Genome manipulation using l-integrase-mediated recombination

Harshyaa Makhija

2017

Harshyaa Makhija. (2017). Genome manipulation using l-integrase-mediated recombination. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/70616>

<https://doi.org/10.32657/10356/70616>

**GENOME MANIPULATION USING
 λ -INTEGRASE-MEDIATED RECOMBINATION**

HARSHYAA MAKHIJA

School of Biological Sciences

**A thesis submitted to the Nanyang Technological University in
partial fulfillment of the requirement for the degree of Doctor
of Philosophy**

2016

Acknowledgements

I take this opportunity to extend my immense gratitude to my supervisor Prof. Peter Dröge, NTU, for his constant support, guidance and his willingness to teach and share his knowledge during my research. His tremendous enthusiasm in science has been a source of inspiration throughout my PhD studies.

I would like to deeply thank Dr. Farid Ghadessy and his team at A*Star, Singapore for their excellent collaboration and providing us with integrase mutants for our experiments. I am very grateful to Prof. Li Jinming and his colleagues for helping us with the bioinformatics analysis.

Special thanks go to Dr Shree Harsha Vijaya Chandra for his support and motivation throughout the project. I would like to thank him and Sabrina Peter for their contribution in cell-based targeting assays and data analyzing.

I am very grateful to Roland Ivanyi-Nagy for providing valuable suggestions for my thesis and my friends Priya Dharshana, Pravesh Gupta and Dharendra Singh for their help and moral support during my research.

My appreciation also goes to other lab members Amir Morshedi, Martin Cai Hui and Martina Stagno D'Alcontres for providing a pleasant environment during my research.

I am grateful for the financial support of scholarship from Nanyang Technological University and for the generous funding by Exploit Pte Ltd, A*Star during my study.

Last but not the least, I would like to thank my friends, for their encouragement and support and my family, for their love and patience during my study.

Page intentionally left blank

Table of Contents

Acknowledgements i

Table of Contents iii

Table Captionsvii

Figure Captionsviii

Abbreviations xiii

Abstract.....xvii

1. IntroductionI

1.1 Overview of genome editing tools.....I

1.2 Genome manipulation.....4

1.2.1 Random integration into the genome.....4

1.2.1.1 Plasmid-based integration4

1.2.1.2 Transposon-based system for gene delivery5

1.2.1.3 Viral-mediated gene delivery6

1.2.2 Site- or sequence-specific transgene insertion into the genome8

1.2.2.1 Gene targeting via homologous recombination (HR)9

1.2.2.2 Enhancing HR via induced double strand breaks (DSBs)11

1.2.2.3 Site-specific endonucleases for genome engineering.....12

1.2.2.3.1 Zinc finger nucleases (ZFNs)12

1.2.2.3.2 Transcription activator like effector nucleases (TALENs)14

1.2.2.3.3 CRISPR/Cas9 genome editing tool16

1.2.2.4 Site-specific recombinases for genome engineering17

1.2.2.4.1 Serine-type phage integrases-mediated recombination18

1.2.2.4.2 Cre-loxP/Flp-FRT system20

1.2.2.4.3 λ-Int-mediated site-specific recombination22

1.3 Aim of the project26

2. Materials and Methods27

2.1 List of primers used in the study27

2.2 Vector constructs 33

2.3 Cell lines40

2.4 Cell culture conditions 41

2.5 Differentiation of hESCs41

2.6 Transfections42

2.7 Antibiotic selection and screening for targeted cell clones 43

2.8 RNA extraction and cDNA synthesis 43

2.9 Identification of sequence-specific recombination events by PCR screening 43

2.10 Inverse PCR44

2.11 Southern blotting 44

2.12 Cell proliferation assay 44

2.13 Karyotyping 45

2.14 Flow cytometry 45

2.15 Transcriptome analysis 45

3. Results	49
3.1 Development of λ -Int mutants	49
3.2 Comparative analysis of various Int mutants	53
3.3 Targeting endogenous <i>attH</i> -MCT5 using mutant Int-h/218	56
3.3.1 Determination of <i>MCT5</i> expression in HT1080 cells	57
3.3.2 An episomal DNA-based intermolecular recombination assay to determine the functionality of the <i>MCT-5</i> locus post-targeting	58
3.3.3 Targeting the genomic <i>attH</i> -MCT5 locus using Int-h/218	59
3.4 Searching for other endogenous <i>att</i> sites using an inverse PCR strategy	61
3.5 Identification of a potential endogenous <i>attH</i> sequence using <i>attP39</i> and Int-h/218	62
3.5.1 Unexpected targeting at chromosome 13 instead of chromosome 6 using Int-h/218	66
3.6 Targeting endogenous <i>attH</i> sequences using wild-type <i>attP</i> along with different Int mutants (Int-h/218, Int-C2, Int-C3)	76
3.7 A functional human <i>attH</i> site	80
3.7.1 Identification of <i>attH4X</i> sequence	80
3.7.2 Targeting of <i>attH4X</i> in <i>LINE-1</i> elements	84
3.7.3 Targeting of the <i>attH4X</i> in hESCs	88
3.7.4 Functionality of targeted <i>attH4X</i> loci in hESCs	94
3.8 Targeting of <i>LINE-1</i> using <i>attL4X</i> and Int-C3CNLS	95
3.9 Summary of the targeting events in hESCs (<i>attH4X</i> X <i>attP4X</i>) and HT1080 cells (<i>attH4X</i> X <i>attL4X</i>)	100
3.10 <i>In vitro</i> intra-molecular recombination for generation of seamless vector	102
3.11 Safety evaluation upon λ -Int-mediated targeting	104
3.11.1 Cell proliferation assay	104

3.11.2 DNA damage response105

3.11.3 Karyotype analysis.....105

3.11.4 *attH4X* as putative safe harbor site.....106

4. Discussion.....111

5. Future work.....121

References123

Appendix139

Publications and Patent.....151

Conference presentations153

Table Captions

Table 1.1 Characteristics of different viral vectors along with some of their applications in gene therapy8

Table 1.2 Characteristics of site-specific recombinase families18

Table 3.1 Summary of Southern analysis for the targeted clones in hESC (A3, E3, K3) using *pattP4X-PGKssPuro-UTF1-EGFP* and *Int-C3CNLS*93

Table 3.2 Summary of Southern analysis for the targeted clones in HT1080 (19, 26, 29, 31, 32) using *pTZ18R-attL4X/attR4X-PGKssPuro-UTF1-EGFP* and *Int-C3CNLS*.....100

Table 3.3 Summary of the targeting events at *attH4X* resulting in recombinant *attL* and *attR* sites in hESCs and HT1080 cells101

Table 3.4 Transcriptome analysis showing expression (log values) for upregulated genes (> log 2) and downregulated genes (< log2) along with their chromosomal location in hESC clones (#3, #24, E3)108

Table 4.1 Targeting in various cell lines115

Page intentionally left blank

Figure Captions

Figure 1.1 Genome manipulation tools.....4

Figure 1.2 Viral-mediated gene delivery7

Figure 1.3 Homologous recombination10

Figure 1.4 Working principles of ZFNs, TALENs and CRISPR/Cas9 endonucleases12

Figure 1.5 An illustration of a ZFN13

Figure 1.6 A schematic diagram of a TALEN15

Figure 1.7 An illustration of a CRISPR/Cas916

Figure 1.8 Steps involved in serine-integrase-mediated recombination19

Figure 1.9 Genome manipulations using the Cre/*loxP* system21

Figure 1.10 Overview of λ -Int-mediated site-specific recombination.....23

Figure 1.11 λ -Int-mediated recombination via a Holliday junction model.....25

Figure 2.1 A vector map of *pCMVssInt-h/218*.....33

Figure 2.2 A vector map of *pattPH-MCT5-cherry*35

Figure 2.3 A vector map of *pattPH-MCT5-cherry-PGKss-puro*35

Figure 2.4 A vector map of *pPGKss-exon2-5 MCT5*36

Figure 2.5 A vector map of *pattP(chr6)-PGKss-neo*37

Figure 2.6 A vector map of *pattP4X-PGKss-puro*38

Figure 2.7 A vector map of *pEF1ass-Int-C3CNLS*.....39

Figure 2.8 A vector map of *pattP4X-PGKsspuro-UTF1-EGFP*39

Figure 2.9 A vector map of *pTZ18R-attL/attR-PGKssPuro-UTF1-EGFP* 40

Figure 3.1 Sequence alignments of WT core *attB/attP* sites with the human *attH/attPH-MCT5* sequence50

Figure 3.2 An illustration showing positions of I43 and H61 in the N-terminal domain of λ -Int51

Figure 3.3 Generation of Int mutants via substrate linked protein evolution (SLiP)52

Figure 3.4 An illustration showing positions of D336 and E319 in the C-terminal domain of λ -Int.....53

Figure 3.5 An episomal DNA-based recombination assay54

Figure 3.6 Comparative analysis of Int-h/218, Int-1 and Int-555

Figure 3.7 Comparative analysis of Int-h/218, Int-C2 and Int-C356

Figure 3.8 Determination of *MCT5* expression57

Figure 3.9 An episomal DNA-based intermolecular recombination test to ensure functionality the of the *MCT5* locus58

Figure 3.10 PCR analysis of episomal DNA-based intermolecular recombination59

Figure 3.11 Targeting of the genomic *attH*-*MCT5* by *pattPH-MCT5-cherry* using *Int-h/218*60

Figure 3.12 Determination of potential genomic integration sites of the target vector using SMART cDNA synthesis kit61

Figure 3.13 Identification of potential endogenous *attH* sequence using *attP39*64

Figure 3.14 Confirmatory PCR for *attH*-chr6 targeting65

Figure 3.15 Targeting at chromosome 1367

Figure 3.16 Sequencing result of *attR* recombinant site on chromosome 1368

Figure 3.17 Confirmatory PCR for *attH*-chr13 targeting69

Figure 3.18 Identification of *attL* recombinant site on chromosome 1371

Figure 3.19 Determination of *attL* and *attR* sites on chromosome 13 in clone K88.72

Figure 3.20 Determination of the *attL* site on in clone K8874

Figure 3.21 Sequencing result showing *attR* and *attL* recombination sites in clone K8875

Figure 3.22 Targeting endogenous *attH* sequences using *attP* (WT)77

Figure 3.23 Identification of *attH*-chr878

Figure 3.24 Confirmatory PCR for *attH*-chr8 targeting79

Figure 3.25 Bioinformatics approach for identification of *attH* sequence.....80

Figure 3.26 Inverse PCR for identification of *attH*82

Figure 3.27 An illustration of a *LINE-1* retrotransposon83

Figure 3.28 Alignment of *attH4X* sequence84

Figure 3.29 Screening for *attH4X* x *attP4X* recombination events in HT1080 clones85

Figure 3.30 Southern blot analysis of targeted HT1080 clones87

Figure 3.31 Targeting of *attH4X* in hESCs.89

Figure 3.32 Screening for *attH4X* X *attP4X* recombination events in hESC clones90

Figure 3.33 Verification of *attL* recombinant site in hESC clones.....91

Figure 3.34 Verification of *attR* recombinant site in hESC clones.....92

Figure 3.35 Southern blot analysis of targeted hESC clones93

Figure 3.36 Functionality of targeted *attH4X* loci in hESC clones95

Figure 3.37 Generation of seamless vector96

Figure 3.38 Targeting of genomic *attH4X* using *attL4X*97

Figure 3.39 PCR screening for unrecombined *attR4X*98

Figure 3.40 Southern blot analysis of *attL4X* X *attH4X* targeted HT1080 clones99

Figure 3.41 *In vitro* intra-molecular recombination for generation of seamless vector102

Figure 3.42 Restriction check after *in vitro* intra-molecular reaction.....103

Figure 3.43 Cell proliferation assay105

Figure 3.44 Karyotyping to verify chromosomal stability.....106

Figure 4.1 *In vitro* generation of seamless vector.....117

Figure A1 Alignment of *attH4X* sequence139

Figure A2 Western Blot analysis for pluripotency markers Oct-4, Nanog and Sox2 in the targeted hESC clones and differentiated derivatives149

Page intentionally left blank

Abbreviations

AAVS1	Adeno-Associated Virus Integration Site 1
ALD	AsadrenoLeukoDystrophy
amp	Ampiciilin resistance gene
attB	Bacteria Attachment Site
attP	Phage Attachment Site
BLIP	β-lactamase inhibitory protein
CAR	Chimeric Antigen Receptor
chr	Chromosome
CMV	CytoMegaloVirus
CNLS	C-terminal Nuclear Localization Signal
Cre	Cyclic Recombinase
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DMEM	Dulbecco’s Modified Eagle Medium
DS	Docking Site
DSB	Double Strand Break
E.coli	Escherichia Coli
EF	Elongation Factor
EGFP	Enhanced Green Fluorescent Protein
ESCs	Embryonic Stem Cells
FACS	Fluorescence-activated cell sorting
FBS	Fetal Bovine Serum
FLP	Flippase
FRT	Flippase Recognition Target
FSC	Forward Scatter Curve
HDR	Homology Directed Repair
HIS	Histidine
HR	Homologous Recombination
HU	Hydroxy Urea

IHF	Integration Host Factor
Ina	Inactive
iPSCs	Induced Pluripotent Stem Cells
ITR	Inverted Terminal Repeats
IVC	<i>In Vitro</i> Compartmentalization
KZ	Kozak
<i>LINE-1</i>	Long INterspersed Elements
LoxP	Locus of X-over P1
M	100bp ladder
M1	1kb ladder
MCT5	MonoCarboxylate Transporter 5
mESCs	Mouse Embryonic Stem Cells
Neo	Neomycin
NHEJ	Non-Homology End Joining Repair
ORF	Open Reading Frame
PAM	Protospacer Adjacent Motif
PCR	Polymerase Chain Reaction
PGK	Phosphoglycerate Kinase
pre-crRNA	Precursor CRISPR RNA
Puro	Puromycin
RA	Retinoic Acid
RMCE	Recombinase Mediated Cassette Exchange
RMDI	Recombinase Mediated DNA Insertion
RT	Reverse Transcription
RVD	Repeat Variable Di-residue
sgRNA	Short Guide RNA
SLiP	Substrate Linked Protein Evolution
SMART	Switching Mechanism at 5' end of RNA Template
SSC	Side Scatter Curve

TALEN	Transcription Activator Like Effector Nucleases
TMAP	Torrent Mapping Alignment Program
tracrRNA	Trans-activating RNA
UTF1	Undifferentiated Transcription Factor 1
WT	Wild Type
XIS	Excisionase
ZFN	Zinc Finger Nuclease
ZFP	Zinc Finger Protein
λ-Int	lambda integrase

Page intentionally left blank

ABSTRACT

Precise and safe genome engineering of mammalian cells plays an important role in biotechnology and molecular medicine. Several random (viral, transposon, plasmid-based) and site-specific (endonuclease-based) genome-editing tools have been previously employed to meet the scientific and translational needs. They could, for example, help to solve fundamental biological problems and be employed for therapeutic purposes. However, inherent safety issues such as genotoxicity, insertional mutagenesis, off-target site activities and other unforeseen risks make existing methods suboptimal for clinical applications. Hence, improved genome manipulation tools are needed. In this context, we have developed a novel site-specific transgene insertion tool for the human genome. Our system is derived from phage lambda integrase (λ -Int) that not only exhibits high target site specificity, but also integrates large DNA molecules into a safe harbor site of a bacterial genome. To this end, a novel, highly active Int variant (Int-C3) has been jointly developed with collaborators, which catalyzes site-specific transgene insertion into a particular sequence (*attH4X*) found in a subset of human Long INterspersed Elements (*LINE-1*). We have validated our system for single-copy integration of transgenes (~8kb) in various cell lines including human embryonic stem cells (hESCs). The integrated transgenes remain stable and functional, perhaps due to a more permissive chromatin structure at the targeted *LINE-1* elements. In addition, our safety profiling data indicates that expression of Int is safe with respect to cell toxicity and genomic integrity. Transcriptome analysis of the targeted clones revealed 20-40/20,000 (~0.2%) differentially expressed genes, suggesting that the global cellular RNA profile remains largely undisturbed. It has also been known that *LINE-1* elements are more prevalent in AT-rich, low-recombining and sparse gene regions of the genome. Thus, based on our studies, at least a subset of *attH4X* can be considered as putative human genome 'safe harbor sites'. Therefore, predetermined targeting using our system will reduce or even eliminate the problems associated with other currently used transgenesis tools. Our tool would thus be an addition to the existing genome editing toolbox. It will find broad applications in stem cell-related therapies, bio-production of therapeutic proteins, CAR-T cell engineering and other human genome engineering applications that require multi-transgenes.

Page intentionally left blank

1. INTRODUCTION

1.1 Overview of genome editing tools

Genome manipulation of eukaryotic system plays an important role in both fundamental and applied research. It can be achieved by both specific and random integration of foreign DNA into the genome. However, specific targeted integration is preferred over the latter to prevent unpredictable transgene expression, insertional mutagenesis, proximal activation of neighboring genes and cell toxicity (Maggio and Goncalves 2015). However, until the 1990s, precise targeting of genomes was restricted to the method of homologous recombination (HR), which was mainly used for generation of knock in/out transgenic mouse models. Owing to the low efficiency of HR in somatic cells and a high rate of random non-homologous chromosomal insertion, researchers resorted to more efficient but non-targeted viral- and transposon-based genome modification methods for somatic cells (Vasquez, Marburger et al. 2001, Maggio and Goncalves 2015).

Different types of viral-based tools derived from retroviral, viral and adenoviral systems are being used to deliver desired transgenes depending on application requirements (Howarth, Lee et al. 2010, Yi, Noh et al. 2011). Viral-mediated transgenesis results in stable and long-term expression of transgenes, but several studies in the past few years have highlighted associated safety issues such as genotoxicity, insertional mutagenesis, oncogenesis and immunotoxicity (Modlich, Bohne et al. 2006, Montini, Cesana et al. 2006, Haccin-Bey-Abina, Garrigue et al. 2008, Nayak and Herzog 2010). In addition, production of viral vectors is expensive, time-consuming and requires extensive biological testing to comply with safety regulations (Modlich, Bohne et al. 2006, Montini, Cesana et al. 2006, Haccin-Bey-Abina, Garrigue et al. 2008, Nayak and Herzog 2010, Silva, Poirot et al. 2011). Another alternative approach to random chromosomal integration is based on DNA transposable elements found in vertebrate system (Miskey, Izsvak et al. 2005, Mates, Chuah et al. 2009, Silva, Poirot et al. 2011). Since transposition is essentially also a random process, attempts have also been made to improve the site specificity of transposons by fusing them with various DNA binding domains. However, its applicability in eukaryotic system has not yet been successfully demonstrated (Wilson, Kaminski et al. 2005, Maragathavally, Kaminski et al. 2006, Yant, Huang et al. 2007, Voigt, Gogol-Doring et al. 2012).

The development of efficient and reliable tools to make precise and controlled genomic modifications in mammalian cells, ranging from knock in/out to single nucleotide variations at specific genomic regions, has been a long standing challenge for biomedical researchers. This was addressed by the development of various sequence specific endonucleases in the past decade, such as zinc finger nucleases (ZFNs), transcription activator like effector nucleases (TALENs) and more recently clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 endonucleases (Sorrell and Kolb 2005, Mahfouz, Piatek et al. 2014).

ZFNs and TALENs are engineered nucleases composed of a customizable DNA binding domain fused to the restriction enzyme *FokI*. The CRISPR/Cas9 technology uses a 20bp single guide RNA with Cas9 endonuclease to induce double strand breaks (DSBs) at a specific site in the genome. These sites are repaired either by homology directed repair (HDR) leading to targeted insertions/mutagenesis (Bibikova, Beumer et al. 2003) or by non-homologous end joining (NHEJ) for knockout purposes (van Gent, Hoeijmakers et al. 2001). These methods have enabled researchers to alter the genome in mammalian systems including human ES cells, thus allowing germ line manipulation of genomes. In contrast to ZFNs and TALENs, which require specialization in protein engineering for target plasmid construction, CRISPR/Cas9 has become a preferred tool as it can be easily adapted to target nearly any genomic sequence by changing the sequence of the 20bp that is a part of guide RNA. In addition, the ease of use of CRISPR/Cas9 has enabled scientists to use it for multiplexing, i.e. to target multiple sites simultaneously (Cong, Ran et al. 2013). However, one of the concerns for the aforementioned nucleases is “off-target” effect, which can lead to genotoxicity, chromosomal aberrations or unwanted mutations due to induction of DSBs at multiple sites apart from targeted site, limiting their applicability for therapeutic purposes (Gaj, Gersbach et al. 2013). Various efforts to improve the specificity of these tools are still under way (Gupta and Musunuru 2014).

Apart from the use of endonucleases, site-specific recombinases (tyrosine and serine-based) also represent an elegant tool in terms of targeting the genome as they carry out autonomous integration, i.e they do not rely on the cellular machinery or homologous recombination for their activity (Silva, Poirot et al. 2011). Tyrosine recombinases such as Cre or Flp are widely used in conjunction with HR in developing conditionally-activated and knock out experimental models. This is done by specific insertion of their cognate sites (*loxP*, *FRT*) as docking sites into the genome (by providing homology arms) that can

be repeatedly targeted by their compatible sites for DNA insertion/deletion/inversion purposes (Thyagarajan, Olivares et al. 2001). Another tyrosine recombinase, lambda integrase (λ -Int) which catalyzes insertion of the phage genome into the bacterial genome via the phage attachment site (*attP*) and the bacterial attachment site (*attB*) (Landy 1989) has also been exploited for manipulating eukaryotic genomes. In this context, various Int mutants such as Int-h/218 (Lorbach, Christ et al. 2000, Tan and Droge 2005) and Int-C3 (Siau, Chee et al. 2015) have been generated to carry out recombination independently of co-factors. A serine-based site-specific recombinase, PhiC31 integrase has been used to target endogenous so-called pseudo *attP* sites in the human genome, but low specificity and efficiency has limited its use for targeting (Thyagarajan, Olivares et al. 2001, Liu, Skjorring et al. 2009).

Even though considerable progress has been made in the field of genome engineering, delivery of large constructs, efficacy and specificity of genome manipulation tools are some of the concerns that need to be addressed for various downstream applications. However, given the rapid pace of technological developments being made in this field, it is likely that improvements of current tools will pave the way for better alternatives for genome engineering.

The next section will introduce a more detailed view of the different genome manipulation tools that are used in basic science and clinical applications.

1.2 Genome manipulation

Genome manipulation technologies play a central role in a wide range of applications, such as stem cell engineering, gene therapy, synthetic biology and biotechnological research. Currently, various genome editing tools are available spanning from non-targeted recombination tools (plasmids, viral vectors, transposons) to site-specific recombinases (Cre, Flp, PhiC31, λ -Int) and recently developed endonucleases (ZFNs, TALENs and CRISPR/Cas9) (Sorrell and Kolb 2005, Mahfouz, Piatek et al. 2014). These tools, although being widely used to achieve stable and safe transgene knock-in into mammalian genomes, still have serious limitations.

Genome-engineering tools can be broadly divided into two categories: random and site-specific (as illustrated in **Figure 1.1**).

Figure 1.1 Genome manipulation tools. An illustration of the various available genome engineering tools for carrying out either random or specific integration of a transgene into the genome

1.2.1 Random integration into the genome

1.2.1.1 Plasmid-based integration

Plasmid-mediated introduction of foreign DNA into cellular genomes is one of the most widely used methods. Plasmid transfection can be done either by chemical (cationic or polymer lipids) or physical means (electroporation or microinjection) (Niidome and Huang 2002). Due to its simplicity and feasibility, plasmid-mediated integration is routinely used in basic molecular biology research.

The transfected plasmid initially remains as episomal DNA in the cytoplasm as its passage into the nucleus mainly depends on the cellular processes (Wurtele, Little et al. 2003). A large proportion of plasmid DNA is rapidly degraded or diluted with subsequent

cell divisions. Hence, a very low amount of plasmid eventually reaches the nucleus and integrates into the genome (Wurtele, Little et al. 2003). Randomly integrated DNA can possibly lead to unpredictable and uncontrolled transgene expression, insertional mutagenesis or even activation of oncogenes (Yew, Zhao et al. 2002). Since plasmids are produced in bacteria, their unmethylated CpGs can generate host immune responses in eukaryotic systems and often lead to epigenetic silencing of transgenes (Niidome and Huang 2002, Hodges, Taylor et al. 2004). Even though plasmid production and transfection is cost-effective, the randomness, unstable transgene expression and other risks have limited its use for genetic manipulation (Yew, Zhao et al. 2002).

1.2.1.2 Transposon-based system for gene delivery

To overcome the limitation of unstable gene expression, another genetic tool based on DNA transposons was developed to improve the functionality of introduced foreign DNA by enhancing the rate of integration into the host cell genome (Hackett, Ekker et al. 2005, VandenDriessche, Ivics et al. 2009).

Transposons are naturally occurring, mobile genetic elements composed of two major components, the inverted terminal repeats (ITRs) and a gene encoding the enzyme transposase. The transposase recognizes the ITRs to induce the excision and integration of the intervening DNA, a process known as transposition (Meir, Lin et al. 2013). The mechanistic principle of transposition was exploited and used as a vehicle for transporting genetic materials into the host genome. The gene of interest is flanked by ITRs and co-expressed with transposase for random integration into the genome (Rio 1991).

Using this system, many transposition-based studies were carried out in plants and invertebrates, but the development of the sleeping beauty (SB) transposon system (derived from salmonid fish) in 1997 made its use possible in vertebrates (Ivics, Hackett et al. 1997, Hackett, Ekker et al. 2005). Since then, many active transposon systems were developed for use in mammalian system (Dupuy 2010), among them the piggyBac transposon derived from the cabbage looper moth, *Trichoplusia ni*, has gained significant importance for manipulating genomes in various mammalian cell lines including human iPSCs and ESCs for pre-clinical studies (Wu, Meir et al. 2006, Yusa, Rad et al. 2009, Meir, Lin et al. 2013).

Since the process of transposition is essentially random, attempts were made to fuse the

SB transposase to E2C zinc finger protein for specific genome integration (Yant, Huang et al. 2007), however this modified transposase did not show target site specificity. Therefore, safety issues such as insertional mutagenesis, oncogenesis and proximal activation of neighboring genes cannot be excluded with transposon systems (Mates, Chuah et al. 2009). Transposase expression can also lead to genotoxic effects since extensive jumping of transgenes can cause mutations and dysfunctionality of endogenous genes (Mates, Chuah et al. 2009, Furushima, Jang et al. 2012).

1.2.1.3 Viral-mediated gene delivery

The random integration of foreign DNA into eukaryotic genomes has become a very common and standard technique in molecular biology. In order to increase the efficiency of integration and expression of transgenes, recombinant viral-based gene transfer replaced plasmid-based methods.

Viral-based gene delivery is normally done with vectors that are devoid of virulent genes and retaining those sequences in the vector backbone that are required for either packaging the viral genome into the capsid or integrating the vector into the host genome (Thomas, Ehrhardt et al. 2003). The desired transgene is cloned in the viral vector backbone and is co-expressed along with the packaging constructs (consisting of structural genes needed for the formation of the viral particles) (Thomas, Ehrhardt et al. 2003). This results in the assembly of viral particles incorporating the desired transgene (**Figure 1.2**). Infection with these viral particles results in more stable and long term transgene expression compared to the plasmid method.

Adapted from (Thomas, Ehrhardt et al. 2003)

Figure 1.2 Viral-mediated gene delivery. Co-expression of packaging constructs and vector genome consisting of desired transgene leads to the production to viral vectors that infects and incorporates transgene into the host genome

Viral vectors have become a very important tool for transgene delivery to cells for biomedical, biotechnological and gene therapy purposes. The various attributes associated with different viral systems (Table 1.1) determine their suitability for various therapeutic applications (Thomas, Ehrhardt et al. 2003). For example, cancer gene therapies or vaccinations that require short-term gene expression in specific cell types mainly utilizes adenoviruses whereas adeno-associated virus (AAV) has become the viral vector of choice particularly in clinical trials where long term correction or enhancement of genetic defect is required, but is limited by its packaging capacity. On the other hand, viral systems such as retroviruses and lentiviruses have the advantage of stable integration leading to long-term and persistent expression, making them more suitable for therapies such as hematopoietic gene therapy, where lentiviruses have become more popular recently due to their ability to infect both dividing and non-dividing cells (Glorioso 2000, Thomas, Ehrhardt et al. 2003) (Table 1.1).

Even though viral-mediated transgenesis has led to many clinical relevant applications, several limitations such as immunogenicity due to expression of viral components, low cargo capacity, genotoxicity and insertional mutagenesis associated with random integration cannot be overlooked and need to be overcome to achieve safer genome engineering

Table 1.1 Characteristics of different viral vectors along with some of their applications in gene therapy

Viral vectors	Packaging capacity	Tropism	Inflammatory potential	Vector genome forms	Main limitations	Main advantages	Applications in gene therapy
Adenovirus	8kb* 30kb*	Broad	High	Episomal	Capsid mediates potent inflammatory response	High transduction efficiency	*Cancer gene therapies *Stem cell differentiation *Gene therapies related to liver, cardiovascular, lung diseases
Adeno-associated virus (AAV)	<5kb	Broad, except hematopoietic stem cells	Low	Episomal (>90%) Integrated (<10%)	Small packaging capacity	Non-inflammatory and non-pathogenic	*Cancer gene therapies *Monogenic diseases
Retrovirus	8kb	Dividing cells only	Low	Integrated	Only transduces dividing cells; integration might induce oncogenesis in some applications	Persistent gene expression	*Hematopoietic gene therapies such as for X-SCID, *Cancer immunotherapy *Tissue engineering
Lentivirus	8kb	Broad	Low	Integrated	Integration might induce oncogenesis in some applications	Persistent gene expression in most tissues	*Anti-HIV therapy *Cancer immunotherapy

(Thomas, Ehrhardt et al. 2003, Henckaerts and Linden 2010, Howarth, Lee et al. 2010, Thaci, Ulasov et al. 2011, Vannucci, Lai et al. 2013)

1.2.2 Site- or sequence-specific transgene insertion into the genome

The site at which transgene integration occurs in a genome can have profound effects on transgene expression or on the surrounding genomic regions (Adamson, Jackson et al. 2011). Integration of the transgene into dense heterochromatin regions can lead to little or no transgene expression whereas transgenes integrated near open chromatin regulatory regions may have high level of expression (Dillon and Festenstein 2002). On the other hand, integration in a regulatory region may disrupt expression of endogenous genes leading to insertional mutagenesis and genotoxicity (Bushman, Lewinski et al. 2005, Fischer, Haccin-Bey-Abina et al. 2010). Therefore, the site of integration of a transgene in the genome is critical for correct control of gene expression and for avoidance of uncertainties associated with illegitimate integration.

Hence, targeting at a specific pre-determined endogenous site or sequence would not only help generate stable, functional transgenic cells but would also reduce the adverse effects that are caused otherwise by random integration (Adamson, Jackson et al. 2011, Silva, Poirot et al. 2011). In this context, one of the approaches involves targeting of a gene at human genomic “safe harbor sites” to prevent interference with transgene expression and avoid adverse effects such as insertional mutagenesis, activation of neighboring genes and cell toxicity (Sadelain, Papapetrou et al. 2012, Holkers, Maggio et al. 2014). One of the most widely used safe harbor sites in the human genome is the AAVS1 locus that not only allows efficient and stable transgene expression but also is safe in terms of normal

cell function and gene regulation (Henckaerts and Linden 2010, Tay, Tan et al. 2013). Apart from using the safe harbor approach, another desirable strategy could be targeting a specific sequence in the human genome. In this regard, utilization of sequence-specific recombinases that utilize short specific sequences could provide a promising way to target similar safe sequences present in the human genome (Keravala, Lee et al. 2009).

To this end, various efforts are under way to achieve stable and specific insertion into the genome. However, identification and validation of more such specific and targetable endogenous sites/sequences would be highly useful to express reporter genes, therapeutic genes and even functional multi-transgene cassettes (Sadelain, Papapetrou et al. 2012). This specific and targeted integration into the human genome would ultimately advance human cell engineering and various related cell based therapies.

Some of the specific genome-modification tools along with their working principle and applications are described in the following sections.

1.2.2.1 Gene targeting via homologous recombination (HR)

HR is the most commonly used method to manipulate genomes and achieve site-specific targeting. This is based on the naturally occurring process of exchanging DNA fragments between two homologous chromosomes during meiosis. HR has been utilized to exchange genetic material between the genome on the one hand and homologous donor DNA carrying exogenous transgenes provided, for example, in form of plasmids, linear DNA or oligonucleotides (Reh and Vasquez 2001). Therefore, a transgene can be specifically targeted at any known endogenous locus by providing the homologous sequences flanking the transgene (**Figure 1.3**). This approach can be used to generate mutations or for knock in/out purposes at defined locus.

dapted from (Wach, Brachat et al. 1994)

Figure 1.3 Homologous recombination. A schematic diagram of transgene targeting via homologous recombination at a specific genomic locus of interest by providing linear transgenic DNA flanked by left and right homology arms

Mario Capecchi, Martin Evans and Oliver Smithies were awarded the Nobel Prize for the application of HR in mouse ESCs, which was a scientific breakthrough in the area of mouse genome engineering. This has been very successful in generating various knock in/out reporter mouse models for developmental studies, embryogenesis and disease pathogenesis. Generating knock in/outs were further improved by site-specific recombinases such as the *Cre/Lox* and the *Flp/Frt* systems to add, remove or invert sequences at the targeted site (Reh and Vasquez 2001).

However, applicability of HR in human cell-culture systems is limited due to low targeting efficiency in hESCs and somatic cells (i.e. one event per 10^5 to 10^7 treated cells) (Reh and Vasquez 2001), which makes it very time consuming and labor intensive to screen for a correct genetically modified clone. In addition, the frequency of random targeting of exogenous DNA is much higher than specific integration (Roth and Wilson 1986, Thomas and Capecchi 1987). Even though positive/negative selection methods helped to increase the ratio of targeted over randomly integrated clones, extensive screening remains a limitation (Vasquez, Marburger et al. 2001). These challenges demand more effective approaches for sequence-specific modification, in particular for the human genome. In this direction, extensive research efforts eventually led to the development of various site-specific genome-editing tools beyond HR.

1.2.2.2 Enhancing HR via induced double strand breaks (DSBs)

HR based gene targeting has been a main focus to study gene function using reverse genetics approaches. However, its application and implementation in various cell types and eukaryotic systems have not been very efficient, mainly because of inefficient cellular homology-directed repair (HDR). Therefore, researchers focused on enhancing the efficiency of HDR by inducing DSBs, which according to several reports highly stimulate the cellular repair process. (Reh and Vasquez 2001, Miller, Holmes et al. 2007, Gaj, Gersbach et al. 2013). This led to the development of site-specific endonucleases, composed of a specific DNA binding module and an endonuclease domain (Sorrell and Kolb 2005, Mahfouz, Piatek et al. 2014).

The first site-specific endonucleases discovered were ZFNs in 1996 followed by TALENs in 2009 and, more recently, CRISPR-Cas9 in 2012 (Doudna and Charpentier 2014). These tools have enabled precise targeting of endogenous genomic sequences, which potentially holds great promise also for clinical applications (Yang, Mali et al. 2014). These programmable endonucleases have different sources of origin, structures and recognition mechanisms, but they share the same working principle, i.e. induction of DSBs at a known locus of the genome. The introduction of DSBs activates the cellular host repair machinery, which repairs DSBs either by non-homologous end joining (NHEJ) or HDR (**Figure 1.4**) (Wijshake, Baker et al. 2014). In the absence of a repair template (donor DNA), DSBs are generally repaired by error prone NHEJ leading to deletions, mutations or small insertions (indels) at the targeted site frequently resulting in gene knockout by either disrupting the gene or by causing a frameshift (van Gent, Hoeijmakers et al. 2001, Rothkamm, Kruger et al. 2003, Wijshake, Baker et al. 2014). On the other hand, HDR-mediated repair may occur when a repair template (donor) is provided and copied, resulting in gene knock-ins or correction at the site of DSBs (**Figure 1.4**) (Sommer, Peters et al. 2015).

Adapted from (Finotti, Breda et al. 2015)

Figure 1.4 Working principles of ZFNs, TALENs and CRISPR/Cas9 endonucleases. A) ZFNs, B) TALENs and C) CRISPR/Cas9 bound to double-stranded target DNA. D) Induction of DSBs by ZFNs, TALENs and CRISPR/Cas9 at a specific site can be repaired either by HDR repair mechanism or NHEJ. HDR pathway repair uses a donor template in the form of plasmids or linear DNA resulting in point mutations or insertions of larger DNA segments at the site of targeting. NHEJ-mediated repair produces random indels, often leading to disruption of coding region (gene knockout)

1.2.2.3 Site-specific endonucleases for genome engineering

1.2.2.3.1 Zinc finger nucleases (ZFNs)

ZFNs are site-specific endonucleases designed by fusing a zinc finger DNA-binding domain to a nuclease domain (Li, Wu et al. 1992, Carroll 2011) for precise targeting within complex genomes (**Figure 1.5**). The DNA binding domain comprises of a varying number of zinc finger proteins (ZFP), each made up of ~30 amino acids and characterized by a beta-alpha-beta motif where a zinc atom is bound to Cys2His2 residues (Carroll 2011). The nature of amino acids present at specific positions in the alpha helix determines specificity of the each ZFP to the DNA sequence (Kandavelou and Chandrasegaran 2009). The DNA cleavage domain of ZFN is derived from the *FokI* restriction enzyme, and has no apparent sequence specificity (Li, Wu et al. 1992) (**Figure 1.5**).

Each ZFP recognizes a distinct 3bp DNA sequence and 3-6 fused ZFPs are assembled to a *FokI* cleavage domain to redirect the specific cleavage of DNA. Depending on the number of ZF motifs, a ZFN pair can recognize a full target site of 18-36bp of DNA

sequence (Handel and Cathomen 2011), i.e. two independent ZFNs are constructed, directed to bind to adjacent genomic sites in an inverted orientation, which leads to functional dimerization of *FokI* and cleavage at the specific genomic site (**Figure 1.5**) (Wijshake, Baker et al. 2014). The introduction of a DSB at a predetermined genomic locus then induces the activation of the DNA repair host machinery that leads to either error prone NHEJ repair or donor DNA-mediated HDR (as shown in **Figure 1.4**).

Adapted from (Klug 2010)

Figure 1.5 An illustration of a ZFN. A ZFN is composed of two functional domains: A DNA binding domain and a DNA cleavage domain. A typical DNA binding domain consists of three ZFPs (in red and blue), each recognizing 3bp sequence of DNA. Three-four zinc finger proteins are linked together to recognize 9-12bp of specific DNA sequence. A DNA cleavage domain consists of *FokI* restriction enzyme, which requires dimerization of two ZFNs in order to cleave the genome at a specific site

ZFN-mediated site-specific genome manipulation works efficiently in various cell types that can have therapeutic implications. Recently, it has been used to disrupt the *CCR5* gene (encoding a HIV co-receptor) via NHEJ repair, and the approach is currently in Phase II clinical trials for HIV treatment (Perez, Wang et al. 2008, Didigu, Wilen et al. 2014). ZFN-induced HDR is also being considered to treat various genetic diseases (Sebastiano, Maeder et al. 2011, Yusa, Rashid et al. 2011). In this context, ZFNs have been used to target a therapeutic gene ($gp91^{phox}$) at the safe harbor AAVS1 site in human iPS cells derived from a patient with X-linked chronic granulomatous disease (X-CGD) (Zou, Sweeney et al. 2011).

However, there are concerns regarding ZFN technology; it is technically challenging, time consuming and expensive to assemble ZFN modules (Kandavelou and Chandrasegaran 2009). In addition, the ZFN modules sometimes lack sufficient sequence specificity which increases the risk for potential off-target effects leading to cytotoxicity

and chromosomal aberrations (Cornu, Thibodeau-Beganny et al. 2008, Urnov, Rebar et al. 2010, Tong, Huang et al. 2011, Gaj, Gersbach et al. 2013). This calls for further improvements in design and construction of ZFN, or even more cost-effective and simplified genome editing tools.

1.2.2.3.2 Transcription activator like effector nucleases (TALENs)

Due to high costs, off-target effects and complexity in the modular assembly of ZFNs, a progressive search was undertaken for new methods which led to the development of an alternative genome editing tool; TALENs and was named as 'method of the year' in 2011 (Nemudryi, Valetdinova et al. 2014). TALENs are functionally similar to ZFNs as both consist of the *FokI* nuclease domain fused to customizable DNA binding modules (Wood, Lo et al. 2011).

The binding domain of TALENs is derived from the plant pathogenic bacteria *Xanthomonas* genus, which infects crop plants such as rice, pepper and tomato (Bogdanove, Schornack et al. 2010). These bacteria secrete effector proteins, TALEs, that bind to plant DNA and activate expression of target genes, making them more susceptible to infection (Bogdanove, Schornack et al. 2010, Jankele and Svoboda 2014). The TALE proteins have a N-terminal activation domain for initiation of transcription, a central DNA binding domain for specific binding and a nuclear localization signal at the C-terminal end (**Figure 1.6**) (Joung and Sander 2013). The central DNA binding domain of the TALEs comprise of 7-34 highly homologous monomers. Each monomer is comprised of 33-34 amino acid residues. The amino acids at position 12 and 13 are polymorphic and form the repeat variable di-residue (RVD), which are responsible for the specific recognition of a single nucleotide in the target DNA sequence (**Figure 1.6**) (Boch, Scholze et al. 2009, Morbitzer, Romer et al. 2010, Jankele and Svoboda 2014). This sequence-specific property of DNA binding domains provided an opportunity for TALENs to be used not only for gene editing, but also for targeted gene regulation in conjunction with transcriptional activators and repressors (Joung and Sander 2013).

Adapted from (Joung and Sander 2013)

Figure 1.6 A schematic diagram of a TALEN. TALENs consist of a N-terminal domain; middle DNA binding domain (TALE repeat domains) and a C-terminal domain attached to a DNA cleavage domain (*FokI* endonuclease). Each TALE repeat is represented by a colored disc comprising 33-34 amino acids with two hypervariable amino acid residues at position 12-13, highlighted in orange and bold text. These two amino acids highlighted in each colored disc of TALE repeat recognize a specific nucleotide in the target DNA sequence. Two TALENs bind adjacent to each other for *FokI* to dimerize and cleave the DNA at a specific site

TALENs have been successfully used for genome editing in various human cells including hESCs and *in vivo* in model organisms such as yeast, zebrafish and mice (Wood, Lo et al. 2011, Panda, Wefers et al. 2013, Hwang, Peterson et al. 2014, Sommer, Peters et al. 2015). The ability of TALENs to achieve specific targeting provides avenues for the use in gene disruption, correction and addition for treating various genetic diseases (Nemudryi, Valetdinova et al. 2014).

Even though there has been widespread use of TALEN technology over the past three years, some of the issues still need to be addressed for its use in research and therapeutics (Nemudryi, Valetdinova et al. 2014). This system requires considerable expertise in designing and cloning of plasmid constructs, making it relatively less user friendly. In addition, the process is labor intensive, time consuming and has low efficiency (Gaj, Gersbach et al. 2013). Since DSBs induced by TALENs rely on the host repair machinery, off target NHEJ-mediated mutagenesis can lead to cytotoxicity and chromosomal aberrations. In terms of gene addition, TALENs are also being explored to target safe harbor sites (Luo, Liu et al. 2014), however it has been reported that the efficiency of integration of large transgene constructs (>5kb) is only 0.1-0.6%. This

highlights the limitation of TALENs to target large transgenes efficiently at the targeted locus (Wu, Gao et al. 2014).

1.2.2.3.3 CRISPR/Cas9 genome editing tool

Recently, a new genome-editing tool derived from the bacterial adaptive immune system known as clustered regularly interspaced short palindromic repeats, (CRISPR) and CRISPR associated (Cas) has been added to the genetic engineering toolbox. CRISPR-Cas-mediated immunity involves integration of short foreign DNA fragments into the spacer region of the bacterial CRISPR locus, known as protospacers (Ma, Zhang et al. 2014). This locus is transcribed into precursor CRISPR RNA (pre-crRNA) and in the presence of trans-activating RNA (tracrRNA) forms a target RNA complex. The complex complements the DNA sequence of invading viral or plasmid sequences and directs the Cas9 endonuclease to the specific site for cleavage (**Figure 1.7**) (Ma, Zhang et al. 2014). This has been exploited for eukaryotic genome engineering wherein a synthetic single guide RNA (sgRNA) (~20bp long) together with Cas9 endonuclease are expressed to target any genomic locus that is adjacent to a PAM sequence for gene editing, knockout or transgene addition. The simplicity of this system has allowed multiplexing, enabling targeting of multiple loci using different sgRNAs simultaneously in the same genome (Cong, Ran et al. 2013).

Adapted from (Kondo and Ueda 2013)

Figure 1.7 An illustration of a CRISPR/Cas9. A Cas9 nuclease, programmed by the tracrRNA:crRNA complex cutting both strands of genomic DNA 5' of the PAM (red)

CRISPR/Cas9 represents a simple, versatile and efficient tool to edit eukaryotic genomes thereby holds promise as a therapeutic intervention strategy for gene therapy (Doudna and Charpentier 2014). To this end, it is widely exploited to alter genomes for germ-line modification studies in various animal models. For example, it has been used to correct

mutant genes responsible for causing eye cataract in mouse models (Wu, Liang et al.) and in another study, it has also been used to correct a mutation in the *CFTR* gene in isolated human intestinal stem cells to treat cystic fibrosis, thus providing evidence that CRISPR/Cas9 (though in early stage) has great potential in gene therapy-related applications (Schwank, Koo et al.).

Even though significant progress has been made using CRISPR/Cas9 in the genome engineering field lately, off-target effects remain a hurdle (Gaj, Gersbach et al. 2013). Off-targeting results in the activation of the cellular repair machinery that leads to mutagenesis (Fu, Foden et al. 2013, Mali, Yang et al. 2013, Pellagatti, Dolatshad et al. 2015). In addition, the requirement of PAM sequence for Cas9 to cleave DNA limits the selection of the sites that can be targeted (Pellagatti, Dolatshad et al. 2015). Another limitation is integration of large transgene (>5kb), as the efficiency of targeting declines very rapidly with larger insert size (Byrne, Ortiz et al. 2015).

Nevertheless, considerable efforts are being made to improve targeting specificity in order to reduce the risks of unwanted off targeting (Ma, Zhang et al. 2014). Cas9 nucleases from different organisms with alternative PAMs have been characterized and successfully used for targeting sites that were previously thought to be non-targetable using the CRISPR system (Shah, Erdmann et al. 2013, Kleinstiver, Prew et al. 2015). Furthermore, a mutant of Cas9 (nickase) has been generated which introduces single-strand breaks or nicks on opposite strands thus preventing DNA damage and activation of host repair machinery (Cong, Ran et al. 2013). Thus, the CRISPR Cas9 system is still undergoing significant improvements for efficient and desired modification of eukaryotic genome (Harrison, Jenkins et al. 2014).

Note: To facilitate the integration at the safe harbor AAVS1 site, many companies such as System Biosciences, Genecopoeia and Sigma Aldrich have formulated site-specific endonucleases (ZFN, TALEN, CRISPR/Cas) along with the donor templates in the form of a kit (Source: Company websites).

1.2.2.4 Site-specific recombinases for genome engineering

Site-specific recombinases represent a family of enzymes, mainly derived from bacteria and bacteriophages that carry out faithful recombination between two specific DNA sequences, resulting in DNA integration, excision or inversion (Hirano, Muroi et al. 2011, Turan and Bode 2011). However, most of these recombinases are also functional outside their organism of origin, thus broadening their applicability in other prokaryotic and

eukaryotic systems with respect to genome manipulation (Brown, Lee et al. 2011). Recombinases can be classified into two groups based on the amino acids used to hydrolyze the phosphodiester bond of DNA during recombination reactions; serine- and tyrosine-based recombinases (Groth and Calos 2004). Some of the distinguishing features of each of the site-specific recombinase families are listed below (**Table 1.2**).

Table 1.2 Characteristics of site-specific recombinase families (Hirano, Muroi et al. 2011)

	Tyrosine-type recombinase		Serine-type recombinase	
Catalytic residue	Tyrosine (forming 3'phosphotyrosine linkage)		Serine (forming 5'phosphoserine linkage)	
Recombination mechanism	Catalyze recombination via Holliday junction intermediate		180° rotation and rejoining of cleaved DNA	
Recombinase family	Phage integrase	Simple recombinase	Resolvase/ Invertase	Phage integrase
Representative members	Lambda HK022	Cre Flp	g8, Tn3/ Hin,Gin	PhiC3, Bxb1
Recognition site	~21bp attB & ~250bp attP	LoxP FRT	114bp res (G8, Tn3) 26bp HixL & HixR(Hin) 34bp gix (Gin)	~40bp attB & ~40bp attP
Host factor requirements	IHF & Fis	None	Not required (γ8 and Tn3) Fis and HU (Hin and Gin)	None
Topological requirements	Supercoiling	None	Supercoiling	None

Due to the high specificity and efficiency of site-specific recombinases in catalyzing recombination, they have become powerful tools for genome engineering. The recombination mechanism, functions and applications of each of the recombinases have been intensively studied; however, here we focus on a few of the most well characterized recombinases that have been mainly used for transgene integration in the human genome.

1.2.2.4.1 Serine-type phage integrases-mediated recombination

Serine phage integrases catalyze unidirectional site-specific recombination between target sites (both of which are ~40bp in size) without the need of any co-factors and DNA supercoiling (Smith and Thorpe 2002, Groth and Calos 2004). The N-terminal catalytic domain of serine integrases is highly conserved but due to the diverse nature of the C-terminal DNA binding domain, each member of serine integrases has the ability to recognize distinct sites (Rowley, Smith et al. 2008). However, these recombinases are mechanistically and evolutionary distinct from tyrosine recombinase family (Thorpe and Smith 1998, Grindley, Whiteson et al. 2006). For integration to occur, two serine recombinase dimers form a synaptic complex between their recognition sites, followed by strand cleavage using the catalytic serine residue (**Figure 1.8**). The subsequent step of strand exchange in the synapse involves 180° rotation of two of four recombinase molecules with respect to the other two recombinase molecules which is followed by the

re-ligation of the cleaved strands (**Figure 1.8**) (Rice, Mouw et al. 2010, Smith, Brown et al. 2010).

Adapted from (Brown, Lee et al. 2011)

Figure 1.8 Steps involved in serine-integrase-mediated recombination. A Schematic diagram-showing steps involved in the serine integrase-mediated recombination. Two serine recombinase dimers (indicated in red and green) bind to their recognition sites forming a synapse followed by a concerted cleavage of all four DNA strands, subunit rotation and ligation of recombined DNA strands

Serine-type phage integrases, such as BxB1 and PhiC31, have been explored in terms of gene integration in mammalian genomes (Brown, Lee et al. 2011). These integrases are often used to catalyze site-specific recombination between phage attachment (*att*) site(s) pre-inserted in the genome with the bacterial attachment (*attB*) site provided on a plasmid (Hirano, Muroi et al. 2011). Furthermore, several *attP* like sequences (so-called pseudo *attP* sites) that bind to PhiC31 integrase have been found in the human genome. The pseudo *attP* sequence show 40% sequence identity to the wild-type *attP* sequence, but PhiC31 can mediate integration at these sites at 10 fold higher rate than at random sites (Thyagarajan, Olivares et al. 2001). A typical mammalian genome contains 100-1000 copies of pseudo *attP* sites and some of them have been shown to be targeted in mouse, bovine, zebrafish and human genomes (Thyagarajan, Olivares et al. 2001, Allen and Weeks 2005, Bolusani, Ma et al. 2006, Ou, Huang et al. 2009). It was observed that PhiC31 shows preference to certain pseudo *attP* sites depending on the cell type and chromatin accessibility (Thyagarajan, Olivares et al. 2001, Chalberg, Genise et al. 2005).

For the past decade, several improvements have been made in the PhiC31 recombination system along with the generation of variants of PhiC31 with better specificity (Keravala, Lee et al. 2009). However, the safety profile of PhiC31-mediated recombination has

limited its use in gene therapy. It has been known to activate the DNA cellular repair machinery leading to chromosomal abnormalities such as indels, translocations and rearrangements at/near the targeted locus (Thyagarajan, Olivares et al. 2001, Liu, Skjorring et al. 2009). In addition, several reports have also proposed that PhiC31 could recombine native pseudo *attP* sites leading to deletions (Ehrhardt, Engler et al. 2006, Liu, Skjorring et al. 2009).

Even though a number of serine type recombinases and their derivatives are being used till date, problems with respect to safe, precise and efficient integration of transgenes at predetermined site still persists. Therefore other integrating systems such as Cre-*loxP*/Flp-*FRT* and λ -Int for site-directed transgenesis have also been explored.

1.2.2.4.2 Cre-*loxP*/Flp-*FRT* system

The Cre and Flp recombinases belong to the integrase family of site-specific recombination system prevalent in bacteriophage P1 and yeast, respectively. Cre and Flp identify 34bp sequences called *loxP* and *FRT*, respectively, each comprised of two flanking 13bp palindromic sequence and a 8bp spacer sequence (Nagy 2000). Unlike serine recombinases, the Cre/*loxP* and Flp/*FRT* systems utilize tyrosine as a catalytic residue and carry out recombination in a sequential manner via formation and resolution of Holliday junction (similar to λ -Int in **Figure 1.11**) (Ghosh and Van Duyne 2002) to yield the recombinant products. However, in comparison to Cre, Flp has lower affinity to its *FRT* sites and is also less thermostable. To overcome this, various mutants of Flp have been generated to increase its stability and tolerability (Buchholz, Angrand et al. 1998).

These recombinases have been extensively used for generating global and conditional knockouts in mammalian cell lines, yeasts, drosophila and mice (Sauer 1987, Gu, Zou et al. 1993, Irion, Luche et al. 2007). This is done by insertion of *loxP*/*FRT* sites flanking the key region to be deleted via HR, followed by expression of Cre/Flp (**Figure 1.9 A**). They are also being used for gene insertion via recombinase-mediated DNA insertion (RMDI) (**Figure 1.9 A**) (Turan, Zehe et al. 2013). In addition, recombinase-mediated cassette exchange (RMCE) is an alternative strategy wherein a genomically integrated cassette flanked by two distinct *loxP* (or *FRT*) sites which cannot recombine with each other is exchanged with another cassette also flanked by the corresponding *loxP* sites present on a plasmid (**Figure 1.9 C**) (Feng, Seibler et al. 1999).

Adapted from (Trinh and Morrison 2000)

Figure 1.9 Genome manipulations using the Cre/loxP system. (A) Integration (RMDI) or excision (knockout) of the insert (in blue) via sites (in green) oriented in same direction (in green), upon expression of Cre. **(B)** Inversion of integrated inserts (in a genome) flanked by inversely oriented *loxP* sites using Cre. **(C)** A genomic integrated cassette 1 (in blue) flanked by two distinct mutant *loxP* (in yellow)/ *loxP* (in green) sites is exchanged with cassette 2 (in green) also flanked by two distinct compatible *loxP* sites incorporated in a plasmid

Another way of using Cre/Flp recombinase in the context of genomic integration is to target native *loxP* or *FRT* like sequences in the genome. Recently, Cre-mediated insertion of a transgene into a genomic *loxP*-like sequence has been documented in mouse genome (Ito, Yamanouchi et al. 2011). Since the exact 34bp *loxP*/*FRT* recognition sequences are not present in the human genome, considerable efforts were made either to identify pseudo-recognition sites that could act as substrates for wild-type Cre/Flp recombinases or to generate variants with altered target site specificity. To this end, a variant of Cre recombinase was generated to target a *loxP* like sequence termed as *loxH*, found on chromosome 22 of the human genome (Thyagarajan, Guimaraes et al. 2000, Buchholz and Stewart 2001). Similarly, better Flp variants were also generated to target two *FRT* like sequences identified in the human interleukin 10 gene (Bolusani, Ma et al. 2006, Shultz, Voziyanova et al. 2011). But none of these recombinases have been able to carry out recombination at pre-existing specific genomic sequences that could otherwise have been a great advancement in the field of gene targeting.

In order to address the long-standing issue of achieving sequence specific targeting, our lab has focused on another tyrosine recombinase, λ -integrase (λ -Int), to explore the possibility of targeting the human genome.

1.2.2.4.3 λ -Int-mediated site-specific recombination

λ -Int-mediated site-specific recombination has gained significant importance for two reasons: i) it represents a *bona fide* naturally evolved transgene adding system and ii) it exhibits high target specificity even with large circular transgenic constructs (up to 48 kb in the case of the λ genome itself).

The wild-type phage λ integration/excision system in the context of the lysogenic and lytic phage life cycles, respectively, is rather complex and can briefly be summarized as follows (**Figure 1.10**) (Landy 1989). The integrase protein, encoded by bacteriophage λ , catalyzes the integration of the phage DNA into the *Escherichia coli* (*E.coli*) genome using a pair of attachment (*att*) sites: *attP* (in the phage DNA) and *attB* (in the bacterial genome). The resulting prophage, flanked by *attL* and *attR* sites (left and right, respectively) generated as a result of recombination are fully competent for excisive recombination to splice out its genome from the host *E.coli* genome and enter the lytic pathway under unfavorable physiological conditions (**Figure 1.10**) (Landy 1989).

Note: Although the *attB* and *attP* target sequence is different for PhiC31 and Int, the nomenclature for these recognition sites is same.

The 21bp *attB* and the 241bp *attP* sites sequences include a pair of 7bp inverted core-binding sites separated by a 7bp overlap (also called spacer) region. This is involved in the actual strand exchange during recombination. In addition, the *attP* site has arm-binding sites (at about 150bp towards the right and 80bp towards the left from the overlap) for binding of Int. The Int enzyme has a N-terminal domain (1-64 aa residues), a C-terminal core binding domain (65-169 aa residues) and a C-terminal catalytic domain (170-356aa residues). The N-terminal domain recognizes several arm binding sites, which when DNA-bound increases the affinity of the C-terminal domain to the core binding sites. It is this higher affinity that allows Int to form synaptic complexes between *att* sites and to catalyze strand exchange via Holliday junction intermediates (Nunes-Duby, Kwon et al. 1998). The wild-type Int requires accessory protein co-factors, such as IHF (integration host factor), FIS and Xis (excisionase) to bind at the *attP* arm sites in order to form functional recombinogenic complexes on *attP* (Landy 2015).

The IHF protein is a heterodimer encoded by *E.coli* genes *himA* and *himD* that binds to the *attP* arm sites and induces bends in DNA. This allows the same Int monomers to simultaneously contact arm and core binding sites to catalyze both integrative and excisive recombination (Landy 2015). Xis is a phage-encoded protein and it also induces a sharp bend in DNA to promote excision by carrying out *attL/attR* recombination in conjunction with Int and IHF (Landy 2015). Another host-encoded protein, FIS, also stimulates excision reaction 20-fold by bending DNA under limiting conditions of Xis (Papagiannis, Sam et al. 2007). In addition, electron microscopy has revealed that the *attP* site is supercoiled facilitating interaction of a large protein complex including 4-8 monomers of Int with other co-factors (Thompson, Moitoso de Vargas et al. 1987).

Adapted from (Landy 2015)

Figure 1.10 Overview of λ-Int-mediated site-specific recombination. Integration of bacteriophage λ genome into the *E. coli* genome via *attB* X *attP* recombination generating a prophage flanked by *attL* and *attR* sites that act as substrates for excisive recombination during the lytic cycle. Both integration and excision of phage genome requires λ-Int protein along with other accessory co-factors such as IHF, FIS, Xis (Landy 2015)

The process of λ-Int-mediated recombination involves strand cleavage, exchange via Holliday junction and re-ligation of the recombining DNA partners. Firstly, for integration, negatively supercoiled *attP* in association with Int and IHF forms an intasome (Richet, Abcarian et al. 1988). This intasome contacts protein free *attB*, presumably

through random collision (Mizuuchi, Gellert et al. 1980, Christ, Corona et al. 2002) and forms a synaptic complex. In the second step, Int induces cleavage of the top strand at each *att* site and the tyrosine residue forms a covalent phospho-tyrosine linkage at the 3'end of DNA. This linkage is subsequently attacked by the OH group of the partner DNA strand resulting in a Holliday junction intermediate structure (Burgin and Nash 1992). The Holliday structure is subsequently resolved by cleavage, exchange and ligation of the bottom strands (**Figure 1.11**) (Nunes-Duby, Matsumoto et al. 1987). The switch from top to bottom strand cleavage involves isomerization of the Holliday junction followed by a process that can be labeled as proof reading at the homologous overlap region in both strands (Nunes-Duby, Azaro et al. 1995). Completed recombination leads to the formation of *attL* and *attR* sites, which serve as substrates for excisive recombination for Int, IHF and XIS proteins via the same process as mentioned above (**Figure 1.11**).

Therefore, Int executes an ordered, sequential pair of strand exchange (Nunes-Duby, Matsumoto et al. 1987) with a kind of proof reading intermediate step limiting recombination at off target sites.

Adapted from (Christ and Droge 1999)

Figure 1.11 λ -Int-mediated recombination via a Holliday junction model. λ -Int-mediated strand cleavage, exchange and ligation takes place in the following four steps. (a) *attB* (BOB') and *attP* (POP') sites each bound to two Int monomers align in anti-parallel orientation and Int initiates a nucleophilic attack using its tyrosine residue within each site. (b) Tyrosine cleaves the top strand at the left boundary of the overlap in each *att* site resulting in strand exchange and formation of a Holliday junction intermediate structure. (c) Similarly, the tyrosine residue cleaves the bottom strand at the right end of the overlap and forms a covalent phospho-tyrosine linkage that is attacked by the OH-group of the partner strand (d) Exchange of bottom strands followed by resolution of the Holliday junction intermediate giving recombinant *attL* (BOP') and *attR* (POB') sites. Excisive recombination is the exact reversal of integrative recombination

The bacteriophage λ -Int is probably the best described recombinase (Landy 1993, Yang and Mizuuchi 1997), but its use in eukaryotic systems has been limited because of the requirements of accessory proteins and the negative supercoiling of *attP*, which is distinct from the Cre/Flp systems (Praznovszky 2011). In this context, various mutants of λ -Int such as Int-h (Miller, Mozola et al. 1980), Int-h/218 (Lorbach, Christ et al. 2000), Int-1, Int-5 (Tay, Ho et al. 2010), Int-C2, Int-C3 (Siau, Chee et al. 2015) have been generated to promote recombination in the absence of accessory proteins. In addition, these mutants have been shown to have higher affinity to wild type as well as non-cognate *att* sites compared to the wild type λ -Int (Tay, Ho et al. 2010, Siau, Chee et al. 2015). However,

none of them have been able to reliably target native sequences in the human genome, suggesting a need for improved strategies or approaches to harness the potential λ -Int-mediated recombination for targeting the human genome.

1.3 Aim of the project

The λ -Int system represents a naturally evolved transgene addition tool, which exhibits high target site specificity, even with large circular transgenic constructs (up to 48 kb in the case of the λ -genome itself). Taking cues from the natural design of the system, we hypothesized that the λ -Int-based recombination system could have potential to achieve site-specific transgene insertion into the human genome.

With the focus on more efficient mutant λ -Ints and the capacity to recognize non-cognate sequences other than wild-type *att* sequences, the aim was to identify an endogenous targetable *att*-like sequences in the human genome allowing for homogeneous and long-term expression of multiple transgene cassettes.

2. MATERIALS AND METHODS

2.1. List of primers used in the study

Primer name	Sequence (5'-3')
Int_fwd_PstI	CCAAGTGCAGCTCGAGGCCACCATGGGAAGAAGGCGAAG
Int_rev_XbaI	GAGCTCTAGATCATTATTTGATTTC
CNLS_XbaI_Int	GAGCTCTAGATTATCATACCTTTCTTCTTCTTAGGGCCGCCTCCGGATTTGATTTC
KpnI_UTF1_fwd	GATTGGTACCTTAGGACCCGGCGGGCGGGG
SDM-Int-Y342A-F	CGGTCACCATGGCATCACAGGCTCGTGATGACAGAGGCAGGGA
SDM-Int-Y342A-R	TCCCTGCCTCTGTCATCACGAGCCTGTGATGCCATGGTGACCG
attP(4X)_sdm (For)	CGTTTCTCGTTCAGCTTTATTTTCAATTAAGTTGGCATTAATAAAAAAGC
attP(4X)_sdm (Rev)	GCTTTTTTATAATGCCAACTTAATGAAATAAAGCTGAACGAGAAACG
INT-HIS-XbaI Rev	TATCCGAGATCTCTAGATTAGTGGTGATGGTGATGATGTTTGATTTC
NLS-HIS-XbaI Rev	TATCCGAGATCTCTAGATTAGTGGTGATGGTGATGATGTACCTTTCTTCTTCTTAGG
EF1Fw_EcoRI	ATCGCCGAATTCGGCTCCGGTGCCCGTCAGTGGGCAG
EF1Rev-NheI	CGAATTGCTAGCTCACGACACCTGAAATGGAAGAA
EF1aRe-NsiI	GTGTATATGCATCCACCACACTGGACTAGTGGATC
EcoRV_EF_fwd	TCGGGATATCGGCTCCGGTGCCCGTCAGTG
ClaI_bgh_bpa_rev	GTGTATCGATGCCGCTACAGGGCGCGTGGGGAT
C3 Int-CNLS-BglII Rev	TATCCGAGATCTGCTAGCTTATCATACCTTTCTTCTCTC

Primer name	Sequence (5'-3')
ClaI_UTF1 enhancer_rev	GTGTATCGATGCCACCACACCCGGCTAATTTTTG
attL(BOP') Fwd (KpnI)	AAGTGGTACCTGAAGCCTGCTTTATTTTATTAAGTTGG
attL(BOP') Rev (NotI)	GTGAGCGGCCCGCCCGGGAGGGAGTGGGACAAAATTG AAATC
attR (POB') Fwd (ClaI)	GTCTATCGATTCTGTTACAGGTCACATAATACCAT
attR (POB') Rev (HindIII)	CTTGAAGCTTGTTTCGCTCAAGTTAATGAAATAAAGCT GA
attP_rev	GACTGATAGTGACCTGTTCGTTGC
Puro rev24	CACCGTGGGCTTGTAATCGGTC
Puro rev303	GGAACCGCTCAACTCGGCCATGC
Puro fw509	CTACGAGCGGCTCGGCTTCACC
Puro fw318	CAACAGATGGAAGGCCTCCTGG
Puro 127	GAGCTGCAAGAACTCTTCCTCAC
Pgk_rev	CAGGTGAATATCAAATCCTCCTCG
cs_attH4X_F1	GAGTGTTTTCCAACCTGGTTCCATT
cs_attH4X_R1	AAAACACAGCACGAGAACTTCGTGA
cs_attH4X_F2	CCTGTCTTGCTAGGTTGGGAAGT
cs_attH4X_R2	TAGAGGAATTGCTAACTAGAATAACCA
pr21	CACAGGAAACAGCTATGACCATG
pr27	GAACGTCTGGGGTGTGATCCT

Primer name	Sequence (5'-3')
pr49	CCAATACGCAAACCGCCTCTC
pr99	CTGAATGAAGCCATACCAAACGAC
pr199	CGTTTACAATTTCCCATTGCGCCATTC
E3-GF1	TTCAGGAGTTGTTGTAAGGCAGGC
E3-GR1	CTCTGGGCAGGGCATCTCTGAAAG
E3-GF2	CCTTTCTCTCTGGCTGCCCTTAAC
E3-GR2	GCACCAACAGCAAAGACCAAAGG
K3-GF2	GCGCCCCACTCTCTTCTGACTTGTAG
K3-GR2	GACGTCCACACAGAAACCCCACTG
K3-GF1	CTTCAGGAGCTCTTGTAAGGCAGGC
K3-GR1	CCAGGGAAACAGGGTCTAGAGTGG
A3-GF2	GGCATGGGTGTTAGGGCCATTATG
A3-GR2	CGTCTAAAGGATCACAACTCCTTGCC
A3-GF1	AGTGGGAAATCTGGGGTCTCACTG
A3-GR1	GGAAGGAACGGGCAGCAATCTTTG
EGFP fwd	CCACATGAAGCAGCACGACTTC
EGFP rev	GGGTGTTCTGCTGGTAGTGTC
attPH_ fwd (HindIII)	GTCTAAGCTTTCTGTTACAGGTCACCTAATACCATC

Primer name	Sequence (5'-3')
attPH_rev (BamHI)	ATAAGGATCCAATAATGATTGTATTTGACTGATAG
M-cherry_fwd (BamHI)	AACGGGATCCGTGAGCAAGGGCGAGGAGG
M-cherry_rev (NheI)	CATTGCTAGCTTACTTGTACAGCTCGTCCATG
MCT5_exon2 CS (PstI)	GGTACTGCAGATGGGGATGGATGATTGTGATT
MCT5_exon5 end (XbaI)	TATCTCTAGACTATCAGGAATTTTGTAAGGGTGCC
MCT5_fwd_1	TTGGGGCTTTCGTTGTTACT
MCT5_rev_1	GATGGCAAAGCAGATGGTGTAG
MCT5_fwd_2	CTTTGTGTGACTATGGGACTTCTACC
MCT5_rev_2	CCGTCTCAAGGATACCTGTCCAG
attP_chr6_ SDM_fwd	CGTTTCTCGTTCAGCTTTATGAAAATAAGTTGGC
attP_chr6_ SDM_rev	GCCAACTTATTTTCATAAAGCTGAACGAGAAACG
SDM_attP_ ch8_fwd	CGTTTCTCGTTCAGCTTTTGTATACTAAGTTGGCATTATAAAAAAGC
SDM_attP_ ch8_rev	GCTTTTTTATAATGCCAACTTAGTATACAAAAGCTGACGAGAAACG
cherry rev	CTTGGCCTTGTAGGTGGTCTTG
F1_attL_MCT5	CGGATCCAGGAGATTATCTTTTCTG
R1_attL_MCT5	CTGTTCACGATGGTGTAGTCCTC
F2_attR_MCT5	CCAATACGCAAACCGCCTCTC
attP_fwd_1	CTGAATGAAGCCATACCAAACGAC

Primer name	Sequence (5'-3')
attP_fwd_2	GATTCATAGTGA CTGCATATGTTGTG
neo rev 81	GCCACAGTCGATGAATCCAGAAA
neo rev 44	CCGCATTGCATCAGCCATG
chr 6 fwd 1	CGATTCTCCAACCTCAGACTTCC
chr 6 fwd 2	CTTGCGTGTTCTTGAAAGCAGTATG
chr6 rev 1	GAACCTGCTAGGAGAAGTGTTGCC
chr 6 fwd 2	CTTGCGTGTTCTTGAAAGCAGTATG
chr6 rev 1	GAACCTGCTAGGAGAAGTGTTGCC
chr6 rev 2	GACTCAGATCTCCCTGCAACCTTTA
chr13 fwd 1	CTGATGATTAGTGATGTTCCGCATT
chr13 fwd 2	GTGCTGCAAAA ACTGGATATTCACAC
chr13 rev 1	GTCATTTGTCCTGTGGAAGAATAGG
chr13 rev 2	GCAGCAAGGTAAGAATATGAGGAA
chr13 K88 fwd 1	CTGGGTGAAGGCATAGATGATACA
chr13 K88 fwd 2	CTGAGCTTGCAGTGAGATGAGATC
Chr13 K88 rev 1	GATAGGAATGGGGCAATATACAGG
Chr13 K88 rev 2	CTGCTGCAATGTTACTGTGAACTG
attP pgk fwd	GTCCCACTCCCTCCCGAATTCTA

Primer name	Sequence (5'-3')
chr8 fwd 1	GCAGTAATCGCTCACACTAATGGC
chr8_fwd_2	GGTTTCACTATGTGGGTAAGACTGGTC
chr8_rev_1	CCAGAGATTATGTTTGACTGTGGCAC
chr8_rev_2	GCTGAGTGGTGTGACTGAGATTTG
pr22	GTTCACCTTGATGCCGTTCTTC
pr111	AAGTCGTGCTGCTTCATGTGGT

2.2 Vector constructs

Standard molecular cloning techniques were employed to generate plasmids for this study. PCR amplifications for cloning purposes was done using High fidelity Pfu polymerase (Thermo Scientific) and plasmids amplification were done using *E. coli* DH5 α .

The construction of Int expression vector *pCMVssInt-h/218* has been described in (Lorbach, Christ et al. 2000), wherein Int-h/218 expression is driven by the CytoMegalovirus (CMV) promoter (**Figure 2.1**).

Figure 2.1 A vector map of *pCMVssInt-h/218* expressing Int-h/218 driven by CMV promoter

pCMVssInt-1 and *pCMVssInt-5* were generated by PCR amplification of Int-1 and Int-5 coding sequence from *pET-Int-1* and *pET-Int-5* (Tay, Ho et al. 2010) using the primers Int_fwd_PstI and Int_rev_XbaI. The PCR product was then cloned into the *pCMVssInt-h/218* between *PstI* and *XbaI* sites (**Figure 2.1**), replacing the Int-h/218 sequence with Int-1 and Int-5 coding sequences.

pCMVssInt-C3 was generated by PCR amplification of Int-C3 coding sequence from *pET-Int-C3* (Siau, Chee et al. 2015) using the primers Int_fwd_PstI and Int_rev_XbaI. The PCR product was then cloned into the *pCMVssInt-h/218* between *PstI* and *XbaI* sites (**Figure 2.1**), replacing the Int-h/218 sequence with Int-C3 sequence.

pCMVssInt-Opt was constructed by PCR amplification of Int-Opt (codon optimized

version of Int-h/218) coding sequence from pUC57-IntOpt (generated by GenScript) using the primers Int_fwd_PstI and Int_rev_XbaI. The PCR product was then cloned into the *pCMVssInt-h/218* between *PstI* and *XbaI* sites (**Figure 2.1**), replacing the Int-h/218 sequence with Int-Opt sequence.

pCMVssInt-C3-CNLS was generated by inserting the SV40 nuclear localization signal (NLS) sequence at the 3' end of Int-C3 coding sequence in *pCMVssInt-C3*. This was done by PCR amplification of Int-C3CNLS sequence from *pCMVssInt-C3* using the primers CNLS_XbaI_Int (which provides the NLS sequence) and Int_fwd_PstI. The Int-C3CNLS PCR product was then cloned into *pCMVssInt-C3* restricted with *PstI* and *XbaI*, replacing the Int-C3 sequence with Int-C3CNLS. Similarly, NLS sequence was also added to *pCMVssInt-Opt* and *pCMVssInt-h/218* generating *pCMVssInt-OptCNLS* and *pCMVssInt-h/218CNLS*.

pattPH-MCT5-cherry (**Figure 2.2**) was constructed by cloning *attPH-MCT5* sequence and m-cherry coding sequence into the backbone of *pattB-hygro* (Tan and Droge, unpublished) replacing the *attB-hygro* cassette with *attPH-MCT5-cherry* cassette. This was done by PCR amplification of *attPH-MCT5* sequence from *pEF-attPH-MCT5-neo* (in-house) (Tan and Droge, unpublished) using the primers attPH_fwd (HindIII) and attPH_rev (BamHI) followed by PCR amplification of m-cherry gene from *pEXC-mcherry* (in-house) using the primers M-cherry_fwd (BamHI) and M-cherry_rev (NheI). *pattB-hygro* vector was restricted with *HindIII* and *NheI* and three step ligation was carried out with *attPH-MCT5* and m-cherry restricted fragments to generate *pattPH-MCT5-cherry* (**Figure 2.2**).

Figure 2.2 A vector map of *pattPH-MCT5-cherry* consisting of *attPH-MCT5* sequence followed by coding region of m-cherry gene

pattPH-MCT5-cherry-PGKss-puro (**Figure 2.3**) was generated by using *attPH-MCT5-cherry* cassette (restricted with *HindIII* and *NheI*) from *pattPH-MCT5-cherry* (**Figure 2.2**) cloned into the vector backbone of *pPGKss-puro*, also restricted with *HindIII* and *NheI*.

Figure 2.3 A vector map of *pattPH-MCT5-cherry-PGKss-puro* consisting of *attPH-MCT5-cherry* cassette in the opposite orientation with respect to the *PGKss-puro* cassette

pPGKss-exon2-5 MCT5 (**Figure 2.4**) was constructed by cloning cDNA of MCT5 gene starting from ATG in exon 2 till *attH* sequence present in exon 5. PCR amplification of

cDNA of MCT5 from exon2 to exon5 (exon(2-5)-MCT5-(attH)) was done using total cDNA of HT1080 cells as template with the primers MCT5_exon2_CS (PstI) and MCT5_exon5_end (XbaI). The PCR product was then cloned into *pPGKss-puro* between *PstI* and *XbaI* sites replacing puromycin resistant gene with cDNA of *MCT5* from exon2 to exon5, generating *pPGKss-exon2-5 MCT5* (Figure 2.4).

Figure 2.4 A vector map of *pPGKss-exon2-5 MCT5* consisting of cDNA of MCT5 gene starting from ATG in exon 2 till *attH* sequence present in exon 5, driven by PGK promoter

pattP(chr6)-PGKss-neo (Figure 2.5) was generated through PCR-based site directed mutagenesis of the overlap sequence of *attP39* sequence in *pattP39-PGKss-neo* (Tan and Droge, unpublished data) using the primers *attP_chr6_SDM_fwd* and *attP_chr6_SDM_rev*. PCR-based site-directed mutagenesis using mutagenic primers was performed based on protocol described before (Zheng, Baumann et al. 2004) with modifications. PCR was performed with the template plasmid and mutagenic primers using the high fidelity *Pfu* polymerase (Thermo Scientific) in a 50µl reaction. The thermal cycling parameters used for PCRs were: initial denaturation step at 95°C for 5 min, 25 cycles of denaturation at 95°C for 30 s, annealing at 57°C for 1 min and extension at 68°C for 12 min, and a final step of 68°C for 5 min. The PCR amplified product was purified using PCR purification kit (Qiagen, GmbH) and subjected to digestion with *DpnI* (New England Biolabs) to selectively digest the methylated parental DNA template. An aliquot of the reaction was transformed into *E.coli* DH5α. Plasmid DNA was isolated from the transformants and analyzed by sequencing to verify the

introduced mutation and integrity of the plasmid.

Figure 2.5 A vector map of *pattyP(chr6)-PGKss-neo* consisting of *attP*(chr6) sequence followed by PGK promoter driving the expression of neomycin resistance gene

pattyP(chr8)-PGKss-neo was generated through PCR-based site-directed mutagenesis (as described above) of the overlap sequence of *attP*(WT) present in *pattyP(WT)-PGKss-neo* (Tan and Droge, unpublished) using the primers SDM_attP_ch8_fwd and SDM_attP_ch8_rev.

pattyP4X-PGKss-puro (**Figure 2.6**) was generated through PCR-based site-directed mutagenesis (as described above) of the overlap sequence of *attP*39 in *pattyP39-PGKss-puro* (Tan and Droge, unpublished data) using the primers attP(4X)sdm (For) and attP(4X)sdm (Rev).

Figure 2.6 A vector map of *pattiP4X-PGKss-puro* consisting of *attP4X* sequence followed by PGK promoter driving the expression of puromycin resistance gene

pCMVssInt-Ina (plasmid expressing Int with an inactivating mutation wherein the amino acid residue tyrosine at sequence position 342 is replaced by the amino acid alanine) was generated by with a similar PCR based site-directed mutagenesis protocol as above using *pCMVssInt-h/218* (**Figure 2.1**) as template with the primers SDM-Int-Y342A-F and SDM-Int-Y342A-R (contributed by Dr. Shree Harsha).

Plasmids for Int-C3 expression under EF1 α promoter in hESCs (Wen, Zhang et al. 2014). To construct *pEF1ass-Int-C3CNLS* (**Figure 2.7**), the EF1 α promoter sequence was PCR-amplified from *pEF1 α -EGFP* using the primers EF1Fw_EcoRI and EF1Rev-NheI and cloned between *EcoRI* and *NheI* sites in *pCMVssInt-C3* and (replacing only the CMV promoter sequence but retaining the “ss” hybrid intronic sequence (Petitclerc, Attal et al. 1995) (contributed by Dr. Shree Harsha).

Figure 2.7 A vector map of *pEF1αssInt-C3CNLS* expressing Int-C3 under EF1α promoter

pattP4X-PGKsspuro-UTF1-EGFP targeting vector (Figure 2.8) was generated by cloning UTF1-EGFP cassette in the opposite orientation with respect to PGKsspuro cassette. This was done by PCR amplification of UTF1-EGFP cassette from *pTZ-UTF1-EGFP* (using primers *EcoRV_UTF1_fwd* and *ClaI_UTF1_enhancer_rev*) and cloning into the *EcoRV* and *ClaI* sites of *pattP4X-PGKsspuro* (Figure 2.6) in the opposite orientation with respect to the PGKsspuro cassette.

Figure 2.8 A vector map of *pattP4X-PGKsspuro-UTF1-EGFP* consisting of *attP4X* sequence followed by PGKsspuro cassette in the opposite orientation with respect to UTF1-EGFP cassette

pTZ18R-attL4X/attR-PGKssPuro-UTF1-EGFP was constructed by cloning *attL4X*, *attR4X* sequences and PGKssPuro-UTF1-EGFP cassette in the *pTZ-18R* vector backbone. The *attL4X* fragment was PCR amplified from *pCMVssattL* (in-house) using primers attL(BOP')Fwd (KpnI) and attL(BOP') Rev(NotI). The *attR4X* fragment was PCR amplified from *pCMVssattR* (in-house) as template using the primers attR (POB') Fwd (ClaI) and attR (POB') Rev (HindIII). *pTZ18R* along with PCR-amplified *attL4X* and *attR4X* were restricted with *KpnI* and *HindIII* enzymes and a three fragment ligation generated *pTZ18R-attL/attR* flanked by *NotI* and *ClaI* sites. The three fragment ligated product *pTZ18R-attL4X/attR4X* flanked by *NotI* and *ClaI* sites and the *p(-attP4X)pgksspuro-UTF1-EGFP* fragment were cleaved with *NotI* and *ClaI* enzymes (**Figure 2.8**) and ligated to generate the *pTZ18R-attL/attR-PGKssPuro-UTF1-EGFP* target vector (**Figure 2.9**).

Figure 2.9 A vector map of *pTZ18R-attL/attR-PGKssPuro-UTF1-EGFP* consisting of UTF1-EGFP and PGKss-puro cassettes flanked by *attL4X* and *attR4X* sequences

All plasmids were confirmed by sequencing analysis.

2.3 Cell lines

This study used human embryonic stem cell (hESC) line [Genea 047 (46, XX)] (provided by Genea Biocells, Sydney, Australia <http://www.geneastemcells.com.au/>) and cancer cell lines of human origin A549 (lung epithelial carcinoma), HT1080 (fibrosarcoma), HeLa (cervical epithelial adenocarcinoma) and NEB-1 (immortalized neonatal foreskin keratinocytes).

2.4 Cell culture conditions

HT1080, A549 and HeLa cell lines were cultured in Dulbecco's Modified Eagle Medium (DMEM) growth medium supplemented with 10% FBS, 1% L-glutamine and 100 Units/ml of Penicillin and Streptomycin each (Gibco, Life technologies) at 37°C under 5% CO₂ in humidified condition. NEB-1 cells were cultured in RM+ medium {DME high glucose (326.25 ml), HAMS F12 (108.75 ml), Fetal Bovine Serum (10%), L-glutamine (1%), Penicillin / Streptomycin (100 Units/ml each) and RM+ Supplement (1%). RM+ Supplement is composed of Hydrocortisone (0.4 µg/ml), Insulin (5 g/ml), Adenine (1.8×10^{-4} M), Epidermal Growth Factor (10 ng/ml), Cholera Toxin (10⁻¹⁰ M), Transferrin (5 µg/ml), Liothyronine (2×10^{-11} M)}. For selection of puromycin-resistant recombinants, puromycin (Gibco, Life technologies) was added in the growth medium (1 µg/ml). Trypsin-EDTA (Gibco, Life technologies) was used for detaching the adherent cells for passaging.

hESCs were cultured at 37°C under 5% CO₂ and 5% O₂ in Collagen I-coated cell culture dishes (Biocoat, Corning) in Genea M2 Medium, (Genea Biocells, Sydney, Australia), supplemented with Penicillin and Streptomycin at 25 Units/ml each (Gibco, Life technologies). For selection of recombinants and maintenance of targeted clones, Neomycin (100-200 µg/ml) or Puromycin (300 ng/ml) (both from Gibco, Life technologies) was included in the growth medium. For passaging or preparing cell suspension for reverse transfections, adherent hESCs were rinsed with 1xPBS, detached by incubating at 37°C for 3 min with passaging solution (Genea Biocells) (with a volume of 100 µl per well of a 6 well plate or 1 ml per 10 cm dish), dislodging cells by tapping and suspending the cells with at least 3x volume of Neutralization solution (Genea Biocells). After counting the cells in a haemocytometer (Neubauer), they were pelleted by centrifuging at 300 x g for 4 min and resuspended in Genea M2 Medium to the required cell density and added drop-wise to Collagen I-coated dishes.

2.5 Differentiation of hESCs

Retinoic acid (RA) induced differentiation of the hESCs, was based on an established protocol (Savatier, Lapillonne et al. 1996). Briefly, hESCs were grown in 6 well plates to reach a confluence of 60-70% and M2 medium was replaced with DMEM (supplemented with 20% FBS, 1% L-glutamine and 100 Units/ml of Penicillin and Streptomycin each) containing Retinoic acid (R 2625, Sigma) at a final concentration of 1 µM and cultured

for 48 h at 37°C under 5% CO₂ in humidified condition. Thereafter the cells were grown in DMEM (supplemented with 20% FBS, 1% L-glutamine and 100 Units/ml of Penicillin and Streptomycin each). Puromycin (at 300 ng/ml) was included in the growth medium after RA treatment in experiments testing functionality of the UTF1 reporter cassettes in hESCs clones and differentiated progenies. Microscopy data acquisition and analysis was done using OLYMPUS IX71 microscope with OLYMPUS DP70 camera and DP Controller.exe software tool (OLYMPUS, Japan) and CorrSight™ FEI microscope, Oregon, USA.

2.6 Transfections

For transfections in HT1080, A549 and HeLa cell lines, 3×10^5 or 3×10^6 cells were seeded per well of 6 well plate (IWAKI, Japan) or per 10 cm tissue culture dishes (TPP, Switzerland), respectively, in DMEM growth medium a day before transfection to obtain 70-90% confluence at the time of transfection. Transfections were employed Lipofectamine2000 (Invitrogen, Life technologies) with DNA/mRNA to Lipofectamine2000 ratio of 1 µg: 2 µl. For every transfection per well, complexes were prepared by mixing DNA/mRNA and Lipofectamine 2000 reagent separately diluted in 100 µl of Opti-MEM medium (Life technologies) and incubating for 20 min at room temperature. The transfection mix was added drop wise onto the cells (under DMEM growth medium without antibiotics) and transfection was allowed to proceed for 4-6 h before complexes were removed by fresh DMEM medium.

For transfections in hESCs, FuGENE HD Transfection reagent (Promega) was used in a reverse transfection protocol. DNA to FuGENE ratio of 1 µg: 3 µl was used. Transfection mixes was prepared by first diluting plasmid DNA in 100 µl of Opti-MEM and 5 min later the FuGENE reagent was added to the DNA dilution, mixed and incubated for 15 min at room temperature for the complexes to form. During the incubation period, hESCs in culture were harvested (as described above) and resuspended in Genea M2 Medium (without antibiotics). The transfection complexes were added drop-wise to Collagen 1-coated plates and incubated at culturing conditions for 5 min after which the harvested cells were gently pipetted to the dishes at 5×10^5 cells per well of 6 well plate and 5×10^6 cells per 10 cm dish. Transfections were allowed overnight under standard culture conditions for hESCs, and media containing transfection complex was replaced with fresh M2 media.

2.7 Antibiotic selection and screening for targeted cell clones

48 h post-transfection, selection with the respective antibiotic in growth medium at the concentrations indicated above was initiated. Selection medium was replaced once in two days until colonies expanded to about 0.3 - 0.4 cm in diameter. At this stage, the colonies were picked from the plates by carefully scraping patches of cells with a pipette tip and transferred to 96 well plates for clonal expansion. The clones were sequentially expanded from 96 wells to 24 wells and subsequently in 6 well plates. Genomic DNA was extracted using DNeasy Blood & Tissue Kit (Qiagen, GmbH) as per manufacturer's protocol.

2.8 RNA extraction and cDNA synthesis

RNA extraction was performed using RNeasy Kit (Qiagen). HT1080 cells were trypsinized and extraction was carried using the manufacturer's protocol. The extracted RNA was reverse transcribed into cDNA using the Agilent First Strand kit.

After targeting HT1080 cells with *pattPH-MCT5-cherry-PGKss-puro*, RNA was extracted from puromycin resistant cells and cDNA was generated using SMART cDNA Synthesis Kit that consists of 3' SMARTer CDS primer (a modified oligo dT primer) and 5' SMARTer oligo. The tagged cDNA was used as a template for subsequent PCR amplifications.

2.9 Identification of sequence-specific recombination events by PCR screening

PCR was performed using GoTaq Flexi DNA polymerase (Promega) to amplify *attL* or *attR* junctions using primers listed in the figure descriptions and 200 ng of genomic DNA from each recombinant clone or parental cells as template in 50 reactions. The thermal cycling parameters used for PCRs was as follows: initial denaturation at 95°C for 5 min, 35 cycles of denaturation at 95°C for 1 min, annealing at 57°C for 30 s and extension at 72°C for 1 min, and a final step of 72°C for 5 min. The PCR samples were analyzed by electrophoresis in 0.8% agarose (Seakem Agarose, Lonza) gels in 1xTBE (Tris-Boric acid-EDTA buffer) containing 0.5 µg/ml ethidium bromide and PCR-amplified products were compared with DNA standard markers and digitally documented under UV illumination (Gel Doc 2K System, BioRad). PCR- amplified products were analyzed by sequencing.

2.10 Inverse PCR

Inverse PCR and corresponding nested PCRs were performed using Long Range PCR kit (Qiagen GmbH). Genomic DNA from the samples (obtained through co- transfection of target vectors and Int expressing plasmids) or parental cell line was restricted with *Pst*I restriction enzyme, followed by ligation (250 ng of DNA) favoring self-ligation of linearized DNA (as described by the NEB protocol). The purification of ligated DNA was done using PCR purification kit (Qiagen GmbH). DNA was subsequently used as template for inverse PCR with the primers in the vectors in opposite orientation (as described in figures in the Result Sections). 1-2 µl of the purified inverse PCR sample was used as template for the subsequent nested PCR. The thermal cycling parameters were: an initial denaturation step at 93°C for 3 min, 35 cycles of denaturation at 93°C for 15 s, annealing at 60°C for 30 s and extension at 68°C for 5 min and a final extension step at 68°C for 5 min. PCR-amplified products were analyzed by sequencing.

2.11 Southern blotting

Genomic DNA was purified using DNeasy Blood & Tissue Kit (Qiagen, GmbH). 15 µg of genomic DNA was subjected to restriction digestion using 50 U of the respective enzyme in 200 µl overnight at 37°C. DNA was ethanol precipitated and dissolved in 20 µl TE buffer (pH 8.0). Targeting vectors were linearized with single cutter restriction enzyme and diluted to 10^7 , 10^8 , 10^9 copies per µl. Digested genomic DNA samples were resolved overnight on a 1% agarose gel in 1x TAE (Tris-Acetate- Boric acid) buffer, with 1 kb DNA ladder (New England Biolabs) and 1 µl of positive control samples. Southern blotting employing the respective probes, as indicated, was performed using the DIG-High Prime DNA Labeling and Detection Starter Kit II (Roche) as per the manufacturers' protocol. The probe- target hybrids on the blots were detected by chemiluminescent assay followed by exposure to an X-ray film (Kodak MXG film, Kodak) and developed on a Kodak X-OMAT 2000 Processor.

2.12 Cell proliferation assay

HT1080 cells were co-transfected with 1 µg each of *pCMV-EGFP* and *pCMVssInt-Ina* or *pCMVssInt-C3CNLS* in 10 cm plates. 48 h post-transfection, GFP⁺ cells from each sample were FACS sorted and plated in triplicates, at a density of 5×10^3 per well in 24 well plates and allowed to attach overnight. Un-transfected HT1080 cells were included as control.

Cell proliferation rates were measured every day for 9 days using MTT-based In Vitro Toxicology Assay Kit (Sigma-Aldrich) as per manufacturer's protocol. Briefly, 3 h post-incubation with MTT dye-containing DMEM growth medium, cells were lysed with solubilization solution and 100 μ l sample from each well was transferred to 96 well plate. The absorbance was measured at 570 nm in a microtiter plate reader (Infinite 200Pro, Tecan). The average values from triplicate readings were determined and values for the blanks subtracted from the average. Absorbance values were plotted against time points in days.

2.13 Karyotyping

The targeted hESC clones and the parental hESC cell lines were karyotyped by G-banding of metaphase chromosomes as previously described (Lim, Lim et al. 2013). For each cell line 20 GTG-banded cells were scored and at least 5 GTG-banded cells were analyzed. Analysis and interpretation of data were in accordance to the International System for Human Cytogenetic Nomenclature (ISCN 2013). This analysis was carried out at the Cytogenetics Lab, Department of Pathology and Singapore General Hospital, Singapore.

2.14 Flow cytometry

FACS Calibur Flow Cytometer (Becton Dickson) and CELL quest software (Becton Dickson) were used to analyze and quantify GFP⁺ cells in a cell population. The cells were trypsinized, centrifuged and suspended in the corresponding media. Dot plot of side scatter (SSC) versus forward scatter (FSC) was used to gate live cells in order to separate them from aggregated and dead cells. For gated cells, a dot plot of GFP versus FSC was constructed for further analyses. Data were analyzed with FlowJo software and GFP⁻ and GFP⁺ cells for each sample were indicated in the lower right and upper right quadrant respectively.

2.15 Transcriptome analysis

Whole transcriptome analysis for three targeted hESC clones (#3, #29 and E3) and parental hESCs cells was carried out by AIT Biotech and Life Technologies, Singapore. The whole transcriptome workflow includes library construction, template preparation, sequencing and data analyzing (as shown in the flowchart) which is described below (in detail).

100 ng of total RNA was used as input for all clones (#3, #24, E3) for library construction using the Ion AmpliSeq™ Transcriptome Human Gene Expression Kit, according to the manufacturer's protocol. Briefly, 100 ng of total RNA was reverse-transcribed using random priming, and target genes were amplified using the Human Gene Expression Core Panel with the Ion AmpliSeq™ Library Kit Plus. Following target amplification, the resulting amplicons were treated with FuPA reagent for partial digestion of the primers and phosphorylation of the amplicons. Amplicons were then ligated to the Ion Xpress™ Barcode Adapters for barcoding. The 8 barcoded libraries were then quantified by qPCR (Ion Library Quantitation Kit) normalized and combined into a single library prior to template preparation and enrichment on the Ion OneTouch™ 2 System. During template preparation, libraries were clonally amplified on Ion Sphere Particles™ by emulsion PCR using the Ion PI™ Template OT2 200 Kit v3 on the Ion OneTouch™ 2 Instrument. Enrichment was then carried out on the Ion OneTouch™ ES (enrichment system) to isolate the template-positive Ion Sphere Particles™. The template-positive Ion Sphere Particles™ was then loaded onto the Ion PI™ Chip for subsequent sequencing using the Ion PI™ Sequencing 200 Kit v3 chemistry. Raw data were then processed on the Ion Proton™ Sequencer and transferred to the Ion Proton™ Torrent Server for primary data analysis with gene-level transcript quantification from sequence read data performed using the associated Torrent Suite™ analysis plugin, ampliSeqRNA. Briefly, sequence reads were aligned to the Ion AmpliSeq™ Transcriptome reference file in Torrent Suite™ Software using the Ion Torrent Mapping Alignment Program (TMAP). The reference file contains the entire set of RefSeq transcripts from which all 20,802 Ion AmpliSeq™ Transcriptome panel primers were designed. Following alignment, the ampliSeqRNA plugin examines the number of reads mapping to the expected amplicon ranges and assigns counts per gene for reads which align to these regions as defined in the BED file. The Ion AmpliSeq™ Transcriptome BED is a formatted file containing the nucleotide positions of each amplicon per transcript in the mapping reference. Reads aligning to the expected amplicon locations are referred to as 'on target' reads and are reported as a

percentage of total reads by the plugin. The gene expression data obtained was further analyzed using Microsoft Excel (MS Office) wherein, a cut-off was set at 4-fold to determine the differential gene expression between hESC clones and hESC parental cell lines.

Page intentionally left blank

3. RESULTS

In order to exploit the natural λ -Int-mediated sequence-specific recombination system for targeting a eukaryotic genome, various Int mutants (Christ and Droge 1999, Lorbach, Christ et al. 2000, Tay, Ho et al. 2010, Siau, Chee et al. 2015) were developed. These mutants not only work independently of co-factors but also have higher affinity to both wild type and non-cognate *att* sites, in comparison to the wild type λ -Int. The development of each of these mutants is summarized in the following section (**Section 3.1**). We carried out episomal-DNA based recombination assays to determine the efficiency of various Int mutants (**Section 3.2**) and their potential utility in targeting (**Section 3.3-3.8**) the human genome.

3.1 Development of λ -Int mutants

The first Int variant, Int-h, was identified in a screen for phage mutants that were able to integrate efficiently in *E.coli* strains lacking the IHF protein (Miller, Mozola et al. 1980). The Int-h protein has a lysine at codon 174 (E174K) in the catalytic domain, instead of a glutamate. Subsequent studies revealed that Int-h is able to carry out both integrative and excisive recombination in the absence of various co-factors and negative supercoiling of *attP* (Lange-Gustafson and Nash 1984). The improved activity of Int-h compared to wild type λ -Int was attributed to its higher affinity to the core binding sites. Int-h could also carry out recombination between non-cognate *att* sites (Patsey and Bruist, 1995).

Another mutant, Int-h/218, was derived from Int-h, by replacing a glutamate with a lysine at codon 218 (E218K) (Lorbach, Christ et al. 2000, Christ and Droge 1999). The lysine at this position is thought to further increase the affinity of Int-h to the core binding sites (Wu, Gumport et al. 1997). Thus, Int-h/218 (E174K, E218K) has enhanced activity to carry out recombination between wild type as well as non cognate *att* sites in the absence of accessory co-factors and supercoiling (Christ and Droge 1999).

Two additional mutants Int-1 (I43A and H61R) and Int-5 (I43A) were subsequently generated from Int-h/218 (Tay, Ho et al. 2010). These variants were evolved to recombine *attB* like sequence termed as *attH* (in human genome), which differs at one position in the overlap sequence and three positions at the right core site (B') (**Figure 3.1**). The *attH* sequence is found on the chromosome 1 in the exon 5 of the *MCT5* (MonoCarboxylate Transporter 5; accession number SLC16A4) gene, serving as a potential site in the human

genome for specific targeting by the corresponding phage sequence *attPH* (Tay, Ho et al. 2010).

(Tay, Ho et al. 2010)

Figure 3.1 Sequence alignments of WT core *attB/attP* sites with the human *attH/attPH-MCT5* sequence. The underlined letters indicate the positions where *attH-MCT5* differs from WT *attB*. The region highlighted in grey represents the 7bp overlap sequence, which must be a perfect match in the respective *attP/attPH* recombining partners. The asterisks below the sequences indicate nucleotides that may interact with integrase

The mutations I43F and H61R are found in the N-terminal domain of Int, which are responsible for binding to *attP* arm sites (**Figure 3.2**). The mutation I43F incorporates the bulkier phenylalanine residue in an alpha-helix, that helps in strengthening the interaction of the N-terminal tail and β -sheets with the minor and major groove of arm sites of *attP*, respectively, thus facilitating the recombination between non-cognate *att* sites (Tay, Ho et al. 2010). On the other hand, H61R lies in a loop region about which the N-terminal domain can rotate. It has been postulated that H61R along with I43F can affect the dynamics of allosteric rearrangements in the N-terminal domain that can drive the resolution of Holliday junction conformations potentiating the recombination activity of the mutant Int (Tay, Ho et al. 2010). However, the exact reasons why these mutations enhance Int recombinogenic activity are still not known.

Adapted from (Tay, Ho et al. 2010)

Figure 3.2 An illustration showing positions of I43 and H61 in the N-terminal domain of λ -Int. A structure of three different domains of λ -Int: N; N-terminal domain, CB; core-binding domain and C; catalytic domain. The box (in black) shows I43 (in alpha-helix) and H61 (in loop region) harbored in the N-terminal domain of λ -Int bound to an arm-binding site

Further efforts were made to develop more active variants of Int. In this context, a novel selection method was devised wherein the Int activity leads to the deletion of a gene encoding the inhibitor of an enzyme resulting in a selectable phenotype (**Figure 3.3**) This methodology is called *in vivo* substrate linked protein evolution (SLiP), in which a library of randomized mutant Int genes was cloned into an expression/selection vector consisting of the *attP/attPH* and *attB/attH* sequence (where *attB* or an *attH* sequence is used as a linker between TEM1 (β -lactamase) and BLIP (β -lactamase inhibitory protein) encoding gene) (**Figure 3.3**). Expression of the mutant Int in *E.coli* recombines *att* sequences leading to the deletion of the BLIP gene and thus conferring ampicillin resistance to the cells. By increasing selection pressure, two Int-h/218-derived mutants, Int-C2.1 (I43F, D336V) and Int-C3 (I43F, D336V, E319G), were identified which showed higher activity compared to the parental Int-h/218 (Siau, Chee et al. 2015).

Adapted from (Siau, Chee et al. 2015)

Figure 3.3 Generation of Int mutants via substrate linked protein evolution (SLiP). The selection plasmid encodes the mutated λ -Int and a β -lactamase (TEM1)–BLIP fusion protein. The shorter *attH* site (yellow) is positioned as a linker connecting TEM1 and BLIP. After expression of λ -Int from the same plasmid, *attH* X *attPH* (purple) recombination results in *attL1* (orange) and deletion of the BLIP gene leading to expression of TEM1 from resultant plasmid and antibiotic resistance (Siau, Chee et al. 2015)

The D336V mutation, found in both Int-C2 and -C3 lies in the vicinity of tyrosine 342 (involved in DNA cleavage) in the catalytic domain and has been proposed to increase the flexibility of the region near the catalytic domain for efficient cleavage (**Figure 3.4**) (Siau, Chee et al. 2015). On the other hand, the E319G mutation, also found in the catalytic domain (**Figure 3.4**) is known to be involved in relaxing the substrate specificity of Int due to which it has been proposed to tolerate the mismatches at the core-binding sites for the formation and resolution of Holliday junction (Dorgai et al, 1995, Han et al, 1994; Kazmierczak et al, 2002).

Adapted from (Biswas *et al.*, 2005) (Siau, Chee *et al.* 2015)

Figure 3.4 An illustration showing positions of D336 and E319 in the C-terminal domain of λ -Int. A magnified image of C-terminal domain of λ -Int, demonstrating the proximity of D336 and E319 to the Y342 nucleophile. D336 is in the highly flexible region as Y342 (yellow), whereas E319 is located near to both Y342 nucleophile and C-terminal region (cyan), responsible for controlling Int catalytic activity. Two catalytic domains (green and magenta) are shown with CB domain (in grey)

3.2 Comparative analysis of various Int mutants

An established episomal-based recombination assay (Christ, Corona *et al.* 2002) was employed to compare the efficiencies of various Int mutants in human fibrosarcoma HT1080 cells. An episomal substrate plasmid called *pLIR* (Christ, Corona *et al.* 2002) containing *attB/attP* and *attH/attPH* sites (as described in **Section 3.1, Figure 3.1**) was used to correlate λ -Int-mediated recombination to EGFP expression (**Figure 3.5**) (Christ, Corona *et al.* 2002). The proportion of EGFP⁺ cells (positive recombination events) was quantified using flow cytometry and normalized to *pCMV-EGFP*-transfected cells.

Figure 3.5 An episomal DNA-based recombination assay. An episomal substrate plasmid *pLIR* consisting of *attB/attH*-neomycin resistance gene cassette under the CMV promoter and *attP/attPH*-EGFP cassette is co-transfected in combination with various Int expressing plasmids in HT1080 cells. After the Int-driven intra-molecular recombination between *attB/attH* and *attP/attPH*, EGFP is expressed from the CMV promoter, thus enabling quantification of EGFP⁺ cells by FACS

Mutants Int-1 (*pCMVssInt-1*) and Int-5 (*pCMVssInt-5*) (Tay, Ho et al. 2010) were compared to Int-h/218 (*pCMVssInt-h/218*) using *attB/attP* and *attH/attPH* as recombination substrates. Our results indicate that Int-h/218 is more active than Int-1 and Int-5 mutants (**Figure 3.6**). Here, *pCMV* transfected cells (Mock) served as negative control.

Figure 3.6 Comparative analysis of Int-h/218, Int-1 and Int-5. Approximately 3×10^5 HT1080 cells were co-transfected with the indicated Int (Int-h/218, Int-1 and Int-5) expression plasmid and substrate plasmid *pLIR*, linking Int activity to EGFP expression. Activity of Int is expressed as a percentage of EGFP⁺ cells, determined by FACS analysis. Bars represent the percentage of EGFP⁺ cells obtained in *attB* X *attP* (blue) and *attH* X *attPH* (red) recombination assay using *pLIR* with Int-h/218, Int-1 and Int-5. EGFP; *pCMV-EGFP* transfected cells used as positive control. Mock; *pCMV* co-transfected with *pLIR*, used as negative control. N=3, Mean +/- SD. Statistical significance between Int-h/218 Int-1 and Int-5 was performed using Unpaired Student t test. P* < 0.05, P** < 0.01. Results are cumulative of three independent experiments

In a second set of experiments, Int-C2 (*pCMVssInt-C2*) and Int-C3 (*pCMVssInt-C3*) (Siau, Chee et al. 2015) were compared to Int-h/218, where cells transfected with Int-INA (*pCMVssInt-Ina*, expressing Int with an inactivating mutation (Y342A) wherein the tyrosine at sequence position 342 is replaced by alanine) were used as negative control (Figure 3.7). The results show that Int-C3 is more active in catalyzing *attH* X *attPH* recombination than Int-h/218 and Int-C2, but for *attB* X *attP* recombination, activity of Int-C3 is similar to Int-C2, but higher than Int-h/218.

Figure 3.7 Comparative analysis of Int-h/218, Int-C2 and Int-C3. Approximately 3×10^5 HT1080 cells were co-transfected with the indicated Int expression plasmid and episomal substrate plasmid *pLIR* linking Int activity to EGFP expression. Activity is expressed as a percentage of EGFP⁺ cells, determined by FACS analysis. Bars represent the percentage of EGFP⁺ cells obtained in *attB* X *attP* (blue) and *attH* X *attPH* (red) recombination assay using *pLIR* with Int-h/218, Int-C2 and Int-C3. EGFP (positive control); *pCMV-EGFP* transfected cells. Int-INA; Int-INA co-transfected with *pLIR* used as negative control. N=3, Mean +/- SD. Statistical significance was performed between Int-h/218 Int-C2 and Int-C3 using Unpaired Student t test. P* < 0.05, P** < 0.01. Results are cumulative of three independent experiments

Combining the results of the episomal recombination assays (**Figure 3.6 and 3.7**), it can be concluded that the activity of Int-C3 outperforms the other Int variants for *attH* X *attPH* recombination. However, for *attB* X *attP* recombination, the activity of Int-C3 was similar to Int-C2, but higher than the activity of Int-h/218.

3.3 Targeting endogenous *attH*-MCT5 using mutant Int-h/218

In order to exploit the Int variants for human genome manipulations, we initially targeted a 21bp *attH* sequence on chromosome 1 in exon 5 of the *MCT5* (MonoCarboxylate Transporter 5) gene. It differs from *attB* at one position in the overlap sequence and three positions in the right arm (B') core-binding sequence (**Figure 3.1**) (Tay, Ho et al. 2010). This sequence was termed as *attH*-MCT5 and its corresponding phage sequence was termed as *attPH*-MCT5 (**Figure 3.1**).

3.3.1 Determination of *MCT5* expression in HT1080 cells

The *attH*-*MCT5* sequence has been shown to recombine episomally with its corresponding *attPH* sequence (**Figure 3.5**). However, prior to targeting the endogenous *attH*-*MCT5* sequence, it was important to check the expression of the *MCT5* gene in HT1080 cells, which was done using RT-PCR.

Note: *MCT5* gene is known to have 12 transcript variants (Kirat, Sallam et al. 2009) but due to the primer positions (in exon 4 and exon 7) only three transcript variants (*MCT5* isoforms 1, 3 and 5) should be detected.

To carry out RT-PCR, RNA was isolated from HT1080 cells and cDNA was generated via reverse transcription. PCR amplification using total cDNA of HT1080 cells as template with the primer pair (*MCT5* fwd 1 & *MCT5* rev 1) resulted in expected size of 1000bp in *MCT5* isoforms 1 and 3 whereas 417bp in *MCT5* isoform 5 (**Figure 3.8b** *MCT5* lane 1). Similarly, PCR amplification with primer pair (*MCT5* fwd 2 & *MCT5* rev 2) resulted in expected size of 710bp in *MCT5* isoforms 1 and 3 whereas 207bp in *MCT5* isoform 5 (**Figure 3.8b** *MCT5* lane 2). The sequencing of all three isoforms confirmed the presence of *attH* sequence in the exon 5 of *MCT5* gene.

Figure 3.8 Determination of *MCT5* expression. (a) Schematic diagram of transcript variants 1, 3 and 5 generated from the *MCT5* gene. Position of *attH* in exon 5 of the *MCT5* gene, location of primers in exon 4 and exon 7 (*MCT5* fwd 1 & *MCT5* rev 1; *MCT5* fwd 2 & *MCT5* rev 2) along with the expected sizes are indicated. (b) PCR amplification using total cDNA from HT1080 cells with primers *MCT5* fwd 1 and *MCT5* rev 1 (as shown in figure 3.8a) resulted in expected size of 1000bp in *MCT5* isoforms 1 and 3 whereas 417bp in *MCT5* isoform 5 (indicated in red box and in *MCT5* lane 1). PCR amplification using total cDNA from HT1080 cells with primers *MCT5* fwd 2 and *MCT5* rev 2 (as shown in figure 3.8a) resulted in expected size of 710bp in *MCT5* isoform 1 and 3 whereas 207bp in *MCT5* isoform 5 (indicated in blue box and *MCT5* lane 2). Lanes. M, 100bp marker; *MCT5* lane 1, PCR products from *MCT5* isoform 1, 3 and 5 with indicated primers; *MCT5* lane 2, PCR products from *MCT5* isoform 1, 3 and 5 with indicated primers

3.3.2 An episomal DNA-based intermolecular recombination assay to determine the functionality of the *MCT-5* locus post-targeting

To determine the functionality of the *MCT5* locus, the target vector *pattPH-MCT5-cherry* was constructed (**Figure 2.2**; Materials and Methods). After λ -Int-mediated recombination at the endogenous *attH*-*MCT5* site, the m-cherry gene should be expressed in fusion with the *MCT5* gene (*MCT5*-cherry).

However, prior to targeting the chromosomal *attH*-*MCT5* site, an episomal DNA-based intermolecular recombination test was performed between two plasmids to ensure functionality of the *MCT5*-cherry fusion protein post-recombination. One of the plasmids (*pPGKss-exon2-5MCT5*) has the *MCT5* coding region from the start codon until exon 5 (which contains the *attH* sequence) driven by the PGK promoter, while the other plasmid (*pattPH-MCT5-cherry*) consists of the *attPH*-*MCT5* sequence followed by the cherry gene (**Figure 3.9**). Co-transfection of these two plasmids with an Int-h/218 expression vector into HT1080 cells resulted in expression of cherry gene (data not shown). A *pCMV* vector co-transfected with the substrate recombination plasmids was used as negative control.

Figure 3.9 An episomal DNA-based intermolecular recombination test to ensure functionality the of the *MCT5* locus. A schematic diagram showing Int-h/218-mediated episomal DNA-based recombination between *pPGKss-exon2-5MCT5* and *pattPH-MCT5-cherry* recombination plasmids

Genomic and episomal DNA was extracted 36 h post-transfection from cherry positive cells (not shown) and PCR was performed using primers *MCT5* fwd 1 & *cherry* rev (**Figure 3.9**). The PCR analysis resulted in the predicted product size (~900bp) with Int-h/218-transfected cherry positive cells, which was not detectable in *pCMV*-transfected cells (**Figure 3.10**). DNA sequencing of PCR products confirmed the expected *attH*-*MCT5* X *attPH*-*MCT5* recombination product, only with Int-h/218.

Figure 3.10 PCR analysis of episomal DNA-based intermolecular recombination. PCR amplification using genomic DNA from *pCMV* transfected cells (CMV) and Int-h/218 transfected cherry positive cells (Int-h/218) with primers (MCT5 fwd 1& Cherry rev; Figure 3.9), resulted in the expected ~900 bp product with Int-h/218. Lanes. M1; 1kb ladder, CMV; Genomic DNA from *pCMV* co-transfection with *pPGKss-exon2-5MCT5* and *pattPH-MCT5-cherry* in HT1080 cells, Int-h/218; Genomic DNA from Int-h/218 co-transfection with *pPGKss-exon2-5MCT5* and *pattPH-MCT5-cherry* in HT1080 cells

3.3.3. Targeting the genomic attH-MCT5 locus using Int-h/218

pattPH-MCT5-cherry target vector (Figure 3.11) was used to target the genomic attH-MCT5 using the Int-h/218 (as the activity of Int-h/218 is higher than Int-1 and Int-5 mutants as shown in Figure 3.6).

Note: The Int-C3 mutant had not been developed at this time.

For endogenous targeting, HT1080 cells were split into six 10cm plates and the following co-transfections were carried out.

Plate#	<i>pattPH-MCT5-cherry</i>	Int-h/218
1.1	0.5 µg	10 µg
2.1	1 µg	10 µg
3.1	2 µg	10 µg
4.1	0.5 µg	10 µg
5.1	1 µg	10 µg
6.1	2 µg	10 µg

Cherry positive cells obtained from Int-h/218 co-transfections were sorted using flow cytometry (BD FACSaria; data not shown) and the genomic DNA was extracted. The *attH* X *attPH* recombination events were analyzed by PCR using genomic DNA from cherry positive cells as template with the primer pairs (F1_ *attL*_MCT5 & R1_ *attL*_MCT5) and (F2_ *attR*_MCT5 & R2_ *attR*_MCT5) (**Figure 3.11**). However, none of the PCR results showed the expected recombination events (data not shown).

Figure 3.11 Targeting of the genomic *attH*-MCT5 by *pattPH-MCT5-cherry* using Int-h/218. A potential docking site (DS) already present in exon 5 of the MCT5 gene (*attH*-MCT5) is targeted by *pattPH-MCT5-cherry* using Int-h/218, which should result in recombinant MCT5-cherry fusion protein. Positions of relevant primers (F1_ *attL*_MCT5 & R1_ *attL*_MCT5 and F2_ *attR*_MCT5 & R2_ *attR*_MCT5) used for PCR amplification of recombination junctions *attL* and *attR* are shown

In order to rule out the possibility of λ -Int-mediated recombination at a different (unknown) *att* site (other than MCT5), a SMART (Switching Mechanism at 5' end of RNA Template) cDNA synthesis kit, which amplifies full-length cDNAs by incorporating adaptors at both 5' and 3' ends, was employed (**Figure 3.12a**). RNA was isolated from

Int-h/218-transfected cherry positive cells (not shown) and cDNA tagged with 5' and 3' adaptors was generated. Assuming *pattPH-MCT5-cherry* integrated elsewhere in the genome resulting in a fusion between genomic coding region and the cherry gene, a PCR was performed with tagged cDNA using 5' complementary Smarter Oligo V (provided with the kit) and 3' Cherry rev primer (Figure 3.12a). The PCR resulted in a few DNA bands that were sequenced (Figure 3.12b).

Figure 3.12 Determination of potential genomic integration sites of the target vector using SMART cDNA synthesis kit. (a) Reverse transcription of RNA isolated from Int-h/218 transfected cherry positive cells into full length cDNA using the 5' complementary Smarter Oligo V and in-fusion SMARTer CDS primer (provided with the kit), generating full length cDNA tagged with 5' and 3' adaptors. **(b)** PCR analysis using full-length enriched cDNA as the template with 5' complementary Smarter Oligo V and cherry rev primers, as illustrated in (a). Gel extraction of expected cherry fusion PCR products obtained with Int-h/218. Lanes. M; 100bp ladder, cherry-fusion (Smart Kit); PCR product obtained using full-length tagged cDNA as template with Smarter Oligo V and cherry rev primers, M1; 1 kb ladder

The sequencing results suggested that an unexpected episomal intermolecular recombination occurred between *pCMVssInt-h/218* (Int-h/218) and *pattPH-MCT5-cherry* vectors driving the expression of the cherry gene through the CMV promoter.

3.4 Searching for other endogenous att sites using an inverse PCR strategy

Since λ -Int-mediated targeting of the endogenous *attH-MCT5* locus with *pattPH-MCT5-cherry* was not successful, other target vectors (consisting of different *attP* variants) were explored to identify *att* like sequences in the human genome via an inverse PCR strategy. Co-transfection of the target vector and an Int expressing plasmid was followed by

selection and enrichment of recombinants using fluorescence and antibiotic-selection markers present in the target vector. The genomic DNA from selected clones was extracted, cleaved with a restriction enzyme (*PstI*, in most of our cases), which specifically cleaves the vector at a predetermined known site and at sites in the genome, followed by self-ligation and purification of linearized DNA (Section 2.10 of Materials and Methods). These DNA molecules were subjected to inverse PCR using oppositely orientated primers at the known vector sequences to amplify the recombinant sites (*attL* and *attR*) as described below (Section 3.5). 1-2 µl of the purified inverse PCR sample was used as template for subsequent nested PCRs, if necessary.

3.5 Identification of a potential endogenous *attH* sequence using *attP39* and Int-h/218

A variant of *attP* sequence termed *attP39* (already available in the lab) was used to target the human genome

attP(WT): CAGCTTTTTTTATACTAAGTTG

attP39: CAGCTTTTTTAAATAAGTTG

In this context, a target vector *pattP39-PGKss-neo* was generated wherein the *attP39* sequence was cloned upstream of the PGK promoter and a neomycin resistance gene (Figure 3.13a), thus enabling selection of targeting events by the use of geneticin. To target a potential *attH* site, HT1080 cells were split into six 10 cm plates and the following co-transfections were done using Int-h/218 and *pCMV* (negative control).

Plate#	<i>pattP39-PGKss-neo</i>	Int-h/218
1.2	0.5 µg	10 µg
2.2	1 µg	10 µg
3.2	2 µg	10 µg
Plate#	<i>pattP39-PGKss-neo</i>	<i>pCMV</i>
4.2	0.5 µg	10 µg
5.2	1 µg	10 µg
6.2	2 µg	10 µg

Note: Int-C3 mutant had not been developed at this time.

After two weeks of selection at 500µg/ml geneticin, neomycin resistance colonies were pooled and bulk genomic DNA from each plate was extracted. The genomic DNA from plates #2.2 and #5.2 was digested overnight with *Pst*I that cleaved the vector downstream of the PGK promoter and in the genomic DNA (**Figure 3.13a**). 250ng of *Pst*I restricted linearized DNA (also containing vector sequences) was self-circularized in 250ul of ligation mixture reaction (to prevent concatemer formation). In order to determine the *att*R recombinant site, the circular ligated DNA was subjected to inverse nested PCR with oppositely oriented primers attP fwd 2 & neo rev 44 in the target vector using a template obtained with a primary PCR (attP fwd 1 & neo rev 81; **Figure 3.13a**). A unique PCR product was obtained with bulk genomic DNA of plate #2.2 (with Int-h/218), which was not seen with DNA from plate #5.2 (with *pCMV*) (**Figure 3.13b**). The “specific” PCR product (highlighted in **Figure 3.13b**) was extracted and sequenced, which indicated λ-Int-mediated integration in chromosome 6 (**Figure 3.13c**).

(a)

(b)

Figure 3.13 Identification of potential endogenous *attH* sequence using *attP39*. (a) A schematic diagram showing targeting of a potential *attH* locus using *pattP39-PGK-neo* along with *Int*, leading to the formation of recombinant *attL* and *attR* sites. The recombinant product is digested with *PstI*, followed by self-ligation. Positions of relevant primers used for inverse nested PCR are indicated. (b) PCR analysis using genomic DNA from plate #2.2 (*Int-h/218*) plate and #5.2 (CMV) with primers attP fwd 1 & neo rev 81, resulted in primary PCR products (CMV-1 and *Int-h/218*-1). Inverse nested PCR using (CMV-1 and *Int-h/218*-1) as template with primers attP fwd 2 & neo rev 44 resulted in unique band in *Int-h/218*-2 (highlighted in white box). Lanes. M1, 1kb ladder; W, water control; HT, control genomic DNA from HT1080 cells; CMV-1/2, Primary and nested PCR products using genomic DNA from plate #5.2 respectively; *Int-h/218*-1/2, Primary and nested PCR products using genomic DNA from plate #2.2 respectively. (c) Sequencing result showing the *attR* recombinant site obtained with *attP39* targeting. The underlined sequence indicates the 21bp *attR* site where the 7bp left core-binding site and the 7bp overlap is part of *attP39* (in red) and the right core-binding site comes from human chromosome 6 (in black). The sequence in black indicates the sequence on chromosome 6 until the *PstI* site where it is ligated to the neomycin resistance gene (in blue) of the vector

Based on the sequencing results obtained (Figure 3.13c) and using bioinformatics, an *attH* sequence was identified on chromosome 6, termed as *attH* (chr6). It differs from *attP39* at three nucleotides in the core binding sites and two nucleotides in the overlap as indicated in red below:

attP39: CAGCTTT**TTT**AAAATAAG**TTG**
attH(chr6): CA**ACTTT**ATGAAAATAAT**GTG**

To confirm site-specific integration, genomic primers flanking the *attH* (chr6) sequence were designed and a nested PCR was performed to amplify *attL* and *attR* recombinant sites using the primers illustrated in **Figure 3.14a**. The results showed amplifications of expected sizes for the *attL* site (~2000bp) and the *attR* site (~400bp) (**Figure 3.14b**) and moreover, sequencing results confirmed the precise targeting at *attH* (chr6) site without any mutations.

Figure 3.14 Confirmatory PCR for *attH*-chr6 targeting. (a) A schematic diagram of *pattP39-pgk-neo* targeting vector after integration into *attH* (chr6). Positions of relevant primers used for PCR amplification of recombinant *attL* and *attR* sites in the genome are indicated. (b) PCR analysis using genomic DNA from plate #2.2 (Int-h/218) and plate #5.2 (CMV) with primers *chr6 fwd 1* & *neo rev 81* for the *attL* site and *chr6 rev 1* & *P_21* for the *attR* site, resulted in primary PCR products (CMV-1 and Int-h/218-1) (primary PCR data not shown for *attR*). Inverse nested PCR using CMV-1 and Int-h/218-1 as templates with primers *chr6 fwd 2* & *neo rev 44* for the *attL* site and *chr6 rev 2* & *attP fwd 1* for the *attR* site resulted in unique bands in Int-h/218-2 for both for *attL* and *attR* (highlighted in white box), which was not present in CMV-2. Lanes. M1, 1kb ladder; W, water control; HT (1,2), Primary and nested PCR for the *attL* junction using genomic DNA from parental HT1080 cells; CMV (1,2), Primary and nested PCR for *attL* junction using genomic DNA from plate #5.2 (CMV) respectively; Int-h/218 (1,2), Primary and nested PCR of the *attL* junction using genomic DNA from plate #2.2 (Int-h/218) respectively. HT-2, Nested PCR for the *attR* junction using genomic DNA from parental HT1080 cells; CMV-2, Nested PCR for the *attR* junction using genomic DNA from plate #5.2 (CMV); Int-h/218-2, nested PCR of the *attR* junction using genomic DNA from plate #2.2 (Int-h/218)

Based on these data, a new target vector *pattP(chr6)-PGKss-neo* was constructed (**Figure 2.5**; Materials and Methods), with the aim of increasing the targeting efficiency, where the overlap sequence of *attP39* was adjusted to the sequence of the presumed target site in *attH* (chr6). A targeting experiment was subsequently carried out in HT1080 cells in two 10 cm dishes as follows:

Plate#	Target vector	Co-transfected with
1.3	1 µg <i>pattP(chr6)-PGKss-neo</i>	10 µg Int-h/218
2.3	1 µg <i>pattP(chr6)-PGKss-neo</i>	10 µg <i>pCMV</i>

48 individual HT1080 colonies were picked from plate #1.3 and recombination was probed using specific primers on chromosome 6 (**Figure 3.14a**). However PCR results suggested that none of these colonies were positive for λ -Int-mediated recombination at *attH* (chr6) (data not shown).

3.5.1. Unexpected targeting at chromosome 13 instead of chromosome 6 using Int-h/218

The colonies obtained from plate #1.3 were subsequently analyzed for targeting events via inverse PCR. Out of 48 picked colonies, 6 colonies were pooled to obtain 8 sample groups (C1-8) for analysis. Bulk genomic DNA from C1-C8 was purified and digested with *PstI*. Self-ligation and subsequent inverse, nested PCR resulted in a unique band of more than 2kb in sample C3, which was not present in the either of the negative controls, *pCMV* transfected cells or HT1080 cells (**Figure 3.15a,b**). Strikingly, sequencing of the PCR product obtained from sample C3 revealed λ -Int-mediated targeting on chromosome 13, instead of chromosome 6 (data not shown).

Figure 3.15 Targeting at chromosome 13. (a) A schematic diagram of *pattP39-PGKss-neo* targeting vector after integration into the putative *attH* site using Int-h/218. The genomic DNA with the integrated recombinant product is digested with restriction enzyme *PstI*, followed by self-ligation. Position of the relevant primers used for inverse PCR are indicated. **(b)** Inverse nested PCR screening for *attH* X *attP39* recombination using primary PCR products (primers attP fwd 1 & neo rev 81) as template with primers attP fwd 2 & neo rev 44 resulted in a unique band in sample C3 (highlighted), Lanes: M, 100bp ladder, W, water control; HT, genomic DNA from HT1080 cells; CMV; pooled genomic DNA from plate #2.3; C1, C2, C3, C4, C5, C6, C7, C8, genomic DNA obtained from pooled colonies from plate #1.3; M1, 1kb ladder

The pooled sample C3 comprised of genomic DNA from 6 colonies (#13, #14, #15, #16, #17, #18). To confirm the correct and specific clone that would have undergone recombination, the 6 individual colonies were expanded and their genomic DNA was isolated. An inverse PCR strategy with *PstI* restricted genomic DNA from clones #13, #14, #15, #16, #17, #18, suggested that the targeting was successful in colony #16 (data not shown). Confirmatory sequencing analysis of colony #16 showed λ -Int-mediated recombination in chromosome 13 (**Figure 3.16**). However, detailed sequencing revealed that there is an addition of 19 nucleotides incorporated upstream of targeted sequence (marked in green after the overlap of *attP39*) (**Figure 3.16**).

Figure 3.16 Sequencing result of *attR* recombinant site on chromosome 13. Sequencing result showing the recombination junction *attR* (P39OH') on chromosome 13, obtained with *attP39* vector. The underlined sequence indicates the 21bp *attR* site where the 7bp left core-binding site and the 7bp overlap is part of *attP39* (in red) whereas the sequence in green indicates an extra 19 nucleotides that have been added at the site of integration. The sequence in black indicates the integration in chromosome 13 until the *PstI* site where it is ligated to the neomycin resistance gene (in blue) of the vector

Nevertheless, PCR with genomic DNA from colony #16 using primers on chromosome 13 and in the target vector (**Figure 3.17a**) was done to amplify the *attL* and *attR* recombinant sites. No PCR product was obtained for the *attL* junction whereas the PCR products with expected sizes (~471bp and 269bp) were obtained for *attR* junction (**Figure 3.17b**). Together these data (absence of the *attL* junction and unexpected addition of extra nucleotides at the *attR* junction), suggested that λ -Int-mediated recombination into this site may have been aborted.

Figure 3.17 Confirmatory PCR for *attH*-chr13 targeting. (a) Schematic diagram of *pattP39-PGKss-neo* targeting vector after integration into *attH*. Positions of relevant primers used for PCR amplification of recombinant *attL* and *attR* sites on chromosome 13 of clone #16 are indicated. (b) PCR analysis using genomic DNA extracted from colony #16 with primers (chr13 fwd 1 & attP rev 1) and (chr13 fwd 2 & attP rev 1) for the *attL* site did not result in any PCR product, whereas primers (chr13 rev 1 & attP fwd 1) and (chr13 rev 2 & attP fwd 1) (figure 3.17a) resulted in the expected 470bp and 260bp bands, respectively, for the *attR* site. These products were not observed with genomic DNA from parental HT1080 cells. Lanes. M, 100bp ladder; W, water control; HT, genomic DNA from parental HT1080 cells; #16, genomic DNA from clone #16; M, 100bp ladder

In parallel with targeting attempts described in section 3.5.1, a similar experiment was set up in 10 cm dishes as follows:

Plate#	<i>pattP39-PGKss-neo</i>	Int-h/218
1.4	50 ng	10 µg
2.4	0.5 µg	10 µg
3.4	1 µg	10 µg
4.4	2 µg	10 µg

After 2 weeks of selection with geneticin, 132 colonies were picked from plate #1.4 and plate #2.4 and of these, 6 colonies were pooled resulting in 22 samples for subsequent analysis. Genomic DNA was extracted and labeled as K1-K22. Furthermore, total bulk genomic DNA extracted from each of the plates (#1.4, #2.4, #3.4, #4.4) were labeled as K23, K24, K25 and K26, respectively, and genomic DNA from *pCMV* transfected plate and HT1080 cells used as negative controls were labeled as K27 and K28, respectively.

Subsequent inverse, nested PCR using genomic DNA from 28 samples (K1-28) with the indicated oppositely-oriented primers in *attP39* and the neomycin resistance gene sequence (**Figure 3.18a**) resulted in unique PCR products in samples K1, K3, K4, K9 and K15, which were not present in the controls K27 and K28 (**Figure 3.18b**). The PCR product obtained with sample K15 was sequenced directly; however, those from K1, K3, K4 and K9 had to be re-amplified, as the amount of DNA obtained was not sufficient for sequencing (**Figure 3.18c**). Based on sequencing data, sample K15 had λ -Int-mediated targeting on chromosome 13 at a locus different from the previously targeted location on chromosome 13 (**Section 3.5.1**) whereas K1 and K9 revealed only random integration in chromosome 2 and 13, respectively (data not shown). The amplifications obtained in K1 and K9 were found to be PCR artifacts due to non-specific amplifications.

Note: The re-amplification of K3 and K4 was not successful (Figure 3.18c).

Figure 3.18 Identification of *attL* recombinant site on chromosome 13. (a) A schematic diagram of *pattP39-PGKss-neo* targeting vector after integration into a putative *attH* using Int-h/218. The genomic DNA with the integrated recombinant product is digested with restriction enzyme *PstI* followed by self-ligation. Positions of the relevant primers used for inverse nested PCR are indicated. (b) Inverse nested PCR screening for putative *attH* X *attP39* recombination resulted in unique PCR products for K1, K3, K4, K9, K15 samples (highlighted) which were not present in K27 and K28 (negative controls). (c) Re-amplification of PCR products obtained with K1, K3, K4 and K9 samples resulted in specific DNA bands in samples K1 and K9, but not for samples K3 and K4. Lanes: M1, 1kb ladder; M, 100bp ladder; K1-K22, genomic DNA from pooled samples; K23, K24, K25 and K26, bulk genomic DNA from plates #1.4, #2.4, #3.4, #4.4 respectively; K27, genomic DNA from *pCMV* transfected plate; K28, genomic DNA from parental HT1080 cells; W, water control

In order to identify the specific targeted clone in pooled sample K15, genomic DNA was extracted from 6 individual colonies (K85, K86, K87, K88, K89, and K90) and PCR amplification with specific primers (chr13 K88 fwd 1 & attP rev 1 and chr13 K88 rev 1 & attP fwd 1) was carried out to amplify *attL* and *attR* (**Figure 3.19a**). Clone K88 was found to be positive for the *attR* recombinant site (**Figure 3.19b**), whereas *attL* junction could not be confirmed (**Figure 3.19b**). Therefore, to determine the breakpoint (discontinuity) in the integrated product, the complete integrated vector was amplified by using various primers in the target vector (attP-pgk fwd, neo-fwd, attP fwd 1) along with chr13 K88 rev 1 (**Figure 3.19a**), which resulted in PCR products with the expected sizes for the *attR* junction (**Figure 3.19c**).

Figure 3.19 Determination of *attL* and *attR* sites on chromosome 13 in clone K88. (a) A schematic diagram of *pattP39-PGKss-neo* targeting vector after integration into *attH*. Positions of the relevant primers used for PCR amplification of *attL* and *attR* on chromosome 13 are indicated. (b) PCR analysis using genomic DNA from clone K88 with primers (chr13 K88 fwd 1 & attP rev 1 and chr13 K88 rev 1 & attP fwd 1) to amplify the *attL* and *attR* sites on chromosome 13, resulted in no specific PCR product for the *attL* site and a specific amplification of ~300bp for *attR* site. (c) Confirmatory PCR of the *attR* site using genomic DNA from clone K88 with primers in target vector (attP-pgk fwd, neo-fwd, attP fwd) along with primer chr13 K88 rev 1 (Figure 3.19a). Lanes. M, 100bp ladder; W, water control; HT, genomic DNA from parental HT1080 cells; K88, genomic DNA from clone K88; M1, 1kb ladder

In order to investigate the *attL* junction, an inverse PCR strategy was applied to *PstI* restricted and ligated genomic DNA of sample K15 and clone K88 with oppositely oriented primers as indicated (**Figure 3.20a**). The inverse PCR resulted in unique bands for both samples, which were absent from the negative controls (**Figure 3.20b**). The inverse PCR product obtained with clone K88 was sequenced and revealed recombination on chromosome 13. Curiously, *attL* was found to be about 70 million bp away from the *attR* site (**Figure 3.20c**). This arrangement was confirmed by performing a PCR with specific primers (**Figure 3.20c**), which amplified the whole target vector along with *attL* and *attR* sites on chromosome 13 (at two different loci) (**Figure 3.20d**).

Figure 3.20 Determination of the *attL* site in clone K88. (a) A schematic diagram of *pattP39-pgk-neo* targeting vector after integration into a putative *attH* site. The genomic DNA with the integrated recombinant product is digested with restriction enzyme *PstI* followed by self-ligation. Positions of relevant primers used in the inverse PCR are indicated. (b) Inverse PCR using genomic DNA from sample K15 and clone K88 resulted in a unique amplification in K15 and K88 (2), which was not present in genomic DNA from parental HT1080 cells. (c) Illustration of λ -Int-mediated targeting at chromosome 13 resulting in deletion of 70 million bp between *attL* and *attR* recombinant sites. (d) PCR analysis using genomic DNA from clone K88 with primers (chr13 K88 fwd 1 & amp rev) resulted in amplification of *attL* (~5kb) site and primers (chr 13 K88 fwd 1 & chr 13 K88 rev 1) resulted in amplification of the full vector including *attL* and *attR* recombinant sites (~8kb). Lanes. W, water control; HT, genomic DNA from HT1080 parental cells; K88, genomic DNA from clone K88; M1, 1 kb ladder; M, 100bp ladder; K15 (1), K15 (2), K15 (3); 200ng, 500ng, 1 μ g of genomic DNA of sample K15; K88 (1), K88 (2), K88 (3), 200ng, 500ng, 1 μ g of genomic DNA of clone K88

Since several targeting attempts with *attP39* variant did not result in any potential targetable *attH* sequence, therefore we next used the wild-type *attP* sequence along with different Int mutants to search for a putative *attH* sequence in the human genome via inverse PCR strategy.

3.6. Targeting endogenous *attH* sequences using wild-type *attP* along with different Int mutants (Int-h/218, Int-C2, Int-C3)

In another targeting attempt, a target vector *pattP(WT)-PGKss-neo* containing the wild-type *attP* sequence was used for targeting unknown endogenous *attH* sites along with the different Int mutants. In order to increase the concentration of Int in the cell nucleus, a nuclear localization signal (NLS) was added to the C-terminus of Int-h/218, Int-C2 and Int-C3. Int-INA (*pCMVssInt-Ina* expressing Int with an inactivating mutation (Y342A) where tyrosine at sequence position 342 is replaced by alanine) was used as a negative control.

HT1080 cells were split into six 10 cm plates, and the following co-transfections were done.

Plate#	<i>pattP(WT)-PGKss-neo</i>	Int expressing plasmids	Labeled as
1.5	50 ng	5 µg <i>pCMVssInt-h/218</i>	Int-h/218
2.5	50 ng	5 µg <i>pCMVssInt-h/218-CNLS</i>	Int-h/218-C
3.5	50 ng	5 µg <i>pCMVssInt-C3</i>	Int-C3
4.5	50 ng	5 µg <i>pCMVssInt-C3-CNLS</i>	Int-C3-C
5.5	50 ng	5 µg <i>pCMVssInt-C2</i>	Int-C2
6.5	50 ng	5 µg <i>pCMVssInt-Ina</i>	Int-INA

After two weeks of selection with 500µg/ml geneticin in the media, a total of 96 colonies (48 from plate # 1.5 and 48 from plate # 4.5) were picked. At the same time, bulk genomic DNA was extracted from all the six plates, and an inverse nested PCR was carried out with the primers as indicated (**Figure 3.22a**). A unique PCR product was obtained with genomic DNA extracted from plates # 1.5, # 2.5 and # 4.5 (Int-h/218, Int-h/218-C and Int-C3-C) (**Figure 3.22b**). The sequencing data obtained from PCR-amplified product derived from plate #4.5 (Int-C3-C), suggested λ-Int-mediated targeting

on chromosome 8 whereas those derived from plate #1.5 (Int-h/218) and plate 2.5 (Int-h/218-C) revealed random integration.

Figure 3.22 Targeting endogenous *attH* sequences using *attP* (WT). (a) Schematic diagram of *pattP(WT)*-*PGKss-neo* targeting vector after integration into putative *attH*. Genomic DNA with the integrated recombinant product was digested with restriction enzyme *PstI* followed by self-ligation. Relative positions of primers used for inverse nested PCR to amplify *attR* are indicated. (b) Inverse, nested PCR screening for *attH* X *attP39* recombination resulted in unique PCR products (highlighted) for Int-h/218, Int-h/218-C and Int-C3-C, not present in INA-Int and HT (negative controls). Lanes. M1, 1kb ladder; W, water control; HT (1,2), Primary and nested PCR products using genomic DNA from parental HT1080 cells; Int-h/218 (1,2), Primary and nested PCR using genomic DNA from plate #1.5, Int-h/218-C (1,2), Primary and nested PCR using genomic DNA from plate #2.5; Int-C3 (1,2), Primary and nested PCR using genomic DNA from plate #3.5; Int-C3-C (1,2), Primary and nested PCR using genomic DNA from plate #4.5; Int-C2 (1,2), Primary and nested PCR using genomic DNA from plate #5.5; Int-INA (1,2), Primary and nested PCR using genomic DNA from plate #6.5

In order to analyze the recombination at chromosome 8, specific genomic primers for this locus were designed. 48 colonies picked from plate #4.5 (Int-C3-C) were analyzed for recombination by pooling 6 colonies giving a total of 8 samples labeled as (C1-8). To amplify the *attL* and *attR* junction, genomic DNA was extracted from C1-C8 samples and PCR analysis was performed using the primers (chr 8 fwd 1 & neo rev) and (chr 8 rev 1

& attP fwd) (**Figure 3.23a**). PCR results depict the expected sizes for the *attL* (~2400bp) and the *attR* (~1050bp) junctions for samples C2 and C8 (**Figure 3.23b**).

Figure 3.23 Identification of *attH*-chr8. (a) Schematic diagram of *pattP(WT)*-*PGKss-neo* targeting vector after integration into *attH* (chr8). Positions of the relevant primers used for PCR amplification of recombinant *attL* and *attR* sites on chromosome 8 are indicated. (b) PCR analysis using genomic DNA extracted from samples C1, C2, C3, C4, C5, C6, C7, C8 with primers (chr 8 rev 1 & attP fwd 1) for the *attR* site and primers (chr 8 fwd 1 & neo rev 81) for the *attL* site, resulted in unique PCR products of sizes ~1000bp and ~2400 bp respectively for the C2 and C7 samples respectively, which was not present in genomic DNA of Int-INA and HT (negative controls). Lanes; M1, 1 kb ladder; M, 100bp ladder; C1, C2, C3, C4, C5, C6, C7, C8, genomic DNA extracted from C1-C8 samples; C3C, bulk genomic DNA from plate #4.5(Int-C3C); INA, bulk genomic DNA from plate #6.5 (Int-INA); HT, genomic DNA from parental HT1080 cells; W, water control

To identify the specific clone that would have undergone recombination, genomic DNA was extracted from individual colonies from sample C2 (C7-12) and C8 (C43-48). PCR results (**Figure 3.24**) and sequencing data confirmed Int-C3-mediated recombination in clones C7 and C43 at chromosome 8.

Figure 3.24 Confirmatory PCR for *attH*-chr8 targeting. PCR analysis using genomic DNA from individual clones from sample C2 (C7, C8, C9, C10, C11, C12) and sample C7 (C43, C44, C45, C46, C47, C48) with primers (chr8 fwd 1 & neo rev; Figure 3.23a) to amplify the *attL* junction and (chr8 rev 1 and attP fwd; Figure 3.23a) to amplify the *attR* junction resulted in unique PCR products for clones C7 and C43 for the *attL* site (~2400bp) (**upper panel**) and *attR* site (~1050bp) (**bottom panel**). Lanes; C7, C8, C9, C10, C11, C12, C43, C44, C45, C46, C47, C48, genomic DNA from colonies obtained from plate #4.5 (Int-C3C); Int-C3C, bulk genomic DNA from plate #4.5 (Int-C3C); Int-INA, bulk genomic DNA from Ina-Int transfected plate; M1, 1kb ladder; M, 100bp ladder; W, water control

Based on the sequencing data, a potential 21bp *attH* sequence was found in chromosome 8 and was labeled as *attH* (chr8). Unlike *attP*(WT), *attH*(chr8) sequence differed by 5 nucleotides in the core binding site (underlined) and one nucleotide in the overlap (in yellow), as shown below.

In order to increase the efficiency of targeting, a new target vector *pattP(chr8)-PGKss-neo* was constructed where the overlap of *attP* (WT) was adjusted to that of *attH*(chr8). The targeting was repeated with *pattP(chr8)-PGKss-neo* and *Int-C3CNLS*, and 30 colonies were picked and analyzed for recombination events. PCR analysis using chromosome 8 specific primers (chr 8 fwd 1 & neo rev 81) and (chr 8 rev 1 & attP fwd 1)

(Figure 3.23a) did not result in any positives, suggesting this targeting cannot be reproduced. Hence, we did not continue to pursue this putative *attH* site.

3.7. A functional human *attH* site

3.7.1 Identification of *attH4X* sequence

Since our initial attempts to identify functional and reproducibly targetable *attH* sequences in the human genome were not successful, we decided on a different approach. The human genome sequence (hs_ref-GRCh38) was searched for targets resembling cores of *att* sites. A sequence encompassing Int core binding sites from *attB* and *attP*, and a degenerate 7bp spacer was used as a query in this bioinformatics search (Figure 3.25). This resulted in 147 hits; of which a sequence termed *attH7X* that appeared seven times in the human genome was selected for further analysis (Figure 3.25) (contributed by Prof. Li Jinming and colleagues). In order to target *attH7X*, the 7bp spacer in the *attP* core site was adjusted to match the spacer in *attH7X*, keeping the core sites, Int arm type and accessory protein DNA binding sites of the *attP* sequence unchanged (not shown). This modified sequence was termed *attP4X*, which would serve as a potential recombination partner for the genomic *attH7X* site identified through bioinformatics (Figure 3.25).

Figure 3.25 Bioinformatics approach for identification of *attH* sequence. Diagram showing the 21 nucleotide sequences comprising the core binding and spacer sequences of various *att* sites including WT *attB* and *attP* sequences. The query sequence was used in a bioinformatics search for targets resembling core *att* sites (contributed by Li Jinming and colleagues)

A target vector containing *attP4X* and puromycin resistance gene driven by PGK promoter termed as *pattP4X-PGKss-puro* was generated to dock into the endogenous *attH7X* site. Furthermore, a codon-optimized version of *Inth-218* (termed *Opt-Int*) (sequence source; GeneScript) was cloned downstream of the CMV promoter generating *pCMVssInt-Opt* (**Section 2.2**; Materials and Methods).

About 3×10^6 HT1080 cells were split in five 10 cm dishes and the following co-transfections were carried out.

Plate#	<i>pattP4X</i> <i>-pgk-puro</i>	Int expressing plasmid	Labeled as
1.6	5 ng	100 ng <i>pCMVssInt-h/218-CNLS</i>	Int-h/218-C1
2.6	50 ng	500 ng <i>pCMVssInt-h/218-CNLS</i>	Int-h/218-C2
3.6	5 ng	100 ng <i>pCMVssInt-Opt</i>	Int-Opt-C1
4.6	50 ng	500 ng <i>pCMVssInt-Opt</i>	Int-Opt-C2
5.6	50 ng	500 ng <i>pCMVssInt-Ina</i>	Int-INA

Note: For the initial screening, Int-C3 was not used.

After 2 weeks of selection at 1µg/ml puromycin, bulk genomic DNA from co-transfected plates was extracted and restricted with *PstI* followed by self-ligation. An inverse, nested PCR was performed using primers as indicated (**Figure 3.26a**) which resulted in an expected PCR product size of more than 3kb (highlighted) in plates #1.6, 2.6 and 3.6 (*Int-h/218-C1*, *Int-h/218-C2* and *Opt-Int-C1*) (**Figure 3.26b**) (contributed by Dr Shree Harsha).

Figure 3.26 Inverse PCR for identification of *attH*. (a) Schematic diagram of *pAttP4X-PGKss-puro* targeting vector after integration into the putative *attH*. The genomic DNA with integrated recombinant product is digested with restriction enzyme *PstI* followed by self-ligation. Position of relevant primers used for inverse PCR are indicated. (b) Inverse PCR analysis using bulk genomic DNA from plates (# 1.6, 2.6, 3.6, 4.6, 5.6, 6.6) with primers indicated in figure (a) resulted in a unique PCR product (more than ~3kb) in plates #1.6, 2.6 and 3.6 (Int-h/218-C1, Int-h/218-C2 and Int-Opt C1). Lanes. M1;1kb ladder,W; water control, HT; genomic DNA from parental HT1080 cells, Int-INA; genomic DNA from plate #5.6, Int-h/218-C1; genomic DNA from plate #1.6, Int-h/218-C2; genomic DNA from plate #2.6, Int-Opt-C1; genomic DNA from plate #3.6, Int-Opt C2; genomic DNA from plate #4.6

Note: PCR product from plate # 1.6 had to be re-amplified, as the amount of DNA obtained was not sufficient for sequencing. However, re-amplification of Int-h/218-C1 was not successful.

The PCR product obtained with genomic DNA from plates #2.6 and #3.6 was sequenced, which confirmed λ -Int-mediated site-specific insertion in chromosome 14 and chromosome 3 respectively (**Figure 3.27b**), which was also confirmed with PCR using specific genomic primers (data not shown) (contributed by Dr Shree Harsha).

The targeting at chromosome 14 and chromosome 3 occurred in a 18bp sequence that was termed *attH4X*. This differed from *attH7X* in the three nucleotides from 5' end (**Figure 3.25**). This *attH4X* was found to be a part of open reading frame 1 (ORF1) of a long

interspersed element 1 (*LINE-1*) (**Figure 3.27**). ORF1 encodes an RNA-binding protein required for ribonucleoprotein assembly during *LINE-1* retro-transposition (Erwin, Marchetto et al. 2014).

Figure 3.27 An illustration of a *LINE-1* retrotransposon. A schematic diagram showing features of the *LINE-1* retro-transposon including the position and orientation of *attH4X* in ORF1 coding sequence and inverse orientation of RNA polymerase II promoters in the 5' untranslated region (UTR)

A BLAST with the *attH4X* sequence was performed with the human genome sequence, which resulted in 935 hits, all belonging to ORF1 of various *LINE-1* subfamilies scattered throughout the genome. These 935 *LINE-1* elements containing the 18bp *attH4X* sequence flanked by 300bp upstream and downstream sequences were aligned (a fraction of alignment is shown below in **Figure 3.28**, for complete alignment refer to appendix **Figure A1**) using Vector NTI, AlignX software. Based on sequence homology, consensus ORF1 primers (cs_attH4X_F1 & cs_attH4X_R1) and (cs_attH4X_F2 & cs_attH4X_R2, not shown) flanking the *attH4X* sequence were designed to enable PCR-screening of integration events (**Figure 3.28**).

Figure 3.28 Alignment of *attH4X* sequence. Alignment of *LINE-1* elements (only a fraction is shown here) containing the 18bp *attH4X* site flanked by 300bp upstream and downstream sequences to design primer pairs (cs_attH4X_F1 & cs_attH4X_R1) and (cs_attH4X_F2 & cs_attH4X_R2, not shown) in the consensus flanking region of attH4X. The respective subfamilies of 935 *LINE-1* elements that are shown in the right column correspond to the serial numbers of the sequences on the left side of the alignment

3.7.2 Targeting of *attH4X* in *LINE-1* elements

For *attH4X* targeting, about 3×10^6 HT1080 cells were co-transfected with *pattP4X-PGKss-puro* along with Int-C3CNLS. 33 puromycin resistant colonies were picked and analyzed by PCR using ORF1 primers (cs_attH4X_F1/ cs_attH4X_F2 & puro rev 24) for the *attL* junction and (cs_attH4X_R1/ cs_attH4X_R2 & P_21) for the *attR* junction as indicated (**Figure 3.29a**). Using our designed ORF1 screening primers in combination with target vector primers, 4 out of 33 puromycin-resistant HT1080 clones identified via PCR/DNA sequencing revealed at least one correct recombination junction (clones 3, 19, 21 for the *attL*) and (clones 11, 19, 21 for the *attR*) (**Figure 3.29b and 3.29c**). The positive control (+) is the bulk genomic DNA from the plate #3.6 (Int-Opt-C1)(**Figure 3.26b**) (contributed by Dr Shree Harsha).

Based on the genomic location of the corresponding four *LINE-1* elements as determined from the sequencing data, specific genomic primers were designed, and subsequent sequencing of PCR products confirmed integration of full-length *pattP4X-PGKssPuro* in

attH4X in all the clones (data not shown). Importantly, both recombination junction sequences revealed precise vector integration as expected for *Int*-mediated catalysis, i.e. without nucleotide additions or deletions (contributed by Dr Shree Harsha).

Figure 3.29 Screening for *attH4X* X *attP4X* recombination events in HT1080 clones. (a) A schematic diagram of *pattP4X-pgk-puro* target vector after integration into *attH4X*. Positions of relevant primers, restriction sites and probe used for Southern blotting are indicated. (b) PCR was performed with the

genomic DNA of all the 33 colonies using primers cs_attH4X_F1 and puro rev 24 (for the *attL* junction) and cs_attH4X_R1 & P_21 (for the *attR* junction). PCR amplified products of the expected size (1100bp; for the *attL*) were detected in clones 3, 19 and 21. (c) PCR amplified products of expected size (~375bp; for the *attR*) were detected in clones 11, 19 and 21. W, water control; HT, genomic DNA from parental HT1080 cells; Ina, genomic DNA from puromycin-resistant clones obtained through co-transfection of *pattP4X-PGKssPuro* and *Int-Ina*; +, positive control, bulk genomic DNA from plate #3.6 (refer to **Figure 3.26b**); M, 100 bp DNA ladder; 1 to 33, genomic DNA from puromycin-resistant HT1080 clones obtained through co-transfection of *pattP4X-PGKss-puro* and *Int-C3CNLS*

Genomic DNA from #3, #11 #19, and #21 positive clones was extracted and restricted with *PstI* and *XbaI* separately (**Figure 3.30**). Southern blotting with an internal puromycin probe (**Figure 3.30**) revealed that clones #19 and #21 showed single-copy transgene integration whereas clones #3 and #11 had extra bands (as shown by arrow in **Figure 3.30**) which could be because of an additional integration event or the clones were not entirely clonal. The genomic restriction pattern of clone #19 matched the predicted intergenic location of the targeted *LINE-1* element on human chromosome 2 (data not shown) (contributed by Dr Shree Harsha).

Figure 3.30 Southern blot analysis of targeted HT1080 clones. Genomic DNA purified from targeted HT1080 clones, as indicated, and parental HT1080 cell line was subjected to digestion with *Pst*I or *Xba*I. A PCR-derived digoxigenin-labelled probe complementary to the puromycin resistance gene was used (Figure 3.29a). Lanes: M1, 1kb DNA ladder; HT, genomic DNA from HT1080; 3, 11, B18, 19 and 21, genomic DNA from targeted clones; pP4X (10⁷, 10⁸, 10⁹), copies of linearized *pattP4X-PGKss-puro* loaded as positive control. The arrows indicate fragments of expected size for clones 3 and 11. *Clone B18 (carrying a single copy transgene) was obtained from a screen of puromycin-resistant HT1080 clones that resulted from a previous co-transfection of *pattP4X-PGKss-puro* and *Int-h/218*

Validation of *attH4X*-targeting was also performed with HeLa and human immortalized keratinocytes (NEB-1 cells). For HeLa cells, PCR/sequencing and Southern blotting, confirmed 5/44 clones positive for targeting and single-copy integration was confirmed in 4/5 clones (contributed by Dr Shree Harsha and Sabrina Peter, data not shown). Genomic restriction fragments from 3 clones matched the predicted locations (data not shown). With NEB-1 cells, 2/26 clones showed *attH4X*-targeting, and single-copy integration was detected in both cases with genomic fragments also matching the predicted loci (data not shown) (contributed by Dr Shree Harsha and Sabrina Peter). Together, these data

revealed that Int-C3CNLS-mediated site-specific transgene addition to *LINE-1* elements occurred in about 10% of analyzed clones, with more than half carrying single-copy transgenes.

Note: Site-specific integration into the *attH4X* could also be achieved using transfected mRNA to express Int-C3CNLS, thus eliminating the possibility of random background integration events arising due to Int expressing plasmids (data not shown).

3.7.3 Targeting of the *attH4X* in hESCs

In order to test the *attH4X*-targeting in human embryonic stem cells (hESCs), employed human Undifferentiated Transcription Factor 1 (*UTF1*) gene-based pluripotency reporter cassette. Expression of this reporter is under the control of pluripotency factors OCT4, SOX2 and, most likely, NANOG (**Figure 3.31**) (Tan, Wang et al. 2007, Morshedi, Soroush Noghabi et al. 2013). Since the expression of UTF1 is known to be one of the most reliable indicators of human and mouse ESCs pluripotency (Buganim, Faddah et al. 2012), target vector *pattP4X-PGKss-puro-UTF1-EGFP* was constructed (**Figure 3.31**). It consists of the *attP4X* site, EGFP reporter gene controlled by UTF1 promoter/enhancer elements, a puromycin gene driven by PGK promoter and bacterial sequences (**Figure 3.31**).

First, we found that transient Int-C3CNLS expression from *pCMVssInt-C3CNLS* is strongly attenuated in hESCs (data not shown) (Wen, Zhang et al. 2014). Therefore, the EF1 α promoter was employed to drive expression of Int from *pEF1 α -ssInt-C3CNLS* for the following targeting experiments.

Figure 3.31 Targeting of *attH4X* in hESCs. A schematic diagram of *pattP4X-PGKss-puro-UTF1-EGFP* recombination with *attH4X*. Positions of the relevant primers (*cs_attH4X_F1*, *cs_attH4X_F2*, *cs_attH4X_R1*, *cs_attH4X_R2*, *puro rev 24*, *PGK rev*, *pr22* and *pr111*) are indicated. Magnified image of UTF1-EGFP cassette showing the *EGFP* gene under the control of UTF1 promoter/ enhancer elements. The black lines indicate the binding sites of OCT4, SOX2 and putative NANOG transcription factor

About 3 x 10⁶ hESCs per 10 cm dish were seeded 24 h prior transfection and the following co-transfections were carried out.

Plate#	<i>pattP4X-pgk-puro-UTF1-EGFP</i>	Int expressing plasmid
1	1 µg	1 µg <i>pEF1α-ssInt-C3CNLS</i>
2	1 µg	1 µg <i>pEF1α-ssInt-C3CNLS</i>

After 2 weeks of puromycin selection, 120 hESC colonies were screened for the *attL* recombinant site by PCR/sequencing using the primers (*cs_attH4x_F1/F2* & *puro rev 24/PGK rev*) (**Figure 3.32a**). 17 clones (A8, A17, G2, G4, A3, A20, D7, K1, K3, K4, L2, M5, M9, B2, B16, E3, J4) were found to be positive for the *attL* junction, indicating successful *attH4X*-targeting (**Figure 3.32b**).

Figure 3.32 Screening for *attH4X* X *attP4X* recombination events in hESC clones. (a) Schematic diagram of *pattP4X-PGKss-puro-UTR1-EGFP* target vector after integration into *attH4X*. Positions of the relevant primers, restriction sites and probe used for Southern blotting are indicated. (b) Semi-nested PCR with primers *cs_attH4X_F1* and *PGK rev* (for the *attL* site) using templates obtained with primary PCR (primers *cs_attH4X_F1* and *Puro rev24*) resulted in PCR products of the expected size (884bp) in genomic DNA of 17 hESC clones (A8, A17, G2, G4, A3, A20, D7, K1, K3, K4, L2, M5, M9, B2, B16, E3, J4). W, no DNA template control; ES, negative control (genomic DNA from parental HT1080 cells); M, 100 bp, DNA ladder; M1, 1 kb DNA ladder; A8, A17, G2, G4, A3, A20, D7, K1, K3, K4, L2, M5, M9, B2, B16, E3, J4, genomic DNA from the puromycin-resistant and GFP⁺ hESC clones obtained through co-transfection of *pattP4X-PGKss-puro-UTR1-EGFP* and *pEF1α-ss-Int-C3CNLS*

To verify the presence of the *attL* junction, semi-nested PCR with genome-specific forward primers and PGK rev (indicated in **Figure 3.32a**) using templates from a primary PCR (with Puro rev 24 and genomic locus-clone specific forward primers; **Figure 3.32a**) was performed, which confirmed the *attL* junctions in genomic DNA of 8/17 clones (A3, A20, E3, K3, K4, L2, M5, M9) (**Figure 3.33**).

Figure 3.33 Verification of *attL* recombinant site in hESC clones. A semi-nested PCR with genomic locus-clone specific forward primers and PGK rev using templates from a primary PCR (primers Puro rev 24 and genomic locus-clone specific forward primer) confirmed *attL* junction in genomic DNA of 8/17 clones (A3, A20, E3, K3, K4, L2, M5, M9). Lanes. W, water control; M1, 1kb ladder; M, 100bp ladder; ES, genomic DNA from parental ES cells; A3, A20, E3, K3, K4, L2, M5, M9; genomic DNA from puromycin-resistant and GFP positive hESC clones obtained through co-transfection of *pattP4X-PGKss-puro-UTF1-EGFP* and *pEF1 α -ss-Int-C3CNLS*

The *attR* junction was confirmed in genomic DNA of 6/17 clones (K4, A20, A3, K3, E3, K1) in a semi-nested PCR with pr111 forward and genome-specific reverse primers using templates from a primary PCR (with pr22 forward and genomic locus-clone specific reverse primers) (**Figure 3.32a** and **Figure 3.34**).

Figure 3.34 Verification of *attR* recombinant site in hESC clones. A semi-nested PCR with primer in EGFP gene (pr111) and genomic locus-clone specific reverse primers using templates from a primary PCR (primer pr22 and genomic locus-clone specific reverse primer) confirmed *attR* junction in genomic DNA of 6/17 (K4, A20, A3, K3, E3, K1) clones. Lanes. W, water control; M1, 1kb ladder; M, 100bp ladder; ES, genomic DNA from parental ES cells; K4, A20, A3, K3, E3, K1, genomic DNA from puromycin-resistant and GFP⁺ hESC clones obtained through co-transfection of *pattP4X-pgk-puro-UTF1-EGFP* and *pEF1 α -ss-Int-C3CNLS*

Note: Two genomic locus-clone specific reverse primers were designed to detect *attR* junction out of which 6/17 clones were positive with either one of the reverse primers (Figure 3.34).

Both the junctions were confirmed for 5 clones (A3, E3, K3, K4, A20) out of which A3, E3 and K3 were chosen because of the intergenic location of transgenes and the level of EGFP expression as a payload. The genomic DNA from A3, E3 and K3 was extracted and restricted with either *NdeI* or *XbaI*. Southern blotting with the internal EGFP probe (Figure 3.32a) revealed that each carried a single-copy transgene construct (Figure 3.35), with clone E3 matching the predicted genomic fragment sizes (Table 3.1).

Figure 3.35 Southern blot analysis of targeted hESC clones. Genomic DNA from three hESC clones (A3.E3, K3) and parental hESC cells were digested with *NdeI* or *XbaI*. A digoxigenin-labelled probe to EGFP was employed (Figure 3.33a). Lanes: M1, 1kb DNA ladder; 10⁸, 10⁹, copies of linearized targeting vector loaded as positive control; ES, parental DNA; A3, E3 and K3, genomic DNA from targeted hESC clones

Table 3.1 Summary of Southern analysis for the targeted clones in hESC (A3, E3, K3) using *pattP4X-PGKssPuro-UTF1-EGFP* and *Int-C3CNLS*

CLONES	<i>attL</i> junction		<i>attR</i> junction	
	Expected Fragment size (bp)	Observed Fragment size (kb)	Expected Fragment size (bp)	Observed Fragment size (kb)
	Restricted with <i>NdeI</i>		Restricted with <i>XbaI</i>	
hESC-A3	10500	~8	6600	~10.5
hESC-E3	10500	~10.5	7400	~7.4
hESC-K3	10300	~5.8	6100	~5.7

3.7.4. Functionality of targeted *attH4X* loci in hESCs

The *UTF1*-driven *EGFP* expression pattern for each of the three hESC clones (A3, E3 and K3) was determined by cell imaging and quantified by FACS. It was observed that clone E3 was more homogenous, whereas clones A3 and K3 were heterogeneous in the *EGFP* expression (**Figure 3.36 a, b**). The undifferentiated clones A3, E3 and K3 comprised of 49%, 89% and 27% of fraction of *EGFP*⁺ cells at 3 weeks whereas 45%, 92% and 28% at 8 weeks of culturing respectively, (**Figure 3.36 a, b**) suggesting stable functional integration of the transgene cassette. Importantly, the fraction of *EGFP*⁺ cells was substantially reduced to 0.5% as a result of RA-induced differentiation in all three clones (**Figure 3.36 a, b**). The loss of *EGFP* expression thus demonstrated the sustained functionality of the *UTF1* reporter cassette integrated at the respective *LINE-1* loci at two different cell states, i.e. pluripotent and differentiated. This was further confirmed by western blotting of pluripotency markers (Oct4, Sox2 and Nanog) for some of the targeted hESC clones (See appendix Figure A2), however *in vivo* teratoma assay for pluripotency was not carried out.

Figure 3.36 Functionality of targeted *attH4X* loci in hESC clones. (a) Functional test for *UTF1* promoter-driven EGFP expression in targeted hESC clones. Fluorescence microscopic analysis of undifferentiated and RA-induced, differentiated parental hES cells and clones A3, E3 and K3. EGFP expression was detected with the undifferentiated hESC clones A3, E3 and K3 (column 2, panels 2, 3 and 4) but disappeared in differentiated progenies (column 4, panels 2, 3 and 4) respectively. Panels in columns 1 and 3 are phase-contrast light micrographs of undifferentiated and differentiated cells, respectively. Magnification 5x; Scale bars 100 μ m. (b) FACS analysis for undifferentiated and differentiated hESCs clones Dot plots representing GFP⁺ cells (upper right quadrant) and GFP⁻ cells (lower right quadrant) for the untargeted hESCs, undifferentiated targeted hESC clones (A3, E3, K3) after 3 weeks (early) and 8 weeks (late) of culturing the cells (left and middle panel) and their differentiated progenies (right panel).

Note: The relatively weak EGFP expression can be attributed to the intrinsically weak *UTF1* promoter (Tan, Wang et al. 2007).

3.8 Targeting of *LINE-1* using *attL4X* and Int-C3CNLS

In order to explore the possibility of elimination of undesired bacterial components in the integration event, a target vector *pTZ18R-attL4X/attR4X-pgksspuro-UTF1-EGFP* was designed. In this context, the corresponding adjusted *attL/R* spacer variants (to match the spacer of *attH4X*), termed *attL4X* and *attR4X*, were cloned as direct repeats into a target vector. Upon *ex vivo* expression of Int-C3, intra-molecular recombination between *attL4X* X *attR4X* would result in generation of a seamless vector containing only the

attP4X recombination sequence and other desired functional sequences (**Figure 3.37**).

Figure 3.37 Generation of seamless vector. A schematic diagram showing intra-molecular recombination between *attL4X* X *attR4X* in *pTZ18R-attL4X/attR4X-pgksspuro-UTF1-EGFP* vector leading to formation of seamless vector *attP4X-pgksspuro-UTF1-EGFP*, eliminating the undesired bacterial sequences

To achieve this, 100ng of target vector *pTZ18R-attL4X/attR4X-pgksspuro-UTF1-EGFP* was co-transfected with 1µg of Int-C3CNLS and Int-INA in two 10 cm dishes of HT1080, respectively. After 2 weeks of puromycin selection, 60 colonies were picked for subsequent analysis. Targeting of *attH4X* was confirmed in genomic DNA of 8/60 clones in a semi-nested PCR with *cs_attH4X_F1* and *Puro fw318* primers using templates from a primary PCR (with primers *cs_attH4X_F1* and *Puro 127*) (**Figure 3.38a**) PCR-amplifications (~850 bp) were obtained for the clones 6, 19, 26, 29, 31, 32, 40 and 58 (**Figure 3.38b**) and products were sequenced. Surprisingly, sequencing analysis of the positive clones revealed that instead of *attP4X* being targeted in the form of seamless vector, *attL4X* had recombined with *attH4X* site in the genome, indicating that *attL4X* can functionally replace *attP4X* for targeting.

Figure 3.38 Targeting of genomic *attH4X* using *attL4X*. (a) A schematic diagram of *pTZ18R-attL4X/attR4X-PGKssPuroEGFP-UTF1* targeting vector after integration into *attH4X*. Relative binding positions of the primers (Puro fw318, Puro rev127, pr27, pr199, pr99 and pr49, cs_attH4X_F1 and cs_attH4X_R1) are indicated (b) Screening for *attL4X/attR4X* x *attH4X* recombination events. Eight clones positive for *attL4X* x *attH4X* recombination events at *LINE-1* loci were identified in an initial screen involving semi-nested PCR with primers cs_attH4X_F1 and Puro fw318 using templates from a primary PCR (primers cs_attH4X_F1 and Puro 127). PCR-amplifications (~850 bp) were detected for clones (6, 19, 26, 29, 31, 32, 40, 58). Lanes. W, no DNA template control; HT, negative control (genomic DNA from parental HT1080 cells); M, 100bp ladder; M1, 1 kb DNA ladder

In order to verify the presence of unrecombined *attR4X*, PCR was performed using genomic DNA from all the 8 clones with primers pr27 and pr199 (Figure 3.38a). PCR products (~1000bp) were detected in all the clones (Figure 3.39a), indicating that *attR* site was intact and did not undergo any recombination. Therefore, to confirm that the regenerated *attH4X* (right junction) has been generated due to recombination of *attL4X* with the genomic *attH4X*, a semi-nested PCR with primers cs_attH4X_R1 and pr49 using samples from a primary PCR (primers cs_attH4X_R1 and pr99; Figure 3.38a) was carried

out which resulted in expected PCR product size (~400bp) in 5/8 clones 19, 26, 29, 32, 40 (Figure 3.39b). Therefore, both *attL4X* and recombinant *attH4X* were confirmed in a total of 5/8 clones for subsequent analysis.

Figure 3.39 PCR screening for unrecombined *attR4X*. (a) PCR products (~1000bp) were detected with genomic DNA of clones (6, 19, 26, 29, 31, 32, 40, 58) using primers pr27 and pr199 (Figure 3.38a). Lanes. W, no DNA template control; HT, negative control (genomic DNA from parental HT1080 cells); M, 100bp ladder; M1, 1 kb DNA ladder. (b) PCR screening of recombinant *attH4X* junction (generated from *attL4X* x *attH4X*). Semi-nested PCR with primers cs_attH4X_R1 and pr49 using samples from a primary PCR (primers cs_attH4X_R1 and pr99; Figure 3.38a) resulted in PCR products of the expected size (~400 bp) with clones 19, 26, 29, 32, and 40. Lanes. W, water control; HT, genomic DNA from parental HT1080 cells (negative control); M, 100bp ladder; M1, 1 kb DNA ladder

Genomic DNA from 5 positive clones (19, 26, 29, 32, and 40) was extracted and restricted with *Bgl*III (Figure 3.38a).

Southern blotting with an internal EGFP probe (Figure 3.38a) confirmed single copy

integration in all the clones (**Figure 3.40**). The genomic restriction pattern of clone 19 and 32 matched the predicted intergenic location of the targeted *LINE-1* element (**Table 3.2**).

Figure 3.40 Southern blot analysis of *attL4X X attH4X* targeted HT1080 clones. Genomic DNA purified from the targeted HT1080 clones and parental HT1080 cells were subjected to digestion with *Bgl*III. PCR-derived Digoxigenin-labelled probe to the EGFP gene was employed. Lanes: 1kb DNA ladder; HT1080, parental DNA; 19, 26, 29, 31, 32 genomic DNA from targeted HT1080 clones; 10⁸, 10⁹ copies of linearized targeting vector as positive control

Table 3.2 Summary of Southern analysis for the targeted clones in HT1080 (19, 26, 29, 31, 32) using *pTZ18R-attL4X/attR4X-PGKssPuro-UTF1-EGFP* and Int-C3CNLS

CLONES	attR junction		Copy number
	Expected Fragment size (bp)	Observed Fragment size (kb)	
	Restricted with <i>Bgl</i> III		
HT1080-19	6600	~6.6	1
HT1080-26	11000	~7.5	1
HT1080-29	15000	~7	1
HT1080-31	8800	~6.3	1
HT1080-32	17000	>15	1

Interestingly, our data implied that intra-molecular *attL4X* x *attR4X* recombination by Int-C3CNLS appeared to be strongly disfavored over inter-molecular *attL4X* X *attH4X* recombination and that the arrangement of Int arm binding sites on the target vector seems to be important for functionality.

3.9. Summary of the targeting events in hESCs (*attH4X* X *attP4X*) and HT1080 cells (*attH4X* X *attL4X*)

Targeting of *attH4X* sequence in hESC clones using target vector (*pattP4X-PGKssPuro-UTF1-EGFP target*) and HT1080 clones using target vector (*pTZ18R-attL4X/attR4X-PGKssPuro-UTF1-EGFP*) has been summarized below (**Table 3.3**). In addition, genomic locations of targeted *attH4X*, along with the length, subfamily and retro-transposition status of corresponding *LINE-1* elements are indicated (**Table 3.3**).

Table 3.3 Summary of the targeting events at *attH4X* (in red) resulting in recombinant *attL* and *attR* sites (in green) in hESCs and HT1080 cells

Targeted Clones	Targeted <i>attH4X</i> sequence and resulting <i>attL</i> and <i>attR</i> sequences	Genomic location of targeted <i>attH4X</i>	Length and subfamily of targeted <i>LINE-1</i> (bp)	Retro-transposition functional status*
hESC A3	TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTGATCTTCAAAC CTGATATCC TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAAAC CTGATATCC	chrX:64787931-64787952	1916 LIPA7	Truncated
hESC E3	TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTGATCTTCAATCTCTTATATCC TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAATCTCTTATATCC	chr1:213884581-213884602	6423 LIPA7	Full length non intact
hESC K4	TTTTCTCTAAACTTGTCTTCATGCT TTATTTCA AGTTGATCTTCAATCTCTCGTATCC TTTTCTCTAAACTTGTCTTCATGCT TTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAATCTCTCGTATCC	chr13:46195229-46195250	6439 LIPA7	Full length non intact
hESC A20	TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTGATCTTCAATCTTGTAGGCC TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAATCTTGTAGGCC	chr5:72103592-72103613	6302 LIPA7	Full length non intact
hESC K1	TTTTCTCTAATCTTGGCTTCACG CTTTATTTCA AGTTGATCTTCAGTCTCTGATATCC TTTTCTCTAATCTTGGCTTCACG CTTTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAGTCTCTGATATCC	chr7:133653822-133653843	6485 LIPA7	Full length non intact
hESC L2	TCCTCTCTAATCTTGCTTCACG CTTTATTTCA AGTTGATCTTCAAAC CTGATATCC TCCTCTCTAATCTTGCTTCACG CTTTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAAAC CTGATATCC	chr1:84910263-84910284	6472 LIPA7	Full length non intact
hESC B16	TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTGATCTTCAGTCTCTGATATCC TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAGTCTCTGATATCC	chr3:111164680-111164701	6423 LIPA7	Full length non intact
hESC K6	TTTTCTCTAATCTTGGCTTCACG CTTTATTTCA AGTTGATCTTCAGTCTCTGATATCC TTTTCTCTAATCTTGGCTTCACG CTTTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAGTCTCTGATATCC	chr7:133653822-133653843	6485 LIPA7	Full length non intact
hESC L1	TTTTCTCTAATCTTGGCTTCACG CTTTATTTCA AGTTGATCTTCAGTCTCTGATATCC TTTTCTCTAATCTTGGCTTCACG CTTTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAGTCTCTGATATCC	chr7:133653822-133653843	6485 LIPA7	Full length non intact
hESC M5	TTTTCTCTAAACTCTTTTC CGCTTTATTTCA ATTGATCTTCAATCACTAATACGCT TTTTCTCTAAACTCTTTTC CGCTTTATTTCA AGTTG//CAGCTTTATTTCA ATTGATCTTCAATCACTAATACGCT	chr1:76244691-76244712(-)	6151 LIPA4	Full length non intact
hESC M9	TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTGGTCTCCATCTCTGATATCC TTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTG//CAGCTTTATTTCA AGTTGGTCTCCATCTCTGATATCC	chr7:80319732-80319753(-)	6371 LIPA7	Full length non intact
hESC A8	TTTTCTCCAATCTTGCTTCATGCT TTATTTCA AGTTGATCTTCAATCTCTGTTATCC TTTTCTCCAATCTTGCTTCATGCT TTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAATCTCTGTTATCC	chr5:9938285-9938306(-)	1029 LIP1	Truncated
hESC A17	TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTGATCTTCAATCTCTGATATCC TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAATCTCTGATATCC	chr1:92719737-92719758(-)	5690 LIPA7	Truncated
hESC G4	TTTTATCTAATCTTGCTTCATGCT TTATTTCA AGTTGATCTTCAATCTCTGATATCC TTTTATCTAATCTTGCTTCATGCT TTATTTCA AGTTG//CAGCTTTATTTCA AGTTGATCTTCAATCTCTGATATCC	chrX:38949476-38949497(-)	2162 LIPA7	Truncated
HT1080 -19	TTTTCTCTAATCTTGCTTCATGCT TTATTTCA AGTTGATCTTCAATCTCTGATATCC TTTTCTCTAATCTTGCTTCATGCT TTATTTCA AGTTG//CTGCTTTATTTCA AGTTGATCTTCAATCTCTGATATCC	chr2:150835444-150835464(-)	5080 LIPA7	Full length non intact
HT1080 26	TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTGATCTTCAGTCTCATATCTTT TTTTCTCTAATCTTGCTTCACAC TTTATTTCA AGTTG//CTGCTTTATTTCA AGTTGATCTTCAGTCTCATATCTTT	chr21:19146794-19146814	6451 LIPA7	Full length non intact
HT108029	TTTTCTCTAATCTTGCTTCACG CTTTATTTCA AGTTGATCTTCAATCTCTGATATCC TTTTCTCTAATCTTGCTTCACG CTTTATTTCA AGTTG//CTGCTTTATTTCA AGTTGATCTTCAATCTCTGATATCC	chr12:2050403-2050423	984 LIPA8A	Truncated
HT1080 -31	TTTCCTTGATCTTGCTTCACG CTTTATTTCA AGTTGATCTTCAATCTCTGATATCC TTTCCTTGATCTTGCTTCACG CTTTATTTCA AGTTG//CTGCTTTATTTCA AGTTGATCTTCAATCTCTGATATCC	chr4:142912896-142912916	5600 LIPA7	Full length non intact
HT1080 32	TTTTTTCTCTAATCTTGCTTC CGCTTTATTTCA TAAGCTGATCTTCAATCTCTGTTATC TTTTTTCTCTAATCTTGCTTC CGCTTTATTTCA TAAGTTG//CTGCTTTATTTCA TAAGCTGATCTTCAATCTCTGTTATC	chr20:33750784-33750805(-)	2947 LIPA7	Truncated

3.10 *In vitro* intra-molecular recombination for generation of seamless vector

The attempt of generating seamless vector failed as it resulted in genomic *attL4X* x *attH4X* recombination, therefore an alternative approach of *in vitro* intra-molecular recombination was used to generate supercoiled seamless vectors.

In this context, as a proof of principle, vector *pTZ18R-attL/attR-pgksspuro-UTF1-EGFP* was used such that after reaction with Int, intra-molecular *attL* X *attR* recombination will yield a DNA catenane containing two rings (Figure 3.41). One of the two-catenane rings comprises a seamless vector with the *attP* and the other ring consists of the *attB* along with bacterial components (Figure 3.41). The *attB* containing ring can be cleaved using a restriction enzyme whereas the remaining seamless, supercoiled vector can be gel purified using a simple affinity chromatography (Qiagen Gel Extraction Kit).

Figure 3.41 *In vitro* intra-molecular recombination for generation of seamless vector A schematic diagram showing intra-molecular recombination between *attL* X *attR* in *pTZ18R-attL/attR-PGKsspuro-UTF1-EGFP* target vector leading to formation of seamless vector *attP-PGKsspuro-UTF1-EGFP*, eliminating the undesired bacterial sequences

To carry out this proof of principle, firstly, Int-h/218 and Int-C3 were obtained in purified form (Siau, Chee et al. 2015) and an Int buffer composed of 50mM Tris-HCl, 50mM NaCl, 0.25mM EDTA, 2.5mM spermidine, 0.2mg/ml BSA was prepared (Siau, Chee et al. 2015). *In vitro* reactions with 5 µg or 2 µg of *pTZ18R-attL/attR-pgksspuro-UTF1-EGFP* and different amounts of Int-C3 (1.6ug, 1ug, 0.5ug and 0.2ug) using the Int buffer and glycerol was set up. After intra-molecular recombination, a 4.2kb seamless vector was obtained, which being in a supercoiled form migrated at about 3kb (Figure 3.42).

The other catenane ring (containing bacterial components) is cleaved twice by *ApaLI*, resulting in 1250bp and 1700bp bands (Figure 3.41 and 3.42). The control reaction comprising of *pTZ18R-attL/attR-pgksspuro-UTF1-EGFP* without Int resulted in 5869bp and 1250bp bands whereas the partially digested, hence linearized plasmid resulted in approximately 7kb band (Figure 3.42). The 3kb supercoiled seamless vector obtained was gel purified but the yield was quite low in order to carry out transfection for *attH4X* targeting in the genome. Therefore, the production of seamless vector needs to be optimized together with the purification of Int protein in a more active form. Eventually, the purified Int protein in complex with the seamless vector can be electroporated in the form of a lambdasome, thus avoiding the random integration of the Int expressing plasmid as well as unwanted bacterial sequences in the genome.

Figure 3.42 Restriction check after *in vitro* intra-molecular reaction. Restriction digest after *in vitro* intra-molecular reaction in *pTZ18R-attL/attR-pgksspuro-UTF1-EGFP* resulted in ~3kb recombined supercoiled seamless vector, 1.7kb and 1.2kb *ApaLI* cleaved products from *attB* containing ring (Figure 3.41). The control reaction comprising of *pTZ18R-attL/attR-pgksspuro-UTF1-EGFP* without Int resulted in 5869bp and 1250bp bands whereas the partially digested linearized plasmid resulted in approximately 7kb band. Lanes. 1.6ug, 1ug, 0.5ug, 0.2ug (different amount of Int) with 5ug and 2ug of substrate DNA plasmid; LR, positive control using LR clonase enzyme (Gateway Cloning Kit) with 2ug of substrate plasmid; C, control reaction (2ug substrate plasmid without Int); M1, 1kb ladder

3.11 Safety evaluation upon λ -Int-mediated targeting

3.11.1 Cell proliferation assay

Since many tyrosine recombinases such as Cre are known to cause cell toxicity and DNA damage when expressed at high levels in human cells (Silver and Livingston 2001), it was important to determine whether the expression of Int-C3-CNLS is safe in human cells with respect to cell toxicity and genome integrity.

In this context, an MTT assay (a colorimetric test) for assessing metabolic activity of cells was performed. The key component in this assay is {3-(4-(5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide}, or MTT, which is yellowish in colour. Mitochondrial dehydrogenases enzymes of viable cells cleave the tetrazolium ring of MTT converting it into purple color crystals that can be dissolved in isopropanol. The amount of purple color crystals formed varies with the increase or decrease in the number of viable cells and can be measured spectrophotometrically, indicating the degree of cytotoxicity.

Based on this principle, cell viability was analysed in Int expressing cells. Two 10 cm dishes of HT1080 cells were co-transfected with 1 μ g of *pCMV-EGFP* + 1 μ g of *pCMVssInt-Ina* (as negative control for an inactive Int) and 1 μ g of *pCMV-EGFP* + 1 μ g of *pCMVssInt-C3CNLS* (expressing Int-C3CNLS). 48 h post-transfection EGFP⁺ cells from each plate were sorted using FACS Aria and plated in triplicates, at a density of 5×10^3 per well in 24 well plates and allowed to attach overnight. Un-transfected HT1080 cells were included as control. Cell proliferation rates were measured every day for 9 days using the MTT-based In Vitro Toxicology Assay Kit (Sigma-Aldrich) as per manufacturer's protocol. The results showed that there was no impact on the cell viability between the parental HT1080 cells, Int-Ina and Int-C3CNLS transfected cells over the period of 9 days (**Figure 3.43**).

Figure 3.43 Cell proliferation assay. A bar chart showing MTT-based cell proliferation assays that was performed to assess effects on cell proliferation rates upon Int expression in human cell lines over the period of 9 days. HT1080 cells un-transfected (HT1080), and FACS sorted GFP+ cells obtained after co-transfection of *pCMV-EGFP* with either *pCMVssInt-Ina* or *pCMVssInt-C3CNLS* (Int-C3CNLS) were analyzed for the effect on cellular proliferation using the colorimetric MTT assays over the indicated time course. Data shows the mean of triplicates and standard deviation of a representative experiment, n=2

3.11.2. DNA damage response

To investigate if Int induces DNA damage at expression levels suitable for genome editing, a Western blot analysis was done to determine phospho- γ -H2AX (DNA damage repair protein) levels as a marker to assess DNA damage induced by expression of Int in HT1080 cells at 48 h and 72h post-transfection (performed and analyzed by Dr Shree Harsha; data not shown). There was no detectable induction of phospho- γ -H2AX upon expression of Int-C3CNLS compared to inactive Int-INA expressing cells and hydroxyurea (HU)-treated cells, indicating that Int expression in human cells is does not induce noticeable levels of DNA damage.

3.11.3. Karyotype analysis

In order to determine if there are any chromosomal aberrations in the targeted clones, karyotype analysis was done on parental hESCs and three single-copy transgenic hESC clones (including E3). For each cell line 20 GTG-banded cells were scored and at least 5 GTG-banded cells were analyzed (Section 2.13;Material and Method) (Lim, A.S. et al.,

2013). A representative karyotype for clone E3 and parental hES cell line is shown (Figure 3.44).

Figure 3.44 Karyotyping to verify chromosomal stability. The targeted hESC lines hESC clone E3 and the parental hESC-047 were karyotyped by G-banding of metaphase chromosomes. A representative karyotype (from 20 scored and 5 analyzed GTG-banded cells) for each cell line is shown. Results indicated no apparent chromosomal abnormalities in the tested cell lines

Together, these data indicated that transient expression of Int-C3CNLS in human cells for *attH4X*-targeting appears to be safe with respect to cell toxicity and host cell genome integrity.

3.11.4 *attH4X* as putative safe harbor site

One important criterion to classify a genomic site as safe harbor site is minimal interaction of the inserted transgene with the flanking chromosomal regions (Sadelain, Papapetrou et al. 2012). To ascertain the possibility that at least some of the *attH4X* sites qualify as a safe harbor site, a global transcriptome analysis on single-copy integrant hES clones #3, #24 (Data not shown) and hES clone E3 (Figure 3.36) was performed, and an expression analysis through targeted next generation sequencing of about 20,000 protein-coding genes using Ion AmpliSeq™ Transcriptome Human Gene Expression Kit in combination with Ion Proton System, which covers >95% of human RefSeq genes (UCSC Ref Gene), was determined. (Section 2.15; Materials and Methods). Subsequent bioinformatics analysis was done to study differential gene expression using Torrent Suite™ Software.

We isolated mRNA from the three single-copy integrant clones #3, #24 (data not shown) and hESC E3 and determined comparative expression levels from cells harvested at early and late cell passages, and from corresponding samples of parental cells.

To carry this out, mRNA from single-copy integrant clones #3, #24 and hESC E3 was isolated. The comparative expression levels from the cells harvested at an early and a late cell passages, and from corresponding samples from parental cells was determined. Using a 4-fold difference in expression levels as cut-off for genes consistently differentially expressed, irrespective of cell passage, we identified expression variations for 22, 2 and 45 out of 20,000 genes for clones #3, #24 and E3, respectively (**Table 3.4**). Moreover, most of the genes were located on different chromosome or at-least at a distance of 10 million bp, (if on the same chromosome) from the targeted *attH4X* sequence. This indicated that, at least for these three *LINE-1* elements, targeting had a rather negligible effect on global transcriptome.

Table 3.4 Transcriptome analysis showing expression (log values) for upregulated genes (> log 2) and downregulated genes (< log2) along with their chromosomal location in hESC clones (#3, #24, E3).

hESC #3 : Targeting at chrX: 43,585.650-43.585.670

Gene Name	Expression (Log value)	Chromosomal location
CSAG1	4.1465	chrX:151,903,228-151,909,518
MAGEA2B	3.1835	chrX:151,918,388-151,922,364
SP140L	3.583	chr2:231,191,894-231,268,445
MAGEA2	2.864	chrX:151,918,388-151,922,364
CSAG3	2.534	chrX:151,923,679-151,928,738
TRIM43B	2.4655	chr2:96,142,715-96,150,479
DMD	3.072	chrX:32,382,699-32,383,325
RPL10L	2.915	chr14:47,120,220-47,121,028
DNER	-2.322	chr2:230,312,032-230,451,615
ZIC2	-2.068	chr13:100,634,026-100,639,019
HIST1H2BB	-2.5265	chr6:26,184,023-26,184,404
LOC729506	-2.578	chr5:8,333,596-8,457,677
ZNF257	-3.5415	chr19:22,235,266-22,273,903
ZNF560	-3.887	chr19:9,577,031-9,609,279
ZNF492	-3.789	chr19:22,817,126-22,850,472
RWDD2B	-3.0945	chr21:30,378,080-30,391,685
XIST	-3.937	chrX:73,040,486-73,072,588
CLC	-4.4425	chr19:40,221,893-40,228,669
ZNF98	-4.966	chr19:22,573,899-22,605,148
ZNF676	-5.843	chr19:22,361,903-22,379,753
ZNF729	-6.006	chr19:22,469,252-22,499,978
HIST1H3C	-6.1145	chr6:26,045,639-26,046,097
ZNF208	-7.899	chr19:22,148,897-22,193,745

hESC #24: Targeting at chr5:12.672.998-12.673.018

Gene Name	Expression (Log value)	Chromosomal location
ZNF560	-2.5195	chr19:9,577,031-9,609,279
XIST	-4.236	chrX:73,040,486-73,072,588

hESC E3: Targeting at chr1: 213,884,581-213,884,602

Gene Name	Expression (Log value)	Chromosomal location
PIK3AP1	3.494	chr10:98,353,069-98,480,279
MAGEA2B	3.4025	chrX:151,918,388-151,922,364
MAGEA2	3.4285	chrX:151,918,388-151,922,364
CSAG1	3.3225	chrX:151,903,228-151,909,518
CSAG3	3.5585	chrX:151,923,679-151,928,738
PCDH10	3.2505	chr4:134,070,470-134,112,732
KHDC3L	3.659	chr6:74,072,400-74,073,898
NPPB	3.1175	chr1:11,917,521-11,918,992
SFTA1P	2.8615	chr10:10,826,402-10,836,877
S100A4	2.8155	chr1:153,516,095-153,518,282
PLAT	2.646	chr8:42,032,236-42,065,194
LOC339535	2.582	chr1:238,643,684-238,649,317
CDKN2B	2.613	chr9:22,002,902-22,009,312
SP140L	2.997	chr2:231,191,894-231,268,445
MMP9	2.503	chr20:44,637,547-44,645,200
HAPLN1	3.389	chr5:82,934,017-83,016,896
MYOF	2.942	chr10:95,066,186-95,242,074
COL5A1	3.0405	chr9:137,533,651-137,736,688
SYTL5	2.6955	chrX:37,892,787-37,988,073
TGFB1	2.9995	chr5:135,364,584-135,399,507
CAPN2	2.2885	chr1:223,900,119-223,963,720
CD46	2.212	chr1:207,925,383-207,968,861
CXCL14	2.2885	chr5:134,906,371-134,914,969
EBI3	2.209	chr19:4,229,540-4,237,524
TGM2	2.477	chr20:36,756,864-36,793,700
NDRG1	2.269	chr8:134,249,414-134,309,547
COL3A1	2.823	chr2:189,839,099-189,877,472
DPEP3	2.254	chr16:68,009,566-68,014,452
ACTA1	2.1155	chr1:229,566,993-229,569,843
FOXD1	2.5285	chr5:72,742,085-72,744,352
WNK4	2.5205	chr17:40,932,649-40,949,084
FRG1B	2.5545	chr20:29,624,035-29,653,908
ANGPT1	2.0485	chr8:108,261,710-108,510,254
ACTC1	2.2675	chr15:35,080,297-35,087,927
ZNF99	-2.3175	chr19:22,934,985-22,966,973
COMTD1	-2.4025	chr10:76,993,729-76,995,770
LOC729506	-3.2075	chr5:8,333,596-8,457,677
CAT	-3.1955	chr11:34,460,472-34,493,607
ZNF560	-3.4215	chr19:9,577,031-9,609,279
ZNF257	-4.171	chr19:22,235,266-22,273,903
XIST	-4.566	chrX:73,040,486-73,072,588
ZNF98	-4.734	chr19:22,573,899-22,605,148
ZNF676	-5.019	chr19:22,361,903-22,379,753
ZNF729	-5.2215	chr19:22,469,252-22,499,978
ZNF208	-5.9805	chr19:22,148,897-22,193,745

Page intentionally left blank

4. DISCUSSION

Over the last two decades, significant efforts have been made to harness the potential of genome engineering tools to establish proof of concept of precise and safe human genome engineering and subsequent translation in the quest for gene therapy and other clinically relevant applications. Most efforts in this direction have failed due to the critical thresholds of efficacy and safety issues such as off-targeting, insertional mutagenesis, chromosomal aberration and cytotoxicity associated with the current transgenesis tools in translational research. In addition, these tools also have the limitation in the cargo size that can be introduced into the genome as the efficiency of targeting declines with larger insert size. This necessitates a refinement of the current editing toolbox or escalated efforts towards development of novel engineering tools that would possibly overcome the current challenges pertaining to the controlled manipulation of the human genome.

For the past 25 years, significant progress has been made to exploit bacteriophage integrases (belonging to the family of site-specific recombinases) as genome manipulation tools for diverse applications. Amongst others, in particular tyrosine recombinases have been a subject of intense study due to its conserved nature (recombination without indels) and high target site specificity, even with large DNA molecules for the purpose of genome modification (Groth and Calos 2004). This guided the researchers to explore site-specific phage integrases for genome manipulation of various bacteria such as *E. coli* (Atlung, Nielsen et al. 1991), *Staphylococcus* (Lee, Buranen et al. 1991), *Actinobacteria*, *Mycobacteria* (Lee, Pascopella et al. 1991, Gargett and Brown 2014, Tey 2014) *Pseudomonas aeruginosa* (Hoang, Kutchma et al. 2000), for applications such as production of recombinant proteins, vaccines and for studying virulence factors (Lee, Pascopella et al. 1991). All these integrases carry out recombination using a dual plasmid system: Int expressing plasmid and *attP*-containing plasmid (encoding a suitable marker and a transgene to be integrated) (Fogg, Colloms et al. 2014).

λ -Int is a member of the large tyrosine-recombinase family and was the first prototypical transgenesis tool derived from bacteriophage λ for manipulation of the *E. coli* genome using *attB/attP* sites (Fogg, Colloms et al. 2014). However, the applicability of λ -Int in eukaryotic system was limited due to obligate requirement of accessory co-factors and supercoiling of *attP* site, unlike the Cre/Flp systems (Groth and Calos 2004). To

overcome this problem, various mutants of λ -Int were generated (Lorbach, Christ et al. 2000; Tay, Ho et al. 2010, Siau, Chee et al. 2015) which could recombine *att* sites in eukaryotic system independent of co-factors and supercoiling requirements. In this context, a strategy was designed to generate engineered cell lines wherein a single copy of *attP* was artificially harbored randomly into the genome for targeting transgenes via *attP* X *attB* recombination (Sabrina Peter and Peter Droge, data not published).

Another site-specific recombination tool, serine integrase, PhiC31, has also been extensively studied to manipulate eukaryotic genomes. This system allows targeting of endogenous *att*-like sequences, termed pseudo-*attP* sites in the genome and has been extensively used to target mouse, bovine, zebrafish and human genome (Bolusani, Ma et al. 2006, Ou, Huang et al. 2009). However, the problems pertaining to safety, efficacy and inconsistency in terms of reproducibility have limited its use for gene therapy (Thyagarajan, Olivares et al. 2001, Malla, Dafhnis-Calas et al. 2005, Ou, Huang et al. 2009).

Considering the fact that λ -Int system can faithfully integrate large DNA molecules, using a short and specific 21 bp sequence as target, we explored the possibility of targeting the human genome. To this effect, a global human genome search was performed to possibly identify an endogenous *attB*-like sequence. However, the search results did not yield a perfect match, rather a similar sequence in the exon 5 of *MCT5* gene was found, termed as *attH*-*MCT5* that was targeted using its corresponding *attPH*-*MCT5* sequence (**Figure 3.1**) (Tay, Ho et al. 2010). In this context, various mutants of Int were generated, of which Int-C3 outperformed all other variants to recombine wild type as well as non-cognate *att* sites (**Figure 3.6 and 3.7**) (Siau, Chee et al. 2015). However, targeting of genomic *attH*-*MCT5* sequence was performed using Int-h/218 as the activity of Int-h/218 was higher than Int-1 and Int-5 mutants (**Figure 3.6**) and the variant Int-C3 was not generated at that time.

Several attempts of targeting *attH*-*MCT5* with *attPH*-*MCT5* sequence using Int-h/218 were not successful. We could only detect inter-molecular recombination between the target vector and Int expressing plasmid followed by random integration into the human genome (**Section 3.3.3**). At this point, we do not understand the reason behind this illegitimate recombination between two vectors, as there was no sequence homology involved in this process.

Simultaneously, another approach for the identification of an *attB*-like (*attH*) sequence in the human genome was used wherein different target vectors comprising of various modified *attP* (*attP39*/ WT *attP*) sequences (**Figure 3.13a and 3.22a**), previously generated in our laboratory, was used for targeting human genome. Using inverse PCR technique, it was observed that integration of *attP39* sequence occurred into a 21bp sequence in chromosome 6 in a λ -Int-mediated manner and this sequence was termed as *attH*-chr6 (**Figure 3.14**). However, this targeting could not be reproduced with *attP39* or an adjusted *attPH*-chr6 sequence, even after screening of a significant number of colonies from multiple transfections. Furthermore, in two separate and independent targeting events with *attP39* sequence, it was observed that integration of *attP39* occurred at two different *attH* sequences in chromosome 13 in a λ -Int-mediated fashion (**Figure 3.16 and 3.21**).

In one of the targeting events, *attL* (HOP') recombinant site could not be determined and *attR* (POH') recombinant site had an insertion of extra 19 nucleotides (in black) after the overlap (highlighted in grey & underlined) of *attP39* sequence, prior to integration into chromosome 13 (in blue & underlined) (as shown below). This integration may have occurred via λ -Int-mediated recombination, as the cleavage seemed to be induced at the end of the overlap in *attP39*. However, the comparison of the 21bp sequence of *attP39* (in red) with the native *attH* locus at chromosome 13 (in blue) shows very little homology at the right core-binding site.

<i>attR</i> (POH'):	<u>CAGCTTTTTTAAAA</u> AAAGCTGAACATAATAAAATAACTTC
<i>attP39</i> (POP'):	CAGCTTT <u>TTTAAAA</u> TAAGTTG
<i>attH</i> (chr13):	TTCAC <u>TTCTGCAG</u> TAACTTC

In the second targeting event, integration of *attP39* occurred at a different locus on the chromosome 13 (**Figure 3.20c**). In this targeting, although *attL* and *attR* junctions could be determined, there was a deletion of about 70 million bp between two recombinant sites and an addition of 19 nucleotides at the *attL* recombinant site (**Figure 3.20c and Figure 3.21**). The observed *attL* sequence had the partial overlap and right core-binding site of *attP39* (in red and underlined) whereas the observed *attR* sequence had the left core-binding site of *attP39* sequence (in red and underlined), indicating λ -Int-mediated recombination. However, the comparison of *attP39* sequence with that of *attH* (based on *attL* sequencing result) shows some homology at the overlap and right core-binding site

whereas comparison of *attP39* sequence with that of *attH* (based on *attR* sequencing result) shows complete homology at the left core-binding site.

<i>attL</i> (HOP'):	CCACTGATCTAAAAATAAGTTG
<i>attP39</i> (POP'):	CAGCTTTTTTAAAAATAAGTTG
<i>attH</i> (based on <i>attL</i>):	TTTAAAGGTGAAAATGATGTG

<i>attR</i> (POH'):	CAGCTTTCAGTTTGATCCCTC
<i>attP39</i> (POP'):	CAGCTTTTTTAAAAATAAGTTG
<i>attH</i> (based on <i>attR</i>):	CAGCTTTCAGTTTGATCCCTC

Thus, our observation in the aforesaid targeting events suggests a compromise in the fidelious λ -Int-mediated recombination. The failure of λ -Int-mediated recombination with *attP39* at chromosome 13 in both the targeting events could be attributed to the complex nature of λ -Int-mediated recombination, which makes it rather more difficult to provide an explanation for such observations in context of human genome. However, based on previous studies, it has been shown that integrase once bound to *attP* sequence can capture the naked *attB* site (in this case *attH*) for synapsis, suggesting complete homology may not be required for capture and cleavage of *attH* site at the end of overlap sequence (Richet, Abcarian et al. 1988). Therefore, it can be speculated that in both targeting events, minimal homology at any of the core-binding site was sufficient for the initiation of λ -Int-mediated recombination. But due to the non-availability of compatible partner, the recombination process could have been aborted (Kitts and Nash 1987, Rajeev, Malanowska et al. 2009). During this process, a compromise in the genome integrity could also have possibly led to the activation of the cellular repair mechanism that would have resulted in indels at the targeted locus.

In yet another case, targeting with another *attP* variant, *attP*(WT), wherein a putative *attH* was found to be in chromosome 8 (*attH*-chr 8), led to predicted recombination without any addition or loss of nucleotides (as described in **Section 3.6**). Here, the overlap of *attH*-chr 8 exactly matched the overlap of its recombining partner *attP*(WT), with the exception of one nucleotide at the 2nd position from the left side. This could have been

tolerated by Int-h/218 while catalyzing the formation and resolution of Holliday junction intermediate. However, this targeting could not be reproduced, perhaps due to low the copy number of *attH*-chr8, not so easily accessible chromatin structure at the *attH*-chr 8 or the inefficiency of generating recombinant products in the presence of the aforementioned single nucleotide mismatch in the overlap.

Since the previous targeting attempts were not very successful, various other experimental and bioinformatics approaches to identify a putative *attH* site in the human genome were explored. With these approaches, a potential 18bp sequence, termed *attH4X* was found that could specifically be targeted with its corresponding recombination partner *attP4X* sequence (in the target vector) (**Figure 3.25**). This *attH4X* sequence was found to be a part of *LINE-1* elements and occurred at 935 different locations (**Figure 3.27 and 3.28**). In order to pick the positive integration events by PCR, consensus primers were designed in the flanking conserved region of *attH4X* (**Figure 3.28**). The targeting of *attH4X* sequence at various *LINE-1* loci in human genome with *attP4X*-based target vector carrying various transgenes was successfully achieved and validated in various cell lines including HT1080, HeLa, keratinocytes (NEB-1) and hESCs with an efficiency ranging between 10-20%. The specificity and the copy number of the target vector integrated in the human genome were determined by Southern blotting, which demonstrates that single-copy transgene (currently comprising > 8kb) insertions in *LINE-1* elements can be achieved in more than half of the targeted clones analyzed (**Table 4.1**). In addition, sequencing results showed that recombination between genomic *attH4X* and *attP4X* occurred as predicted without any unwanted mutations.

Table 4.1 Targeting in various cell lines

Cell lines	Number of positive clones	Single copy integrants
HT 1080	4/33	2/4
Hela	5/44	4/5
NEB-1	2/26	2/2
hESCs	17/120	8/17

It has also been shown that *attL4X* (but not *attR4X*) could functionally replace the *attP4X* sequence for targeting *attH4X* (**Figure 3.38**). Though the mechanism behind *attL* (BOP') recombining with *attB* (BOB') is not known, our finding suggests only half arm site (P')

is required for Int to bind and catalyze recombination between *attL4X* X *attH4X*. The possible reason for this surprising observation could be the arrangement of Int arm binding sites on the target vector that could possibly affect recombinogenic potential of *att* sites in combination with Int-C3, at least in human cells. However, this itself is a very interesting finding that could be exploited to unravel the mechanism by which mutant λ -Int catalyze recombination. Nonetheless, *attL4X* could target *attH4X* with the same efficiency as *attP4X* in HT1080 as 5/60 clones were positive for both (*attL* and *attR*) recombinant sites, with all the clones being single copy integrants (**Figure 3.40**).

In order to determine the functionality of targeted *attH4X* loci, a target vector comprising of puromycin resistance gene under a constitutive promoter and a GFP gene driven by undifferentiated transcription factor-1 (UTF1) promoter (an indicator of pluripotency) was used (**Figure 3.31**). The transgenes targeted at these loci show expected expression both in pluripotent hESCs and differentiated progenies for the clones A3, E3 and K3 (**Figure 3.36**). Upon retinoic acid (RA) induced differentiation, the hESC clones lost the GFP expression (as expected), indicating that our transgene cassettes remained functional and receptive to cellular signaling at targeted *attH4X* loci in hESCs (**Figure 3.36 a, b**). This was further confirmed by western analysis (See **appendix Figure A2**)

It is possible that Int-C3-mediated recombination preferentially occurs at loci with a more accessible or recombination-permissive chromatin structure promoting sustained transgene functionality. This idea is supported by the fact that integrants still contained bacterial sequences, which are prone to genomic silencing, particularly in hESCs (Fontes and Lakshmipathy 2013). In this context it is worth noting that full length yet inactive *LINE-1* elements, particularly those belonging to the L1PA7 subfamily, seem to be overrepresented in the list of targeted *attH4X* sequences (**Table 3.3**). This may also indicate that a more permissive chromatin state is perhaps present at these loci. However, it is also clear that the pattern of transgene expression varied considerably between individual clones. Because there is no *a priori* size limit for Int target vectors as long as they can be generated in circular form, it should be possible in the future to include flanking insulator sequences in order to minimize heterogeneity in transgene expression pattern.

Another interesting future improvement of our technology may include elimination of remaining bacterial sequences from Int target vectors through *in vitro* *attP4X* x *attH4X* intra-molecular Int-mediated recombination, resulting in seamless *attL4X* vectors (**Figure**

3.41), which may also contribute to increased homogeneity in transgene expression pattern (Figure 4.1).

Figure 4.1 *In vitro* generation of seamless vector. A schematic diagram showing *in vitro* intra-molecular Int-mediated *attH4X* X *attP4X* recombination in the substrate vector resulting in seamless *attL4X* vectors (eliminating the undesired bacterial sequences), which can be targeted at *attH4X*

On a different note, about 500,000 *LINE-1* elements sculpted the human genome during mammalian evolution (Burns and Boeke 2012) and their integration sites are generally considered benign with respect to human genome function (Sadelain, Papapetrou et al. 2012, Babatz and Burns 2013). *LINE-1* elements belong to the large family of mobile elements of non-LTR transposons that replicate via RNA intermediate. It mainly comprises of truncated or mutated copies of functional, full-length elements out of which only a subset are active (80-100 *LINE-1* elements) and are capable of retrotransposition (Furano 2000). The full-length *LINE-1* element is 6-7kb long and consists of 5'UTR, two ORFs and a 3' UTR region (Figure 3.27). The 5' UTR region consists of a sense promoter that drives the transcription of *LINE-1* element and an anti-sense promoter that is known to regulate the activity of retrotransposon via RNAi mechanism (Feng, Moran et al. 1996, Furano 2000). The ORF-1 encodes for RNA-binding protein, which is required for ribonucleoprotein particle assembly whereas ORF2 encodes for endonuclease and reverse-transcriptase enzymes, which are required for converting RNA intermediate into DNA and enables retro-transposition (Martin 2006). There are various ways by which

LINE-1 mediated retrotransposition events can impact the human genome (Beck, Garcia-Perez et al. 2011). For example, it can lead to exon disruption directly or may cause exon skipping/ mis-splicing by disrupting introns (Belancio, Hedges et al. 2008). This has been shown in the case of hemophilia where *LINE-1* retrotransposition disrupted exon 14 of the factor IX, thus acting a disease mutagen (Li, Scaringe et al. 2001). In some cases, it is also known to affect the gene expression by variety of mechanism. *LINE-1* retrotransposition can affect both germ and somatic cells at discrete times during development, which could also be responsible for intra-individual genetic variation and also contribute to the genomic diversity (Bennett, Coleman et al. 2004, Belancio, Roy-Engel et al. 2010).

The *attH4X* target sequence resides in ORF1 of the *LINE-1* (Erwin et al., 2014) and the vast majority of ORF1 sequences are non-functional due to mutations and truncations acquired throughout human evolution (Babatz and Burns 2013). Disruption of ORF1 through transgenesis may thus pose little or no risk to human genome function. In particular, *LINE-1* loci bearing the *attH4X* sequence located in long intergenic regions may fulfil this criterion of safe harbor sites (Sadelain, Papapetrou et al. 2012). In fact, our global transcriptome analyses obtained for three hESC clones at different passages in comparison with parental cells revealed that transgenes at the respective *LINE-1* loci can be considered safe as their RNA expression profiles did not show any significant up/down regulation of the proximal genes. Out of 20,000 genes analyzed in each targeted clone, only 20-40 genes (~0.2%) showed differential expression profile compared to parental cells suggesting negligible effect of targeting on the genome (**Table 3.4**). In addition, none of these differentially expressed genes fell in the category of oncogenes and most of them were located on different chromosome or at-least at a distance of 10 million bp, (if on the same chromosome) from the targeted *attH4X* sequence (**Table 3.4**). Since sequence specific transgenesis at *LINE-1* elements leaves global transcriptomes largely undisturbed and also allows functional and sustained expression of transgenes, thus meeting most of the criteria for safe harbor sites based on previous data (Sadelain, Papapetrou et al. 2012). In addition, it has been known that *LINE-1* elements are prevalent more in AT-rich, low-recombining and sparse gene regions of the genome (Graham and Boissinot 2006). Based on our studies, at least some of the *attH4X* sites can be considered as a putative human genome ‘safe harbor site’.

The safety profiling experiments (**Section 3.11**) have shown that cellular toxicity, genotoxicity and chromosomal aberrations were not induced during targeting. Together, these data indicate that a transient expression of Int-C3 in human cells for *attH4X* targeting is safe.

In contrast to our previous targeting with other *attP* variants (such as *attP39*), our *attP4X* based targeting has been specific in the analysed clones (as described in (**Section 3.5 and 3.6**). This could be due to the provision of functional and compatible partner sequence that ensures legitimate recombination at *attH4X*, unlike in the case of *attP39* targeting where *att* sites showed minimal homology. However, it is possible that the *attH4X*-targeted clones carrying multiple copies of transgenes, which could either result from insertion into multiple *LINE-1* elements or from a mixture of targeted and random integration events. However, we did not investigate this possibility further in the current study.

Therefore, we report a novel λ -Int-mediated sequence specific recombination- based transgenesis tool capable of inserting large multi-transgene cassettes at putative safe harbor site without activating cellular host repair machinery. This would really advance the utility of λ -Int system in the field of human genome engineering as our study provides strong evidence that variants of λ -Int in conjunction with modified *attP* can routinely be used for efficiently targeting the human genome, rather than otherwise opting for recombination strategy through artificially introduced *att* sites. This would not only enable targeting of desired gene at specific locus with minimal interference from the other components of the host chromosome but would also avoid safety issues such as genotoxicity, adverse host immune responses and insertional mutagenesis associated with viral and plasmid-mediated tools. In addition, our tool does not introduce DSBs thus greatly reducing the risk of unwanted mutations/truncations, chromosomal rearrangements and cell toxicity, otherwise witnessed with nuclease based genome-editing tools (ZFNs, TALENs and CRISPR/Cas9). Another distinct advantage is that with a single targeting attempt, it is possible to choose from various target sites, some of them may be safe harbor sites, to obtain desired multi-transgene cassette performances.

The applications of our simplified and user-friendly transgenesis tool are plentiful. We have pioneered this system for applications in pluripotent hESCs, therefore we envision that this will be of broader usefulness to the scientific community as well as for biotech industry in general, in particular when large transgene constructs are being employed. For

example, this could be quite beneficial for stem cell engineering where large constructs consisting of multi-cassettes transgene are used to study various developmental stages of the stem cells. In addition, it could also be used to modify other cell lines, thus enable us to tap huge cell line development market.

Gene therapy could be another potential application that can be explored through our novel improved tool. One such example could be insertion of Chimeric Antigen Receptor (CAR) construct for CAR-T based immunotherapy against cancers. The development of efficient and safer CARs would involve introduction of large multi-transgene constructs with regulatory elements and safety switches (iCasp9) (Gargett and Brown 2014, Tey 2014) in the form of seamless vector at specific locus. This would mitigate the safety issues otherwise associated with current CAR-T based immunotherapy approaches using viral platforms. Other promising avenues that can be considered include treatment of monogenic diseases (hemophilia, sickle cell anemia etc.) by insertion of functional copy of therapeutic gene at pre-determined locus.

Bio-production of therapeutic antibodies is also few of the other potential applications that could benefit our technological strengths. Transgene targeted human cell lines have emerged as a powerful and important source for the bio-production of therapeutic proteins as it allows the production of proteins with a similar glycosylation pattern compared to its native form, thus reducing the risks of potential immunogenicity. Our technology can be utilized in this important niche by targeting the therapeutic gene at the safe harbor sites in the human genome and maintaining the quality of the proteins.

Hence we believe that our tool would be an important addition to the current toolbox of transgenesis tools for precise and safe human genome engineering particularly, when applied to clinically relevant applications.

5. FUTURE WORK

Although we have developed a prototype for novel λ -Int-mediated site-specific insertion tool, certain improvements can be done to further simplify the technology.

In this context, firstly a target vector can be modified into the seamless vectors (eliminating the undesired bacterial sequences) and Int can be used in a highly active purified form, rather than expression from plasmid or mRNA. This can be done in the following ways.

- (i) Purified Integrase protein: Int-C3 can be purified in a highly active form in large batches via step-wise protein purification protocols already published (Siau *et al.*, 2015). This protein version of int would prevent random integration of int expressing plasmids into the genome.
- (ii) Generation of seamless target vectors: Target vectors can be engineered into seamless vectors (devoid of non-essential bacterial sequences). It is known that the bacterial sequences such as origin of replication or selection markers in the plasmid are prone to epigenetic silencing after integration into the human genome leading to unstable expression of the transgene and generating adverse host immune responses. To overcome these problems, bacterial sequences can be eliminated from the target vectors through in vitro *attP4X* x *attH4X* intra-molecular recombination, resulting in seamless, circular *attL4X* vector (Figure 4.1) This *attL4X* then can be used to recombine with *attH4X*, integrating the desired sequence into the safe harbor site without any unwanted sequences inserted into the genome (unpublished data).

Secondly, in order to further provide tangible evidence of potential targeting of large transgenes, various progressive attempts can be made to target (~15-20kb) multi-cassettes transgenes at *attH4X* to explore the feasibility of our tool in applications such stem cell engineering, CAR-T engineering and synthetic biology.

Thirdly, the targeting efficiency of our system is about 10%, but the real value might be higher because sequence polymorphism in the 935 *LINE-1* ORF1 might compromise the binding of consensus primer and our inability to detect some of the integration events in first round screens. Therefore, improvements in designing of the primers would facilitate the detection of untapped site- specific integration events.

Page intentionally left blank

REFERENCES

1. Adamson, A. D., D. Jackson and J. R. Davis (2011). "Novel approaches to in vitro transgenesis." *J Endocrinol* 208(3): 193-206.
2. Allen, B. G. and D. L. Weeks (2005). "Transgenic *Xenopus laevis* embryos can be generated using phiC31 integrase." *Nat Methods* 2(12): 975-979.
3. Atlung, T., A. Nielsen, L. J. Rasmussen, L. J. Nellesmann and F. Holm (1991). "A versatile method for integration of genes and gene fusions into the lambda attachment site of *Escherichia coli*." *Gene* 107(1): 11-17.
4. Babatz, T. D. and K. H. Burns (2013). "Functional impact of the human mobilome." *Curr Opin Genet Dev* 23(3): 264-270.
5. Beck, C. R., J. L. Garcia-Perez, R. M. Badge and J. V. Moran (2011). "LINE-1 elements in structural variation and disease." *Annu Rev Genomics Hum Genet* 12: 187-215.
6. Belancio, V. P., D. J. Hedges and P. Deininger (2008). "Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health." *Genome Res* 18(3): 343-358.
7. Belancio, V. P., A. M. Roy-Engel, R. R. Pochampally and P. Deininger (2010). "Somatic expression of LINE-1 elements in human tissues." *Nucleic Acids Res* 38(12): 3909-3922.
8. Bennett, E. A., L. E. Coleman, C. Tsui, W. S. Pittard and S. E. Devine (2004). "Natural genetic variation caused by transposable elements in humans." *Genetics* 168(2): 933-951.
9. Bibikova, M., K. Beumer, J. K. Trautman and D. Carroll (2003). "Enhancing gene targeting with designed zinc finger nucleases." *Science* 300(5620): 764.
10. Boch, J., H. Scholze, S. Schornack, A. Landgraf, S. Hahn, S. Kay, T. Lahaye, A. Nickstadt and U. Bonas (2009). "Breaking the code of DNA binding specificity of TAL-type III effectors." *Science* 326(5959): 1509-1512.
11. Bogdanove, A. J., S. Schornack and T. Lahaye (2010). "TAL effectors: finding plant genes for disease and defense." *Curr Opin Plant Biol* 13(4): 394-401.
12. Bolusani, S., C. H. Ma, A. Paek, J. H. Konieczka, M. Jayaram and Y. Voziyanov (2006). "Evolution of variants of yeast site-specific recombinase F1p that utilize native genomic sequences as recombination target sites." *Nucleic Acids Res* 34(18): 5259-5269.
13. Brown, W. R., N. C. Lee, Z. Xu and M. C. Smith (2011). "Serine recombinases as tools for genome engineering." *Methods* 53(4): 372-379.

14. Buchholz, F., P.-O. Angrand and A. F. Stewart (1998). "Improved properties of FLP recombinase evolved by cycling mutagenesis." *Nat Biotech* 16(7): 657-662.
15. Buchholz, F. and A. F. Stewart (2001). "Alteration of Cre recombinase site specificity by substrate-linked protein evolution." *Nat Biotechnol* 19(11): 1047-1052.
16. Buganim, Y., D. A. Faddah, A. W. Cheng, E. Itskovich, S. Markoulaki, K. Ganz, S. L. Klemm, A. van Oudenaarden and R. Jaenisch (2012). "Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase." *Cell* 150(6): 1209-1222.
17. Burgin, A. B., Jr. and H. A. Nash (1992). "Symmetry in the mechanism of bacteriophage lambda integrative recombination." *Proc Natl Acad Sci U S A* 89(20): 9642-9646.
18. Burns, K. H. and J. D. Boeke (2012). "Human transposon tectonics." *Cell* 149(4): 740-752.
19. Bushman, F., M. Lewinski, A. Ciuffi, S. Barr, J. Leipzig, S. Hannenhalli and C. Hoffmann (2005). "Genome-wide analysis of retroviral DNA integration." *Nat Rev Microbiol* 3(11): 848-858.
20. Byrne, S. M., L. Ortiz, P. Mali, J. Aach and G. M. Church (2015). "Multi-kilobase homozygous targeted gene replacement in human induced pluripotent stem cells." *Nucleic Acids Res* 43(3): e21.
21. Carroll, D. (2011). "Genome engineering with zinc-finger nucleases." *Genetics* 188(4): 773-782.
22. Chalberg, T. W., H. L. Genise, D. Vollrath and M. P. Calos (2005). "phiC31 integrase confers genomic integration and long-term transgene expression in rat retina." *Invest Ophthalmol Vis Sci* 46(6): 2140-2146.
23. Christ, N., T. Corona and P. Droge (2002). "Site-specific recombination in eukaryotic cells mediated by mutant lambda integrases: implications for synaptic complex formation and the reactivity of episomal DNA segments." *J Mol Biol* 319(2): 305-314.
24. Christ, N. and P. Droge (1999). "Alterations in the directionality of lambda site-specific recombination catalyzed by mutant integrases in vivo." *J Mol Biol* 288(5): 825-836.
25. Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini and F. Zhang (2013). "Multiplex genome engineering using CRISPR/Cas systems." *Science* 339(6121): 819-823.

26. Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini and F. Zhang (2013). "Multiplex Genome Engineering Using CRISPR/Cas Systems." *Science* 339(6121): 819-823.
27. Cornu, T. I., S. Thibodeau-Beganny, E. Guhl, S. Alwin, M. Eichtinger, J. K. Joung and T. Cathomen (2008). "DNA-binding specificity is a major determinant of the activity and toxicity of zinc-finger nucleases." *Mol Ther* 16(2): 352-358.
28. Didigu, C. A., C. B. Wilen, J. Wang, J. Duong, A. J. Secreto, G. A. Danet-Desnoyers, J. L. Riley, P. D. Gregory, C. H. June, M. C. Holmes and R. W. Doms (2014). Simultaneous zinc-finger nuclease editing of the HIV coreceptors ccr5 and cxcr4 protects CD4+ T cells from HIV-1 infection.
29. Dillon, N. and R. Festenstein (2002). "Unravelling heterochromatin: competition between positive and negative factors regulates accessibility." *Trends Genet* 18(5): 252-258.
30. Doudna, J. A. and E. Charpentier (2014). "Genome editing. The new frontier of genome engineering with CRISPR-Cas9." *Science* 346(6213): 1258096.
31. Doudna, J. A. and E. Charpentier (2014). "The new frontier of genome engineering with CRISPR-Cas9." *Science* 346(6213).
32. Dupuy, A. J. (2010). "Current applications of transposons in mouse genetics." *Methods Enzymol* 477: 53-70.
33. Ehrhardt, A., J. A. Engler, H. Xu, A. M. Cherry and M. A. Kay (2006). "Molecular analysis of chromosomal rearrangements in mammalian cells after phiC31-mediated integration." *Hum Gene Ther* 17(11): 1077-1094.
34. Erwin, J. A., M. C. Marchetto and F. H. Gage (2014). "Mobile DNA elements in the generation of diversity and complexity in the brain." *Nat Rev Neurosci* 15(8): 497-506.
35. Feng, Q., J. V. Moran, H. H. Kazazian, Jr. and J. D. Boeke (1996). "Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition." *Cell* 87(5): 905-916.
36. Feng, Y. Q., J. Seibler, R. Alami, A. Eisen, K. A. Westerman, P. Leboulch, S. Fiering and E. E. Bouhassira (1999). "Site-specific chromosomal integration in mammalian cells: highly efficient CRE recombinase-mediated cassette exchange." *J Mol Biol* 292(4): 779-785.

37. Finotti, A., L. Breda, C. W. Lederer, N. Bianchi, C. Zuccato, M. Kleanthous, S. Rivella and R. Gambari (2015). "Recent trends in the gene therapy of beta-thalassemia." *J Blood Med* 6: 69-85.
38. Fischer, A., S. Hacein-Bey-Abina and M. Cavazzana-Calvo (2010). "20 years of gene therapy for SCID." *Nat Immunol* 11(6): 457-460.
39. Fogg, P. C., S. Colloms, S. Rosser, M. Stark and M. C. Smith (2014). "New applications for phage integrases." *J Mol Biol* 426(15): 2703-2716.
40. Fontes, A. and U. Lakshmipathy (2013). "Advances in genetic modification of pluripotent stem cells." *Biotechnol Adv* 31(7): 994-1001.
41. Fu, Y., J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, J. K. Joung and J. D. Sander (2013). "High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells." *Nat Biotechnol* 31(9): 822-826.
42. Furano, A. V. (2000). "The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons." *Prog Nucleic Acid Res Mol Biol* 64: 255-294.
43. Furushima, K., C. W. Jang, D. W. Chen, N. Xiao, P. A. Overbeek and R. R. Behringer (2012). "Insertional mutagenesis by a hybrid piggyBac and sleeping beauty transposon in the rat." *Genetics* 192(4): 1235-1248.
44. Gaj, T., C. A. Gersbach and C. F. Barbas, 3rd (2013). "ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering." *Trends Biotechnol* 31(7): 397-405.
45. Gargett, T. and M. P. Brown (2014). "The inducible caspase-9 suicide gene system as a "safety switch" to limit on-target, off-tumor toxicities of chimeric antigen receptor T cells." *Front Pharmacol* 5: 235.
46. Ghosh, K. and G. D. Van Duyne (2002). "Cre-loxP biochemistry." *Methods* 28(3): 374-383.
47. Glorioso, J. C. (2000). "Perspectives on viral vector design and applications." *Adv Virus Res* 55: 403-407.
48. Graham, T. and S. Boissinot (2006). "The genomic distribution of L1 elements: the role of insertion bias and natural selection." *J Biomed Biotechnol* 2006(1): 75327.
49. Grindley, N. D., K. L. Whiteson and P. A. Rice (2006). "Mechanisms of site-specific recombination." *Annu Rev Biochem* 75: 567-605.
50. Groth, A. C. and M. P. Calos (2004). "Phage integrases: biology and applications." *J Mol Biol* 335(3): 667-678.

51. Gu, H., Y. R. Zou and K. Rajewsky (1993). "Independent control of immunoglobulin switch recombination at individual switch regions evidenced through Cre-loxP-mediated gene targeting." *Cell* 73(6): 1155-1164.
52. Gupta, R. M. and K. Musunuru (2014). "Expanding the genetic editing tool kit: ZFNs, TALENs, and CRISPR-Cas9." *J Clin Invest* 124(10): 4154-4161.
53. Haccin-Bey-Abina, S., A. Garrigue, G. P. Wang, J. Soulier, A. Lim, E. Morillon, E. Clappier, L. Caccavelli, E. Delabesse, K. Beldjord, V. Asnafi, E. MacIntyre, L. Dal Cortivo, I. Radford, N. Brousse, F. Sigaux, D. Moshous, J. Hauer, A. Borkhardt, B. H. Belohradsky, U. Wintergerst, M. C. Velez, L. Leiva, R. Sorensen, N. Wulffraat, S. Blanche, F. D. Bushman, A. Fischer and M. Cavazzana-Calvo (2008). "Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1." *J Clin Invest* 118(9): 3132-3142.
54. Hackett, P. B., S. C. Ekker, D. A. Largaespada and R. S. McIvor (2005). "Sleeping beauty transposon-mediated gene therapy for prolonged expression." *Adv Genet* 54: 189-232.
55. Handel, E.-M. and T. Cathomen (2011). "Zinc-Finger Nuclease Based Genome Surgery: It's All About Specificity." *Current Gene Therapy* 11(1): 28-37.
56. Harrison, M. M., B. V. Jenkins, K. M. O'Connor-Giles and J. Wildonger (2014). "A CRISPR view of development." *Genes Dev* 28(17): 1859-1872.
57. Henckaerts, E. and R. M. Linden (2010). "Adeno-associated virus: a key to the human genome?" *Future Virol* 5(5): 555-574.
58. Hirano, N., T. Muroi, H. Takahashi and M. Haruki (2011). "Site-specific recombinases as tools for heterologous gene integration." *Appl Microbiol Biotechnol* 92(2): 227-239.
59. Hoang, T. T., A. J. Kutchma, A. Becher and H. P. Schweizer (2000). "Integration-proficient plasmids for *Pseudomonas aeruginosa*: site-specific integration and use for engineering of reporter and expression strains." *Plasmid* 43(1): 59-72.
60. Hodges, B. L., K. M. Taylor, M. F. Joseph, S. A. Bourgeois and R. K. Scheule (2004). "Long-term transgene expression from plasmid DNA gene therapy vectors is negatively affected by CpG dinucleotides." *Mol Ther* 10(2): 269-278.
61. Holkers, M., I. Maggio, S. F. Henriques, J. M. Janssen, T. Cathomen and M. A. Goncalves (2014). "Adenoviral vector DNA for accurate genome editing with engineered nucleases." *Nat Methods* 11(10): 1051-1057.

62. Howarth, J. L., Y. B. Lee and J. B. Uney (2010). "Using viral vectors as gene transfer tools (Cell Biology and Toxicology Special Issue: ETCS-UK 1 day meeting on genetic manipulation of cells)." *Cell Biol Toxicol* 26(1): 1-20.
63. Hwang, W. Y., R. T. Peterson and J. R. Yeh (2014). "Methods for targeted mutagenesis in zebrafish using TALENs." *Methods* 69(1): 76-84.
64. Irion, S., H. Luche, P. Gadue, H. J. Fehling, M. Kennedy and G. Keller (2007). "Identification and targeting of the ROSA26 locus in human embryonic stem cells." *Nat Biotechnol* 25(12): 1477-1482.
65. Ito, M., K. Yamanouchi, K. Naito, M. P. Calos and H. Tojo (2011). "Site-specific integration of transgene targeting an endogenous lox-like site in early mouse embryos." *J Appl Genet* 52(1): 89-94.
66. Ivics, Z., P. B. Hackett, R. H. Plasterk and Z. Izsvak (1997). "Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells." *Cell* 91(4): 501-510.
67. Jankele, R. and P. Svoboda (2014). "TAL effectors: tools for DNA targeting." *Brief Funct Genomics* 13(5): 409-419.
68. Joung, J. K. and J. D. Sander (2013). "TALENs: a widely applicable technology for targeted genome editing." *Nat Rev Mol Cell Biol* 14(1): 49-55.
69. Kandavelou, K. and S. Chandrasegaran (2009). "Custom-designed molecular scissors for site-specific manipulation of the plant and mammalian genomes." *Methods Mol Biol* 544: 617-636.
70. Keravala, A., S. Lee, B. Thyagarajan, E. C. Olivares, V. E. Gabrovsky, L. E. Woodard and M. P. Calos (2009). "Mutational derivatives of PhiC31 integrase with increased efficiency and specificity." *Mol Ther* 17(1): 112-120.
71. Kirat, D., K. Sallam, H. Hayashi, T. Miyasho and S. Kato (2009). "Presence of ten isoforms of monocarboxylate transporter (MCT) family in the bovine adrenal gland." *Mol Cell Endocrinol* 298(1-2): 89-100.
72. Kitts, P. A. and H. A. Nash (1987). "Homology-dependent interactions in phage lambda site-specific recombination." *Nature* 329(6137): 346-348.
73. Kleinstiver, B. P., M. S. Prew, S. Q. Tsai, V. V. Topkar, N. T. Nguyen, Z. Zheng, A. P. Gonzales, Z. Li, R. T. Peterson, J. R. Yeh, M. J. Aryee and J. K. Joung (2015). "Engineered CRISPR-Cas9 nucleases with altered PAM specificities." *Nature* 523(7561): 481-485.

74. Klug, A. (2010). "The discovery of zinc fingers and their applications in gene regulation and genome manipulation." *Annu Rev Biochem* 79: 213-231.
75. Kondo, S. and R. Ueda (2013). "Highly improved gene targeting by germline-specific Cas9 expression in *Drosophila*." *Genetics* 195(3): 715-721.
76. Landy, A. (1989). "Dynamic, structural, and regulatory aspects of lambda site-specific recombination." *Annu Rev Biochem* 58: 913-949.
77. Landy, A. (1993). "Mechanistic and structural complexity in the site-specific recombination pathways of Int and FLP." *Curr Opin Genet Dev* 3(5): 699-707.
78. Landy, A. (2015). "The lambda Integrase Site-specific Recombination Pathway." *Microbiol Spectr* 3(2).
79. Lange-Gustafson, B. J. and H. A. Nash (1984). "Purification and properties of Int-h, a variant protein involved in site-specific recombination of bacteriophage lambda." *J Biol Chem* 259(20): 12724-12732.
80. Lee, C. Y., S. L. Buranen and Z. H. Ye (1991). "Construction of single-copy integration vectors for *Staphylococcus aureus*." *Gene* 103(1): 101-105.
81. Lee, M. H., L. Pascopella, W. R. Jacobs, Jr. and G. F. Hatfull (1991). "Site-specific integration of mycobacteriophage L5: integration-proficient vectors for *Mycobacterium smegmatis*, *Mycobacterium tuberculosis*, and bacille Calmette-Guerin." *Proc Natl Acad Sci U S A* 88(8): 3111-3115.
82. Li, L., L. P. Wu and S. Chandrasegaran (1992). "Functional domains in Fok I restriction endonuclease." *Proc Natl Acad Sci U S A* 89(10): 4275-4279.
83. Li, X., W. A. Scaringe, K. A. Hill, S. Roberts, A. Mengos, D. Careri, M. T. Pinto, C. K. Kasper and S. S. Sommer (2001). "Frequency of recent retrotransposition events in the human factor IX gene." *Hum Mutat* 17(6): 511-519.
84. Lim, A. S., T. H. Lim, K. H. See, Y. J. Ng, Y. M. Tan, N. S. Choo, S. X. Lim, Y. Yee, L. C. Lau, S. L. Tien, K. Sathish and D. C. Tan (2013). "Cytogenetic and molecular aberrations of multiple myeloma patients: a single-center study in Singapore." *Chin Med J (Engl)* 126(10): 1872-1877.
85. Liu, J., T. Skjorring, T. Gjetting and T. G. Jensen (2009). "PhiC31 integrase induces a DNA damage response and chromosomal rearrangements in human adult fibroblasts." *BMC Biotechnol* 9: 31.
86. Lorbach, E., N. Christ, M. Schwikardi and P. Droge (2000). "Site-specific recombination in human cells catalyzed by phage lambda integrase mutants." *J Mol Biol* 296(5): 1175-1181.

87. Luo, Y., C. Liu, T. Cerbini, H. San, Y. Lin, G. Chen, M. S. Rao and J. Zou (2014). "Stable enhanced green fluorescent protein expression after differentiation and transplantation of reporter human induced pluripotent stem cells generated by AAVS1 transcription activator-like effector nucleases." *Stem Cells Transl Med* 3(7): 821-835.
88. Ma, Y., L. Zhang and X. Huang (2014). "Genome modification by CRISPR/Cas9." *FEBS Journal* 281(23): 5186-5193.
89. Maggio, I. and M. A. Goncalves (2015). "Genome editing at the crossroads of delivery, specificity, and fidelity." *Trends Biotechnol* 33(5): 280-291.
90. Mahfouz, M. M., A. Piatek and C. N. Stewart (2014). "Genome engineering via TALENs and CRISPR/Cas9 systems: challenges and perspectives." *Plant Biotechnology Journal* 12(8): 1006-1014.
91. Mali, P., L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville and G. M. Church (2013). "RNA-guided human genome engineering via Cas9." *Science* 339(6121): 823-826.
92. Malla, S., F. Dafhnis-Calas, J. F. Brookfield, M. C. Smith and W. R. Brown (2005). "Rearranging the centromere of the human Y chromosome with phiC31 integrase." *Nucleic Acids Res* 33(19): 6101-6113.
93. Maragathavally, K. J., J. M. Kaminski and C. J. Coates (2006). "Chimeric Mos1 and piggyBac transposases result in site-directed integration." *FASEB J* 20(11): 1880-1882.
94. Martin, S. L. (2006). "The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition." *J Biomed Biotechnol* 2006(1): 45621.
95. Mates, L., M. K. Chuah, E. Belay, B. Jerchow, N. Manoj, A. Acosta-Sanchez, D. P. Grzela, A. Schmitt, K. Becker, J. Matrai, L. Ma, E. Samara-Kuko, C. Gysemans, D. Pryputniewicz, C. Miskey, B. Fletcher, T. VandenDriessche, Z. Ivics and Z. Izsvak (2009). "Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates." *Nat Genet* 41(6): 753-761.
96. Meir, Y. J., A. Lin, M. F. Huang, J. R. Lin, M. T. Weirauch, H. C. Chou, S. J. Lin and S. C. Wu (2013). "A versatile, highly efficient, and potentially safer piggyBac transposon system for mammalian genome manipulations." *FASEB J* 27(11): 4429-4443.
97. Miller, H. I., M. A. Mozola and D. I. Friedman (1980). "int-h: An int mutation of phage lambda that enhances site-specific recombination." *Cell* 20(3): 721-729.
98. Miller, J. C., M. C. Holmes, J. Wang, D. Y. Guschin, Y. L. Lee, I. Rupniewski, C. M. Beausejour, A. J. Waite, N. S. Wang, K. A. Kim, P. D. Gregory, C. O. Pabo and E. J.

- Rebar (2007). "An improved zinc-finger nuclease architecture for highly specific genome editing." *Nat Biotechnol* 25(7): 778-785.
99. Miskey, C., Z. Izsvak, K. Kawakami and Z. Ivics (2005). "DNA transposons in vertebrate functional genomics." *Cell Mol Life Sci* 62(6): 629-641.
100. Mizuuchi, K., M. Gellert, R. A. Weisberg and H. A. Nash (1980). "Catenation and supercoiling in the products of bacteriophage lambda integrative recombination in vitro." *J Mol Biol* 141(4): 485-494.
101. Modlich, U., J. Bohne, M. Schmidt, C. von Kalle, S. Knoss, A. Schambach and C. Baum (2006). "Cell-culture assays reveal the importance of retroviral vector design for insertional genotoxicity." *Blood* 108(8): 2545-2553.
102. Montini, E., D. Cesana, M. Schmidt, F. Sanvito, M. Ponzoni, C. Bartholomae, L. Sergi Sergi, F. Benedicenti, A. Ambrosi, C. Di Serio, C. Doglioni, C. von Kalle and L. Naldini (2006). "Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration." *Nat Biotechnol* 24(6): 687-696.
103. Morbitzer, R., P. Romer, J. Boch and T. Lahaye (2010). "Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors." *Proc Natl Acad Sci U S A* 107(50): 21617-21622.
104. Morshedi, A., M. Soroush Noghabi and P. Droge (2013). "Use of UTF1 genetic control elements as iPSC reporter." *Stem Cell Rev* 9(4): 523-530.
105. Nagy, A. (2000). "Cre recombinase: The universal reagent for genome tailoring." *genesis* 26(2): 99-109.
106. Nayak, S. and R. W. Herzog (2010). "Progress and prospects: immune responses to viral vectors." *Gene Ther* 17(3): 295-304.
107. Nemudryi, A. A., K. R. Valetdinova, S. P. Medvedev and S. M. Zakian (2014). "TALEN and CRISPR/Cas Genome Editing Systems: Tools of Discovery." *Acta Naturae* 6(3): 19-40.
108. Niidome, T. and L. Huang (2002). "Gene therapy progress and prospects: nonviral vectors." *Gene Ther* 9(24): 1647-1652.
109. Nunes-Duby, S. E., M. A. Azaro and A. Landy (1995). "Swapping DNA strands and sensing homology without branch migration in lambda site-specific recombination." *Curr Biol* 5(2): 139-148.

110. Nunes-Duby, S. E., H. J. Kwon, R. S. Tirumalai, T. Ellenberger and A. Landy (1998). "Similarities and differences among 105 members of the Int family of site-specific recombinases." *Nucleic Acids Res* 26(2): 391-406.
111. Nunes-Duby, S. E., L. Matsumoto and A. Landy (1987). "Site-specific recombination intermediates trapped with suicide substrates." *Cell* 50(5): 779-788.
112. Ou, H. L., Y. Huang, L. J. Qu, M. Xu, J. B. Yan, Z. R. Ren, S. Z. Huang and Y. T. Zeng (2009). "A phiC31 integrase-mediated integration hotspot in favor of transgene expression exists in the bovine genome." *FEBS J* 276(1): 155-163.
113. Panda, S. K., B. Wefers, O. Ortiz, T. Floss, B. Schmid, C. Haass, W. Wurst and R. Kuhn (2013). "Highly efficient targeted mutagenesis in mice using TALENs." *Genetics* 195(3): 703-713.
114. Papagiannis, C. V., M. D. Sam, M. A. Abbani, D. Yoo, D. Cascio, R. T. Clubb and R. C. Johnson (2007). "Fis targets assembly of the Xis nucleoprotein filament to promote excisive recombination by phage lambda." *J Mol Biol* 367(2): 328-343.
115. Pellagatti, A., H. Dolatshad, S. Valletta and J. Boulton (2015). "Application of CRISPR/Cas9 genome editing to the study and treatment of disease." *Arch Toxicol* 89(7): 1023-1034.
116. Perez, E. E., J. Wang, J. C. Miller, Y. Jouvenot, K. A. Kim, O. Liu, N. Wang, G. Lee, V. V. Bartsevich, Y.-L. Lee, D. Y. Guschin, I. Rupniewski, A. J. Waite, C. Carpenito, R. G. Carroll, J. S. Orange, F. D. Urnov, E. J. Rebar, D. Ando, P. D. Gregory, J. L. Riley, M. C. Holmes and C. H. June (2008). "Establishment of HIV-1 resistance in CD4⁺ T cells by genome editing using zinc-finger nucleases." *Nat Biotech* 26(7): 808-816.
117. Petitclerc, D., J. Attal, M. C. Theron, M. Bearzotti, P. Bolifraud, G. Kann, M. G. Stinnakre, H. Pointu, C. Puissant and L. M. Houdebine (1995). "The effect of various introns and transcription terminators on the efficiency of expression vectors in various cultured cell lines and in the mammary gland of transgenic mice." *J Biotechnol* 40(3): 169-178.
118. Praznovszky, T. (2011). "Chromosome engineering with lambda-integrase mediated recombination system: the ACE system." *Methods in molecular biology* 738: 141-149.
119. Rajeev, L., K. Malanowska and J. F. Gardner (2009). "Challenging a paradigm: the role of DNA homology in tyrosine recombinase reactions." *Microbiol Mol Biol Rev* 73(2): 300-309.

120. Reh, W. A. and K. M. Vasquez (2001). Gene Targeting by Homologous Recombination. eLS, John Wiley & Sons, Ltd.
121. Rice, P. A., K. W. Mouw, S. P. Montano, M. R. Boocock, S. J. Rowland and W. M. Stark (2010). "Orchestrating serine resolvases." *Biochem Soc Trans* 38(2): 384-387.
122. Richet, E., P. Abcarian and H. A. Nash (1988). "Synapsis of attachment sites during lambda integrative recombination involves capture of a naked DNA by a protein-DNA complex." *Cell* 52(1): 9-17.
123. Rio, D. C. (1991). "Regulation of Drosophila P element transposition." *Trends Genet* 7(9): 282-287.
124. Roth, D. B. and J. H. Wilson (1986). "Nonhomologous recombination in mammalian cells: role for short sequence homologies in the joining reaction." *Mol Cell Biol* 6(12): 4295-4304.
125. Rothkamm, K., I. Kruger, L. H. Thompson and M. Lobrich (2003). "Pathways of DNA double-strand break repair during the mammalian cell cycle." *Mol Cell Biol* 23(16): 5706-5715.
126. Rowley, P. A., M. C. Smith, E. Younger and M. C. Smith (2008). "A motif in the C-terminal domain of phiC31 integrase controls the directionality of recombination." *Nucleic Acids Res* 36(12): 3879-3891.
127. Sadelain, M., E. P. Papapetrou and F. D. Bushman (2012). "Safe harbours for the integration of new DNA in the human genome." *Nat Rev Cancer* 12(1): 51-58.
128. Sauer, B. (1987). "Functional expression of the cre-lox site-specific recombination system in the yeast *Saccharomyces cerevisiae*." *Mol Cell Biol* 7(6): 2087-2096.
129. Savatier, P., H. Lapillonne, L. A. van Grunsven, B. B. Rudkin and J. Samarut (1996). "Withdrawal of differentiation inhibitory activity/leukemia inhibitory factor up-regulates D-type cyclins and cyclin-dependent kinase inhibitors in mouse embryonic stem cells." *Oncogene* 12(2): 309-322.
130. Schwank, G., B.-K. Koo, V. Sasselli, Johanna F. Dekkers, I. Heo, T. Demircan, N. Sasaki, S. Boymans, E. Cuppen, Cornelis K. van der Ent, Edward E. S. Nieuwenhuis, Jeffrey M. Beekman and H. Clevers "Functional Repair of CFTR by CRISPR/Cas9 in Intestinal Stem Cell Organoids of Cystic Fibrosis Patients." *Cell Stem Cell* 13(6): 653-658.
131. Sebastiano, V., M. L. Maeder, J. F. Angstman, B. Haddad, C. Khayter, D. T. Yeo, M. J. Goodwin, J. S. Hawkins, C. L. Ramirez, L. F. Z. Batista, S. E. Artandi, M. Wernig and J. K. Joung (2011). "In Situ Genetic Correction of the Sickle Cell Anemia Mutation

in Human Induced Pluripotent Stem Cells Using Engineered Zinc Finger Nucleases." *STEM CELLS* 29(11): 1717-1726.

132. Shah, S. A., S. Erdmann, F. J. Mojica and R. A. Garrett (2013). "Protospacer recognition motifs: mixed identities and functional diversity." *RNA Biol* 10(5): 891-899.

133. Shultz, J. L., E. Voziyanova, J. H. Konieczka and Y. Voziyanov (2011). "A genome-wide analysis of FRT-like sequences in the human genome." *PLoS One* 6(3): e18077.

134. Siau, J. W., S. Chee, H. Makhija, C. M. Wai, S. H. Chandra, S. Peter, P. Droge and F. J. Ghadessy (2015). "Directed evolution of lambda integrase activity and specificity by genetic derepression." *Protein Eng Des Sel*.

135. Siau, J. W., S. Chee, H. Makhija, C. M. Wai, S. H. Chandra, S. Peter, P. Droge and F. J. Ghadessy (2015). "Directed evolution of lambda integrase activity and specificity by genetic derepression." *Protein Eng Des Sel* 28(7): 211-220.

136. Silva, G., L. Poirot, R. Galetto, J. Smith, G. Montoya, P. Duchateau and F. Paques (2011). "Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy." *Curr Gene Ther* 11(1): 11-27.

137. Silver, D. P. and D. M. Livingston (2001). "Self-excising retroviral vectors encoding the Cre recombinase overcome Cre-mediated cellular toxicity." *Mol Cell* 8(1): 233-243.

138. Smith, M. C., W. R. Brown, A. R. McEwan and P. A. Rowley (2010). "Site-specific recombination by phiC31 integrase and other large serine recombinases." *Biochem Soc Trans* 38(2): 388-394.

139. Smith, M. C. and H. M. Thorpe (2002). "Diversity in the serine recombinases." *Mol Microbiol* 44(2): 299-307.

140. Sommer, D., A. Peters, A.-K. Baumgart and M. Beyer (2015). "TALEN-mediated genome engineering to generate targeted mice." *Chromosome Research* 23(1): 43-55.

141. Sommer, D., A. E. Peters, A. K. Baumgart and M. Beyer (2015). "TALEN-mediated genome engineering to generate targeted mice." *Chromosome Res* 23(1): 43-55.

142. Sorrell, D. A. and A. F. Kolb (2005). "Targeted modification of mammalian genomes." *Biotechnol Adv* 23(7-8): 431-469.

143. Tan, S. M. and P. Droge (2005). "Comparative analysis of sequence-specific DNA recombination systems in human embryonic stem cells." *Stem Cells* 23(7): 868-873.

144. Tan, S. M., S. T. Wang, H. Hentze and P. Droge (2007). "A UTF1-based selection system for stable homogeneously pluripotent human embryonic stem cell cultures." *Nucleic Acids Res* 35(18): e118.
145. Tay, F. C., W. K. Tan, S. L. Goh, C. J. Ramachandra, C. H. Lau, H. Zhu, C. Chen, S. Du, R. Z. Phang, M. Shahbazi, W. Fan and S. Wang (2013). "Targeted transgene insertion into the AAVS1 locus driven by baculoviral vector-mediated zinc finger nuclease expression in human-induced pluripotent stem cells." *J Gene Med* 15(10): 384-395.
146. Tay, Y., C. Ho, P. Droge and F. J. Ghadessy (2010). "Selection of bacteriophage lambda integrases with altered recombination specificity by in vitro compartmentalization." *Nucleic Acids Res* 38(4): e25.
147. Tay, Y., C. Ho, P. Droge and F. J. Ghadessy (2010). "Selection of bacteriophage lambda integrases with altered recombination specificity by in vitro compartmentalization." *Nucleic acids research* 38(4): e25.
148. Tay, Y., C. Ho, P. Dröge and F. J. Ghadessy (2010). "Selection of bacteriophage λ integrases with altered recombination specificity by in vitro compartmentalization." *Nucleic Acids Research* 38(4): e25.
149. Tey, S. K. (2014). "Adoptive T-cell therapy: adverse events and safety switches." *Clin Transl Immunology* 3(6): e17.
150. Thaci, B., I. V. Ulasov, D. A. Wainwright and M. S. Lesniak (2011). "The challenge for gene therapy: innate immune response to adenoviruses." *Oncotarget* 2(3): 113-121.
151. Thomas, C. E., A. Ehrhardt and M. A. Kay (2003). "Progress and problems with the use of viral vectors for gene therapy." *Nat Rev Genet* 4(5): 346-358.
152. Thomas, K. R. and M. R. Capecchi (1987). "Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells." *Cell* 51(3): 503-512.
153. Thompson, J. F., L. Moitoso de Vargas, C. Koch, R. Kahmann and A. Landy (1987). "Cellular factors couple recombination with growth phase: characterization of a new component in the lambda site-specific recombination pathway." *Cell* 50(6): 901-908.
154. Thorpe, H. M. and M. C. Smith (1998). "In vitro site-specific integration of bacteriophage DNA catalyzed by a recombinase of the resolvase/invertase family." *Proc Natl Acad Sci U S A* 95(10): 5505-5510.

155. Thyagarajan, B., M. J. Guimaraes, A. C. Groth and M. P. Calos (2000). "Mammalian genomes contain active recombinase recognition sites." *Gene* 244(1-2): 47-54.
156. Thyagarajan, B., E. C. Olivares, R. P. Hollis, D. S. Ginsburg and M. P. Calos (2001). "Site-specific genomic integration in mammalian cells mediated by phage phiC31 integrase." *Mol Cell Biol* 21(12): 3926-3934.
157. Tong, C., G. Huang, C. Ashton, P. Li and Q.-L. Ying (2011). "Generating gene knockout rats by homologous recombination in embryonic stem cells." *Nat. Protocols* 6(6): 827-844.
158. Trinh, K. R. and S. L. Morrison (2000). "Site-specific and directional gene replacement mediated by Cre recombinase." *J Immunol Methods* 244(1-2): 185-193.
159. Turan, S. and J. Bode (2011). "Site-specific recombinases: from tag-and-target- to tag-and-exchange-based genomic modifications." *FASEB J* 25(12): 4088-4107.
160. Turan, S., C. Zehe, J. Kuehle, J. Qiao and J. Bode (2013). "Recombinase-mediated cassette exchange (RMCE) - a rapidly-expanding toolbox for targeted genomic modifications." *Gene* 515(1): 1-27.
161. Urnov, F. D., E. J. Rebar, M. C. Holmes, H. S. Zhang and P. D. Gregory (2010). "Genome editing with engineered zinc finger nucleases." *Nat Rev Genet* 11(9): 636-646.
162. van Gent, D. C., J. H. Hoeijmakers and R. Kanaar (2001). "Chromosomal stability and the DNA double-stranded break connection." *Nat Rev Genet* 2(3): 196-206.
163. VandenDriessche, T., Z. Ivics, Z. Izsvak and M. K. Chuah (2009). "Emerging potential of transposons for gene therapy and generation of induced pluripotent stem cells." *Blood* 114(8): 1461-1468.
164. Vannucci, L., M. Lai, F. Chiappesi, L. Ceccherini-Nelli and M. Pistello (2013). "Viral vectors: a look back and ahead on gene transfer technology." *New Microbiol* 36(1): 1-22.
165. Vasquez, K. M., K. Marburger, Z. Intody and J. H. Wilson (2001). "Manipulating the mammalian genome by homologous recombination." *Proc Natl Acad Sci U S A* 98(15): 8403-8410.
166. Voigt, K., A. Gogol-Doring, C. Miskey, W. Chen, T. Cathomen, Z. Izsvak and Z. Ivics (2012). "Retargeting sleeping beauty transposon insertions by engineered zinc finger DNA-binding domains." *Mol Ther* 20(10): 1852-1862.

167. Wach, A., A. Brachat, R. Pohlmann and P. Philippsen (1994). "New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*." *Yeast* 10(13): 1793-1808.
168. Wen, S., H. Zhang, Y. Li, N. Wang, W. Zhang, K. Yang, N. Wu, X. Chen, F. Deng, Z. Liao, J. Zhang, Q. Zhang, Z. Yan, W. Liu, Z. Zhang, J. Ye, Y. Deng, G. Zhou, H. H. Luu, R. C. Haydon, L. L. Shi, T. C. He and G. Wei (2014). "Characterization of constitutive promoters for piggyBac transposon-mediated stable transgene expression in mesenchymal stem cells (MSCs)." *PLoS One* 9(4): e94397.
169. Wijshake, T., D. J. Baker and B. van de Sluis (2014). "Endonucleases: new tools to edit the mouse genome." *Biochim Biophys Acta* 1842(10): 1942-1950.
170. Wilson, M. H., J. M. Kaminski and A. L. George, Jr. (2005). "Functional zinc finger/sleeping beauty transposase chimeras exhibit attenuated overproduction inhibition." *FEBS Lett* 579(27): 6205-6209.
171. Wood, A. J., T. W. Lo, B. Zeitler, C. S. Pickle, E. J. Ralston, A. H. Lee, R. Amora, J. C. Miller, E. Leung, X. Meng, L. Zhang, E. J. Rebar, P. D. Gregory, F. D. Urnov and B. J. Meyer (2011). "Targeted genome editing across species using ZFNs and TALENs." *Science* 333(6040): 307.
172. Wu, S. C., Y. J. Meir, C. J. Coates, A. M. Handler, P. Pelczar, S. Moisyadi and J. M. Kaminski (2006). "piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells." *Proc Natl Acad Sci U S A* 103(41): 15008-15013.
173. Wu, Y., T. Gao, X. Wang, Y. Hu, X. Hu, Z. Hu, J. Pang, Z. Li, J. Xue, M. Feng, L. Wu and D. Liang (2014). "TALE nickase mediates high efficient targeted transgene integration at the human multi-copy ribosomal DNA locus." *Biochem Biophys Res Commun* 446(1): 261-266.
174. Wu, Y., D. Liang, Y. Wang, M. Bai, W. Tang, S. Bao, Z. Yan, D. Li and J. Li (2013). "Correction of a Genetic Disease in Mouse via Use of CRISPR-Cas9." *Cell Stem Cell* 13(6): 659-662.
175. Wu, Z., R. I. Gumport and J. F. Gardner (1997). "Genetic analysis of second-site revertants of bacteriophage lambda integrase mutants." *J Bacteriol* 179(12): 4030-4038.
176. Wurtele, H., K. C. Little and P. Chartrand (2003). "Illegitimate DNA integration in mammalian cells." *Gene Ther* 10(21): 1791-1799.
177. Yang, L., P. Mali, C. Kim-Kiselak and G. Church (2014). "CRISPR-Cas-mediated targeted genome editing in human cells." *Methods Mol Biol* 1114: 245-267.

178. Yang, W. and K. Mizuuchi (1997). "Site-specific recombination in plane view." *Structure* 5(11): 1401-1406.
179. Yant, S. R., Y. Huang, B. Akache and M. A. Kay (2007). "Site-directed transposon integration in human cells." *Nucleic Acids Res* 35(7): e50.
180. Yew, N. S., H. Zhao, M. Przybylska, I. H. Wu, J. D. Tousignant, R. K. Scheule and S. H. Cheng (2002). "CpG-depleted plasmid DNA vectors with enhanced safety and long-term gene expression in vivo." *Mol Ther* 5(6): 731-738.
181. Yi, Y., M. J. Noh and K. H. Lee (2011). "Current advances in retroviral gene therapy." *Curr Gene Ther* 11(3): 218-228.
182. Yusa, K., R. Rad, J. Takeda and A. Bradley (2009). "Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon." *Nat Methods* 6(5): 363-369.
183. Yusa, K., S. T. Rashid, H. Strick-Marchand, I. Varela, P.-Q. Liu, D. E. Paschon, E. Miranda, A. Ordonez, N. R. F. Hannan, F. J. Rouhani, S. Darche, G. Alexander, S. J. Marciniak, N. Fusaki, M. Hasegawa, M. C. Holmes, J. P. Di Santo, D. A. Lomas, A. Bradley and L. Vallier (2011). "Targeted gene correction of [agr]1-antitrypsin deficiency in induced pluripotent stem cells." *Nature* 478(7369): 391-394.
184. Zheng, L., U. Baumann and J. L. Reymond (2004). "An efficient one-step site-directed and site-saturation mutagenesis protocol." *Nucleic Acids Res* 32(14): e115.
185. Zou, J., C. L. Sweeney, B. K. Chou, U. Choi, J. Pan, H. Wang, S. N. Dowey, L. Cheng and H. L. Malech (2011). "Oxidase-deficient neutrophils from X-linked chronic granulomatous disease iPS cells: functional correction by zinc finger nuclease-mediated safe harbor targeting." *Blood* 117(21): 5561-5572.

APPENDIX

Figure A1 Alignment of attH4X sequence. Alignment of 935 *LINE-1* elements containing 18bp attH4X sequence flanked by 300bp upstream and downstream sequences (Vector NTI , AlignX) to design the primer pairs (cs_ attH4X _F1 & cs_ attH4X _R1) and (cs_ attH4X _F2 & cs_ attH4X _R2, not shown) in the consensus flanking region of attH4X. The respective subfamilies of 935 LINE1 elements that are shown in the right column correspond to the serial numbers of the sequences on the left side of the alignment

Figure A2. Western Blot analysis for pluripotency markers Oct-4, Nanog and Sox2 in the targeted hESC clones and differentiated derivatives.

Supplementary Figure S2. Western Blot analysis for pluripotency markers Oct-4, Nanog and Sox2 in the targeted hESC clones and differentiated derivatives. Cell lysates from undifferentiated parental hESCs, the targeted clones (hESC#3, hESC#24, hESC#59 and hESC clone E3) and their respective RA-induced differentiated derivatives were subjected to Western blot analysis and probed with antibodies against Oct-4, Nanog and Sox2 (top panels of a, b and c, respectively). Oct-4, Nanog and Sox2 expression was detected only in hESC clones. β -actin protein was used as loading control in each case (bottom panels of a, b and c). U, undifferentiated; D, differentiated; HT, cell lysate of HT1080 as control for differentiated cells; M, Marker lane; P-16, passage 16; P-32, passage 32 (contributed by Dr. Shree Harsha)

Page intentionally left blank

PUBLICATIONS AND PATENT

- Shree Harsha Vijaya Chandra*, **Harshyaa Makhija***, Sabrina Peter, Cho Mar Myint Wai, Jinming Li, Jindong Zhu, Zhonglu Ren, Martina Stagno D'Alcontres, Siau Jia Wei, Sharon Chee, Farid John Ghadessy & Peter Dröge. Conservative site-specific and single-copy transgenesis in human LINE-1 elements. *Nucleic Acid Research*, (2015)
* equal contribution
- Jia Wei Siau, Sharon Chee, **Harshyaa Makhija**, Cho Mar Myint Wai, Shree Harsha Vijaya Chandra, Sabrina Peter, Peter Dröge and Farid J. Ghadessy . Directed evolution of lambda integrase activity and specificity by genetic derepression. *Protein Engineering, Design and Selection*, (2015).
- UK Patent (Application No. 1414130.3) filed for the technology “Mutants of the Bacteriophage Lambda Integrase”. Co-inventors; Peter Dröge, Shree Harsha Vijaya Chandra, **Harshyaa Makhija**, Sabrina Peter, Siau Jia Wei, Farid J. Ghadessy.

Page intentionally left blank

CONFERENCE PRESENTATION

Oral presentation

- Investor Pitch at “World Stem cells and Regenerative Medicine Congress”, organized by Terrapinn on May 20-22, 2015 in London, UK.

Title: Genome manipulation using λ -integrase mediated site-specific insertion tool

- Talk at Bioprocessing Technology Institute, A*STAR, Singapore.

Title: Site-specific transgenesis in human LINE-1 elements

- Panel discussion at Biotech Connection Singapore and Joyful Frog Digital Incubator, Singapore.

Title: Genome editing

- Startup Pitch at OneStart Europe Biotech Business Bootcamp organized by Oxbridge Roundtable Biotech on February 6-7, 2015 in London, UK.

Title: Genome editing (gene λ^i)

- Talk at 6th Annual Student Retreat organized by Konstanz Research School Chemical Biology on August 13-15, 2014 in Germany.

Title: Human genome editing using λ integrase-mediated site-specific gene insertion

Poster presentation

- 12th Annual Meeting organized by ISSCR on June 18-21, 2014 in Vancouver, BC Canada.

Title: Human genome editing using λ integrase-mediated site-specific gene insertion

Page intentionally left blank