Manipulation of the Tn5 Transpososome Assembly Pathway and its Effect on Reaction Dynamics

George Blundell-Hunter

School of Life Sciences

University of Nottingham



Thesis submitted for the Degree of Doctor of Philosophy

Nottingham, 2017

Abstract

Transposable elements are an almost ubiquitous feature in all three domains of life, and are a significant driver of genomic evolution. These discrete genetic elements are capable of mobilisation and amplification, which allows them to multiply. A subset of these elements, the cut-and-paste transposons, excise themselves entirely from the surrounding DNA and integrate themselves elsewhere in the genome.

Hsmar1, a cut-and-paste element of the *mariner* transposon family, was recently investigated to examine the kinetics of the complex formation pathway. This transposase forms a dimer in solution before binding to first one transposon end, then recruiting the other. This pathway, Synapsis by Naked End Capture (S-NEC), is found in both prokaryotes and eukaryotes. Tn5, of the IS4 family, is an example of the alternative pathway, Synapsis by Protein Dimerisation (S-PD). In this pathway, the transposase is monomeric in solution. Each monomer binds to a transposon end independently, before combining to form a synaptic complex. This pathway has been seen widely in prokaryotes, but no eukaryotic elements tested to date employ it.

To understand what limitations the complex formation pathway might place upon an S-PD transposase in a eukaryotic genome, the Tn5 transposase was manipulated. Forming the transposase into a single-chain dimer allowed it to mimic S-NEC, an alteration that made it hyperactive in prokaryotes and eukaryotes.

As the eukaryotic genome contains chromatin, adapting the Tn5 transposase to this might overcome some of the low activity seen. The fusion of a chromatin binding domain to the transposase was not able to enhance the frequency of transposition, which may suggest that the presence of chromatin is not causing significant inhibition.

The results presented in this work suggest two significant conclusions. The first is that the S-NEC reaction dynamics are significantly preferable to S-PD within eukaryotes. The second is that S-PD itself may be a system of regulation of transposition, to prevent activity levels that could be detrimental to the host.

Acknowledgements

“You just keep on trying till you run out of cake.” GLaDOS

The first person I must thank here is Ronald Chalmers, for the enthusiasm when experiments work and the constant doubting of those same experiments. I would not be the scientist I am now without your help.

All the other members of the Chalmers lab have been a great assistance over the years as well, and deserve thanks. Michael for teaching me so much when I first started and putting up with all my questions, as well as Shivam and Jack for all the good conversations. The members of the Bolt lab have also been fantastic company and help throughout the years.

I would also like to thank my family for all the support for the last four years. In particular, I would like to thank my dad for reading through uncountable copies of this thesis, even though he may not have understood what any of it meant.

I would also like to thank my partner, Juliet, for coping with me throughout all of this PhD, including all the weekends and late nights spent in the lab.

The cohort of the DTP has also been a great support and company throughout all of this.

I also gratefully acknowledge the support of the BBSRC DTP and the University of Nottingham.

Table of Contents

Abstract **2**

Acknowledgements **3**

Table of Contents **4**

Table of Figures **9**

List of Abbreviations **11**

Chapter 1 Introduction **12**

1.1 Background **13**

1.1.1 Retrotransposons 14

1.1.2 DNA transposons 17

1.2 Mechanisms of the RNase-H like fold **18**

1.2.1 Replicative transposition mechanisms 20

1.2.2 Non-replicative transposition mechanisms 21

1.3 Structure of the RNase-H like fold **23**

1.4 How transposases capture the transposon end **25**

1.4.1 Binding sites in the transposons 25

1.4.2 The Helix-Turn-Helix motif 27

1.4.3 Zn-based DNA-binding domains 31

1.4.4 The intertwined binding domain 34

1.5 Secondary DNA-Binding Activity **38**

1.5.1 DNA binding by the catalytic core 38

1.5.2 Target DNA binding 43

1.5.3 DNA binding of host factor and accessory proteins 48

1.6 HUH nuclease fold 51

1.6.1 Mechanism of the IS200/IS605 family 53

1.6.2 Structure of the IS200/IS605 family transposases 53

1.6.3 The DNA-binding activity of IS200/IS605 family transposases 54

1.7 Mechanisms of synapsis 56

1.7.1 Hypothetical model of complex formation pathway 58

1.8 Research aims 61

Chapter 2 Materials and Methods 65

2.1 Materials 66

2.1.1 Chemicals and reagents 66

2.1.2 Strains and cell lines 66

2.1.2.1 *Escherichia coli* strains 66

2.1.2.2 Human cell lines: 67

2.1.3 Plasmids and oligonucleotides 67

2.1.3.1 Plasmid list 67

2.1.3.2 Oligonucleotide list 79

2.4 Methods 82

2.4.1 Cloning and DNA manipulation 82

2.4.1.1 Polymerase Chain Reaction 82

2.4.1.2 Gel electrophoresis 82

2.4.1.3 Plasmid construction 83

2.4.1.4 Chemically competent cells 83

2.4.1.5 Radioactive substrate purification 84

2.4.1.6 Fluorescent oligonucleotide substrate purification 84

2.4.2 Bacterial assays 85

2.4.2.1 Papillation assay 85

2.4.2.2 *E. coli* pre-formed complex integration assay....86
2.4.3 Human cell culture and assays.... 86
2.4.3.1 Cell maintenance....86
2.4.3.2 Cell transfection....87
2.4.3.3 Cell lysis and protein extraction....87
2.4.3.4 Fluorescence microscopy....88
2.4.3.5 Transposition integration assay....88
2.4.3.6 Human cell pre-formed complex integration assay....89
2.4.3.7 Genomic DNA extraction from human cells....89
2.4.3.8 Splinkerette PCR....90
2.4.3.9 Inverse PCR mapping....91
2.4.4 Protein purification and *in vitro* assays.... 92
2.4.4.1 Protein purification....92
2.4.4.2 Western blotting....93
2.4.4.3 *In vitro* Transposition assay....94
2.4.4.4 Electrophoretic-mobility shift assay (EMSA)....95
Chapter 3 The Tn5 Transposase Dimer....96
3.1 Introduction.... 97
3.2 Results and Discussion.... 99
3.2.1 Design and analysis of the dimeric transposase.... 99
3.2.1.1 Testing of the dimeric transposase in bacteria....99
3.2.1.2 Purification and *in vitro* testing of the transposase dimers....106
3.2.2 The transposase dimers in eukaryotic cells.... 113
3.2.3 Characterisation of the dimer complex assembly pathway.... 121
3.2.3.1 The 'long-short' experiment....123
3.2.3.2 Active site mutagenesis of the dimer....127
3.2.3.3 Dimerisation site mutagenesis of the dimer....131
3.2.3.4 Mutant cleavage and binding....134

3.2.3.5 Mechanisms conclusion 137

3.2.4 Pre-formed transposase complexes 138

3.3.4 Conclusions and future work 142

Chapter 4 Enhancing the Stability of the Single End Complex 144

4.1 Introduction 145

4.2 Results 148

4.2.1 Designing and testing the ZFA fusions 148

4.2.2 Optimisation of the ZFA binding sites 152

4.2.3 Analysing binding of the fusion proteins 158

4.3 Discussion 162

4.3.1 The activity of the ZFA fusion proteins 162

4.3.2 The location of the ZBS and its importance to the activity 164

4.3.3 SEC binding stability 167

4.3.3 Conclusions 168

Chapter 5 Adapting the Tn5 Transposase to Bind Chromatin 169

5.1 Introduction 170

5.2 Results 172

5.2.1 Functional analysis of the PHD fusions 172

5.2.2 Testing of the PHD fusion proteins in human cells 178

5.3 Discussion 182

Chapter 6 The Transposition Activity of PGBD5 186

6.1 Introduction 187

6.2 Results 189

6.2.1 Functionality in HEK293 cells 189

6.2.2 Functionality in HeLa cells 191

6.3 Discussion 195

Chapter 7 Conclusions and Perspectives 199

7.1 Summary of the research 200

7.1.1 Modification of the complex formation mechanism 200

7.1.2 Other inhibitors within a eukaryotic genome 204

7.2 Outlook 205

Chapter 8 Appendices 209

8.1 Supplementary information 210

8.1.1 Supplementary figures 210

8.1.2 Protein sequences 213

8.1.3 DNA substrate sequences 227

8.2 Professional internships for PhD student reflection form 228

8.3 Bibliography 230

Table of Figures

Figure 1: Mechanisms Of Retrotransposition.....15

Figure 2: Mechanisms Of DNA Transposition. 19

Figure 3: RNase-H Catalytic Core In Transposases......23

Figure 4: The HTH DNA-Binding Domains. 27

Figure 5: Zn-Based DNA-Binding Domains. 32

Figure 6: Intertwined DNA-Binding Domains......35

Figure 7: Catalytic Core Interactions With DNA. 38

Figure 8: Target DNA Binding Of Transposase Complexes......44

Figure 9: IHF DNA-Binding Complex.. 48

Figure 10: The IS200/IS605 Transposition System. 52

Figure 11: The Complex Formation Pathways Of Transposition......56

Figure 12: Model Of Eukaryotic Transposition Through The Different Complex Assembly Pathways. 59

Figure 13: Schematics And Papillation Assays Of The Transposase Dimers......102

Figure 14: Purification And Titrations Of The Dimeric Transposases. 106

Figure 15: In Vitro Transposition Assay Time Courses Of The Monomeric And Dimeric Transposases. 110

Figure 16: The Tn5 Dimeric Transposases In Eukaryotes. 114

Figure 17: Mapping Of Transposition Integration Into The Human Genome......117

Figure 18: Model Of The Two Complex Formation Mechanisms Of The Dimeric Transposase......122

Figure 19: The Long-Short Experiment......124

Figure 20: Active Site Mutant Transposase Activity......127

Figure 21: Dimerisation Mutant Transposase Activity. 131

Figure 22: Binding And Cleavage Assay. 135

Figure 23: The Activity Of Monomeric And Dimeric Pre-Formed Complexes.. 138

Figure 24: Schematics And Papillation Assay Of The ZFA-Fusion Constructs. 149

Figure 25: Purification And Titrations Of The ZFA-Fusion Proteins. 151

Figure 26: Identification Of Optimal ZBS Location For The ZFA-Fusion Proteins...153

Figure 27: In Vitro Transposition Assay Time Courses Of The ZFA-Fusion Proteins.

.....155

Figure 28: Binding Assays Of The ZFA-Fusion Proteins......159

Figure 29: Model Demonstrating The Location Of The Zinc Finger Respective To The Tn5 Transposase. 164

Figure 30: Schematics Of The Fusion Proteins Of Tn5 Transposase And PHD Domains......173

Figure 31: Initial Analysis Of The PHD Fusion Proteins......174

Figure 32: In Vitro Transposition Assay Time Course Of The Short PHD-Fusion Protein. 176

Figure 33: Transposition Activity Of The PHD Fusion Proteins In Hela Cells......180

Figure 34: PGBD5 Experiments On Hek293 Cells. 189

Figure 35: PGBD5 Experiments On Hela Cells. 193

Supplementary Figure 1: Mutagenesis Screening for Integration Deficient Mutants 210

Supplementary Figure 2: Binding Assay of Dimeric Transposases and Mutants.....211

Supplementary Figure 3: Papillation Assay Comparing the Transposition Activity Between Two Different Vectors 212

List of Abbreviations

aa	amino acid(s)	NLS	Nuclear-Localisation Signal
Ac	Activator	NTD	N-terminal Domain
bp	Base-pair(s)	NTS	Non-Transferred Strand
CTD	C-terminal Domain	OD	Optical Density
DMEM	Dulbeccos Modified Eagles Media	OPI	Over-Production Inhibition
DMSO	Dimethyl Sulfoxide	ORF	Open-Reading Frame
Ds	Dissociator	PAGE	Polyacrylamide Gel Electrophoresis
dsDNA	Double-stranded DNA	PEC	Paired end Complex
DTT	Dithiothreitol	PFV	Prototype Foamy Virus
EDTA	ethilenediaminetetraacetic acid	PGBD	*piggyBac* Derived
EMSA	Electromobility shift assay	PHD	Plant-Homeodomain
FBS	Foetal Bovine Serum	RSS	Recognition Signal Sequence
GFP	Green-Fluorescent Protein	SDC	Single-Dimer Complex
HTH	Helix-Turn-Helix	SDS	Sodium dodecyl sulfate
IBD	Intertwined Binding Domain	SEC	Single end complex
IHF	Integration Host Factor	ssDNA	Single-stranded DNA
IPTG	Isopropyl β-D-1-thiogalactopyranoside	STBS	Sub-Terminal Binding Site
IS	Insertion Sequences	S-NEC	Synapsis by Protein Dimerisation
ITm	IS630-Tc1-*mariner*	S-PD	Synapsis by Naked End Capture
L1	Long-Interspersed Nuclear Element 1	TALE	Transcription Activator-Like Effector
LB	Lysogeny Broth	TBE	Tris-Borate-EDTA
LTR	Long-Terminal-Repeat	TBS	Terminal Binding Site
MDC	Multi-Dimer Complex	TIR	Terminal Inverted Repeat
Mya	Million years ago	TPM	Transcripts Per Million
NBD	Nonamer-Binding Domain	TS	Transferred Strand
NEB	New England Biolabs	ZBS	Zinc Finger-Binding Site
NED	N-terminal Extension Domain	ZFA	Zinc Finger Array

Chapter 1

Introduction

1.1 Background

Transposons were discovered by Barbara McClintock in the 1940s (McClintock 1950). This investigation into the chromosomal rearrangements in maize identified two distinct elements: Dissociator (Ds), and Activator (Ac); both of which could change location within the maize genome, although Ds could only do this when Ac was also present. The Ds and Ac elements were transposons. Ac encoded the protein responsible for the transposition, making it an autonomous element. Ds was a non-autonomous element, requiring the Ac encoded protein to transpose. Since this discovery, transposons have been found to permeate all branches of the tree of life.

Transposons are thought of as selfish elements, their primary function is to replicate and survive which can be deleterious to the host (Hua-Van et al. 2011). The abundance of active transposons could therefore be considered counter-intuitive, unless they potentially provide some advantage to the host. There are many cases where transposons have been utilised by the host for various functions, such as chromosome 'caps' in *Drosophila melanogaster* (Pardue et al. 1997), carrying antibiotic resistance genes (Baquero et al. 2013) or effects in gene regulation (Capy et al. 2000). Under particular circumstances certain transposition systems are activated by environmental stress, and could also be linked to host defence responses (Grandbastien 1998). Furthermore, some transposons are used to activate otherwise silent genes in *Escherichia coli* (Hall 2000). These studies suggest some level of host-transposon interaction, aiding in the survival of both.

Transposons can be divided into two groups. The retrotransposons are those that mobilise with an RNA intermediate step, and generally are the more complex systems, dominant in eukaryotes. DNA transposons, on the other hand, mobilise without the need for RNA intermediates and are the dominant form found in most prokaryotes.

1.1.1 Retrotransposons

Retrotransposons are divided into two subgroups, based on the presence or absence of Long-Terminal-Repeat (LTR) sequences at the transposon ends.

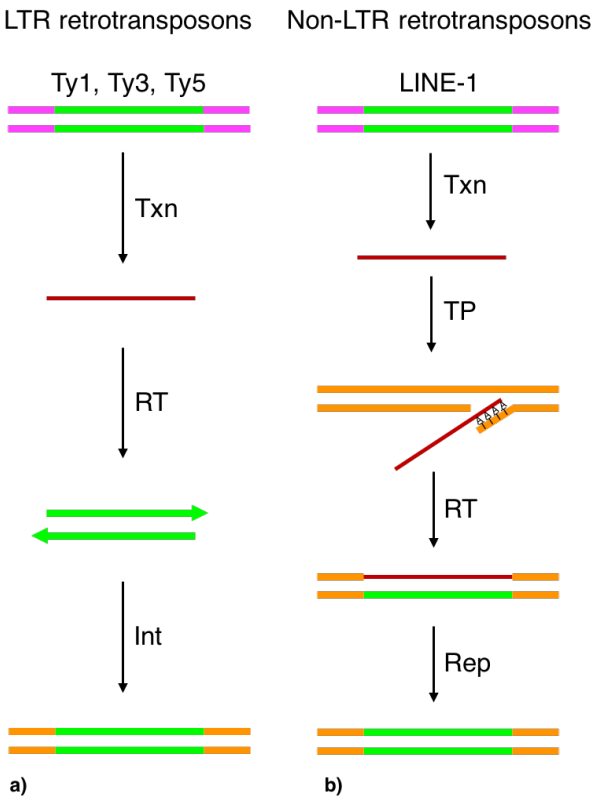


Figure 1: Mechanisms of Retrotransposition. Donor DNA shown in magenta, transposon DNA in green, transposon RNA in red and target DNA in orange. Txn=Transcription, RT=Reverse Transcription, TP=Target Priming, RT=Reverse Transcription, Int=Integration and Rep=Repair. **a)** LTR retrotransposon mechanism. Transposon is transcribed into ssRNA, then reverse-transcribed into dsDNA. This is then integrated into the target site. **b)** Non-LTR retrotransposons mechanism. The transposon is transcribed into ssRNA. This is then integrated and partially reverse transcribed, in target-primed reverse transcription. The RNA is lost and the ssDNA is repaired.

The LTR-retrotransposons are descended from retroviruses (Hughes 2015). They encode the same basic proteins as retroviruses: a structural protein (Gag), an RNA-dependent DNA-polymerase (Pol) and an Integrase (IN). The only common retroviral protein that the LTR-retrotransposons do not encode is Env, the protein for making the viral capsid. Retrotransposition begins when host proteins synthesise an RNA transcript of the entire transposon, initiated from a primer binding site in the 5' LTR. Transcripts are

enclosed by a virus-like particle, constructed from Gag proteins. Pol reverse transcribes the RNA into a single stranded DNA (ssDNA), which is then duplicated into a double-stranded DNA (dsDNA) copy of the full transposon. Subsequently, the DNA leaves the virus-like particle, IN binds this dsDNA transposon and integrates it into the genome (Havecker et al. 2004) (Figure 1a). The best-characterised LTR-retrotransposons are the Ty elements from yeast (Carr et al. 2012; Curcio et al. 2015).

The non-LTR retrotransposons have short direct repeats at either end of the element. The best characterised are the mammalian LINE-1 (L1) elements (S. R. Richardson et al. 2015). These contain two Open Reading Frames (ORFs): ORF1 encodes an RNA-binding protein, while the ORF2 protein functions as both an endonuclease and a reverse transcriptase. Initiation of L1 retrotransposition occurs in the same manner as LTR-retrotransposons; host proteins produce an RNA transcript of the element initiated from a 5' transcription start site. This is then bound by the ORF1 protein in a ribonucleoprotein complex. The ORF2 protein makes a nick in the target site DNA, and the quadruple-thymine sequence adjacent to this acts as a binding site for the retrotransposon RNA strand. Reverse transcription by the ORF2 protein then generates a DNA complement of the RNA, and this dsDNA/RNA integrates into the target site. Host factors then remove the RNA strand and duplicate the DNA, forming a new copy of L1 (Figure 1b).

1.1.2 DNA transposons

DNA transposons are not subdivided as clearly as the retrotransposons, as they can be broken up into a large number of different superfamilies dependent on mechanism of transposition and structure of the element. The nomenclature of DNA transposons is varied dependent on whether they are prokaryotic or eukaryotic in origin. In prokaryotes, DNA transposons are known as Insertion Sequences (IS) when they encode only transposition related proteins. These ISs have a common nomenclature that starts with 'IS' followed by an alphanumeric code (for example, IS10 or ISMac16). The Eukaryotic DNA transposons do not have a common nomenclature and are simply divided into 23 superfamilies.

The DNA transposons share a small set of common features. Each end of the transposon has a Terminal Inverted Repeat (TIR) that varies from relatively short, ~20 base pairs (bp), to hundreds of bp in length. Each transposon contains at least one binding site at the terminus known as a Terminal Binding Sequence (TBS), as well as potentially more binding sites referred to as Sub-Terminal Binding Sequences (STBS). The autonomous transposons also encode at least one protein that is predominantly, if not solely, responsible for the transposition reaction, called the transposase. There are other features that can occur in the transposon such as STBSs outside of the TIR region, as well as additional genes for antibiotic resistance or transposition regulation.

All transposase proteins encoded by the wide array of elements utilise one of four catalytic core domains to coordinate the steps of the transposition reaction. These cores are: the RNase-H like fold; the HUH-nuclease fold; the serine core; and the tyrosine core (Nowotny 2009; Thomas & Pritham 2015; He et al. 2013; Boocock & Rice 2013; Wang et al. 2006; Roberts & Mullany 2009; Waters & Salyers 2013). The only cores that will be discussed here are the RNase-H like fold and the HUH-nuclease fold as the serine and tyrosine cores require a greater degree of study.

1.2 Mechanisms of the RNase-H like fold

The RNase-H like fold is the best-studied catalytic core found in DNA transposases, and is the most commonly observed (Hickman & Dyda 2016; Siguier et al. 2006). This versatile catalytic core can coordinate a variety of different transposition mechanisms, though all initiate with the same reaction step. The common initial reaction step is a hydrolysis reaction, using a water molecule as the nucleophile on one strand at the transposon end, either the Transferred Strand (TS) or Non-Transferred Strand (NTS). The TS is defined as the strand that is first integrated into the target DNA strand. It is also the strand that, after cleavage, has a free 3'OH on the transposon side of the DNA break.

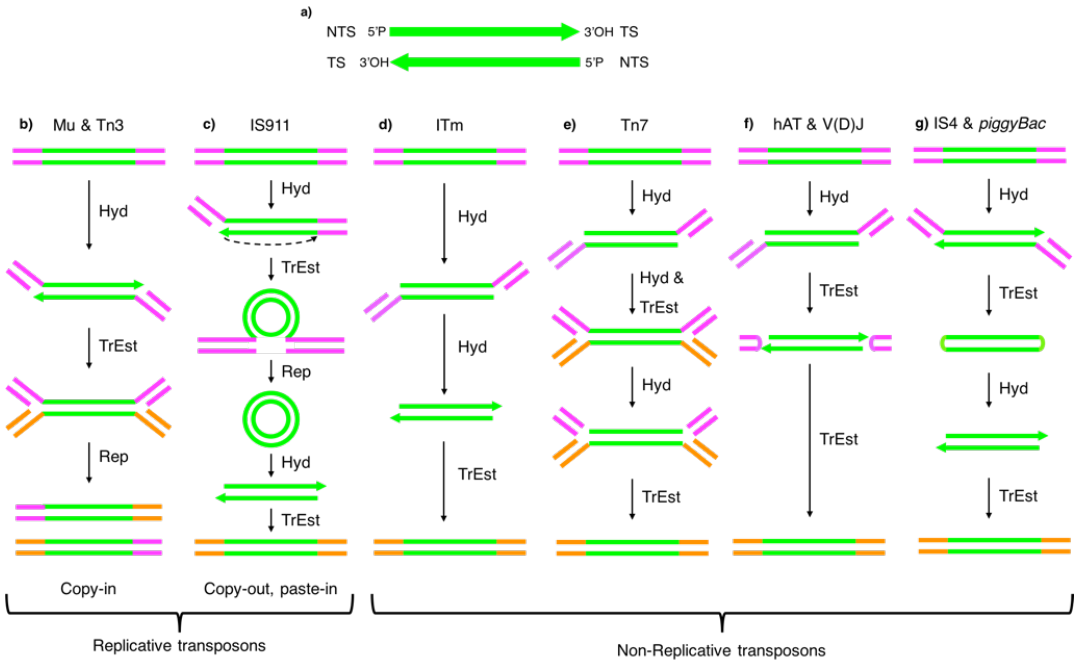


Figure 2: Mechanisms of DNA Transposition. Donor DNA shown in magenta, transposon DNA in green and target DNA in orange. The arrow indicates the 3' end of the DNA strand. Rep=Replication, Hel=Helicase. **a)** Depiction of transposon DNA to show the location of the Transferred Strand and Non-Transferred Strand. **b)** Copy-in transposition, TS is cleaved, then integrated into the target DNA. Replication produces cointegrate which then leaves two transposons. **c)** Copy-out-paste-in transposition, one TS end is cleaved, attacks opposite end of same strand, producing ssDNA circle. Replication produces dsDNA circle which is then cleaved and integrates into target site. **d)** Cut-and-paste transposition of ITm superfamily. NTS is cleaved at both ends, then TS is cleaved, after which the transposon can integrate elsewhere. **e)** Cut-and-paste transposition of Tn7. The NTS is cleaved first. The TS is then cleaved and directly integrated into the target site. The NTS is then integrated into the target site subsequently. **f)** Cut-and-paste transposition of hAT and V(D)J recombination. NTS is cleaved first, the donor 3'OH attacks the other strand forming a hairpin on the donor DNA. The excised transposon can then integrate into the target site. **g)** Cut-and-paste transposition of IS4 and *piggyBac.* The TS is cleaved first, the 3'OH attacks the opposite strand to create a hairpin on the transposon end. Hydrolysis breaks the hairpin and the transposon can then integrate into the target DNA.

1.2.1 Replicative transposition mechanisms

'Copy-in' transposition is the earliest-discovered replicative transposition mechanism (Figure 2a), seen in the bacteriophage Mu and Tn3 families (Shapiro 1979; Harshey 2014; Nicolas et al. 2015). Initially the TS is cleaved, then the 3'OH of the transposon becomes the nucleophile for a transesterification reaction attacking the target DNA. At this stage, each transposon end is known as a Shapiro intermediate, where the NTS is attached to the donor DNA and the TS to the target DNA (Shapiro 1979). DNA replication machinery is actively recruited to the Shapiro intermediate junctions (Kruklitis et al. 1996). The DNA replication produces a cointegrate where the two transposons have the donor DNA on one end and the target DNA on the other.

The other RNase-H based replicative transposition mechanism is 'copy-out-paste-in' (Figure 2b), exemplified by the IS3 family member IS911 (Duval Valentin et al. 2004; Pouget et al. 2006; Chandler et al. 2015). The initial step in the reaction is hydrolysis of the TS at only one end of the transposon. The free 3'-OH then makes a nucleophilic attack on the same strand, at the opposite end of the transposon, which produces a cyclic ssDNA transposon. DNA replication forms a dsDNA circular product, which is then cleaved on both strands via hydrolysis reactions, to yield a linear transposon which can be integrated via a transesterification step into the target site.

1.2.2 Non-replicative transposition mechanisms

The non-replicative transposition mechanisms are commonly known as 'cut-and-paste' as they cleave the entire transposon from the donor site (Tellier et al. 2015). There are four different mechanisms to do this, distinguished by the use of hairpin intermediates in the reaction.

The simplest cut-and-paste reaction is used by the ITm (IS630-Tc1-*mariner*) superfamily of transposons (Figure 2c). The initial hydrolysis reaction occurs on the NTS of each transposon end. The second strand is then directly cleaved using a second hydrolysis reaction, producing a transposon which can then be integrated into the target site through transesterification reactions (J. M. Richardson et al. 2009; Tellier et al. 2015). The other mechanism that avoids the need for a hairpin intermediate is used by the Tn7 transposon (Figure 2d). In this mechanism, the initial hydrolysis reaction cleaves the NTS of each end. The TS is then cleaved in a second hydrolysis reaction, and then directly integrated into the target site. In the final step the free NTS is integrated into the target site (Peters 2014).

The remaining mechanisms for non-replicative transposition make use of a hairpin intermediate in order to cleave the second strand. The simpler of the two mechanisms is the reverse-hairpin mechanism (Figure 2e). Initially the NTS is cleaved with a hydrolysis reaction. The free 3'OH on the donor DNA then acts as the nucleophile in a transesterification reaction attacking the TS. This fully excises the transposon and leaves the donor DNA with a hairpin, which is resolved by host mechanisms. Both the *hAT*

superfamily and RAG1 use this approach (Rubin et al. 2001; Hickman et al. 2014; Atkinson 2015; Fugmann 2001). The forward-hairpin reaction initially cleaves the TS (Figure 2f). A transesterification reaction occurs between the free 3'OH on the transposon and the NTS, cleaving the transposon from the donor site and producing a hairpin between the two strands. As there is no free 3'OH, another hydrolysis reaction is used to break the hairpin apart. This mechanism is used by the IS4 family and also the *piggyBac* superfamily (Reznikoff 2008; Yusa 2015).

1.3 Structure of the RNase-H like fold

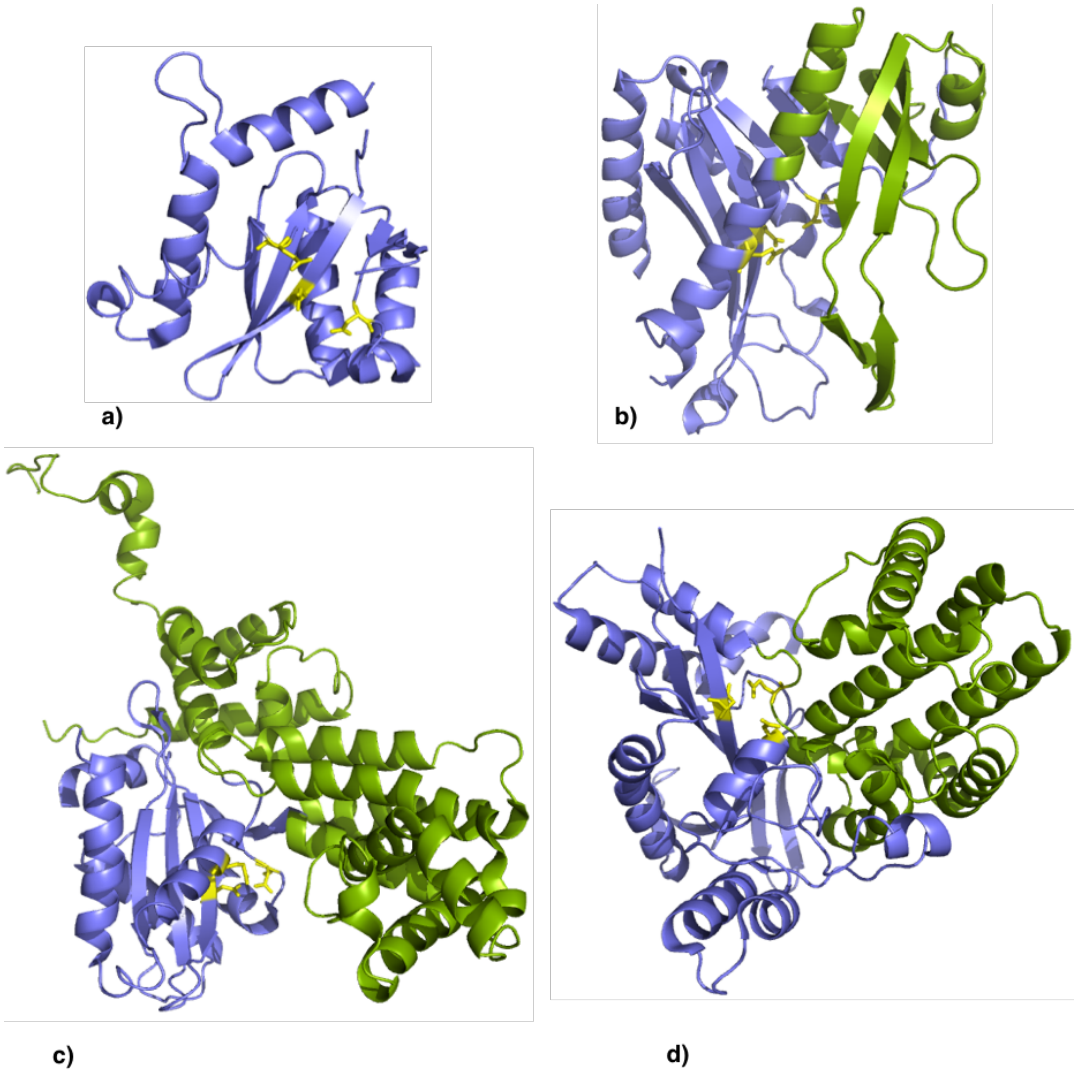


Figure 3: RNase-H Catalytic Core in Transposases. The catalytic core is shown in blue, with the insertion domains shown in green and the three catalytic residues in yellow. **a)** The catalytic core of HIV integrase (PDB: 1EX4). **b)** The catalytic core and insertion domain of the Tn5 transposase (PDB: 1MUH). **c)** The catalytic core and insertion domain of Hermes (PDB: 4D1Q). **d)** The catalytic core and insertion domain of RAG1 (PDB: 3JBX).

The structure of the RNase-H like fold was first defined for the *E. coli* RNase-H1 (Figure 3a) (Katayanagi et al. 1990; Yang et al. 1990). At the centre of the fold is a five-stranded β-sheet, where the first three strands run antiparallel to one another, while the shorter fourth and fifth strands run parallel to the first. This central sheet is the most invariant section of the core. Surrounding this sheet there are several (commonly four or five) α-helices.

Central to the RNase-H fold are the three catalytic residues. The first two residues are invariably aspartates, and are essential to the enzymatic activity. Mutagenesis of these residues removes any activity (Chapados et al. 2001; Naumann & Reznikoff 2002; Peterson & Reznikoff 2003). The first residue is located at the centre of β-strand 1 and the second is at the end of β-strand 4 (Figure 3a). The third catalytic residue is more variable, with either an aspartate or a glutamate depending on the core. This residue is located after the fifth β-sheet, commonly on an α-helix that runs across the front of the β-sheet, which places the three residues is close proximity. The function of the catalytic residues is to coordinate the two metal ions required for catalysis. This mechanism, first proposed by Steitz and Steitz (T. A. Steitz & J. A. Steitz 1993), has the first ion coordinating and activating the nucleophile while the other ion stabilises the oxyanion leaving group. For transposases, the optimal ion is Mg^{2+}, although Mn^{2+} can be used as an alternative (Hickman & Dyda 2016).

In some cases, an additional section is added into the RNase-H like fold known as an insertion domain. This domain is located after the final β-strand in the sheet, but prior to the final catalytic residue, and normally does

not significantly affect the structure of the core itself. The IS4 family has an ~100 residue insertion domain that, in at least the Tn5 transposase, has been shown to have involvement in the DNA binding (Davies et al. 2000) (Figure 3b). Hermes transposase, of the *hAT* superfamily, has a much larger ~300 residue insertion domain that is involved in DNA binding but also has a function with dimerisation of the protein (Hickman et al. 2014) (Figure 3c). The RAG1 insertion domain is ~250 residues and has been shown to be involved in the binding of the accessory protein RAG2 as well as some DNA-binding functionality (Ru et al. 2015) (Figure 3d).

1.4 How transposases capture the transposon end

The initial step of DNA transposition is the locating and binding of the transposase to its appropriate terminal or sub-terminal binding sites (TBS or STBS). This section will focus on the transposase binding and the various domains involved.

1.4.1 Binding sites in the transposons

All transposons have a TBS that is the primary binding site for the transposase directly involved in the catalytic reaction. In some IS4, Tn3 and *mariner* elements, amongst others, this is the only binding site for the

transposase at each terminus (Reznikoff 2008; Nicolas et al. 2015; Tellier et al. 2015). Transposons with longer TIR often have more than just the single binding site at the terminus. P elements and the Tc1 family have a single additional binding site at each end, which are direct repeats of the TBS (Vos 1994; Majumdar et al. 2015). Other systems have even more binding sites for transposases: Mu has three sites in each TIR, and Tn7 has three at the left and four at the right end (Harshey 2014; Peters 2014). The transposases can also bind multiple sites within the transposon at once, via additional interactions with the STBS. These can be single additional binding sites, such as in Tn10 (IS4 family) though many have multiple STBS: Mu has three, while P elements and the *hAT* superfamily demonstrate a widely varying number of binding sites (Jiang et al. 1999; Atkinson 2015; Sharmistha Majumdar 2015).

1.4.2 The Helix-Turn-Helix motif

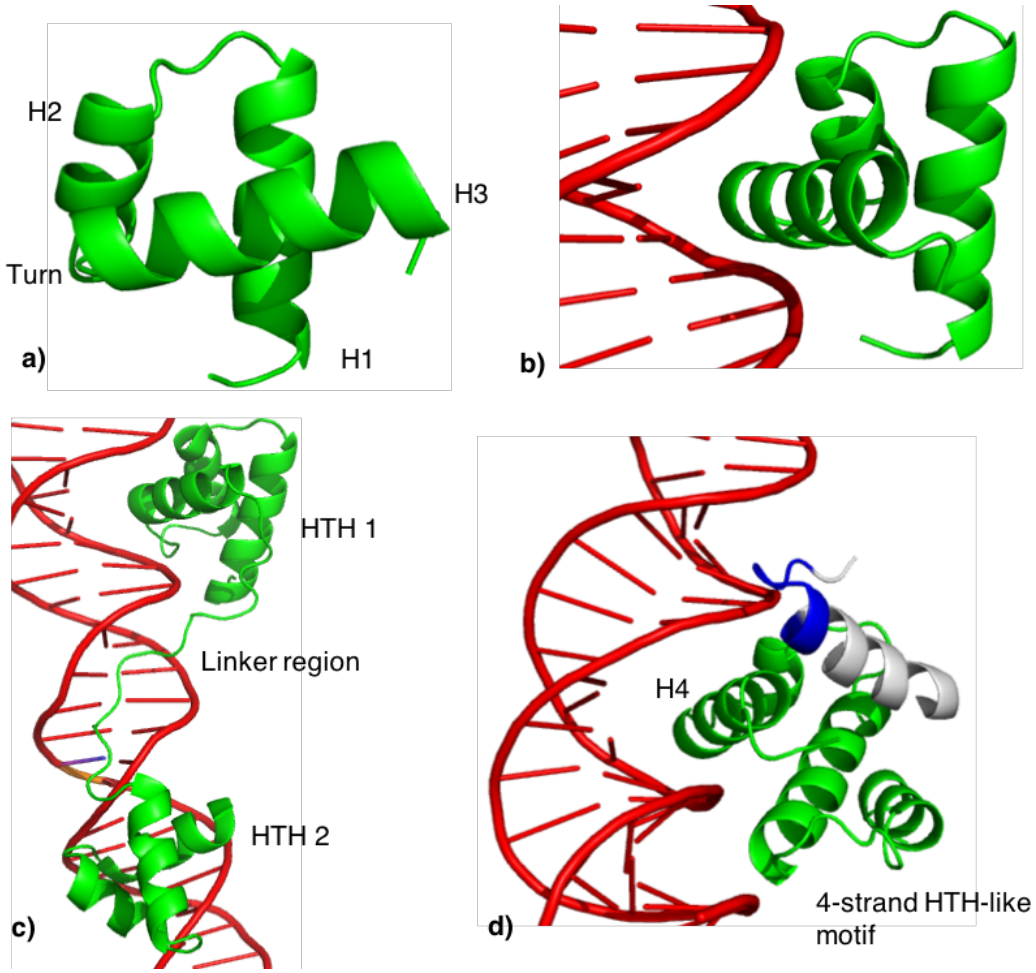


Figure 4: The HTH DNA-Binding Domains. HTH domains are shown in green, DNA in red, non-binding protein in grey and catalytic core DNA binding sections in blue. **a)** a representative HTH motif from the Pax5 transcription factor (PDB: 1k78). **b)** the same HTH motif, shown with DNA present. **c)** the DNA-binding domains from the Mos1 transposase, in complex with DNA (PDB: 3HOT). **d)** The DNA-binding domain of the Tn5 transposase (PDB: 1MUH).

One of the most common DNA-binding domains, seen in a huge variety of proteins, is the Helix-Turn-Helix (HTH) motif. It has been suggested that this motif may originate from the last universal common ancestor given its prevalence in all three domains of life (Aravind et al. 2005). The motif core consists of three α-helices. Helix one rises up the back of the motif, and helix two goes down the left side of the motif. There is a sharp four-residue turn from the end of the second helix, followed by the third helix which cuts across the face of the motif (Figure 4a). Helix three, often known as the recognition helix, forms the majority of the DNA-interactions, binding into the major groove of the DNA strand (Figure 4b). The second helix also has some involvement in the DNA interactions, while helix one performs a structural role in the motif (Aravind et al. 2005). There are many variants of the HTH motif that exist as well as the core tri-helical version. Some HTH motifs consist of four or more helices, generally located after the original three described above (Orth et al. 2000; Kostrewa et al. 2009). Some variants are also known to have a 'wing' on the motif, a two or three stranded β-sheet (Zhang et al. 2010; Chakravartty & Cronan 2013).

The most straightforward example of transposase HTH motifs comes from the ITm superfamily. These transposases have a distinct DNA-binding region that consists of two tri-helical HTH motifs separated by a short linker sequence. The best structural characterization of this superfamily is from the Mos1 transposase (Figure 4c) (van Pouderoyen et al. 1997; Watkins et al. 2004; J. M. Richardson et al. 2009). In Mos1, the first HTH acts as the predominant interaction motif in the region forming both specific, and non-

specific, interactions within the major groove: bases 21-26. This first HTH motif also functions in the dimerisation of the transposase. A 21 amino acid (aa) linker region stretches between the two HTH motifs with no significant secondary structure, but does interact non-specifically with bases 15-18 in the minor groove of the DNA. The second HTH motif is positioned approximately on the opposite side of the DNA strand, several base pairs from the first motif. This motif has multiple specific and non-specific interactions with the DNA at bases 8-13, though there are fewer than the first HTH.

The dual-HTH motif is not exclusively limited to the ITm superfamily. At least some of the replicative 'copy-in' transposases also use this binding domain. Sequence information for the Tn3 family transposases show that they have dual-HTH motifs with a short linker region, though at present no structural data have been produced to confirm this (Nicolas et al. 2015). The transposase for bacteriophage Mu, MuA, has also been shown to have a similar DNA-binding region. In this case there are the two motifs (labelled I-β and I-γ respectively) with a short linking sequence, although the second motif is a Helix-Loop-Helix motif, where the 4-residue turn is replaced by a 6-residue loop. Low-resolution crystal structures of the domain in contact with DNA show that they bind in the same manner as those of the ITm family, with the first and second motifs on opposite sides of the DNA strand. The first helix binds around bases 5-10 and the second binds around bases 15-25 (Montaño et al. 2012).

MuA has a third DNA-binding motif, known as I-α, located N-terminal of I-β and I-γ. This domain is a winged HTH motif which includes three β-strands, one located before helix one and the other two between helices two and three (Clubb et al. 1994). This domain has not been structurally characterised in contact with DNA. However, it has been shown to bind to STBSs, (known as enhancer sites) in the Mu transposon (Leung et al. 1989; M. Mizuuchi & K. Mizuuchi 1989).

Tn7 is an unusual transposition system as its 'transposase' is actually two different proteins, TnsA and TnsB, in complex, though only TnsB contains the actual RNase-H fold. The TnsB DNA-binding domain has not been structurally solved, but structural predictions of the domains suggest it has at least one HTH motif, possibly two in the same format as the dual-HTH motif discussed above (Z. Li et al. 2013). TnsA is part of the PD-(D/E)XK phosphodiesterase superfamily, and has its DNA-binding domain at the C-terminus of the protein (Hickman et al. 2000; Steczkiewicz et al. 2012). The domain consists of a single HTH motif, and forms both DNA and protein interactions (Ronning et al. 2004; Choi et al. 2013).

The binding domain in the IS4 elements is structurally similar to HTH motifs, though it consists of four α-helices as opposed to three (Figure 4d). Helix four is the 'recognition helix' and forms the largest quantity of DNA interactions in the motif, both specific and non-specific. Helices two and three also form some DNA interactions, but these interactions are predominantly non-specific. This single DNA-binding motif is also supplemented by another small set of interactions from residues shortly after the third catalytic residue

in the core domain. Both specific and non-specific DNA interactions come from a loop between two α-helices (Davies et al. 2000). DNA binding from both sections occurs between bases 5 and 17 in Tn5 (Reznikoff 2008). In Tn10, extensive mutagenesis on the TIR showed that bases 6-13 are key to the transposase binding, which strongly indicates base-specific interactions (Huisman et al. 1989). This suggests that these binding domains are broadly conserved across the IS4 family.

1.4.3 Zn-based DNA-binding domains

Unlike the HTH motifs, Zn-based DNA-binding domains are a much more varied sub-group of protein-motifs. However, the capacity to coordinate zinc ions does require some conserved features. The residues commonly associated with Zn-binding are cysteine and histidine, classically as a Cys_2-His_2 motif, that coordinate a single Zn^{2+} ion. Zn^{2+} performs a primarily structural component in the motif, used to stabilise an α-helix that binds to the DNA strand (Laity et al. 2001; Pace & Weerapana 2014).

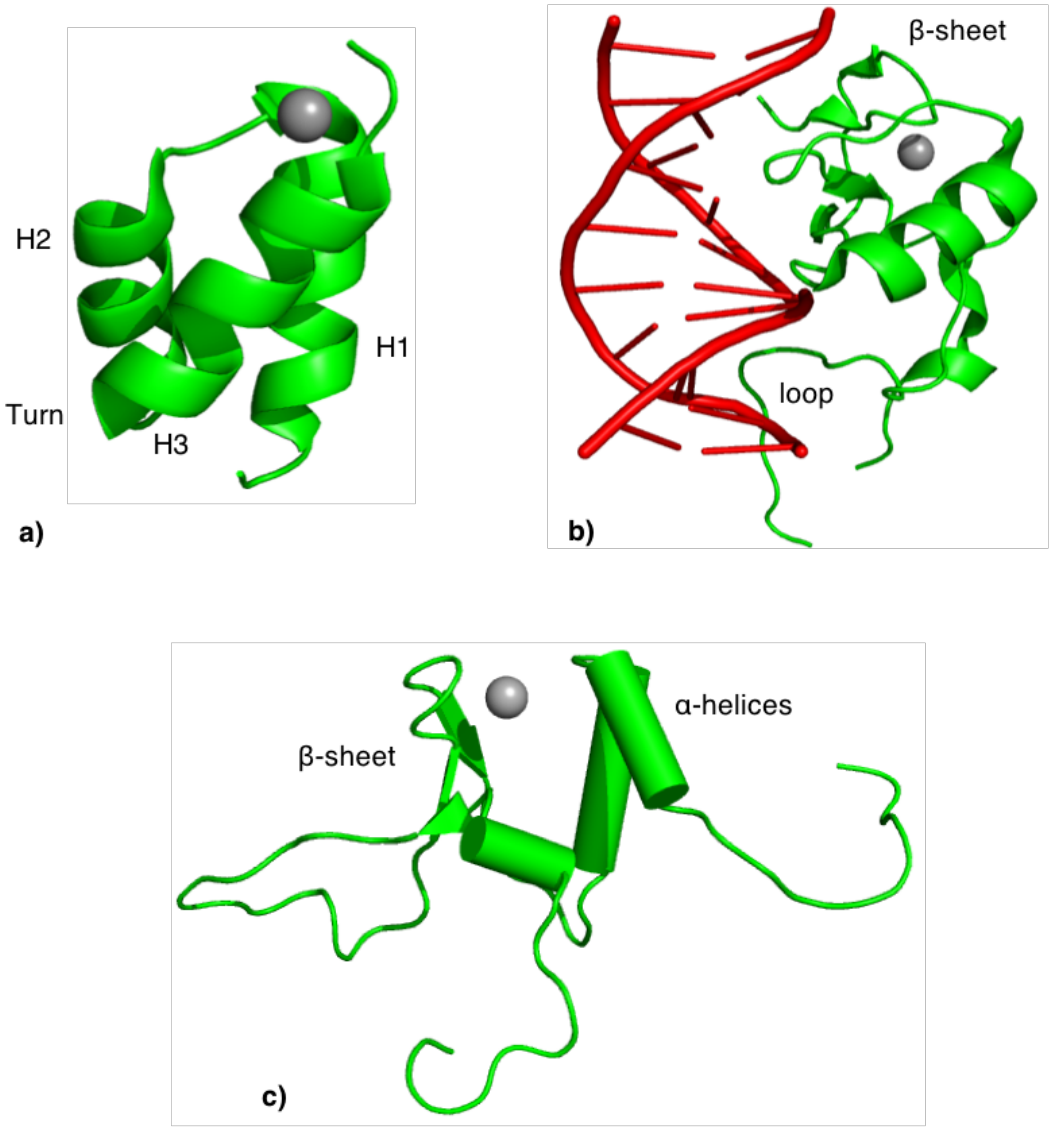


Figure 5: Zn-Based DNA-Binding Domains. The domain itself is shown in green, with the DNA, where available, in red, and the Zn^{2+} ion in grey. **a)** The Zn-HTH motif from HIV integrase, shown without DNA (PDB: 1K6Y). **b)** The THAP domain from the *D. melanogaster* P element, in the presence of DNA (PDB: 3KDE). **c)** BED domain, as predicted for the Hermes transposase, based on a solution structure (PDB: 2CT5).

The simplest Zn-based DNA-binding domain in transposases is found in the retroviral integrase superfamily, best characterised in the Prototype-Foamy-Virus (PFV). The domain is actually an HTH motif that also coordinates a Zn^{2+} ion, known as a Zn-HTH (Figure 5a). The Zn^{2+} ion is coordinated by a classic Cys_2-His_2 motif, with both histidines located at the end of helix one and both cysteines at the end of helix three. In PFV, this domain forms a series of site-specific contacts with the DNA strand between bases 8-13. For members of the spumaviral and gammaretroviral families, there is also an additional domain called the N-terminal Extension Domain (NED) which has a single α-helix and a three-stranded β-sheet, and does not bind Zn (Hare et al. 2010). This domain forms a series of non-specific contacts with the DNA strand further away from the cleavage site at bases 16-18. In PFV for example, this domain is not essential for activity, suggesting that its function is to enhance DNA binding (Pahl & Flügel 1995).

P elements make use of a more complex Zn-binding motif called a THAP domain (Figure 5b). This domain contains a Cys_2-Cys-His Zn-binding motif with a characteristic central α-β-α fold. The central β-sheet formed from the fold binds into the major groove of the DNA, while the minor groove of the DNA is bound by a C-terminal loop structure (Sabogal et al. 2010). In at least some of the P elements there are multiple STBSs for the THAP domain near the TIR, theorised to provide the transposase with greater binding specificity (Marzo et al. 2013).

The last of the well-known transposase Zn-binding motifs is found in the *hAT* superfamily. These transposases all contain a BED domain at the

N-terminus of the protein, though the domain itself has not been crystallised from any *hAT* transposases (Figure 5c) (Arensburger et al. 2011). The domain is predicted to consist of a three-stranded β-sheet followed by at least one α-helix with the residues (Cys_2-His-Cys) coordinating the Zn^{2+} located on β-strands two and three and on the α-helix (Aravind 2000, Hickman et al. 2014). The function of the BED is to bind to essential STBS in the transposon DNA (Kim et al. 2011). These sites provide additional specificity to the transposase binding, as the TIR sequence alone occurs frequently in large genomes and without other factors, there would be many off-target transposition events that would adversely affect the host (Hickman et al. 2014).

1.4.4 The intertwined binding domain

There is one other type of DNA-binding domain used in transposase proteins, here referred to as an Intertwined Binding Domain (IBD). IBDs are distinct as they only bind to DNA after they have formed a dimeric structure with the same domain of another transposase monomer, and both domains interact with both TIR sequences.

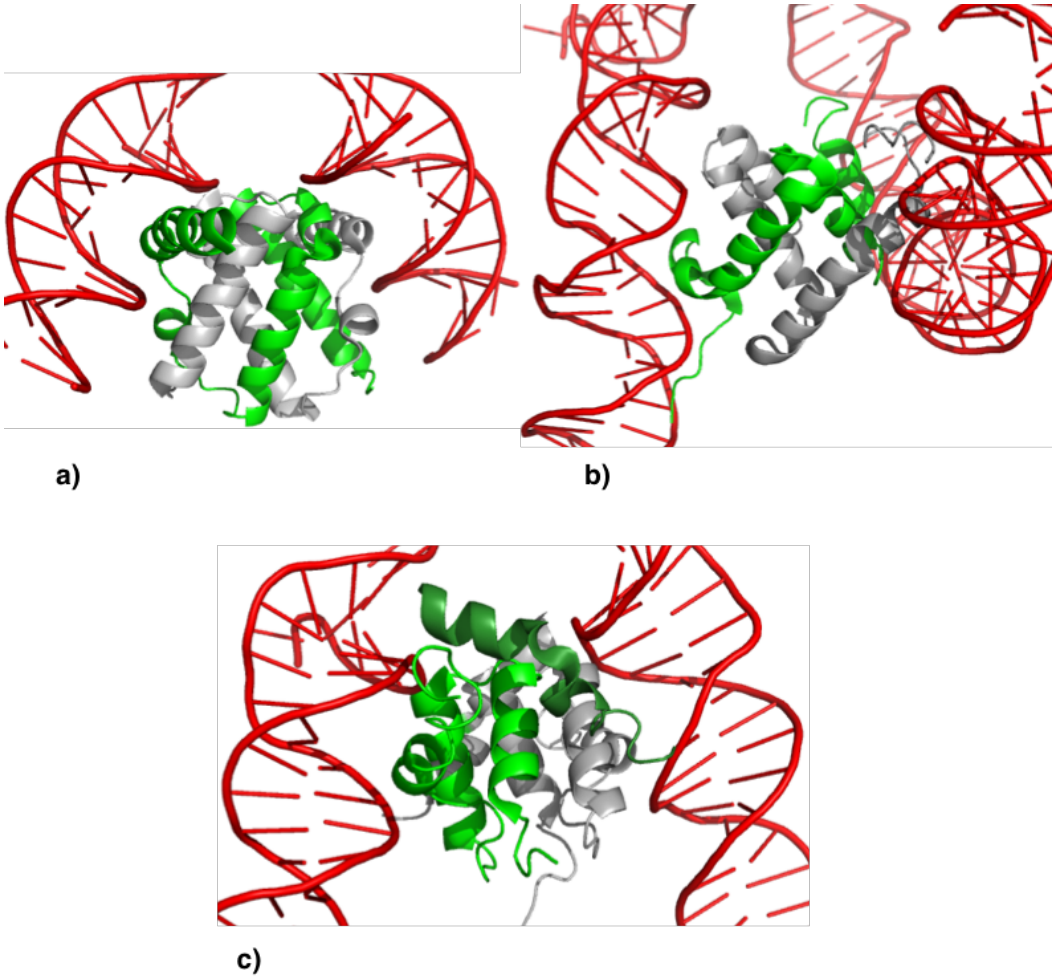


Figure 6: Intertwined DNA-Binding Domains. The domain itself is shown in green for one monomer and grey for the other, with the DNA in red. **a)** The Hermes IBD (PDB: 4D1Q). **b)** The first IBD from RAG1, the Nonamer Binding Domain (PDB: 3JBX). **c)** The second IBD from RAG1, the Dimerisation and DNA Binding Domain. The dark green α-helix is the part of the domain from the C-terminus of the protein (PDB: 3JBX).

The Hermes transposase and other members of the *hAT* superfamily utilise this type of domain (Figure 6a) (Hickman et al. 2005, Hickman et al. 2014). In Hermes, the domain is located near the N-terminus of the protein, just after the BED domain, and consists of five α-helices from each monomer that are intertwined with one another. The fifth α-helix makes a site-specific contact (base 11 of the TS) in the major groove, while residues in the fourth helix coordinate a Na^+ ion that makes non-specific DNA contacts to the phosphate backbone between bases 11-12. Very shortly after the end of helix one there is an additional non-specific interaction with the DNA backbone, between bases 7-8 of the NTS. A section close to the end of the first helix binds to the other TIR, forming nonspecific interactions at the extremity of the major groove between bases 6-7 (Hickman et al. 2014). This domain is broadly conserved across the *hAT* superfamily (Arensburger et al. 2011).

The other well-studied transposase to make use of this type of domain is the RAG1 protein, which is involved in V(D)J recombination. The binding site for the RAG1 transposase known as a Recognition Signal Sequence (RSS). The site consists of a heptamer element and a nonamer element, which flank either a 12 or 23 bp spacer region. The first IBD of RAG1, the Nonamer Binding Domain (NBD), binds to the nonamer element of the RSS (Figure 6b). The domain is a three-helical motif where helices two and three form the protein interface. An AT-hook at the N-terminus (prior to helix one) forms specific interactions in the minor groove of DNA at bases 5-7, while the first part of helix one binds specifically to base two of the nonamer. This

domain also makes non-specific contacts with the spacer region of the RSS through the N-terminal portion of helix one. The other DNA strand is bound through a series of non-specific interactions from the loop region situated between helices two and three, binding the spacer region (F. F. Yin et al. 2009). RAG1 has a second IBD located C-terminal of the first, called the Dimerisation and DNA-Binding Domain (Figure 6c). This consists of four helices; three are located in the primary sequence after the first NBD, while the other originates from the C-terminus of the protein. These domains all interact non-specifically with the RSS spacer region (Ru et al. 2015; Rodgers 2017).

1.5 Secondary DNA-Binding Activity

1.5.1 DNA binding by the catalytic core

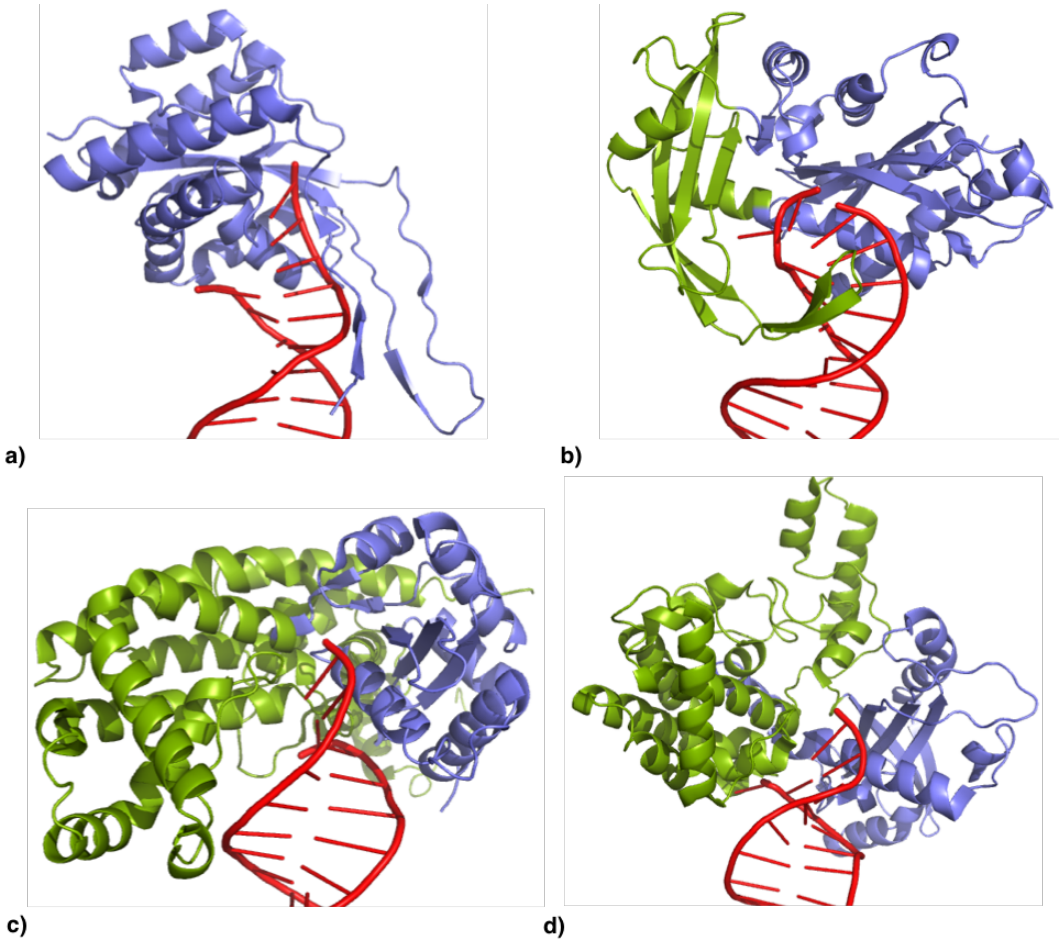


Figure 7: Catalytic Core Interactions with DNA. The catalytic core of one protein monomer is shown in blue, with the insertion domain shown in green and the DNA helix in red. **a)** The Mos1 transposase (PDB: 3HOT), with the DNA showing a 3 bp overhang on the TS. **b)** The Tn5 transposase (PDB: 1MUH), with the DNA showing no overhanging base pairs but a flipped-out base at the penultimate position on the NTS. **c)** The Hermes transposase (PDB: 4D1Q), with the DNA showing a 1 bp overhang on the TS. **d)** The RAG1 transposase (PDB: 3JBX), with the DNA showing no overhanging bases but the final base on the NTS is flipped out of the helix.

The catalytic core domains contribute a high proportion of the transposase interactions with the DNA, in particular the interactions close to the TIR cleavage site. This section will look at these interactions with the TIR, by both the core, and the insertion domains if they are present. Only five families of RNase-H transposases have solved X-ray crystallography structures. In all these complexes, each DNA strand is bound *in trans*, which means that the DNA-binding domains of the protein are attached to one TIR, while the catalytic core interacts with the other. All interactions in this section should be assumed to be with the TIR that is being reacted on by the catalytic core, unless otherwise stated.

The IS4 family has well-characterised interactions with the TIR sequence close to the site of cleavage (Figure 7b). Extensive biochemical analysis in Tn10 showed that bases 1-3 and 6-13 were bound by the transposase in a site-specific manner (Huisman et al. 1989). The crystal structure of the Tn5 transposase (Davies et al. 2000) demonstrated that the core forms interactions with predominantly the first three bases of the TS. The insertion domain forms contacts mostly further away from the cleavage site, bases 5-8, though it does play a key role with base 2 of the NTS. This base is flipped out of the DNA by a tryptophan probe residue, during first strand cleavage, and is retained in an extrahelical position until after the transposon is fully excised. This essential mechanism provides the flexibility to generate the hairpin intermediate, as well as providing an ‘anchor’ to maneuver the DNA strand. The flipped-out base is seen in multiple IS4 elements, therefore is likely a conserved feature of the family (Davies et al.

2000; Ason & Reznikoff 2002; Bischerour & Chalmers 2007; Bischerour & Chalmers 2009). Another highly-conserved motif in IS4 elements, the (R)YREK motif, serves to move the 5' NTS away from the TS after the hairpin has been cleaved, through interactions with the 5' phosphate. This is to allow space for the initial integration reaction (Klenchin et al. 2008).

The ITm superfamily have a much smaller RNase-H core than the IS4 elements, with no insertion domain within it and overall fewer interactions with the DNA (Figure 7a). The post-excision DNA strand has a 3 bp overhang on the TS. This overhang is base-specifically bound by the linker region that connects the DNA-binding domains and the core domain. The core itself forms interactions around the fourth base of both strands, the first base-pair after the overhang, reinforced by what is called the 'clamp loop'. The clamp loop is a region of the catalytic core domain that stretches up and sits alongside the TIRs in the complex forming interactions with both, particularly the DNA bound to the opposing monomer (J. M. Richardson et al. 2009). The first base of the conserved flanking DNA dinucleotide, a thymine, has been shown to be key to the TS cleavage reaction in at least the Hsmar1 element (Claeys Bouuaert & Chalmers 2010; Claeys Bouuaert et al. 2014). A recent crystal structure of Mos1, which included the flanking DNA, showed that two conserved motifs, WVPHEL (in the linker) and YSPDL (in the core between active site residues two and three) form a pocket that the flanking thymine base fits into, in order to position the strand for cleavage (Dornan et al. 2015).

The *hAT* superfamily of transposons make use of a 5-27 bp TIR. The crystal structure of the Hermes transposase shows that its catalytic core and the large insertion domain bind to the first 8 bp of this region (Figure 7c) (Hickman et al. 2014). The site-specific binding is predominantly from the catalytic core itself, while the insertion domain forms mostly non-specific interactions to the TS of the TIR. The 1 bp overhang of the excised DNA is derived from the donor DNA, not the transposon itself. The overhanging nucleotide is known to be removed from the transposon by the time it has successfully integrated into the target site (L. Zhou et al. 2004; Atkinson 2015).

The RAG1 complex interactions around the cleavage site have been well characterised (Figure 7d). The heptamer of the RSS is cleaved with no overhanging bases, with the terminal base on the NTS flipped out of the helix. The flanking DNA, which is known as the coding DNA in V(D)J recombination, is cleaved, leaving a blunt end. The catalytic core region of RAG1 makes extensive contacts to the first four bases of the heptamer region, both specific and non-specific. These contacts are primarily on the TS of the heptamer though there are some on the NTS bases 2-4. The terminal base of the TS is flipped out of the helix and held in place by the first region of the insertion domain (the ZnH2 region). This region is responsible for a number of interactions around the flipped base, as well as base two of the NTS. This strand interaction continues across the site of cleavage into the coding DNA, where the ZnH2 region forms non-specific interactions up to base six. The other part of the insertion domain (the ZnC2

region) has a very limited role in DNA binding, but has a number of mostly non-specific interactions with the coding DNA up to base six (Ru et al. 2015).

Retroviral integrases have two domains that are responsible for DNA interactions near the site of catalysis, before the integration step of the reaction occurs: the RNase-H catalytic core, and the C-terminal Domain (CTD). The majority of the interactions are with the NTS, particularly the terminal 3 bp, two of which overhang the TS. These 3 bp are flipped away from the helix, and held there by base-specific interactions from both the catalytic core and CTD. Two of the key residues for the flipped bases are from the catalytic core, which act as a 'plough' to push the NTS bases out of the DNA helix. The catalytic core has multiple specific and non-specific interactions up to base eight of the NTS, as well as two non-specific interactions with the TS. The CTD has interactions with both the TIRs in the complex. Its interactions *in cis* are with the three flipped-out bases of the NTS, predominantly non-specific. *In trans*, the interactions are both specific and non-specific, with the TS (bases 2-11) and NTS (bases 8-10) (Hare et al. 2010). Amongst those we have discussed here it is unusual how few interactions are on the TS compared with the NTS. As this is retroviral DNA, it is not excised at the start of the reaction from flanking DNA, however there is 3' end processing that removes the two additional nucleotides from the TS. These two additional nucleotides have only Van der Waals interactions with the retroviral integrase core (Hare et al. 2012).

These structures point to a general consensus for the catalytic region. The RNase-H fold sits mostly at one side of the DNA strand, and requires

some other protein segment to form interactions with the other section. This is the linker from Mos1 or the insertion domains in Tn5, Hermes, and RAG1. Both PFV and retroviral integrases use DNA interactions, *in trans,* to stabilise the other side of the DNA from the predominant catalytic core binding.

1.5.2 Target DNA binding

The binding of the target DNA is not as well characterised as that of the TIR, mostly due to its quite transient nature. The best understood example is probably the Mos1 transposase. When the target integration site was predicted based on the post-excision complex of Mos1, the target site was located into a cleft at the base of the RNase-H core, perpendicular to the TIR DNA (J. M. Richardson et al. 2009). A recent structure has shown the complex after the TS has integrated into target DNA (Morris et al. 2016). What this demonstrated was that the target DNA did fit into the cleft at the base of the core as predicted. It also showed that the TA dinucleotide, which is the essential target site, becomes significantly distorted (147^{o}) in the reaction, due to the lower base-base interactions of the TA. The adenines from both strands of the TA dinucleotide site is flipped out in the structure, into a base-specific pocket of the catalytic core domain. This flipping out is assisted by a conserved phenylalanine which acts a probe to displace the base. The conservation of the residues involved suggests that base-flipping and target distortion are conserved among ITm members (Morris et al. 2016).

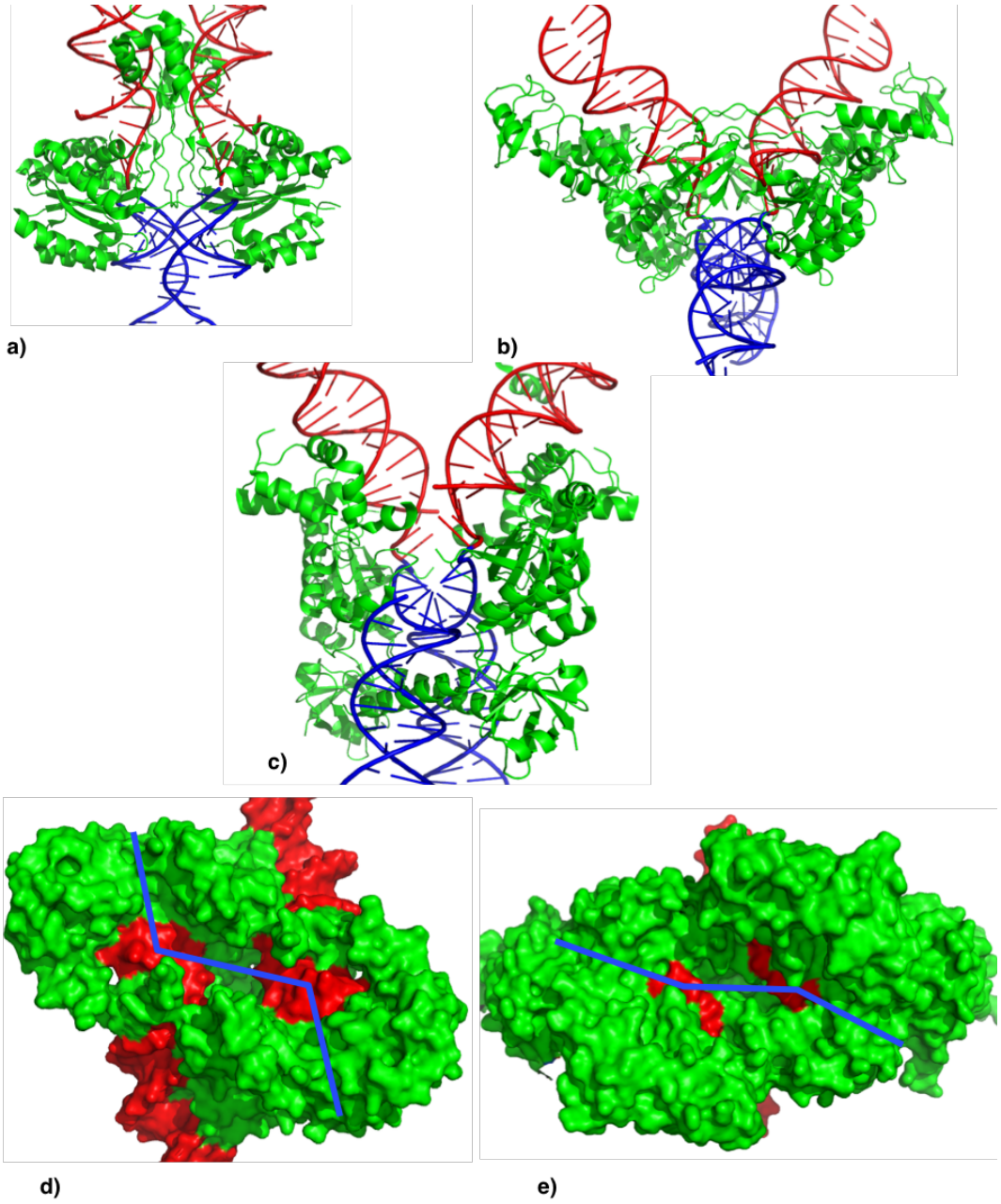


Figure 8: Target DNA Binding of Transposase Complexes. The transposases are shown in green, with the TIR DNA in red and the target DNA in blue. **a-c)** Crystal structures of the partial integration complex of the transposases, shown in cartoon form. **a)** Mos1 complex (PDB: 5HOO. **b)** PFV complex (PDB: 4BAC). **c)** MuA complex (PDB: 4FCY). **d-e)** The crystal structures of transposition systems, shown as surface structure, that have the target DNA-binding region predicted (shown by the blue line). **d)** Tn5 complex (PDB: 1MUH). **e)** Hermes complex (PDB: 4D1Q).

The only other complex with significant structural information on the strand-transfer complex is the retroviral integrase PFV (Figure 8b). This showed a partial strand transfer structure that bears some similarities to the ITm structures. The target sequence fits into the cleft between the far side of the catalytic cores (away from where the TIR runs into the complex), running close, and perpendicular, to the excised TIR ends. Target DNA is bent 55^{o}, using the central TA of the CTAG target sequence to be the apex of the curvature (Maertens et al. 2010; Krishnan & Engelman 2012; Z. Yin et al. 2012). This TA is not essential to the reaction as it is in ITm however, and other base conformations are known to be viable, possibly as a result of the lower distortion required on the helix (compared with the Mos1 transposase).

The Bacteriophage Mu transposase strand-transfer complex is not as well defined as the Mos1 and PFV structures, due to low-resolution crystallography. What is shown however, is a 140^{o} bend in the target DNA, a severe distortion similar to the ITm superfamily (Montaño et al. 2012). This is maintained by multiple sub-domains of the MuA transposase. The RNase-H core, subunit II-α, has interactions mainly through a 20-residue loop that extends under the target DNA, while II-β has a positively charged β-barrel structure that binds the target DNA (Rice & Kiyoshi 1995; Krementsova et al. 1998). The III-α subunit, known as a BAN domain (Binding and Nuclease), is positioned in the middle of the bent target DNA loop, preventing the DNA leaving the site, and also stabilising the complex by reducing repulsion from the arms of the DNA loop (Z Wu 1995; Montaño et al. 2012; Harshey 2014).

The target binding clefts seen in the ITm and retroviral integrase families are a common occurrence, though in some transposases it has not been structurally confirmed to hold the target DNA. The Tn5 transposase shows a cleft that runs along between the two TIR DNA ends in the post-excision complex, and the distance between the two ends is sufficient to accommodate the 9 bp target site. Based on the cleft, there is likely significant bending of the target DNA in order for it to be accommodated, particularly around the site of integration at each end. Nicking of these sites can enhance the target capture and integration steps of the reaction (Whitfield et al. 2006; Reznikoff 2008). The cleft of the Tn5 transposase has been shown to bind to target DNA through significant mutagenesis of the region (Gradman et al. 2008).

The Hermes transposase of the *hAT* superfamily also has a cleft at the outside of the RNase-H core, which can comfortably accommodate the 8 bp target sequence if it is bent. It is suggested that the DNA is bent at the T2 and A7 sites in the target sequence which are the only significant target site preferences of Hermes (Hickman et al. 2014). The *hAT* superfamily does not show a conserved target sequence preference at bases two and seven which suggests that the target DNA bending varies significantly within the superfamily.

The Tn7 transposition system has two possible integration pathways with two separate types of target DNA, both of which differ from those seen previously. The first is the highly sequence-specific integration mechanism, that uses the TnsD protein in complex with TnsA-C. TnsD binds to a 36 bp

region in *glmS*, which is an extremely well-conserved gene (Milewski 2002). Unlike many other transposition systems, the target binding site is not the same as the integration site; this is located downstream, in the transcriptional terminator region of *glmS*, which prevents the transposon integration damaging the essential gene and therefore killing the host (Gay et al. 1986). TnsD binding has not been characterised at present. However, it is known to have at least one binding domain, a zinc-finger, at the N-terminus of the protein (Mitra et al. 2010). The alternative integration pathway uses the TnsE protein in place of TnsD, and occurs at a much lower frequency. TnsE integrates in a non-sequence specific manner, preferentially into conjugative plasmids using the 3' recessed ends that occur at replication forks (Wolkow et al. 1996). This preference increases the frequency of the transposon spreading between species. The DNA-binding domains of the TnsE are unknown at present, though they are known to be located at the C-terminus of the protein (Z. Li et al. 2013).

1.5.3 DNA binding of host factor and accessory proteins

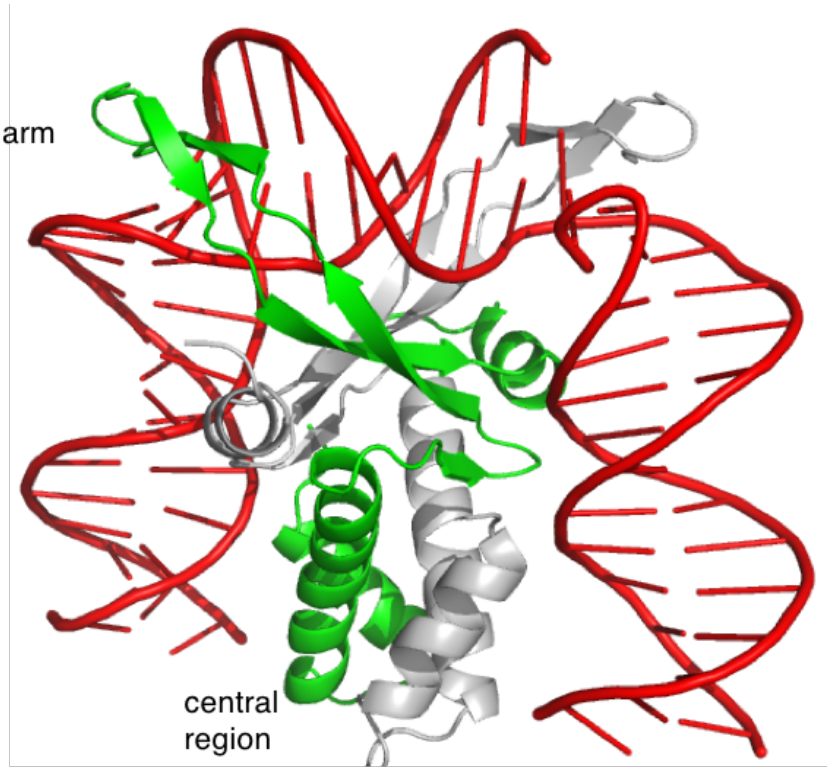


Figure 9: IHF DNA-Binding Complex. IHF-β is shown in green, while IHF-α is in grey. The DNA is shown in red. (PDB: 1IHF).

Some transposition systems make use of host proteins in order to facilitate additional DNA-binding functions, without needing to encode more proteins themselves. The most common example of the host proteins that are involved are those of the IHF/HU family, predominantly the Integration Host Factor (IHF) (Figure 9). IHF is a heterodimeric protein complex, ~21 kDa in size. Each protein (IHF-α and IHF-β) consists of a central region, responsible for the protein-protein interactions, and an 'arm' region that wraps around the DNA helix, inserting into the minor groove. A proline residue at the tip of the arm loop becomes intercalated between base pairs in the groove,

forming a kink which bends the DNA sharply. The other protein in the complex does the same at the minor groove downstream and between them they can bend the DNA ~180^{o} (Rice et al. 1996).

The IS4 family element, Tn10, makes use of the IHF protein. There is an IHF-binding site situated in the 'outside' transposon end, which is located shortly after the TIR (Sakai et al. 2000). The bend generated by IHF allows Tn10 to interact with the DNA at a STBS, which appears to help stabilise the synaptic complex (J Sakai 1995; R M Chalmers 1996; Haniford & Ellis 2015). The other well-known example of the IHF/HU family in transposition is in the Bacteriophage Mu system. There is an IHF binding site located in the enhancer region of the transposon, which is posited to optimise the interactions of the enhancer sites with the transposon end complex (Harshey 2014). There is also a site for the HU protein (a homolog of IHF (Bonnefoy & Rouvière-Yaniv 1991)) between the first and second transposase-binding sites in the left transposon end. The left transposon end has a large gap between transposase-binding sites L1 and L2, so HU bends the DNA, orientating the L2 site back into proximity to form part of the synaptic complex (Montaño et al. 2012; Harshey 2014).

There is another set of host proteins seen in transposition that are responsible for DNA bending, the HMGB proteins. HMGB1/2 are found in the synaptic complex of V(D)J recombination with RAG1 and RAG2. These HMGB proteins are made up of two tri-helical HMGB domains, which are responsible for binding the 23 bp spacer region through both specific and non-specific interactions with the minor groove (Stott et al. 2006; Ciubotaru

et al. 2013). Both HMGB1 and HMGB2 are functionally identical in this context (Schatz & Swanson 2011).

Some transposition systems also have accessory proteins that are involved in the synaptic complex formation pathway. Bacteriophage Mu produces one such protein, called MuB. The function of MuB is to bind to the transposon DNA to prevent integration into itself and also to bind target DNA and capture it for the transposition reaction. MuB belongs to the AAA+ ATPase family of proteins and consists of two domains, an N-terminal appendage, and the ATPase domain. The protein forms an almost continuous polymer around DNA, interacting through a loop region in the ATPase domain. ATP hydrolysis causes MuB to dissociate from the DNA. The presence of DNA reduces ATPase hydrolysis. However, when MuA interacts with MuB, there is an increase in ATPase hydrolysis, and MuB then dissociates from the DNA, giving the MuA synaptic complex access to the DNA as a target site (Mizuno et al. 2013).

Another accessory protein is RAG2, which occurs in V(D)J recombination. RAG2 forms part of the synaptic complex, binding to the insertion domain of RAG1. The RAG2 protein is made up of six Kelch-like motifs that form a six-bladed β-propeller structure (Rodgers 2017). Within one of these Kelch-like motifs there is also a PHD domain, which is known to bind to histone-3 lysine-4 tri-methylation (Elkin et al. 2005; Matthews et al. 2007; Liu et al. 2007). RAG2 in known to be responsible for binding to the flanking coding DNA, bases 3-6 and 9-16 away from the cleavage site,

through a series of both specific and non-specific interactions (Ru et al. 2015).

1.6 HUH nuclease fold

The only non-RNase-H containing transposases that have been well characterised are those with an HUH nuclease fold. In this core, instead of water molecules as the nucleophile, specific tyrosines are used, and form 5' phosphotyrosine intermediates. There are two types of these cores, Y1 and Y2, depending on the number of tyrosines that can form the intermediate in the active site. Transposases with Y1 cores, such as the IS200/IS605 family, use the 'peel-and-paste' mechanism (Figure 10a) (He et al. 2013). Those that use Y2 cores operate using the 'rolling circle' pathway, such as in Helitrons and IS91 (Kapitonov & Jurka 2001; Thomas & Pritham 2015; Mendiola & la Cruz 1992). Only the IS200/IS605 family will be mentioned here due to limited study of the other elements.

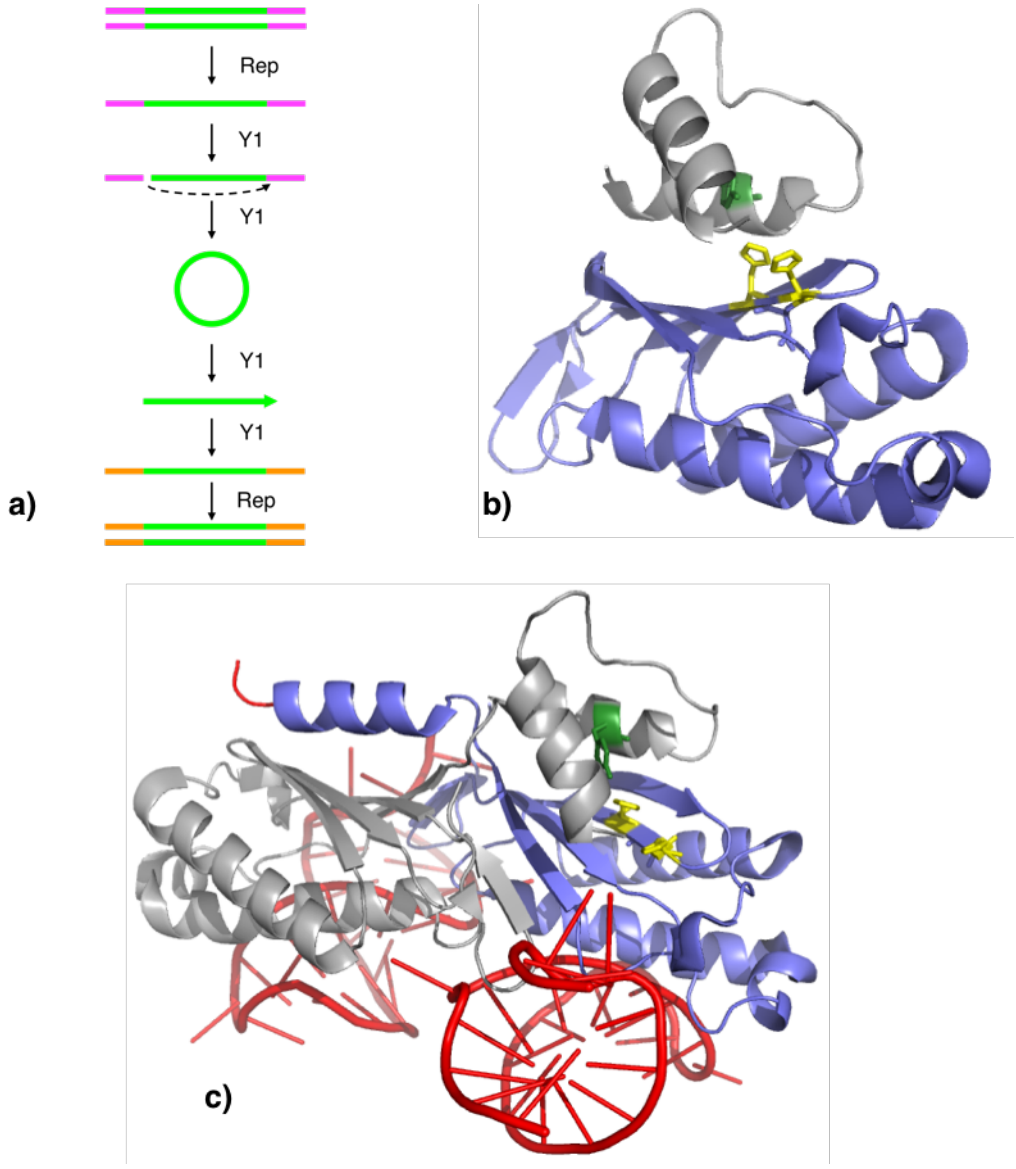


Figure 10: The IS200/IS605 Transposition System. a) Diagram representing the 'peel and paste' transposition mechanism. Donor DNA shown in magenta, transposon DNA in green and target DNA in orange. The arrow indicates the 3' end of the DNA strand. Rep=Replication. The ssDNA transposon is excised and forms a circular intermediate. This is then split into a linear form and integrated into ssDNA at a new locus. **b-c)** Protein structures of the IS608 transposase complex (PDB: 2VIH). One monomer of the transposase is shown in blue, with the other shown in grey, and DNA shown in red. The HUH motif is shown in yellow, with the catalytic tyrosine shown in dark green. **b)** The catalytic core, with the central domain of one transposase, and the tyrosine arm of the other. **c)** The full crystal structure with DNA present.

1.6.1 Mechanism of the IS200/IS605 family

The mechanism used by IS200/IS605 is the replicative system known as 'peel-and-paste' (Figure 10a). This pathway requires ssDNA which is located in the lagging strand of a DNA replication fork. The transposase dimer binds to each end of the transposon, cleaving the ssDNA using the tyrosine nucleophile to form a 5' phosphotyrosine intermediate. One phosphotyrosine intermediate is formed with the flanking donor DNA, at the 3' end of the transposon. This will then attack the free end of the donor DNA at the 5' end of the transposon, which repairs the excision site. The other phosphotyrosine intermediate, at the 5' end of the transposon, attacks the free 3'OH on the opposite end, forming a circular intermediate with the two transposon ends joined head-to-head in an inverted repeat configuration. The circular transposon intermediate, and the target site, are then cleaved in the same manner, before integration into single-stranded target DNA (Barabas et al. 2008; He et al. 2013).

1.6.2 Structure of the IS200/IS605 family transposases

The transposases of this family are some of the smallest-known of their type, only 155 aa in the case of IS608 (Figure 10b) (Guynet et al. 2008). Each monomer consists of two subdomains. The first has a 4-stranded β-sheet which is flanked on one side by a three-helical section. This β-sheet contains

the HUH motif which coordinates the single Mg^{2+} ion on the third β-strand. The other subdomain contains a two-stranded β-sheet and two α-helices, with the catalytically active tyrosine on one of the α-helices (Ronning et al. 2005). These proteins are obligatory dimers in solution, and each active site is made from the HUH-containing subdomain of one monomer and a tyrosine-containing subdomain. The two tyrosine-containing subdomains have been shown to swap between the active sites throughout the transposition process (Barabas et al. 2008; He et al. 2013).

1.6.3 The DNA-binding activity of IS200/IS605 family transposases

The DNA binding of IS200/IS605 elements is not similar to any of the previously discussed systems. Instead of binding to a specific sequence, the transposases bind to a stem-loop structure that forms at each end of the ssDNA transposon (Figure 10c). The stem-loop was first characterised in IS608 where it consists of a 22-base sequence at the right-hand end of the transposon (Ronning et al. 2005), though more recent structures have shown the left-hand end to be 23 bases in length (Barabas et al. 2008). The stem of both structures is an imperfect complement, as it contains an additional T base on one side, which is recognised by the transposase and flipped out in the various structures. The loop structure at the top of the stem-loop has either two or three thymine bases within it, that are also flipped out of the

structure. There are a multitude of other non-specific interactions with the stem loop; the top four base pairs and the loop are contacted by the HUH-core subdomain of the transposase, while the bottom six base pairs and the un-paired thymine are contacted by the tyrosine-containing subdomain from the other transposase.

As the binding site does not have the cleavage site within it, like many other transposons, there is a mechanism to recruit this site into the proximity of the transposase. The cleavage site is denoted in IS608 by a tetranucleotide sequence, which has a 75% homology to a sequence at the 5'base of the hairpin. This homology leads the two sites to pair up, which places the scissile phosphate into the active site, allowing cleavage at the 3' of the sequence (He et al. 2015).

The cleavage at the 3' ends of both tetranucleotide sequences means that the sequence at the 5' end of the transposon remains in the donor DNA, because this is the sequence used as the target for integration. The target sequence does need to be located on ssDNA in order for integration events to occur, so the lagging strand of a replication fork is needed once again. This is a highly-conserved feature amongst the IS200/IS605 family (Kersulyte et al. 2002; Guynet et al. 2008).

There are other members of the IS200/IS605 family that have also been characterised, such as ISDra2 (Hickman et al. 2010). This system has a distinctly different stem-loop structure, that has a G-T mismatch formed part-way up the stem, similar to what was in IS608 but in this case causing deformation of both strands.

1.7 Mechanisms of synapsis

Although the DNA-protein interactions of the transposases are important, the dynamics of the synaptic complex assembly pathway also play a key role. There seem to be only two general pathways that have been identified: Synapsis-by-Protein-Dimerisation (S-PD) and Synapsis-by-Naked-End-Capture (S-NEC) (Claeys Bouuaert et al. 2013).

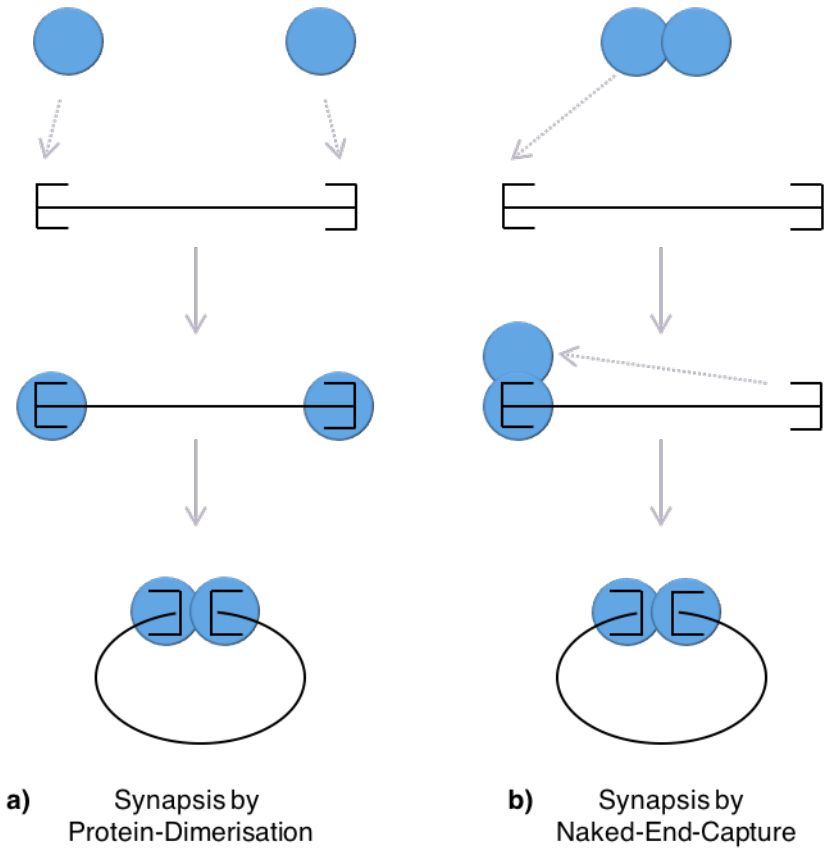


Figure 11: The Complex Formation Pathways of Transposition. The transposase proteins are shown in blue, the transposon DNA in black, with brackets to indicate the TIR at each end. **a)** Synapsis by Protein-Dimerisation. Monomeric transposases bind to each end independently, then combine to form the PEC. **b)** Synapsis by Naked-End-Capture. A Transposase multimer binds to one TIR, and then the other TIR is recruited into this complex to form the PEC.

In S-PD (Figure 11a), the transposase is monomeric in solution. Each monomer must locate the TIR sequence independently and bind to it, to form the Single End Complexes (SECs). Once both TIRs of a transposon are bound, the two SECs can combine to form the full Paired End Complex (PEC), which can then proceed with the reaction steps. The most well-known examples of this mechanism are the Tn5 and Tn10 elements of the IS4 family (York & Reznikoff 1996; Haniford & Ellis 2015). Another example is likely to be bacteriophage Mu as the transposase is known to be monomeric in solution (Kuo et al. 1991), though as there up to six transposase proteins in the complex, the dynamics of the reaction may be more complicated.

The alternative pathway, S-NEC (Figure 11b), initiates with the transposases forming multimers in solution. This multimer will then bind a single TIR, to form the S-NEC version of a SEC. Instead of the second TIR being bound by other proteins, it is recruited into the SEC to form the full PEC that can proceed with the reaction. This pathway is used by the members of the *mariner* family, where the transposases exist as dimers in solution (Claeys Bouuaert et al. 2017). The other transposase that is known to use this pathway is Hermes, which forms an octameric ring (composed of four dimer subunits) in solution, though it is predicted that only a single dimer is responsible for binding the TIRs and the transposition reaction steps (Hickman et al. 2014)

To date only a relatively small number of transposition systems have been characterised well enough to confirm the assembly pathway, so the two pathways may not be the only ones possible. The assembly pathways

for those systems that have multiple binding sites within the TIR, such as the Tc1 elements like Sleeping Beauty, would be of particular interest.

1.7.1 Hypothetical model of complex formation pathway

In the relatively limited systems thus far characterised, none have been identified in eukaryotes to operate using the S-PD pathway. In prokaryotes however, both assembly pathways have been identified, suggesting a potential reason for the specificity of one system and the universal nature of the other. Our group has developed a model to explain why these differences might exist.

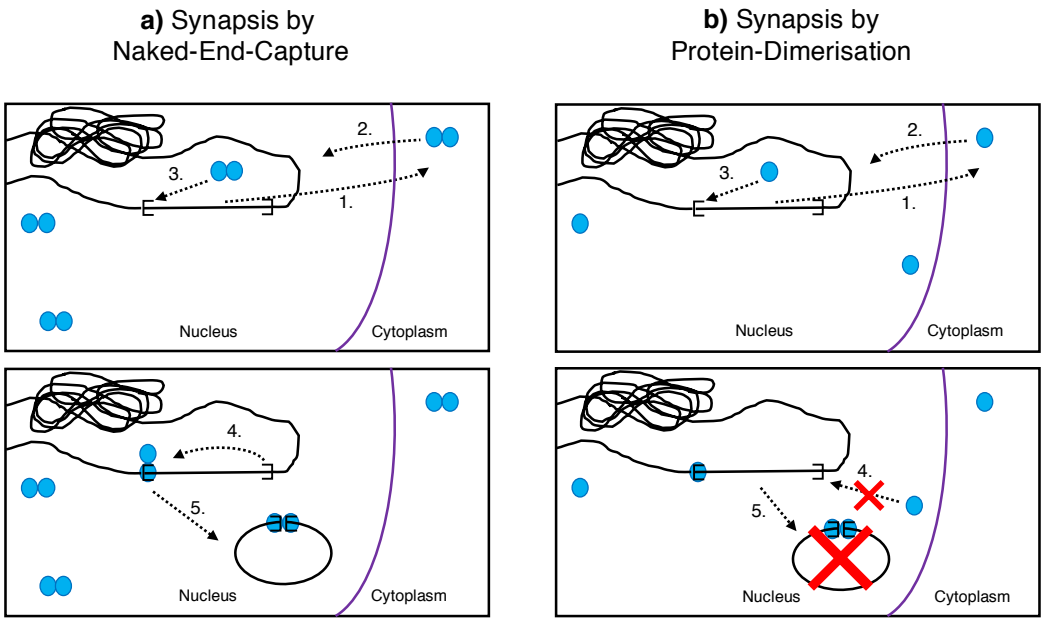


Figure 12: Model of Eukaryotic Transposition Through the Different Complex Assembly Pathways. The DNA is shown as black lines, and the transposon TIR appear as black brackets. Transposase protein is shown as blue ovals, and the nuclear membrane is the purple curved line. **a)** The model of how Synapsis by Naked-End-Capture (S-NEC) may function in eukaryotes. The transposase is translated in the cytoplasm, where it will form dimeric complex in solution (step 1), and will be transported into the nucleus (step 2). A dimeric transposase will bind one TIR (step 3). The other TIR is recruited into the SEC (step 4), which will form the PEC, and transposition can proceed (step 5). **b)** the model of how Synapsis by Protein-Dimerisation may function in eukaryotes. The transposase is translated in the cytoplasm (step 1), and is then transported back into the nucleus (step 2). Monomeric transposase may bind the TIR (step 3), however the frequency of the second end being bound at the same time (step 4) will be very low. Therefore, the PEC will be unable to form except at a very low frequency (step 5).

In prokaryotes, the transposase is transcribed and translated within a very small distance from the transposon itself, which generates a high local concentration of the transposase. For the S-PD based systems, this is advantageous as, in order for transposition to occur, each TIR has to be bound by a transposase independently, and these SECs can be transient complexes. Thus, the high local concentration increases the likelihood of both ends being bound, and therefore dramatically increases the frequency of transposition. For transposition systems that use S-NEC the high local concentration is not as beneficial. There is an increased chance of the ends being bound by the multi-transposase complexes; however, this comes at a cost. If both ends are bound simultaneously by separate multi-transposase complexes, the PEC cannot form, and therefore the reaction cannot proceed. This is known as Over-Production-Inhibition (OPI), which leads to a low concentration of the transposase to have high activity, but higher concentrations show greatly decreased activity. It is particularly prevalent in *mariner* elements, where it was first identified (Lohe & Hartl 1996a).

In eukaryotes, the transposase is translated outside of the nucleus and is subsequently transported back in. This means there is no high local concentration of the transposase near the transposon ends, just a low concentration throughout the genome. For S-NEC this is not a problem, as, although it may be infrequent that it forms a SEC, once it has, the free TIR has a relatively high chance of interacting which will lead to the PEC formation (Figure 12a). By contrast, the S-PD based systems would not function well (Figure 12b). The odds of a single end being bound would not

be any worse than for S-NEC-based systems, but the second end has to be bound in the exact same way, and both complexes have to come together to form the PEC. This would all have to occur before one of the transposases became unbound again. The chances of this sequence of events happening would be very low, so it is possible that, if any of the S-PD-based transposition systems got into eukaryotes, the rate of transposition would be so low that they would be removed through the process of genetic drift relatively quickly. Eukaryotes also generally have larger genomes than prokaryotes, which decreases the frequency of randomly locating the binding sites within the genome.

1.8 Research aims

The model in section 1.7.1 provides a potential explanation for why S-PD-based transposition systems do not function in eukaryotes. While the model fits with the data already available, no experimental work has been done to test it directly. The primary aim of this research is to test this model, through manipulation of the transposase complex-assembly dynamics. The possible routes to test this model are either to alter an S-NEC-based transposase and try and see which modifications prevent activity in only eukaryotes, or to take an S-PD-based transposase and manipulate it to allow function in eukaryotes. The latter approach is the one taken, as it provides much clearer

results and could also open up possibilities for further experimental work with transposases in eukaryotes.

In order to test the model, a candidate transposase was needed. This would have to operate through the S-PD mechanism, and needed to be well characterised. There are three well characterised systems that operate through S-PD, MuA, Tn10 and Tn5. MuA has at least four transposase monomers in its final complex, and also makes use of the copy-in transposition pathway, which is complicated to analyse. Tn10 provides a much simpler system; only two transposases are needed in the final complex, and its mechanism of action is the forward-hairpin cut-and-paste, which is easier to characterise. Despite these advantages, Tn10 transposition has an additional host factor, IHF, which is involved in the complex formation, and the lack of a Tn10 transposase crystal structure makes manipulation more difficult. The last option, Tn5, represents the best candidate transposase. It has been well characterised, through its crystal structure, as well as multiple biochemical studies to identify the various domains and mechanisms involved. It also does not require IHF as a host factor in order to transpose efficiently, which reduces the complexity of the system that is to be manipulated. Prior work has suggested that Tn5 transposition is not functional in higher eukaryotes such as humans (Reznikoff 2013, pers comm); however, when pre-formed complexes were supplied to yeast cells, some limited transposition was detected (Goryshin et al. 2000).

As mentioned before, the primary aim of this project was to test the model in section 1.7.1. There are two possible approaches that can be used to do this.

The first approach is to try modifying the assembly pathway entirely. If the Tn5 transposase can be modified so that it mimics S-NEC, then based on the model, it should be capable of transposition in a eukaryotic genome. Chapter 3 explores this idea, through the construction of a single-chain dimer of the Tn5 transposase. This attaches two Tn5 monomers together, so they mimic the S-NEC assembly dynamics, although they do not actually dimerise prior to DNA binding.

The other approach to test this model is through strengthening the binding of the SEC. Based on the model, the SEC would occur for S-PD based transposases, but would not endure long enough for the other end to be bound as well. Chapter 4 explores this idea of strengthening the SEC, through the use of a zinc finger-transposase fusion.

While testing the model was the primary aim, a secondary objective of this project was to identify what other factors might be inhibiting Tn5 transposition in eukaryotes. One possibility was that chromatin could be a leading inhibitor, as the Tn5 transposase is known to be unable to bind to it (Sos et al. 2016). Chapter 5 explores this, by fusing the transposase with a predicted histone binding domain from *piggyBac*, the Plant-HomeoDomain (PHD) motif.

A small side-project was also performed, alongside the main body of work shown here. This project focused on a recently discovered human-

derived *piggyBac* element, PGBD5, that had been domesticated by its host ~500 Mya, though it still retains the capacity to transpose (Henssen et al. 2015). Chapter 6 is based around repeating and expanding the work from the original paper.

Chapter 7 summarises the work done in this thesis as well as suggesting possible future directions for the research.

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Chemicals and reagents

Chemicals were supplied by Sigma-Aldrich unless otherwise stated. All enzymes for Polymerase Chain Reaction (PCR) and plasmid construction were supplied by New England Biolabs (NEB). DNA purification kits: QIAquick Miniprep Kit, QIAquick PCR purification kit and QIAquick gel extraction kit were all supplied by Qiagen™. The human genomic extractions were done using a Sigma-Aldrich GenElute™ Mammalian Genomic DNA Miniprep Kit.

2.1.2 Strains and cell lines

2.1.2.1 *Escherichia coli* strains

Strain name	Genotype
NEB 5-alpha	*fhuA2 (argF-lacZ)U169 phoA glnV44 80 (lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17.* Sourced from NEB.
RC5045	*fhuA2 lacZ::T7 gene1 [lon] ompT gal sulA11 R(mcr-73::miniTn10—TetS)2 [dcm] R(zgb-210::Tn10—TetS) endA1 Δ(mcrC-mrr)114::IS10.* Also known as ER2566, sourced from NEB.

RC5095	*F- fhuA2 Δ(lacZ)r1 glnV44 e14-(McrA-) trp-31 his-1 rpsL104 xyl-7 mtl-2 metB1 Δ(mcrC-mrr)114::IS10.* Also known as ER1793, sourced from NEB.
RC5123	*F- fhuA2 Δ(lacZ)r1 glnV44 e14-(McrA-) trp-31 his-1 rpsL104 xyl-7 mtl-2 metB1 Δ(mcrC-mrr)114::IS10 argE::Tn5-lacZ'-kanR.* Developed from RC5095.

2.1.2.2 Human cell lines:

All human cell work was done with either the HeLa cell line or the HEK293 cell line, both of which were kindly donated by Professor Peter Shaw. All experiments were conducted within a 10-passage limit to ensure the reproducibility of the experiments.

2.1.3 Plasmids and oligonucleotides

2.1.3.1 Plasmid list

Plasmid name	Plasmid description	Relevant Sections
pBS	pBlueScript SKII+, sourced from Stratagene. Expression vector with a T7 promoter	3.2.2, 5.2, 6.2

	controlled by a LacO site. Contains ampicillin resistance cassette.	
pMAL-C2X	pMAL-C2X expression vector, sourced from NEB. Plasmid designed for protein expression, has a T7 promoter controlled by two LacO sites. Also contains the sequence for a maltose binding protein tag, which is removed when adding proteins in using NdeI site.	3.2.1.1, 5.2.1
pTYB2	pTYB2 expression vector, sourced from NEB. Plasmid designed for protein expression, has a T7 promoter controlled by a LacO site. Also contains the sequence for an intein tag, which is located on the C-terminus of any genes added to the plasmid.	3.2.1, 3.2.3, 4.2.1, 5.2.1
pCDNA3.1+	pCDNA3.1+ human expression vector, sourced from Invitrogen. Contains a cytomegalovirus promoter for human expression, upstream of location for gene sequence insertion.	3.2.2, 5.2.2, 6.2

pRecLV103-GFP-PGBD5	Mammalian expression vector containing the sequence coding for the GFP-PGBD5 fusion. A gift from Alex Kentsis (Addgene plasmid # 65409)	6.2
pRC415	Plasmid with the sequence coding for kanamycin resistance between two Tn5 TIRs. At each end of the TIRs are PvuII restriction enzyme sites, so the entire 1221 bp transposon can be excised. Made by Nicolas Buisine, formerly part of the Chalmers laboratory.	3.2.4
pRC1240	*Hsmar1* transposon reporter plasmid, sourced from Miskey et al. 2007, original name is pHsmar1-Neo. The plasmid contains the sequence coding for Neomycin resistance, controlled by an SV40 promoter. This is located between two TIR for the Hsmar1 transposon. The total transposon is 1807 bp.	5.2.2
pRC1241	Sequence coding for the Hsmar1 transposase monomer, in pCDNA3.1+, between the EcoRI and NotI sites. Made by	5.2.2

	Danzu Liu, formerly part of the Chalmers laboratory.	
pRC1542	Sequence coding for mouse codon-optimised *piggyBac,* in pCDNA3.1+, between the EcoRI and NotI sites. Made by Ronald Chalmers.	5.2.2, 6.2
pRC1551	Reporter plasmid for *piggyBac*. Made by Ronald Chalmers. The plasmid contains the sequence coding for Neomycin resistance gene with a PGK promoter. *piggyBac* TIR are located shortly outside of the resistance gene cassette, making a total transposon size of 3231 bp.	6.2
pRC1568	The sequence coding for the *piggyBac* transposase, codon optimised for expression in bacteria (made by Geneart™).	5.2.1
pRC2101	The sequence coding for the Tn5 transposase large dimer, inserted into the pTYB2 vector, between NruI and XmaI. Has intein tag attached to transposase large dimer C-terminus. See section 3.2.1 for details.	3.2.1

pRC2102	The sequence coding for the Tn5 transposase small dimer, inserted into the pTYB2 vector, between NruI and XmaI. Has intein tag attached to the transposase small dimer C-terminus. See section 3.2.1 for details.	3.2.1, 3.2.3
pRC2106	Reporter construct with Tn5 TIR, surrounding the sequence for Neomycin resistance. Made through inverse PCR on pRC1240. Transposon total size is 1767 bp.	3.2.1.1, 4.2.1, 5.2.1
pRC2107	pRC2102 with the sequence coding for the Tn5 transposase small dimer. The second Tn5 transposase monomer is mutated with the D97A mutation.	3.2.3.2
pRC2109	The sequence coding for the Tn5 transposase small dimer inserted into the pMAL-C2X vector, between NdeI and SbfI.	3.2.1.1, 4.2.1, 5.2.1
pRC2110	pMAL-C2X vector with the sequence coding for chloramphenicol resistance inserted at the SwaI site.	3.2.1.1

pRC2113	The sequence coding for the Tn5 transposase small dimer, inserted into pRC2110, between NdeI and SbfI.	3.2.1.1
pRC2114	The sequence coding for the Tn5 transposase monomer, inserted into pRC2110, inserted between NdeI and SbfI.	3.2.1.1
pRC2116	The sequence coding for GFP, inserted into pQE trisystem™ His-Strep 1 at the NcoI site.	6.2
pRC2117	The sequence coding for the Tn5 transposase monomer, inserted into pMAL-C2X between NdeI and SbfI.	3.2.1.1, 4.2.1, 5.2.1
pRC2122	The sequence coding for the Tn5 transposase large dimer, codon optimised for expression in human cells, in pMK-RQ vector between NcoI and KpnI.	3.2.2
pRC2126	The sequence coding for the Tn5 transposase large dimer, inserted into the pMAL-C2X vector between NdeI and SbfI.	3.2.1.1
pRC2128	The sequence coding for the Tn5 transposase monomer, inserted into the pTYB2 vector between NdeI and EcoRI.	3.2.1.2, 3.2.3, 4.2, 5.2.1, 8.1

pRC2143	pRC2102 with the sequence coding for the Tn5 transposase small dimer. The first Tn5 transposase monomer is mutated with the D97A mutation.	3.2.3.2
pRC2146	The codon-optimised sequence for the Tn5 transposon large dimer, inserted into the pCDNA3.1+ vector between NheI and XbaI.	3.2.2
pRC2147	The codon-optimised sequence for the Tn5 transposase small dimer, inserted into the pCDNA3.1+ vector between NheI and XbaI sites.	3.2.2
pRC2148	The codon-optimised Tn5 transposase monomer, inserted into the pCDNA3.1+ vector between NheI and XbaI sites.	3.2.2
pRC2149	pRC2102, with the sequences of both Tn5 transposase monomers of the small dimer mutated with D97A.	3.2.3.2
pRC2152	pRC2102 with the sequence coding for the Tn5 transposase small dimer. The first Tn5 transposase monomer is mutated with the G462D mutation.	3.2.3.3

pRC2153	pRC2102 with the sequence coding for the Tn5 transposase small dimer. The second Tn5 transposase monomer is mutated with the G462D mutation.	3.2.3.3
pRC2154	pRC2128, where the sequence coding for the Tn5 transposase monomer is mutated with D97A.	3.2.3.2
pRC2162	pRC2146 with the sequence coding for GFP inserted at the HpaI site, to express GFP tagged Tn5 transposase large dimer.	3.2.2
pRC2163	pRC2147 with the sequence coding for GFP inserted at the HpaI site, to express GFP tagged Tn5 transposase small dimer.	3.2.2
pRC2164	pRC2148 with the sequence coding for GFP inserted at the HpaI site, to express GFP tagged Tn5 transposase monomer.	3.2.2
pRC2165	pRC2128 where the sequence coding for the Tn5 transposase monomer is mutated with G462D.	3.2.3.3
pRC2166	pRC2102 with the sequence coding for the Tn5 transposase small dimer. Both the Tn5	3.2.3.3

	transposase monomer are mutated with the G462D mutation.	
pRC2169	pRC2106 with zinc-finger binding sites added 1 bp inside of each TIR.	4.2.1, 4.2.2
pRC2170	The sequence coding for the zinc-finger array fused to the Tn5 transposase monomer sequence. The zinc finger array sequence is inserted into pRC2102 between NdeI and SpeI.	4.2
pRC2171	The sequence coding for the zinc-finger array fused to the Tn5 transposase small dimer sequence. The zinc finger array sequence is inserted into pRC2102 at the NdeI site.	4.2
pRC2172	pRC2128 with the sequence coding for the Tn5 transposase monomer mutated with K160A.	8.1
pRC2173	pRC2128 with the sequence coding for the Tn5 transposase monomer mutated with H194A.	8.1

pRC2174	pRC2128 with the sequence coding for the Tn5 transposase monomer mutated with K212A.	8.1
pRC2175	pRC2128 with the sequence coding for the Tn5 transposase monomer mutated with K249A.	8.1
pRC2179	pRC2106 with zinc-finger binding sites added 2 bp inside of each TIR.	4.2.2
pRC2180	pRC2106 with zinc-finger binding sites added 3 bp inside of each TIR.	4.2.2
pRC2181	pRC2106 with zinc-finger binding sites added 4 bp inside of each TIR.	4.2.2
pRC2182	pRC2106 with zinc-finger binding sites added 5 bp inside of each TIR.	4.2.2
pRC2187	pRC2170 with sequence for an additional 10 aa inserted at the start of the linker region.	4.2
pRC2192	The sequence coding for the zinc-finger array, inserted into the pTYB2 vector between NdeI and EcoRI.	4.2
pRC2194	Reporter construct for human cell work, with Tn5 TIRs surrounding the sequence coding	3.2.2, 5.2.2

	for neomycin resistance, and a ColE1 origin of replication. A Thymidine Kinase-Hygromycin fusion cassette (sourced from the Rob Klose lab) was added outside the transposon section at the PciI site. Total transposon region is 2589 bp.	
pRC2195	pRC2128 with the sequence coding for the Tn5 transposase monomer, mutated with K160D.	3.2.1.2, 8.1
pRC2196	pRC2128 with the sequence coding for the Tn5 transposase monomer, mutated with K160E.	8.1
pRC2199	pRC2128 with the sequence coding for a nuclear localisation signal, added to the N-terminus of the Tn5 transposase monomer.	3.2.4
pRC2200	pRC2102 with the sequence coding for a nuclear localisation signal, added to the N-terminus of the Tn5 transposase small dimer.	3.2.4
pRC2501	pRC1551 with a point mutation made (G to A at base 4002) in the *piggyBac* 3' TIR, to	5.2.2

	make the sequence of the TIRs identical to that used in Henssen et al. 2015.	
pRC2506	The sequence coding for the 82 aa PHD domain, fused to the C-terminus of the Tn5 transposase monomer. Inserted from pRC1568 into pRC2102 at the EcoRI and XmaI sites.	5.2.1
pRC2507	The sequence coding for the 90 aa PHD domain fused to the C-terminus of the Tn5 transposase monomer. Inserted from pRC1568 into pRC2102 at the EcoRI and XmaI sites.	5.2.1
pRC2508	The sequence coding for the 82 aa PHD domain fused to the C-terminus of the Tn5 transposase monomer from pRC2148, codon-optimised for human expression. Inserted from pRC1542 between the BamHI and XhoI sites.	5.2.2
pRC2509	The sequence coding for the 90 aa PHD domain fused to the C-terminus of the Tn5 transposase monomer from pRC2148, codon-optimised for human expression.	5.2.2

	Inserted from pRC1542 between the BamHI and XhoI sites.	
pRC2510	The sequence coding for the 82 aa PHD domain fused to the C-terminus of the Hsmar1 transposase monomer in pRC1241, codon-optimised for human expression. Inserted from pRC1542 at the NheI site.	5.2.2
pRC2511	The sequence coding for the 90 aa PHD domain fused to the C-terminus of the Hsmar1 monomer in pRC1241, codon-optimised for human expression. Inserted from pRC1542 at the NheI site.	5.2.2

2.1.3.2 Oligonucleotide list

Oligo name	Oligo sequence	Oligo description	Relevant sections
Pr129	CGAAGAGTAACCGTTGCTAGGAGA GACCGTGGCTGAATGAGACTGGT GTCGACACTA	Top strand of the splinkerette adapter.	2.4.3.8, 3.2.2

Pr130	CGCCACTAGTGTCGACACCAGTCTCTAATTTTTTTTTTCAAAAAAA	Bottom strand of the splinkerette adapter.	2.4.3.8, 3.2.2
Pr131	GGTACCAGTCTAAAATAAACATAAAATAAACAGC	Splinkerette reverse PCR primer for the first round of PCR.	2.4.3.8, 3.2.2
Pr132	CGAAGAGTAACCGTTGCTAGG	Splinkerette forward PCR primer for the first round of PCR.	2.4.3.8, 3.2.2
Pr134	GCAAAAAAACGTAAAGCGAGAAATGCGC	Splinkerette reverse PCR primer for the second round of PCR.	2.4.3.8, 3.2.2
Pr135	GTGGCTGAATGAGACTGGTGTCGAC	Splinkerette forward PCR primer for the second round of PCR.	2.4.3.8, 3.2.2
Pr177	CTGTCTCTTATACACATCTTGCATTGTTGG	Forward primer to amplify the Tn5 transposon.	2.4.3.9, 3.2.2
Pr178	CTGTCTCTTATACACATCTGGACCGTG	Reverse primer to amplify the Tn5 transposon.	2.4.3.9, 3.2.2

Pr206	(cy5)TCAACGCGTGGGCGAGCTTAGGGTATACCCTAAAGGTGATGCTTAGTG	Labelled top strand of oligo with zinc-finger binding site (ZBS).	2.4.4.4, 4.3.3
Pr207	CACTAAGCATCACCTTTAGGGTATACCCTAAGCTCGCCCACGCGTTGA	unlabelled bottom strand of oligo with ZBS.	2.4.4.4, 4.3.3
Pr240	(cy5)CACGGGGTGATGCTCAGATGTGTATAAGAGACAGAAGCTTGGCGGATCGC	Labelled top strand of oligo with ZBS and Tn5 TIR.	2.4.4.4, 4.3.3
Pr241	GCGATCCGCCAAGCTTCTGTCTCTTATACACATCTGAGCATCACCCCGTG	unlabelled bottom strand of oligo with ZBS and Tn5 TIR.	2.4.4.4, 4.3.3
Pr242	(cy5)CGCAATTCACGGTCCAGATGTGTATAAGAGACAGAAGCTTGGCGGATCGC	Labelled top strand of oligo with Tn5 TIR.	2.4.4.4, 3.2.3.1, 3.2.3.4, 4.3.3
Pr243	GCGATCCGCCAAGCTTCTGTCTCTTATACACATCTGGACCGTGAATTGCG	Unlabelled bottom strand of oligo with Tn5 TIR.	2.4.4.4, 3.2.3.1, 3.2.3.4, 4.3.3

2.4 Methods

2.4.1 Cloning and DNA manipulation

2.4.1.1 Polymerase Chain Reaction

All reactions were done with Q5™ DNA polymerase, unless otherwise stated, with 1x Q5 reaction buffer, 200 nM of each primer, 200 µM dNTPs, 10 ng of template, and 1 unit of enzyme per 50 µl. Standard reaction conditions were: initial denaturation for 5 minutes at 95 ^{o}C, denaturation for 30 seconds at 95 ^{o}C, annealing for 30 seconds at the temperature calculated from the NEB website, extension at 72 ^{o}C for 1 kb per 30 seconds, for 30-35 cycles, and a final extension step at 72 ^{o}C for 5 minutes. Products were separated by agarose gel electrophoresis (see section 2.4.1.2), then purified.

2.4.1.2 Gel electrophoresis

Gel electrophoresis was done using a BioRad gel electrophoresis system. All gels were 0.5-2% agarose content, made with 0.5x Tris-Borate-EDTA (TBE) buffer. The gels were run at 10 V/cm. For accurate analysis of the

DNA samples, 1 kb DNA ladder or 100 bp DNA ladder were used. Both were supplied from NEB.

For polyacrylamide gels, gels were made using 1x TBE buffer, 8% polyacrylamide, 0.01% APS and 0.001% TEMED. Gels were run at 8 V/cm, in a BioRad PROTEAN-II tank.

2.4.1.3 Plasmid construction

Restriction digests were run with the various restriction enzymes, as recommended by the manufacturer, NEB. DNA fragments were purified by agarose gel electrophoresis (see section 2.4.1.2). The vector backbone fragment was dephosphorylated using Calf-Intestinal Phosphatase (CIP). Ligations were done overnight at 16 ^{o}C using T4 DNA ligase, following the manufacturer's guidelines. The ligations were then transformed into chemically competent NEB 5-alpha cells using the manufacturer's heat-shock protocol. The various plasmids were confirmed through analytical restriction digests and sequencing (Source Bioscience).

2.4.1.4 Chemically competent cells

To make the cells competent, the frozen cell line was first streaked out on 1% agar LB plates and grown overnight at 37 ^{o}C. A single colony from the plate was used to inoculate a 5 ml LB culture which was grown overnight at

37 ^{o}C rotating gently. This 5 ml was then added to 500 ml LB and grown at 37 ^{o}C, shaking at 160 rpm. Once the OD (optical density) 600 reached 0.6, the cells were centrifuged at 3000 xg, 4 ^{o}C, 20 minutes. The cells were resuspended in 167 ml TFB1 buffer (30 mM potassium acetate, 50 mM $MgCl_2$, 100 mM KCl_2, 10 mM $CaCl_2$, 15% glycerol), and put on ice for 1 hour. The cells were then centrifuged again, before being resuspended in 40 ml TFB2 buffer (10 mM MOPS pH 7.0, 75 mM $NaCl_2$, KCl_2, 15% glycerol). 100 μl aliquots were frozen down at -80 ^{o}C for later use.

2.4.1.5 Radioactive substrate purification

Radioactively labelled substrates were prepared by digesting pRC916 with either BamHI and XbaI for the 'short' DNA substrate (93 bp), or KpnI and NotI for the 'long' DNA substrate (154 bp). They were labelled at each end with [α-P^{32}] dCTP using the Klenow enzyme (sourced from NEB), according to manufacturer's instructions. To purify the substrates, they were run on an 8% Tris-Borate-EDTA (TBE) polyacrylamide gel (see section 2.4.1.2) and isolated using the 'Crush and Soak' method (Z. Chen & Ruffner 1996).

2.4.1.6 Fluorescent oligonucleotide substrate purification

The fluorescent substrates are derived from oligonucleotides. Each oligonucleotide pair was mixed, 5 nM labelled oligonucleotide and 6 nM unlabelled oligonucleotide, in TE (10 mM Tris-HCl, pH 7.5, 1 mM Na2-EDTA,

pH 8.0), and heated to 95 ^{o}C for five minutes, then slowly cooled. The substrate was run on an 8% (TBE) polyacrylamide gel (2.4.1.2), and isolated using the 'Crush and Soak' method (Z. Chen & Ruffner 1996).

2.4.2 Bacterial assays

2.4.2.1 Papillation assay

For plasmids containing the gene for ampicillin resistance, 100 µl of competent RC5123 (see section 2.4.1.4) was transformed with 1 ng of plasmid DNA using a heat shock at 42 ^{o}C for 45 seconds, before addition of LB (Lysogeny broth) up to 1 ml. After one hour of incubation at 37 ^{o}C, ampicillin was added to a final concentration of 100 µg/ml and the cells were left for a further four hours at 37 ^{o}C. Cells were then serially diluted 10^{-4}, 100 µl was then plated out on agar plates containing 100 µg/ml ampicillin, 40 µg/ml X-gal, as well as varying quantities of lactose and IPTG (Isopropyl β-D-1-thiogalactopyranoside). For plasmids with the chloramphenicol resistance gene, the same protocol was followed, except that after the 1 hour incubation at 37 ^{o}C, 100 µl of cells were plated out directly onto agar plates containing 25 µg/ml chloramphenicol, 40 µg/ml X-gal, as well as varying quantities of lactose and IPTG. All plates were incubated for 5 days at 37 ^{o}C, then full plates and the individual colonies were imaged.

2.4.2.2 *E. coli* pre-formed complex integration assay

To produce the pre-cleaved transposon substrate, pRC415 was digested with PvuII, leaving blunt ends at the cleavage site. The transposon DNA was then purified using agarose gel electrophoresis. 2.5 nM (40 ng) was mixed with the transposase (at various concentrations) in transposition reaction buffer (100 mM potassium glutamate, 0.5 µg/ml BSA, 25 mM Tris.HCl pH 7.5, 0.5 µM β-mercaptoethanol) in a total of 20 µl, before incubation at 37 ^{o}C for one hour. All 20 µl was then transformed into RC5095 using standard heat shock protocols. The cells (100 µl) were then plated out on LB-agar plates containing 50 µg/ml Kanamycin, and left overnight at 37 ^{o}C, after which the colonies per plate were quantified.

2.4.3 Human cell culture and assays

2.4.3.1 Cell maintenance

Both HeLa and HEK293 cells were cultured using low glucose Dulbeccos Modified Eagles Media (DMEM, Sigma™). This was supplemented with 10% heat inactivated Foetal Bovine Serum (FBS), penicillin (100 units per ml) and streptomycin (100 µg/ml). The cells were grown in a humidified environment at 37 ^{o}C, with 5% (HeLa) or 10% (HEK293) CO_2.

Passaging of the cells lines was performed at ~90% confluency, using trypsin/EDTA (Ethylenediaminetetraacetic acid) treatment. ~10% of the cells were seeded into a new flask and left to grow. Stocks of the cells were stored in vapour phase of liquid nitrogen, in freezing media (50% DMEM, 40% FBS, 10% DMSO (Dimethyl sulfoxide)).

2.4.3.2 Cell transfection

Cells were seeded into 6-well plates, and grown to 70-80% confluence. 3 ml media was changed to DMEM + 10% FBS, with no penicillin or streptomycin. Up to 2 µg of plasmid DNA was transfected in using Lipofectamine® 2000 (Invitrogen), according to the manufacturer's instructions.

2.4.3.3 Cell lysis and protein extraction

The cells were transfected with 1 µg of protein expression plasmid, as described in section 2.4.3.2. 24 hours after transfection, the cells on the 6-well plates were taken for protein extraction. The cells were put on ice for 5 minutes, then washed twice with 1.5 ml PBS per well, before being resuspended in 750 µl PBS. These samples were then centrifuged at 3000 x g, 4 ^{o}C, for 5 minutes, to pellet the cells. Each pellet was resuspended in 100 µl of Radioimmunoprecipitation assay (RIPA) buffer (50 mM Tris pH 7.5, 150 mM NaCl, 1 mM Na2-EDTA, 1% Triton X-100, 0.5% sodium deoxycholate and 0.1% SDS) and 40 µl protease inhibitor cocktail (Roche

Applied Science). These were incubated on ice for 30 minutes and vortexed every 10 minutes, before centrifugation at 15,000 x g, 4 ^{o}C, for 15 minutes. The supernatant was removed, quantified using a Bradford assay (Bradford 1976), and then aliquoted and frozen at -80 ^{o}C.

2.4.3.4 Fluorescence microscopy

The cells were transfected with 1 µg of expression plasmid, as described in section 2.4.3.2. After 24 hours, the live cells were imaged, in the 6-well plate, using a Carl Zeiss Axiovert S100 TV Inverted Microscope with an HBO 100 illuminator.

2.4.3.5 Transposition integration assay

The cells were transfected with 1 µg of a transposon reporter plasmid, along with 1 ug of the transposase expression plasmid, following the protocol described in section 2.4.3.2. 24 hours after transfection, the cells were treated with trypsin/EDTA to resuspend then, and then counted using a Hawklsey Neubauer counting chamber. A known quantity of cells (stated for each experiment) was seeded out onto 100 mm^2 plates, with 10 ml media that was supplemented with G418 (800 µg/ml for HeLa, 2 mg/ml for HEK293). This media was changed every three days for two weeks to allow selection for neomycin resistance. The cells were then fixed to the plates, using 10 ml 10% formaldehyde in PBS, for 15 minutes. 10 ml methylene blue

buffer (1% methylene blue, 70% ethanol) was then used to stain the cells on the plates for 30 minutes, before they were washed with water, air-dried and imaged.

2.4.3.6 Human cell pre-formed complex integration assay

To generate the pre-cleaved transposon, PCR was done on pRC2106 using pr177 and pr178, amplifying the transposon with blunt ends. The PCR product was then purified using agarose gel electrophoresis. This was incubated with 80 nM monomeric transposase or 40 nM dimeric transposase, for one hour, in the transposition reaction buffer (100 mM potassium glutamate, 0.5 µg/ml BSA, 25 mM Tris.HCl pH 7.5, 0.5 µM β-mercaptoethanol) in a total volume of 20 µl. Wells of a 6-well plate that were prepared at 70-80% confluency were transfected with the entire 20 µl, using Lipofectamine® 2000, as was previously outlined for Mos1 (Trubitsyna et al. 2017). The cells were then left for 24 hours, before $4x10^5$ cells were seeded onto a 100 mm^2 dish, in standard media supplemented with G418 at 800 µg/ml. The media was changed every three days for 14 days, at which point the cells were fixed and stained, as described in section 2.4.3.5.

2.4.3.7 Genomic DNA extraction from human cells

The transposition integration assay was performed, as in section 2.4.3.4, the cells were then grown for an additional three days to ensure a sufficient

quantity of cells. The genomic DNA was extracted using the GenElute™ mammalian genomic DNA miniprep Kit (Sigma-Aldrich), according to the manufacturer's protocols, and stored in TE buffer at 4 ^{o}C.

2.4.3.8 Splinkerette PCR

This protocol was adapted from published versions (Uren et al. 2009). 2 µg of genomic DNA was digested with TaqαI overnight at 65 ^{o}C, in a 30 µl volume, then heat-inactivated at 80 ^{o}C for 20 minutes. 1.25 µM splinkerette adapters (annealed Pr129 and Pr130 oligos) were ligated to 300 ng genomic DNA at 4 ^{o}C overnight, in 40 µl total volume. This ligation was heat-inactivated at 65 ^{o}C for 20 minutes. The entire 40 µl was used in a 100 µl total volume restriction digest, using BsmI overnight at 65 ^{o}C. The digestion reaction was heat-inactivated at 80 ^{o}C for 20 minutes before the DNA was purified and eluted in 30 µl EB buffer. A PCR reaction was then run with 5 µl DNA, using primers Pr131 and Pr132, with the Q5 polymerase, at 63 ^{o}C annealing temperature, and a 1 minute extension time. The second PCR reaction was run directly after, using 1 µl of the first PCR reaction as template. This reaction was done with Taq polymerase, using Pr134 and Pr135, with the following reaction steps: 95 ^{o}C for 30 seconds, then 25 cycles of (95 ^{o}C for 30 seconds, 61^{o}C for 60 seconds, 68 ^{o}C for 3 minutes), and finally 68 ^{o}C for 5 minutes. The DNA was then purified and eluted in 30 µl EB buffer.

1 µg vector pBS was digested with EcoRV at 37 °C, then purified and eluted in 30 µl EB buffer. The purified pBS was treated with Terminal Transferase, in a 50 µl reaction volume at 37 °C for one hour, with 1 mM dideoxy-thymidine triphosphate to create single thymidine overhangs at each end. The pBS was then purified, eluted in 30 µl EB buffer. A ligation was set up in 50 µl total volume, using a 3:1 molar ratio of PCR product to pBS backbone, overnight at 16 °C. This was then heat-inactivated at 65 °C for 20 minutes. The ligated product was then transformed into NEB 5-alpha using the manufacturers heat-shock transformation protocol, and plated on 100 µg/ml ampicillin agar plates overnight at 37 °C. Overnight cultures were grown at 37 °C from the colonies, from which the plasmids were isolated, and samples sequenced.

2.4.3.9 Inverse PCR mapping

2 µg of genomic DNA was digested overnight with HindIII-HF, at 37 °C, before heat inactivation, 80 °C for 20 minutes. The DNA was purified and eluted in 30 µl EB buffer. 100 ng of the DNA was then self-ligated overnight at 16 °C in a 200 µl reaction, before ethanol precipitation, and the pellet was resuspended in 30 µl TE buffer. 15 µl (50 ng) was used as the template DNA for a PCR reaction with pr276 and pr277, using an annealing temperature of 67 °C and an extension time of 2 minutes. The product of the PCR reaction was then purified and eluted in 30 µl EB buffer, before 100 ng was self-ligated overnight at 16 °C in a 50 µl reaction. This was then transformed into

NEB 5-alpha, using the standard heat shock protocol, and plated out on LB-agar plates with 50 µg/ml Kanamycin. After the plates were incubated at 37 ^{o}C overnight, plasmid DNA was isolated from individual colonies and then sequenced to identify the integration events.

2.4.4 Protein purification and *in vitro* assays

2.4.4.1 Protein purification

The expression plasmids (variants of pTYB2) were transformed into RC5045 (ER2566, NEB). From this a 20 ml culture was set up and grown at 37 ^{o}C for 16 hours, before being used to inoculate 2 litres of LB supplemented with 100 µg/ml ampicillin. This was grown to OD 600 of 0.4, at 37 ^{o}C, before the addition of 0.5 mM IPTG. The culture was then put at 23 ^{o}C for four hours to overexpress the protein, before being centrifuged, 5000 x g, 4 ^{o}C for 20 minutes. The pelleted cells were then resuspended in 40 ml column buffer (20 mM Tris pH 7.5, 700 mM NaCl, 0.1% Triton™X-100, 10% glycerol) and frozen at -80 ^{o}C. The cells were thawed out and a protease inhibitor cocktail tablet (Roche Applied Science) was added. The cells were lysed using a French Pressure cell (American Instrument Co., Inc., Silver Spring, Md.), each run at a minimum of 10,000 Psi, done three times per sample. The lysate was then centrifuged at 12,000 x g, 4 ^{o}C, for 1 hour. The supernatant was diluted to 100 ml total volume using column buffer, and then run through

a 2 ml bed volume of chitin (NEB) resin. The resin was then washed with 30 ml column buffer, followed by 6 ml cleavage buffer (column buffer + 50 mM dithiothrietol (DTT)). The column was then capped and loaded with 1 ml cleavage buffer, before being stored for 16 hours at 4 ^{o}C. Ten 0.5 ml samples were eluted, using additional cleavage buffer once the residual buffer had run out. These were run on a 10% SDS-PAGE (SDS-polyacrylamide gel electrophoresis) gel, and the largest fractions were pooled before quantification through a Bradford assay. The samples were then aliquoted and frozen at -80 ^{o}C.

2.4.4.2 Western blotting

Each sample consisted of either 30 µg or 60 µg cell extract, mixed with 4x SDS loading buffer (200 mM Tris.HCl pH 6.8, 400 mM DTT, 8% SDS, 0.4% bromophenol blue, 40% glycerol). These samples were denatured by boiling for 5 minutes, before being loaded on either 10% SDS-PAGE gels or 4-20% gradient SDS-PAGE gels, run using either a Mini-Protean 3 Cell (BioRad) or a GeneFlow vertical gel tank respectively, in 1x SDS running buffer (3.03 g/l Tris base, 14.4 g/l glycine, 1% SDS). The protein was then transferred from the gel to a PVDF (Polyvinylidene fluoride) membrane, using a Mini Trans-Blot Cell (BioRad), in Transfer buffer (25 mM Tris, 192 mM glycine, and 20% ethanol). The membrane was incubated in Ponceau Red stain for 5 minutes, and then transferred to water for a further 5 minutes. The transfer was checked on the membrane, before the remaining stain was washed off using

TBST (50 mM Tris, pH 7.6, 150 mM NaCl, 0.1% Tween® 20). The membrane is then incubated in TBST + 5% nonfat dry milk, for at least one hour at 4 °C, to block the membrane. 1 ml of primary antibody in TBST + 5% nonfat dry milk was then incubated with the membrane for at least one hour at 4 °C. The membrane was then washed three times with TBST, before 1 ml secondary antibody (also in TBST + 5% nonfat dry milk) was incubated with it, for a minimum of two hours at 4 °C. The membrane was then washed a further three times in TBST. The detection was done using Chemiluminescence (ECL system, Promega), and imaged using the LAS-3000 imaging system (Fujifilm). Primary antibodies were either anti-6x His tag (Monoclonal mouse anti-histidine tag, 1/1000, Serotec) or anti-GFP (Diclonal mouse anti-GFP, 1/500, Roche). The secondary antibody used was polyclonal rabbit anti-mouse, 1/2000, Dako.

2.4.4.3 *In vitro* Transposition assay

Each transposition reaction used transposition reaction buffer (100 mM potassium glutamate, 0.5 µg/ml BSA, 25 mM Tris.HCl pH 7.5, 0.5 µM β-mercaptoethanol, 10 mM $MgCl_2$). The substrate for the transposition reaction was pRC2106 (unless otherwise specified), which contains two 19 bp Tn5 TIRs with a total of 1767 bp in the transposable segment, and 2641 bp outside of the transposon sequence. The reactions were incubated at 37 °C, before addition of 0.2% SDS, then heat-inactivated for 5 minutes at 68 °C. 6x SDS loading dye (4 µl) was added, and 328 ng (20 out of 24 µl) of the

DNA was loaded onto a 1% TBE agarose gel, run overnight at 60 V. The gel was then stained with ethidium bromide.

2.4.4.4 Electrophoretic-mobility shift assay (EMSA)

Each reaction was set up with transposition binding buffer (100 mM potassium glutamate, 0.5 µg/ml BSA, 25 mM Tris.HCl pH 7.5, 0.5 µM β-mercaptoethanol), in a total volume of 20 µl. The complexes were assembled, as stated for each experiment, and the products separated on an 8% TBE polyacrylamide gel. Imaged using a FLA-3000 (FujiFilm).

Chapter 3

The Tn5 Transposase Dimer

3.1 Introduction

DNA transposons occur throughout the three domains of life. Some transposon superfamilies, such as ITm, are widespread across the kingdoms whereas others, such as IS4 or P elements, have a smaller range of hosts in which they have been identified (Tellier et al. 2015). While some cases of limited range can be explained by the necessity of host proteins to transpose (such as for Sleeping Beauty transposition (Ivics & Izsvák 2015)), not all transposition systems require them, and therefore it is likely that there are other factors that affect the transposition activity.

There is a huge variability in genome sizes between different host organisms, from $1.8x10^6$ (*Pelagibacter ubique* (Giovannoni 2005)) up to $2.22x10^{11}$ base pairs (*Paris japonica* (Pellicer et al. 2010)). The transposase must locate the specific binding site within these genomes, which is a significant challenge. In bacteria, the problem of the genome size is, to some extent, circumvented as the transposase protein is transcribed and translated in close proximity to the transposon DNA. This produces a high local concentration of the protein that increases the frequency at which the transposon ends (TIRs) are bound. A high local concentration is particularly beneficial for those transposases that are monomeric in solution as each TIR must be bound independently and sufficiently close in time that the full synaptic complex can be produced. This pathway is known as S-PD. Eukaryotes, by comparison, translate the protein outside the nucleus and import it back in, therefore there will be no high local concentration of the

transposase. Due to this, the frequency of both ends being bound independently would be extremely low, particularly given the generally larger genome sizes of eukaryotes. It is possible that this is why no S-PD based transposition systems have yet been identified in higher eukaryotes. The eukaryotic transposase proteins characterised thus far form multimeric complexes in solution that initially bind one end then recruit the other into the complex, a pathway known as S-NEC (J. M. Richardson et al. 2009, Hickman et al. 2014, Claeys Bouuaert et al. 2017). Complex formation would therefore require only one TIR to be located across the whole genome, which is likely to increase the frequency of transposition dramatically compared with S-PD (Claeys Bouuaert et al. 2013).

This chapter explores the hypothesis that the requirement for two independent monomeric binding events is the root cause of S-PD not being found and, when tested, not being functional in higher eukaryotes (Reznikoff 2013, pers comm). Two single-chain dimeric transposase constructs were designed, and their activity tested in both prokaryotes and higher eukaryotes. The complex formation pathway of the dimeric transposases was also analysed to test if S-NEC is being effectively mimicked. These mechanistic alterations provide significant insight into what mechanistic requirements exist for transposase function in different organisms.

3.2 Results and Discussion

3.2.1 Design and analysis of the dimeric transposase

3.2.1.1 Testing of the dimeric transposase in bacteria

The dimeric transposase was constructed using two monomers of hyperactive Tn5 transposase that each contain the E54K, M56A and L372P mutations (Goryshin & Reznikoff 1998). The E54K mutation affects a residue in close proximity to the DNA strand, altering the residues charge from negative to positive, allowing it to interact directly with the DNA. The M56A mutation prevents expression of the 'inhibitor' protein, an N-terminally truncated form of Tn5 transposase, which forms complexes with the transposase in solution and prevents it binding to DNA. The Tn5 transposase C-terminus has been shown to interfere with the DNA binding of the N-terminal region; the L372P mutation reduces the flexibility of a key α-helix, reducing the inhibition (Wiegand & Reznikoff 1992; Weinreich et al. 1994; Zhou & Reznikoff 1997; Davies et al. 2000). The two Tn5 transposase monomers require a protein linker to connect the C-terminus of one monomer to the N-terminus of the other. In the post-excision crystal structure this distance is known to be 13.1 Å (Davies et al. 2000), although this is the distance at only one stage in a multi-step reaction. Therefore, the linker must be significantly longer than 13.1 Å to ensure that the proteins have sufficient

flexibility to allow transposition to proceed uninhibited. To this end, two different linkers were designed. The first has only a short 22 aa sequence (GSSRGGGSEGGGSEGGSGTSEF) which is predominantly glycine-serine rich, in line with other flexible linkers (Reddy Chichili et al. 2013; X. Chen et al. 2013). This would reach ~79.2 Å based on an average length of 3.6 Å per residue, which should be sufficient to allow transposition. The second linker is longer at 187 aa. This was previously used to fuse an Hsmar1 dimer (Claeys Bouuaert et al. 2017). It contains the short 22 aa sequence used in the other construct, followed by a TrxA tag, a 6x Histidine tag, a thrombin site, a streptavidin tag, and an enterokinase recognition site. The longer linker allows significantly more freedom of movement of the two transposases respective to each other, which may be necessary during complex formation or the later reaction steps. The various tags and sites are likely to prevent this linker from fully extending, though theoretically it has a total reach of up to 673 Å. The Tn5 dimers using these two linkers shall henceforth be referred to as the 'small dimer' and 'large dimer' respectively. Schematics are shown in Figure 13a and full protein sequences can be seen in section 8.1.2.

Initial analysis of transposition activity was carried out through a papillation assay (see section 2.4.2.1). This assay uses a reporter transposon, located in the genome of *E. coli,* that has a *lacZ* gene which is missing a promoter. When transposition occurs, the transposon may integrate downstream of an endogenous promoter, which can lead to expression of LacZ. As the plates contain X-gal, these events will cause the

cells to turn blue, and the frequency of the blue spots, known as papillae, occurring on single colonies indicates the frequency of transposition (Foster et al. 1981). This assay was carried out on the monomeric and both dimeric transposase constructs (sequences of which can be found in section 8.1.2), using various different concentrations of lactose and/or IPTG (Figure 13b). For this assay, the vector was pMAL-C2X, which maintains ~20 copies per cell, and uses the p_{TAC} promoter under the control of the Lac operator. Therefore, increasing the levels of IPTG, and to a lesser extent lactose, increases the expression of the transposase.

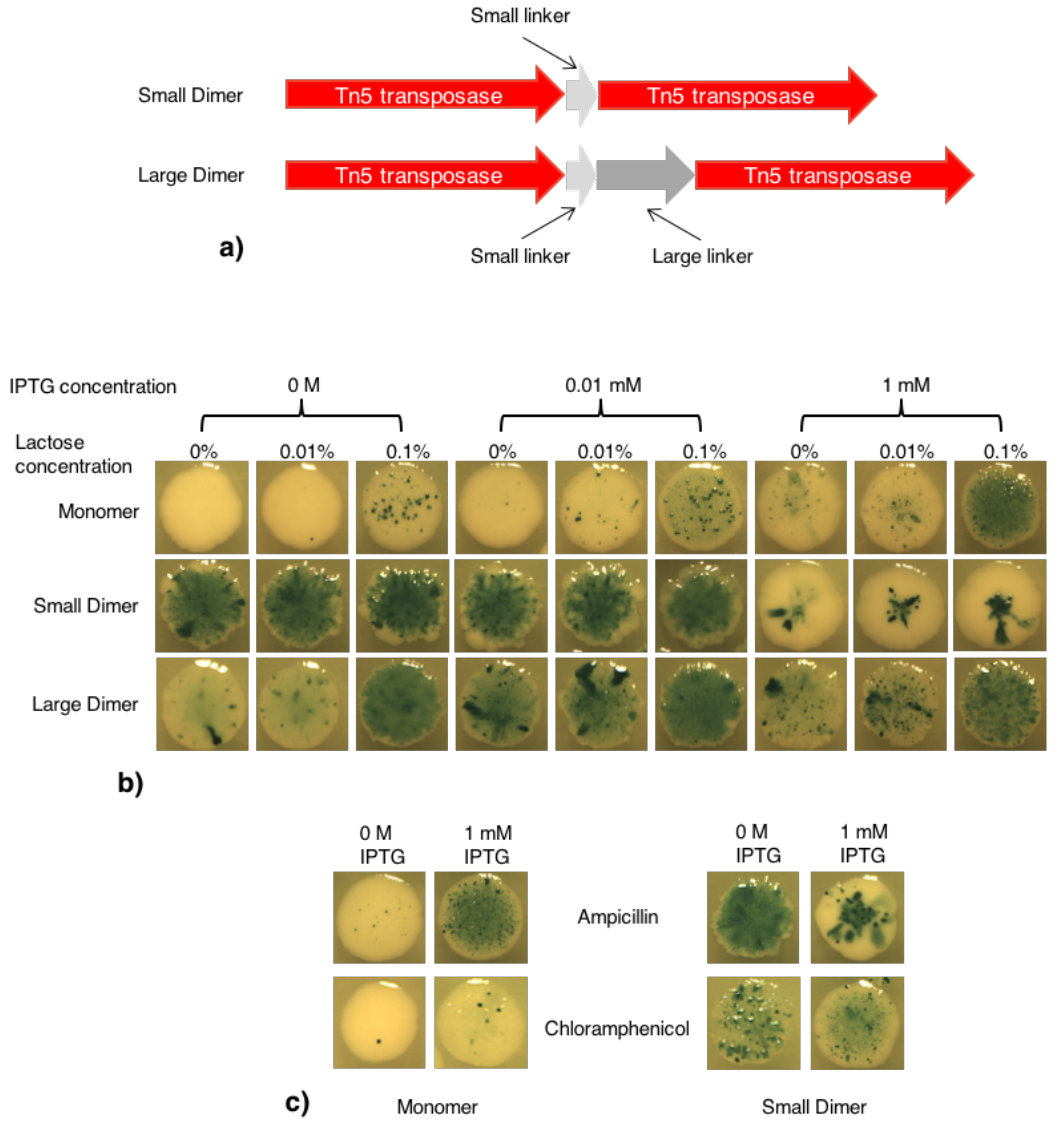


Figure 13: Schematics and Papillation Assays of the Transposase Dimers. a) Schematics of the Small Dimer and Large Dimer. The transposases are shown as red arrows, the small linker in light grey and the large linker in dark grey. Schematics are not to scale. **b-c)** The strain used in these assays was RC5123. The different plates were supplemented with lactose and/or IPTG as well as ampicillin and X-gal. The colonies shown are representative of those on the plates. Blue papillae indicate LacZ expression. **b)** Assay of the various expression levels of the Tn5 monomer (pRC2117), small dimer (pRC2109) and large dimer (pRC2126), using the pMAL-C2X expression vector. **c)** Assay comparing papillation when using the monomer or small dimer with ampicillin (pRC2117 or pRC2109) and when using chloramphenicol (pRC2114 or pRC2113). All these plates were supplemented with 0.1% lactose as well as the X-gal, antibiotics and in some cases IPTG.

The monomeric transposase showed steadily higher activity as the expression level increased, producing more individual papillae per colony. Larger papillae were seen on plates with higher lactose concentrations, due to the LacZ-expressing cells being capable of metabolising lactose, which led to blue sections of the colony growing after other carbon sources have been exhausted. At the 1 mM IPTG and 0.1% lactose level, the colony is mostly blue in colour. Under these conditions, the transposase expression will be at its highest. As the transposition events happen very frequently in the colony, many papillae will be produced. Due to the large quantity of lactose, the papillae will grow significantly more during the later stages of the assay, becoming indistinct from one another and covering the majority of the colony in blue cells.

Both the dimers appeared significantly more active than the monomer, particularly at lower expression levels. However, the behaviour of the dimers was not consistent with the monomer in this assay, as there were frequently fewer individual papillae and many more whole sub-sections (or sectors) of the colony that turned blue. Blue sectors indicated activity earlier in the colony growth, as cells from earlier in colony development produced a larger proportion of the final colony. For example, a LacZ-expressing cell occurring among only eight cells would be expected to produce ~1/8 of the colony as blue. This suggests the dimeric versions have an increased rate of activity or, at least, that transposition events are likely to occur much earlier than with the monomeric transposase. The monomer, even at the highest expression levels, showed very few sectors. This observation

indicated that the large quantity of transposase, that would yield higher transposition rates, requires time to accumulate in the cells, therefore transposition would not occur early enough to produce a lot of sectors.

The small dimer demonstrated very unusual behaviour at the 1 mM IPTG levels. Each colony showed a few sectors as well as some papillae close to the centre of the colony, while the outer areas had no new papillae. Samples taken from the centre and edge of the colonies all demonstrated resistance to kanamycin, indicating the transposon was still present. The samples from the edge were unable to grow in ampicillin however, and the plasmid could not be recovered from them. Therefore, the plasmid was lost from these cells. The hypothesis to explain these findings is that the presence of the small dimer expression cassette is sufficiently detrimental to the cell's survival that it acts as a negative selection marker. After the first few hours, the ampicillin around the initial cells was degraded by the secreted β-lactamase, therefore the positive selection from the plasmid's antibiotic resistance was no longer as significant a factor. The plasmid however, was still being selected against due to the small dimer expression cassette. Any cells that lost the transposase expression plasmid would therefore have been at a significant growth advantage, and would have out-competed cells containing the plasmid. This leads to them replicating faster and forming the outer region of the colony. Should transposition have already occurred when the plasmid was lost, the cells that spread out to the edge of the colony would have formed a blue sector. If this explanation was correct, alternative antibiotics that are not degraded around the resistant cells should

have papillation throughout the colony. To test this reasoning, a papillation assay was conducted using plasmids that encoded the gene for chloramphenicol resistance instead of ampicillin (Figure 13c). The *cat* gene encodes chloramphenicol acetyltransferase, which inactivates the antibiotic within the cell instead of being secreted. Under these conditions, the papillation effect was not limited to the central region of the colony and instead showed a lower overall level of transposition, with many small papillae (Figure 13c). The small dimer did appear to be less active, or at least took longer to initiate transposition, when under chloramphenicol selection compared with ampicillin. The exact cause of this is unknown, though it did occur with both the monomer and small dimer. One possibility is that there was reduced expression of the transposase under this selection, although this hypothesis has not been tested.

Mutations in the ORF of the transposase, such as a premature stop codon, causing inactivation, would likely be significantly more beneficial to the cell than loss of the entire plasmid, as the antibiotic resistance would be maintained. Transposase mutation events would occur at a much lower frequency compared with the frequency of plasmid loss, which occurs once in every 10,000 cell divisions (Boe et al. 1987). It might be possible to add a stability locus, such as *sop*, which should reduce the rate of plasmid loss. This could hypothetically reduce the phenotype seen in the colonies, where the outer portion of the colony shows no new papillae, although this has yet to be tested.

3.2.1.2 Purification and *in vitro* testing of the transposase dimers

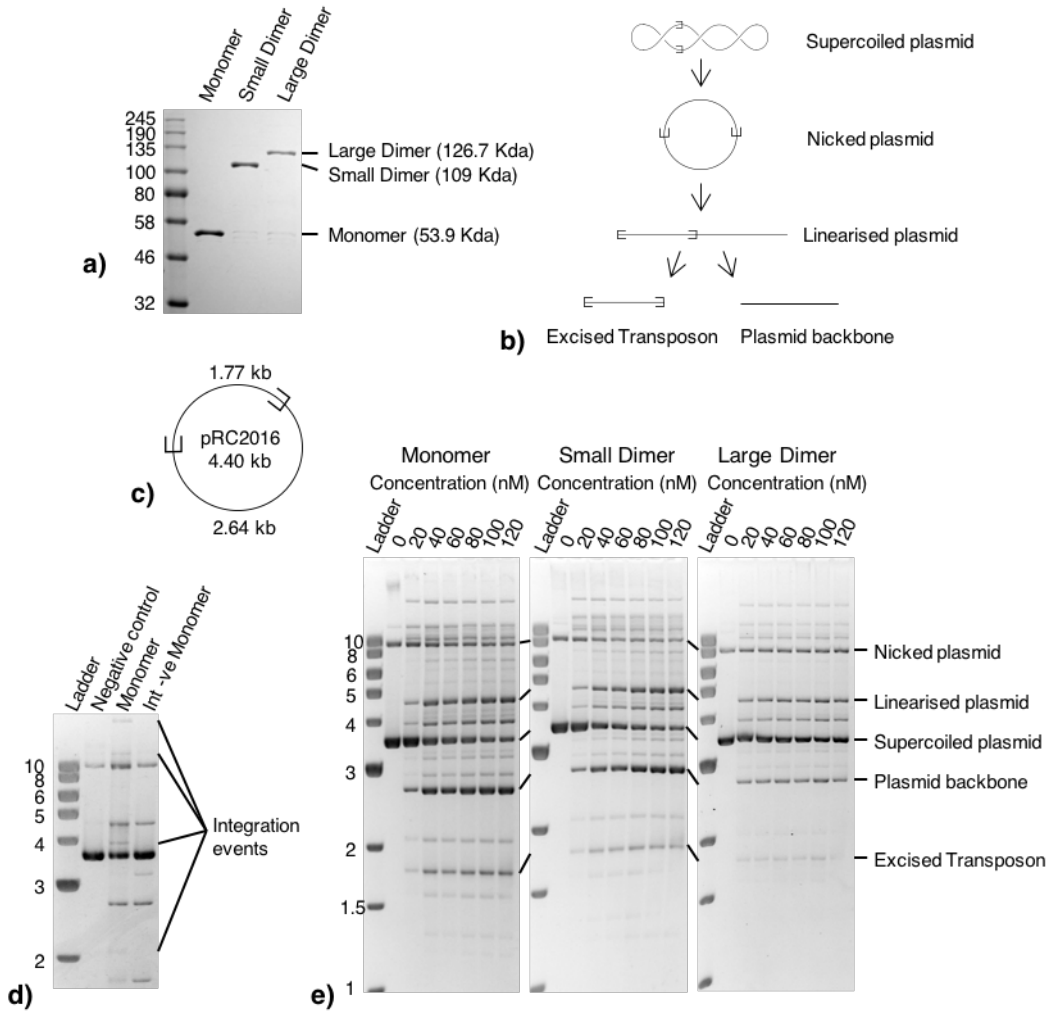


Figure 14: Purification and Titrations of the Dimeric Transposases. a) 10% Denaturing gel showing the purification of the different transposase constructs. Monomer was derived from pRC2128, small dimer from pRC2102 and large dimer from pRC2101. Each lane was loaded with 1 μg of protein. **b)** Diagram demonstrating the steps of the transposition reaction, up to the fully-excised transposon step. All the DNA shown is dsDNA, and the brackets indicate the transposon ends, and their orientation. **c)** Schematic of the plasmid substrate used for parts d and e. **d)** Transposition assay gel showing the integration-deficient mutant transposase compared with functional transposase monomer, using 7 nM pRC2106 as the plasmid substrate. 328 ng of plasmid was loaded per lane. Bands linked with integration of the transposon are indicated. **e)** Transposition assays of the monomer, small dimer and large dimer, titrating the concentration (nM) of the protein. Incubated for 4 hours with 7 nM pRC2106 substrate plasmid. 328 ng of plasmid was loaded per lane. The various bands from the transposition reaction are indicated.

Tn5 transposase purification has been well documented through the use of an intein tag (Naumann & Reznikoff 2002). Briefly, the intein tag is fused to the C-terminus of the protein (see section 8.1.2 for exact sequence). The intein tag contains a chitin-binding domain, which binds to chitin resin on a column. When a thiol reagent is applied, such as dithiothreitol (DTT), it induces the self-cleavage activity of the intein, leaving the protein without a tag when eluted from the column. The monomer, small dimer and large dimer were purified using this system (Figure 14a).

Analysis was done using an *in vitro* transposition assay (see section 2.4.4.3), where a transposon reporter plasmid is mixed with the protein and incubated, before it is run on an agarose gel. The reporter plasmid contains two TIRs separated by ~1.7 kb of DNA. The TIRs used are the Tn5 mosaic ends (agatgtgtataagagacag) which show optimal binding for the transposase (M. Zhou et al. 1998). Initially the plasmid substrate (Figure 14b) is predominantly supercoiled, though a small proportion appears in the nicked (or relaxed) form. The transposition reaction begins with a nick in the substrate, then one TIR is fully cleaved. This produces linearised plasmid, which appears on the gel approximately half-way between the supercoiled and nicked bands. Transposition proceeds by cleaving at the other TIR, which leaves two bands, the excised transposon, and the plasmid backbone (Figure 14b). After the transposon has been excised, it can then be integrated into any DNA in the suspension. These integration steps are more complex to identify due to the wide variety of potential targets, which can include the transposon DNA itself, which is known as autointegration

(Chalmers & Kleckner 1996). The specific bands associated with integration of the transposon could be identified through the use of an integration-deficient mutant transposase. A series of monomeric mutants, that had been previously linked to deficiency in target-DNA binding, were produced and tested (Gradman et al. 2008) (Supplementary Figure 1). K160D demonstrated low activity in the papillation assay, and when tested using the *in vitro* transposition assay, excision appeared unaffected while the bands associated with integration were no longer visible on the gel (Figure 14c).

To test for optimal concentration, the three transposase proteins (monomer, small dimer and large dimer) were subjected to *in vitro* transposition assays, titrated from 0 to 120 nM, with each reaction left to proceed for four hours (Figure 14d). The simplest method to identify the activity levels of the various proteins was to compare the intensity of the band that corresponds to the plasmid backbone. This was because the plasmid backbone should not be consumed by the transposition reaction after it has been generated (apart for a small amount of integration events) as there is no transposon present within the backbone. It was expected that for this titration the dimeric transposases would show optimal activity at approximately half the concentration of the monomeric form, as each dimer contains two transposase proteins. Against expectations, all three of the proteins showed optimal activity at the 100 nM concentration.

Transposases that are dimeric in solution (such as the *mariner* family Mos1 and Hsmar1) often demonstrate OPI (Lohe & Hartl 1996b; Claeys Bouuaert et al. 2013). Under OPI, the activity peaks at a relatively low

concentration, then decreases significantly as the concentration rises further. The cause of OPI is that transposase dimers bind to a single TIR (forming a SEC) at high affinity, and are tightly bound. The transposase binding affinity for the second TIR is much lower, there is therefore a significant delay in second end recruitment, during which time another dimer can bind the second TIR, and prevent the PEC forming (Claeys Bouuaert et al. 2013). Both Tn5 transposase dimers did not show any decrease in substrate conversion after the concentration peaked, instead staying approximately constant, in line with what was seen for the monomer, which indicates an absence of OPI occurring. The Tn5 transposase dimer is expected to not have any lower affinity for the second-end binding than it does for the first TIR. Therefore, there would be a very limited time-span for a second dimer to block the other TIR, reducing any OPI. The Tn5 transposase SEC has a relatively weak interaction (evidence for this is shown in section 3.2.3.4), so if a second dimer blocked PEC formation, it would not remain bound long enough to significantly inhibit transposition.

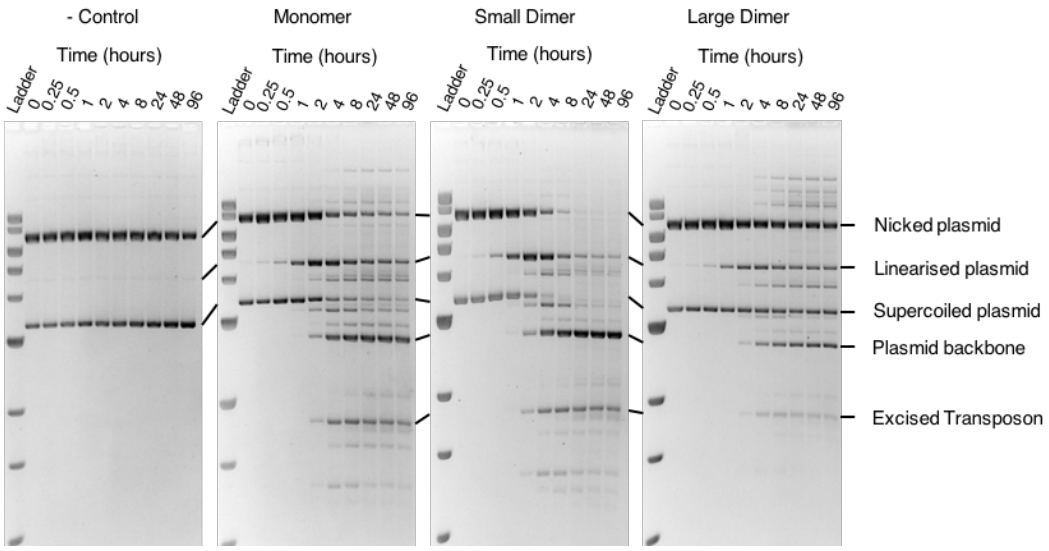


Figure 15: *In Vitro* Transposition Assay Time Courses of the Monomeric and Dimeric Transposases. All assays were done with 7 nM pRC2106 and 100 nM of the protein. 328 ng of plasmid substrate was loaded per lane. The negative control was done by supplementing the reaction with the protein storage buffer.

Based on the titration data, the three proteins were subjected to a time course transposition assay (see section 2.4.4.3), at 100 nM transposase, for 96 hours (Figure 15). The negative control assay showed very minor variation over the time-course, as the quantity of supercoiled substrate slowly increased and the nicked substrate slowly decreased. This effect was unavoidable, and was due to the storage conditions which caused some nicking of supercoiled plasmid after the reaction terminated. The early reactions were stored for longer under these conditions which led to increased nicked plasmid.

Monomeric Tn5 transposase demonstrated a steady conversion through the transposition reaction stages. The nicked substrate peaked at 0.25 hours, the linear at two hours and the plasmid backbone at 24 hours, while the integration bands were clearly visible from the four-hour time point

onwards. While the supercoiled plasmid substrate was still present at the 96-hour time point, most of the substrate had been depleted. The rate at which the supercoiled and nicked substrate are depleted reduced in the later stages of the reaction. This was because, of the pool of transposons available in the reaction, a smaller and smaller proportion are within the plasmids of the supercoiled and nicked substrate bands.

The small dimer demonstrated the highest rate of excision among the three proteins. The supercoiled band was absent by the eight-hour time point and the nicked portion was almost entirely depleted after 24 hours. The linear substrate, while never fully consumed, was almost entirely depleted by the 96-hour time point. The excised plasmid backbone was also more intense at its peak than the other transposases. Despite this high excision rate, the bands that corresponded to integration events were almost entirely absent throughout the assay, only visible in the two, four and eight-hour time points, and at much lower intensity than the monomer. Given the very high rate of transposition *in vivo* (Figure 13) and the high rate of excision in the early stages of the assay, the small dimer was likely to be integrating the transposon into its target site, and then subsequently excising it back out.

The large dimer appeared to show a much lower overall rate of excision compared with either of the other two proteins. The supercoiled, nicked and linear substrates were all depleted at a much slower rate, and the excised plasmid backbone was less intense by 96 hours than the other proteins are, even at four hours. The bands that were indicative of integration events, particularly those that are located above the nicked substrate, were

significantly more intense for the large dimer. Overall this suggested that the large dimer has a lower rate of excision of the transposons but a higher rate of successful transposition events when the complex has formed, in comparison with the monomeric transposase. The 187-aa linker may be long enough to interfere with the transposases during the DNA binding stage, which reduces the amount of transposition events. Once it has formed however, the complex may be more stable than with monomeric transposase. Consequently, the linker may allow the complex to reform if disassembly occurs, which could lead to a higher rate of integration.

Overall this assay indicated that the dimers act in very different ways *in vitro*. The small dimer was found to be hyperactive during early stages of transposition, while the large dimer appeared to be hyperactive during the later stages. The end result was similar however, and both dimers appeared to be more active than the monomeric transposase, which was consistent with the *in vivo* results (Figure 13a).

3.2.2 The transposase dimers in eukaryotic cells

Analysis of transposition activity in bacteria demonstrated that the dimers were more active than the monomer, the next stage therefore was to test the transposase constructs in human cells. The monomer, small dimer and large dimer were all codon-optimised for human expression. As the Tn5 transposase cannot typically localise to the nucleus, an SV40 nuclear localisation signal (NLS) was attached to the N-terminus of the protein (Kalderon et al. 1984), and a 6x-His tag was fused to the C-terminus to test expression levels. Full sequences of the proteins can been seen in section 8.1.2.

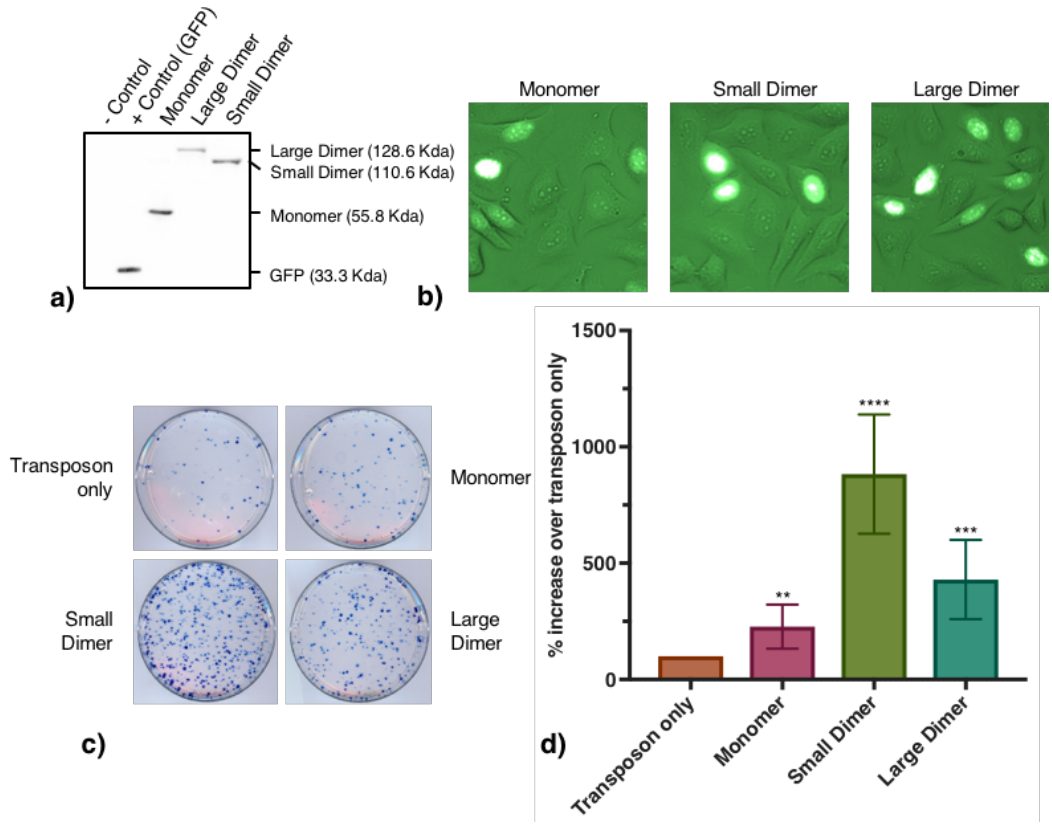


Figure 16: The Tn5 Dimeric Transposases in Eukaryotes. a) A Western blot, identifying the level of expressed protein in full cell extracts of HeLa cells. Cell extracts were taken after 24 hours, and 60 μg was loaded per lane. The primary antibody was anti-6x His tag. Sizes were determined by comparing with the ladder that is not shown. **b)** Fluorescent microscopy images of the cells expressing the different transposase constructs fused with GFP. The images were taken after 24 hours, and false colour was added subsequently. **c)** Human cell transposition assay plates, after 2 weeks of selection with G418. Cells were transfected with 1 μg reporter plasmid (pRC2194) and 1 μg transposase expression plasmid (pBS for transposon only, pRC2148 for monomer, pRC2147 for small dimer and pRC2146 for large dimer). $4x10^5$ cells were loaded per plate at the start of the selection. **d)** The frequency of cell survival with the different transposase constructs. The graph depicts the percentage increase over the transposon-only baseline. Error bars are standard deviation. ** means P<0.01, *** means P<0.001 **** means P<0.0001. Statistical test used was a ratio paired T-test, n=8.

Initially, the expression levels in HeLa cells were analysed through a western blot (see section 2.4.4.2) of cell lysate samples from 24 hours after transfection with the expression plasmid (Figure 16a). All three proteins were present on the blot, at the correct sizes, which indicates both successful transfection and expression. It was unlikely the proteins suffered significant degradation due to the lack of protein bands below the expected size. However, degradation of the 6x His tag on the C-terminus would render the product no longer visible. The monomeric transposase appeared to show slightly increased expression compared with either of the dimeric constructs, which is consistent with expression levels seen in *E. coli* (data not shown).

Although the three transposase constructs were expressed in HeLa cells, it was important to confirm successful localisation to the nucleus of the cells. To this end, a Green-Fluorescent-Protein (GFP) ORF was fused into each of the three constructs, at the N-terminal end between the NLS and the transposase, between the valine and asparagine residues (see section 8.1.2 for full sequence before GFP addition). These constructs were transfected into HeLa cells, and the live cells were examined (see section 2.4.3.4) after 24 hours (Figure 16b). All three proteins were found to be capable of localising into the nucleus, with minimal fluorescence visible in the cytoplasm. This indicated that localisation would not be a limiting factor in the activity of the transposase constructs.

As the proteins were found to be capable of expressing and localising to the nucleus of the cells, transposition integration assays (see section 2.4.3.5) were performed with each protein. In this assay, the HeLa cells are

transfected with a transposon reporter plasmid (that contains a neomycin resistance gene between two TIRs) and either a control or transposase expression plasmid. After two weeks of G418 selection the plates are stained and the surviving cells can be seen as small foci (Figure 16c). The plasmids cannot be replicated, so in order for the antibiotic resistance to spread to the daughter cells of the one that was initially transfected, it must become integrated into the genome.

In this assay, the reporter plasmid alone was capable of integration at a low frequency (average of 81.3) through host mechanisms. When the monomeric transposase is present, there was a statistically significant increase in the rate of integration (average of 182.7, 227% increase, P=0.0062) (Figure 16d). Both the dimers showed a statistically significant increase in integration frequency, 883% for the small dimer (average of 726.6, P=<0.0001) and 430% for the large dimer (average of 350.6, P=0.0001). These results demonstrated that the transposase is capable of activity in the human cells, both when linked as a single-chain dimer and in the original monomeric form, though the monomeric form is significantly less active. It is important to note that while the monomer was active in these experiments, the expression level was far above what would be seen in a normal transposition system, where it is likely that the S-PD pathway would show a more significant detriment. Experiments with other transposition systems, such as *piggyBac* and Sleeping Beauty, show much higher rates of transposition than was seen here, so Tn5 may have other factors that inhibit its transposition within a eukaryotic genome (X. Li et al. 2013; Y. Wang

et al. 2014; Yusa 2015; Grabundzija et al. 2013; Ivics & Izsvák 2015; F. Voigt et al. 2016).

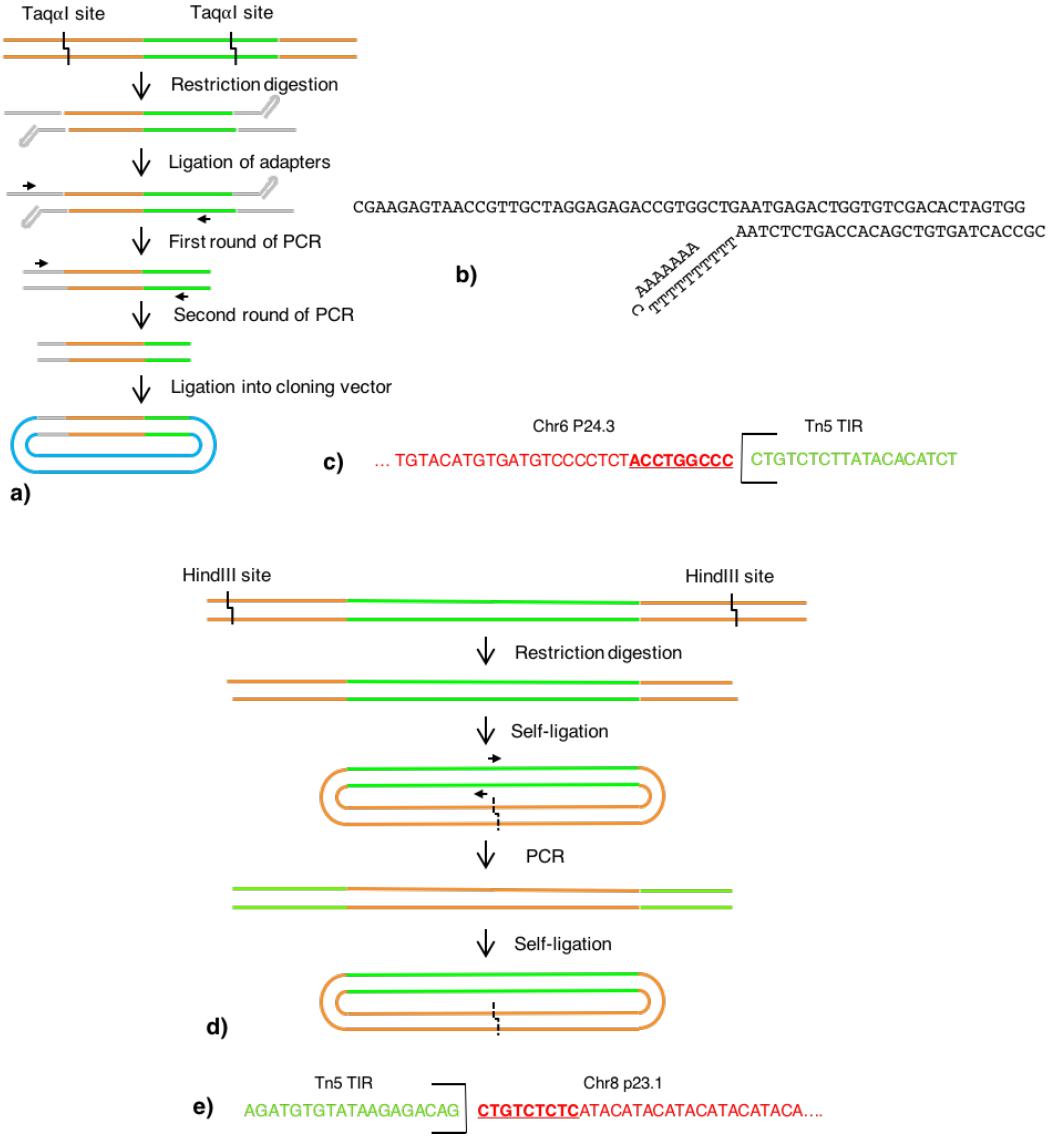


Figure 17: Mapping of Transposition Integration into the Human Genome. a) A model of the Splinkerette PCR technique. The transposon DNA is shown in green, the DNA it is integrated into is shown in red, the splinkerette adapters in grey and the cloning vector DNA in blue. The black arrows indicate the sites or primers for the PCR reactions. **b)** The sequence of the splinkerette adapter. The GC overhang on the bottom strand allows it to ligate to TaqαI digested DNA. c) The sequence of the integration event that was found using Splinkerette PCR. The transposon DNA is shown in green and the human genomic DNA in orange. The underlined sequence is the target sequence for the Tn5 transposase in this case. d) A diagram showing how the inverse PCR approach works. The transposon DNA is shown in green, and the target DNA in red. The black arrows indicate the sites of primers for the PCR reaction step. e) The sequence of the integration event that was identified using the inverse PCR approach. The transposon DNA is in green and the human genomic DNA in orange. The underlined sequence is the target sequence for the Tn5 transposase in this case.

Mapping individual integration events into the human genome is a clear way to demonstrate that the increase in surviving cells is due to transposition events instead of any side-effects of the transposase being present. As the small dimer appeared to be the most hyperactive in HeLa cells, genomic DNA was extracted from those cells.

The first method used to identify the integration sites was Splinkerette PCR (Figure 17a) (Potter & Luo 2010). In this technique, the genomic DNA is cut with a restriction enzyme that does not cut within the transposon DNA, and leaves a 5' overhanging sticky end. Compatible splinkerettes (Figure 17b) are ligated onto the sticky ends. These splinkerettes are double-stranded dephosphorylated oligonucleotides that have a hairpin loop structure on the 5' overhanging strand. Two rounds of nested PCR are done, each round using primers, one located in the splinkerette and one within the transposon DNA. The final product is then sequenced. This approach was done with genomic DNA from the small dimer-transfected cells (that had used the pRC2106 reporter plasmid), though only a single transposition event was recovered (Figure 17c). This event was mapped at one end of the transposon, where it had integrated into the short arm of chromosome 6. Despite significant effort the other TIR could not be mapped, though it is possible it was only one TIR integrated and the other remained in the plasmid.

The second approach for identifying transposition integration sites was inverse PCR mapping (Figure 17d) (Trubitsyna et al. 2017). In this approach, the genomic DNA is digested with a restriction enzyme that does

not cleave within the transposon. After digestion, the DNA is self-ligated, and then inverse PCR is performed, originating from within the transposon DNA. This product is self-ligated and transformed into bacteria where the transposon-containing DNA will be able to propagate. Sequencing is then used to identify the integration sites. This was attempted using a fresh genomic DNA sample (that had been transfected with pRC2194 as the reporter plasmid) and through this approach another single-end transposition event was mapped, this time into a site in the short arm of chromosome 8 (Figure 17e). This was only a one sided-integration event, as the other end was still attached to the plasmid (verified through sequencing of the region), so the transposase must have formed a complex between TIRs derived from two different plasmids in order to integrate. These two single-end events showed that the small dimer is capable of transposition in human cells, and integration into the genomic DNA is possible.

The increase in surviving small dimer foci observed strongly suggests that there is a large number of transposition events occurring. While the single-end transposition events showed the occurrence of transposition, the lack of mapped double-ended events was not entirely in line with this finding, as two TIRs located on the same plasmid are more likely to be bound in a single transposition complex than ends located on different plasmids. As the two events mapped were derived from the two different TIRs, this shows that both were functional (although in slightly different versions of the reporter construct). It is possible that the increased survival observed in the assay was due, in part, to increased genomic instability caused by the small dimer

expression. The 19 bp Tn5 mosaic end TIR does occur within the human genome (100% match found 20 times using BLAST alignment with the human RefSeq genome (Altschul et al. 1990)), which suggests that the transposase might be causing damage to the genome. This damage potentially increases the integration of the reporter plasmid through means other than transposition. The results of single-ended transposition events however, do suggest that some of the increased survival is due to transposition.

These findings showed that Tn5 transposition was possible in eukaryotes, particularly through the use of a dimeric transposase. This suggests that the Tn5 monomeric transposase is also capable of catalysing transposition in human cells, this occurs at such a low frequency due to the two independent binding events, that it is less detectable. Each small dimer locates the TIRs as frequently as the monomers but only requires one binding event at that low frequency. The second binding event likely occurs at a much higher as a result of the close proximity of the second monomer. It is possible that the reason the large dimer shows lower transposition frequency than the small dimer is that the linker region of the large dimer is significantly longer. This increases the distance between the two monomers, so they will collide less frequently and therefore are less likely to form the PEC once the second DNA strand is bound.

3.2.3 Characterisation of the dimer complex assembly pathway

The dimeric transposases were shown to be hyperactive compared with the monomeric form, and both are capable of transposition activity in human cells. The hypothesis of this section of the work is that the transposase dimer will be active because it can mimic the complex formation mechanism, S-NEC. While the transposition activity in eukaryotes supports this hypothesis, it is important to prove this is the pathway being used, particularly as the dimeric Tn5 transposases do not demonstrate OPI (see section 3.2.1.2), a hallmark of at least some S-NEC systems. In standard S-NEC transposition systems, the transposase dimerises prior to binding the DNA. The single-chain Tn5 transposase dimer system however, has the two transposases only joined by a protein linker to the other monomer, which leads to two possible mechanisms of action.

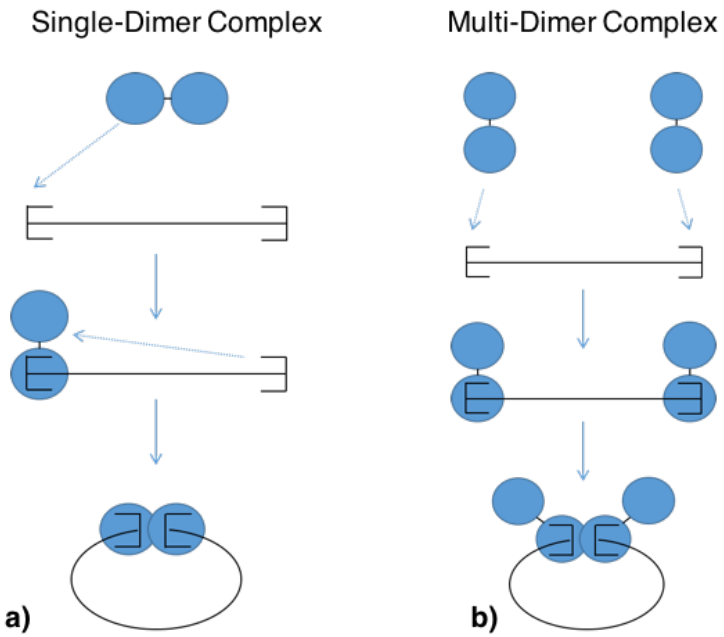


Figure 18: Model Of The Two Complex Formation Mechanisms Of The Dimeric Transposase. The transposase is shown as blue circles and the transposon as a black line with brackets at either end. **a)** Diagram showing the formation of a Single-Dimer Complex (SDC). One monomer from a single dimeric transposase binds the transposon end. The other transposon end is then recruited into the complex, forming the PEC. The transposon can then excise from the donor DNA. **b)** Diagram of the alternative formation mechanism, the Multi-Dimer Complex (MDC). Monomers from separate transposase dimers bind the two transposon ends independently. The two bound monomers form the PEC together. The two additional monomers that are covalently linked to the PEC reside in close proximity.

The first mechanism of action is that it functions like S-NEC; the dimer binds one end and then recruits the other naked end before forming the PEC. This would be a Single-Dimer Complex (SDC) (Figure 18a). Alternatively, separate dimers could bind to each end of the transposon independently. The monomers that are bound to the TIRs could then dimerise with each other, leaving the other monomers unbound, here referred to as a Multi-Dimer Complex (MDC) (Figure 18b). It is important to identify which of the two complex formation pathways is used before conclusions can be formed.

The more active of the two dimeric constructs, the small dimer, was investigated further.

3.2.3.1 The ‘long-short’ experiment

The first approach taken to test the complex formation pathway was the ‘long-short’ experiment. This experiment is used to identify the number of separate DNA fragments that are bound in a complex, by using two different labelled substrates, the short (93 bp, see section 8.1.3 for full sequence) and the long (154 bp, see section 8.1.3 for full sequence). These each contain a single Tn5 TIR within them, and sequence on either side to allow effective binding. Both the monomer and small dimer were tested with each substrate independently and also together (Figure 19).

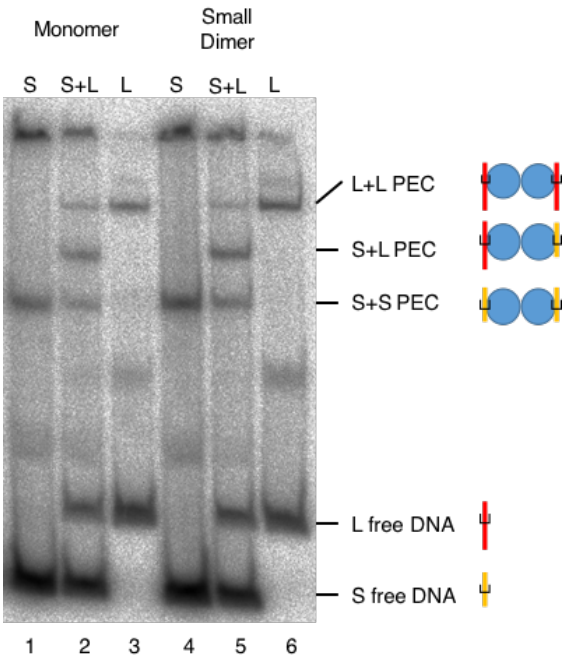


Figure 19: The Long-Short Experiment. This demonstrates the number of DNA fragments bound by the transposase complex. Reactions were incubated for 15 minutes at room temperature then run on an 8% TBE gel. Both DNA strands were labelled using α-^{32}P dCTP. S=Short, L=Long. The various different complexes formed are shown on the left. The blue circles indicate the transposase, the red line indicates the long substrate, the yellow line indicates the short substrate and the black bracket on the DNA strand indicates the transposon end sequence.

In the presence of the short substrate (lane 1), the monomer formed a single well-defined band which corresponded to the short-only PEC. Similarly, when the monomer was run with the long substrate alone (lane 3), there was a single well-defined band corresponding to the long-only PEC. When the two substrates were supplied together (lane 2), the short-only and long-only PEC bands formed, as well as a band midway between them, which corresponded to a PEC formed from one short and one long substrate. This demonstrates the complex is only able to bind two strands of DNA. If only one strand is bound the lane would not show the middle band, and if more

than two strands are bound there would be at least four bands instead of three.

There were some additional bands that appeared on the gel that can be accounted for. Above the long-only PEC band there was a faint band. This is known to be the 'slow' version of the PEC that exists prior to the step where the W323 disrupts the DNA helix, the predominant form seen in the assay (Bischerour & Chalmers 2007). There were also faint, broad bands below the PECs in every lane; these were the SECs. As the protein-DNA interactions are much weaker in the SEC, the transposase dissociates more frequently, causing a smearing effect on the band as it runs on the gel. These were not accounted for in previous Tn5 work (Bischerour & Chalmers 2007), but further evidence was presented to define them in later work (see section 3.2.3.4).

The small dimer showed the same bands as the monomeric transposase, which suggests that it too can only bind two DNA strands. If the small dimer behaved as an MDC, the difference in protein quantity in the complex may make the complex run slower on the gel. The MDC may also produce additional bands that correspond to substrate being bound by the two extra transposases in the complex.

This assay does have some limitations however. The overall size of the protein is significantly smaller than the DNA strands that it is bound to, so an MDC may not migrate down the gel significantly slower than the monomeric PEC. The additional bound substrate in an MDC would also be unlikely to form a PEC. SEC formation is a weak interaction, so unless a

second PEC was formed within an MDC, it would be unlikely to show up as a visible band. The evidence here does support the SDC hypothesis, although it is not conclusive.

3.2.3.2 Active site mutagenesis of the dimer

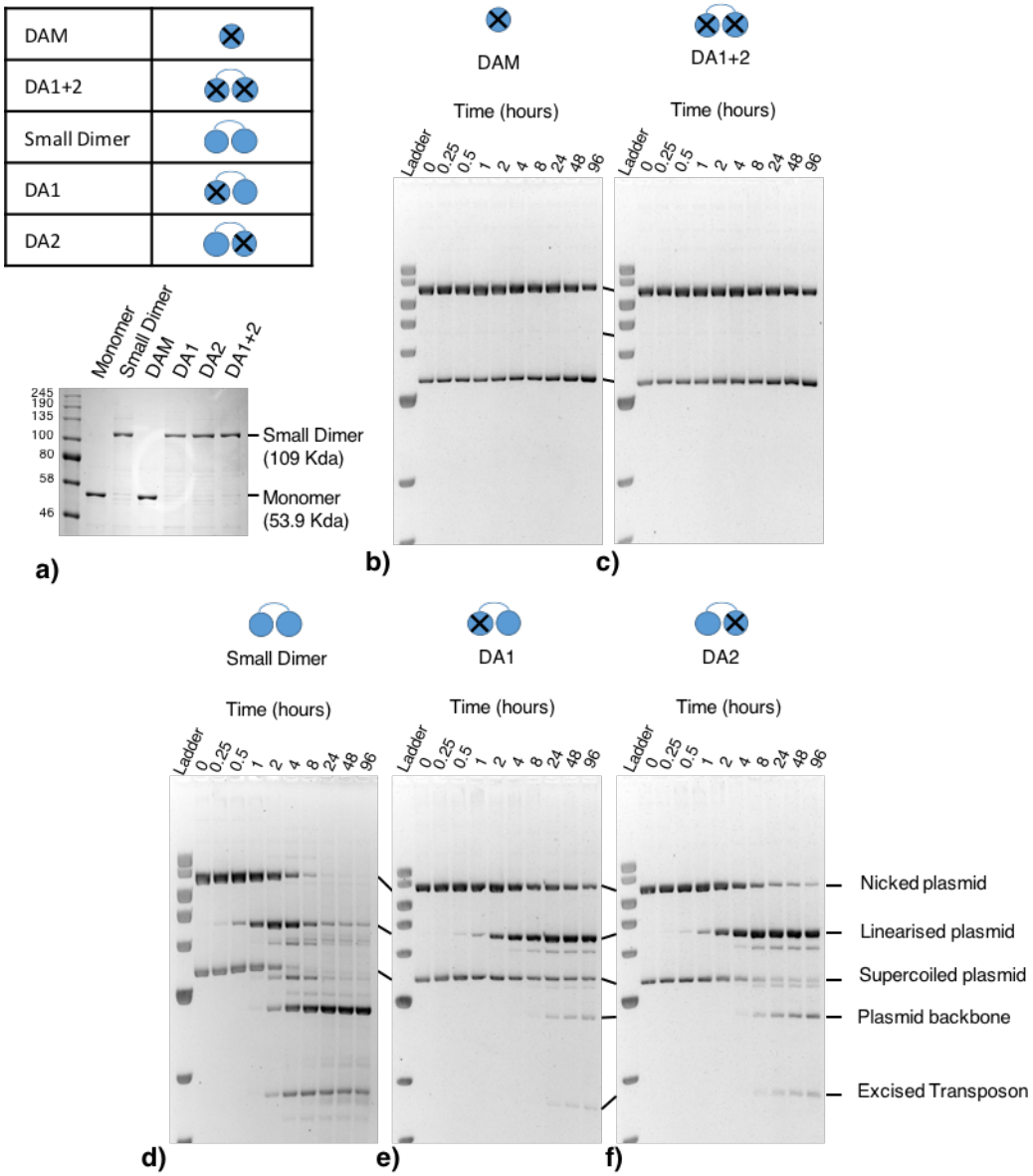


Figure 20: Active Site Mutant Transposase Activity. a) A 10% SDS-PAGE gel of the purified active site mutant transposases, with the functional monomer and small dimer as points of comparison. DAM was derived from pRC2154, DA1 from pRC2143, DA2 from pRC2107 and DA1+2 from pRC2149. Each lane was loaded with 1 μg protein. **b-f)** Transposition assay time courses over 96 hours, run on 1% agarose gels. All assays were done with 7 nM pRC2106 and 100 nM of each transposase. 328 ng of plasmid substrate was loaded per lane. The diagrams above indicate the transposase construct. Each monomer is a blue circle, and the cross indicates the presence of the mutation in that monomer. **b)** The mutant monomeric transposase (DAM). **c)** The small dimer with both transposases mutated (DA1+2). **d)** The fully functional small dimer, the same assay as used in Figure 14. **e)** The small dimer with the first monomer mutated (DA1). **f)** The small dimer with the second monomer mutated (DA2).

To try and determine whether the small dimer formed an SDC or an MDC, the next experiment was to inactivate one of the transposases in the dimer. The hypothesis for this was that if the dimer were only capable of acting as an SDC, then having one inactive transposase in the complex would prevent the full transposition reaction from occurring, whilst an MDC would show full transposition, although at a lower rate. The first mutation tested was D97A, which mutates the first residue of the catalytic triad (see section 1.3). This mutation has been shown to inactivate the transposition reaction, though it does not affect complex formation (Naumann & Reznikoff 2002). Through site-directed mutagenesis, four versions of the mutant transposases were constructed (Figure 20): monomeric (DAM), small dimer with both active sites mutated (DA1+2), small dimer with only the first active site mutated (DA1) and the small dimer with only the second active site mutated (DA2).

Each mutant was purified, using the same intein tag method (see section 2.4.4.1) that was used previously (Figure 20a). 96-hour transposition assays were then run with each of them to identify the level of transposition activity (Figure 20b-f). DAM (Figure 20b) and DA1+2 (Figure 20c) demonstrated no transposition activity throughout the entire assay, confirming that the mutation prevents the transposase from proceeding with the reaction.

DA1 (Figure 20e) and DA2 (Figure 20f) did show some transposition activity. There was a steady increase over time in the linearisation of the plasmid DNA in both cases. This is consistent with an SDC, as one of the active sites in each complex would still be active and can cleave at its TIR

even if the other protein cannot. The rate of conversion of the linearised plasmid to the fully excised transposon and plasmid backbone was greatly decreased, in both cases, compared with rate of linearisation. The plasmid backbone band became visible after eight hours for DA1 and four hours for DA2, compared with the fully functional small dimer (Figure 20d) where this band was seen after one hour. Theoretically, if the pathway is as an SDC, there should be no transposon excision occurring as the mutant protein in the complex cannot cleave. However, if the transposition reaction has stalled at the single end cleavage stage, eventually the complex will disassemble. At this point a new complex would be able to form with the linearised DNA, and this would have a 50% chance of having the functional active site at the uncleaved end. The transposon would then be fully excised from the plasmid backbone.

This experiment did suggest that the cleavage reactions occur in a specific order in the complex, as the rate of linearisation was approximately half the rate of the fully functional dimer. While this rate reduction could be the result of the complex assembly being slowed due to the mutated monomer, previous work has shown complex assembly to be unaffected by the lack of metal ions in the catalytic site (which is similar to the effect of the mutation) (Bischerour & Chalmers 2007). The other explanation is that once the complex has assembled, one of the monomers may have to perform at least the first step of the cleavage reaction before the other monomer, and the mutated monomer has a 50% chance of being in that position. Therefore,

only 50% of the assembled complexes would be able to perform the single end cleavage events.

The evidence shown here supported the hypothesis that the complex forms as SDCs; however, the possibility of MDCs was not ruled out. The reduced rate of the reaction steps could have been due to MDCs forming, as only one of the four possible combinations of the monomers, from two separate dimers, would have been fully functional and able to proceed with the reaction.

3.2.3.3 Dimerisation site mutagenesis of the dimer

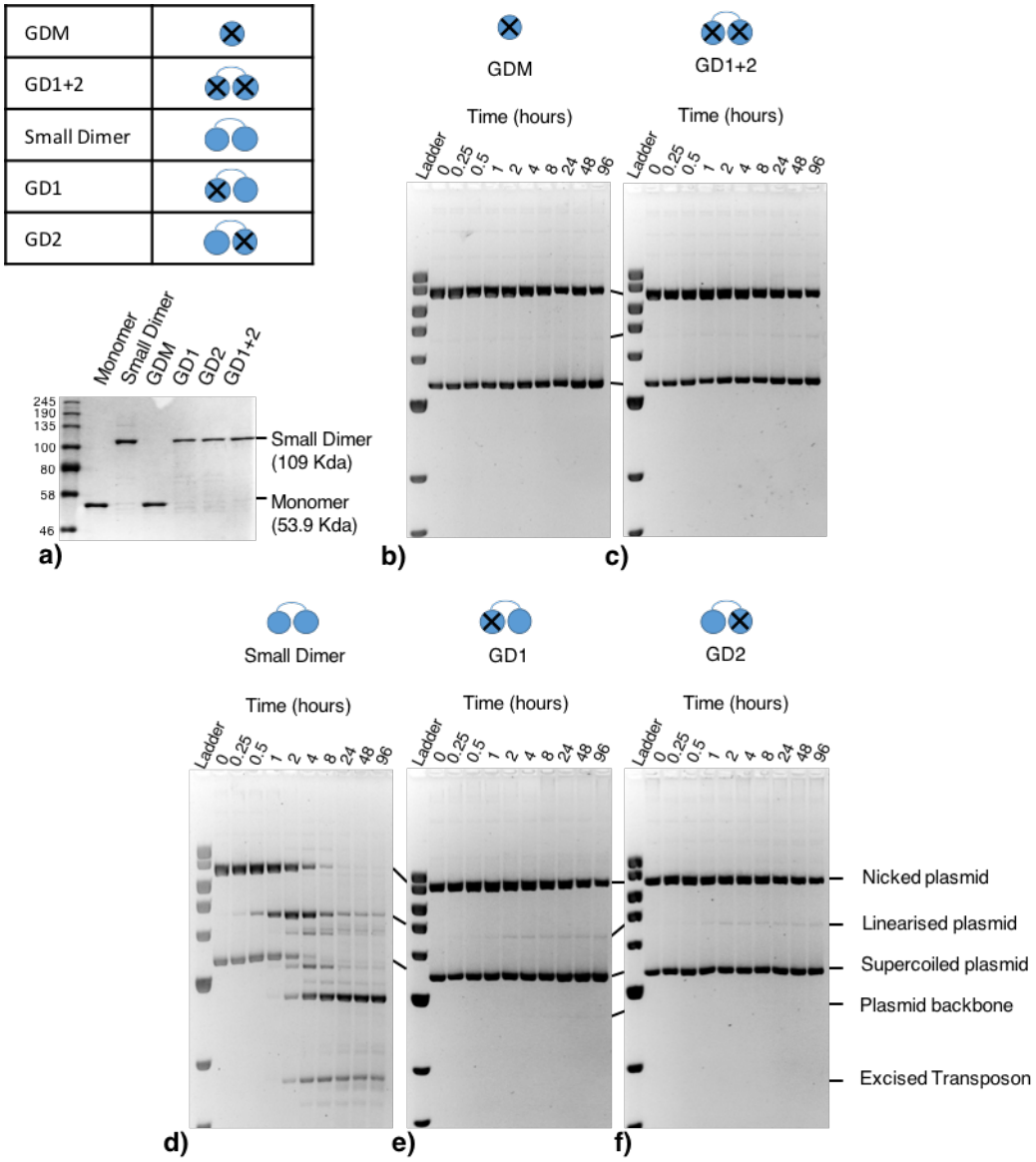


Figure 21: Dimerisation Mutant Transposase Activity. a) A 10% SDS-PAGE gel of the purified active site mutant transposases, with the functional monomer and small dimer as points of comparison. GDM was derived from pRC2165, GD1 from pRC2152, GD2 from pRC2153 and GD1+2 from pRC2166. Each lane was loaded with 1 μg protein. **b-f)** Transposition assay time courses over 96 hours, run on 1% agarose gels. All assays were done with 7 nM pRC2106 and 100 nM of the transposase. 328 ng of plasmid substrate was loaded per lane. The diagrams above indicate the transposase construct. Each monomer is a blue circle, and the cross indicates the presence of the mutation in that monomer. **b)** The mutant monomeric transposase, (GDM). **c)** The small dimer with both transposases mutated (GD1+2). **d)** The fully functional small dimer, the same assay as used in Figure 14. **e)** The small dimer with the first monomer mutated (GD1). **f)** The small dimer with the second monomer mutated (GD2).

The active-site mutagenesis did not prevent the complex formation and therefore a functional active site could still have been capable of reaction steps. This made the results less conclusive. A mutation that inactivated the DNA binding or the dimerisation of the transposase would prevent the PEC from forming, which would allow clearer identification of the complex formation pathway. The G462D mutation is located in the C-terminal dimerisation domain, and it has been shown to prevent synaptic complex formation and, thus, any transposition activity (Steiniger-White & Reznikoff 2000). Using site-directed mutagenesis, four constructs were generated (Figure 21): monomeric mutant transposase (GDM), double-mutant dimer (GD1+2), mutant only in the first monomer of the dimer (GD1) and mutant only in the second monomer of the dimer (GD2). Using the intein tag method that was used previously (see section 2.4.4.1), these proteins were purified (Figure 21a).

A 96-hour transposition assay was run on the mutants to test activity levels (Figure 21b-f). It was clear that GDM (Figure 21b) was completely incapable of transposition activity as there was no change in the bands over the course of the assay. GD1+2 (Figure 21c) was almost identical to the monomeric mutant. The quantity of linear substrate by the end of the assay did show a very slight increase however. This indicated that the mutation may not entirely prevent dimerisation of the transposase, instead significantly reducing the overall strength of the protein-protein interactions. That would mean that complexes could form with this mutation, albeit at a greatly-reduced frequency, and would be more likely to disassemble during

the reaction. In the dimer, the likelihood of complex formation should be greatly increased over the monomeric transposase, and this may counteract a small amount of the negative effect from the mutation, leading to some activity being visible.

The GD1 (Figure 21e) and GD2 (Figure 21f) dimers were largely similar to GDM and GD1+2; over time there was very little change in the bands on the gel. The most noticeable difference was that the linearised band showed up with greater intensity than for GDM and GD1+2, though it was still only a minimal fraction of that seen in the fully-functional dimer (Figure 21d). GD1 demonstrated a slightly greater quantity of linearised plasmid at the end of the assay, compared with GD2. There was also an extremely faint band that corresponded to the plasmid backbone for GD1, which was not visible for GD2, that indicated a small amount of transposon excision. This demonstrated that the two monomers in the small dimer are not equal to one another. It is likely that the effects of having the linker and other transposase attached to either the N-terminus or C-terminus have different effects on the protein.

The extremely low activity that was seen in the GD1 and GD2 assays does indicate that the small dimer was acting as an SDC. If it were acting exclusively as an MDC you would expect that 25% of the time a fully-active complex would form in the GD1 and GD2 assays, therefore the various bands of the reaction steps would have been produced at ~25% of the rate seen in the unmutated dimer (Figure 21d). The actual rate of band production was much lower than this, which produces two possibilities. One possibility

is that the phenotype of the G462D mutation may be partially rescued by the presence of one fully functional monomer in the dimer, which allows a small proportion of transposition reactions to occur. This would correspond with the finding that GD1 and GD2 were more active than GD1+2. The other possibility is that it is capable of forming an MDC a small fraction of the time, which leads to transposition activity being seen.

3.2.3.4 Mutant cleavage and binding

The mutagenesis experiments indicated that the dimer was forming SDCs rather than MDCs. However, this was not entirely conclusive as both the data sets left open the possibility of at least some MDC formation occuring. In order to clarify the data further, analysis of the various complexes formed in the process of transposon excision would be appropriate, through a binding and cleavage assay (Figure 22). This assay was run with the fully-functional monomer and small dimer, as well as all the various mutants from sections 3.2.3.2 and 3.2.3.3.

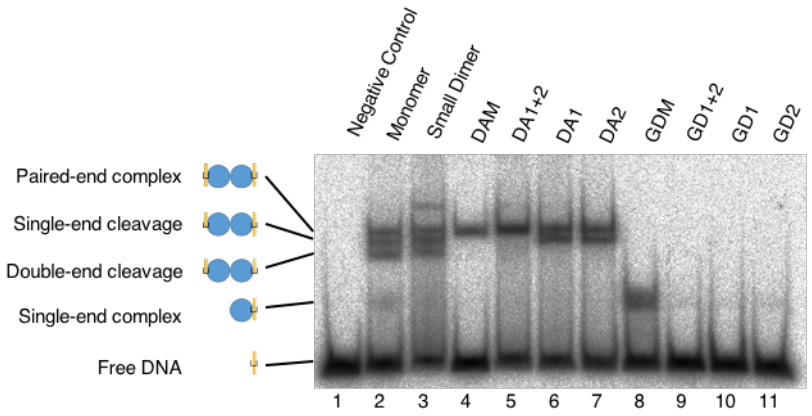


Figure 22: Binding and Cleavage Assay. Samples were incubated for 15 minutes at room temperature then 2 hours at 37 ^{o}C, before running on an 8% TBE gel. The substrate that was used was the short substrate used in Figure 19, labelled using α-^{32}P dCTP. The various different complexes formed are shown on the left. The blue circles indicate the transposase, the orange line indicates the substrate and the black bracket indicates the transposon end sequence.

At the base of the gel the free DNA showed up for every reaction, confirmed by the negative control (lane 1). The monomeric transposase (lane 2) showed four bands above the free DNA. The uppermost band was the full PEC with two uncleaved DNA strands, as this is the largest complex possible. Below this were first the single and then the double-cleaved PECs, as the lower quantity of DNA in the complexes would run faster down the gel. The very faint band below the various PECs was the SEC, which has very transient binding so was not well defined on the gel. The small dimer (lane 3) showed most of the same bands as the monomer, though the SEC was no longer visible. The ‘slow’ PEC appeared in this lane as well, though the reason for this bands prevalence in the small dimer but not the monomer is unknown.

The active site mutants demonstrated the defect in strand cleavage. DAM (lane 4) and DA1+2 (lane 5) were capable of forming the PEC, at a comparable level with the fully functional transposases. They were incapable of producing the two bands that corresponded to the strand cleavage products, as it is believed neither active site in the complex would have been able to coordinate the Mg^{2+} ions necessary for the reaction. Both DA1 (lane 7) and DA2 (lane 8) showed the band for the single end cleavage event, but not the double end cleavage event. This means that in the individual complexes, only one strand could be cleaved. As this assay was over a much shorter time period compared with that of the 96-hour transposition assays in section 3.2.3.2, the complexes did not have time to disassemble and reassemble as was suggested earlier to be the reason for full transposon excision. Therefore, the cleavage seen in those assays was due to the complex collapse and reassembly, and not due to the formation of MDCs.

The dimerisation mutants did not demonstrate any PEC formation or cleavage events with any of the four variants. GDM (lane 9) did present a strong, though still quite broad, band that corresponded to the SEC. This was significantly more intense than in the monomer. Accumulation of the SEC was not due to any increased binding capacity, it was simply because GDM is not capable of forming a PEC, which would have depleted the pool of SECs. The mutant dimers, GD1+2 (lane 9), GD1 (lane 10) and GD2 (lane 11), were only capable of producing a very faint SEC band, though in the binding-only assay using the same proteins the SEC was more clearly defined (Supplementary Figure 2). This supports the hypothesis that the

dimeric transposase is less stable than its monomeric version, at least at the level of the SEC. The cleavage assay data for the G462D mutant dimers was consistent with the results of the 96-hour transposition assays, as the very low level of activity that was seen for the dimeric transposase complexes would suggest an insufficient amount of complex formation to show up on a native gel such as this. If MDCs were forming, it would be expected that there would be a band, approximately one quarter of the intensity, compared with the fully functional versions.

3.2.3.5 Mechanisms conclusion

In order to confirm the initial hypothesis that the dimeric Tn5 transposase mimics the S-NEC pathway, the presence of MDCs had to be ruled out. The work in this section has demonstrated that the SDC complex is not only capable of forming for the small dimer, but is the dominant pathway. It is still possible that the MDC is forming in the reactions a very small fraction of the time, though this is infrequent enough that it should not affect the results seen in other sections.

3.2.4 Pre-formed transposase complexes

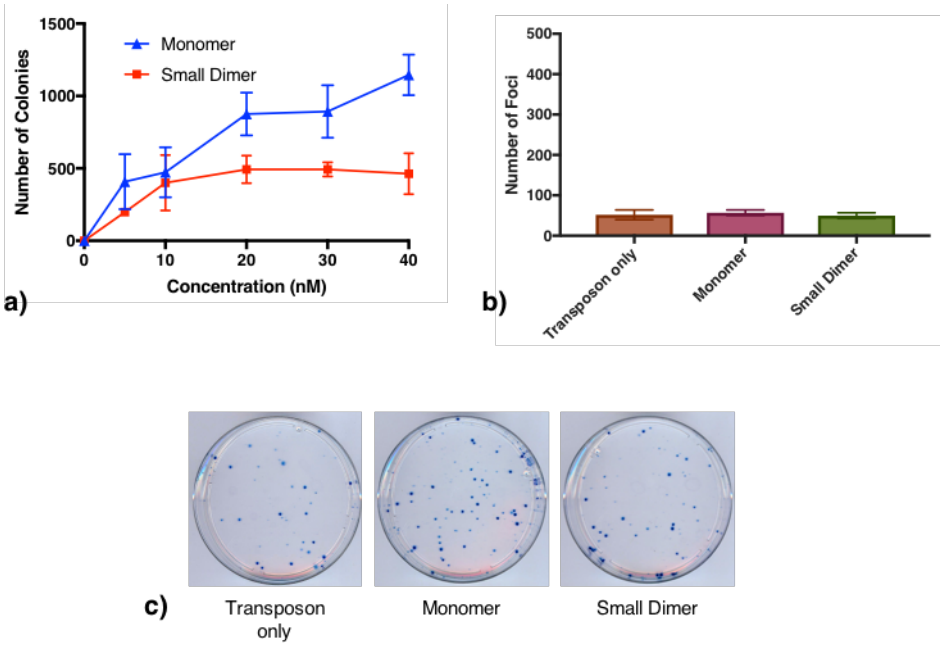


Figure 23: The Activity of Monomeric and Dimeric Pre-Formed Complexes. a) The protein concentration titration for the pre-formed complex integration assay in bacteria. Formation of pre-formed complexes occurred during 1 hour incubation at 37^{o}C, before heat shock transformation into RC5095 cells. Transposon DNA was derived from pRC415 by PvuII digestion. 2.7 nM (40 ng) was used per reaction. The total number of colonies counted was 1/10th of the total number produced per reaction. The total number was estimated from the initial count. Monomer is shown as blue triangles, small dimer is shown as red squares. Statistical test was 2-way ANOVA, n=4. Error bars are standard deviation. **b)** Pre-formed complex integration assay in HeLa cells. Pre-formed complexes were formed in the same manner as for bacteria, but with 40 ng transposon DNA PCR amplified from pRC2106. The cells were transfected with the complexes using Lipofectamine® 2000. The total number of surviving colonies was counted after 2 weeks G418 selection. The statistical test used was a one-way ANOVA, n=3. Error bars are standard deviation. **c)** Example plates from the pre-formed complex integration assay in HeLa cells, post-staining.

The small dimer was shown to be active in human cells, when supplied as a DNA construct and expressed by the cell. The transposase therefore had to form the PEC, excise the transposon and integrate it into the host genome. Purified small dimer protein theoretically might be capable of pre-forming complexes *in vitro*, that could then be transfected into the cells, and the transposon could be integrated into the genome. It is also possible that the monomeric transposase could integrate transposon DNA into the genome through this method. The hypothesis of this section was that the complex assembly pathway is the limiting factor for functionality in eukaryotes, so pre-forming complexes *in vitro* and then supplying them into the cells might be effective. Previously, work on Tn5 transposition has shown that pre-formed complexes were functional in multiple bacterial species and the yeast strain *Saccharomyces cerevisiae* (Goryshin et al. 2000). The Mos1 transposase has also been shown to function when supplied as pre-formed complexes to human cells (Trubitsyna et al. 2017).

The first experiment was to test both the monomer and small dimer in *E. coli*, where monomeric Tn5 pre-formed complexes have been shown to be active previously (Goryshin et al. 2000). The monomer and small dimer of Tn5 transposase were assembled with a set quantity of the pre-cleaved kanamycin resistant transposon substrate, at different concentrations of protein, before being transformed into the bacteria and selected for on agar plates (Figure 23a).

The monomer showed increasing numbers of surviving colonies as the transposase concentration increased. The small dimer however, showed

an increase that peaked at 20 nM transposase, above which it remained relatively consistent. Given the low concentration of cleaved transposon substrate that was used in the assay (2.5 nM), it is possible that there was an overabundance of transposase (above 20 nM) available to bind the DNA, and the presence of more protein did not speed up the complex formation. This is consistent with what was seen in the transposition assay titrations, where the peak rate of transposition at 100 nM was only minimally different from the 80 nM and 120 nM reactions. The small dimer showed fewer transposition events than the monomer at every concentration, though the difference at 5 nM and 10 nM was not statistically significant (P=0.17 and P=0.97 respectively). Given the high activity in papillation assays and *in vitro,* the initial hypothesis for this experiment was that the small dimer would show significantly higher integration of the transposon compared with the monomeric form. The lack of transposition events could be due to potential instability of the small dimer complex. It has to remain complexed through the heat-shock transformation process and then successfully integrate into the genome, so if the PEC was less stable it may not survive as frequently as the PEC of the monomer.

The small dimer was not as active as the monomer in bacteria but was still capable of transposition using pre-formed complexes. The next step was therefore to test pre-formed complexes in human cells. Constructs were made of Tn5 transposase with the SV40 NLS attached (pRC2199 for the monomeric form and pRC2200 for the small dimer form, see section 8.1.2 for full protein sequences), and the protein was purified. Complexes were

then formed *in vitro*, using a Tn5 transposon substrate that contained the neomycin-resistance gene. These were transfected into the HeLa cells following the method used for Mos1 (Trubitsyna et al. 2017). After selection and staining, the numbers of surviving foci clearly indicated that that there was no integration occurring with the monomer (P=0.5517) or the small dimer (P=0.9539) above the level seen without transposase (Figure 22b&c).

In previously published data, this approach worked (with some differences in the methodology) in *S. cerevisiae* (Goryshin et al. 2000), therefore it should have functioned within a higher eukaryote as well, with both monomer and small dimer. It is possible that the process of transfection into the cells cannot accommodate Tn5 pre-formed complexes effectively, so other transfection methods, such as GeneJuice®, X-fect™ or Calcium Phosphate (Jordan 2004), might produce better results. Another possible optimisation to this experiment would be the addition of 5' phosphates to the substrate DNA, as previous work on Tn5 transposition has shown that without that the transposition activity is reduced significantly (Klenchin et al. 2008). This is unlikely to be the only factor inhibiting integration in the current experiment, as a lack of 5' phosphates would have reduced the transposition frequency partially, whereas this assay showed no detectable transposition at all.

3.3.4 Conclusions and future work

The initial hypothesis for this chapter was that the mechanism of complex assembly was a key underlying factor for transposition system functionality in eukaryotes. To this end a single-chain dimer of the Tn5 transposase was produced that is hyperactive in bacterial cells and HeLa cells. This transposase dimer actively mimics the S-NEC pathway, which shows that the pathway of complex assembly is at least one of the major factors influencing the ability of Tn5 in eukaryotes.

The work in eukaryotic systems left open the possibility of genomic instability being a factor in the frequency of integration. If the small dimer were causing DNA double-strand breaks in the genome, the frequency of plasmid integration would increase, and there would be a higher number of surviving foci on the plates. In order to identify if this was a contributory factor, the best experiment would be to mutate or entirely remove the TIRs in the reporter construct, doing each individually and both at once. The DNA integration assays could be run with these, and if the frequency of foci survival was increased even without TIRs being present, then the small dimer would be causing genomic damage and instability.

The stability of small dimer PEC was brought into question several times in the work presented here, though this was not directly tested. A potential experiment to compare the stability of the monomer or small dimer complex would be to allow PEC formation using a labelled substrate. After the PEC has formed, the reaction would be supplemented with an excess of

unlabelled substrate and then samples would be taken over a time course and run on a native-PAGE gel to identify the rate at which the PEC disassembles. The hypothesis for this experiment would be that the band would dissipate faster for the small dimer than for the monomer, which would show that the complex is not as stable.

Chapter 4

Enhancing the Stability of

the Single End Complex

4.1 Introduction

As was discussed in section 1.7, the SEC of the Tn5 transposase is unstable, only becoming stabilised when two SECs combine to form the PEC. This stabilisation occurs because the catalytic domains form multiple interactions with the opposite DNA strand in the complex (see section 1.5.1 for details). The weak DNA-protein interaction in the SEC may be an issue if the transposase is attempting to bind in a eukaryotic genome. Under these circumstances, the transposase would be at a low concentration throughout the nucleus and, as a consequence, SEC formation would be a rare event. As both TIRs have to be bound simultaneously in order for the PEC to form, transposition events would be extremely scarce. If the SEC was much more stable however, the frequency of both TIRs being bound would be significantly increased and transposition may occur at a much higher frequency.

In order to increase the stability of the SEC, there are two possible approaches: mutating the transposase itself to enhance DNA-interactions, or fusing another DNA-binding protein to the transposase. Mutating the transposase itself would be the ideal method, as this would increase DNA-protein interactions with the actual TIR sequence. The E54K mutation has already been shown to improve DNA binding of the transposase, through changing the charge on the residue to a positive one, which can contact the negatively charged DNA better (Davies et al. 2000). As was demonstrated in the previous chapter, the SEC formation with this mutation was still a

relatively weak interaction (compared with the PEC interactions, see section 3.2.3.4). While further mutagenesis is a possible way to optimise DNA binding, this would be unlikely to produce any significant increases to the stability as the best candidates would have been identified through earlier screens (Wiegand & Reznikoff 1992; Zhou & Reznikoff 1997).

The alternative approach for increasing the duration of the SEC would be to fuse the transposase to another DNA-binding protein. This would not alter the binding of the transposase itself, but it would mean that when it dissociates from the TIR, it is kept in very close proximity and will likely bind again. The DNA-binding protein selected for the fusion will be required to form strong interactions with a specific DNA sequence, and these interactions must be maintained over an extended period of time. There are three well-known DNA-binding protein systems that might be possible candidates: Zinc-fingers (Lombardo et al. 2007; Foley et al. 2009), Transcription Activator-Like Effectors (TALE) (Moore et al. 2014), and the nuclease-deficient Cas9 nucleoprotein complex (H. Wang et al. 2016). Both TALEs and the Cas9 complex are relatively large proteins: TALEs require 33-35 aa per base pair they bind to, while the nuclease-deficient Cas9 protein is 1369 aa (Gaj et al. 2013; Qi et al. 2013). The zinc-fingers on the other hand are quite small, only requiring ~30 aa per motif, which can base-specifically bind three base-pairs. Zinc-fingers have also previously been fused to transposases, though only for the purposes of targeting the integration reaction step (K. Voigt et al. 2012). The zinc-fingers therefore represent the best candidate for extending the duration of the SEC.

A zinc-finger is a common type of DNA-binding motif, found across a wide array of proteins (Laity et al. 2001). There are multiple varieties of zinc-fingers but the simplest are those which have a Cys_2-His_2 Zn-coordinating motif. The general structure of these is a short, two-stranded, anti-parallel β-sheet which contains the two cysteine residues, followed by an α-helix where the two histidine residues are located. Together these four key residues coordinate a single Zn^{2+} ion, which provides a structural role in the motif (Pace & Weerapana 2014). Each zinc-finger makes a series of specific contacts with three bases of the DNA, through the N-terminal end of the α-helix, which sits in the major groove of the DNA. Commonly, three of these zinc-fingers are linked together in an array, which forms a C-shaped structure that follows the major groove round the double-helix, binding a total of nine consecutive base-pairs (Pavletich & Pabo 1991; Elrod-Erickson et al. 1996).

This chapter focuses on whether the presence of a Zinc-Finger Array (ZFA) can extend the duration of the SEC by providing additional binding stability, and thereby improve transposition activity. In order to test this, three ZFA-transposase fusions were constructed: the ZFA-monomer, extended ZFA-monomer and ZFA-small dimer (see section 4.2.1 for details). The protein activity of all three was analysed both *in vivo* and *in vitro*, and significant work was done to optimise the system to test if the SEC stability can be enhanced.

4.2 Results

4.2.1 Designing and testing the ZFA fusions

The various fusion proteins were constructed from hyperactive Tn5 transposases and a ZFA. Any Tn5 transposase used contained the E54K, M56A and L372P mutations that were described in section 3.2.1. The ZFA (designated 153086_L) was derived from a pool of custom-built arrays, and had been tested in a bacterial two-hybrid assay and shown to have 11.9 fold activation, which means it bound relatively tightly to its ZF Binding Sequence (ZBS, sequence GGTGATGCT) (Maeder et al. 2008). Three constructs were designed using this array, the ZFA-monomer, the extended ZFA-monomer and the ZFA-small dimer (pRC2170, pRC2187 and pRC2171 respecitvely, Figure 24a). Full protein sequences of the fusions can be found in section 8.1.2. For the ZFA-monomer and the ZFA-small dimer, the ZFA was attached to the N-terminus, using a slightly shortened 20 aa version (SRGGGSEGGGSEGGSGTSEF) of the 22 aa short linker that was used in the small dimer (see chapter 3). In the case of the extended ZFA-monomer, an additional 10 aa was added onto the start of the linker (GGGSGGGSGG…), to provide additional length and flexibility between the proteins.

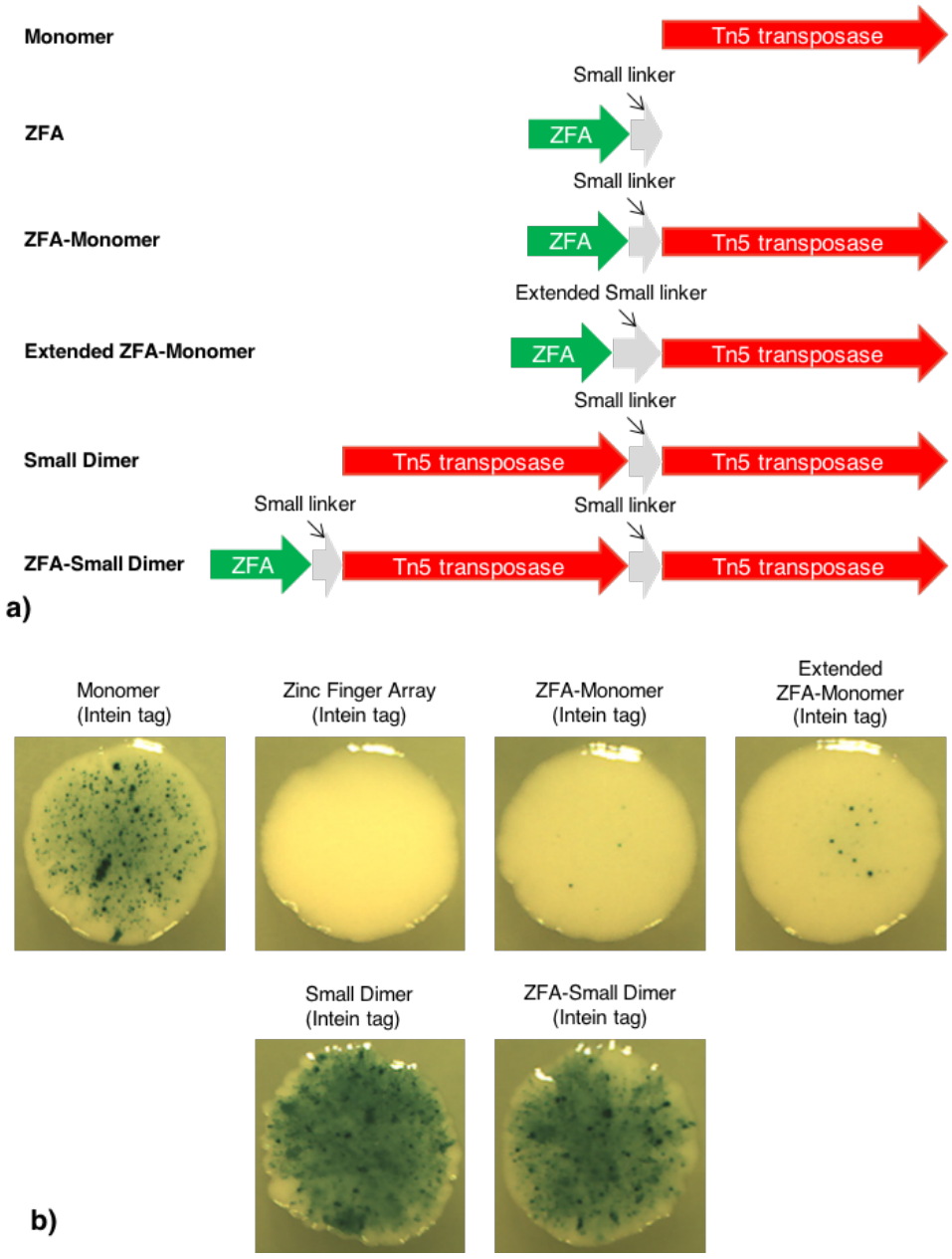


Figure 24: Schematics and Papillation Assay of the ZFA-Fusion Constructs. a) Schematics of the Tn5 transposase Monomer, ZFA, ZFA-Monomer, extended ZFA-Monomer, Small Dimer and ZFA-Small Dimer. The ZFA is shown in green, the small linker in light grey and the transposases in red. All forms also had an intein tag on when not purified, not shown in this schematic. The schematics shown are not to scale. **b)** The Tn5 transposase monomer (pRC2128), zinc finger (pRC2192), zinc finger monomer (pRC2170), extended zinc finger monomer (pRC2187), small dimer (pRC2109) and zinc finger small dimer (pRC2171) were tested using the pTYB2 expression vector. The strain used in this assay was RC5123. The agar plates contained 0.01% lactose as well as X-gal and Ampicillin. The colonies shown are representative of those on the plates. Blue papillae are indicative of LacZ expression.

The first test of activity for these different constructs was a papillation assay, to look at transposition activity *in vivo* (Figure 24b). The papillation assay is explained in more detail in section 3.2.1.1. All constructs tested in this particular assay were in the pTYB2 vector, where the protein of interest is fused to an intein tag, whereas previously the proteins were untagged (for a direct comparison of the transposition activity of the two vectors, see Supplementary Figure 3). From this assay, it was clear that the ZFA alone is not capable of any transposition. The ZFA-monomer showed extremely low activity compared with the unfused monomer control, suggesting the presence of the ZFA inhibited transposition. This inhibition was slightly relieved in the extended ZFA-monomer, though the activity was still very low. The ZFA-small dimer showed very high levels of activity, comparable to the small dimer itself, though at this high level of activity direct comparisons of transposition were much more difficult.

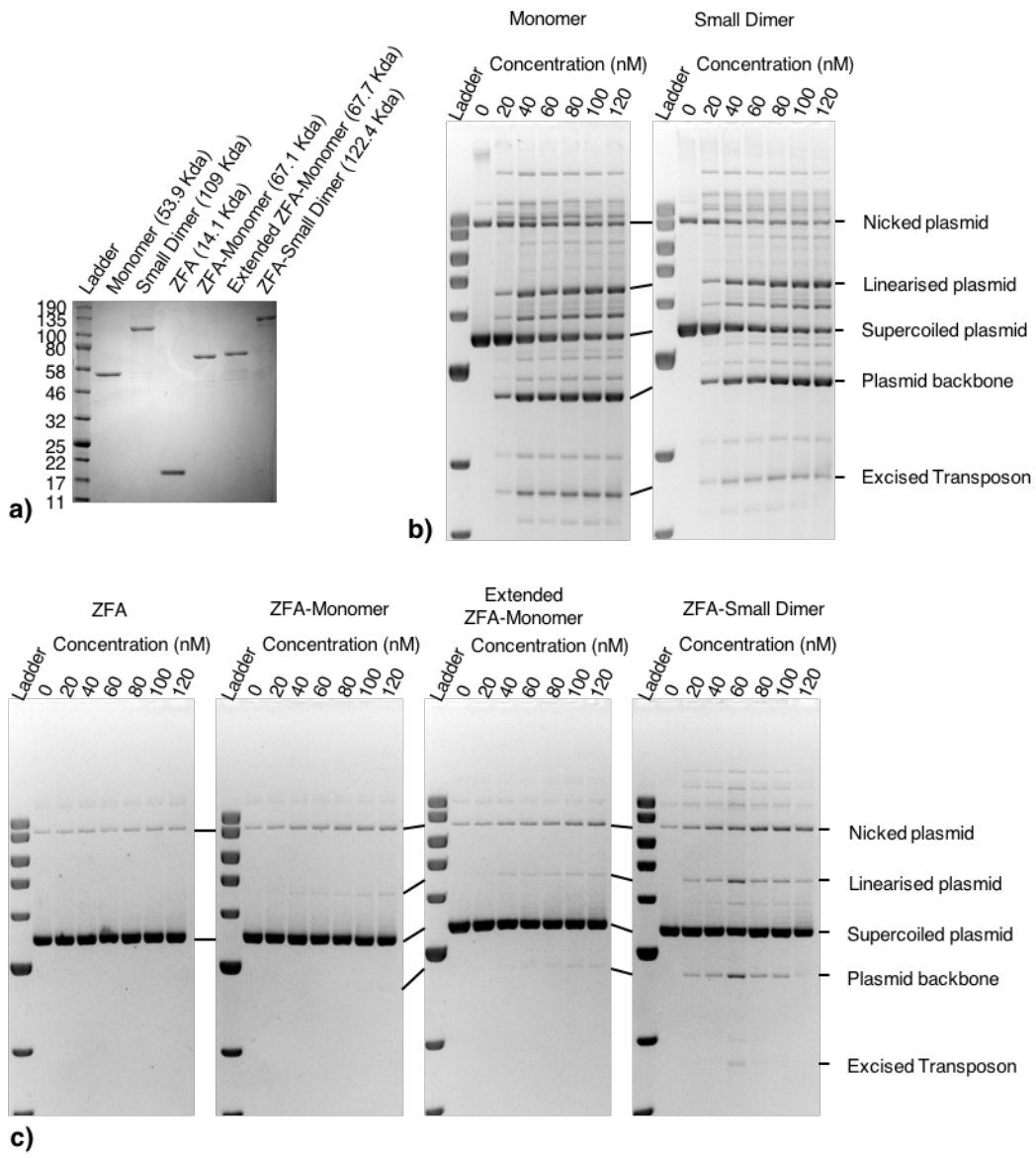


Figure 25: Purification and Titrations of the ZFA-Fusion Proteins. a) 10% Denaturing gel showing the purification of the different transposase constructs. Tn5 transposase monomer was derived from pRC2128, the small dimer from pRC2102, the ZFA from pRC2192, the ZFA-monomer from pRC2170, the extended ZFA-monomer from pRC2187 and the ZFA-small dimer from pRC2171. Each lane was loaded with 1 μg of protein. **b-c)** Transposition assays, titrating the concentration (nM) of the protein. Incubated for 4 hours with 7 nM of the pRC2106 substrate plasmid. 328 ng of plasmid was loaded per lane. The various bands from the transposition reaction are indicated. **b)** The proteins without and ZFA fusion. These assays are the same ones seen in Figure 14e. **c)** The various ZFA containing proteins.

All three of the proteins were overexpressed and purified using the intein tag approach (see section 3.2.1.2 for details, Figure 25a). There was one difference in the technique; 100 μM Zn acetate was supplemented at every step to ensure that the ZFA remained functional. The proteins were then subjected to *in vitro* transposition assays to identify the optimum concentration for each (Figure 25b). The ZFA protein alone was incapable of any transposition activity at any concentration, in line with what was seen in the papillation assay. The ZFA-monomer and extended ZFA-monomer showed very limited activity at all concentrations, though the extended ZFA-monomer was the more active of the two. The optimal concentration (based on the intensity of the linearised transposon and plasmid backbone bands) of transposase was 100 nM for both the monomeric forms, and 60 nM for the ZFA-small dimer. This assay was done using pRC2106, which has no ZBSs, so the presence of the ZFA would not have helped the rate of transposition.

4.2.2 Optimisation of the ZFA binding sites

Although the fusion proteins were largely inactive in the previous experiments, this may have been due to the lack of ZBSs close to the TIR. Therefore, a series of reporter constructs were developed with the 9 bp ZBS present. These reporter plasmids had the ZBS located inside the transposon, at distances of 1-5 bp from the TIR sequence (Figure 26a). The ZBS was oriented so the C-terminus of the ZFA was directed towards the TIR and the transposase, to minimise the length the linker had to bridge.

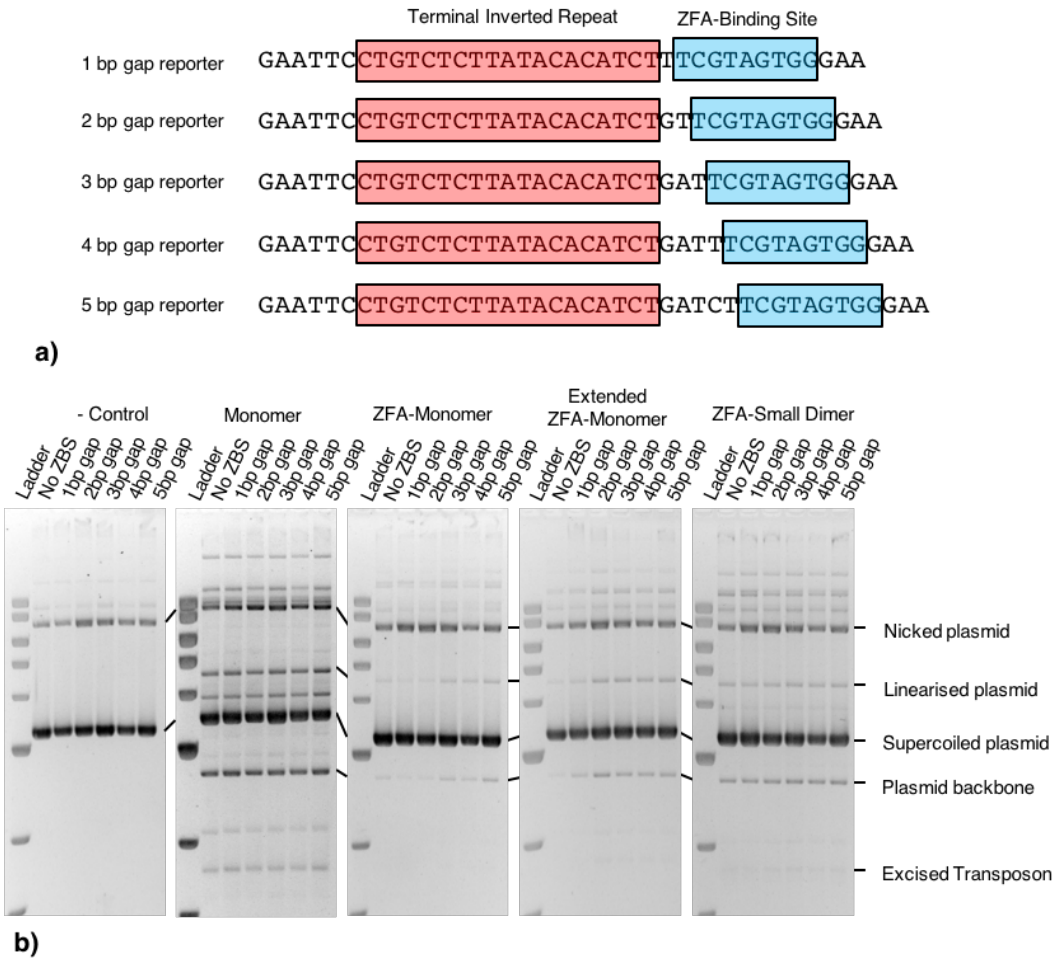


Figure 26: Identification of Optimal ZBS location for the ZFA-Fusion Proteins. a) Diagram showing the different reporter constructs, and how the ZFA-binding site is situated respective to the Terminal Inverted Repeat (TIR, transposase binding site). In each case, both the transposon ends in the reporter plasmid were constructed in the same manner, though, for simplicity, only one is shown here. The transposase binding site is shown in a red box while the ZFA-binding site is shown in a blue box. **b)** Transposition assays demonstrating the effect of the presence of the Zinc-Binding Site (ZBS) on activity. Each 50 μl reaction was incubated for four hours at 37^{o}C. Transposase concentration was 100 nM for all the monomeric transposase forms and 60 nM for the ZFA-Small Dimer. Assays done with 7 nM plasmid substrate, 328 ng of which was loaded per lane. No ZBS = pRC2106, 1 bp gap = pRC2169, 2 bp gap= pRC2179, 3 bp gap = pRC2180, 4 bp gap = pRC2181, 5 bp gap = pRC2182. Negative control was done using protein elution buffer.

The various ZFA-transposase fusion proteins were tested against all these reporter plasmids, using an *in vitro* transposition assay (Figure 26b). The negative control demonstrated that each reporter showed broadly similar levels of the supercoiled and nicked bands, while none demonstrated any linearised plasmid in the absence of transposase. Monomeric transposase showed minimal differences between the different reporter constructs; any slight variations could be explained as minor differences in the quality of the plasmids.

The ZFA-monomer again demonstrated very low activity without a ZBS. In the reactions with the one or two base-pair gap reporter plasmids, there was actually less activity than seen without a ZBS. The activity levels increased when there was at least a 3 bp gap, though the optimal was with the 5 bp gap reporter plasmid. This was still greatly decreased compared with the unfused monomer however, suggesting that activity was still inhibited by the ZFA fusion. The extended ZFA-monomer showed an increase in activity when the ZBS was present, regardless of its location relative to the TIR. Of the various reporter plasmids, the optimal distance was the 2 bp gap, where the rate of excision exceeded any produced by the ZFA-monomer, though it was still significantly lower than the unfused monomer. The ZFA-small dimer presented a higher level of activity than the two ZFA-monomer variants, but showed no variation when the ZBSs were present in the reporter plasmid, suggesting that the ZFA was having little or no effect on transposition.

The data demonstrated that the presence of a ZBS improved the activity of both the ZFA-monomer and the extended ZFA-monomer. It was therefore important to identify the extent to which the activity is enhanced. To this end, 96-hour time courses were run on both the monomeric fusion proteins, both without a ZBS, and with one at the distance shown to be optimal in Figure 25b, 5 bp and 2 bp respectively (Figure 26).

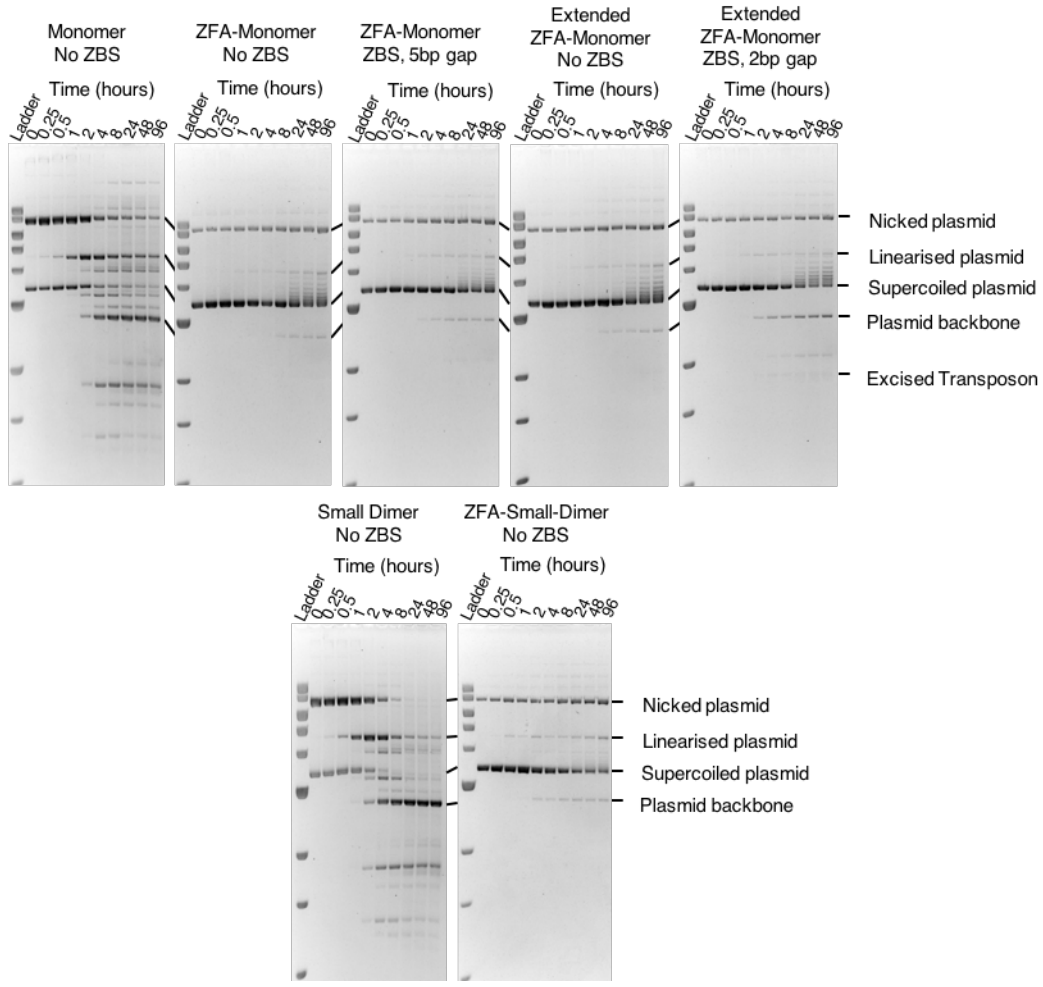


Figure 27: *In Vitro* Transposition assay time courses of the ZFA-Fusion Proteins. All assays were done with 7 nM plasmid substrate, pRC2106 for no ZBS, pRC2179 for 2 bp gap and pRC2181 for 5 bp gap. 328 ng of plasmid substrate loaded per lane. Both the monomer and small dimer experiments were run at a separate time to the other experiments. All proteins were at 100 nM final concentration in the reactions.

The ZFA-monomer activity increased slightly in the presence of the ZBS, separated from the TIR by a 5 bp gap (Figure 27). This is based on the intensity of the plasmid backbone band (which indicates full transposon excision) as it became visible after only two hours, compared with eight hours without a ZBS. While the plasmid backbone band was more intense with a ZBS present, the linear DNA substrate band showed fairly consistent levels of intensity. This suggests that there is a higher frequency of transposition events starting and, of the greater number that started, a greater proportion were fully excised.

In the case of the extended ZFA-monomer the overall rate of transposition, both with and without the ZBS, was higher than was seen for the ZFA-monomer (Figure 27), consistent with the previous data (Figure 26). In the assay with the 2 bp gap reporter construct, the rate of excision was significantly higher than without the ZBS, based on the intensity of the plasmid backbone band. The excised plasmid backbone first appeared at two hours, rather than the four hours that was seen in the absence of the ZBS, and by the 96-hour timepoint it was a significantly more intense band in the reactions with a ZBS. As was seen for the ZFA-monomer, the rate of linearisation showed very little difference with or without the ZBS being present, which suggests that the presence of the ZFA and the ZBS had the same effect for both forms of the ZFA-monomer. The excised transposon band was visible in the reactions with the extended ZFA-monomer in the presence of the ZBS, however, it was a proportionally much fainter band compared with the reactions catalysed by the unfused monomer. This

suggested that once the transposon has been excised it integrates with higher efficiency when the ZFA is fused on.

The ZFA-small dimer fusion protein showed very low activity compared with the unfused small dimer. The rate of substrate consumption of the ZFA-small dimer was significantly less than the unfused protein, with the supercoiled and nicked plasmid bands never being completely consumed, even after 96 hours. The linearised plasmid band was not seen until 30 minutes into the ZFA-small dimer assay, compared with only 15 minutes for the small dimer to achieve the same. Both the linearised and excised transposon band reached a much lower intensity in the assay, which suggested much lower levels of excision overall. This experiment contradicts the data shown in the papillation assay (Figure 24), which showed high levels of transposition occurring at a similar level to that of the unfused small dimer. The frequency of integration is harder to identify accurately in the *in vitro* transposition assay, so it is possible that integration was still occurring to a relatively high level as some bands associated with integration appeared from two hours onwards. As there was no difference between the reporter constructs with and without the ZBS in the previous experiments (Figure 25b), the ZFA-small dimer was not tested with any ZBS-containing reporter constructs.

In all the ZFA-fusion protein assays (but not in any of the unfused transposase assays) there was a distinct series of bands between the supercoiled and linearised plasmids. These bands were very close together near to the supercoiled substrate, but became more spread out as they went

up the gel towards the linear substrate. They appeared from the eight-hour time point and increased in intensity by the end of the assay. This effect was most noticeable in the reaction products of both the ZFA-monomer proteins, though it was also visible in the reactions with the ZFA-small dimer to a lesser extent. A similar effect had not been seen in any other transposition assays done with the Tn5 transposase, so the effect was not related to the transposase itself and must have had some relation to the presence of the ZFA.

In all the ZFA-transposase fusion proteins, the decrease in activity was not specific to a single step in the reaction. This fits with the initial hypothesis, that the loss of transposition activity was due to lower binding affinity of the transposase.

4.2.3 Analysing binding of the fusion proteins

In order to identify what effect the fusion was having on the capacity of the different transposase constructs to bind DNA, binding assays were performed. These assays were also a way of identifying whether fusing the ZFA to the transposase provided additional stability to the SEC. To test the effect of the different components of the fusion constructs, three substrates were used, one with only a Tn5 TIR, one with only a ZBS, and one with both the TIR and ZBS, set up so that the ZBS was 1 bp away from the inside of the TIR (the sequences of these substrates can be seen in section 8.1.3). For this final substrate, the ZBS was oriented so the C-terminus of the ZFA

would face towards the TIR sequence, as was done for the reporter plasmids earlier (see section 4.2.2).

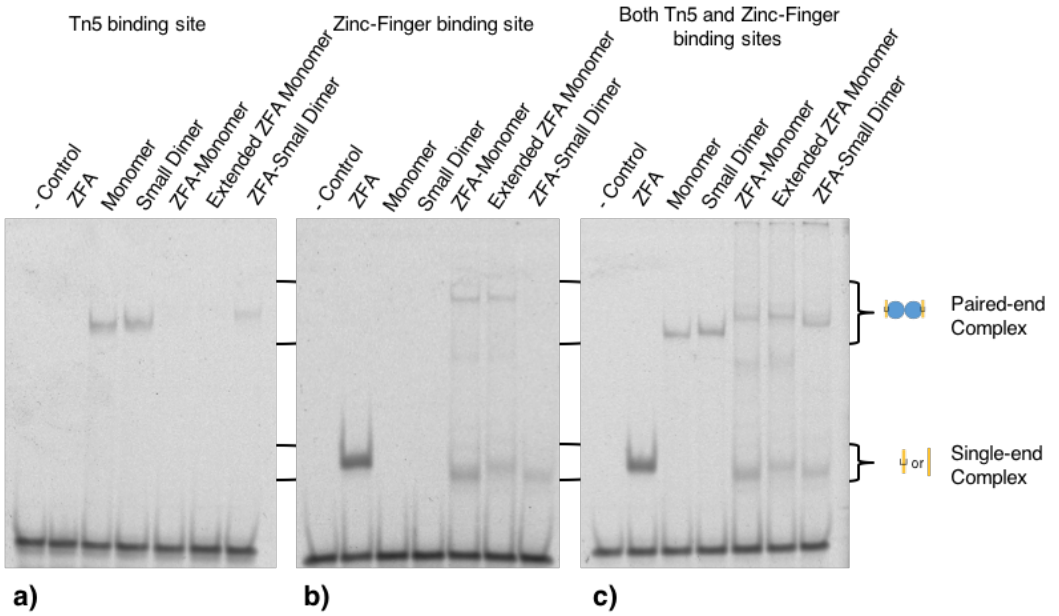


Figure 28: Binding assays of the ZFA-Fusion Proteins. The samples were incubated for 15 minutes at room temperature, before running on an 8% TBE gel. The negative control was the protein elution buffer. The ZFA was used at 100 nM, the Tn5 transposase Monomer at 80 nM, the Small Dimer at 40 nM, the ZFA-Monomer at 80 nM, the Extended ZFA-Monomer at 100 nM and the ZFA-Small Dimer at 40 nM. The different complexes formed are shown on the right. The blue circles indicate the transposase, the orange line indicates the substrate and the black bracket on the substrate indicates the transposon end sequence. All the substrates that were used at 100 nM and were labelled on one strand with a 5' Cy5 fluorescent label. **a)** The substrate that was used was 'Tn5 only', which contained a binding site for the Tn5 Transposase. **b)** The substrate that was used was 'ZBS only', which contained a zinc finger binding site. **c)** The substrate that was used was 'Tn5 and ZBS', which contained a binding site for the Tn5 Transposase and a zinc finger binding site.

When there was only a Tn5 TIR (Figure 28a), the ZFA protein alone was unable to bind, while the monomer and small dimer demonstrated a single band that corresponded to the PEC. Neither the monomer nor the small dimer showed the band that corresponded to the SEC (see Figures 19 & 22).

This may have been due to slightly lower sensitivity when using the fluorescently labelled substrates instead of radioactively labelled. The ZFA-monomer and extended ZFA-monomer showed no binding in this situation, which suggested that the ZFA being present significantly inhibited the initial binding capacity of the transposase, in line with what was seen in Figure 27. The ZFA-small dimer was capable of binding and forming the PEC, although it was a much fainter band than was seen for the unfused small dimer. This pointed to weakened binding. As the second transposase in the dimer is not fused to a ZFA directly (it is bound to another transposase which is bound to the ZFA), it could still have bound sufficiently to be visible in this assay.

When the substrate that only has a ZBS was used (Figure 28b), the ZFA protein was capable of producing a strong band that corresponded to its SEC, as the ZFA protein alone cannot dimerise. The monomer and small dimer alone were not capable of binding this substrate as there was no TIR and the non-specific binding of Tn5 transposase was insufficient to produce any bands under these conditions. Both the ZFA-monomer and extended ZFA-monomer produced two distinct bands and one very faint band. The lowest band likely corresponded to the SEC as it was at a similar size to the band seen for the ZFA protein alone. At the top of the gel the other distinct band was visible. While the transposase had no binding sites for itself, it is capable of binding to DNA non-specifically to some extent (Adams et al. 2006). Therefore, a possible, though untested, explanation is that the ZFA was binding to its site, which placed the transposase in close proximity to the DNA strand where it bound non-specifically. Then another transposase that

had done the same comes into contact with it when both were weakly bound to the DNA and they formed a PEC which was stable enough to be visible on a gel. The very faint band below the PEC is at present unidentified. The ZFA-small dimer shows the SEC band at a slightly lower intensity than the ZFA-monomer versions. The band for the PEC and the band below the PEC were both absent from the gel. This suggested the ZFA-small dimer might have been less capable of forming a PEC with non-specific DNA, or it was just less stable and the complexes became too infrequent to present a band on the gel.

The final substrate tested had the Tn5 TIR and a ZBS, with a 1 bp gap between the two (Figure 28c). As with the other substrates, the ZFA-protein was able to form the SEC, while both the monomer and small dimer could produce only the PEC. SEC formation for all three ZFA fusion proteins was present, at a similar intensity to that seen with the ZBS-only substrate. The two ZFA-monomer variants produced the PEC band, at a slightly higher intensity than seen with the ZFA-only substrate, which suggests that the presence of both binding sites improved the overall DNA binding. This was also the case for the ZFA-small dimer, which showed greater intensity with this final substrate than seen in the substrate with only a Tn5 TIR. The faint band below the PEC for the ZFA-monomer variants was visible, also at a higher intensity, though this band was not present for the ZFA-small dimer, despite the PEC being displayed with this substrate.

4.3 Discussion

4.3.1 The activity of the ZFA fusion proteins

From the data presented here it is apparent that the fusion of a ZFA onto the transposase results in significant inhibition of the transposition activity, both *in vivo* (Figure 24) and *in vitro* (Figures 25-28). From the work in chapter 3, it is clear that fusions to the N-terminus of the transposase are not inherently detrimental as both dimeric transposase constructs were hyperactive. There is a possibility that the ZFA is binding to unintended sequences in these assays, as imperfect binding is known to occur for ZFAs (Pattanayak et al. 2011). In the *in vitro* assays, the reporter plasmid contains no sequences that are identical to the 9 bp ZBS. There are however, two sequences that are 89% (8/9 bp) identical, which could have acted as a weaker binding site for the ZFA. This would have created some competition for the transposase fusion protein that could have led to it being bound at the TIRs less frequently, causing the reduced levels of transposition. In order to rule out this potential issue, mutations to the ZFA could be done that would prevent DNA-protein interactions. These mutants could be tested and any inhibition to the transposition activity would be due to the ZFA-fusion as opposed to off-target binding effects. If the fusion of the ZFA to the Tn5 transposase N-terminus is inherently detrimental, then a potential alternative would be to fuse it to the C-terminus instead. The Tn5 transposase C-terminus is located only 28.9Å from the N-terminus and is on the outside of the protein, close to

the DNA helix (based on PDB: 1MUH). The papillation assays shown here (Figure 24 and Supplementary Figure 3) indicated that the proteins were still functional when the intein tag was attached to the C-terminus, so a fusion at this end of the transposase might be viable.

In Figure 27, there were a series of unidentified bands, particularly for the ZFA-monomer and extended ZFA-monomer, between where the supercoiled and linear substrate could be seen. There are two possible explanations for these bands. The first is that the ZFA was not being unbound from the DNA by the addition of SDS and the heat shock at 68 ^{o}C, and therefore the supercoiled DNA that it was bound to ran higher than the unbound DNA, a phenomenon seen with the p53 protein (Jagelská et al. 2008). The other possibility is that there was some topoisomerase activity occurring within the assay, possibly due to a low-level contaminant protein that had been purified along with the various fusions (Nitiss et al. 2001). At present, neither of these hypotheses has been tested, though testing the fusion proteins with a plasmid that lacks TIRs or ZBSs, such as pBS, would isolate the possibilities, as if the bands are seen then it must be topoisomerase activity that occurs as the ZFA is not binding.

4.3.2 The location of the ZBS and its importance to the activity

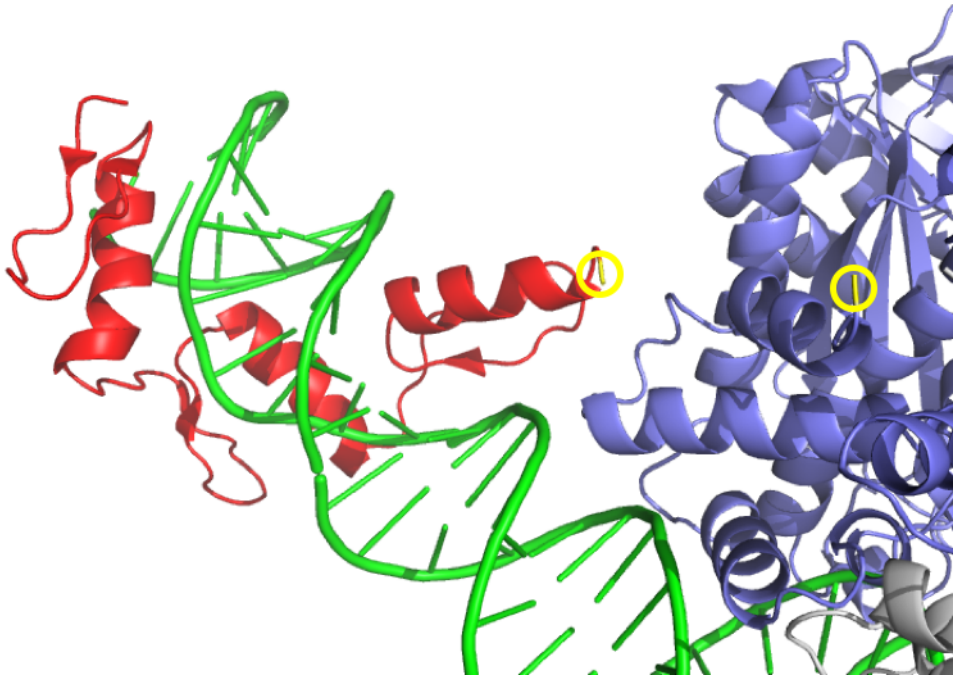


Figure 29: Model Demonstrating the Location of the Zinc Finger Respective to the Tn5 Transposase. The Tn5 transposase is shown in blue, the Zinc Finger Array in red, the DNA in green and the two residues that would be at either end of the linker in yellow with a yellow circle surrounding them. This model places the ZBS with a 1 bp gap from the Tn5 TIR. The model is based on two separate crystal structures. The Tn5 transposase post-cleavage complex is 1MUH (Davies et al. 2000), and the zinc finger array comes from the Zif268 crystal structure 1AAY (Elrod-Erickson et al. 1996).

As can be seen in Figures 26 & 27, the location of the ZBS relative to the TIR was important for the activity levels of the fusion proteins. The further away the ZBS is, the less impact it will have on the reaction, as the transposase will not be held in such close proximity to the TIR, or may be out of range to bind entirely. However, if it is too close to the TIR, the ZFA

may interfere with the spatial requirements of transposase binding, and instead prevent rather than enhance binding.

In order to get a better understanding of where the transposase and the ZFA would sit respective to each other, a model was created (Figure 29) using crystal structures of the Tn5 transposase (Davies et al. 2000) and the Zif268 ZFA. The Zif268 ZFA was used in the model as it is of a similar size to the ZFA used in the experimental work, and has a crystal structure solved in the presence of its own binding site (Pavletich & Pabo 1991; Elrod-Erickson et al. 1996). This model shows the approximate location of the ZFA if the ZBS was positioned with a 1 bp gap between it and the end of the TIR. This position puts the C-terminus of the ZFA in close proximity to the transposase once both are bound. If the ZFA does bind before the transposase, this proximity may inhibit the transposases capacity to bind to the TIR, and this could explain the very low activity for the ZFA-monomer with the 1 bp gap reporter (Figure 26b). The extended ZFA-monomer did show greater transposition activity with the 1 bp gap reporter plasmid, so it is possible that the additional length of the linker allowed the transposase sufficient maneuverability to bind even while the ZFA was in such close proximity.

The optimal distances for the ZFA-monomer and the extended ZFA-monomer fusion proteins were 5 bp and 2 bp from the TIR, respectively. For the ZFA-monomer the 5 bp gap positions the ZFA C-terminus (where the linker connects) on the opposite side of the DNA strand to the transposase. This may position the ZFA at the furthest distance from the transposase that

is still close enough to allow both proteins to bind (though this distance would still be within the maximum distance that the linker can stretch). It is also possible that this distance puts the ZBS sufficiently far away from the TIR that if the transposase binds first, the ZFA cannot bind at all and does not inhibit the reaction through binding to DNA. Exactly what is occurring is difficult to isolate, though testing reporter plasmids with the ZBS even further from the TIR might provide some insight. If the activity of the transposase remains at the level seen for the 5 bp gap reporter, even at distances that would definitely be beyond what the linker would be capable of binding to, then the increase in activity would be due to transposase binding preventing the ZFA binding as a result of the distance.

For the extended ZFA-monomer, the optimal distance between the sites was 2 bp, which places the C-terminus of the ZFA slightly 'below' the transposase (assuming that the image for Figure 29 is taken from the 'top'). This may be positioning the ZFA far enough from the transposase that both can occupy their sites simultaneously. The increased linker length likely serves to improve overall manoeuvrability so the transposase can access the TIR, which may explain why the transposition is more active regardless of where the ZBS is located. While it was possible to improve activity by relocating the ZBS, neither ZFA-monomer variant was able to replicate more than a small proportion of the activity of the unfused monomer. It is possible that further variation of the length of the linker, and possibly a greater range of gap size, might yield increased activity of the transposase, above the levels seen for the unfused monomer.

4.3.3 SEC binding stability

The initial hypothesis of this work was that the presence of a ZFA could improve the overall stability of the SEC. In the binding assays (Figure 28), this hypothesis was tested. The presence of a ZBS in the substrates produced a band that corresponded to the SEC. This band occurs regardless of the presence of a TIR sequence in the substrate. It is possible that the binding could be from the ZFA alone, and not the transposase being stabilised at the site. The substrate used in Figure 28c has the ZBS located only 1 bp away from the TIR, which was not optimal for either fusion construct (Figure 26b). Testing the binding capacity when the optimal binding distances were used would provide some clarification on the potential SEC stability. In order to quantify SEC formation, the G462D mutation could be used in the fusion proteins (see sections 3.2.3.3 and 3.2.3.4) which would prevent PEC formation. Then only the SEC would form, which would make direct comparisons easier, as well as further clarifying which bands are associated with the dimerisation capacity. From this, a test of the stability of the SEC could then be performed, by allowing the SEC to form with labelled substrate for both the unfused monomer and the two monomeric transposase fusion constructs, then supplying an excess of the unlabelled substrate and observing the rate at which the SEC band dissipates. If the SEC is does have increased stability for the ZFA-monomer variants, the band would dissipate at a slower rate compared with the band formed by the unfused monomer.

In all the experiments performed for this chapter, the transposase was present at high relative concentrations compared with what would be present in human cells. To test low concentrations, a papillation assay could be done where the reporter transposon contains ZBSs, at the optimal distances for the fusion proteins. Expression levels could be kept low by using plasmids that exist at only five-copy or one-copy in *E. coli*, instead of the ~20 copy ones used. At low protein concentration, the theoretically stable SEC should provide a significant advantage, allowing time for the second TIR to be bound, therefore increased activity would be seen compared with unfused monomer.

4.3.3 Conclusions

The focus of this chapter was to improve the stability of the SEC of the Tn5 transposase, to test the hypothesis that this was a significant limitation for functional transposition in eukaryotes. This was approached through the use of a Tn5 transposase fusion with a ZFA domain, that would bind to a site adjacent to the TIR and prevent the transposase dissociating and moving away from its binding site. The fusions created were significantly less active than the original unfused transposase, although, with the ZBS at specific distances, some activity was restored. Therefore, this suggests that the fusion is more active when the ZFA binds close to the TIR. Further work may provide hyperactive variants compared with the unfused variant. These variants would then be viable to test the original hypothesis: whether SEC stability is a limitation of transposition in higher eukaryotes.

Chapter 5

Adapting the Tn5 Transposase to Bind Chromatin

5.1 Introduction

One of the most significant differences between the genomes of bacteria and eukaryotes is chromatin packaging. Eukaryotic cells have much of their DNA wrapped around an octameric core of histone proteins, in what is known as a nucleosome. An additional histone 1 protein binds to the complex, at which point it is known as a chromatosome, and this will bind ~165 bp of DNA. This is the form that compacted heterochromatin DNA takes when packaged (McGinty & Tan 2015). The presence of histones and their various modifications presents a challenge for a transposase, as they will have to bind to a specific sequence, which may be partially or wholly inaccessible.

The presence of heterochromatin is not necessarily an inhibiting factor for some eukaryotic transposases. Members of the ITm superfamily are able to function with chromatin present, although they do not have a dedicated chromatin-binding domain. For those elements with a single transposase binding site per transposon end (Tc1, *Himar* and *Hsmar1*), the chromatin has been shown not to affect the overall transposition activity of the protein. Those elements that had more than one transposase binding site per transposon end (Sleeping Beauty, Frog Prince and Minos) actually show enhanced transposition activity when the DNA is packaged as chromatin (Jursch et al. 2013). The model for this enhanced activity states that the two binding sites are located in very close proximity when the DNA is condensed in heterochromatin, which promotes the synaptic complex formation.

Some transposition systems have a specific domain that is capable of binding to histone modifications and can thereby aid function in eukaryotic systems. This domain is called a zinc-binding Plant HomeoDomain (PHD) finger, and is found in the V(D)J recombination protein RAG2 and the *piggyBac* transposase from *Trichoplusia ni*. PHD fingers are between 50 and 80 aa in length, with a core two-stranded anti-parallel β-sheet, in some cases with an additional α-helix at the C-terminus. These domains coordinate two Zn^{2+} ions between a Cys_4-His-Cys_3 motif, which stabilizes the domain (Bienz 2006; Sanchez & M.-M. Zhou 2011).

RAG2 is an accessory protein to the actual transposase, RAG1, but RAG2 forms part of the synaptic complex. This PHD finger has been shown to bind to histone 3, when its lysine 4 site is di- or tri-methylated (H3K4me2/me3) (Matthews et al. 2007; Liu et al. 2007), and is involved in efficient V(D)J recombination *in vivo*. The *piggyBac* transposase has a predicted PHD domain in the C-terminus of the protein. Removal of part, or all, of this domain produced no overall reduction in transposition activity when there was an absence of chromatin on the DNA. This suggests that this domain may only be necessary in the presence of chromatin, though the effect of its removal with chromatin present is currently unknown (Mitra et al. 2008). The exact binding sites of this PHD domain are also unknown at present, as is its exact size within the protein.

The Tn5 transposase evolved to function in bacterial genomes, where no chromatin is present. If the DNA that the transposase will bind to is wrapped up by histones, then Tn5 does not seem able to interact with it (Sos

et al. 2016). The aim of the work described in this chapter is to ascertain whether the transposase's inability to bind to its sites in chromatin might be a significant limitation for its activity in human cells, or if transposition can function in spite of these limitations. In order to test this, a *piggyBac* PHD domain was fused onto the Tn5 C-terminus, and its activity was investigated.

5.2 Results

5.2.1 Functional analysis of the PHD fusions

Based on the general sizes of the PHD fingers in other proteins, the domain in the *piggyBac* transposase could be up to 80 aa, although all the conserved residues are located within the C-terminal 50 residues (Mitra et al. 2008). Structural predictions of the *piggyBac* protein, using the JPRED prediction software (Drozdetskiy et al. 2015), showed the C-terminus to have three β-strands within it (in which all the conserved residues are located). Preceding that was a 30-40 aa stretch with no secondary structures predicted. The section prior to the region without secondary structure likely belongs to the catalytic core domain, as it was a large region with many α-helices predicted, and the conserved DDD(D) catalytic residues were located within it. The 80 aa size prediction placed the start of the PHD domain within the last predicted α-helix.

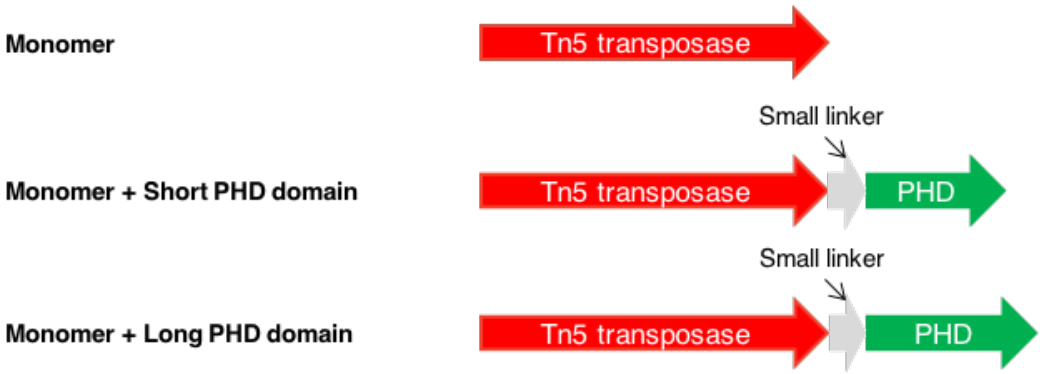


Figure 30: Schematics of the fusion proteins of Tn5 transposase and PHD domains. Schematics of the Tn5 transposase Monomer, the Monomer fused to the short (82 aa) PHD domain and the Monomer fused to the long (90 aa) PHD domain. The Tn5 transposases are shown in red, the small linker in light grey and the PHD domain in green. All forms also had an intein tag attached to the C-terminus prior to protein purification, not shown in this schematic. The schematics shown are not to scale.

Two transposase-PHD domain fusion constructs were designed (Figure 30). Each construct had a hyperactive Tn5 transposase with the E54K, M56A and L372P mutations, for details on the exact function of these mutations, see section 3.2.1.1. On the C-terminus of the transposase, there was the same 22 aa small linker (GSSRGGGSEGGGSEGGSGTSEF) used in chapter 3, which had one of the two versions of the PHD domain attached at the end. The two PHD domains were either 82 or 90 aa in length. These are referred to as the short and long PHD domains, respectively (full protein sequences can be found in section 8.1.2). The large amount of the *piggyBac* transposase that was taken is likely more than the necessary amount for the domain, but ensured that the PHD finger is extracted in full, including the unstructured sections that are known to form a part of it (Sanchez & M.-M. Zhou 2011).

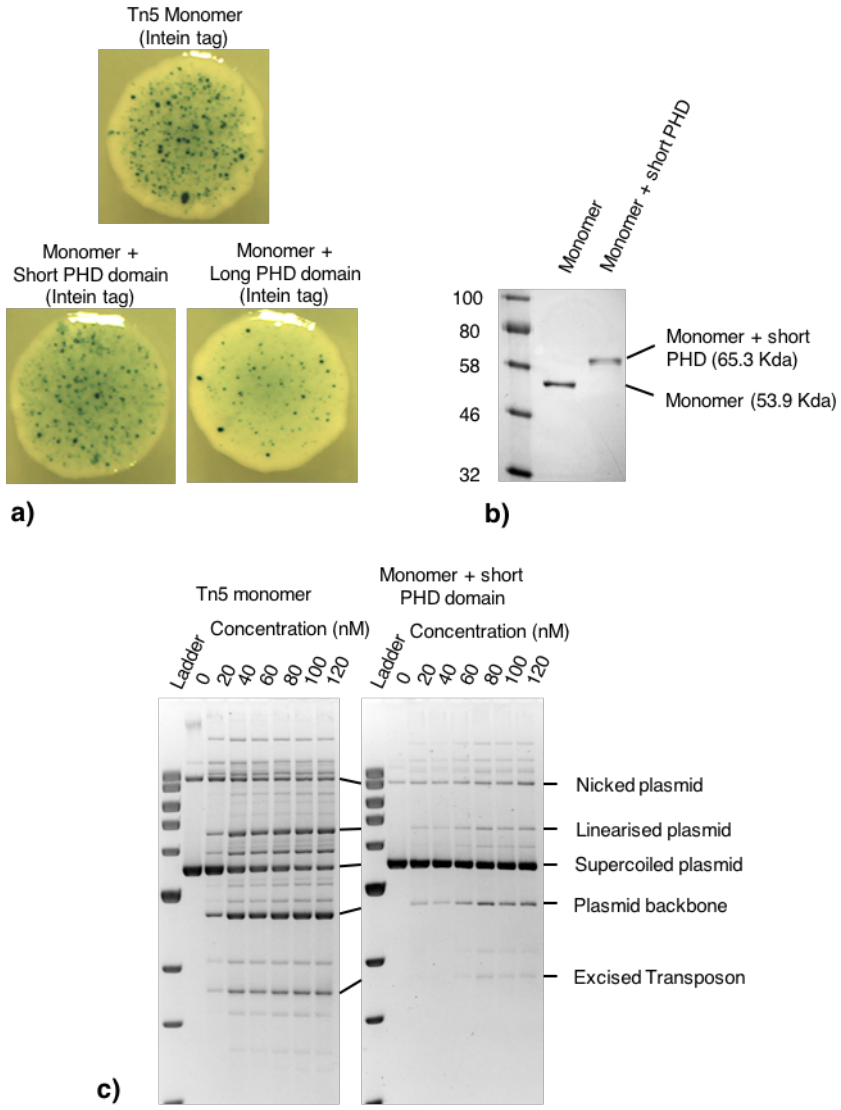


Figure 31: Initial Analysis of the PHD Fusion Proteins. a) Papillation assay of the Tn5 transposase monomer (pRC2128), monomer + short PHD domain (pRC2506) and monomer + short PHD domain (pRC2507). These plasmid constructs are all derived from the pTYB2 vector. The strain used in this assay was RC5123. The agar plates contained 0.01% lactose as well as X-gal and ampicillin. The colonies shown are representative of those on the plates. Blue papillae are indicative of LacZ expression. **b)** 10% denaturing gel showing the purification of the different transposase constructs. Monomer was derived from pRC2128 and the monomer + short PHD domain was derived from pRC2506. Each lane was loaded with 1 μg of protein per lane. **c)** Transposition assays, titrating the concentration (nM) of the protein. Incubated for 4 hours with 7 nM pRC2106 substrate plasmid. 328 ng of plasmid was loaded per lane. The various bands from the transposition reaction are indicated. These two gels were run at different times.

Initially, a papillation assay was run on the PHD fusion proteins, to determine their transposition activity (Figure 31a). The papillation assay is explained in detail in section 3.2.1.1. All constructs tested in this particular assay were in the pTYB2 vector, where the protein of interest is fused to an intein tag, which does alter the activity levels slightly; for details on the effects, see section 4.2.1 and Supplementary Figure 3. This assay demonstrated that the short PHD domain fusion did not significantly alter the transposition activity *in vivo*, though a small variation in activity would be difficult to detect. The long PHD domain appeared to affect the overall activity more significantly, where it reduced activity by at least half (exact quantification is not possible given the qualitative nature of the assay).

As the papillation assay showed that both proteins were functional in bacteria, the proteins needed to be purified in order to characterise the activity levels further. When purification was attempted for the long PHD protein, the intein tag did not successfully cleave off of the protein, so it was unable to be purified. It is possible that the domain interfered with the self-cleavage action of the intein tag, though this possibility has not been investigated further. The short PHD fusion protein was able to be eluted however (Figure 31b).

The purified short PHD fusion was subjected to a transposition assay, titrating the protein to test for optimal concentration, with each reaction run for four hours (Figure 31c). The results showed that its optimal concentration was 80 nM, similar to the unfused transposase. The overall activity of the protein appeared to be much lower than for the unfused transposase. It is

possible that the fused PHD domain is interfering with either the DNA/protein interactions or specific steps in the reactions. Even if the PHD domain is functioning, there was an absence of any chromatin to bind to in this assay. Therefore it would be unable to assist the transposase binding in any form, consistent with what was seen with the *piggyBac* transposase when tested *in vitro* (Mitra et al. 2008)

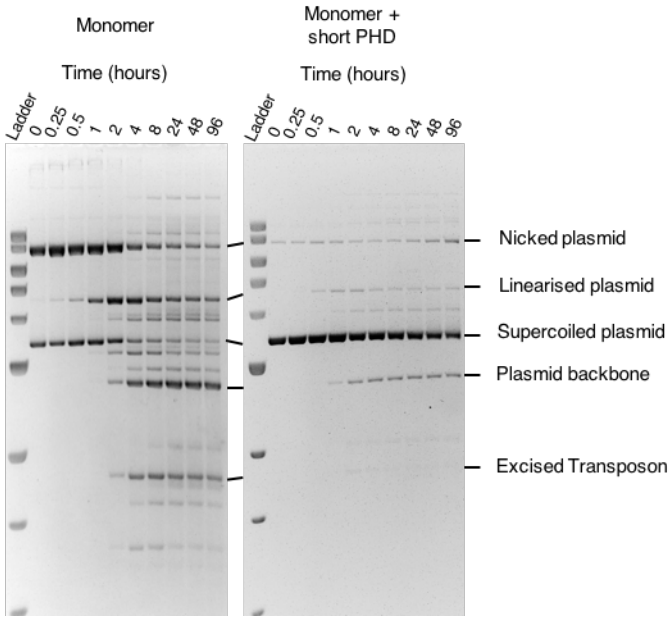


Figure 32: *In Vitro* Transposition assay Time Course of the Short PHD-Fusion Protein. All assays done with 7 nM pRC2106. 328 ng of plasmid substrate loaded per lane. These assays were run at separate times to each other. The Tn5 transposase Monomer was at 100 nM concentration and the Monomer + short PHD domain was at 80 nM.

To investigate what effect the PHD domain was having on the activity of the transposase, a 96-hour time-course transposition assay was performed in order to identify the rates of various steps of the transposition reactions (Figure 32). The two assays shown here were run at different times, so the variation in the quantity of nicked plasmid is due to different levels of plasmid

degradation. Consistent with what was seen in the transposition titration assay (Figure 31c), the fusion protein was significantly less active than its unfused counterpart. The overall rate of the reaction for the fused protein appeared consistent with the unfused protein. For example, it produced a peak of linearised substrate around the two-hour mark, though this peak was a lot fainter than for the unfused transposase monomer. The rate at which the supercoiled band was being consumed was a lot slower as well, as the reaction products with the unfused transposase had almost entirely lost that band at the 96 hour time point, whereas it was still a strong band at that point in the products of the fusion protein.

The only band in the transposition assay that showed a difference in intensity when comparted to the unfused protein was the plasmid backbone. The plasmid backbone band could be seen earlier in the reaction for the fused protein, after one hour, compared with two hours for the unfused protein. Therefore, the presence of the PHD domain could be somehow beneficial to the rate of second strand cleavage, though the overall amount of transposition that occurred was still much lower than the unfused protein. The rate of integration was noticeably lower for the fused protein (based on the intensity of the group of bands above the nicked substrate band), which was not consistent with what was seen in the papillation assay (Figure 31a). This suggests either that the papillation assay was not sensitive enough to show the difference in transposition for these proteins, or that the PHD domain fusion functioned better *in vivo* than it does *in vitro*.

5.2.2 Testing of the PHD fusion proteins in human cells

The PHD domain should only benefit the transposase when there is chromatin to bind to (assuming that the predicted PHD domain does target methylated histones). To test the functionality with chromatin present, new constructs were designed to test the transposase-PHD fusion in human cells. All the proteins (the monomer, pRC2148, short PHD fusion, pRC2508 and the long PHD fusion, pRC2509) were codon-optimised to ensure optimal expression of the proteins. The fusion proteins also lacked the small linker region used in the bacterial and *in vitro* work. All the constructs had an SV40 NLS attached at the N-terminus, in the same manner as was done for the Tn5 dimer (see section 3.2.1). As a control for human expression, constructs were designed fusing the PHD domain onto the Hsmar1 transposase (pRC2506 and pRC2507). Full sequences of the proteins can be found in section 8.1.2. The Hsmar1 transposase is native to human cells, and has been shown to be capable of effective transposition within them. Therefore, the addition of the PHD domain may show more clearly whether the PHD domain is having any impact on the transposition activity.

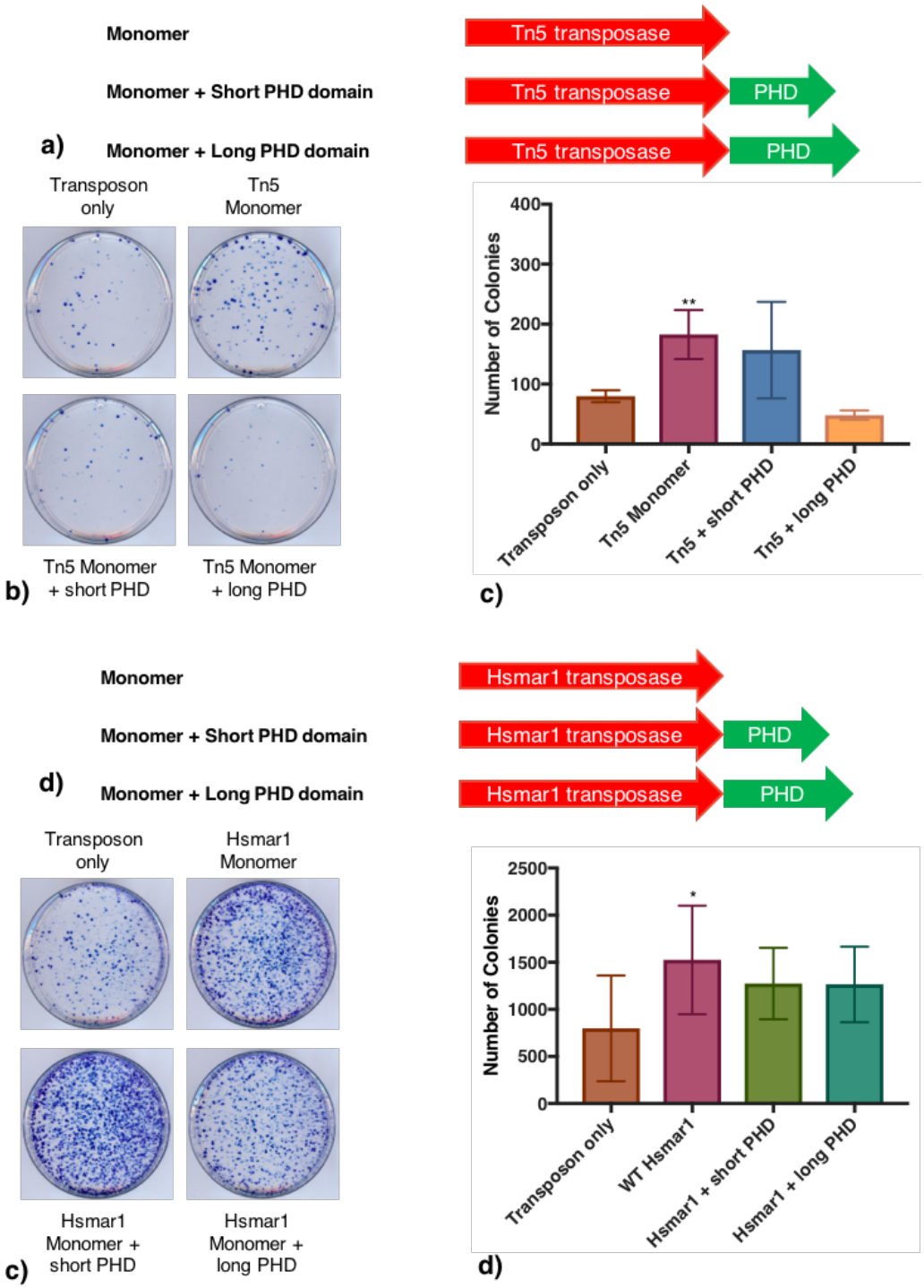

Monomer
Monomer + Short PHD domain
a)
Monomer + Long PHD domain
Transposon only
Tn5 Monomer
Tn5 Monomer + short PHD
Tn5 Monomer + long PHD
b)
Tn5 transposase
Tn5 transposase
PHD
Tn5 transposase
PHD
Number of Colonies
400
300
200
100
0
**
Transposon only
Tn5 Monomer
Tn5 + short PHD
Tn5 + long PHD
c)
Monomer
Monomer + Short PHD domain
d)
Monomer + Long PHD domain
Transposon only
Hsmar1 Monomer
Hsmar1 Monomer + short PHD
Hsmar1 Monomer + long PHD
c)
Hsmar1 transposase
Hsmar1 transposase
PHD
Hsmar1 transposase
PHD
Number of Colonies
2500
2000
1500
1000
500
0
*
Transposon only
WT Hsmar1
Hsmar1 + short PHD
Hsmar1 + long PHD
d)

Figure 33.

Figure 33: Transposition Activity of the PHD Fusion Proteins in HeLa Cells. a & d) The transposases are shown in red and the PHD domain in green. The schematics shown are not to scale. **a)** Schematics of the Tn5 transposase monomer, the Tn5 transposase monomer fused to the short (82 aa) PHD domain and the Tn5 transposase monomer fused to the long (90 aa) PHD domain. **d)** Schematics of the Hsmar1 transposase monomer, the Hsmar1 transposase monomer fused to the short (82 aa) PHD domain and the Hsmar1 transposase monomer fused to the long (90 aa) PHD domain. **b & e)** Human cells transposition assay plates, after 2 weeks of selection with G418. Cells were transfected with 1 μg of reporter plasmid (pRC2194) and 1 μg of transposase expression plasmid. $4x10^5$ cells were loaded per plate at the start of the selection. **b)** Plates with Tn5 transposition system used. The reporter plasmid was pRC2194, the Tn5 monomer expression plasmid was pRC2148, the short PHD fusion expression plasmid was pRC2508, and the long PHD fusion expression plasmid was pRC2509. **e)** Plates with the Hsmar1 transposition system used. The reporter plasmid that was used was pRC1240, the Hsmar1 monomer expression plasmid was pRC1241, the short PHD fusion expression plasmid was pRC2510 and the long PHD fusion expression plasmid was pRC2511. **c & f)** The frequency of cell survival under the different transposase constructs. The graph depicts the percentage increase over the transposon-only baseline. * means $P<0.05$, ** means $P<0.01$. Statistical tests were ratio paired T tests. n=4. **c)** A graph depicting the frequency of integration of the Tn5 transposon. **f)** A graph depicting the frequency of integration of the Hsmar1 transposon.

Transposition integration assays were run on the both the Tn5 and Hsmar1 variants. The assay uses a neomycin resistance gene that is located between two TIRs on a plasmid. This is transfected along with a transposase expression plasmid, and the neomycin resistance is selected for. Surviving foci are quantified, and increases in the number of foci indicate transposition activity (see section 2.4.3.5 for method details).

In the assays using the Tn5 transposase variants (Figure 33a-c), the unfused monomeric transposase (~183 per plate) showed increased activity over transposon only (~80 per plate, ~227% increase), which was

statistically significant, P=0.0062. When the short PHD domain is fused, the activity level decreased; it only showed a 152% increase over the transposon-only control (~157 per plate), which was not statistically significant, P=0.5545. The long PHD domain fusion actually showed a decrease in integration frequency, although this was not significant (~48 per plate, P=0.0539), compared with the transposon only control.

The Hsmar1 assay (Figures 33d-f) showed a higher level of background integration than in the Tn5 assay with ~10-fold more foci (~798 per plate). The WT Hsmar1 showed a significant increase over the transposon only control, ~340% more (~1525 per plate). This increase was statistically significant, P=0.0341. The short and long PHD fusions to the Hsmar1 transposase demonstrated a percentage increase over the transposon only plates, of ~318% (~1274 per plate), and ~315% (~1265 per plate) respectively. These averages were only slightly below the WT Hsmar1, though both demonstrated a much higher sample variation and were not statistically significant over the transposon-only control (P=0.0939 and 0.0967 respectively). This suggests that the fusion of a PHD domain was detrimental to the Hsmar1 transposase to some degree, though not as significant as was seen for the Tn5 transposase.

5.3 Discussion

The focus of this chapter was to ascertain whether the inability to bind to chromatin is a severe limitation for the activity of the Tn5 transposase in eukaryotes. To achieve this, the transposase was fused with a predicted chromatin-binding PHD domain from a eukaryotic *piggyBac* element. This fusion protein showed significantly inhibited activity, particularly *in vitro*.

From the evidence seen with the work in both bacterial and human cells (Figures 31a & 33), the long PHD domain acts as a stronger inhibiting factor for the transposase, compared with its short counterpart. The additional eight amino acids, NLYMSLTS, may form some unwanted secondary structure, particularly given that region is predicted to form an α-helix in *piggyBac*. If additional secondary structure forms around the C-terminus of the transposase, it may impact its ability to function. This issue is particularly pertinent to the fusion in human cells, where the fusion proteins did not have the small linker in order to separate the transposase and the PHD domain. Of the additional eight amino acids, three are hydrophobic (two leucines and a methionine), which may have the effect of positioning the start of the PHD domain closer to the transposase where it could cause inhibitory effects.

The short PHD domain fused to the transposase demonstrates better transposition function than the long PHD domain, both in the presence and absence of the small linker. The transposase itself was still significantly less active than unmodified transposase however (Figure 31c and 32). These

data suggest that the frequency of transposition reactions is significantly reduced rather than any single step of the reaction being inhibited, which implies that the PHD domain is interfering with the DNA-binding or complex formation steps that precede any transposition reaction steps. To test this hypothesis further, binding assays could be run on both the unfused Tn5 transposase and the short PHD fusion protein to identify any inhibition in the DNA-protein interactions, either at the level of SEC formation or PEC formation (for details on these complexes see chapter 3). If the SEC band is significantly stronger, and the PEC band weaker, then this would suggest that the PHD domain is interfering with protein-protein interactions which are required for the PEC formation step of the reaction. However, if the PEC and SEC were both reduced, by an approximately equal amount, then the PHD domain is likely interfering with the DNA-protein interactions. This would be because the initial DNA binding (which forms the SEC) is reduced which then reduces the number of available SECs that can form PECs.

The important outstanding questions of this chapter are what function the PHD domain actually plays in *piggyBac* transposition and whether the fused domain is actually functional at all once attached to the Tn5 transposase. As the domain is located on the C-terminus of the protein, it is likely that it binds to either flanking or target DNA. This is because, in all transposases with an RNase-H catalytic core that have been characterised to date, those DNA-binding domains that interact with sites in the transposon itself are located N-terminal of the catalytic core (Davies et al. 2000; J. M. Richardson et al. 2009; Montaño et al. 2012; Hickman et al. 2014). The best

way to start analysing the function of the PHD domains directly however would be to test it in the *piggyBac* transposase itself. One initial experiment to test the impact of the PHD domain would be to remove it from *piggyBac*, and then analyse the level of transposition in human cells, to see if it was essential under those circumstances. The transposase could be purified using methods described by Mitra et al. 2008, and then tested using a Histone Peptide Microarray (Matthews et al. 2007; Rothbart et al. 2012). These provide a wide selection of histone modifications, to which the transposase could bind, and antibodies are used to identify which are bound. The Tn5-PHD domain fusions could also be tested using the same system to identify whether the domain was still functioning in the same manner as in *piggyBac*. It would be possible to isolate the exact size of the PHD domain using this technique as well, as small sections of the *piggyBac* C-terminus could be tested until the minimum size for activity was found. This would help to avoid any unnecessary side effects. This might also improve the functionality of the whole fusion protein, allowing the underlying theory to be better investigated.

In the human cell assays used here the transposon would, at least initially, not be packaged in chromatosomes, as it is supplied in a bacterial plasmid. This means that the only limitation the chromatin can provide is the available target sites for transposition, which would not be such a strong inhibition for Tn5 due to its relatively low target site preference (Reznikoff 2008). Integration of a transposon into the genome would allow chromatin to form on the transposon DNA itself, which could inhibit early stages of the

reaction. This could potentially demonstrate different results and may make the PHD domain more beneficial to the transposase.

The initial hypothesis of this chapter was that chromatin might provide a significant limitation on Tn5 transposition. The evidence in this chapter suggests that chromatin is not a significant limitation for the Tn5 transposase, at least for the integration steps of the reaction. More work needs to be done in order to make this evidence more conclusive. This result does fit with the currently published data, which shows that Tn5 transposition happens only into sites that do not have any chromatin (Buenrostro et al. 2013). With more data, it may also be possible to get a more complete understanding on both the original hypothesis and why *piggyBac* needs the dedicated chromatin binding domain.

Chapter 6

The Transposition Activity of PGBD5

6.1 Introduction

Over millennia many transposons have been domesticated by their host organism. A large number of these domesticated transposases lead to changes in gene regulation, though some have been adapted for more complex roles (Britten 1996). Perhaps the most well-known domestications are RAG1 and SETMAR. RAG1 is believed to be descended from the *Transib* transposon family (Kapitonov & Jurka 2005; Kapitonov & Koonin 2015) and forms a key component of V(D)J recombination. SETMAR is related to the Hsmar1 transposase from the *mariner* transposon family, and is involved in histone methylation and DNA repair pathways (Shaheen et al. 2010).

One of the families of domesticated transposons in *H. sapiens* are the five PGBD (*piggyBac* derived) elements. To date only two of these elements have been examined, PGBD3 and PGBD5. PGBD3 is found in intron five of the Cockayne syndrome B (CSB) gene, where it inserted itself ~43 Mya. The transposase now expresses as a fusion to a truncated form of the CSB protein itself. This fusion is speculated to be involved in gene regulation or possibly some defence against transposons (Newman et al. 2008).

PGBD5 is the product of a significantly more ancient domestication event, ~500 Mya, that predates vertebrates, and has been highly conserved in genomes since (Pavelitz et al. 2013). The PGBD5 gene is found in chromosome one, consisting of seven exons spread across ~112,000 bp. This is distinct from the other PGBD proteins which are enclosed in a single

exon for the length of the transposase, though the proteins they are fused to often have multiple exons. The exact function of PGBD5 is not known, though it was believed to be inactive for transposition as, in sequence alignments with other *piggyBac* superfamily members, the catalytic triad was mostly altered (Pavelitz et al. 2013).

A recent paper, 'Genomic DNA transposition induced by human PGBD5' (Henssen et al. 2015), investigated the function of PGBD5. It demonstrated that a GFP-PGBD5 fusion was capable of transposition in HEK293 cells, and was able to do this with the Terminal Inverted Repeats (TIRs) derived from *T. ni*, the source of the first identified *piggyBac* transposon (Cary et al. 1989). Transposon ends are normally highly specific to the transposase, so active transposition after divergence of at least 500 million years is unexpected and warrants further investigation.

This chapter focuses on a repeat of the experimental work from the 2015 paper that demonstrated transposition activity for the PGBD5 protein. To do this, the GFP-PGDB5 fusion protein was tested in both HEK293 cells (the cell line used by the original paper) and HeLa cells (an alternative cell line that is commonly used). The expression and localisation were tested, as well as the transposition activity of the protein, compared with the original *piggyBac* transposase.

6.2 Results

6.2.1 Functionality in HEK293 cells

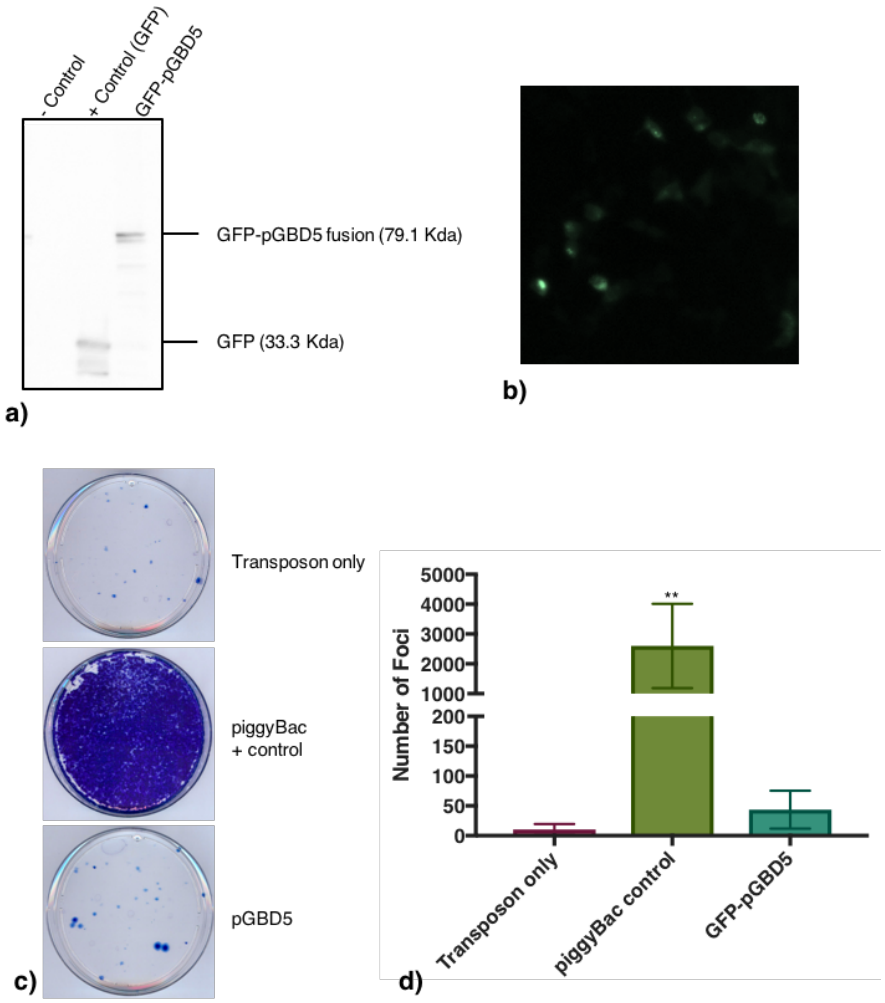


Figure 34: PGBD5 experiments on Hek293 cells. a) Western blot of the expressed proteins. Cell extracts were taken after 24 hours, and 30 μg was loaded per lane. The primary antibody was anti-GFP. Sizes were determined by comparing with the ladder (not shown). Cells were transfected with the following plasmids: Transposon only had pBS, *PiggyBac* control was pRC2116 and GFP-PGBD5 was pRecLV103-GFP-PGBD5. **b)** Fluorescence microscopy done on live cells 24 hours after transfection, using cells transfected with pRecLV103-GFP-PGBD5. The fluorescence level was too low to show a visible light combined with fluorescence image. **c)** Human cells transposition assay plates, after 2 weeks of selection with G418. Cells were transfected with 1 μg reporter plasmid (pRC2501) and 1 μg transposase expression plasmid (pBS for transposon only control, pRC1542 for *piggyBac* control and pRecLV103-GFP-PGBD5 for GFP-PGBD5). $4x10^5$ cells were loaded per plate at the start of the selection. **d)** The frequency of cell survival under the different conditions. The graph depicts number of foci per 100 mm^2 plate. ** means $P<0.01$. Statistical test was a one-way ANOVA. n=3.

The first step in assessing the reproducibility of the experimental work from the paper was to ensure that the GFP-PGBD5 was expressed in the cells. To ensure consistency, the expression construct used was the same as that in the paper (pRecLV103-GFP-PGBD5). A western blot was used to identify the expression of the fusion protein, along with the GFP protein alone as a control (Figure 34a). This showed expression of both the GFP and the GFP-PGBD5 fusion protein, to similar levels, with some degradation products produced for both. Unfortunately, in this western blot there was no direct way to compare with the expression levels seen in the paper, due to the variability in antibodies and detection equipment.

PGBD5 was shown to localise to the nucleus in previous work, so it was important to ensure that the fusion construct was doing the same (Pavelitz et al. 2013). The fluorescence level was quite low so it was impossible to show the rest of the cell using visible light, but GFP expression could be clearly seen localising to the nucleus of the cells (Figure 34b). There was a small amount of fluorescence seen in the rest of the cell. This could be due to the fusion protein degrading, as there was a faint band that corresponds to GFP alone (33.3 Kda) in the western blot (Figure 34a).

The most important experiment to replicate was the transposition integration assay. The reporter construct used here (pRC2501) was not identical to the one used in the paper; however, the transposon end sequences were identical throughout the known transposase binding sites of *piggyBac*, the TIR and the STBS (Yusa 2015). In the paper, the transposition activity assay seeded only 1000 cells per plate but when this

was done only the *piggyBac* positive control plates showed any foci at all. Instead, $4x10^5$ cells were seeded out per plate which showed colonies on all the plates (Figure 34c). Foci were counted and the number of surviving foci are shown (Figure 34d).

The transposon-only plates showed very low foci numbers, averaging ~10 per plate. As a positive control, the *piggyBac* transposase was extremely active, averaging 2600 foci per plate, and had strong statistical significance over transposon only, P=0.0047. According to the data from the paper, the GFP-PGBD5 fusion protein should show activity that equates to ~4.5-fold less than the *piggyBac* control, ~578 foci per plate. The GFP-PGBD5 plates only showed an average of ~44 foci per plate however, which equated to ~60-fold decrease from the *piggyBac* control. When compared to the transposon only, the GFP-PGBD5 still demonstrated a 4.4-fold increase, though this was not statistically significant (P=0.2818).

6.2.2 Functionality in HeLa cells

While the function of PGBD5 in HEK293 cells compared directly with the work done in the paper, it was also important to identify the activity in other cell lines, as a point of comparison. The HeLa cell line is perhaps the best known human cell line used in research, and is straightforward to work with, so therefore provided a viable alternative for testing.

Before any experimental work could be done, it was important to identify any endogenous expression of PGBD5 in HeLa cells, as the HEK293 cells were not reported to have any (Henssen et al. 2015). Any endogenous expression of PGBD5 might affect the transposition activity seen for HeLa cells, as transposition could potentially occur in the transposon-only control plates. HeLa cells are derived from a cervical carcinoma. While normal cervical cells were known to have relatively high transcription of PGBD5 compared with other tissues, (~29 TPM ,transcripts per million), HeLa cells have differentiated significantly from them and demonstrated only ~0.1 TPM (Uhlén et al. 2015). This provided strong evidence that PGBD5 was not endogenously expressed in HeLa cells, which was reinforced by RNA-Seq data from a collaborator which suggested that what RNA is expressed was then degraded by the cell prior to translation (Nojima et al. pers comm).

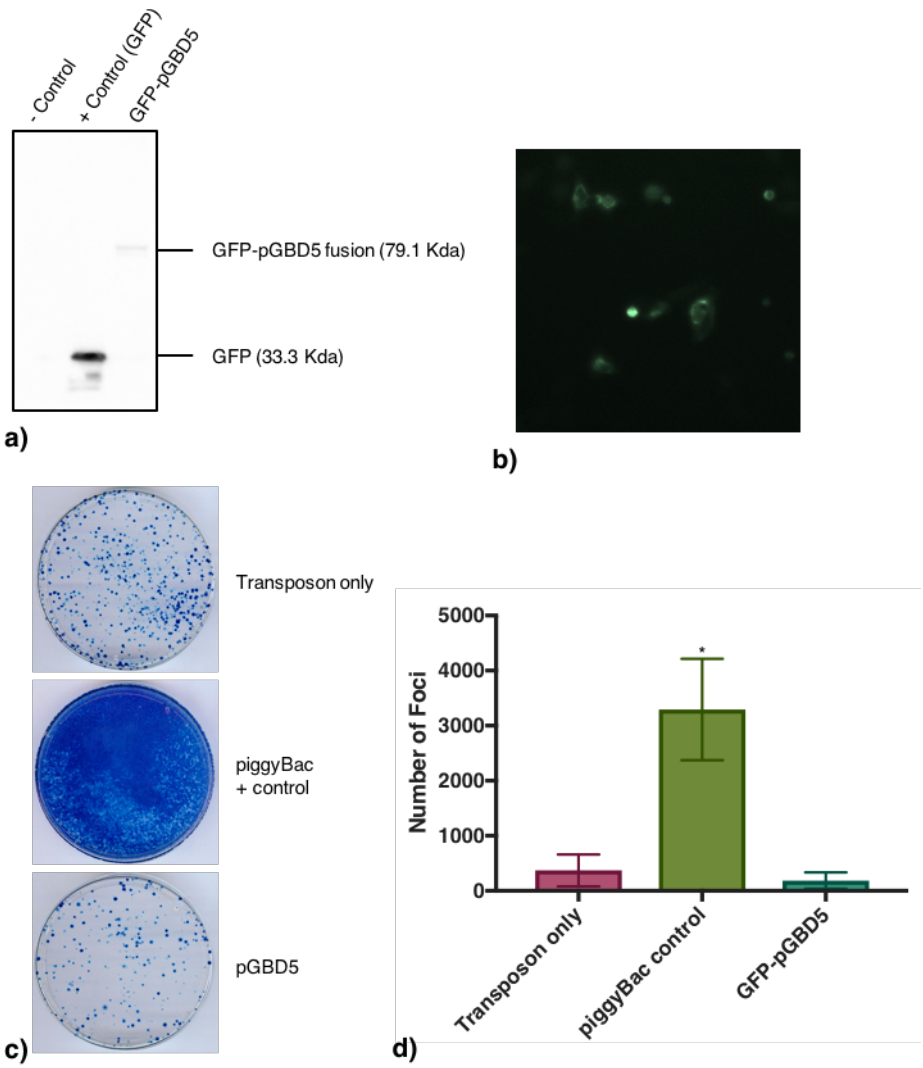


Figure 35: PGBD5 experiments on HeLa cells. a) Western blot of the expressed proteins. Cell extracts were taken after 24 hours, and 30 μg was loaded per lane. The primary antibody was anti-GFP. Sizes were determined by comparing with the ladder (not shown). Cells were transfected with the following plasmids: Negative control was pBS, positive control was pRC2116 and GFP-PGBD5 was pRecLV103-GFP-PGBD5. **b)** The fluorescence microscopy done on live cells 24 hours after transfection, using cells transfected with pRecLV103-GFP-PGBD5. The fluorescence of the cells was too low to show a visible light combined with fluorescence image. **c)** Human cells transposition assay plates, after 2 weeks of selection with G418. Cells were transfected with 1 μg reporter plasmid (pRC2501) and 1 μg transposase expression plasmid (pBS for transposon only control, pRC1542 for *piggyBac* control and pRecLV103-GFP-PGBD5 for GFP-PGBD5). $2x10^5$ cells were loaded per plate at the start of the selection. **d)** The frequency of cell survival under the different conditions. The graph depicts number of foci per 100 mm^2 plate. * means $P<0.05$. Statistical test was a one-way ANOVA. n=3.

As with the HEK293 cells, the initial experimental work was to identify the plasmid-based expression levels of the GFP-PGBD5 fusion protein in the HeLa cells. A western blot was performed from both the GFP, and GFP-PGBD5 fusion-expressing cells (Figure 35a), which showed high levels of GFP in the control cells. The GFP-PGBD5 protein had much lower expression than was seen with the GFP control, different from the HEK293 cells where the expression levels were approximately consistent with one another (Figure 34a).

The live cell fluorescence imaging was used to test the localisation of the GFP-PGBD5 fusion (Figure 35b). The HeLa cells showed very similar results to the HEK293 cells, where the majority of the fluorescence was located within the nucleoplasm of the cells, though again the expression level was relatively low so visible light could not be used to show the rest of the cell.

The transposition integration assay in HeLa cells made use of the same reporter construct as in the HEK293 assays. The plates were seeded with 1000 cells initially, but again there were too few colonies on the transposon only and PGBD5 plates to be viable for accurate quantification. Therefore, $2x10^5$ cells were seeded per plate, and viewed after two weeks (Figure 35c). Cells were counted and the number of foci per plate were displayed (Figure 35d).

The transposon-only plates showed a higher foci count than in the HEK293 cells, averaging 372 cells per plate. The *piggyBac* control again had significantly more foci than the transposon only (P=0.0458), averaging 3288

foci per plate. This equated to an increase of ~10-fold, much smaller than the ~200-fold increase seen in HEK293 cells. GFP-PGBD5 did not show an increase in foci count as would be expected, its average actually fell to 187 foci per plate, approximately half that seen in the transposon-only plates. This decrease was not statistically significant however (P=0.2861), so it was likely to be merely a by-product of the experimental conditions used.

6.3 Discussion

In the original 2015 paper PGBD5 was shown to be a functional transposase, integrating using the *T. ni* transposon ends, in HEK293 cells (Henssen et al. 2015). PGBD5 was ~4.5 times less active than the original *piggyBac* transposase from *T. ni,* although it was still significantly more active than the negative control. The work in this chapter attempted to replicate these findings, however, here the PGBD5 transposition frequency was not significantly higher than the negative control and was ~60 fold less than for *piggyBac*. The slight upwards trend that was seen for PGBD5 does suggest that it may have been having some effect in the cells and the discrepancy could have been due to some conditions that are different between the assays here and the ones from the published work. Given that the expression and localisation of the protein appear to be functioning correctly (Figure 34a-b), it is not likely that there was a problem at those stages of the assay. As cell lines differentiate over time, the same cell lines in two different

laboratories may be slightly different, and have different protein expression profiles. It is possible then that the cells themselves were sufficiently different that some factors were repressing the activity of PGBD5. The only other major difference between the work from the paper and here is to do with the transposon only control. In the published work the transposon only control used the same expression plasmid as was used for the other samples, with only GFP being expressed, which was expressed but was unable to cause any transposition. In the experiments shown here, the transposon only control used an empty plasmid, pBS. It is possible that the alternative plasmid affected the overall quantity of integration occurring, or that the expression of GFP in the published work reduced the overall level of reporter integration. It would be worthwhile testing this in the future, in order to ensure the accuracy of these assays.

In HeLa cells, the PGBD5 protein showed no transposition activity at all, in fact it showed a slight decrease, though this decrease wasn't statistically significant (Figure 35c-d). As the expression level of PGBD5 was significantly lower than the GFP control (Figure 35a), it is possible that PGBD5 may have been degraded by the cell, or simply not transcribed as well as in HEK293 cells. If the protein was active in the HeLa cells, lower protein expression would reduce any increase in integration frequency compared with the transposon only control. The GFP-PGBD5 cells seemed to show an overall decrease, even if it is only a slight trend, which implied the protein was not showing any transposition activity. A possible hypothesis for these data is that there is a protein that represses the PGBD5 activity,

which shows higher expression in HeLa cells compared with HEK293 cells. This might produce the absence of activity seen under these circumstances.

Since the original 2015 paper, two more publications have been produced concerning PGBD5. The first developed a forward genetic screen for PGBD5 activity and identified that PGBD5 was capable of recombination using PGBD5-specific signal (PSS) sequences (Henssen et al. 2016). The PSS was shown to be 12 bp long, with significant variability in the sequence. In the paper they identified a series of translocation and deletion events using the screen that had various PSS sequences flanking them. The other, more recent publication, looked at PGBD5 and its relationship with rhabdoid tumours (Henssen et al. 2017). The paper demonstrated that PGBD5 was associated with multiple rearrangements that are connected to the formation of rhabdoid tumours, therefore it is likely to be an oncogenic mutator in these cells.

There are two rather significant questions that are unanswered regarding PGBD5. The first is its non-oncogenic function in cells. It is known to be well expressed in the brain and central-nervous system in humans, and has been well conserved throughout the last ~500 Mya. This conservation means it must serve some function, though at present none has been identified.

The second major question concerning PGBD5 involves the PSS sequence. If PGBD5 performs rearrangements in the genome, using these sequences, why does it seem limited to certain genes in rhabdoid tumours and why in normal cells does it not cause large scale rearrangements in the

cell? A single example of the PSS sequence (CTGGGAGGCAGA) occurs 4494 times in the human genome (based on a BLAST search against the human RefSeq genome (Altschul et al. 1990)), and this is just for that particular sequence. In order for it not to be a detrimental factor in the cells, there must be other factors that are involved in controlling and possibly targeting it. Exactly what these factors are remains to be seen.

Chapter 7

Conclusions and Perspectives

7.1 Summary of the research

The underlying question for this entire body of work was whether there are mechanistic limitations that affect the capacity of some transposases to function within a higher eukaryotic organism. To undertake this analysis, a transposase was needed that was both well characterised and could be relatively easily manipulated. The Tn5 transposase fitted this brief, and was believed to be inactive in human cells (Reznikoff 2013, pers comm) with only limited activity in lower eukaryotes under certain conditions (Goryshin et al. 2000).

7.1.1 Modification of the complex formation mechanism

The mechanisms of complex formation were the target of the majority of the manipulations in this work. The hypothesis was that the S-PD pathway was the root cause of the Tn5 transposase being almost entirely inactive in eukaryotes. In this pathway, the transposase itself exists only as a monomer in solution, which means the binding of one TIR happens entirely independently of the other. These SECs are held together through quite weak DNA-protein interactions (see Chapter three) that do not endure for an extended period. Eukaryotic genomes are almost always larger than their prokaryotic counterparts which means that transposase binding to the TIRs will be a significantly less frequent event and, given the transient nature of the SEC, two of these events occurring independently, sufficiently close in

time, will be extremely uncommon. The two SECs would also have to combine to form the PEC in order to proceed with the reaction, which reduces the frequency of transposition even further.

If the S-PD pathway is the significant limitation for eukaryotic Tn5 activity, then mimicking the S-NEC pathway was a possible way to enhance the transposition. In most cases with S-NEC, the transposase exists as a dimer prior to any DNA binding (Cuypers et al. 2012; Claeys Bouuaert et al. 2013). One TIR is bound by the transposase dimer, and the second TIR is recruited into this to form the PEC. To mimic this pathway effectively, a single-chain dimer of the Tn5 transposase was used, with an amino acid linker between the monomers (Chapter three). Although the two Tn5 transposases were not actually in a complex with each other prior to the second-strand binding, they still mimicked the reaction dynamics of the S-NEC, and demonstrated significantly greater activity in eukaryotes, as well as hyperactivity in bacteria. Given that the Tn5 transposase forms weak interactions in its SEC, the dimer can bind one end and put the second transposase in close proximity to its TIR, which then leads to PEC formation occurring before the SEC disassembles. This work implies that S-NEC as the pathway of complex formation is significantly more effective in both prokaryotic and eukaryotic settings.

If S-NEC is inherently a more effective form of complex assembly than S-PD, as is suggested by Chapter three, it raises the question of why transposition systems would function through S-PD in the first place. The first answer is linked to regulation of transposition activity. Unregulated highly

active transposition activity in an organism would lead to the death of the host and, by extension, the loss of the transposons within it (Le Rouzic et al. 2007). Therefore, transposition systems are inherently less active than theoretically possible, in order to prevent death of the host. The lower levels of activity by the S-PD pathway could therefore be beneficial to transposon and host survival in the long term. If the transposase has developed S-PD as its pathway, there would be very limited selection pressure on it that would normally lead to higher efficiency mechanisms evolving over time. Those transposases that use S-NEC demonstrate their own mechanisms of regulation. Many, at least in the *mariner* family, exhibit OPI which regulates the frequency of transposition and ensures it is kept at a low level.

The other explanation for the prevalence of the S-PD pathway is concerned with which transposon is the intended target. For prokaryotic autonomous elements, the transposase is translated in close proximity to the TIRs. This produces a high local concentration of transposase, which increases the frequency of the two TIRs being bound. For monomeric transposases, the rate of transposition with the local transposon (*in cis*) will be high. If a monomeric transposase does bind to another transposon in the genome (*in trans*) the second TIR would also have to be bound for transposition, which would be rare event. Dimeric transposase would be much more likely to function well *in trans* as, once a dimer has bound to one TIR, the unbound transposase is effectively at a high local concentration and will be more likely to bind the opposite end. Therefore, having monomers in

solution for some systems may be an adaptation to avoid mobilising other transposons.

There is another significant difference between those S-NEC based systems characterised to date and the S-PD based systems such as Tn5 and Tn10. That difference is the stability of the SEC. In S-PD transposition systems, the SEC has weak DNA-protein interactions (at least for Tn5 and Tn10 (Bischerour & Chalmers 2007; Bischerour & Chalmers 2009)). This will lead to the SEC quickly disassembling, a significant limitation in a large eukaryotic genome where no high local concentration would be possible, due to translation occurring outside the nucleus. The SEC formed in the S-NEC pathway by comparison, is much more stable, allowing a longer period of time in which the second TIR can be recruited into the complex. This is particularly important for S-NEC as the affinity for the second TIR is much lower than for the first, in at least some cases (Claeys Bouuaert et al. 2013). Therefore, a feature that may be necessary for effective transposition in eukaryotes could be a stable SEC. If the SEC of the S-PD pathway was more stable, then the length of time available in which a second transposase could bind would be increased (Chapter four). Increasing the binding time of the SEC, using a fused ZF domain, would be likely to result in improved activity for the transposase within a eukaryotic genome. The side effect of the ZF-fusion, however, was significant inhibition of the transposition activity, as it reduced the DNA-binding capacity of the transposase. The presence of a ZBS did improve the binding of the fusion protein overall though, so this hypothesis might be testable with some additional work.

7.1.2 Other inhibitors within a eukaryotic genome

While the complex formation pathway may explain the lack of Tn5 transposition within a eukaryotic genome, there are also other factors that may inhibit its functionality. One major difference between bacterial and eukaryotic genomes is the presence of chromatin which could potentially severely limit the access for, and therefore the activity of, the Tn5 transposase. As the Tn5 transposase is prokaryotic in origin, it does not have any adaptations to cope with the presence of chromatin, and has been shown to be unable to integrate into heterochromatic DNA (Buenrostro et al. 2013; Sos et al. 2016). The chromatin may complex with the transposon TIRs, which may prevent transposase binding, and thus inhibit transposition. Hypothetically, fusing the chromatin-binding PHD domain from the *piggyBac* transposase could enhance transposition (Chapter five) (Yusa 2015). When this fusion construct was tested however, there was no enhancement of the transposition activity in the human cells. This does suggest that, if the PHD domain was functioning correctly, the presence of chromatin may not a significant disadvantage for Tn5 transposition for at least the integration steps of the reaction. This fits with the published evidence that shows that Tn5 does not integrate into DNA that is bound into chromatin (Buenrostro et al. 2013). The possibility that the PHD domain itself was non-functional has not been ruled out however, so the strength of this evidence is limited.

There are other factors that could affect Tn5 transposition in eukaryotes. The transposase could be degraded by the cellular host factors, preventing any transposition. This possibility is ruled out by the work in Chapter three, as the monomeric and dimeric constructs were able to transpose, so there was not any significant degradation occurring. Similarly, the successful eukaryotic transposition in Chapter three also ruled out bacterial-specific host factors being required by Tn5 when *in vivo*, as well as any eukaryotic host factors that inhibit the activity.

7.2 Outlook

While this work has provided some significant evidence that complex formation is the major limiting factor for Tn5 transposition in eukaryotes, more investigation needs to be done, both expanding on the research done here and exploring potential new avenues of inquiry.

The single-chain dimer that mimics S-NEC (Chapter three) showed substantial evidence for function in eukaryotes; however, there are some additional pieces of data that would strengthen the body of work. One outstanding question was whether the activity in eukaryotes might be partially due to the dimeric transposase causing genomic instability, which would increase the frequency of reporter integration. Testing the dimeric transposase in human cells, using a reporter plasmid that did not contain Tn5 TIR sequences would be an effective way to test this hypothesis. The other unanswered question of this chapter was whether the small dimer PEC

was significantly less stable than that of the monomeric transposases, which would be tested through a PEC stability assay. In this assay, labelled TIR substrate would form a PEC with the transposase, then excess unlabeled TIR substrate would be added and the intensity of the PEC band could be examined through an EMSA over time.

The hypothesis that a more stable SEC would enhance eukaryotic transposition was not tested due to the unfortunate side-effects of the ZF fusion to the transposase (Chapter four). As N-terminal fusions produced these unwanted side effects, fusing the ZF to the C-terminus might produce a more active transposase protein. Once the side-effects have been overcome, the system could be tested at low protein concentrations in both prokaryotes and eukaryotes, to see if any increase in activity can be demonstrated.

If it is possible, combining the dimeric transposase with the ZF fusion could, in theory, produce a transposase that is capable of extremely high levels of activity in eukaryotes. A possible disadvantage of this approach might be that strong binding of the SEC, when using a dimeric transposase, could lead to the occurrence of OPI (OverProduction Inhibition). This could occur because, at higher protein concentrations, both TIRs would be bound by seperate ZF-dimeric transposase, they would likely be very slow to dissociate, due to the high stability of the ZF-DNA complex. Therefore, fewer PECs would be able to form, and less transposition would occur. This idea was tested in Chapter four; however, no improvement in transposition activity was seen when the ZBSs were present. If the ZF system were optimised, as

discussed in the previous paragraph, it may be a more advantageous combination.

Given that the dimeric transposase demonstrates hyperactivity both *in vitro* and *in vivo*, this raises the possibility of it being used as a genetic tool. The Tn5 transposase monomer has already seen significant use as a tool for random mutagenesis (Goryshin et al. 2000; Petzke & Luzhetskyy 2009), and in the generation of genomic libraries for next-generation sequencing projects (Picelli et al. 2014). The single-chain transposase dimer would be particularly beneficial for both of these techniques, if it were combined with the recently characterised Tn5 mutant transposase that shows a reduced target site preference (Kia et al. 2017). Before the single-chain transposase dimer could be used in these applications however, the potential issue with stability would need to be resolved as instability has been shown to make it less efficient for these applications (Chapter three, Figure 22). The best way to resolve this potential problem would be through optimisation of the linker length and composition, which might produce a stable, and more applicable transposase dimer.

As the work presented here suggests that the complex formation pathway is a limiting factor for the Tn5 transposase, a potential new line of inquiry might be whether this remains true with other transposable elements. The Tn10 transposase would be an ideal candidate for investigation as it is well characterised, and also very similar to the Tn5 transposase, as both are members of the IS4 family (Haniford & Ellis 2015). Another possible candidate transposase might be MuA, however this forms at least a

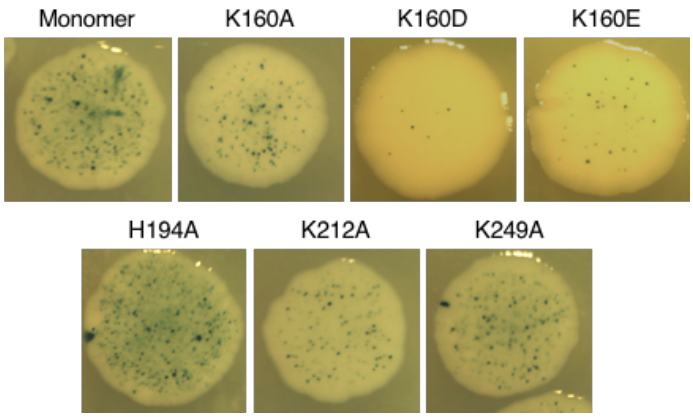
tetrameric complex, so it would be more of a challenge to analyse (Craig 2002). It might also be valuable to examine those transposition systems that do not use the RNase-H core, although none characterised sufficiently have demonstrated the S-PD pathway. Testing of other systems would make it possible to identify whether the model (Figure 12) is a common feature of transposable elements.

Chapter 8

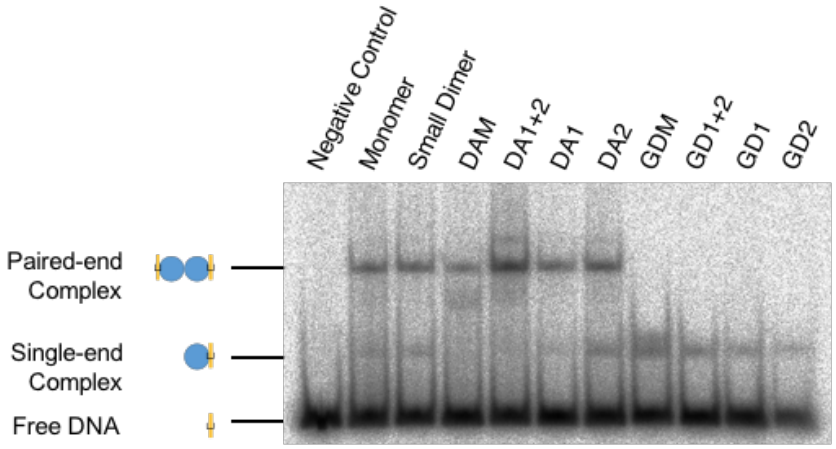
Appendices

8.1 Supplementary information

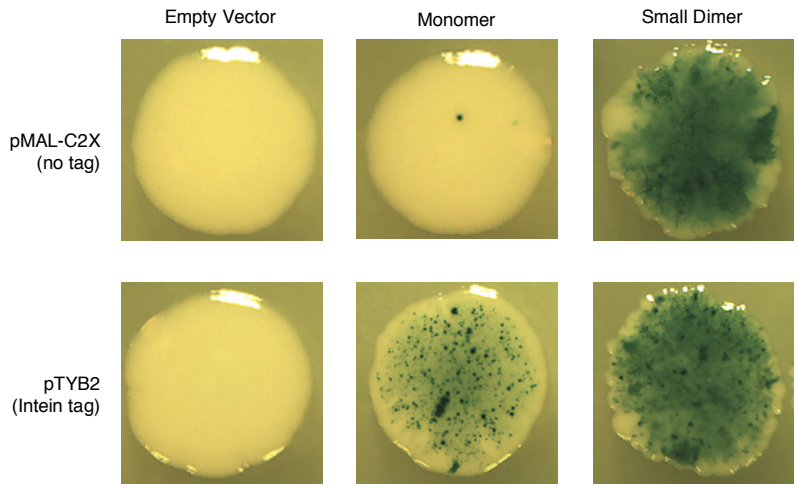
8.1.1 Supplementary figures



Supplementary Figure 1: Mutagenesis Screening for integration Deficient Mutants. This was done using a papillation assay, in RC5123. The plasmids involved were pRC2128 for the Tn5 transposase monomer, pRC2172 for K160A mutant, pRC2195 for K160D mutant, pRC2196 for K160E mutant, pRC2173 for H194A mutant, pRC2174 for K212A mutant and pRC2175 for K249A mutant. Plates were made without IPTG and with 0.01% lactose. Refer to section 3.2.1.2 for details.



Supplementary Figure 2: Binding Assay of Dimeric Transposases and Mutants. Samples incubated for 15 minutes at RT, before running on an 8% TBE gel. Substrate used was the short substrate used in Figure 17, labelled using α-^{32}P dCTP. The various different complexes formed are shown on the left. The blue circles indicate the transposase, the orange line indicates the substrate and the black bracket indicates the transposon end sequence. DAM= Transposase monomer mutated with D97A. DA1+2= Transposase small dimer with both monomers mutated with D97A. DA1= Transposase small dimer with the first monomer mutated with D97A. DA2= Transposase small dimer with the second monomer mutated with D97A. GDM= Transposase monomer mutated with G462D. GD1+2= Transposase small dimer with both monomers mutated with G462D. GD1= Transposase small dimer with the first monomer mutated with G462D. GD2= Transposase small dimer with the second monomer mutated with G462D. Refer to section 2.3.2.4 for details.



Supplementary Figure 3: Papillation Assay Comparing the Transposition Activity Between Two Different Vectors. The vectors tested were the pMAL-C2X and pTYB2. Plasmids used were pMAL-C2X empty vector, Tn5 transposase monomer in pMAL-C2X (pRC2117), Tn5 transposase small dimer in pMAL-C2X (pRC2109), pTYB2 empty vector, Tn5 transposase monomer in pTYB2 (pRC2128) and Tn5 transposase small dimer in pTYB2 (pRC2102). The strain used in this assay was RC5123. The agar plates contained 0.01% lactose as well as X-gal and Ampicillin. The colonies shown are representative of those on the plates. Blue papillae are indicative of LacZ expression. Refer to section 4.2.1 for details.

8.1.2 Protein sequences

Sequence of the Tn5 transposase monomer (red) with the intein tag (blue) on its C-terminus. Shown as it was expressed, from pRC2128:

MITSALHRAADWAKSVFSSAALGDPRRTARLVNVAAQLAKYSGKSITISS
EGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLAQEFPELLAIEDTTSL
SYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATTFRTVGLLHQEWWM
RPDDPADADEKESGKWLAAAATSRLRMGSMMSNVIAVCDREADIHAYLQ
DKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPELGGYQISIPQKGVV
DKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLAEEINPPKGETPLK
WLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTGAGAERQRMEEPD
NLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEAEHVESQSAETVL
TPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGGFMDSKRTGIAS
WGALWEGWEALQSKLDGFLAAKDLMAQGIKIEFLEPGCFAKGTNVLMAD
GSIECIENIEVGNKVMGKDGRPREVIKLPRGRETMYSVVQKSQHRAHKS
DSSREVPELLKFTCNATHELVVRTPRSVRRLSRTIKGVEYFEVITFEMGQ
KKAPDGRIVELVKEVSKSYPISEGPERANELVESYRKASNKAYFEWTIEA
RDLSLLGSHVRKATYQTYAPILYENDHFFDYMQKSKFHLTIEGPKVLAYLL
GLWIGDGLSDRATFSVDSRDTSLMERVTEYAEKLNLCAEYKDRKEPQVA
KTVNLYSKVVRGNGIRNNLNTENPLWDAIVGLGFLKDGVKNIPSFLSTDNI
GTRETFLAGLIDSDGYVTDEHGIKATIKTIHTSVRDGLVSLARSLGLVVSVN
AEPAKVDMNVTKHKISYAIYMSGGDVLLNVLSKCAGSKKFRPAPAAAFAR
ECRGFYFELQELKEDDYYGITLSDDSDHQFLLGSQVVVHACGGLTGLNS
GLTTNPGVSAWQVNTAYTAGQLVTYNGKTYKCLQPHTSLAGWEPSNVP
ALWQLQ

Sequence of the Tn5 transposase small dimer with the intein tag (blue, the … indicates full sequence, as seen before) on its C-terminus. The two Tn5 transposase monomers (red) are joined by the small linker (light grey). Shown as it was expressed, from pRC2102:

MASSRMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAAQLAKYSGK
SITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLAQEFPELLAIE
DTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATTFRTVGLLHQ
EWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNVIAVCDREADI
HAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPELGGYQISIPQ
KGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLAEEINPPKGE
TPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTGAGAERQRM
EEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEAEHVESQSA
ETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGGFMDSKRTG
IASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGSSRGGGSEGGGSE
GGSGTSEFMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAAQLAKY
SGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLAQEFPEL
LAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATTFRTVGL
LHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNVIAVCDR
EADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPELGGYQI
SIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLAEEINPP
KGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTGAGAER
QRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEAEHVES
QSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGGFMDSK
RTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIPGCFAK…

Sequence of the Tn5 transposase large dimer with the intein tag (blue) on its C-terminus. The two Tn5 transposase monomers (red) are joined by first the small linker (light grey) and then the large linker (dark grey). Shown as it was expressed, from pRC2101:

MASSRMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAAQLAKYSGK
SITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLAQEFPELLAIE
DTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATTFRTVGLLHQ
EWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNVIAVCDREADI
HAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPELGGYQISIPQ
KGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLAEEINPPKGE
TPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTGAGAERQRM
EEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEAEHVESQSA
ETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGGFMDSKRTG
IASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGSSRGGGSEGGGSE
GGSGTSMSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILD
EIADEYQGKLTVAKLNIDQNPGTAPKYGIRGIPTLLLFKNGEVAATKVGAL
SKGQLKEFLDANLAGSGSGHMHHHHHHSSGLVPRGSGMKETAAAKFER
QHMDSPDLGTDDDDKAMADITSEFMITSALHRAADWAKSVFSSAALGDP
RRTARLVNVAAQLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRK
AGAMQTVKLAQEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWW
VHSVLLLEATTFRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSR
LRMGSMMSNVIAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGL
YLYDHLKNQPELGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITL
KQGNITLNAVLAEEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRW
RIEEFHKAWKTGAGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPP
QALRAQGLLKEAEHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSL
QWAYMAIARLGGFMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLM
AQGIKIPGCFAK…

Sequence of the Tn5 transposase monomer (red) as it was expressed, from pRC2117:

MASSRVDGGREFMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAA
QLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLA
QEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATT
FRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNV
IAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPE
LGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLA
EEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTG
AGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEA
EHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGG
FMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIP

Sequence of the Tn5 transposase small dimer. The two Tn5 transposase monomers (red) are joined by the small linker (light grey). Shown as it was expressed, from pRC2109:

MASSRMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAAQLAKYSGK
SITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLAQEFPELLAIE
DTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATTFRTVGLLHQ
EWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNVIAVCDREADI
HAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPELGGYQISIPQ
KGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLAEEINPPKGE
TPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTGAGAERQRM
EEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEAEHVESQSA
ETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGGFMDSKRTG
IASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGSSRGGGSEGGGSE
GGSGTSEFMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAAQLAKY
SGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLAQEFPEL
LAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATTFRTVGL
LHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNVIAVCDR
EADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPELGGYQI
SIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLAEEINPP
KGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTGAGAER
QRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEAEHVES
QSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGGFMDSK
RTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIP

Sequence of the Tn5 transposase large dimer. The two Tn5 transposase monomers (red) are joined by first the small linker (light grey) and then the large linker (dark grey). Shown as it was expressed, from pRC2126:

MASSRMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAAQLAKYSGK
SITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLAQEFPELLAIE
DTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATTFRTVGLLHQ
EWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNVIAVCDREADI
HAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPELGGYQISIPQ
KGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLAEEINPPKGE
TPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTGAGAERQRM
EEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEAEHVESQSA
ETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGGFMDSKRTG
IASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGSSRGGGSEGGGSE
GGSGTSMSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILD
EIADEYQGKLTVAKLNIDQNPGTAPKYGIRGIPTLLLFKNGEVAATKVGAL
SKGQLKEFLDANLAGSGSGHMHHHHHHSSGLVPRGSGMKETAAAKFER
QHMDSPDLGTDDDDKAMADITSEFMITSALHRAADWAKSVFSSAALGDP
RRTARLVNVAAQLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRK
AGAMQTVKLAQEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWW
VHSVLLLEATTFRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSR
LRMGSMMSNVIAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGL
YLYDHLKNQPELGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITL
KQGNITLNAVLAEEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRW
RIEEFHKAWKTGAGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPP
QALRAQGLLKEAEHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSL
QWAYMAIARLGGFMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLM
AQGIKIP

Sequence of the Tn5 transposase monomer (red) with an N-terminal NLS (yellow). Shown as it was expressed, from pRC2148:

MEPPKKKRKVRVNMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAA
QLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLA
QEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATT
FRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNV
IAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPE
LGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLA
EEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTG
AGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEA
EHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGG
FMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIHHHHHHL

Sequence of the Tn5 transposase small dimer with an N-terminal NLS (yellow). The two Tn5 transposase monomers (red) are joined by the small linker (light grey). Shown as it was expressed, from pRC2147:

MEPPKKKRKVRVNMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAA
QLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLA
QEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATT
FRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNV
IAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPE
LGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLA
EEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTG
AGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEA
EHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGG
FMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGPSRGG
GSEGGGSEGGSGTSQLMITSALHRAADWAKSVFSSAALGDPRRTARLV
NVAAQLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQT
VKLAQEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLL
EATTFRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSM
MSNVIAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLK
NQPELGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITL
NAVLAEEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKA
WKTGAGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQG
LLKEAEHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAI
ARLGGFMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIHH
HHHHL

Sequence of the Tn5 transposase large dimer with an N-terminal NLS (yellow). The two Tn5 transposase monomers (red) are joined by first the small linker (light grey) and then the large linker (dark grey). Shown as it was expressed, from pRC2146:

MEPPKKKRKVRVNMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAA
QLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLA
QEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATT
FRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNV
IAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPE
LGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLA
EEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTG
AGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEA
EHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGG
FMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGPSRGG
GSEGGGSEGGSGTSMSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGP
CKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYGIRGIPTLLLFKNGEV
AATKVGALSKGQLKEFLDANLAGSGSGHMHHHHHHSSGLVPRGSGMKE
TAAAKFERQHMDSPDLGTDDDDKAMADITSQLMITSALHRAADWAKSVF
SSAALGDPRRTARLVNVAAQLAKYSGKSITISSEGSKAAQEGAYRFIRNP
NVSAEAIRKAGAMQTVKLAQEFPELLAIEDTTSLSYRHQVAEELGKLGSIQ
DKSRGWWVHSVLLLEATTFRTVGLLHQEWWMRPDDPADADEKESGKW
LAAAATSRLRMGSMMSNVIAVCDREADIHAYLQDKLAHNERFVVRSKHP
RKDVESGLYLYDHLKNQPELGGYQISIPQKGVVDKRGKRKNRPARKASL
SLRSGRITLKQGNITLNAVLAEEINPPKGETPLKWLLLTSEPVESLAQALRV
IDIYTHRWRIEEFHKAWKTGAGAERQRMEEPDNLERMVSILSFVAVRLLQ
LRESFTPPQALRAQGLLKEAEHVESQSAETVLTPDECQLLGYLDKGKRK
RKEKAGSLQWAYMAIARLGGFMDSKRTGIASWGALWEGWEALQSKLDG
FLAAKDLMAQGIKIHHHHHHL

Sequence of the Tn5 transposase monomer (red) with an N-terminal NLS (yellow) and a C-terminal intein tag (blue). Shown as it was expressed, from pRC2199:

MPKKKRKVMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAAQLAKY
SGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLAQEFPEL
LAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATTFRTVGL
LHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNVIAVCDR
EADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPELGGYQI
SIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLAEEINPP
KGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTGAGAER
QRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEAEHVES
QSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGGFMDSK
RTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGEFLEPGCFAK…

Sequence of the Tn5 transposase small dimer with an N-terminal NLS (yellow) and a C-terminal intein tag (blue). The two Tn5 transposase monomers (red) are joined by the small linker (light grey). Shown as it was expressed, frompRC2200:

MPKKKRKVMASSRMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAA
QLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLA
QEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATT
FRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNV
IAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPE
LGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLA
EEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTG
AGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEA
EHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGG
FMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGSSRGG
GSEGGGSEGGSGTSEFMITSALHRAADWAKSVFSSAALGDPRRTARLV
NVAAQLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQT
VKLAQEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLL
EATTFRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSM
MSNVIAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLK
NQPELGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITL
NAVLAEEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKA
WKTGAGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQG
LLKEAEHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAI
ARLGGFMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIPG
CFAK…

Sequence of the Tn5 transposase monomer (red), with a ZFA (green) fused to its N-terminus via a small linker (light grey), and also an intein tag (blue) at the C-terminus. Shown as it was expressed, from pRC2170:

MDYKDDDDKSRPGERPFQCRICMRNFSKHQTLQRHTRTHTGEKPFQCR
ICMRNFSLLHNLTRHLRTHTGEKPFQCRICMRNFSVNHHLDRHLKTHLRG
SSRGGGSEGGGSEGGSGTSEFMITSALHRAADWAKSVFSSAALGDPRR
TARLVNVAAQLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAG
AMQTVKLAQEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVH
SVLLLEATTFRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSRLR
MGSMMSNVIAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYL
YDHLKNQPELGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQ
GNITLNAVLAEEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIE
EFHKAWKTGAGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQA
LRAQGLLKEAEHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQW
AYMAIARLGGFMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLMAQ
GIKIPGCFAK…

Sequence of the Tn5 transposase monomer (red), with a ZFA (green) fused to its N-terminus via an extended small linker (light grey), and also an intein tag (blue) at the C-terminus. Shown as it was expressed, from pRC2187:

MDYKDDDDKSRPGERPFQCRICMRNFSKHQTLQRHTRTHTGEKPFQCR
ICMRNFSLLHNLTRHLRTHTGEKPFQCRICMRNFSVNHHLDRHLKTHLRG
SGGGSGGGSGGSRGGGSEGGGSEGGSGTSEFMITSALHRAADWAKSV
FSSAALGDPRRTARLVNVAAQLAKYSGKSITISSEGSKAAQEGAYRFIRN
PNVSAEAIRKAGAMQTVKLAQEFPELLAIEDTTSLSYRHQVAEELGKLGSI
QDKSRGWWVHSVLLLEATTFRTVGLLHQEWWMRPDDPADADEKESGK
WLAAAATSRLRMGSMMSNVIAVCDREADIHAYLQDKLAHNERFVVRSKH
PRKDVESGLYLYDHLKNQPELGGYQISIPQKGVVDKRGKRKNRPARKAS
LSLRSGRITLKQGNITLNAVLAEEINPPKGETPLKWLLLTSEPVESLAQALR
VIDIYTHRWRIEEFHKAWKTGAGAERQRMEEPDNLERMVSILSFVAVRLL
QLRESFTPPQALRAQGLLKEAEHVESQSAETVLTPDECQLLGYLDKGKR
KRKEKAGSLQWAYMAIARLGGFMDSKRTGIASWGALWEGWEALQSKLD
GFLAAKDLMAQGIKIPGCFAK…

Sequence of the Tn5 transposase small dimer, with a ZFA (green) fused to its N-terminus via an extended small linker (light grey), and also an intein tag (blue) at the C-terminus. The two Tn5 transposase monomers (red) are joined by the small linker (light grey) Shown as it was expressed, from pRC2171:

MDYKDDDDKSRPGERPFQCRICMRNFSKHQTLQRHTRTHTGEKPFQCR
ICMRNFSLLHNLTRHLRTHTGEKPFQCRICMRNFSVNHHLDRHLKTHLRG
SSRGGGSEGGGSEGGSGHMASSRMITSALHRAADWAKSVFSSAALGD
PRRTARLVNVAAQLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIR
KAGAMQTVKLAQEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGW
WVHSVLLLEATTFRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATS
RLRMGSMMSNVIAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESG
LYLYDHLKNQPELGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRIT
LKQGNITLNAVLAEEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRW
RIEEFHKAWKTGAGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPP
QALRAQGLLKEAEHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSL
QWAYMAIARLGGFMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLM
AQGIKIGSSRGGGSEGGGSEGGSGTSEFMITSALHRAADWAKSVFSSAA
LGDPRRTARLVNVAAQLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSA
EAIRKAGAMQTVKLAQEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSR
GWWVHSVLLLEATTFRTVGLLHQEWWMRPDDPADADEKESGKWLAAA
ATSRLRMGSMMSNVIAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDV
ESGLYLYDHLKNQPELGGYQISIPQKGVVDKRGKRKNRPARKASLSLRS
GRITLKQGNITLNAVLAEEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYT
HRWRIEEFHKAWKTGAGAERQRMEEPDNLERMVSILSFVAVRLLQLRES
FTPPQALRAQGLLKEAEHVESQSAETVLTPDECQLLGYLDKGKRKRKEK
AGSLQWAYMAIARLGGFMDSKRTGIASWGALWEGWEALQSKLDGFLAA
KDLMAQGIKIPGCFAK…

Sequence of the ZFA (green) attached to an intein tag (blue) at the C-terminus, via the small linker (light grey). Shown as it was expressed, from pRC2192:

MDYKDDDDKSRPGERPFQCRICMRNFSKHQTLQRHTRTHTGEKPFQCR
ICMRNFSLLHNLTRHLRTHTGEKPFQCRICMRNFSVNHHLDRHLKTHLRG
SSRGGGSEGGGSEGGSGTSEFLEPGCFAK…

Sequence of the Tn5 transposase monomer (red), with an 82 aa PHD domain (green) attached to the monomer C-terminus via a small linker (light grey). There is also an intein tag (blue) on its C-terminus. Shown as it was expressed, from pRC2506:

MASSRMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAAQLAKYSGK
SITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLAQEFPELLAIE
DTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATTFRTVGLLHQ
EWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNVIAVCDREADI
HAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPELGGYQISIPQ
KGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLAEEINPPKGE
TPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTGAGAERQRM
EEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEAEHVESQSA
ETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGGFMDSKRTG
IASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGSSRGGGSEGGGSE
GGSGTSEFSFMRKRLEAPTLKRYLRDNISNILPNEVPGTSDDSTEEPVMK
KRTYCTYCPSKIRRKANASCKKCKKVICREHNIDMCQSCFPGCFAK…

Sequence of the Tn5 transposase monomer (red), with a 90 aa PHD domain (green) attached to the monomer C-terminus via a small linker (light grey). There is also an intein tag (blue) on its C-terminus. Shown as it was expressed, from pRC2507:

MASSRMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAAQLAKYSGK
SITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLAQEFPELLAIE
DTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATTFRTVGLLHQ
EWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNVIAVCDREADI
HAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPELGGYQISIPQ
KGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLAEEINPPKGE
TPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTGAGAERQRM
EEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEAEHVESQSA
ETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGGFMDSKRTG
IASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGSSRGGGSEGGGSE
GGSGTSEFNLYMSLTSSFMRKRLEAPTLKRYLRDNISNILPNEVPGTSDD
STEEPVMKKRTYCTYCPSKIRRKANASCKKCKKVICREHNIDMCQSCFPG
CFAK...

Sequence of the Tn5 transposase monomer (red), with a NLS (yellow) attached at the N-terminus, and an 82 aa PHD domain (green) attached to the C-terminus. Shown as it was expressed, from pRC2508:

MEPPKKKRKVRVNMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAA
QLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLA
QEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATT
FRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNV
IAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPE
LGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLA
EEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTG
AGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEA
EHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGG
FMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGSSFMRK
RLEAPTLKRYLRDNISNILPNEVPGTSDDSTEEPVMKKRTYCTYCPSKIRR
KANASCKKCKKVICREHNIDMCQSCFLEHHHHHHL

Sequence of the Tn5 transposase monomer (red), with a NLS (yellow) attached at the N-terminus, and a 90 aa PHD domain (green) attached to the C-terminus. Shown as it was expressed, from pRC2509:

MEPPKKKRKVRVNMITSALHRAADWAKSVFSSAALGDPRRTARLVNVAA
QLAKYSGKSITISSEGSKAAQEGAYRFIRNPNVSAEAIRKAGAMQTVKLA
QEFPELLAIEDTTSLSYRHQVAEELGKLGSIQDKSRGWWVHSVLLLEATT
FRTVGLLHQEWWMRPDDPADADEKESGKWLAAAATSRLRMGSMMSNV
IAVCDREADIHAYLQDKLAHNERFVVRSKHPRKDVESGLYLYDHLKNQPE
LGGYQISIPQKGVVDKRGKRKNRPARKASLSLRSGRITLKQGNITLNAVLA
EEINPPKGETPLKWLLLTSEPVESLAQALRVIDIYTHRWRIEEFHKAWKTG
AGAERQRMEEPDNLERMVSILSFVAVRLLQLRESFTPPQALRAQGLLKEA
EHVESQSAETVLTPDECQLLGYLDKGKRKRKEKAGSLQWAYMAIARLGG
FMDSKRTGIASWGALWEGWEALQSKLDGFLAAKDLMAQGIKIGSNLYMS
LTSSFMRKRLEAPTLKRYLRDNISNILPNEVPGTSDDSTEEPVMKKRTYC
TYCPSKIRRKANASCKKCKKVICREHNIDMCQSCFLEHHHHHHL

Sequence of the Hsmar1 transposase monomer (red). Shown as it was expressed, from pRC2141:

MEMMLDKKQIRAIFLFEFKMGRKAAETTRNINNAFGPGTANERTVQWWF
KKFRKGDESLEDEERSGRPSEVDNDQLRAIIEADPLTTTREVAEELNVDH
STVVRHLKQIGKVKKLDKWVPHELSENQKNRRFEVSSSLILRNNNEPFLD
RIVTCDEKWILYDNRRRSAQWLDREEAPKHFPKPNLHQKKVMVTVWWS
AAGVIHYSFLNPGETITSEKYCQQIDEMHRKLQRLQPALVNRKGPILLHDN
ARPHVAQPTLQKLNELGYEVLPHPPYSPDLSPTDYHFFKHLDNFLQGKR
FHNQQDAENAFQEFVESRSTDFYATGINKLISRWQKCVDCNGSYFD

Sequence of the Hsmar1 transposase monomer (red) with an 82 aa PHD domain (green) attached at the C-terminus. Shown as it was expressed, from pRC2510:

MEMMLDKKQIRAIFLFEFKMGRKAAETTRNINNAFGPGTANERTVQWWF
KKFRKGDESLEDEERSGRPSEVDNDQLRAIIEADPLTTTREVAEELNVDH
STVVRHLKQIGKVKKLDKWVPHELSENQKNRRFEVSSSLILRNNNEPFLD
RIVTCDEKWILYDNRRRSAQWLDREEAPKHFPKPNLHQKKVMVTVWWS
AAGVIHYSFLNPGETITSEKYCQQIDEMHRKLQRLQPALVNRKGPILLHDN
ARPHVAQPTLQKLNELGYEVLPHPPYSPDLSPTDYHFFKHLDNFLQGKR
FHNQQDAENAFQEFVESRSTDFYATGINKLISRWQKCVDCNGSYFDASS
FMRKRLEAPTLKRYLRDNISNILPNEVPGTSDDSTEEPVMKKRTYCTYCP
SKIRRKANASCKKCKKVICREHNIDMCQSCF

Sequence of the Hsmar1 transposase monomer (red) with a 90 aa PHD domain (green) attached at the C-terminus. Shown as it was expressed, from pRC2511:

MEMMLDKKQIRAIFLFEFKMGRKAAETTRNINNAFGPGTANERTVQWWF
KKFRKGDESLEDEERSGRPSEVDNDQLRAIIEADPLTTTREVAEELNVDH
STVVRHLKQIGKVKKLDKWVPHELSENQKNRRFEVSSSLILRNNNEPFLD
RIVTCDEKWILYDNRRRSAQWLDREEAPKHFPKPNLHQKKVMVTVWWS
AAGVIHYSFLNPGETITSEKYCQQIDEMHRKLQRLQPALVNRKGPILLHDN
ARPHVAQPTLQKLNELGYEVLPHPPYSPDLSPTDYHFFKHLDNFLQGKR
FHNQQDAENAFQEFVESRSTDFYATGINKLISRWQKCVDCNGSYFDASN
LYMSLTSSFMRKRLEAPTLKRYLRDNISNILPNEVPGTSDDSTEE

8.1.3 DNA substrate sequences

Short substrate containing Tn5 TIR sequence (red), used in section 3.2.3:

```
GATCCAGAATTCGTGAGCGTGGGTCTCGCGGTCTGTCTCTTATACACATCTCAAC
    GTCTTAAGCACTCGCACCCAGAGCGCCAGACAGAGAATATGTGTAGAGTTG

CATCATCGACAAGCTTCTCGAGCCTAGGCTAGCTC
GTAGTAGCTGTTCGAAGAGCTCGGATCCGATCGAGATC
```

Long substrate containing Tn5 TIR sequence (red), used in section 3.2.3:

```
   CCACGCATGCTGCAGACGCGTTACGTATCGGATCCAGAATTCGTGAGCGTGG
CATGGTGCGTACGACGTCTGCGCAATGCATAGCCTAGGTCTTAAGCACTCGCACC

GTCTCGCGGTCTGTCTCTTATACACATCTCAACCATCATCGACAAGCTTCTCGAG
CAGAGCGCCAGACAGAGAATATGTGTAGAGTTGGTAGTAGCTGTTCGAAGAGCTC

CCTAGGCTAGCTCTAGACCACACGTGTGGGGGCCCGAGCTCGCGGCC
GGATCCGATCGAGATCTGGTGTGCACACCCCCGGGCTCGAGCGCCGG
```

Fluorescently tagged substrate with Tn5 TIR (red), used in section 4.2.3:

```
Cy5CGCAATTCACGGTCCAGATGTGTATAAGAGACAGAAGCTTGGCGGATCGC
  GCGTTAAGTGCCAGGTCTACACATATTCTCTGTCTTCGAACCGCCTAGCG
```

Fluorescently tagged substrate with ZBS (green), used in section 4.2.3:

```
Cy5TCAACGCGTGGGCGAGCTTAGGGTATACCCTAAAGGTGATGCTTAGTG
  AGTTGCGCACCCGCTCGAATCCCATATGGGATTTCCACTACGAATCAC
```

Fluorescently tagged substrate with Tn5 TIR (red) and a ZBS (green), used in section 4.2.3:

```
Cy5CACGGGGTGATGCTCAGATGTGTATAAGAGACAGAAGCTTGGCGGATCGC
  GTGCCCCACTACGAGTCTACACATATTCTCTGTCTTCGAACCGCCTAGCG
```

8.2 Professional internships for PhD student reflection form

Name of Organisation

New England Biolabs (NEB)

Details of Placement

The placement involved working in a laboratory environment, performing predominantly independent research looking into transposase enzymes. The work was mostly confidential in nature, but during my time I was able to move the project forward significantly. Along with the laboratory work, I was expected to present frequently at group meetings.

Placement Achievements

Within my time at NEB, I gave two full extended presentations to the whole company on my research for my PhD and on my research there, both of which were well received. I also managed to develop a good relationship with the group I was in, that facilitated further NEB placements for members of my PhD group.

Skill development

This placement significantly improved my laboratory skills, as I was working in a new environment and utilising several different techniques. My ability to present in front of an audience was also greatly enhanced, helped with a lot of feedback from both my supervisors there.

Future career directions

This placement has strongly tempted me to return to NEB in the future to work full time, and has also helped me to develop an interest in working in a non-academic setting elsewhere, though both academic and non-academic appeal at present.

8.3 Bibliography

Adams, C.D. et al., 2006. Tn5 transposase loops DNA in the absence of Tn5 transposon end sequences. *Molecular Microbiology*, 62(6), pp.1558–1568.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403–410.

Aravind, L., 2000. The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends in Biochemical Sciences*, 25(9), pp.421–423.

Aravind, L. et al., 2005. The many faces of the helix-turn-helix domain: Transcription regulation and beyond. *FEMS Microbiology Reviews*, 29(2), pp.231–262.

Arensburger, P. et al., 2011. Phylogenetic and Functional Characterization of the *hAT* Transposon Superfamily, *Genetics*, 188(1), pp.45–57.

Ason, B. & Reznikoff, W.S., 2002. Mutational Analysis of the Base Flipping Event Found in Tn5 Transposition. *Journal of Biological Chemistry*, 277(13), pp.11284–11291.

Atkinson, P.W., 2015. *hAT* Transposable Elements. *Microbiology Spectrum*, 3(4).

Baquero, F., Tedim, A.P. & Coque, T.M., 2013. Antibiotic resistance shaping multi-level population biology of bacteria. *Frontiers in Microbiology*, 4, pp.1-15.

Barabas, O. et al., 2008. Mechanism of IS200/IS605 Family DNA Transposases: Activation and Transposon-Directed Target Site Selection. *Cell*, 132(2), pp.208–220.

Bienz, M., 2006. The PHD finger, a nuclear protein-interaction domain. *Trends in Biochemical Sciences*, 31(1), pp.35–40.

Bischerour, J. & Chalmers, R., 2009. Base Flipping in Tn10 Transposition: An Active Flip and Capture Mechanism A, *PLoS ONE*, 4(7).

Bischerour, J. & Chalmers, R., 2007. Base-flipping dynamics in a DNA hairpin processing reaction. *Nucleic Acids Research*, 35(8), pp.2584–2595.

Boe, L., Gerdes, K. & Molin, S., 1987. Effects of genes exerting growth inhibition and plasmid stability on plasmid maintenance. *Journal of Bacteriology*, 169(10), pp.4646–4650.

Bonnefoy, E. & Rouvière-Yaniv, J., 1991. HU and IHF, two homologous histone-like proteins of *Escherichia coli*, form different protein-DNA complexes with short DNA fragments. *The EMBO journal*, 10(3), pp.687–696.

Boocock, M.R. & Rice, P.A., 2013. A proposed mechanism for IS607-family serine transposases. *Mobile DNA*, 4(1), p.1.

Bradford, M.M., 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, 72(1–2), pp.248–254.

Britten, R.J., 1996. Cases of Ancient Mobile Element DNA Insertions That Now Affect Gene Regulation. *Molecular Phylogenetics and Evolution*, 5(1), pp.13–17.

Buenrostro, J.D. et al., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12), pp.1213–1218.

Capy, P. et al., 2000. Stress and transposable elements: co-evolution or useful parasites? *Heredity*, 85(2), pp.101–106.

Carr, M., Bensasson, D. & Bergman, C.M., 2012. Evolutionary Genomics of Transposable Elements in *Saccharomyces cerevisiae, PLoS ONE*, 7(11).

Cary, L.C. et al., 1989. Transposon mutagenesis of baculoviruses: Analysis of *Trichoplusia ni* transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology*, 172(1), pp.156–169.

Chakravartty, V. & Cronan, J.E., 2013. The Wing of a Winged Helix-Turn-Helix Transcription Factor Organizes the Active Site of BirA, a Bifunctional Repressor/Ligase. *Journal of Biological Chemistry*, 288(50), pp.36029–36039.

Chalmers, R. & Kleckner, N., 1996. IS10/Tn10 transposition efficiently accommodates diverse transposon end configurations. *The EMBO journal*, 15(18), pp.5112–240.

Chandler, M. et al., 2015. Copy-out–Paste-in Transposition of IS911: A Major Transposition Pathway. *Microbiology Spectrum*, 3(4).

Chapados, B.R. et al., 2001. Structural biochemistry of a type 2 RNase H: RNA primer recognition and removal during DNA replication. *Journal of Molecular Biology*

Chen, X., Zaro, J.L. & Shen, W.-C., 2013. Fusion protein linkers: Property, design and functionality. *Advanced Drug Delivery Reviews*, 65(10), pp.1357–1369.

Chen, Z. & Ruffner, D.E., 1996. Modified crush-and-soak method for recovering oligodeoxynucleotides from polyacrylamide gel. *Biotechniques*, 21(5), pp.850-852.

Choi, K.Y. et al., 2013. Direct interaction between the TnsA and TnsB subunits controls the heteromeric Tn7 transposase. *Proceedings of the National Academy of Sciences*, 110(22), pp.E2038–E2045.

Ciubotaru, M. et al., 2013. RAG and HMGB1 create a large bend in the 23RSS in the V(D)J recombination synaptic complexes. *Nucleic Acids Research*, 41(4), pp.2437–2454.

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 Research*, 38(1), pp.190–202.

Claeys Bouuaert, C. et al., 2014. Crosstalk between transposase subunits during cleavage of the mariner transposon. *Nucleic Acids Research*, 42(9), pp.5799–5808.

Claeys Bouuaert, C. et al., 2013. The autoregulation of a eukaryotic DNA transposon. *eLife*, 2(0).

Claeys Bouuaert, C. & Chalmers, R., 2017. A single active site in the mariner transposase cleaves DNA strands of opposite polarity, *Nucleic Acids Research*.

Clubb, R.T. et al., 1994. A novel class of winged helix–turn–helix protein: the DNA–binding domain of Mu transposase. *Structure*, 2(11), pp.1041–1048.

Craig, N.L. et al., 2002. Mobile DNA II, *American Society for Microbiology*.

Curcio, M.J., Lutz, S. & Lesage, P., 2015. The Ty1 LTR-retrotransposon of budding yeast, *Saccharomyces cerevisiae*. *Microbiology Spectrum*, 3(2), pp.1-35.

Cuypers, M.G. et al., 2012. Solution conformations of early intermediates in Mos1 transposition. *Nucleic Acids Research*, 41(3), pp.2020–2033.

Davies, D.R. et al., 2000. Three-Dimensional Structure of the Tn5 Synaptic Complex Transposition Intermediate. *Science*, 289(5476), pp.77–85.

Dornan, J., Grey, H. & Richardson, J.M., 2015. Structural role of the flanking DNA in mariner transposon excision. *Nucleic Acids Research*, 43(4), pp.2424–2432.

Drozdetskiy, A. et al., 2015. JPred4: a protein secondary structure prediction server. *Nucleic Acids Research*, 43(W1), pp.W389–W394.

Duval Valentin, G., Marty Cointin, B. & Chandler, M., 2004. Requirement of IS911 replication before integration defines a new bacterial transposition pathway. *The EMBO Journal*, 23(19), pp.3897–3906.

Elkin, S.K. et al., 2005. A PHD Finger Motif in the C-Terminus of RAG2 Modulates Recombination Activity. *Journal of Biological Chemistry*, 280(31), pp.28701–28710.

Elrod-Erickson, M. et al., 1996. Zif268 protein–DNA complex refined at 1.6å: a model system for understanding zinc finger–DNA interactions. *Structure*, 4(10), pp.1171–1180.

Foley, J.E. et al., 2009. Targeted mutagenesis in zebrafish using customized zinc-finger nucleases. *Nature Protocols*, 4(12), pp.1855–1868.

Foster, T.J. et al., 1981. Three Tn10-associated excision events: Relationship to transposition and role of direct and inverted repeats. *Cell*, 23(1), pp.215–227.

Fugmann, S.D., 2001. RAG1 and RAG2 in V(D)J recombination and transposition. *Immunologic Research*, 23(1), pp.23–39.

Gaj, T., Gersbach, C.A. & Barbas, C.F., III, 2013. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, 31(7), pp.397–405.

Gay, N.J., Tybulewicz, V.L. & Walker, J.E., 1986. Insertion of transposon Tn7 into the *Escherichia coli glmS* transcriptional terminator. *Biochemical Journal*, 234(1), pp.111–117.

Giovannoni, S.J., 2005. Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science*, 309(5738), pp.1242–1245.

Goryshin, I.Y. & Reznikoff, W.S., 1998. Tn5 *in Vitro* Transposition. *Journal of Biological Chemistry*, 273(13), pp.7367–7374.

Goryshin, I.Y. et al., 2000. Insertional transposon mutagenesis by electroporation of released Tn5 transposition complexes. *Nature Biotechnology*, 18(1), pp.97–100.

Grabundzija, I. et al., 2013. Sleeping Beauty transposon-based system for cellular reprogramming and targeted gene insertion in induced pluripotent stem cells. *Nucleic Acids Research*, 41(3), pp.1829–1847.

Gradman, R.J. et al., 2008. A bifunctional DNA binding region in Tn5 transposase. *Molecular Microbiology*, 67(3), pp.528–540.

Grandbastien, M.-A., 1998. Activation of plant retrotransposons under stress conditions. *Trends in Plant Science*, 3(5), pp.181–187.

Guynet, C. et al., 2008. *In Vitro* Reconstitution of a Single-Stranded Transposition Mechanism of IS608. *Molecular Cell*, 29(3), pp.302–312.

Hall, B.G., 2000. Transposable elements as activators of cryptic genes in *E. coli*, *Genetica*, 107(181), pp. 181–187.

Haniford, D.B. & Ellis, M.J., 2015. Transposons Tn10 and Tn5. *Microbiology Spectrum*, 3(1).

Hare, S. et al., 2010. Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature*, 464(7286), pp.232–236.

Hare, S., Maertens, G.N. & Cherepanov, P., 2012. 3′-Processing and strand transfer catalysed by retroviral integrase in crystallo. *The EMBO Journal*, 31(13), pp.3020–3028.

Harshey, R.M., 2014. Transposable phage Mu. *Microbiology Spectrum*, 2(5).

Havecker, E.R., Gao, X. & Voytas, D.F., 2004. The diversity of LTR retrotransposons. *Genome Biology*, 5(6), p.225.

He, S. et al., 2013. IS200/IS605 family single-strand transposition: mechanism of IS608 strand transfer. *Nucleic Acids Research*, 41(5), pp.1-12.

He, S. et al., 2015. The IS200/IS605 Family and "Peel and Paste" Single-strand Transposition Mechanism. In *Mobile DNA III*. *American Society of Microbiology*, pp. 609–630.

Henssen, A.G. et al., 2016. Forward genetic screen of human transposase genomic rearrangements. *BMC Genomics*, 17(1), p.548.

Henssen, A.G. et al., 2015. Genomic DNA transposition induced by human PGBD5, *eLife*, 4.

Henssen, A.G. et al., 2017. PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nature Genetics*, 49(7), pp.1005–1014.

Hickman, A.B. & Dyda, F., 2016. DNA Transposition at Work. *Chemical Reviews*, 116(20), pp.12758–12784.

Hickman, A.B. et al., 2010. DNA recognition and the precleavage state during single-stranded DNA transposition in *D. radiodurans*. *The EMBO Journal*, 29(22), pp.3840–3852.

Hickman, A.B. et al., 2005. Molecular architecture of a eukaryotic DNA transposase. *Nature Structural & Molecular Biology*, 12(8), pp.715–721.

Hickman, A.B. et al., 2014. Structural Basis of *hAT* Transposon End Recognition by Hermes, an Octameric DNA Transposase from *Musca domestica*. *Cell*, 158(2), pp.353–367.

Hickman, A.B. et al., 2000. Unexpected Structural Diversity in DNA Recombination. *Molecular Cell*, 5(6), pp.1025–1034.

Hua-Van, A. et al., 2011. The struggle for life of the genome's selfish architects. *Biology Direct*, 6(1), pp.1-29.

Hughes, S.H., 2015. Reverse Transcription of Retroviruses and LTR Retrotransposons. *Microbiology Spectrum*, 3(2).

Huisman, O. et al., 1989. Mutational analysis of IS10's outside end. *The EMBO Journal*, 8(7), pp.2101–2109.

Ivics, Z. & Izsvák, Z., 2015. Sleeping Beauty Transposition. *Microbiology Spectrum*, 3(2).

J Sakai, R.M.C.N.K., 1995. Identification and characterization of a pre-cleavage synaptic complex that is an early intermediate in Tn10 transposition. *The EMBO Journal*, 14(17), p.4374.

Jagelská, E.B. et al., 2008. DNA topology influences p53 sequence-specific DNA binding through structural transitions within the target sites. *Biochemical Journal*, 412(1), pp.57–63.

Jiang, H., Yang, J.Y. & Harshey, R.M., 1999. Criss-crossed interactions between the enhancer and the att sites of phage Mu during DNA transposition. *The EMBO Journal*, 18(13), pp.3845–3855.

Jordan, M., 2004. Transfection of adherent and suspended cells by calcium phosphate. *Methods*, 33(2), pp.136–143.

Jursch, T. et al., 2013. Regulation of DNA transposition by CpG methylation and chromatin structure in human cells. *Mobile DNA*, 4(1), p.15.

Kalderon, D. et al., 1984. Sequence requirements for nuclear location of simian virus 40 large-T antigen. *Nature*, 311(5981), pp.33–38.

Kapitonov, V.V. & Jurka, J., 2005. RAG1 Core and V(D)J Recombination Signal Sequences Were Derived from *Transib* Transposons, *PLoS Biology*, 3(6), p.e181.

Kapitonov, V.V. & Jurka, J., 2001. Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, 98(15), pp.8714–8719.

Kapitonov, V.V. & Koonin, E.V., 2015. Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biology Direct*, 10(1), p.1035.

Katayanagi, K., Miyagawa, M. & Matsushima, M., 1990. Three-dimensional structure of ribonuclease H from *E. coli*. *Nature*, 307, pp. 306-309.

Kersulyte, D. et al., 2002. Transposable Element ISHp608 of *Helicobacter pylori*: Nonrandom Geographic Distribution, Functional Organization, and Insertion Specificity. *Journal of Bacteriology*, 184(4), pp.992–1002.

Kia, A. et al., 2017. Improved genome sequencing using an engineered transposase. *BMC Biotechnology*, 17(1), pp.1-10.

Kim, Y.J. et al., 2011. DNA sequence requirements for hobo transposable element transposition in *Drosophila melanogaster*. *Genetica*, 139(8), pp.985–997.

Klenchin, V.A. et al., 2008. Phosphate coordination and movement of DNA in the Tn5 synaptic complex: role of the (R)YREK motif. *Nucleic Acids Research*, 36(18), pp.5855–5862.

Kostrewa, D. et al., 2009. RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature*, 462(7271), pp.323–330.

Krementsova, E. et al., 1998. Mutational Analysis of the Mu Transposase contributions of two distinct regions of domain II to recombination. *Journal of Biological Chemistry*, 273(47), pp.31358–31365.

Krishnan, L. & Engelman, A., 2012. Retroviral Integrase Proteins and HIV-1 DNA Integration. *Journal of Biological Chemistry*, 287(49), pp.40858–40866.

Kruklitis, R., Welty, D.J. & Nakai, H., 1996. ClpX protein of *Escherichia coli* activates bacteriophage Mu transposase in the strand transfer complex for initiation of Mu DNA synthesis. *The EMBO Journal*, 15(4), pp.935-944.

Kuo, C.F. et al., 1991. DNA-protein complexes during attachment-site synapsis in Mu DNA transposition. *The EMBO Journal*, 10(6), pp.1585-1595.

Laity, J.H., Lee, B.M. & Wright, P.E., 2001. Zinc finger proteins: new insights into structural and functional diversity. *Current Opinion in Structural Biology*, 11(1), pp.39–46.

Leung, P.C., Teplow, D.B. & Harshey, R.M., 1989. Interaction of distinct domains in Mu transposase with Mu DNA ends and an internal transpositional enhancer. *Nature*, 338(6217), pp.656–658.

Le Rouzic, A., Boutin, T.S. & Capy, P., 2007. Long-term evolution of transposable elements. *Proceedings of the National Academy of Sciences*, 104(49), pp.19375–19380.

Li, X. et al., 2013. *piggyBac* transposase tools for genome engineering. *Proceedings of the National Academy of Sciences*, 110(25), pp.E2279–E2287.

Li, Z., Craig, N.L. & Peters, J.E., 2013. 'Transposon Tn7', in Roberts, P. et al., Bacterial Integrative Mobile Genetic Elements. pp.1-32.

Liu, Y. et al., 2007. A Plant Homeodomain in RAG-2 that Binds Hypermethylated Lysine 4 of Histone H3 Is Necessary for Efficient Antigen-Receptor-Gene Rearrangement. *Immunity*, 27(4), pp.561–571.

Lohe, A.R. & Hartl, D.L., 1996a. Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Molecular Biology and Evolution*, 13(4), pp.549–555.

Lohe, A.R. & Hartl, D.L., 1996b. Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Molecular Biology and Evolution*, 13(4), pp.549–555.

Lombardo, A. et al., 2007. Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nature Biotechnology*, 25(11), pp.1298–1306.

Maeder, M.L. et al., 2008. Rapid "Open-Source" Engineering of Customized Zinc-Finger Nucleases for Highly Efficient Gene Modification. *Molecular Cell*, 31(2), pp.294–301.

Maertens, G.N., Hare, S. & Cherepanov, P., 2010. The mechanism of retroviral integration through X-ray structures of its key intermediates. *Nature*, 468(7321), pp.326–329.

Marzo, M. et al., 2013. Identification of multiple binding sites for the THAP domain of the Galileo transposase in the long terminal inverted-repeats. *Gene*, 525(1), pp.84–91.

Matthews, A.G.W. et al., 2007. RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. *Nature*, 450(7172), pp.1106–1110.

McClintock, B., 1950. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6), pp.344–355.

McGinty, R.K. & Tan, S., 2015. Nucleosome Structure and Function. *Chemical Reviews*, 115(6), pp.2255–2273.

Mendiola, M.V. & la Cruz, de, F., 1992. IS91 transposase is related to the rolling-circle-type replication proteins of the pUB110 family of plasmids. *Nucleic Acids Research*, 20(13), p. 3521.

Milewski, S., 2002. Glucosamine-6-phosphate synthase—the multi-facets enzyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, 1597(2), pp.173–192.

Miskey, C. et al., 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. *Molecular and Cellular Biology*, 27(12), pp.4589–4600.

Mitra, R. et al., 2010. Characterization of the TnsD- attTn7 complex that promotes site-specific insertion of Tn7. *Mobile DNA*, 1(1), pp.1-14.

Mitra, R., Fain Thornton, J. & Craig, N.L., 2008. *piggyBac* can bypass DNA synthesis during cut and paste transposition. *The EMBO Journal*, 27(7), pp.1097–1109.

Mizuno, N. et al., 2013. MuB is an AAA+ ATPase that forms helical filaments to control target selection for DNA transposition. *Proceedings of the National Academy of Sciences*, 110(27).

Mizuuchi, M. & Mizuuchi, K., 1989. Efficient Mu transposition requires interaction of transposase with a DNA sequence at the Mu operator: Implications for regulation. *Cell*, 58(2), pp.399–408.

Montaño, S.P., Pigli, Y.Z. & Rice, P.A., 2012. The Mu transpososome structure sheds light on DDE recombinase evolution. *Nature*, 491(7424), pp.413–417.

Moore, R., Chandrahas, A. & Bleris, L., 2014. Transcription Activator-like Effectors: A Toolkit for Synthetic Biology. *ACS Synthetic Biology*, 3(10), pp.708–716.

Morris, E.R. et al., 2016. A bend, flip and trap mechanism for transposon integration. *eLife*, 5.

Naumann, T.A, & Reznikoff, W., 2002. Tn5 Transposase Active Site Mutants. *Journal of Biological Chemistry*, 277(20), pp.17623–17629.

Newman, J.C. et al., 2008. An Abundant Evolutionarily Conserved CSB-*PiggyBac* Fusion Protein Expressed in Cockayne Syndrome, *PLoS genetics*, 4(3).

Nicolas, E. et al., 2015. The Tn3-family of Replicative Transposons. *Microbiology Spectrum*, 3(4).

Nitiss, J. L. et al., 2012. Topoisomerase Assays. *Current Protocols in Pharmacology*.

Nojima, T, & Tellier, M. 2017. Personal Communication.

Nowotny, M., 2009. Retroviral integrase superfamily: the structural perspective. *EMBO reports*, 10(2), pp.144–151.

Orth, P. et al., 2000. Structural basis of gene regulation by the tetracycline inducible Tet repressor-operator system. *Nature Structural & Molecular Biology*, 7(3), pp.215–219.

Pace, N. & Weerapana, E., 2014. Zinc-Binding Cysteines: Diverse Functions and Structural Motifs. *Biomolecules*, 4(2), pp.419–434.

Pahl, A. & Flügel, R.M., 1995. Characterization of the Human Spuma Retrovirus Integrase by Site-directed Mutagenesis, by Complementation Analysis, and by Swapping the Zinc Finger Domain of HIV-1. *Journal of Biological Chemistry*, 270(7), pp.2957–2966.

Pardue, M.L., Danilevskaya, O.N. & Traverse, K.L., 1997. Evolutionary links between telomeres and transposable elements. *Genetica*, 100(1-3), pp.73-84.

Pattanayak, V. et al., 2011. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature Methods*, 8(9), pp.765–770.

Pavelitz, T. et al., 2013. PGBD5: a neural-specific intron-containing *piggyBac* transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. *Mobile DNA*, 4(1), pp.1-17.

Pavletich, N.P. & Pabo, C.O., 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, 252(5007), pp.809–817.

Pellicer, J., Fay, M.F. & Leitch, I.J., 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, 164(1), pp.10–15.

Peters, J.E., 2014. Tn7. *Microbiology Spectrum*, 2(5).

Peterson, G. & Reznikoff, W., 2003. Tn5 Transposase Active Site Mutations Suggest Position of Donor Backbone DNA in Synaptic Complex. *Journal of Biological Chemistry*, 278(3), pp.1904–1909.

Petzke, L. & Luzhetskyy, A., 2009. In vivo Tn5-based transposon mutagenesis of *Streptomycetes*. *Applied Microbiology and Biotechnology*, 83(5), pp.979–986.

Picelli, S. et al., 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research*, 24(12), pp.2033–2040.

Potter, C.J. & Luo, L., 2010. Splinkerette PCR for Mapping Transposable Elements in *Drosophila, PLoS ONE*, 5(4).

Pouget, N., Turlan, C. & Destainville, N., 2006. IS911 transpososome assembly as analysed by tethered particle motion. *Nucleic Acids Research,* 34(16), pp.4313-4323.

Qi, L.S. et al., 2013. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell*, 152(5), pp.1173–1183.

Roberts, A.P. & Mullany, P., 2009. A modular master on the move: the Tn916 family of mobile genetic elements. *Trends in microbiology*, 17(6), pp.251–258.

Reddy Chichili, V.P., Kumar, V. & Sivaraman, J., 2013. Linkers in the structural biology of protein–protein interactions. *Protein Science*, 22(2), pp.153–167.

Reznikoff, W.S., 2013. Personal communication.

Reznikoff, W.S., 2008. Transposon Tn5. *Annual Review of Genetics*, 42(1), pp.269–286.

Rice, P. & Kiyoshi, M., 1995. Structure of the bacteriophage Mu transposase core: A common structural motif for DNA transposition and retroviral integration. *Cell*, 82(2), pp.209–220.

Rice, P.A. et al., 1996. Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U-Turn. *Cell*, 87(7), pp.1295–1306.

Richardson, J.M. et al., 2009. Molecular Architecture of the Mos1 Paired-End Complex: The Structural Basis of DNA Transposition in a Eukaryote. *Cell*, 138(6), pp.1096–1108.

Richardson, S.R. et al., 2015. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiology Spectrum*, 3(2).

Rodgers, K.K., 2017. Riches in RAGs: Revealing the V(D)J Recombinase through High-Resolution Structures. *Trends in Biochemical Sciences*, 42(1), pp.72–84.

Ronning, D.R. et al., 2005. Active Site Sharing and Subterminal Hairpin Recognition in a New Class of DNA Transposases. *Molecular Cell*, 20(1), pp.143–154.

Ronning, D.R. et al., 2004. The carboxy-terminal portion of TnsC activates the Tn7 transposase through a specific interaction with TnsA. *The EMBO Journal*, 23(15), pp.2972–2981.

Rothbart, S.B. et al., 2012. Peptide Microarrays to Interrogate the “Histone Code” In Nucleosomes, Histones & Chromatin Part A. *Methods in Enzymology*, 512, pp.107–135.

Ru, H. et al., 2015. Molecular Mechanism of V(D)J Recombination from Synaptic RAG1-RAG2 Complex Structures. *Cell*, 163(5), pp.1138–1152.

Rubin, E., Lithwick, G. & Levy, A.A., 2001. Structure and Evolution of the *hAT* Transposon Superfamily. *Genetics*, 158(3), pp.949–957.

Sabogal, A. et al., 2010. THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves. *Nature Structural & Molecular Biology*, 17(1), pp.117–123.

Sakai, J.S. et al., 2000. Tn10 transpososome assembly involves a folded intermediate that must be unfolded for target capture and strand transfer. *The EMBO Journal*, 19(4), pp.776–785.

Sanchez, R. & Zhou, M.-M., 2011. The PHD finger: a versatile epigenome reader. *Trends in Biochemical Sciences*, 36(7), pp.364-372.

Schatz, D.G. & Swanson, P.C., 2011. V(D)J Recombination: Mechanisms of Initiation. *Annual Review of Genetics*, 45(1), pp.167–202.

Shaheen, M. et al., 2010. Metnase/SETMAR: a domesticated primate transposase that enhances DNA repair, replication, and decatenation. *Genetica*, 138(5), pp.559–566.

Shapiro, J.A., 1979. Molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. *Proceedings of the National Academy of Sciences*, 76(4), pp.1933–1937.

Sharmistha Majumdar, D.C.R., 2015. P transposable elements in *Drosophila* and other eukaryotic organisms. *Microbiology Spectrum*, 3(2).

Siguier, P. et al., 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research*, 34(1), pp.D32-36

Sos, B.C. et al., 2016. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biology*, 17(1), pp.1-15

Steczkiewicz, K. et al., 2012. Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. *Nucleic Acids Research*, 40(15), pp.7016–7045.

Steiniger-White, M. & Reznikoff, W.S., 2000. The C-terminal α Helix of Tn5 Transposase Is Required for Synaptic Complex Formation. *Journal of Biological Chemistry*, 275(30), pp.23127–23133.

Steitz, T.A. & Steitz, J.A., 1993. A general two-metal-ion mechanism for catalytic RNA. *Proceedings of the National Academy of Sciences*, 90(14), pp.6498–6502.

Stott, K. et al., 2006. Structure of a Complex of Tandem HMG Boxes and DNA. *Journal of Molecular Biology*, 360(1), pp.90–104.

Tellier, M., Bouuaert, C.C. & Chalmers, R., 2015. *Mariner* and the ITm Superfamily of Transposons. *Microbiology Spectrum*, 3(2).

Thomas, J. & Pritham, E.J., 2015. Helitrons, the Eukaryotic Rolling-circle Transposable Elements. *Microbiology Spectrum*, 3(4).

Trubitsyna, M. et al., 2017. Use of *mariner* transposases for one-step delivery and integration of DNA in prokaryotes and eukaryotes by transfection. *Nucleic Acids Research*, 45(10).

Uhlén, M. et al., 2015. Tissue-based map of the human proteome. *Science*, 347(6220), pp.1260419–1260419.

Uren, A.G. et al., 2009. A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nature protocols*, 4(5), pp.789–798.

van Pouderoyen, G. et al., 1997. Crystal structure of the specific DNA-binding domain of Tc3 transposase of *C.elegans* in complex with transposon DNA. *The EMBO Journal*, 16(19), pp.6044–6054.

Voigt, F. et al., 2016. *Sleeping Beauty* transposase structure allows rational design of hyperactive variants for genetic engineering. *Nature Communications*, 11126.

Voigt, K. et al., 2012. Retargeting Sleeping Beauty Transposon Insertions by Engineered Zinc Finger DNA-binding Domains. *Molecular Therapy*, 20(10), pp.1852–1862.

Vos J C, R.H.P., 1994. Tc1 transposase of Caenorhabditis elegans is an endonuclease with a bipartite DNA binding domain. *The EMBO journal*, 13(24), pp.6125-6132.

Wang, H., Smith, M.C.M. & Mullany, P., 2006. The Conjugative Transposon Tn5397 Has a Strong Preference for Integration into Its Clostridium difficile Target Site. *Journal of Bacteriology*, 188(13), pp.4871–4878.

Wang, H., La Russa, M. & Qi, L.S., 2016. CRISPR/Cas9 in Genome Editing and Beyond. *Annual Review of Biochemistry*, 85(1), pp.227–264.

Wang, Y. et al., 2014. Suicidal Autointegration of Sleeping Beauty and *piggyBac* Transposons in Eukaryotic Cells, *PLoS Genetics*, 10(3).

Waters, J.L. & Salyers, A.A., 2013. Regulation of CTnDOT Conjugative Transfer Is a Complex and Highly Coordinated Series of Events. *MBio*, 4(6)

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 Research*, 32(14), pp.4306–4312.

Whitfield, C.R., Wardle, S.J. & Haniford, D.B., 2006. Formation, Characterization and Partial Purification of a Tn5 Strand Transfer Complex. *Journal of Molecular Biology*, 364(3), pp.290–301.

Wiegand, T.W. & Reznikoff, W.S., 1992. Characterization of two hypertransposing Tn5 mutants. *Journal of Bacteriology*, 174(4), pp.1229–1239.

Weinreich, M.D., Gasch, A. & Reznikoff, W.S., 1994. Evidence that the cis preference of the Tn5 transposase is caused by nonproductive multimerization. *Genes & Development*, 8(19), pp.2363–2374.

Wolkow, C.A., DeBoy, R.T. & Craig, N.L., 1996. Conjugating plasmids are preferred targets for Tn7. *Genes & Development*, 10(17), pp.2145–2157.

Yang, S. et al., 1990. Structure of ribonuclease H phased at 2 Angstrom resolution by MAD analysis of the selenomethionyl protein. *Science*, 249(4975), pp.1398–1406.

Yin, F.F. et al., 2009. Structure of the RAG1 nonamer binding domain with DNA reveals a dimer that mediates DNA synapsis. *Nature Structural & Molecular Biology*, 16(5), pp.499–508.

Yin, Z. et al., 2012. Assembly of prototype foamy virus strand transfer complexes on product DNA bypassing catalysis of integration. *Protein Science*, 21(12), pp.1849–1857.

York, D. & Reznikoff, W.S., 1996. Purification and Biochemical Analyses of a Monomeric Form of Tn5 Transposase. *Nucleic Acids Research*, 24(19), pp.3790–3796.

Yusa, K., 2015. *piggyBac* Transposon. *Microbiology Spectrum*, 3(2).

Z Wu, G.C., 1995. A novel DNA binding and nuclease activity in domain III of Mu transposase: evidence for a catalytic region involved in donor cleavage. *The EMBO Journal*, 14(15), p.3835.

Zhang, A.P.P., Pigli, Y.Z. & Rice, P.A., 2010. Structure of the LexA-DNA complex and implications for SOS box measurement. *Nature*, 466, pp.883–886.

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

Zhou, M., Bhasin, A. & Reznikoff, W.S., 1998. Molecular genetic analysis of transposase-end DNA sequence recognition: cooperativity of three adjacent base-pairs in specific interaction with a mutant Tn5 transposase. *Journal of Molecular Biology*, 276(5), pp.913–925.

Zhou, M. & Reznikoff, W.S., 1997. Tn5 Transposase Mutants that Alter DNA Binding Specificity. *Journal of Molecular Biology*, 271(3) pp.1–12.