies for all conditions was normalized to the untreated sample (no rapamycin; corresponding to 100% transposition rates). To evaluate statistical significance, two-tailed one-sample *t*-test was performed.

5.8.4 *In vivo* transposition assays by hsSB protein delivery in HeLa cells, CHO cells and mESCs

hsSB protein was prepared for electroporation as follows: purified hsSB was dialysed in two steps in buffer containing 20 mM HEPES, 300 mM NaCl, 1 mM $MgCl_2$ and 0.2 mM TCEP, pH 7.5 and concentrated to 20 mg/ml using Vivaspin 2 concentrators (Sartorius). Protein samples were filtered using 0.22 µm centrifugal filters (Millipore) prior to electroporation into cells. The activity of hsSB was quantified by drug selection or FACS using modified protocols from Ivics and colleagues (Ivics *et al.*, 1997). HeLa cells, CHO cells and mESCs were plated one day prior to transfection in a 6-well plate at a density of 3 × 10^5 cells per well. Transfection was performed with 1 µg of the SB_PGK-neo plasmid or the SB_T2/Venus plasmid (containing a neomycin resistance gene and a Venus gene,

respectively, flanked by the SB transposon ends). The NT (non-transfected) control samples were treated with transfection reagent, but no transposon plasmids. HeLa cells were transfected using JetPei® (Polyplus) transfection reagent. CHO and mESCs were transfected using FuGENE® (Invitrogen), according to the manufacturer's instructions, to improve transfection efficiency. After a recovery period of 4.5 h, cells were trypsinized, washed with PBS and resuspended in R buffer (provided with the Neon® system) at a cell density of 3–3.5 × 10^7 cells per ml. For each condition, 10 µl of the cell suspension was transfected with indicated amounts of hsSB using a Neon® transfection system (10 µl kit, Thermo Fisher Scientific) following settings listed in Table 5-15. Following electroporation, cells were plated onto 6-well plates containing medium and incubated at 37 °C in a 5% CO_2 atmosphere.

Table 5-15: Settings used for electroporation of different cell lines by the Neon® transfection system.

Cell lines	Voltage (V)	Time (ms)	Pulse number
HeLa	1005	35	2
CHO	1650	10	3
mESCs	1200	20	2

HeLa cells and CHO cells transfected with the SB_PGK-neo transposon donor plasmid were trypsinized after a recovery period of 24 h in a total volume of 1 ml, and 100 µl of the cell suspension was plated onto 10 cm dishes containing medium. Selection, staining and quantification of the surviving colonies were performed as previously described in section 5.8.2. HeLa cells, CHO cells and mESCs transfected with the SB_T2/Venus plasmid were sorted by flow cytometry 2 days after electroporation to obtain a pure population of Venus-expressing cells, which had acquired the SB_T2/Venus plasmid. Twenty-one days after electroporation, the percentage of Venus-positive cells (in whose genome the Venus cassette was stably integrated by hsSB) was quantified by flow cytometry to determine the transposition efficiency (as described in the following section).

5.8.5 Fluorescence-activated cell sorting (FACS)

Two days post-transfection, HeLa cells, CHO cells and mESCs were dissociated to a single-cell suspension with 0.05% Trypsin-EDTA (Invitrogen), resuspended in D-PBS containing 2% BSA and 2.5 mM EDTA, strained through a 40 µm cell strainer (BD Biosciences) and sorted on a BD FACSMelody™ or a BD FACSAria Fusion™ flow cytometer (BD Bioscience) in the EMBL Flow Cytometry Core Facility. Only $Venus^+$ cells

were collected for further culturing. After 21 days, the cells were measured for Venus expression and stained with propidium iodide (PI, at a final concentration of 1 µg/ml) or 4',6-diamidino-2-phenylindole (DAPI, at a final concentration of 0.1 µg/ml) to determine the ratio between living and dead cells. Flow cytometry measurements were performed on a BD LSRFortessa, BD Accuri or Beckman Coulter Cyan flow cytometer. When necessary, GFP/PI fluorescent spillover was compensated with the according controls. The results were analysed using FlowJo software or the FCS Express 4 Flow Cytometry software (De Novo Software).

5.8.6 Western blot analysis of HeLa cell lysate

Hela cells were treated for *in vivo* transposition using the protocols described above [using SB proteins (section 5.8.4) or SB proteins-encoding plasmids (section 5.8.2)]. At the indicated time points after electroporation, cells were harvested, washed twice with PBS and sonicated (using a VialTweeter sonifier, Hielscher Ultrasound Technology, set to 65% duty cycle and 50% output control, in 1 minute) in RIPA buffer (50 mM TrisHCl pH 7.4, 150 mM NaCl, 2 mM EDTA, 1% NP-40, 0.1% SDS) for lysis. Total protein extracts (20 µg) were separated by gel electrophoresis [using precast NuPAGE 4-12% Bis-Tris Gels (Invitrogen)] and transferred to a nitrocellulose membrane (Whatman® Protran®) using a semi-dry transfer system (Trans-Blot® SD Semi-Dry Electrophoretic Transfer Cell, Bio-Rad). Membranes were probed with Sleeping Beauty transposase antibody, polyclonal goat IgG (R&D Systems). The internal loading control was glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and was detected with an anti-GAPDH antibody (GAPDH Monoclonal Antibody, Cusabio). The intensities of the bands were measured using the Image Lab Software (Bio-Rad).

5.8.7 Cell surface immunostaining

To test the maintenance of pluripotency after transfection, mESCs were tested for surface expression of the self-renewal marker Oct4. Cells were fixed for 10 min with 4% Paraformaldehyde (PFA) in PBS and quenched with 150 mM glycine in PBS for 5 min at 25 °C. Cells were blocked for 30 min in PBS buffer containing 5% BSA and incubated overnight at 4 °C with anti-Oct4 antibody (Oct-3/4, Santa Cruz). Cells were washed and incubated with an anti-mouse secondary antibody [F(ab')2-goat anti-mouse IgG (H+L)

secondary antibody, Pacific blue conjugate from Invitrogen], for 2 h at at 25 °C. Cells were then analysed on a LSRFortessa flow cytometer (BD BioScience) (see section 5.8.5).

5.8.8 Sequence analysis of *SB* insertions in the HeLa cell genome

Genomic DNA from neomycin-resistant homogenous HeLa clones was purified using the GenElute™ Mammalian Genomic DNA Miniprep Kit (Sigma) according to the manufacturer's instructions. Neomycin transgene insertions were PCR amplified using specific primers for the left (Left_end) and right (Right_end) transposon ends and a random primer (Random_primer) (Table 5-16). The thermocycling conditions are shown in Table 5-3. All primers contained a specific barcode (underlined in the primer sequences) used for sequencing. Sequence analysis of the resulting PCR products was performed in the EMBL Genomics Core Facility (T. Rausch and V. Benes, EMBL Heidelberg).

Table 5-16: Primers used for amplification of *SB* insertions in the HeLa cell genome.

Primer	**Sequence***
Left_end	CTGGAGTTCAGACGTGTGCTCTTCCGATCTAGTAGATGTCCTAACTGACTTGC
Right_end	CTGGAGTTCAGACGTGTGCTCTTCCGATCTTCCTAACTGACCTAAGACAGG
Random_primer	CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNCTCAG

* Nucleotides belonging to the barcodes are underlined

Table 5-17: Thermocycling conditions for amplification of *SB* insertions in the HeLa cell genome.

Step	**Temperature**	**Time**	
1. Initial Denaturation	95 °C	3 minutes	
2. Denaturation	94 °C	20 seconds	
3. Annealing	63 °C	45 seconds	repeated 9 times
4. Extension	72 °C	3 minutes	
5. Denaturation	94 °C	20 seconds	
6. Annealing	63 °C	45 seconds	
7. Extension	72 °C	3 minutes	
8. Denaturation	94 °C	20 seconds	
9. Annealing	63 °C	45 seconds	repeated 12 times
10. Extension	72 °C	3 minutes	
11. Denaturation	94 °C	20 seconds	
12. Annealing	44 °C	1 minute	
13. Extension	72 °C	3 minutes	
14. Final Extension	72 °C	7 minutes	
15. Hold	4 °C	hold	

6 REFERENCES

Alekshun, M. N., & Levy, S. B. (2007). Molecular mechanisms of antibacterial multidrug resistance. *Cell, 128*(6), 1037-1050. doi:10.1016/j.cell.2007.03.004

Alford, S. C., Abdelfattah, A. S., Ding, Y., & Campbell, R. E. (2012). A fluorogenic red fluorescent protein heterodimer. *Chem Biol, 19*(3), 353-360. doi:10.1016/j.chembiol.2012.01.006

Ammar, I., Izsvak, Z., & Ivics, Z. (2012). The Sleeping Beauty transposon toolbox. *Methods Mol Biol, 859*, 229-240. doi:10.1007/978-1-61779-603-6_13

Aronovich, E. L., Bell, J. B., Belur, L. R., Gunther, R., Koniar, B., Erickson, D. C., Schachern, P. A., Matise, I., McIvor, R. S., Whitley, C. B., & Hackett, P. B. (2007). Prolonged expression of a lysosomal enzyme in mouse liver after Sleeping Beauty transposon-mediated gene delivery: implications for non-viral gene therapy of mucopolysaccharidoses. *J Gene Med, 9*(5), 403-415. doi:10.1002/jgm.1028

Auge-Gouillou, C., Brillet, B., Germon, S., Hamelin, M. H., & Bigot, Y. (2005). Mariner Mos1 transposase dimerizes prior to ITR binding. *J Mol Biol, 351*(1), 117-130. doi:10.1016/j.jmb.2005.05.019

Avramopoulou, V., Mamalaki, A., & Tzartos, S. J. (2004). Soluble, oligomeric, and ligand-binding extracellular domain of the human alpha7 acetylcholine receptor expressed in yeast: replacement of the hydrophobic cysteine loop by the hydrophilic loop of the ACh-binding protein enhances protein solubility. *J Biol Chem, 279*(37), 38287-38293. doi:10.1074/jbc.M402533200

Aziz, R. K., Breitbart, M., & Edwards, R. A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res, 38*(13), 4207-4217. doi:10.1093/nar/gkq140

Baker, T. A., Kremenstova, E., & Luo, L. (1994). Complete transposition requires four active monomers in the mu transposase tetramer. *Genes Dev, 8*(20), 2416-2428.

Balasubramanian, S., Wurm, F. M., & Hacker, D. L. (2016). Multigene expression in stable CHO cell pools generated with the piggyBac transposon system. *Biotechnol Prog, 32*(5), 1308-1317. doi:10.1002/btpr.2319

Banerjee, A., Santos, W. L., & Verdine, G. L. (2006). Structure of a DNA glycosylase searching for lesions. *Science, 311*(5764), 1153-1157. doi:10.1126/science.1120288

Barry, E. G., Witherspoon, D. J., & Lampe, D. J. (2004). A bacterial genetic screen identifies functional coding sequences of the insect mariner transposable element Famar1 amplified from the genome of the earwig, Forficula auricularia. *Genetics, 166*(2), 823-833.

Basner-Tschakarjan, E., & Mingozzi, F. (2014). Cell-Mediated Immunity to AAV Vectors, Evolving Concepts and Potential Solutions. *Front Immunol, 5*, 350. doi:10.3389/fimmu.2014.00350

Baudry, C., Malinsky, S., Restituito, M., Kapusta, A., Rosa, S., Meyer, E., & Betermier, M. (2009). PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. *Genes Dev, 23*(21), 2478-2483. doi:10.1101/gad.547309

Bermejo-Rodriguez, C., & Perez-Mancera, P. A. (2015). Use of DNA transposons for functional genetic screens in mouse models of cancer. *Curr Opin Biotechnol, 35*, 103-110. doi:10.1016/j.copbio.2015.05.005

Bire, S., Casteret, S., Arnaoty, A., Piegu, B., Lecomte, T., & Bigot, Y. (2013). Transposase concentration controls transposition activity: myth or reality? *Gene, 530*(2), 165-171. doi:10.1016/j.gene.2013.08.039

Bouuaert, C. C., Tellier, M., & Chalmers, R. (2014). One to rule them all: A highly conserved motif in mariner transposase controls multiple steps of transposition. *Mob Genet Elements, 4*(1), e28807. doi:10.4161/mge.28807

Brillet, B., Bigot, Y., & Auge-Gouillou, C. (2007). Assembly of the Tc1 and mariner transposition initiation complexes depends on the origins of their transposase DNA binding domains. *Genetica, 130*(2), 105-120. doi:10.1007/s10709-006-0025-2

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods, 10*(12), 1213-1218. doi:10.1038/nmeth.2688

Cai, Y., Bak, R. O., Krogh, L. B., Staunstrup, N. H., Moldt, B., Corydon, T. J., Schroder, L. D., & Mikkelsen, J. G. (2014). DNA transposition by protein transduction of the piggyBac transposase from lentiviral Gag precursors. *Nucleic Acids Res, 42*(4), e28. doi:10.1093/nar/gkt1163

CAR T cells - what have we learnt? (2017). *Nat Rev Clin Oncol, 15*(1), 1. doi:10.1038/nrclinonc.2017.196

Carpentier, C. E., Schreifels, J. M., Aronovich, E. L., Carlson, D. F., Hackett, P. B., & Nesmelova, I. V. (2014). NMR structural analysis of Sleeping Beauty transposase binding to DNA. *Protein Sci, 23*(1), 23-33. doi:10.1002/pro.2386

Carpentier, G., Jaillet, J., Pflieger, A., Adet, J., Renault, S., & Auge-Gouillou, C. (2011). Transposase-transposase interactions in MOS1 complexes: a biochemical approach. *J Mol Biol, 405*(4), 892-908. doi:10.1016/j.jmb.2010.11.032

Cary, L. C., Goebel, M., Corsaro, B. G., Wang, H. G., Rosen, E., & Fraser, M. J. (1989). Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology, 172*(1), 156-169.

Cesana, D., Sgualdino, J., Rudilosso, L., Merella, S., Naldini, L., & Montini, E. (2012). Whole transcriptome characterization of aberrant splicing events induced by lentiviral vector integrations. *J Clin Invest, 122*(5), 1667-1676. doi:10.1172/JCI62189

Chen, X., Shen, Y., Draper, W., Buenrostro, J. D., Litzenburger, U., Cho, S. W., Satpathy, A. T., Carter, A. C., Ghosh, R. P., East-Seletsky, A., Doudna, J. A., Greenleaf, W. J., Liphardt, J. T., & Chang, H. Y. (2016). ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat Methods, 13*(12), 1013-1020. doi:10.1038/nmeth.4031

Chenais, B., Caruso, A., Hiard, S., & Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene, 509*(1), 7-15. doi:10.1016/j.gene.2012.07.042

Cheng, C. Y., Vogt, A., Mochizuki, K., & Yao, M. C. (2010). A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in Tetrahymena thermophila. *Mol Biol Cell, 21*(10), 1753-1762. doi:10.1091/mbc.E09-12-1079

Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet, 18*(2), 71-86. doi:10.1038/nrg.2016.139

Chusainow, J., Yang, Y. S., Yeo, J. H., Toh, P. C., Asvadi, P., Wong, N. S., & Yap, M. G. (2009). A study of monoclonal antibody-producing CHO cell lines: what makes a stable high producer? *Biotechnol Bioeng, 102*(4), 1182-1196. doi:10.1002/bit.22158

Cipriani, F., Rower, M., Landret, C., Zander, U., Felisaz, F., & Marquez, J. A. (2012). CrystalDirect: a new method for automated crystal harvesting based on laser-induced photoablation of thin films. *Acta Crystallogr D Biol Crystallogr, 68*(Pt 10), 1393-1399. doi:10.1107/S0907444912031459

Claeys Bouuaert, C., & Chalmers, R. (2010). Transposition of the human Hsmar1 transposon: rate-limiting steps and the importance of the flanking TA dinucleotide in second strand cleavage. *Nucleic Acids Res, 38*(1), 190-202. doi:10.1093/nar/gkp891

Claeys Bouuaert, C., & Chalmers, R. (2013). Hsmar1 transposition is sensitive to the topology of the transposon donor and the target. *PLoS One, 8*(1), e53690. doi:10.1371/journal.pone.0053690

Claeys Bouuaert, C., & Chalmers, R. (2017). A single active site in the mariner transposase cleaves DNA strands of opposite polarity. *Nucleic Acids Res, 45*(20), 11467-11478. doi:10.1093/nar/gkx826

Claeys Bouuaert, C., Lipkow, K., Andrews, S. S., Liu, D., & Chalmers, R. (2013). The autoregulation of a eukaryotic DNA transposon. *Elife, 2*, e00668. doi:10.7554/eLife.00668

Claeys Bouuaert, C., Liu, D., & Chalmers, R. (2011). A simple topological filter in a eukaryotic transposon as a mechanism to suppress genome instability. *Mol Cell Biol, 31*(2), 317-327. doi:10.1128/MCB.01066-10

Claeys Bouuaert, C., Walker, N., Liu, D., & Chalmers, R. (2014). Crosstalk between transposase subunits during cleavage of the mariner transposon. *Nucleic Acids Res, 42*(9), 5799-5808. doi:10.1093/nar/gku172

Clark, K. J., Carlson, D. F., Leaver, M. J., Foster, L. K., & Fahrenkrug, S. C. (2009). Passport, a native Tc1 transposon from flatfish, is functionally active in vertebrate cells. *Nucleic Acids Res, 37*(4), 1239-1247. doi:10.1093/nar/gkn1025

Clayes Bouuaert, C., & Chalmers, R. (2010). Transposition of the human Hsmar1 transposon: rate-limiting steps and the importance of the flanking TA dinucleotide in second strand cleavage. *Nucleic Acids Res, 38*(1), 190-202. doi:10.1093/nar/gkp891

Collier, L. S., Carlson, C. M., Ravimohan, S., Dupuy, A. J., & Largaespada, D. A. (2005). Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature, 436*(7048), 272-276. doi:10.1038/nature03681

Cordaux, R., Udit, S., Batzer, M. A., & Feschotte, C. (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci U S A, 103*(21), 8101-8106. doi:10.1073/pnas.0601161103

Cozzolino, M., Amori, I., Pesaresi, M. G., Ferri, A., Nencini, M., & Carri, M. T. (2008). Cysteine 111 affects aggregation and cytotoxicity of mutant Cu,Zn-superoxide dismutase associated with familial amyotrophic lateral sclerosis. *J Biol Chem, 283*(2), 866-874. doi:10.1074/jbc.M705657200

Craig, N. (2002). *Mobile DNA II*.

Cui, Z., Geurts, A. M., Liu, G., Kaufman, C. D., & Hackett, P. B. (2002). Structure-function analysis of the inverted terminal repeats of the sleeping beauty transposon. *J Mol Biol, 318*(5), 1221-1235.

Cuypers, M. G., Trubitsyna, M., Callow, P., Forsyth, V. T., & Richardson, J. M. (2013). Solution conformations of early intermediates in Mos1 transposition. *Nucleic Acids Res, 41*(3), 2020-2033. doi:10.1093/nar/gks1295

Daboussi, M. J., Langin, T., & Brygoo, Y. (1992). Fot1, a new family of fungal transposable elements. *Mol Gen Genet, 232*(1), 12-16.

Davis, R. P., Nemes, C., Varga, E., Freund, C., Kosmidis, G., Gkatzis, K., de Jong, D., Szuhai, K., Dinnyes, A., & Mummery, C. L. (2013). Generation of induced pluripotent stem cells from human foetal fibroblasts using the Sleeping Beauty transposon gene delivery system. *Differentiation, 86*(1-2), 30-37. doi:10.1016/j.diff.2013.06.002

Dawson, A., & Finnegan, D. J. (2003). Excision of the Drosophila mariner transposon Mos1. Comparison with bacterial transposition and V(D)J recombination. *Mol Cell, 11*(1), 225-235.

de la Rosa, J., Weber, J., Friedrich, M. J., Li, Y., Rad, L., Ponstingl, H., Liang, Q., de Quiros, S. B., Noorani, I., Metzakopian, E., Strong, A., Li, M. A., Astudillo, A., Fernandez-Garcia, M. T., Fernandez-Garcia, M. S., Hoffman, G. J., Fuente, R., Vassiliou, G. S., Rad, R., Lopez-Otin, C., Bradley, A., & Cadinanos, J. (2017). A single-copy Sleeping Beauty transposon mutagenesis screen identifies new PTEN-cooperating tumor suppressor genes. *Nat Genet, 49*(5), 730-741. doi:10.1038/ng.3817

Dornan, J., Grey, H., & Richardson, J. M. (2015). Structural role of the flanking DNA in mariner transposon excision. *Nucleic Acids Res, 43*(4), 2424-2432. doi:10.1093/nar/gkv096

Dupuy, A. J., Akagi, K., Largaespada, D. A., Copeland, N. G., & Jenkins, N. A. (2005). Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature, 436*(7048), 221-226. doi:10.1038/nature03691

Dupuy, A. J., Jenkins, N. A., & Copeland, N. G. (2006). Sleeping beauty: a novel cancer gene discovery tool. *Hum Mol Genet, 15 Spec No 1*, R75-79. doi:10.1093/hmg/ddl061

Dyda, F., Chandler, M., & Hickman, A. B. (2012). The emerging diversity of transpososome architectures. *Q Rev Biophys, 45*(4), 493-521. doi:10.1017/S0033583512000145

Emmons, S. W., Yesner, L., Ruan, K. S., & Katzenberg, D. (1983). Evidence for a transposon in Caenorhabditis elegans. *Cell, 32*(1), 55-65.

Emsley, P., & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr, 60*(Pt 12 Pt 1), 2126-2132. doi:10.1107/S0907444904019158

Escobar, H., Schowel, V., Spuler, S., Marg, A., & Izsvak, Z. (2016). Full-length Dysferlin Transfer by the Hyperactive Sleeping Beauty Transposase Restores Dysferlin-deficient Muscle. *Mol Ther Nucleic Acids, 5*, e277. doi:10.1038/mtna.2015.52

Eyjolfsdottir, H., Eriksdotter, M., Linderoth, B., Lind, G., Juliusson, B., Kusk, P., Almkvist, O., Andreasen, N., Blennow, K., Ferreira, D., Westman, E., Nennesmo, I., Karami, A., Darreh-Shori, T., Kadir, A., Nordberg, A., Sundstrom, E., Wahlund, L. O., Wall, A., Wiberg, M., Winblad, B., Seiger, A., Wahlberg, L., & Almqvist, P. (2016). Targeted delivery of nerve growth factor to the cholinergic basal forebrain of Alzheimer's disease patients: application of a second-generation encapsulated cell biodelivery device. *Alzheimers Res Ther, 8*(1), 30. doi:10.1186/s13195-016-0195-9

Feng, X., & Colloms, S. D. (2007). In vitro transposition of ISY100, a bacterial insertion sequence belonging to the Tc1/mariner family. *Mol Microbiol, 65*(6), 1432-1443. doi:10.1111/j.1365-2958.2007.05842.x

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet, 9*(5), 397-405. doi:10.1038/nrg2337

Feschotte, C., & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet, 41*, 331-368. doi:10.1146/annurev.genet.40.110405.090448

Fesnak, A. D., June, C. H., & Levine, B. L. (2016). Engineered T cells: the promise and challenges of cancer immunotherapy. *Nat Rev Cancer, 16*(9), 566-581. doi:10.1038/nrc.2016.97

Fischer, S. E., van Luenen, H. G., & Plasterk, R. H. (1999). Cis requirements for transposition of Tc1-like transposons in C. elegans. *Mol Gen Genet, 262*(2), 268-274.

Fontana, A., de Laureto, P. P., Spolaore, B., Frare, E., Picotti, P., & Zambonin, M. (2004). Probing protein structure by limited proteolysis. *Acta Biochim Pol, 51*(2), 299-321. doi:035001299

Franz, G., & Savakis, C. (1991). Minos, a new transposable element from Drosophila hydei, is a member of the Tc1-like family of transposons. *Nucleic Acids Res, 19*(23), 6646.

Fu, A., Tang, R., Hardie, J., Farkas, M. E., & Rotello, V. M. (2014). Promises and pitfalls of intracellular delivery of proteins. *Bioconjug Chem, 25*(9), 1602-1608. doi:10.1021/bc500320j

Gaj, T., Guo, J., Kato, Y., Sirk, S. J., & Barbas, C. F., 3rd. (2012). Targeted gene knockout by direct delivery of zinc-finger nuclease proteins. *Nat Methods, 9*(8), 805-807. doi:10.1038/nmeth.2030

Galla, M., Schambach, A., Falk, C. S., Maetzig, T., Kuehle, J., Lange, K., Zychlinski, D., Heinz, N., Brugman, M. H., Gohring, G., Izsvak, Z., Ivics, Z., & Baum, C. (2011). Avoiding cytotoxicity of transposases by dose-controlled mRNA delivery. *Nucleic Acids Res, 39*(16), 7147-7160. doi:10.1093/nar/gkr384

Goodwin, K. D., He, H., Imasaki, T., Lee, S. H., & Georgiadis, M. M. (2010). Crystal structure of the human Hsmar1-derived transposase domain in the DNA repair enzyme Metnase. *Biochemistry, 49*(27), 5705-5713. doi:10.1021/bi100171x

Gorecka, K. M., Komorowska, W., & Nowotny, M. (2013). Crystal structure of RuvC resolvase in complex with Holliday junction substrate. *Nucleic Acids Res, 41*(21), 9945-9955. doi:10.1093/nar/gkt769

Grabundzija, I., Irgang, M., Mates, L., Belay, E., Matrai, J., Gogol-Doring, A., Kawakami, K., Chen, W., Ruiz, P., Chuah, M. K., VandenDriessche, T., Izsvak, Z., & Ivics, Z. (2010). Comparative analysis of transposable element vector systems in human cells. *Mol Ther, 18*(6), 1200-1209. doi:10.1038/mt.2010.47

Grabundzija, I., Wang, J., Sebe, A., Erdei, Z., Kajdi, R., Devaraj, A., Steinemann, D., Szuhai, K., Stein, U., Cantz, T., Schambach, A., Baum, C., Izsvak, Z., Sarkadi, B., & Ivics, Z. (2013). Sleeping Beauty transposon-based system for cellular reprogramming and targeted gene insertion in induced pluripotent stem cells. *Nucleic Acids Res, 41*(3), 1829-1847. doi:10.1093/nar/gks1305

Greenfield, N. J. (2006). Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc, 1*(6), 2876-2890. doi:10.1038/nprot.2006.202

Haapa, S., Taira, S., Heikkinen, E., & Savilahti, H. (1999). An efficient and accurate integration of mini-Mu transposons in vitro: a general methodology for functional genetic analysis and molecular biology applications. *Nucleic Acids Res, 27*(13), 2777-2784.

Hackett, P. B., Largaespada, D. A., & Cooper, L. J. (2010). A transposon and transposase system for human application. *Mol Ther, 18*(4), 674-683. doi:10.1038/mt.2010.2

Hackett, P. B., Largaespada, D. A., Switzer, K. C., & Cooper, L. J. (2013). Evaluating risks of insertional mutagenesis by DNA transposons in gene therapy. *Transl Res, 161*(4), 265-283. doi:10.1016/j.trsl.2012.12.005

Hare, S., Gupta, S. S., Valkov, E., Engelman, A., & Cherepanov, P. (2010). Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature, 464*(7286), 232-236. doi:10.1038/nature08784

Hare, S., Maertens, G. N., & Cherepanov, P. (2012). 3'-processing and strand transfer catalysed by retroviral integrase in crystallo. *EMBO J, 31*(13), 3020-3028. doi:10.1038/emboj.2012.118

Hargrove, P. W., Kepes, S., Hanawa, H., Obenauer, J. C., Pei, D., Cheng, C., Gray, J. T., Neale, G., & Persons, D. A. (2008). Globin Lentiviral Vector Insertions Can Perturb the Expression of Endogenous Genes in beta-thalassemic Hematopoietic Cells. *Mol Ther, 16*(3), 525-533. doi:10.1038/sj.mt.6300394

Hartl, D. (2001). Discovery of the transposable element mariner. *Genetics, 157*(2), 471-476.

Henssen, A. G., Henaff, E., Jiang, E., Eisenberg, A. R., Carson, J. R., Villasante, C. M., Ray, M., Still, E., Burns, M., Gandara, J., Feschotte, C., Mason, C. E., & Kentsis, A. (2015). Genomic DNA transposition induced by human PGBD5. *Elife, 4*. doi:10.7554/eLife.10565

Henssen, A. G., Koche, R., Zhuang, J., Jiang, E., Reed, C., Eisenberg, A., Still, E., MacArthur, I. C., Rodriguez-Fos, E., Gonzalez, S., Puiggros, M., Blackford, A. N., Mason, C. E., de Stanchina, E., Gonen, M., Emde, A. K., Shah, M., Arora, K., Reeves, C., Socci, N. D., Perlman, E., Antonescu, C. R., Roberts, C. W. M., Steen, H., Mullen, E., Jackson, S. P., Torrents, D., Weng, Z., Armstrong, S. A., & Kentsis, A. (2017). PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nat Genet, 49*(7), 1005-1014. doi:10.1038/ng.3866

Heras, B., & Martin, J. L. (2005). Post-crystallization treatments for improving diffraction quality of protein crystals. *Acta Crystallogr D Biol Crystallogr, 61*(Pt 9), 1173-1180. doi:10.1107/S0907444905019451

Hickman, A. B., & Dyda, F. (2016). DNA Transposition at Work. *Chem Rev, 116*(20), 12758-12784. doi:10.1021/acs.chemrev.6b00003

Houck, M. A., Clark, J. B., Peterson, K. R., & Kidwell, M. G. (1991). Possible horizontal transfer of Drosophila genes by the mite Proctolaelaps regalis. *Science, 253*(5024), 1125-1128.

Huang, C. R., Burns, K. H., & Boeke, J. D. (2012). Active transposition in genomes. *Annu Rev Genet, 46*, 651-675. doi:10.1146/annurev-genet-110711-155616

Huang, S., Tao, X., Yuan, S., Zhang, Y., Li, P., Beilinson, H. A., Zhang, Y., Yu, W., Pontarotti, P., Escriva, H., Le Petillon, Y., Liu, X., Chen, S., Schatz, D. G., & Xu, A. (2016). Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. *Cell, 166*(1), 102-114. doi:10.1016/j.cell.2016.05.032

Huang, X., Guo, H., Kang, J., Choi, S., Zhou, T. C., Tammana, S., Lees, C. J., Li, Z. Z., Milone, M., Levine, B. L., Tolar, J., June, C. H., Scott McIvor, R., Wagner, J. E., Blazar, B. R., & Zhou, X. (2008). Sleeping Beauty transposon-mediated engineering of human primary T cells for therapy of CD19+ lymphoid malignancies. *Mol Ther, 16*(3), 580-589. doi:10.1038/sj.mt.6300404

Huang, X., Haley, K., Wong, M., Guo, H., Lu, C., Wilber, A., & Zhou, X. (2010). Unexpectedly high copy number of random integration but low frequency of persistent expression of the Sleeping Beauty transposase after trans delivery in primary human T cells. *Hum Gene Ther, 21*(11), 1577-1590. doi:10.1089/hum.2009.138

Huang, X., Wilber, A. C., Bao, L., Tuong, D., Tolar, J., Orchard, P. J., Levine, B. L., June, C. H., McIvor, R. S., Blazar, B. R., & Zhou, X. (2006). Stable gene transfer and expression in human primary T cells by the Sleeping Beauty transposon system. *Blood, 107*(2), 483-491. doi:10.1182/blood-2005-05-2133

Hudecek, M., Izsvak, Z., Johnen, S., Renner, M., Thumann, G., & Ivics, Z. (2017). Going non-viral: the Sleeping Beauty transposon system breaks on through to the clinical side. *Crit Rev Biochem Mol Biol, 52*(4), 355-380. doi:10.1080/10409238.2017.1304354

Ihry, R.J., Worringer, K.A. , Salick, M.R., Frias, E., Ho, D., Theriault, K., Kommineni, S., Chen, J., Sondey, M., Ye, C., Randhawa, R., Kulkarni, T., Yang, Z., McAllister, G., Russ, C., Reece-Hoyes, J., Forrester, W., Hoffman, G.R., Dolmetsch, R., & Kaykas, A. (2017). P53 toxicity is a hurdle to CRISPR/CAS9 screening and engineering in human pluripotent stem cells. *BioRxiv (preprint)*.

Imbeault, M., Helleboid, P. Y., & Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature, 543*(7646), 550-554. doi:10.1038/nature21683

Ivics, Z. (2016). Endogenous Transposase Source in Human Cells Mobilizes piggyBac Transposons. *Mol Ther, 24*(5), 851-854. doi:10.1038/mt.2016.76

Ivics, Z., Hackett, P. B., Plasterk, R. H., & Izsvak, Z. (1997). Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell, 91*(4), 501-510.

Ivics, Z., Izsvak, Z., Minter, A., & Hackett, P. B. (1996). Identification of functional domains and evolution of Tc1-like transposable elements. *Proc Natl Acad Sci U S A, 93*(10), 5008-5013.

Ivics, Z., Li, M. A., Mates, L., Boeke, J. D., Nagy, A., Bradley, A., & Izsvak, Z. (2009). Transposon-mediated genome manipulation in vertebrates. *Nat Methods, 6*(6), 415-422. doi:10.1038/nmeth.1332

Izsvak, Z., Chuah, M. K., Vandendriessche, T., & Ivics, Z. (2009). Efficient stable gene transfer into human cells by the Sleeping Beauty transposon vectors. *Methods, 49*(3), 287-297. doi:10.1016/j.ymeth.2009.07.001

Izsvak, Z., Ivics, Z., & Plasterk, R. H. (2000). Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *J Mol Biol, 302*(1), 93-102. doi:10.1006/jmbi.2000.4047

Izsvak, Z., Khare, D., Behlke, J., Heinemann, U., Plasterk, R. H., & Ivics, Z. (2002). Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in Sleeping Beauty transposition. *J Biol Chem, 277*(37), 34581-34588. doi:10.1074/jbc.M204001200

Izsvak, Z., Stuwe, E. E., Fiedler, D., Katzer, A., Jeggo, P. A., & Ivics, Z. (2004). Healing the wounds inflicted by sleeping beauty transposition by double-strand break repair in mammalian somatic cells. *Mol Cell, 13*(2), 279-290.

Jacobson, J. W., Medhora, M. M., & Hartl, D. L. (1986). Molecular structure of a somatically unstable transposable element in Drosophila. *Proc Natl Acad Sci U S A, 83*(22), 8684-8688.

Jin, Z., Maiti, S., Huls, H., Singh, H., Olivares, S., Mates, L., Izsvak, Z., Ivics, Z., Lee, D. A., Champlin, R. E., & Cooper, L. J. (2011). The hyperactive Sleeping Beauty transposase SB100X improves the genetic modification of T cells to express a chimeric antigen receptor. *Gene Ther, 18*(9), 849-856. doi:10.1038/gt.2011.40

Johnen, S., Djalali-Talab, Y., Kazanskaya, O., Moller, T., Harmening, N., Kropp, M., Izsvak, Z., Walter, P., & Thumann, G. (2015). Antiangiogenic and Neurogenic Activities of Sleeping Beauty-Mediated PEDF-Transfected RPE Cells In Vitro and In Vivo. *Biomed Res Int, 2015*, 863845. doi:10.1155/2015/863845

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res, 110*(1-4), 462-467. doi:10.1159/000084979

Kabsch, W. (2010). Xds. *Acta Crystallogr D Biol Crystallogr, 66*(Pt 2), 125-132. doi:10.1107/S0907444909047337

Karaca, E., & Bonvin, AM. (2011). A multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes. *Structure, 19*, 555-565.

Kebriaei, P., Izsvak, Z., Narayanavari, S. A., Singh, H., & Ivics, Z. (2017). Gene Therapy with the Sleeping Beauty Transposon System. *Trends Genet, 33*(11), 852-870. doi:10.1016/j.tig.2017.08.008

Kebriaei, P., Singh, H., Huls, M. H., Figliola, M. J., Bassett, R., Olivares, S., Jena, B., Dawson, M. J., Kumaresan, P. R., Su, S., Maiti, S., Dai, J., Moriarity, B., Forget, M. A., Senyukov, V., Orozco, A., Liu, T., McCarty, J., Jackson, R. N., Moyes, J. S., Rondon, G., Qazilbash, M., Ciurea, S., Alousi, A., Nieto, Y., Rezvani, K., Marin, D., Popat, U., Hosing, C., Shpall, E. J., Kantarjian, H., Keating, M., Wierda, W., Do, K. A., Largaespada, D. A., Lee, D. A., Hackett, P. B., Champlin, R. E., & Cooper, L. J. (2016). Phase I trials using Sleeping Beauty to generate CD19-specific CAR T cells. *J Clin Invest, 126*(9), 3363-3376. doi:10.1172/JCI86721

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc, 10*(6), 845-858. doi:10.1038/nprot.2015.053

Kidwell, M. G. (1992). Horizontal transfer. *Curr Opin Genet Dev, 2*(6), 868-873.

Kim, S., Kim, D., Cho, S. W., Kim, J., & Kim, J. S. (2014). Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res, 24*(6), 1012-1019. doi:10.1101/gr.171322.113

Klebanoff, C. A., Rosenberg, S. A., & Restifo, N. P. (2016). Prospects for gene-engineered T cell immunotherapy for solid cancers. *Nat Med, 22*(1), 26-36. doi:10.1038/nm.4015

Kolacsek, O., Erdei, Z., Apati, A., Sandor, S., Izsvak, Z., Ivics, Z., Sarkadi, B., & Orban, T. I. (2014). Excision efficiency is not strongly coupled to transgenic rate: cell type-dependent transposition efficiency of sleeping beauty and piggyBac DNA transposons. *Hum Gene Ther Methods, 25*(4), 241-252. doi:10.1089/hgtb.2013.149

Konarev, Petr V., Volkov, Vladimir V., Sokolova, Anna V., Koch, Michel H. J., & Svergun, Dmitri I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *Journal of Applied Crystallography, 36*, 1277-1282.

Konnova, T. A., Singer, C. M., & Nesmelova, I. V. (2017). NMR solution structure of the RED subdomain of the Sleeping Beauty transposase. *Protein Sci, 26*(6), 1171-1181. doi:10.1002/pro.3167

Kuerten, D., Johnen, S., Harmening, N., Souteyrand, G., Walter, P., & Thumann, G. (2015). Transplantation of PEDF-transfected pigment epithelial cells inhibits corneal neovascularization in a rabbit model. *Graefes Arch Clin Exp Ophthalmol, 253*(7), 1061-1069. doi:10.1007/s00417-015-2954-x

Lampe, D. J. (2010). Bacterial genetic methods to explore the biology of mariner transposons. *Genetica, 138*(5), 499-508. doi:10.1007/s10709-009-9401-z

Lampe, D. J., Churchill, M. E., & Robertson, H. M. (1996). A purified mariner transposase is sufficient to mediate transposition in vitro. *EMBO J, 15*(19), 5470-5479.

Lampe, D. J., Grant, T. E., & Robertson, H. M. (1998). Factors affecting transposition of the Himar1 mariner transposon in vitro. *Genetics, 149*(1), 179-187.

Lampe, D. J., Walden, K. K., & Robertson, H. M. (2001). Loss of transposase-DNA interaction may underlie the divergence of mariner family transposable elements and the ability of more than one mariner to occupy the same genome. *Mol Biol Evol, 18*(6), 954-961. doi:10.1093/oxfordjournals.molbev.a003896

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., & International Human Genome Sequencing, Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062

Langin, T., Capy, P., & Daboussi, M. J. (1995). The transposable element impala, a fungal member of the Tc1-mariner superfamily. *Mol Gen Genet, 246*(1), 19-28.

Latella, M. C., Di Salvo, M. T., Cocchiarella, F., Benati, D., Grisendi, G., Comitato, A., Marigo, V., & Recchia, A. (2016). In vivo Editing of the Human Mutant Rhodopsin Gene by Electroporation of Plasmid-based CRISPR/Cas9 in the Mouse Retina. *Mol Ther Nucleic Acids, 5*(11), e389. doi:10.1038/mtna.2016.92

Lee, S. I., & Kim, N. S. (2014). Transposable elements and genome size variations in plants. *Genomics Inform, 12*(3), 87-97. doi:10.5808/GI.2014.12.3.87

Li, M., Mizuuchi, M., Burke, T. R., Jr., & Craigie, R. (2006). Retroviral DNA integration: reaction pathway and critical intermediates. *EMBO J, 25*(6), 1295-1304. doi:10.1038/sj.emboj.7601005

Liang, Q., Kong, J., Stalker, J., & Bradley, A. (2009). Chromosomal mobilization and reintegration of Sleeping Beauty and PiggyBac transposons. *Genesis, 47*(6), 404-408. doi:10.1002/dvg.20508

Liang, X., Potter, J., Kumar, S., Zou, Y., Quintanilla, R., Sridharan, M., Carte, J., Chen, W., Roark, N., Ranganathan, S., Ravinder, N., & Chesnut, J. D. (2015). Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. *J Biotechnol, 208*, 44-53. doi:10.1016/j.jbiotec.2015.04.024

Liu, D., Bischerour, J., Siddique, A., Buisine, N., Bigot, Y., & Chalmers, R. (2007). The human SETMAR protein preserves most of the activities of the ancestral Hsmar1 transposase. *Mol Cell Biol, 27*(3), 1125-1132. doi:10.1128/MCB.01899-06

Liu, D., & Chalmers, R. (2014). Hyperactive mariner transposons are created by mutations that disrupt allosterism and increase the rate of transposon end synapsis. *Nucleic Acids Res, 42*(4), 2637-2645. doi:10.1093/nar/gkt1218

Liu, J., Gaj, T., Patterson, J. T., Sirk, S. J., & Barbas, C. F., 3rd. (2014). Cell-penetrating peptide-mediated delivery of TALEN proteins via bioconjugation for genome engineering. *PLoS One, 9*(1), e85755. doi:10.1371/journal.pone.0085755

Locke, F. L., Neelapu, S. S., Bartlett, N. L., Siddiqi, T., Chavez, J. C., Hosing, C. M., Ghobadi, A., Budde, L. E., Bot, A., Rossi, J. M., Jiang, Y., Xue, A. X., Elias, M., Aycock, J., Wiezorek, J., & Go, W. Y. (2017). Phase 1 Results of ZUMA-1: A Multicenter Study of KTE-C19 Anti-CD19 CAR T Cell Therapy in Refractory Aggressive Lymphoma. *Mol Ther, 25*(1), 285-295. doi:10.1016/j.ymthe.2016.10.020

Lohe, A. R., Moriyama, E. N., Lidholm, D. A., & Hartl, D. L. (1995). Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. *Mol Biol Evol, 12*(1), 62-72. doi:10.1093/oxfordjournals.molbev.a040191

Lohe, A. R., Sullivan, D. T., & Hartl, D. L. (1996). Subunit interactions in the mariner transposase. *Genetics, 144*(3), 1087-1095.

Luo, G., Ivics, Z., Izsvak, Z., & Bradley, A. (1998). Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *Proc Natl Acad Sci U S A, 95*(18), 10769-10773.

Maertens, G. N., Hare, S., & Cherepanov, P. (2010). The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature, 468*(7321), 326-329. doi:10.1038/nature09517

Makarova, O., Kamberov, E., & Margolis, B. (2000). Generation of deletion and point mutations with one primer in a single cloning step. *Biotechniques, 29*(5), 970-972.

Mates, L., Chuah, M. K., Belay, E., Jerchow, B., Manoj, N., Acosta-Sanchez, A., Grzela, D. P., Schmitt, A., Becker, K., Matrai, J., Ma, L., Samara-Kuko, E., Gysemans, C., Pryputniewicz, D., Miskey, C., Fletcher, B., VandenDriessche, T., Ivics, Z., & Izsvak, Z. (2009). Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet, 41*(6), 753-761. doi:10.1038/ng.343

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A, 36*(6), 344-355.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., & Read, R. J. (2007). Phaser crystallographic software. *J Appl Crystallogr, 40*(Pt 4), 658-674. doi:10.1107/S0021889807021206

Medhora, M., Maruyama, K., & Hartl, D. L. (1991). Molecular and functional analysis of the mariner mutator element Mos1 in Drosophila. *Genetics, 128*(2), 311-318.

Menendez-Arias, L. (2009). Mutation rates and intrinsic fidelity of retroviral reverse transcriptases. *Viruses, 1*(3), 1137-1165. doi:10.3390/v1031137

Miskey, C., Izsvak, Z., Kawakami, K., & Ivics, Z. (2005). DNA transposons in vertebrate functional genomics. *Cell Mol Life Sci, 62*(6), 629-641. doi:10.1007/s00018-004-4232-7

Miskey, C., Papp, B., Mates, L., Sinzelle, L., Keller, H., Izsvak, Z., & Ivics, Z. (2007). The ancient mariner sails again: transposition of the human Hsmar1 element by a reconstructed transposase and activities of the SETMAR protein on transposon ends. *Mol Cell Biol, 27*(12), 4589-4600. doi:10.1128/MCB.02027-06

Mitra, R., Fain-Thornton, J., & Craig, N. L. (2008). piggyBac can bypass DNA synthesis during cut and paste transposition. *EMBO J, 27*(7), 1097-1109. doi:10.1038/emboj.2008.41

Moldt, B., Miskey, C., Staunstrup, N. H., Gogol-Doring, A., Bak, R. O., Sharma, N., Mates, L., Izsvak, Z., Chen, W., Ivics, Z., & Mikkelsen, J. G. (2011). Comparative genomic integration profiling of Sleeping Beauty transposons mobilized with high efficacy from integrase-defective lentiviral vectors in primary human cells. *Mol Ther, 19*(8), 1499-1510. doi:10.1038/mt.2011.47

Monjezi, R., Miskey, C., Gogishvili, T., Schleef, M., Schmeer, M., Einsele, H., Ivics, Z., & Hudecek, M. (2017). Enhanced CAR T-cell engineering using non-viral Sleeping Beauty transposition from minicircle vectors. *Leukemia, 31*(1), 186-194. doi:10.1038/leu.2016.180

Montano, S. P., Pigli, Y. Z., & Rice, P. A. (2012). The mu transpososome structure sheds light on DDE recombinase evolution. *Nature, 491*(7424), 413-417. doi:10.1038/nature11602

Montano, S. P., & Rice, P. A. (2011). Moving DNA around: DNA transposition and retroviral integration. *Curr Opin Struct Biol, 21*(3), 370-378. doi:10.1016/j.sbi.2011.03.004

Moriarity, B. S., Otto, G. M., Rahrmann, E. P., Rathe, S. K., Wolf, N. K., Weg, M. T., Manlove, L. A., LaRue, R. S., Temiz, N. A., Molyneux, S. D., Choi, K., Holly, K. J., Sarver, A. L., Scott, M. C., Forster, C. L., Modiano, J. F., Khanna, C., Hewitt, S. M., Khokha, R., Yang, Y., Gorlick, R., Dyer, M. A., & Largaespada, D. A. (2015). A Sleeping Beauty forward genetic screen identifies new genes and pathways driving osteosarcoma development and metastasis. *Nat Genet, 47*(6), 615-624. doi:10.1038/ng.3293

Morris, E. R., Grey, H., McKenzie, G., Jones, A. C., & Richardson, J. M. (2016). A bend, flip and trap mechanism for transposon integration. *Elife, 5*. doi:10.7554/eLife.15537

Munoz-Lopez, M., & Garcia-Perez, J. L. (2010). DNA transposons: nature and applications in genomics. *Curr Genomics, 11*(2), 115-128. doi:10.2174/138920210790886871

Munoz-Lopez, M., Siddique, A., Bischerour, J., Lorite, P., Chalmers, R., & Palomeque, T. (2008). Transposition of Mboumar-9: identification of a new naturally active mariner-family transposon. *J Mol Biol, 382*(3), 567-572. doi:10.1016/j.jmb.2008.07.044

Narayanavari, S. A., Chilkunda, S. S., Ivics, Z., & Izsvak, Z. (2017). Sleeping Beauty transposition: from biology to applications. *Crit Rev Biochem Mol Biol, 52*(1), 18-44. doi:10.1080/10409238.2016.1237935

Nenortas, D. B. E. (1995). Advances in biophysical chemistry, volume 5.

Nesmelova, I. V., & Hackett, P. B. (2010). DDE transposases: Structural similarity and diversity. *Adv Drug Deliv Rev, 62*(12), 1187-1195. doi:10.1016/j.addr.2010.06.006

Nowacki, M., Higgins, B. P., Maquilan, G. M., Swart, E. C., Doak, T. G., & Landweber, L. F. (2009). A functional role for transposases in a large eukaryotic genome. *Science, 324*(5929), 935-938. doi:10.1126/science.1170023

Nowotny, M. (2009). Retroviral integrase superfamily: the structural perspective. *EMBO Rep, 10*(2), 144-151. doi:10.1038/embor.2008.256

Omasa, T., Onitsuka, M., & Kim, W. D. (2010). Cell engineering and cultivation of chinese hamster ovary (CHO) cells. *Curr Pharm Biotechnol, 11*(3), 233-240.

Page, R., Grzechnik, S. K., Canaves, J. M., Spraggon, G., Kreusch, A., Kuhn, P., Stevens, R. C., & Lesley, S. A. (2003). Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the Thermotoga maritima proteome. *Acta Crystallogr D Biol Crystallogr, 59*(Pt 6), 1028-1037.

Petoukhov, M. V., Konarev, P. V., Kikhney, A. G. , & Svergun, D. I. . (2007). ATSAS 2.1 - towards automated and web-supported small-angle scattering data analysis. . *J. Appl. Crystallogr., 40* s223–s228

Petrakis, S., Rasko, T., Mates, L., Ivics, Z., Izsvak, Z., Kouzi-Koliakou, K., & Koliakos, G. (2012). Gateway-compatible transposon vector to genetically modify human embryonic kidney and adipose-derived stromal cells. *Biotechnol J, 7*(7), 891-897. doi:10.1002/biot.201100471

Pfeifer, A., Brandon, E. P., Kootstra, N., Gage, F. H., & Verma, I. M. (2001). Delivery of the Cre recombinase by a self-deleting lentiviral vector: efficient gene targeting in vivo. *Proc Natl Acad Sci U S A, 98*(20), 11450-11455. doi:10.1073/pnas.201415498

Piegu, B., Bire, S., Arensburger, P., & Bigot, Y. (2015). A survey of transposable element classification systems--a call for a fundamental update to meet the challenge of their diversity and complexity. *Mol Phylogenet Evol, 86*, 90-109. doi:10.1016/j.ympev.2015.03.009

Plasterk, R. H. (1996). The Tc1/mariner transposon family. *Curr Top Microbiol Immunol, 204*, 125-143.

Plasterk, R. H., Izsvak, Z., & Ivics, Z. (1999). Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends Genet, 15*(8), 326-332.

Poirot, L., Philip, B., Schiffer-Mannioui, C., Le Clerre, D., Chion-Sotinel, I., Derniame, S., Potrel, P., Bas, C., Lemaire, L., Galetto, R., Lebuhotel, C., Eyquem, J., Cheung, G. W., Duclert, A., Gouble, A., Arnould, S., Peggs, K., Pule, M., Scharenberg, A. M., & Smith, J. (2015). Multiplex Genome-Edited T-cell Manufacturing Platform for "Off-the-Shelf" Adoptive T-cell Immunotherapies. *Cancer Res, 75*(18), 3853-3864. doi:10.1158/0008-5472.CAN-14-3321

Porod, G. (1982). *General Theory Small Angle X-ray Scattering*.

Putyrski, M., & Schultz, C. (2012). Protein translocation as a tool: The current rapamycin story. *FEBS Lett, 586*(15), 2097-2105. doi:10.1016/j.febslet.2012.04.061

Ramakrishna, S., Kwaku Dad, A. B., Beloor, J., Gopalappa, R., Lee, S. K., & Kim, H. (2014). Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. *Genome Res, 24*(6), 1020-1027. doi:10.1101/gr.171264.113

Richardson, J. M., Colloms, S. D., Finnegan, D. J., & Walkinshaw, M. D. (2009). Molecular architecture of the Mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote. *Cell, 138*(6), 1096-1108. doi:10.1016/j.cell.2009.07.012

Richardson, J. M., Dawson, A., O'Hagan, N., Taylor, P., Finnegan, D. J., & Walkinshaw, M. D. (2006). Mechanism of Mos1 transposition: insights from structural analysis. *EMBO J, 25*(6), 1324-1334. doi:10.1038/sj.emboj.7601018

Robertson, H. M., & Lampe, D. J. (1995). Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. *Mol Biol Evol, 12*(5), 850-862. doi:10.1093/oxfordjournals.molbev.a040262

Robertson, H. M., & MacLeod, E. G. (1993). Five major subfamilies of mariner transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. *Insect Mol Biol, 2*(3), 125-139.

Robertson, H. M., & Zumpano, K. L. (1997). Molecular evolution of an ancient mariner transposon, Hsmar1, in the human genome. *Gene, 205*(1-2), 203-217.

Ru, R., Yao, Y., Yu, S., Yin, B., Xu, W., Zhao, S., Qin, L., & Chen, X. (2013). Targeted genome engineering in human induced pluripotent stem cells by penetrating TALENs. *Cell Regen (Lond), 2*(1), 5. doi:10.1186/2045-9769-2-5

Ruf, S., Symmons, O., Uslu, V. V., Dolle, D., Hot, C., Ettwiller, L., & Spitz, F. (2011). Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat Genet, 43*(4), 379-386. doi:10.1038/ng.790

Rupp, B. (2010). Biomolecular crystallography.

Russo Krauss, I., Sica, F., Mattia, C. A., & Merlino, A. (2012). Increasing the X-ray diffraction power of protein crystals by dehydration: the case of bovine serum albumin and a survey of literature data. *Int J Mol Sci, 13*(3), 3782-3800. doi:10.3390/ijms13033782

Sadelain, M., Riviere, I., & Riddell, S. (2017). Therapeutic T cell engineering. *Nature, 545*(7655), 423-431. doi:10.1038/nature22395

Singh, H., Manuri, P. R., Olivares, S., Dara, N., Dawson, M. J., Huls, H., Hackett, P. B., Kohn, D. B., Shpall, E. J., Champlin, R. E., & Cooper, L. J. (2008). Redirecting specificity of T-cell populations for CD19 using the Sleeping Beauty system. *Cancer Res, 68*(8), 2961-2971. doi:10.1158/0008-5472.CAN-07-5600

Singh, H., Moyes, J. S., Huls, M. H., & Cooper, L. J. (2015). Manufacture of T cells using the Sleeping Beauty system to enforce expression of a CD19-specific chimeric antigen receptor. *Cancer Gene Ther, 22*(2), 95-100. doi:10.1038/cgt.2014.69

Sinzelle, L., Izsvak, Z., & Ivics, Z. (2009). Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell Mol Life Sci, 66*(6), 1073-1093. doi:10.1007/s00018-009-8376-3

Skipper, K. A., Andersen, P. R., Sharma, N., & Mikkelsen, J. G. (2013). DNA transposon-based gene vehicles - scenes from an evolutionary drive. *J Biomed Sci, 20*, 92. doi:10.1186/1423-0127-20-92

Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet, 8*(4), 272-285. doi:10.1038/nrg2072

Slusarczyk, H., Felber, S., Kula, M. R., & Pohl, M. (2000). Stabilization of NAD-dependent formate dehydrogenase from Candida boidinii by site-directed mutagenesis of cysteine residues. *Eur J Biochem, 267*(5), 1280-1289.

Sultana, T., Zamborlini, A., Cristofari, G., & Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat Rev Genet, 18*(5), 292-308. doi:10.1038/nrg.2017.7

Svergun, D. I., Petoukhov, M. V., & Koch, M. H. (2001). Determination of domain structure of proteins from X-ray solution scattering. *Biophys J, 80*(6), 2946-2953. doi:10.1016/S0006-3495(01)76260-1

Svergun, D., & Knoch, M.H.J. (2003). Small-angle scattering studies of biological macromolecules in solution. *Rep. Prog. Phys., 66*(1735).

Svergun, D.I. , Barberato, C. , & Koch, M.H.J. . (1995). CRYSOL– a Program to Evaluate X-ray Solution Scattering of Biological
Macromolecules from Atomic Coordinates. *J. Appl. Cryst., 28*, 768-773.

Takeda, H., Rust, A. G., Ward, J. M., Yew, C. C., Jenkins, N. A., & Copeland, N. G. (2016). Sleeping Beauty transposon mutagenesis identifies genes that cooperate with mutant Smad4 in gastric cancer development. *Proc Natl Acad Sci U S A, 113*(14), E2057-2065. doi:10.1073/pnas.1603223113

Talluri, T. R., Kumar, D., Glage, S., Garrels, W., Ivics, Z., Debowski, K., Behr, R., Niemann, H., & Kues, W. A. (2015). Derivation and characterization of bovine induced pluripotent stem cells by transposon-mediated reprogramming. *Cell Reprogram, 17*(2), 131-140. doi:10.1089/cell.2014.0080

Tellier, M., Bouuaert, C. C., & Chalmers, R. (2015). Mariner and the ITm Superfamily of Transposons. *Microbiol Spectr, 3*(2), MDNA3-0033-2014. doi:10.1128/microbiolspec.MDNA3-0033-2014

Thumann, G., Harmening, N., Prat-Souteyrand, C., Marie, C., Pastor, M., Sebe, A., Miskey, C., Hurst, L. D., Diarra, S., Kropp, M., Walter, P., Scherman, D., Ivics, Z., Izsvak, Z., & Johnen, S. (2017). Engineering of PEDF-Expressing Primary Pigment Epithelial Cells by the SB Transposon System Delivered by pFAR4 Plasmids. *Mol Ther Nucleic Acids, 6*, 302-314. doi:10.1016/j.omtn.2017.02.002

Tickle, I.J., Flensburg, C., Keller, P., Paciorek, W., Sharff, A.,, & Vonrhein, C., Bricogne, G. . (2017). STARANISO. .

Trounson, A., & DeWitt, N. D. (2016). Pluripotent stem cells progressing to the clinic. *Nat Rev Mol Cell Biol, 17*(3), 194-200. doi:10.1038/nrm.2016.10

Trubitsyna, M., Michlewski, G., Finnegan, D. J., Elfick, A., Rosser, S. J., Richardson, J. M., & French, C. E. (2017). Use of mariner transposases for one-step delivery and integration of DNA in prokaryotes and eukaryotes by transfection. *Nucleic Acids Res, 45*(10), e89. doi:10.1093/nar/gkx113

Trubitsyna, M., Morris, E. R., Finnegan, D. J., & Richardson, J. M. (2014). Biochemical characterization and comparison of two closely related active mariner transposases. *Biochemistry, 53*(4), 682-689. doi:10.1021/bi401193w

Turchiano, G., Latella, M. C., Gogol-Doring, A., Cattoglio, C., Mavilio, F., Izsvak, Z., Ivics, Z., & Recchia, A. (2014). Genomic analysis of Sleeping Beauty transposon integration in human somatic cells. *PLoS One, 9*(11), e112712. doi:10.1371/journal.pone.0112712

Turtle, C. J., Hanafi, L. A., Berger, C., Gooley, T. A., Cherian, S., Hudecek, M., Sommermeyer, D., Melville, K., Pender, B., Budiarto, T. M., Robinson, E., Steevens, N. N., Chaney, C., Soma, L., Chen, X., Yeung, C., Wood, B., Li, D., Cao, J., Heimfeld, S., Jensen, M. C., Riddell, S. R., & Maloney, D. G. (2016). CD19 CAR-T cells of defined CD4+:CD8+ composition in adult B cell ALL patients. *J Clin Invest, 126*(6), 2123-2138. doi:10.1172/JCI85309

Turunen, S., Huhtakangas, J., Nousiainen, T., Valkealahti, M., Melkko, J., Risteli, J., & Lehenkari, P. (2016). Rheumatoid arthritis antigens homocitrulline and citrulline are generated by local myeloperoxidase and peptidyl arginine deiminases 2, 3 and 4 in rheumatoid nodule and synovial tissue. *Arthritis Res Ther, 18*(1), 239. doi:10.1186/s13075-016-1140-9

van Luenen, H. G., Colloms, S. D., & Plasterk, R. H. (1994). The mechanism of transposition of Tc3 in C. elegans. *Cell, 79*(2), 293-301.

van Zundert, G. C. P., Rodrigues, Jpglm, Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., Melquiond, A. S. J., van Dijk, M., de Vries, S. J., & Bonvin, Amjj. (2016). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol, 428*(4), 720-725. doi:10.1016/j.jmb.2015.09.014

Van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C., & Saccheri, I. J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature, 534*(7605), 102-105. doi:10.1038/nature17951

Varnai, P., Thyagarajan, B., Rohacs, T., & Balla, T. (2006). Rapidly inducible changes in phosphatidylinositol 4,5-bisphosphate levels influence multiple regulatory functions of the lipid in intact living cells. *J Cell Biol, 175*(3), 377-382. doi:10.1083/jcb.200607116

Verdine, G. L., & Norman, D. P. (2003). Covalent trapping of protein-DNA complexes. *Annu Rev Biochem, 72*, 337-366. doi:10.1146/annurev.biochem.72.121801.161447

Vigdal, T. J., Kaufman, C. D., Izsvak, Z., Voytas, D. F., & Ivics, Z. (2002). Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements. *J Mol Biol, 323*(3), 441-452.

Voigt, F., Wiedemann, L., Zuliani, C., Querques, I., Sebe, A., Mates, L., Izsvak, Z., Ivics, Z., & Barabas, O. (2016). Sleeping Beauty transposase structure allows rational design of hyperactive variants for genetic engineering. *Nat Commun, 7*, 11126. doi:10.1038/ncomms11126

Vos, J. C., De Baere, I., & Plasterk, R. H. (1996). Transposase is the only nematode protein required for in vitro transposition of Tc1. *Genes Dev, 10*(6), 755-761.

Vos, J. C., & Plasterk, R. H. (1994). Tc1 transposase of Caenorhabditis elegans is an endonuclease with a bipartite DNA binding domain. *EMBO J, 13*(24), 6125-6132.

Watkins, S., van Pouderoyen, G., & Sixma, T. K. (2004). Structural analysis of the bipartite DNA-binding domain of Tc3 transposase bound to transposon DNA. *Nucleic Acids Res, 32*(14), 4306-4312. doi:10.1093/nar/gkh770

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet, 8*(12), 973-982. doi:10.1038/nrg2165

Wiehe, J. M., Ponsaerts, P., Rojewski, M. T., Homann, J. M., Greiner, J., Kronawitter, D., Schrezenmeier, H., Hombach, V., Wiesneth, M., Zimmermann, O., & Torzewski, J. (2007). mRNA-mediated gene delivery into human progenitor cells promotes highly efficient protein expression. *J Cell Mol Med, 11*(3), 521-530. doi:10.1111/j.1582-4934.2007.00038.x

Wilber, A., Frandsen, J. L., Geurts, J. L., Largaespada, D. A., Hackett, P. B., & McIvor, R. S. (2006). RNA as a source of transposase for Sleeping Beauty-mediated gene insertion and expression in somatic cells and tissues. *Mol Ther, 13*(3), 625-630. doi:10.1016/j.ymthe.2005.10.014

Wilber, A., Wangensteen, K. J., Chen, Y., Zhuo, L., Frandsen, J. L., Bell, J. B., Chen, Z. J., Ekker, S. C., McIvor, R. S., & Wang, X. (2007). Messenger RNA as a source of transposase for sleeping beauty transposon-mediated correction of hereditary tyrosinemia type I. *Mol Ther, 15*(7), 1280-1287. doi:10.1038/sj.mt.6300160

Wong, S.S. . (1993). CRC Chemistry of Protein Conjugation and Crosslinking. *CRC Press, Boca Raton, Florida*.

Woodard, L. E., & Wilson, M. H. (2015). piggyBac-ing models and new therapeutic strategies. *Trends Biotechnol, 33*(9), 525-533. doi:10.1016/j.tibtech.2015.06.009

Wu, Y., Zhang, T. Y., Fang, L. L., Chen, Z. X., Song, L. W., Cao, J. L., Yang, L., Yuan, Q., & Xia, N. S. (2016). Sleeping Beauty transposon-based system for rapid generation of HBV-replicating stable cell lines. *J Virol Methods, 234*, 96-100. doi:10.1016/j.jviromet.2016.04.010

Xie, X., Pashkov, I., Gao, X., Guerrero, J. L., Yeates, T. O., & Tang, Y. (2009). Rational improvement of simvastatin synthase solubility in Escherichia coli leads to higher whole-cell biocatalytic activity. *Biotechnol Bioeng, 102*(1), 20-28. doi:10.1002/bit.22028

Yang, G., Weil, C. F., & Wessler, S. R. (2006). A rice Tc1/mariner-like element transposes in yeast. *Plant Cell, 18*(10), 2469-2478. doi:10.1105/tpc.106.045906

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat Methods, 12*(1), 7-8. doi:10.1038/nmeth.3213

Yang, W., Lee, J. Y., & Nowotny, M. (2006). Making and breaking nucleic acids: two-Mg2+-ion catalysis and substrate specificity. *Mol Cell, 22*(1), 5-13. doi:10.1016/j.molcel.2006.03.013

Yant, S. R., & Kay, M. A. (2003). Nonhomologous-end-joining factors regulate DNA repair fidelity during Sleeping Beauty element transposition in mammalian cells. *Mol Cell Biol, 23*(23), 8505-8518.

Yant, S. R., Meuse, L., Chiu, W., Ivics, Z., Izsvak, Z., & Kay, M. A. (2000). Somatic integration and long-term transgene expression in normal and haemophilic mice using a DNA transposon system. *Nat Genet, 25*(1), 35-41. doi:10.1038/75568

Yant, S. R., Park, J., Huang, Y., Mikkelsen, J. G., & Kay, M. A. (2004). Mutational analysis of the N-terminal DNA-binding domain of sleeping beauty transposase: critical residues for DNA binding and hyperactivity in mammalian cells. *Mol Cell Biol, 24*(20), 9239-9247. doi:10.1128/MCB.24.20.9239-9247.2004

Zayed, H., Izsvak, Z., Khare, D., Heinemann, U., & Ivics, Z. (2003). The DNA-bending protein HMGB1 is a cellular cofactor of Sleeping Beauty transposition. *Nucleic Acids Res, 31*(9), 2313-2322.

Zayed, H., Izsvak, Z., Walisko, O., & Ivics, Z. (2004). Development of hyperactive sleeping beauty transposon vectors by mutational analysis. *Mol Ther, 9*(2), 292-304. doi:10.1016/j.ymthe.2003.11.024

Zeineddine, D., Hammoud, A. A., Mortada, M., & Boeuf, H. (2014). The Oct4 protein: more than a magic stemness marker. *Am J Stem Cells, 3*(2), 74-82.

Zhou, L., Mitra, R., Atkinson, P. W., Hickman, A. B., Dyda, F., & Craig, N. L. (2004). Transposition of hAT elements links transposable elements and V(D)J recombination. *Nature, 432*(7020), 995-1001. doi:10.1038/nature03157

ACKNOWLEDGEMENTS

I wish to thank Dr. Orsolya Barabas for accepting me first as a Master's student in her lab (when I desperately wanted to work on transposons, without knowing exactly why) and then as a PhD student (when I knew why, but not exactly how). Thank you Orsolya for guiding me through the wonderful world of transposition, but also for teaching me how to drive in the much less fabulous (for me) land of scientific writing (my license is still far away). And finally, thanks for encouraging me to follow alternative routes in my project that, as an incorrigible biotechnologist, I really wanted to explore.

I would like to thank the members of my Thesis Advisory Committee, Dr. Kiran Raosaheb Patil, Prof. Dr. Irmgard Sinning and Dr. Marco Marcia, for all the support, advice, and guidance given over the entire course of my PhD. Many thanks also to Dr. Christoph Müller and Prof. Dr. Rob Russell for joining my Thesis Defense Committee. My sincere thanks also go to our collaborators from the Ivics Laboratory, Dr. Zoltán Ivics, Dr. Csaba Miskey and Dr. Esther Grueso, and from the Hudecek Laboratory, Dr. Michael Hudecek and Dr. Andreas Mades, for their contributions to the SBprotAct project.

I am very grateful to the EMBL International PhD Programme for funding my studies and to the EMBLEM staff for translating the SBprotAct invention into a patent application. I want to acknowledge also the EMBL Core Facilities for providing top quality services and invaluable advices during the experiments; in particular, Dr. Vladimir Benes and Dr. Tobias Rausch from the Genomics Core Facility and Dr. Malte Paulsen, Dr. Diana Ordonez and Dr. Neda Tafrishi from the Flow Cytometry Core Facility. Many thanks go to Dr. Brice Murciano for setting up crystallization trials at EMBL Heidelberg and to Vincent Mariaule and Dr. José Márquez for the crystallization and automatic crystal harvesting experiments performed at EMBL Grenoble.
I am particularly grateful to Dr. Andrew McCarthy for all his help with crystallographic data analysis and structure determination. Many thanks also to the ESRF and PetraIII staff for providing assistance during data collection at the synchrotron beamlines. I want to also thank Dr. Anna Berteotti, Andrea Graziadei and Leo Nesme for their help with SAXS data collection and analysis; Dr. Ezgi Karaca for teaching me the basics of structural modelling; Dr. Vladimir Rybin for helping me with the CD spectroscopy measurements; Dr. Alexander Aulehla, Sebastian Henkel, Dr. Dmytro Yushchenko, Dr. Amparo Andres-Pons and Gemma Estrada Girona for sharing materials, protocols and expertise for the cell-based experiments.

Special thanks go to all the former and current members of the Barabas group, who made my PhD the more unpredictable, irreproducible and unconventional experiment I have ever carried out. Thanks to Franka for teaching me everything in the lab with patience and

enthusiasm when I could barely say a full sentence in English. Thank you for being the first one who welcomed me as a lab mate and as a friend at EMBL. Thanks to Cecilia for her fundamental contributions to all my PhD work and for being such an irreplaceable source of advices on writing, fashion and life styles. Thanks to Ola for making our "flatmateship" so awesome and for teaching me to constantly fight for what I want (I admire you a lot). Thanks to Anna for sharing with me her explosive way of loving science and life and for always caring about her Raffaello. Thanks to Natalia for showing me that perseverance and calmness are signs of an incredibly strong personality. Thanks to Gera for being such a tolerant office mate and for all our philosophical and politically incorrect discussions about our place in science and society. Thanks to Lotte for sharing with me her considerable experience in the worlds of academia and industry: this helped me think out of the box! Thanks to Carlos for reminding me that a good colleague is the one who makes your day with a kind word…or a flower. Thanks to Buse for constantly igniting my curiosity for science with all her insightful questions and answers. Thanks to Aukie for being such a smart student and a cool "buddy" to work with. Thank you guys for putting up with my crazy ideas (ginger wig included) and sharing with me a fully "-omic" experience at EMBL (tragicomic, interactomic and sometimes gastronomic).

I wish to also thank Sachi for giving me the bible of crystallography as a gift; Alex for all his support during my M.Sc. thesis writing and my moving from flat to flat in Heidelberg; Enrico for being such an extraordinarily positive person inside and outside the lab; Banu for being such a talented and passionate young researcher; again Andrea and Leo for occasionally leading me to the dark side of the force (NMR and SAXS). Thanks to Jose, Raffaele and Annamaria for smiling to me every time we meet in the corridors. Special thanks go to Loredana and Giorgia for making me discover the best indoor cycling classes of all time and the related anti-stress benefits.

Ed, infine, lo sapevate che non avrei risparmiato nessuno. Grazie a mia madre per avermi sostenuto in qualsiasi mia scelta, nonostante ciò mi abbia spesso portato lontano (ma mai dal suo amore). Grazie a mio padre per essere e mostrarsi a noi come un uomo pieno di tantissimi sentimenti. Grazie a mia sorella Ilaria per essere l'essere umano più complicato e perfetto che conosca. Ora ringraziate me per non aver usato i vostri simpatici soprannomi. Grazie alla nostra Annamaria per occuparsi sempre delle sue "creature". E, per completare, grazie a Matteo per essere la scoperta più incredibile della mia vita. Grazie a tutti voi, famiglia, la mia ricerca della felicità si è conclusa con successo. Per quella scientifica, come sapete, ho ancora tanta strada da fare.