Targeted transposition of the *Sleeping Beauty* transposon using zinc finger proteins

Dissertation zur Erlangung des akademischen Grades des

Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie

der Freien Universität Berlin

vorgelegt von Diplom-Biologin

Katrin Voigt

geboren in Berlin

2010

Die vorliegende Arbeit wurde in der Zeit von Februar 2005 bis Mai 2010 im Labor von Dr. Zoltán Ivics am Max-Delbrück-Centrum für Molekulare Medizin in Berlin Buch angefertigt.

1. Gutachter: Professor Dr. Udo Heinemann

2. Gutachter: Professor Dr. Reinhard Kunze

Disputation am: 02.11.2010

Table of contents

1. Introduction

1.1. Gene therapy

Gene therapy can be described as the use of genes as medicine. Inherited primary immunodeficiencies have become a center field of clinical gene therapy. Treatment of choice in these genetic diseases is haemapoetic stem cell transplantation from a human leucocyte antigen-matched donor. If no compatible donor is found gene therapy can be considered. In inherited diseases introducing a functional copy of the defective gene will improve the disease phenotype or even fully cure the patient. In diseases where a defective gene product acts dominantly over the correct version inactivation or knock-out of the defective gene is desired. In the case of cancer either healthy cells can be targeted to enhance their ability to fight malignant cells or cancer cells can be genetically modified to destroy themselves or prevent their growth. Some cell types like circulating blood cells can be extracted rather easily from patients, manipulated outside the individual (*ex vivo*) and transferred back into the patient´s blood. Other cell types like cells from solid organs like the liver can not be withdrawn and reinserted so easily. Here vectors need to be used that deliver their genetic cargo specifically to the desired cell type and insertion of the transgene into the genome occurs within the patient (*in vivo*). So far only somatic cells are targeted in human gene therapy, so that genetic modifications by gene therapy are not transmitted to the progeny. For successful gene therapy the therapeutic transgene has to be delivered efficiently to the cells of concern. Integration at a suitable site in the genome should provide long-term gene expression at a suitable level. Regulatory or promoter elements delivered with the transgene should not affect expression of endogenes.

Before a new therapy receives permission to be applied generally clinical trials of different phases have to be conducted. Preclinical studies involve *in vitro* studies in test-tube and *in vivo* experiments on laboratory animals to examine efficacy and toxicity. In Phase I trials about 5-80 healthy individuals or patients with poor prognosis are treated to gain information about safety and tolerability of the therapy. For gene therapy fate of genetically modified cells in the patient and expression levels of the therapeutic transgene over time are examined. Phase I trials mainly focus on the safety of the treatment, not on curing the disease. Phase II trials evaluate efficacy of the new gene therapy. About 20-300 patients are enrolled in phase II trials. In Phase III trials a bigger group of patients (300-3,000) partake. Therapeutic achievements of the new therapy are compared to the gold-standard treatment for this disease

at the time. It is monitored whether the new treatment is more effective or has less side effects than the standard treatment. Phase IV trials are conducted post-permission. They examine interactions of the new drug with other drugs, explore the use of the new drug for other markets and monitor tolerance and safety for groups of patients yet not tested such as pregnant woman or children and monitor therapy achievements over a longer time period.

1.2. Delivery vectors in gene therapy

For successful gene therapy therapeutic transgenes have to be delivered to their target cell and inserted into the genome. Currently viral vectors are predominantly used in clinical trials. About 24 % of clinical trials conducted to date use adenoviruses for gene delivery (http://www.wiley.co.uk/genmed/clinical/). Adenoviruses can transfect dividing as well as undividing cells, however their genetic cargo is generally not integrated into the genome. Therefore transgene expression is only transient, making it necessary to readminister the vector in dividing cells. Most adenovirus serotypes enter cells by binding to the coxsackievirus adenovirus receptor (CAR). Cells expressing only little amounts of this protein like the endothelium, brain tissue, hemapoetic cells or primary tumors are thus hard to infect with adenovirus. In order to be able to target other cell types adenovirus variants with altered tropism have been developed (Mizuguchi and Hayakawa 2002). In 1999 one patient suffering from partial deficiency of ornithine transcarbamylase died from a systemic adenovirus vector-induced shock syndrome after receiving gene therapy treatment. A high dosage of the adenoviral vector, delivered by infusion directly to the liver, probably saturated CAR receptors in the liver and spilled to circulary system as well as other organs and the bone marrow inducing the systemic immune response. The adenoviral capsid proteins elicited a humoral immune response that results in the generation of anti-Adenovirus antibodies (NIH report 2002). Other patients receiving low dosage gene therapy treatment with the same vector did not develop such fatal immune response. Adeno-associated viruses (AAV) are also able to infect dividing as well as undividing cells. In the absence of a so-called helper virus such as adenovirus or herpesvirus AAV integrates into the genome rather site-specifically into the *AAVS1* locus on chromosome 19. For site-specific integration a virus-encoded replication protein (rep) is needed. In the absence of rep viral DNA is integrated in a random fashion. AAV vectors used in gene therapy are devoid of any viral proteins. Hence for site-specific integration the rep protein has to be provided in *trans*. Persistance of the rep protein however leads to chromosomal instability and remobilisation of the transgene (Young and Samulski

2001). Though AAV vectors are generally less immunogenic than adenoviral vectors, preexisting anti-AAV antibodies that could neutralize AAV gene therapy vectors can be a problem. A solution to this problem could be the identification of immunogenic capsid domains and development of AAV vectors with modified capsid structure. About 21% of gene therapy clinical trials conducted so far have used gamma-retroviral or lentiviral vectors based on the murine leukemia virus (MLV), the avian sarcoma-leucosis virus (ASLV), or the human immunodeficiency virus (HIV). These vector systems are very efficient in gene delivery and in providing sustained expression of the transgene in dividing cells. However some problems are connected to their use in gene therapy like recombination of the virus *in vivo*, immunological complication (Follenzi et al. 2007) and insertional mutagenesis. Possible mutagenic consequences can arise through a transgene insertion into an exon of a gene resulting in gene truncation, insertion into intronic regions where a vector encoded enhancer could lead to upregulation of the endogenous gene and/or ectopic expression or insertion into the upstream regulatory sequence also resulting in upregulation of the endogenous gene and/or ectopic expression.

MLV has been shown to have a strong tendency to insert into transcription start sites of genes (Wu et al. 2003). HIV has a bias towards insertion into transcription units (Schroder et al. 2002). ASLV shows the weakest bias towards integration into active genes in this group but still at a frequency higher than that of random integration (Mitchell et al. 2004). Abnormal expression patterns can have devastating consequences for the cell including the development of cancers. This was the case in gene therapy trials held in Paris and London where patient suffering from X-linked severe combined immunodeficiency (SCID-X1) could be cured upon *ex vivo* transfer of a gene construct encoding the γ chain of the common cytokine receptor (γc) into autologous cluster of differentiation (CD)34+ bone marrow cells by replication deficient MLV vector (Hacein-Bey-Abina et al. 2002). Unfortunately some years later some of theses patients developed T-cell leukemia. In these patients the transgene inserted close to the LIM domain only 2 (*LMO2*) gene (Hacein-Bey-Abina et al. 2003). Retrovirus-driven enhancer activity on the *LMO2* promotor lead to deregulated cell proliferation. Despite these tragic events it is to say that a number of patients who received γ-retroviral vector-mediated gene therapy treatment profit from improvement or even cure of their disease. This shows that one crucial point in successful gene therapy is the insertion of the therapeutic transgene at a safe site in the genome that ensures sustained expression but does not alter expression of genes adjacent to the integration site. Even though highly efficient in gene delivery viral vectors can elicit immune responses of the innate or acquiered immune system which can lead to

elimination of transduced cells and to acute systemic toxicity. Their tendency for integration into transcribed regions or regulatory elements bears the risk of altered expression of endogenous genes which can result in the formation of cancer. Furthermore production of viral vectors is elaborate and costly. Transposable elements (TE) could offer an alternative to viral-based gene delivery systems. They are relatively easy to produce and show, depending on the transposon system used, near random integration pattern in the human genome. Recent development of variants with improved performance results in integration efficiencies similar to that of viral vector systems (Mates et al. 2009) (Cadinanos and Bradley 2007).

1.3. Discovery and history of transposable elements

TEs were first discovered by the cytogeneticist Barbara McClintock in the 1940s. Studying the breakage-fusion-bridge cycles in maize, she observed frequent chromosome breakage at characteristic sites. She called the element located at these sites *Dissociator* (*Ds*). She discovered that *Ds* not only caused the chromosomes to break but was also able to move from one site in the genome to another. Movement of *Ds* however only occurred in the presence of another element which she termed *Activator* (*Ac*). *Ac* by itself was also able to move within the genome (McClintock 1946; McClintock 1947; McClintock 1948). Since *Ds* and *Ac* were able to modify gene expression as a result of insertion near or within genes Barbara McClintock named them “controlling elements”. The idea of mobile DNA was not valued at all at this time when scientist saw genes as static structures in the nucleus. It took until the 1970s when molecular biologists could verify McClintock´s genetic insights that her discovery of mobile DNA was fully accepted and honoured by awarding her the Nobel Prize in physiology and medicine in 1983. Even though the concept of mobile DNA was finally accepted, its role in or potential for the genome was still not appreciated as terms like “junk DNA” (Ohno 1972) or “selfish DNA” (Orgel and Crick 1980) disclose. TEs were believed to serve no other purpose than to propagate themselves. No significance for the host organism was presumed and transposon research was limited to the discovery of new elements. This changed in the 1990s when the mechanism of transposition, the impact of transposition on the host genome and usage of transposon components for cellular functions came into focus of researchers. In addition, transposons started to be utilised as tools in functional genomics. Today the application fields for TEs cover a wide range of purposes. TEs are nowadays routinely used for transgenesis or insertional mutagenesis in invertebrates like *Caenorhabditis elegans* (Bazopoulou and Tavernarakis 2009) or *Drosophila* (Thibault et al. 2004) and

different vertebrates like mouse (Dupuy et al. 2005), fish (Kawakami et al. 2000), frog (Hamlet et al. 2006) or rat (Lu et al. 2007). Recently, the first gene therapy clinical trial using a transposon as a gene delivery tool in humans was launched (Williams 2008).

1.4. Classification of transposable elements

About 45 % of the human genome consists of TEs (Lander et al. 2001); most of them are inactive due to mutations or deletions. TEs have contributed to the cell´s protein repertoire. In humans at least 4 % of protein coding regions are believed to contain TEs or at least parts of them (Nekrutenko and Li 2001). In most of such cases the TE inserted into a non-coding region like an intron and was then recruited as an additional exon through splicing. In an evolutionary process called "molecular domestication" an enzymatic component of a TE is adopted by the host and applied to serve the cell´s needs (Sinzelle et al. 2009). In humans prominent examples for "domesticated" transposon components are the recombination-activating gene product RAG1 which performs immunoglobulin and T-cell receptor (TCR) gene recombination or centromere protein B (CENP-B) a protein involved in centromeric chromatin assembly.

TEs can be divided into two main classes according to their mechanism of transposition. The majority of TEs, about 42 %, belongs to the class I or retrotransposons. Transposition of class I retrotransposons is performed by a "copy and paste" mechanism which requires an RNA intermediate. The RNA intermediate is reverse transcribed into a cDNA which is then integrated at a new donor site as double stranded cDNA. "Copy and paste" transposition leaves the TE at the donor site intact and leads to insertion of an additional copy of the TE at a new target site. About 3 % of the human genome comprises of class II or DNA transposons which use a "cut and paste" mechanism for transposition. The transposon is hereby removed from the donor site before integrating at the target site. For both classes autonomous and non-autonomous elements exist, the later depend on the enzymatic machinery of autonomous elements for transposition. Class I retrotransposons can be further divided into long terminal repeat (LTR) and non-LTR retrotransposons (Fig. 1). In all LTR retrotransposons, the inner coding region of the element is flanked by long terminal direct repeats which carry transcriptional regulatory elements. Autonomous LTR retrotransposons contain group-specific-antigen (*gag*) and polymerase (*pol*) genes in the inner region which encode proteins necessary for retrotransposition: a protease (PR), reverse transcriptase (RT), RNAse H and an integrase (IN). The two main groups of LTR retrotransposons are named *Ty1/copia* and

Ty3/*gypsy* after their first members identified in yeast and *Drosophila*. Some elements additionally contain envelope (*env*)-like genes. *Env* genes are used by retroviruses to infect other cells. One example for a LTR retrotransposon containing *env*-like gene is the *gypsy* element which has been shown to be infectious (Song et al. 1994) (Kim et al. 1994). Within class I transposons the most abundant TEs, comprising 34 % of the human genome, are non-LTR retrotransposons. Long interspersed nuclear elements (LINEs) represent the autonomous form, whereas short interspersed nuclear elements (SINEs) are devoid of intact open reading frames (ORFs) and depend on the enzymatic machinery of LINEs for transposition. From the three distantly related LINE families (LINE1-3) found in the human genome only LINE1, also referred to as L1 or L1H, is still active. Five different families of LINE1 exist in the human genome (L1PA1-5). All of these families except L1PA1 are inactive today due to 5′-truncations or debilitating mutations. The L1PA1 family is therefore also called Ta family for transcriptionally active (Skowronski et al. 1988) and can be further classified into subfamilies Ta-0 and Ta-1. Ta-1 seems to be the youngest subfamily with the most elements that are still able to retrotranspose. Ta-1 is further subdivided into Ta-1d and Ta-1nd subsets which have a number of sequence differences including a deletion in the 5′-untranslated region (UTR) (Boissinot et al. 2000). Today the Ta-1d seems to be the subset that is most active within all LINE1 elements. Full length LINE1s are about 6 kb in size, harbour an internal polymerase II promoter and two ORFs which are both indispensible for retrotransposition. ORF1 encodes a protein of yet unknown function with RNA binding capacity and ORF2 codes for a protein which acts as reverse transcriptase (RT) and endonulease (EN). LINE1 elements contain a promoter in their 5′-UTR which initiates transcription by RNA–polymerase II. Upon translation proteins attach to the LINE1 derived mRNA. The endonuclease cuts the DNA at a new target site often at an oligo(dT) sequence creating a free 3′-OH which serves as primer during reverse transcription. The LINE1 mRNA anneals to the oligo(d)T sequence and serves as template during reverse transcription. Often the process of reverse transcription is terminated early leading to 5′-truncation of the new LINE1 copy. Hence for classification of a LINE1 element mainly the 3′-region of ORF2 or the 3′-UTR of the element is used. LINEs are widely distributed throughout the human genome. The X chromosome is particulary packed with LINE elements whereas chromosomes 17, 19 and especially 22 show a lesser extent of LINE1 insertions. LINE elements are found preferably (about fourfold enriched (Lander et al. 2001)) in AT-rich regions. Targeting gene-poor, AT-rich regions may prevent fatal impact on the host genome and help the propagation the mobile element. Nevertheless LINE1 insertions are cause of genetic defects. About 14 human diseases are known which

arose by reason of a LINE1 insertion. Examples are integration of LINE1 into the factor VIII gene leading to development of haemophilia A (Kazazian et al. 1988) and disruption of the adenomatosis polyposis coli (*APC*) gene found in a colon cancer (Miki et al. 1992). One in a thousand new spontaneous mutations in humans is caused by a LINE insertion (Kazazian and Moran 1998).

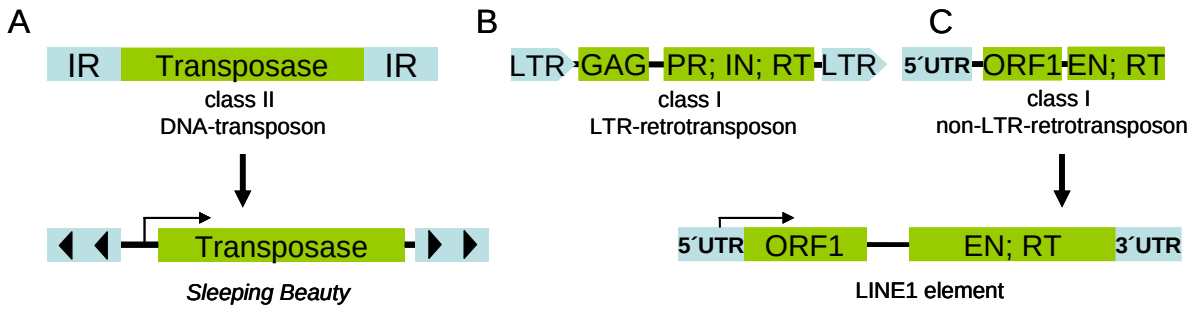


Figure 1. Schematic structure of main transposon types in eukaryotes. (A) Class II or DNA transposons contain two (left and right) inverted repeats (L-IR and R-IR respectively) flanking an ORF encoding the enzymatic component of the transposon, the transposase. Below, the *Sleeping Beauty* transposon is shown as an example of a class II transposon from the Tc1/*mariner* family. Black arrowheads represent transposase binding sites within the IRs. **(B)** LTR-retrotransposons belong to class I of TEs. Two LTRs in direct orientation flank two ORFs; ORF1 encodes the group-specific antigen (gag); ORF2 encodes a protease (PR), an integrase (IN) and a reverse transcriptase (RT). **(C)** Class I non-LTR retrotransposons contain two ORFs; ORF1 encodes a DNA-binding protein and ORF2 encodes an endonuclease (EN) and an RT. Below, a LINE1 element, a well-studied class I non-LTR-retrotransposon is depicted. The 5´-UTR contains a promotor which drives transcription of ORF1 and ORF2. A polyadenylation sequence (pA) is found in the 3´-UTR.

Class II transposons can be divided into different groups. The simplest elements are the so-called insertion sequences (IS). IS elements contain a single open reading frame coding for the enzymatic component the transposase which is flanked by two terminal inverted repeats (IRs). Specific sites in the IRs are recognized by its transposase protein which upon binding excises the transposon and inserts it elsewhere in the genome. IS elements typically also encode a second ORF encoding a regulatory protein that can either enhance or inhibit transposition. IS elements are predominantly found in prokaryotes; however, some IS elements also exist in eukaryotes. Though inactive class II transposon fossils make up about 3% of the human genome (Lander et al. 2001) no active copy has been found in humans until present. In 1997 Ivics et al. (1997) resurrected a Tc1/*mariner* type transposon from inactive transposon copies in fish which showed to be active in a number of vertebrate cells including human. Class II transposons have been reclassified several times. Currently two subclasses based on generation of single or double stranded cuts at the DNA donor site during transposition are distinguished. Elements from the first subclass who´s transposition requires double stranded cuts at the DNA donor site can be further subdivided by sequence motifs

within their IR and length of their target site duplications. Target site duplications are generated by duplication of short host-derived DNA sequences flanking the transposon after insertion. A complete list of class II transposon subclasses and families and their features can be viewed in (Sinzelle et al. 2009) (Feschotte and Pritham 2007). Transposon systems frequently used as tools for vertebrate genomics are the *Sleeping Beauty* (*SB*) transposon system resurrected from salmoid fish (Ivics et al. 1997), *piggyBac* (*PB*) from cabbage looper moth (Ding et al. 2005) and *Tol2* from medaka (Koga et al. 1996). These three elements belong to the Tc1/*mariner*, *piggyBac* and *hAT* superfamilies respectively.

1.5. A brief introduction to transposons in prokaryotes

Apart from the IS elements introduced earlier which represent a simple form of TEs more complex transposons are found in bacteria (Fig. 2). Two major classes of mobile elements exist that are distinguished by their mode of transposition. In recent years more diverse types of transposons have been found see a recent more comprehensive classification in Roberts et al. (2008). Prokaryotic transposons that belong to the prokaryotic class I typically encode antibiotic resistance genes, a transposase and a regulatory protein that can control transposition activity flanked by short IRs at both ends. A well studied example for a class I TE is the *Tn3* transposon. The *Tn7* transposon mentioned later in this work also belongs to class I. Prokaryotic transposons that belong to class II contain various antibiotic resistance genes flanked by two variants of the an IS element. In the well-studied *Tn5* transposon two *IS50* elements termed *IS50L* and *IS50R*, for left and right, respectively, flank the antibiotic resistance genes to kanamycin (kan), bleomycin and streptomycin. *IS50R* provides the transposon with a functional transposase protein as well as a regulatory protein. *IS50L* encodes two inactive truncated versions of transposase and inhibitor because of a DNA mutation causing a premature stop codon. Because of their assembled structure these elements are also called composite transposons. Note that definition of class I and class II is defined differently for prokaryotic and eukaryotic TEs. Bacteriophages represent a third class of mobile elements in bacteria. They contain genes for integration and replication, genes responsible for cell lysis and genes encoding phage coat proteins flanked by terminal IRs. Well-known examples from this class are bacteriophage Mu and ΦC31.

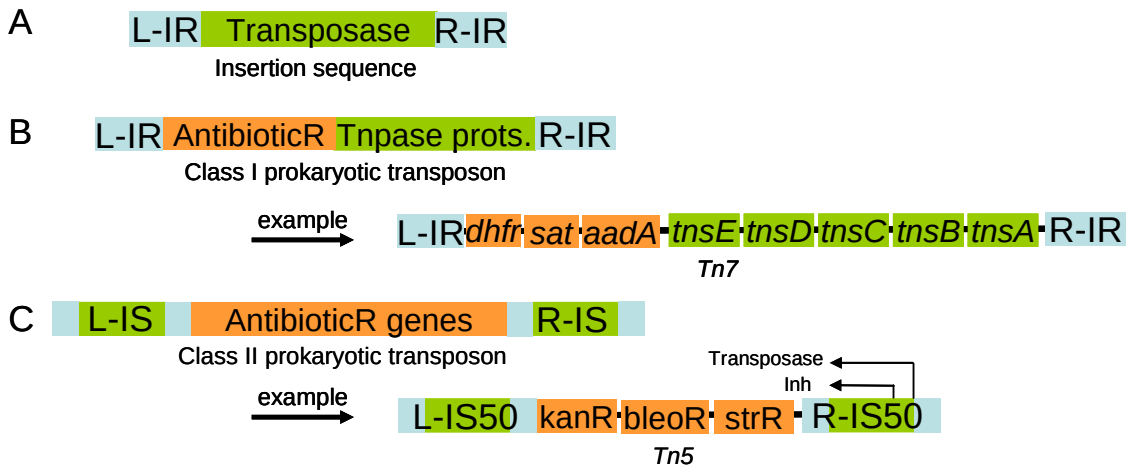


Figure 2. Schematic structure of prokaryotic TEs. (A) Depicted is a IS with a single ORF encoding a transposase flanked by two IRs. **(B)** Prokaryotic class I TEs encode a tranposase protein along with other regulatory proteins and antibiotic resistance genes flanked by two short IRs. An example for a prokaryotic class I transposon is shown on the right. The *Tn7* transposon contains three anibiotic resistance genes: *dhfr*, *sat* and *aadA* which confer antibiotic resistance to trimethoprim, streptothricin and streptomycin or spectinomycin respectively. The transposition proteins TnsA and TnsB form a transposase protein that binds to *Tn7*. TnsC, TnsD and TnsE interact with target DNA. *Tn7* is flanked by two short IRs. **(C)** Class II or composite transposons carry antibiotic resistance genes flanked by two variants of the same IS. On the right as an example the *Tn5* transposon is depicted. Resistance genes to kanamycin, bleomycin and streptomycine are flanked by two *IS50* elements. Due to a premature stop codon proteins deriving from the left *IS50* element are truncated and non-functional. Transposase and inhibitor are provided by the right *IS50* element.

1.6. The *Sleeping Beauty* transposon

Until the discovery of the *Tol2* transposon in medaka in 1996 no active class II transposon was known in vertebrates (Koga et al. 1996). Envisioning the potential power of class II transposons as molecular tools in vertebrate genetics, Ivics *et al*. resurrected an active transposon from ancient transposon fossils in salmonid fish (Ivics et al. 1997). This revived transposon was called *Sleeping Beauty* since it was "kissed" alive after a long inactive "sleep". As described above the *SB* transposon belongs to the Tc1/*mariner* superfamily. Members of this superfamily are widespread throughout different kingdoms, other examples include the name-giving *Tc1* from *Caenorhabditis elegans* or the *mariner* element from *Drosophila*. They are found in protozoa as well as in fungi, nematodes, arthropodes and chordates including fish and human. Until discovery of *Passport* in flatfish (Clark et al. 2009) no active member of this superfamily had been identified in vertebrates (reviewed in (Miskey et al. 2005)). The originally resurrected *SB* transposon is 1639 bp in length with one terminal IR of about 230 bp at each end. Each IR contains two imperfect direct repeats (DR) of about 32 bp. One DR is found directly on the outer end of the IR the other 165-166 bp inwards the transposon. Each DR contains a core transposase-binding sequence resulting in four transposase binding sites per transposon, two within each IR. Sequences within the transposon

adjacent to these core transposase-binding sites vary in sequence and also contribute to transposase binding. DRs from the different positions in the IRs can not be exchanged without loss of transpositional activity (Cui et al. 2002). Between the two IRs the enzymatic component of the *SB* transposon, the transposase, is encoded in a 1023 bp long ORF. The N-terminus of SB transposase mediates DNA binding and interaction with transposon IRs whereas the catalytic domain responsible for cleavage and joining reactions is located at the C-terminus of the protein (Fig. 3B). For active transposition both functional domains are required. The N-terminus of the transposase contains two helix-turn-helix domains resembling the paired domain of Pax proteins. The pai subdomain is believed to be responsible for contacting the DRs, binding of enhancer-like sequences within the left IR as well as for protein-protein interaction via a leucine zipper between transposase subunits. Between the pai and red subdomain lies a GRRR-motif (AT hook) which contributes to transposon binding. The function of the red domain is still not entirely clear but is believed to also contribute to DNA binding (Izsvak et al. 2002). The red subdomain overlaps with a nuclear localisation signal (NLS). At the C-terminus the catalytically active DDE motif is found which is present in most transposases and INs. The DDE motif is responsible for cleavage reactions and strand transfer during transposition. The first 57 amino acids of the SB transposase (from now on termed N57) that carry the pai subdomain are in part responsible for protein-protein interaction necessary for transposase multimerisation. N57 is able to form dimers and tetramers *in vitro*, and transfection of N57 together with full length transposase did not impair transposition (Izsvak et al. 2002). During transposition SB transposase is believed to form a tetramer with one SB transposase molecule binding to each of the transposase binding sites on the transposon. Most amino acid changes in the SB transposase protein lead to complete inactivation or at least a decrease in transposition activity. However some transposase mutants transposed more efficient than the original version (Mates et al. 2009) (Geurts et al. 2003) (Zayed et al. 2004). The hyperactive SB transposase variant SB100x (Mates et al. 2009) can transpose up to a 100-fold more efficiently in mammalian cells than SB10, the transposase originally resurrected from fish, and the hyperactive SB transposase mutant M3a transposes about 7-fold more efficiently than SB10. SB100x transposase has nine amino acid changes compared to SB10 (K14R, K33A, R115H, RKEN214-217DAVQ, M243H, T314N). M3a transposase differs in seven amino acids from the original SB10 sequence (K13A, K33A, T83A, RKEN214-217DAVQ). The hyperactive transposase variant HSB3 mediates transposition about 9-fold as efficient as SB10 and even 13-fold when combined with a hyperactive transposon (Yant et al. 2004). SB transposase acts

in *trans* and can thus mobilize DNA as long as it is flanked by *SB* specific IRs. For use as a molecular tool the two components of the *SB* transposon system: the catalytic element, the transposase, and the scaffold for gene transfer, the transposon, can be physically separated. The ORF of the transposase can be removed from the transposon and replaced with a gene of interest (GOI). SB transposase can either be provided from a separate plasmid, as mRNA or as protein (Fig. 3A). The amount of transposase expressed in cells is critical for the transposition process since efficiency of transposition decreases in the presence of excess of SB transposase, a phenomenon called "overproduction inhibition". Fusion of the SB transposase to other domains especially fusions to the C-terminus of the transposase often leads to decreased transposition activity or renders the transposase completely inactive (Wu et al. 2006) (Ivics et al. 2007). However a zinc finger (ZF)/SB transposase fusion protein which was functional in transposition also showed attenuated overproduction inhibition (Wilson et al. 2005). Even though transposition efficiency also decreases with increasing transposon size (Izsvak et al. 2000) up to 10 kb can be delivered using the *SB* transposon system. With a so-called sandwich transposon where a transgene is flanked by two complete *SB* transposons even bigger cargo can be delivered (Zayed et al. 2004). Rearrangements, deletions or mutations of or within left and right IR of the transposon in the most of the cases leads to reduction of transposition events. However a transposon flanked by two left IRs showed nearly three times as many transposition events as wild type (wt) transposon. Including a transposase binding half-site which is only present in the left IR in the right IR doubles transposition events compared to wt *SB* transposon (Izsvak et al. 2002).

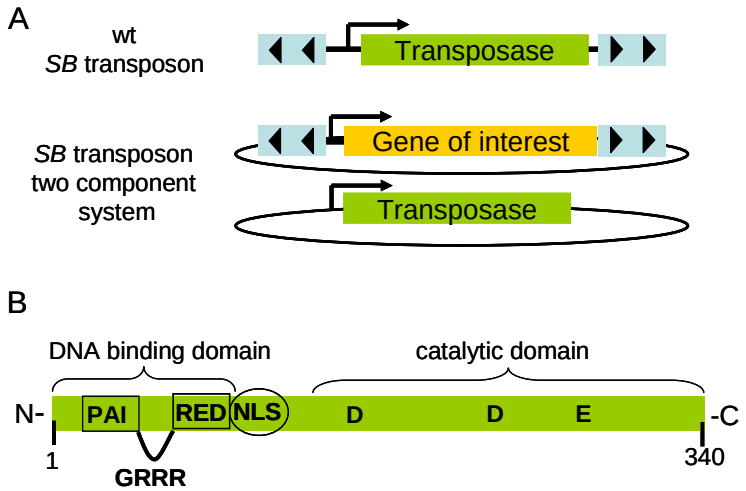


Figure 3. The *SB* transposon system. (A) The *SB* transposon system consists of two components which are both indispensable for transposition: the two IRs which provide the scaffold for transposition and the enzymatic component, the transposase. For use as DNA transfer tool the ORF encoding the transposase is removed from in between the IRs and replaced by a GOI. Transposase can be provided on a separate plasmid, as mRNA or protein. **(B)** Schematic of the SB transposase protein. The SB transposase is 340 aa long. The DNA-binding domain with a pai and a red subdomain is found at the N-terminus. The red domain partially overlaps with a NLS. The catalytically active DDE motif is found at the C-terminal part.

1.6.1. SB transposition

The process of *SB* transposition is executed in three main steps: excision of the transposon from the donor locus, joining of the transposon to target DNA and host-mediated repair of the double strand break (DSB) at the donor site. During the excision process the transposase binds to the transposon ends. DSBs are created by hydrolysis of phosphodiesterbonds at the 3´-end of the transposon. Cuts at the 5´-end are made three nucleotides inside the transposon. Target site DNA is cleaved at the phosphodiesterbonds at the 3´-end of a TA dinucleotide by the attack of the free 3´-hydroxyl groups from each transposon end. Staggered cuts at the target site TA dinuleotide and transposon ends result in a five nucleotide single-stranded gap after integration of the transposon at the target site. This gap is believed to be filled by proteins of the host cell (Fig. 4).

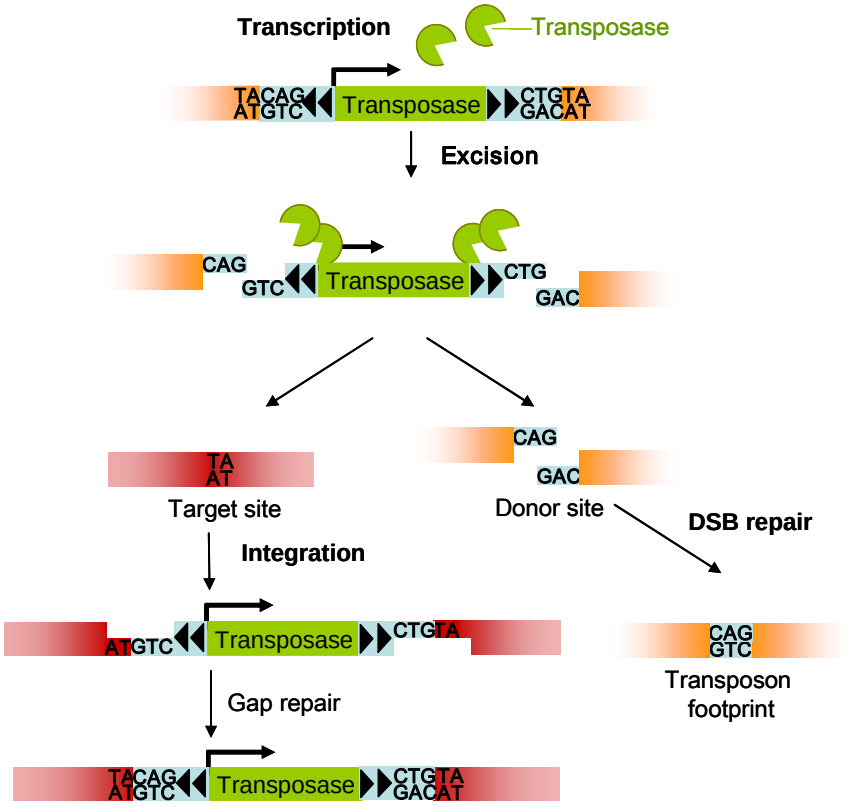


Figure 4. Cut-and-paste transposition of the SB transposon. In the first step of transposition the transposase binds the IRs. Lysis of phosphodiester bonds at the 3´-transposon ends excises the transposon from its donor site. Due to staggered cuts a 3´-overhang of three transposon-derived nucleotides is left behind at the donor site. The DSB is repaired by proteins of the host cell leaving behind a characteristic three nucleotide transposon footprint. At the target site phosphodiester bonds at the 3´-ends of a TA dinucleotide are attacked by the 3´-hydroxyl groups of the excised transposon. Single-stranded gaps at the transposon ends are believed to be repaired by proteins of the host cell.

The *SB* transposon system shows efficient transposition in different somatic tissue of a wide range of vertebrate species including humans, as well as in the germline of fish, frog, mouse and rat (reviewed in (Mates et al. 2007)). On genome-wide level *SB* shows a slight bias toward integration into genes and their upstream regulatory sequences (Yant et al. 2005); however this tendency is not as pronounced as for viral vectors currently applied in gene therapy and no insertion preference of *SB* for transcribed genes has been detected so far. *SB* transposon vectors were shown to be fairly inert in their transcriptional activities and insulator

elements incorporated into the next generation of transposon elements can make them even safer for application in gene therapy (Walisko et al. 2008).

On the nucleotide level, *SB* requires a TA dinuleotide for insertion. Apart from this, target site preferences of *SB* are based on physical properties of the DNA rather than on DNA sequence (Vigdal et al. 2002) (Liu et al. 2005). Probably due to its physical structure TA-rich DNA seems a preferred target for *SB* insertions compared to GC-rich DNA.

Therapeutic genes relevant for gene therapy treatment of a variety of genetic diseases have been integrated successfully in relevant cell types including human CD34+ hemapoetic stem/progenitor cells followed by engraftment and differentiation in animal disease models using the *SB* transposon system (Xue et al. 2009) (Sumiyoshi et al. 2009).

Unlike viruses, transposons lack the ability to infect and thus to actively enter cells. Transposons can be delivered into cell types extracted from the patient prior manipulation, such as cells from the hematopoietic system, using methods like electroporation, microinjection or complexing with polyethylenimin (PEI) reviewed in (Ivics and Izsvak 2006). However, for gene delivery to cells *in vivo,* obviously different methods for transposon delivery have to be applied. Hydrodynamic injection has been widely used for gene delivery to liver cells in animals but its clinical application is questionable. Coupling the integration machinery of transposons with the cell infection machinery of a virus such as adenovirus, lentivirus or herpes simplex virus could provide a solution to this problem (Yant et al. 2002) (Bowers et al. 2006) (Staunstrup et al. 2009) (Vink et al. 2009). In recent work nanocapsules coated with natural or endogenous ligands to target specific cell types in mouse were used (Kren et al. 2009). In rat proteoliposomes coated with a cell type specific ligand combined with a fusogenic viral protein delivered DNA to the desired cell type (Wang et al. 2009).

1.6.2. SB as a delivery vector in gene therapy

From the TEs described above only some have the potential to be applied for gene therapy at present. Use of the *SB* transposon system for gene delivery in animals and human cell culture has shown great promise for use in human gene therapy. Alternatives to currently used viral vector systems are in demand but until start of this year no in-human clinical trial using a transposon as gene delivery vector had been conducted. The first-in-human clinical trial involving a transposon uses the *SB* transposon system (Singh et al. 2008) (Williams 2008). Peripheric blood mononuclear cells (PBMCs) from patients suffering from $CD19^+$ B lymphoid malignancies will be genetically modified by inserting a chimeric antigen receptor which recognizes the lineage-specific tumor antigen CD19 via a murine single-chain variable

fragment linked to the CD28 endodomain fused with the CD3-ζ cytoplasmic domain (Kowolik et al. 2006). Specific binding of CD19 occurs via the murine single-chain variable fragment. CD3-ζ chains are responsible for signal transduction involving immunoreceptor tyrosin-based activation motifs (ITAMs). Modification of the phosphorylation status of ITAMs leads to T lymphocyte activation. In mammals ζ chains together with CD3 and the TCR generate an activation signal in T lymphocytes. Activated CD28 produces a co-stimulatory signal in T lymphocytes which leads to interleukin (IL) production. TCR activated T lymphocytes without CD28 co-stimulation are anergic. After full myeloablative chemotherapy, a peripheral blood stem cell autologous transplantation and treatment with monoclonal antibody to CD20 genetically modified autologous T lymphocytes will be infused into patients. This clinical study is executed to examine safety and feasibility of gene therapy using the *SB* transposon system and persistence of T lymphocytes *in vivo*.

1.7. Targeted DNA insertion in the genome

Uncontrolled integration of a therapeutic transgene into the genome presents the risk of insertional mutagenesis, overexpresssion of adjacent endogenous genes and can lead to long-term silencing of the integrated transgene. Viral vectors and non-viral vectors with natural integration biases and targeting strategies exist. Depending on the life cycle of the mobile DNA element different strategies are pursued. Viruses which will ultimately lyse and therefore destroy the host cell profit from insertion into genomic regions where they can take advantage of promotor or enhancer elements that ensure good expression. Other elements which spread vertically rather than horizontally depend on the integrity of their host cell in order to be transmitted to following generations. These elements insert into regions or sites of the genome that will not cause harm to the host since elimination of the host cell will result in elimination of the element.

1.7.1. Naturally occurring targeting in mobile element integration

Viruses

Biased integration patterns are observed for a number of viral gene delivery systems. Preferences for certain genomic regions such as upstream regulatory sequences, active transcription units or certain chromatin states of DNA may be due to better accessibility of these regions at the time of transgene integration by the viral IN. In the case of HIV IN however no preference for DNase I hypersensitive sites was observed. Insertion preferences

rather seem to emanate from interaction of the IN with a cellular protein called lens epithelium-derived growth factor (LEDGF)/p75 (Ciuffi et al. 2005). LEDGF/p75 is a transcriptional co-activator that interacts with components of the basal transcriptional machinery (Ge et al. 1998). It binds to HIV IN and drives it into the nucleus when both proteins are expressed at high levels (Llano et al. 2004).

As mentioned before AAV integrates in the absence of a helper virus to a high percentage into the so-called *AAVS1* locus on human chromosome 19. Integration is mediated by rep68 and rep78 two of the virus-encoded proteins via a replicative recombination mechanism. Rep68/78 proteins bind to specific sites of the viral genome and also to a specific site within the *AAVS1* locus resembling the rep68/78 binding sites in the viral genome. This brings the viral genome to close proximity with the *AAVS1* locus. Integration occurs by template strand switch during unidirectional DNA synthesis. Hybrid vectors using the site specific integration of AAV and the amplicon of herpes simplex virus or adenovirus use favourable traits of both vector systems (Recchia et al. 2004) (Cortes et al. 2008).

Recombinases

Site-specific DNA integration is also mediated by recombinases. There are two main types of recombinases: the serine and tyrosine recombinases which differ in their mechanism of integration. Recombination steps mediated by tyrosine recombinases are reversible with excision favoured over integration whereas recombination of serine recombinases is directional. Well known tyrosine recombinases include *Cre* and *Flp*. *Cre* descends from the bacteriophage P1 and mediates recombination events between so-called *loxP* sites. DNA flanked by two *loxP* sites in direct orientation will be excised, DNA flanked by two *loxP* sites in inverted orientation will be inverted and recombination between two single *loxP* sites will lead to DNA strand exchange. *Cre* is active in eukaryotic, including human, cells and is widely used in genome engineering in mice (Yu and Bradley 2001). Recently a new recombinase termend *Dre* closely related to *Cre* was found in bacteriophage P1. *Dre* mediates recombination between so-called *rox* sites. *Cre* and *Dre* are heterospecific, *Cre* does not mediate recombination at *rox* sites and *Dre* does not recognize *loxP* sites (Sauer and McDermott 2004) (Anastassiadis et al. 2009). Using directed evolution a new Cre recombinase specific for sequences in the LTRs of integrated HIV proviruses has been created. This novel enzyme was termend *Tre* and could mediate excision of the HIV provirus from genomic DNA (Sarkar et al. 2007).

The Flp recombinase from *Saccharomyces cerevisiae* recombines so-called *FRT* sites in a similar mechanism like *Cre* though less efficiently. Attempts to improve *Flp* recombination efficiency include creation of a thermooptimized version for the use in mammalian cells called *Flpe* that performs four-fold better at 37 °C than the wt version (Buchholz et al. 1998). A mouse–codon-optimized version of *Flp* called *Flpo* was developed that shows recombination efficiency similar to that of *Cre* (Raymond and Soriano 2007).

The ΦC31 IN originally found in *Streptomyces lividans* is a representative of serine recombinases (Thorpe and Smith 1998). ΦC31 excises and inverts DNA between the heterotypic sites *attP* and *attB* in direct and indirect orientation respectively. Upon recombination *attP* and *attB* sites form *attL* and *attR* sites which are not further recombined by ΦC31. Pseudo *att* sites exist in humans (*psA*) as well as in mouse (*mpsA*) (Chalberg et al. 2006) (Thyagarajan et al. 2001). In human 293 cells harbouring an inserted *attP* site , 15 % of integrations were detected at the inserted *attP* site, 5 % of the rest of insertion occurred at *psA* , 5-10 % were distributed randomly and the rest was believed to be distributed over the other ~100 pseudo sites in the human genome (Thyagarajan et al. 2001). However a significant level of toxicity and inter-chromosomal recombination in the human genome has been reported for ΦC31 in human cells (Chalberg et al. 2006) (Liu et al. 2006). Some cases of toxicity upon usage of the *Cre*/*loxP* systems have also been observed (Loonstra et al. 2001) (Forni et al. 2006) (Schmidt et al. 2000) which may be due to recombination at pseudo *loxP* sites as found in the mouse genome (Thyagarajan et al. 2000). However the *Cre*/*loxP* system has been used widely for a long time and toxicity rather seems to be the exception than the rule.

Transposases

Since transposons lack an extracellular phase their fate is closely linked to that of the host cell. Insertion into a coding region resulting in fatal consequences for the cell will also be fatal for the transposon. In organisms like *Saccharomyces cerevisiae* with a compact genome containing a high proportion of coding regions transposition should be highly regulated to avoid insertion into genes and their regulatory sequences. *Ty* retrotransposons are structurally and functionally related to retroviruses. Integration of *Ty1*, *Ty3* and *Ty5* retrotransposons is tethered to certain sites in the yeast genome by host proteins. *Ty1* insertions are found preferably upstream of genes transcribed by RNA polymerase III (Pol III) like tRNA genes (Kim et al. 1998) or 5S RNA genes (Bryk et al. 1997). Components of the Pol III machinery were found to be essential for *Ty1* targeting (Bachman et al. 2005) but other factors like

chromatin structures or physical properties of DNA at the target site may also play a role in the choice of the integration site. *Ty3* is recruited to Pol III transcription start sites by TFIIIB and TFIIIC two factors important for assembly of Pol III complexes (Kirchner et al. 1995). TFIIIB is sufficient to target *Ty3* (Yieh et al. 2002) whereas TFIIIC orients binding of TFIIIB to the TATA box and weakly interacts with Ty3 IN (Aye et al. 2001). Ty5 interacts with the host protein Sir4p (Xie et al. 2001) which directs transposon insertions into heterochromatic regions of the genome such as telomers and silent mating loci (Zou et al. 1996).

The bacterial transposon *Tn7* is able to target two particular sites in the bacterial genome depending on proteins involved in the transposition process (Peters and Craig 2001). *Tn7* encodes five different transpositional proteins: TnsA, B, C, D and E. During bacterial conjugation TnsE seems to recognize DNA structures with recessed 3´-ends during lagging strand DNA synthesis, and directs integration of the transposon to this site. This transposon targeting seems to depend on DNA structure rather than on DNA sequence. TnsD however recognizes a specific DNA site called *attTn7* at the 3´-end of the bacterial glutamate synthetase gene (*glmS*). TnsD directed transposon insertion occurs several base pairs downstream of *glmS*. However *Tn7* transposition has not yet been reported in human cells.

Another strategy to avoid disruption of host genes is pursued by the DIRS LTR-retrotransposon family in *Dictylostelium discoideum*. Even though this familiy shows no initial target site selectivity they are found in clusters of several copies of themselves (Loomis et al. 1995) at centromeric and telomeric regions of the chromosomes. Another group of elements, the non-LTR retrotransposons called TRE[1] (tRNA gene-targeting retrotransposable elements) insert at regions free from protein-coding sequences in the genome (Winckler et al. 2002). *TRE5* elements insert preferentially about 50 bp upstream of tRNA genes whereas *TRE3* elements favour the integration about 100-150 bp downstream of tRNA genes. Interaction of TFIIIB with the *TRE5* element has been reported (Chung et al. 2007).

1.7.2. Artificial targeting

Targeting abilities of the vector systems mentioned above that are currently applied in preclinical experiments or clinical trials are not specific enough to be used for targeted gene therapy. Different strategies for targeted integration exist. One possibility is fusion of a DNA binding domain (DBD) to the enzymatic factor of the gene delivery system with the intention to tether the cargo DNA/enzyme complex to a certain DNA sequence. Bound to this DNA sequence, the enzymatic component is expected to integrate its cargo DNA nearby (Fig. 5A). If direct modification of the enzymatic factor is not desired the enzymatic component can be

tethered to a particular site in the genome by protein-protein interaction. A protein known to interact with the enzymatic component fused to a DBD will lead the enzymatic component to a specific site in the genome and cargo DNA insertion will occur at nearby sequences (Fig. 5C). Targeted gene insertion can theoretically also be achieved be tethering the DNA component of the gene delivery system to a certain site in the genome. A DNA sequence that is specifically recognized and bound by a DBD domain is included in the DNA component. This DBD domain is fused to a DBD that binds a specific DNA sequence in the genome. After bringing the DNA component close to the specified site in the genome the enzymatic component of the delivery system is expected to integrate it in the adjacent region (Fig. 5B).

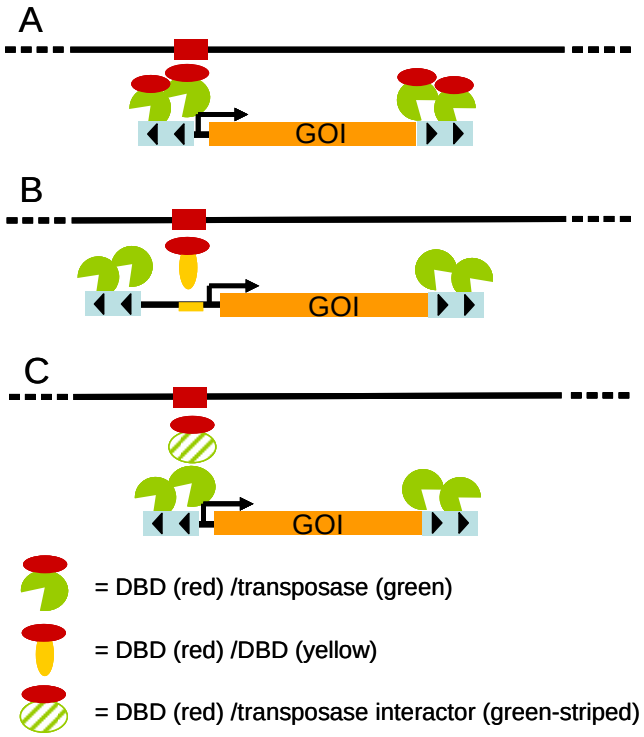


Figure 5. Different strategies for targeted transposon integration into a specific region in the genome. The transposon is depicted by two IRs (blue) flanking a GOI (orange). The black line represents target DNA; the red square a DBD binding site. **(A)** Transposase is fused to a DBD. Upon binding of the DBD to its recognition site the transposase inserts the transposon into adjacent DNA. **(B)** DBD1 is fused to DBD2. DBD1 specifically binds a sequence in the genome and DBD2 specifically binds a sequence incorporated in the transposon. Upon tethering of the transposon to a specific genomic site by the fusion protein, the transposase can integrate the transposon into adjacent DNA. **(C)** A protein interacting with the transposase is fused to a DBD. Upon binding of the DBD to its recognition site the transposase/transposon complex is tethered to this site by protein-protein interaction and transposon insertions will occur in the adjacent region.

1.7.3. Targeting by DBD fusion proteins

Apart from serine recombinases where DNA-binding and catalytic domain are spatially separated in most of the other INs/recombinases/transposases catalytic and DNA-binding domain is structurally interwoven making it difficult to alter one property without effecting the other. Altering target specificities of such enzymes is challenging, but can be achieved using approaches like directed evolution (a random mutagenesis followed by activity screening under selective conditions) or mutation of key amino acid implicated in target recognition (reviewed in (Collins et al. 2003)). Unfortunately, mutations in the amino acid sequence leading to altered site-specificity may not only affect binding properties but may also alter catalytic performance of the enzyme. Fusion of specific DBD to enzymes seems to

offer an easier approach. However fusion of a foreign domain to the enzyme may also cause problems. Due to altered folding of the chimeric enzyme features like intrinsic DNA binding or catalytic activity may be impaired or abolished. Furthermore, for adequate target site selection apart from choosing the right DBD, requirements of the catalytic domain on target DNA sequence or structure have to be met.

First successful approaches to target DBD/IN proteins were done *in vitro* by fusing the IN of avian sarcoma virus (ASV) to the DBD of the *E.coli* LexA protein. The fusion protein showed altered integration pattern compared to wt protein and an integration hot spot near a tandem *LexA* operator (Katz et al. 1996). Similarly, the HIV IN was fused to the DBD of the LexA repressor protein (Goulaouic and Chow 1996) or the DBD of phage λ repressor protein (Bushman 1994). Both fusion proteins showed enriched integration near respective target sites *in vitro*.

Numerous transcription factors (TFs) are active in the human genome binding to diverse DNA sequences. TFs specifically recognize and bind DNA and recruit other proteins to these sites. The DBDs of TFs are spatially separated from other functions of the protein and can thus be used as an independent modul in fusions with other proteins. TFs can be classified according to the structure of their DBD, such as ZFs, helix-turn-helix, helix-loop-helix and high-mobility group boxes. Using the DBD of a known TF to target a specific DNA sequence thus seems promising. The TF Gli1 found in vertebrates has a six finger ZF as DBD. Either the DBD of Gli1 or the cI repressor of phage λ was fused to the bacterial IS element *IS30*. Fusion proteins were able to target integration into plasmid targets in *E.coli* and zebrafish (Szabo et al. 2003). This study was the first to show targeted integration of a modified TE *in vivo*.

The DBD of the Gal4 TF from yeast is a ZF of the Zn_2C_6 type that binds to a 17-bp DNA sequence called upstream activating sequence (*UAS*). The DBD of Gal4 was fused to either the Mos1 (a Tc1/*mariner* transposon from *Drosophila mauritiana*) or PB transposase. Fusion proteins were examined for transpositional activity and targeting abilities in plasmid-based transposition assays in mosquito embryos (Maragathavally et al. 2006). Efficient targeting was observed for both fusion proteins. For the Gal4 DBD/Mos1 fusion protein 96 % of insertions were detected at the same TA dinucleotide 954 bp away from the targeted *UAS*. For the Gal4 DBD/PB fusion protein 67 % of insertions were found at the same TTAA 1,103 bp away from the *UAS*.

Other independent studies with Gal4 DBD fusions to transposases have been conducted. Gal4 DBD fusions to Tol2 and SB11 (an early-generation hyperactive mutant of SB) completely

abolished transpositional activity, whereas Gal4 DBD/PB fusions showed transpositional activity similar to that of unfused PB transposase. The targeting potential of these fusion proteins has not been examined in this study (Wu et al. 2006). In another study it was shown that C-terminal fusions of the Gal4 DBD to the SB transposase abolished transpositional activity, whereas N-terminal fusions retained 26 % of transpositional activity of unfused transposase. The SB transposase used in this study was HSB5, a third generation hyperactive mutant of SB. *SB* transposon insertions were 11-fold increased in a 443-bp window around a 5-mer *UAS* site in the target plasmid for the Gal4 DBD/SB transposase fusion compared to the integration pattern for unfused SB transposase (Yant et al. 2007).

The Tn3 resolvase belongs to the serine recombinase group. DBD and catalytic domain are spatially separated in this group and function independent of each other. The naturally occurring three finger ZF domain from the TF Zif286 originally discovered in mouse was fused to the catalytic domain of Tn3 transposase. Fusion proteins successfully targeted two inverted Zif286 binding sites flanking a *Tn3 res* site in *E.coli* (Akopian et al. 2003). The Zif286 DBD was also fused to HIV IN. This fusions protein showed a bias for insertions near specific binding sites *in vitro* (Bushman and Miller 1997).

Naturally occurring DBDs have some limitations for their use as targeting agents. Some of the DBDs mentioned so far do not have physiological targets in the human genome. For effective targeting their recognition site has to be introduced into the genome before delivery of the transgene. Other DBDs have numerous physiological targets in the human genome. For example, the DBD of Zif286 recognizes a 9-bp DNA sequence. A 9-bp DNA sequence is expected to be present >10,000 times in the human genome by chance.

ZF nucleases (ZFNs) are fusions between a ZF protein, typically a three or four finger ZF, and the FokI cleavage domain. Two ZFNs need to heterodimerize in order to cleave DNA at the target site. Upon introduction of a DNA DSB mediated by the ZFN potential gene repair or introduction of DNA can occur via homologous recombination with a homologous DNA template. ZFNs were first used in 2003 (Porteus and Baltimore 2003) to introduce a DSB at a target site in the human genome. In 2005 Urnov et al. (2005) were the first to use ZFNs for gene correction for therapeutic proposes. Off-target cleavage through unspecific binding of the ZF or homodimerization of the FokI nuclease domain lead to cytotoxic effects in cells. Charges introduced at the FokI dimerization interface that promote heterodimer formation and reduce homodimer formation decreased off-target cleavage and thus cytotoxicity (Szczepek et al. 2007). Integration-deficient viral delivery systems based on lentivirus (Lombardo et al.

2007) or adenovirus (Perez et al. 2008) have been used to transport ZFNs and DNA templates into different cell types.

Fusions of the artificial multifinger ZF protein E2C to the HIV IN were able to target retroviral integrations near the E2C binding site *in vitro* (Tan et al. 2004). Retroviral insertions near the E2C binding site in the human genome were found to be ten-fold enriched compared to unfused HIV IN (Tan et al. 2006). However, virions containing the fusion protein showed low infectivity ranging from 1 to 24 % compared to virions carrying unfused IN.

The E2C ZF was also fused to HSB5 (a hyperactive mutant of the SB transposase). Fusion of E2C to the transposase reduced transpositional activity. However, fusion proteins with a glycine/serine linker between fusion partners and a human codon-optimized E2C gene retained ~10 % activity of the unfused transposase protein. An enrichment of *SB* transposon insertions was observed in a 443 bp window around a 5-mer repeat of the E2C binding site on a plasmid as compared to unfused transposase. In genomic context no targeted transposon insertions were detected (Yant et al. 2007).

The engineered three finger ZF Jazz binds a 9-bp sequence in the promoter region of the human utrophin gene (Corbi et al. 2000). Fusions of Jazz to SB transposase yielded a fusion protein with about 15 % transpositional activity compared to unfused transposase. However targeted transposon integrations near the Jazz binding site could not be discovered in genomic context (Ivics et al. 2007).

One possible explanation for the lack of targeted transposon insertions could be physical constraints imposed on the transposase part of the fusion protein upon binding of the ZF part to its recognition site that make it impossible for the transposase to interact with a TA dinucleotide required for transposon insertion. This would especially hold true for GC-rich genomic regions where TA dinucleotides are rare and/or lie in a DNA context which make them poor targets. Summarizing these studies it appears that fusing a DBD directly to the IN/transposase produces problems for these vector systems. This manifests in production of virions of low infectivity in the case of viral vectors and in reduction of transposition activity in case of transposase fusion proteins. Even though targeted DNA integration could be established on plasmid level to some extent, targeted DNA insertions in genomic context poses a challenge. Furthermore, transposases fused to foreign DBDs retain their intrinsic ability to bind DNA. Such enzymes used for targeting experiments do not depend on their DBD fusion partner to bind DNA and thus also show off-target transposon insertions. Insertion frequencies near the targeted DNA sequence have been achieved to some extent

(Tan et al. 2006) (Ivics et al. 2007) by such proteins so far, still insertions still occur elsewhere in the genome independent of DBD binding.

Invertases like Gin can catalyze resolution or insertion of DNA. However most serine invertases are responsible for inversions that generate one of two heritable traits and rare incidences of conversions are due to aberrant reactions. Their strong bias for inversion of DNA is probably due to the involvement of host cell factors and enhancer sequences during the process of recombination. Mutant Gin proteins have been established that are able to recombine in the absense of such host factors (Klippel et al. 1988). These Gin mutants promote inversions as well as deletion and intermolecular recombination. ZF/Gin fusion proteins based on one of these Gin mutants were selected for enhanced resolution activity using a directed evolution approach called substrate-linked protein evolution (SliPE) (Gordley et al. 2007) (Buchholz and Stewart 2001). Proteins that emerged from this procedure bound specifically to target sites containing the correct ZF binding site but showed relaxed binding specificity for the catalytic Gin domain sequence. In contrast to the wt serine recombinases the mutated catalytic Gin domains show a loss of site orientation specificity. Another ZF/Gin fusion protein showed targeted insertion of 98.5 % to an artifically introduced target site in human cells (Gordley et al. 2009). This 20 bp core sequence is not naturally found in the human genome and thus needs to be introduced into the genome for successful DNA insertion a fact problematic for applications like gene therapy.

1.7.4. Targeting by interaction with DNA-binding proteins

Another approach to target transgene insertion leaves the IN/transposase/recombinase unfused and rather tethers the DNA component of the delivery system or a protein interacting with the IN/transposase/recombinase to a specific site in the genome. As mentioned before the HIV IN interacts with a protein called LEDGF/p75. Full length LEDGF/p75 or the domain of LEDGF/p75 interacting with HIV IN were fused to the DBD of the λ repressor protein. HIV integrations occurred near λ repressor binding sites in reactions containing the artificial interaction partner *in vitro* (Ciuffi et al. 2006).

Sir4p, a protein directing integration of the *Ty5* retrotransposon in yeast, fused to the *E.coli* LexA DBD lead *Ty5* integrations near a *LexA* operator (Zhu et al. 2003). Similarly exchanging the domain of Ty5 IN interacting with Sir4p with a heterologous domain interacting with a protein fused to LexA also lead to *Ty5* integration near *LexA* operators (Zhu et al. 2003).

One approach to target *SB* transposon insertions without directly fusing a DBD to the transposase was conducted by including the *LexA* operator into the *SB* transposon. The LexA protein was fused to a scaffold attachment factor (SAF)-box. SAF-box domain was first identified in the human scaffold attachment factor A that specifically binds to scaffold/matrix attachment regions (S/MARs) (Kipp et al. 2000). S/MARs are structural components that organize chromosomes into separate domains by binding to the nuclear matrix. They are required for transcription, replication, recombination and chromosome condensation. They are important for assembly of proteins involved in the transcription process and transgenes flanked by S/MARs and have shown expression independent of their insertion site (Harraghy et al. 2008). Thus insertion of a transgene into an S/MAR could reduce the possibility of silencing. In this approach integration of the transposon should be tethered to S/MARs through binding of LexA to the *LexA* operator on the transposon and interaction of the SAF-box with S/MARs (Ivics et al. 2007). An enrichment of transposon integrations within 1 kb of genomic S/MAR sequences were found in the presence of the LexA/SAF-box fusion protein. Similarly the tetracycline repressor (TetR), a highly specific DNA binding protein, was fused to LexA. In a cell line containing a the tetracycline response element (TRE^2)-driven *EGFP* gene two transposon insertions downstream of the TRE^2 could be recovered in the presence of the fusion protein, whereas none was discovered in the absence of the fusion protein. A different targeting strategy involving protein-protein interactions was applied in the same publication. N57, the first 57 amino acids of the SB transposase, contains a helix-turn-helix motif responsible for protein-protein interactions that is sufficient for transposase dimer formation. Coexpression of N57 with full length SB transposase had no dominant negative effect on transposition (Izsvak et al. 2002). A fusion protein of N57 and TetR was believed to tether full length, transpositionally active, SB transposase to the TRE^2 followed by transposon integration into adjacent regions. Again the cell line containing a TRE^2-driven *EGFP* gene was used for targeting experiments. On average > 10 % of cells that contained a transposition event showed a transposon insertion within 2.5 kb of the TRE^2 for experiments including the N57/TetR fusion protein.

1.8. Targeting strategy

Several proteins with DNA-binding capacity exist and some of them have already been applied in fusion proteins to target integration of DNA to certain sites. However most of these DNA binding proteins descent from bacteria or yeast and their recognition sequence often is

not found in the human genome. In order to use such a DNA binding protein for targeting purposes the DBD recognition sequence would first need to be introduced into the human genome. Also if DNA binding at a selected site in the genome is demanded insertion of a DNA binding site at exact this position could present additional obstacles. For applications where a defined region or locus should be targeted like in some cases of gene therapy treatments a DBD binding a DNA sequence in this exactly defined genomic region is desirable. ZF proteins of the C_2H_2 type could offer a possible solution for this demand.

1.8.1. Zinc finger proteins

In 1985 ZF protein were first discovered in the transcription factor IIIA of Xenopus laevis by Miller et al. (1985). ZF proteins coordinate zinc ions with a combination of cysteine and in some cases also histidine residues which together with hydrophobic interactions stabilize their structure (Miller et al. 1985) (Lee et al. 1989). Different types of ZF proteins can be classified by type and order of these residues, for example C_4, C_6, and the most common type C_2H_2. Because of their modular structure ZFs of the C_2H_2 type can theoretically be designed to bind any sequence. For this property these proteins have emerged to be the most commonly used domain for sequence specific binding. A well-studied representative for a C_2H_2 type ZF is the mammalian transcription factor Sp1. This three finger protein naturally binds to GC box promoter elements. The similar named Sp1C ZF is a consensus protein made from a database of 131 ZF sequences which serves as a framework for artificial ZF design (Desjarlais and Berg 1993). Another very thoroughly studied C_2H_2 type ZF stems from Zif286 a transcription factor originally discovered in mouse (Fig. 6). Three-dimensional solution structures by nuclear magnetic resonance spectroscopy gave insights into ZF residues involved in DNA binding (Lee et al. 1989). Also the crystal structure of the three individual ZFs of Zif268 bound to their recognition DNA has been solved and published in 1991 greatly enhancing the understanding of ZF binding (Pavletich and Pabo 1991). Each C_2H_2 type ZF binds 3-4 bp of DNA. For binding of longer sequences individual fingers are arranged in a tandem repeat with adjacent fingers connected by canonical linkers (TGQKP and TGEKP). Each individual ZF is about 30 aa in length with a consensus amino acid sequence of (F/Y)-X-C-X(2-5)-C-X3-(F/Y)-X5-ψ-X2-H-X(3-5)-His where X stands for any amino acid and ψ for a hydrophobic residue. Through binding of zinc by the two histidines and cysteins the finger forms a stable ßßα fold. The α-helix fits into the major groove of the DNA double helix. Contact to the primary strand of DNA is established through amino acids at positions -1, 3 and 6 of the α-helix to nucleotides 3, 2 and 1 of its ZF recognition sequence respectively with nucleotide 3 at

the 3´-end and nucleotide 1 at the 5´-end of the primary strand. The first finger of a polydactyl ZF protein would thus bind the base triplet at the 3´-end of its recognition sequence (Fig. 6). The amino acid at position 2 of the α-helix can make contact to the complementary strand of DNA to nucleotide -1 of the ZF recognition site. An aspartic acid at position 2 of the ZF α-helix binds either cytosine or adenosine on the complementary DNA strand at the 5´-position of the base pair triplet of the preceding finger. Hence the preceding finger could only bind 5´-TGG-3´or 5´-GGG-3´ triplets. In other words, ZF domains with aspartic acid at position 2 at the α-helix would exclude 5´-ANN-3´ and 5´-CNN-3´ triplets for the preceding finger. This phenomenon called target site overlap may cause problems when arranging individual ZF in a tandem array. This is one reason why design of ZF domains binding 5´-GNN-3´ type recognition sites, which are quite frequent in naturally occurring ZF proteins, are considered more promising than for base triplets of the 5´-ANN-3´type.

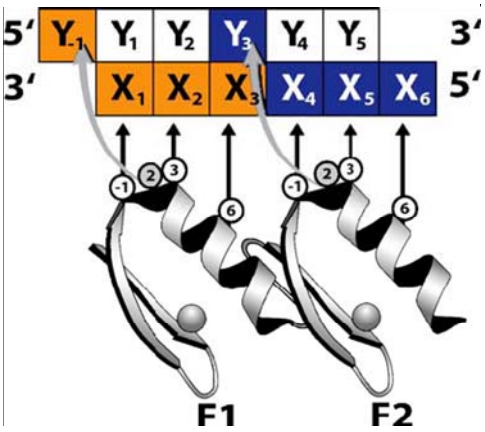


Figure 6. Model of a two finger C_2H_2 type ZF binding to its recognition site. Each ZF consists of two ß-stands and an α-helix. The structure of the protein is stabilized by chelation of a zinc ion. Amino acids at position -1, 3 and 6 of the ZF α-helix bind to the 3´-, middle and 5´-position of their recognition site. An aspartic acid at position 2 of the ZF α-helix binds A or C at the 5´-position of the complementary DNA strand, a phenomenon called target site overlap. The figure was published in (Papworth et al. 2006).

1.8.2. Engineered ZF proteins

First approaches to engineer ZFs with altered binding preferences include helix grafting, ZF shuffling and site directed mutagenesis. In helix grafting naturally occuring ZF such as Zif268 are modified for amino acids believed to be important for binding using a ZF database created from known ZFs. In ZF shuffling individual fingers of different known ZF proteins are stiched together in a new order. Alternatively, amino acids of one finger important for binding are randomly mutated and thus change their preferences for binding sites (Desjarlais and Berg 1992) (Desjarlais and Berg 1993) (Thukral et al. 1992) (Shi and Berg 1995). However, these approaches limit the design of new ZF variants to amino acids considered important for binding. However, crystallographic studies of ZF-DNA interactions also suggest contacts between neighboring fingers and subsites (Pavletich and Pabo 1991) (Fairall et al. 1993;) (Pavletich and Pabo 1993). Analysis of ZF proteins varying in amino acids adjacent to amino

acids known for binding would take these contacts into account. By using methods like phage display it is possible to simultaneously screen numerous ZF variants for best binding to certain DNA sequences. Phage libraries with ZF variants were created and purified over DNA of fixed sequence. Sequences of ZF proteins binding to the DNA were compared and further conclusions about positions crucial for binding were drawn (Choo and Klug 1994) (Jamieson et al. 1994) (Jamieson et al. 1996) (Rebar and Pabo 1994) (Wu et al. 1995). To adress the question of how interdigital contacts affect ZF binding, different laboratories developed different strategies. In sequential selection approaches naturally occuring three finger ZF proteins are object to phage display. One finger at a time gets randomised starting with finger 1. For each additionaly randomized finger a new library of ZF proteins is created. ZF libraries get selected by DNA targets encoding part of the wt binding site followed by the new binding site (Greisman and Pabo 1997). Thus, for the sequential selection approach, creation of three phage libraries is required. In a parallel selection approach, only the second finger of a three finger protein is randomized and selected for binding to a new DNA sequence. Individual fingers selected by this mode are then simply stitched together to bind longer sequences. This modular assembly promises fast and easy design of ZF proteins for virtually every DNA sequence once a well-binding ZF for each base pair triplet is found (Segal et al. 1999) (Segal and Barbas 2000). However, modular assembly of individual ZFs poses the risk of possible target site overlap issues. This problem is addressed in the so-called bipartite complementary library where one and a half finger is selected from a phage library at a time. Combination of two of these one-and-a-half fingers, each binding to compatible DNA, then results in a three finger protein (Isalan et al. 2001).

Artificial ZF proteins designed to bind specific sites in the genome have already successfully been used for regulation of gene expression (Beerli et al. 1998) and targeted gene insertion (Tan et al. 2006) (Porteus and Baltimore 2003).

1.8.3. Engineering ZF proteins with the "Zinc finger tools" website

The „Zinc finger tools" website from the laboratory of Carlos Barbas III can aid to choose apt ZF binding sites and suggests ZF proteins that would potentially bind to these sites. For design of ZF proteins it uses the modular assembly strategy stitching together individual ZFs from a database of 64 ZF domains each binding to a certain base pair triplet. For creation of this 64 ZF domain database, ZF database phage libraries were created and screened for DNA binding. For binding of 5´-GNN-3´ triplets, two phage libraries were constructed. Both libraries used C7, a variant of the Zif268 ZF with higher affinity and specificity to the Zif268

binding site than the wt protein as framework (Wu et al. 1995). Randomizations were introduced in the second finger either for positions -1, 1, 2, 3, 5 and 6 or positions -2, -1, 1, 2, 3, 5 and 6 allowing all amino acid combinations or all amino acids except tyrosine, phenylalanine, cysteine and all stop codons. ZFs that recognized a number of DNA sequences were refined using site directed mutagenesis. Directed mutations introduced into ZF domains were hereby guided by phage display data as well as structural information. Domains designed by this method show binding in subnanomolar range and good specificity. A number of these domains show a >100-fold loss of affinity to sequences with a single base change (Segal et al. 1999). The database of ZF domains binding to 5′-GNN-3′ triplets obtained by the latter publication was further refined by rational design (Dreier et al. 2000). ZF domains recognizing 5′-ANN-3′ base triplets are very rare in naturally occurring ZF proteins. The problem of target site overlap imposes an additional difficulty for creation of a ZF domain for such a triplet. Like for 5′-GNN-3′ triplets, for selection of ZF domains binding 5′-ANN-3′ triplets phage display was used. Instead of the three finger protein C7 the variant C7.GAT was selected for randomization of the middle finger. In contrast to C7 C7.GAT has no target site overlap issues for the middle finger. ZF domains selected by phage display were again further refined using site directed mutagenesis. Six finger ZF proteins with various numbers of 5′-ANN-3′ base triplets in their recognition site were created by stitching together domains from the established database. These proteins bound their recognition sites with picomolar to low nanomolar affinity. Fused to regulatory elements they were able to regulate transcription of a reporter as well as endogenous genes (Dreier et al. 2001).

Generation of a database for ZF domains binding 5′-CNN-3′ triplets seemed especially challenging since no structural information for a ZF binding a 5′-CNN-3′ base triplet existed and in finger 4 of YY1 that binds to 5′-CAA-3′ no direct interaction with the 5′-cytosine was observed (Houbaviy et al. 1996). For creation of a database for ZF domains binding 5′-CNN-3′-triplets phage display and rational design was applied. Four ZF domains emerged from these phage display experiments some of them were further refined by site-directed mutagenesis. Eleven domains were generated by site-directed mutagenesis or de-novo design. No ZF domain specifically recognizing the 5′-CTC-3′ base triplet could be created. A six finger protein was created by modular assembly recognizing an 18 bp binding site containing three 5′-CNN-3′ base triplets. Fusions of this polydactyl protein to the VP64 activation domain and to the KRAB repressor domain were successfully tested for transcriptional regulation *in vivo* (Dreier et al. 2005).

The only 5′-TNN-3′ base triplet present in natural occurring ZF recognition sites is 5′-TGG-3′ found for example in the Zif268 binding site. However the 5′-T in this triplet is not recognized by the amino acid at position 6 of the ZF alpha helix which explains why Zif268 can also bind 5′-GGG-3′ with this domain. Some ZF domains binding 5′-TNN-3′ base triplets were established (Dreier et al. 2001).

Large combinatorial libraries of artificial transcription factors comprising of three and six finger ZF domains were generated by modular assembly of characterized ZF domains. These multidigital ZF proteins were mostly exspected to bind 5′-$(RNN)_3$-3′ or 5′-$(RNN)_6$-3′ triplets (R = G or A) rather than 5′-$(NNN)_3$-3′ or 5′-$(NNN)_6$-3′ triplets. Assembled ZF proteins were then fused to regulatory domains and screened for alteration of target gene expression on the cell surface (Blancafort et al. 2003).

In nature mainly three finger ZF proteins occur, but Liu et al. (1997) showed that proteins consisting of six ZFs are also able to bind their 18 bp sequence with affinity in the nanomolar range (Liu et al. 1997). This suggests that designing ZF proteins able to bind unique sites in the human genome should be feasible.

1.8.4. E2C ZF

Beerli et al. (1998) designed an artificial polydactyl ZF called E2C. The 18 bp recognition site (5′-GGG GCC GGA GCC GCA GTG-3′) of the E2C ZF lies on chromosome 17 in the upstream regulatory region of the *erbB-2* gene and is unique in the human genome. For the development of E2C, a strategy called “helix grafting” was applied. As described earlier, certain amino acids at position -2 to 6 in the helical part of the ZF are responsible for DNA binding. Amino acids at these positions from ZF proteins predicted to bind the selected nucleotide triplets were grafted into the context of the framework of the three finger ZF Sp1C mentioned earlier. By helix grafting, two three finger proteins, each binding to the two half sites of the E2C recognition site, were established. The full length six finger E2C protein was created by combining the two three finger proteins. The E2C recognition site in genomic context (with an adenine following the final recognition site triplet) is bound at a K_d of 0.75 nM. As a reference, the Kd of Zif268 in these experiments was determined to be 10 nM (Beerli et al. 1998). Affinities of the E2C ZF to its first or second half site were K_d = 100 nM and K_d = 65 nM, respectively.

E2C ZF fused to effector domains was able to regulate gene expression in a luciferase reporter assay. Fusion proteins of E2C and the HIV IN bind to the E2C specific 18 bp binding site in the human genome as shown by DNase I footprinting (Tan et al. 2006).

2. Material and Methods

2.1. Material

2.1.1. Chemicals, antibodies, membranes

Unless described elsewise all chemicals were purchased from Sigma.

Adenosine 5´-triphosphate	Amersham/GE Healthcare
Agarose genetic technology grade	MP
Ampicillin	Serva
Antibiotic-Antimycotic (100X), liquid	Invitrogen
Bovine serum albumin	Serva
Chlorphenol red- ß-D-galactopyranosid	Roche
Complete Mini, protease inhibitor cocktail	Roche
dATP, dTTP, dGTP, dCTP (100mM each)	New England Biolabs
Dulbecco´s modified Eagle medium	Gibco/Invitrogen
EDTA disodium salt	Serva
Ethanol	Merck
Fetal calf serum (FCS)	PAA
G418-BC sulfate	Biochrom
α-^{32}P-dCTP, 3000 Ci/mmol, 10 mCi/ml	Perkin Elmer
Gene Ruler 1 kb ladder	Fermentas
Gene Ruler 100 bp ladder	Fermentas
Glycerin	Roth
Goat Anti-Mouse IgG peroxidase conjugated	PIERCE
Hybond-C Extra	Amersham/GE Healthcare
Hybond-XL	Amersham/GE Healthcare
InviTaq DNA Polymerase	Invitek
jetPEI	Polyplus-transfection
Kanamycin	Serva
Klenow fragment	Fermentas
Klenow fragment Exo (-)	New England Biolabs
LB-medium, Lennox	Q-BIOgene
Methanol	Merck

$MgCl_2$	Merck
$MgSO_4$	Merck
Monoclonal mouse Anti-Actin pan Ab5 antibody	dianova
Mouse Anti-Goat IgG peroxidase conjugated	PIERCE
Goat Anti-Mouse IgG peroxidase conjugated	PIERCE
Phosphat buffered saline (PBS)	PAA
Phusion High-Fidelity DNA polymerase	New England Biolabs
PfuUltra Fusion HS DNA polymerase	Stratagene
Polyclonal Anti-SB Transposase antibody	R&D systems
Precision Plus Protein Standard All Blue	Biorad
Proteinase K	Invitrogen
ProtoGel 30%, 37.5:1 acryl-/0.8% bisacrylamid solution	National Diagnostics
Rapid-hyb	Amersham/GE Healthcare
Restriction enzymes	New England Biolabs, Fermentas, Amersham/GE Healthcare
Sodium chloride (NaCl)	Fluka
Sodium dodecyl sulphate (SDS)	Serva
T4 DNA Ligase	New England Biolabs
Trypsin-EDTA, 5 %	Gibco/Invitrogen
Tween	CALBIOCHEM
Whatman GB005	Schleicher & Schuell

2.1.2. Bacterial strains and tissue culture cells

E. coli DH5α	labstock
E.coli ElectroMAX DH10B	Invitrogen
Homo sapiens HeLa (Henrietta Lacks), epithelial cell line from cervix, adenocarcinoma	labstock
Homo sapiens HeLa E2CTA	labstock

2.1.3. Kits

Big Dye Terminator v3.1 Cycle Sequencing Kit	Applied Biosystems
BCA Protein Assay Reagent Kit	BioRad
DNeasy Blood and Tissue Kit	QIAGEN

Dynabeads kilobaseBINDER Kit	Invitrogen
ECL Plus Western blotting detection system	Amersham/GE Healthcare
End It DNA End-Repair Kit	Epicentre Biotechnologies
Fast-Link DNA Ligation Kit	Epicentre Biotechnologies
Hexanucleotide Kit	Roche
illustra Microspin G-50 colums	Amersham/GE Healthcare
p-GEM T vector System I	Promega
Prime-It II Random prime labelling kit	Stratagene
QIAGEN Plasmid Mini Kit	QIAGEN
QIAprep Spin Miniprep Kit	QIAGEN
QIAquick Gel Extraction Kit	QIAGEN

2.1.4. Equipment

Applied Biosystems 3730 DNA Analyzer	Applied Biosystems
Covaris S2	Covaris
FLA-3000	Fujifilm
Genome Analyzer IIx	Illumina (Solexa)
Lumat LB 9507	Berthold
MS Imaging Plates	Fujifilm
NanoDrop ND-1000	PeqLab
Optimax	Protec
Peltier Thermal Cycler-200/225	MJ Research
Ultrospec 3000	Pharmacia Biospec
UV Stratalinker 1800	Stratagene

2.1.5. Primer

A p at the beginning of a DNA sequence denotes a phosphor group added to the oligonucleotide. Amino written at the end of a DNA sequence denotes an amino group added to the oligonucleotide, Biotin at the 5´-end denotes a biotin molecule coupled to this site.

Primer name	**Sequence 5´→ 3´**
BalRev3	AAAGCCATGACATCATTTTCTGGAATT
BalRev	CTTGTCATGAATTGTGATACAGTGAATTATAAGTG
CAT fw	GGC CTC ACG TAG TGA CCC GAC GCA CTT TGC GCC GAA T
CAT rv	CTG GAT CGA TCC ACC ATA CCC ACG CCG AAA CAA
erbB2/5	CGA TGT GAC TGT CTC CTC CCA AA
erbB2ΔE2C_fw	p-GAC AGC ACC AAG CTT GGC ATT CC
erbB2ΔE2C_rv	p-GAC ATG GCT CCG GCT GGA CCC GG
fw_NheI/SalI-SB	GCT GCT GCT AGC TGC TGT CGA CGG AAA ATC AAA AGA AAT CAG CCA AGA C
Illumina Pr1	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T
Illumina Pr2	CAA GCA GAA GAC GGC ATA CGA GCT CTT CCG ATC T
LAM SBL-20 hmr	ACT TAA GTG TATGTA AAC TTC CGA CT
LAM-SB/L-50Bio	Biotin-AGT TTT AAT GAC TCC AAC TTA AGT G
LAM SBleft II	ACA AAG TAG ATG TCC TAA CTG ACT
LAM-SBleft-Bio	Biotin-TGT AAA CTT CTG ACC CAC TGG AAT TG
Linker primer	GTA ATA CGA CTC ACT ATA GGG C
Nested primer	AGG GCT CCG CTT AAG GGA C
pUC2	GCG AAA GGG GGA TGT GCT GCA AGG
pUC5	TCT TTC CTG CGT TAT CCC CTG ATT C
SB-Not-rv	CTG AAT GCG GCC GCT AGT ATT TGG TAG CAT TGC CTT TAA ATT G
T-Jobb1	TTTACTCGGATTAAATGTCAGGAATTG
T-Jobb2	TGAGTTTAAATGTATTTGGCTAAGGTG
XhoI ohne ATG ZF4 fw	CGG TCT CGA GCT GGA ACC CGG CGA GAA GCC
ZF4 STOP ApaI rv	GTA CCG GGC CCT CAG CTG GTC TTT TTG CCA GTA TGG
19-3F	GTT TTC CCA GTC ACG ACG TT
19-3R	TGT GGA ATT GTG AGC GGA TA
Bar coded primers:	
SB III AAAA_OVH	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAAAGTAAACTTC CGACTTCAACTGTA
SB III TACC_OVH	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTACCGTAAACTTC CGACTTCAACTGTA
SB III CCCA_OVH	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCCAGTAAACTTC CGACTTCAACTGTA
SB III ATGC_OVH	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATGCGTAAACTTC CGACTTCAACTGTA
SB III CGTC_OVH	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTCGTAAACTTC CGACTTCAACTGTA
SB III TTTA_OVH	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTTAGTAAACTTC CGACTTCAACTGTA
SB III GGGA_OVH	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGGAGTAAACTTC CGACTTCAACTGTA

2.1.6. Oligonucleotides

Oligonucleotide name	**Sequence 5´→ 3´**
Linker 1A +	CTA GCG GGC AGC GGA GGG AGC GGC GGA AGT GGG GGC AGC GGC GGA AGT GGC G
Linker 1A -	TC GAC GCC ACT TCC GCC GCT GCC CCC ACT TCC GCC GCT CCC TCC GCT GCC CG
Linker 2A +	CTA GCG GGC ACC AGC AGT GGC GGA AGT GGG AGC AGT GGC AGT GGG GGC AGT G
Linker 2A -	TC GAC ACT GCC CCC ACT GCC ACT GCT CCC ACT TCC GCC ACT GCT GGT GCC CG
Linker 2B +	CTA GCG GGC ACC AGC AGT GGC GGA AGT GGG AGC AGT GGC AGT GGG AGT GGC GGA AGT GGG GGC AGT
Linker 2B -	TC GAC ACT GCC CCC ACT TCC GCC ACT CCC ACT GCC ACT GCT CCC ACT TCC GCC ACT GCT GGT GCC CG
Linker 3A +	CTA GCG CTG GCC GAG GCC GCT GCA AAG GAG GCC GCT GCA AAG GCA GCC GCT G
Linker 3A -	TCG ACA GCG GCT GCC TTT GCA GCG GCC TCC TTT GCA GCG GCC TCG GCC AGC G
Linker 3B +	CTA GCG CTG GCC GAG GCC GCT GCA AAG GAG GCC GCT GCA AAG GAG GCC GCT GCA AAG GCA GCC GCT G
Linker 3B -	TC GAC AGC GGC TGC CTT TGC AGC GGC CTC CTT TGC AGC GGC CTC CTT TGC AGC GGC CTC GGC CAG CG
Linker (KLGGGAPAVGGGPKAADK)+	GGC ATG CTA GCT AAG CTG GGC GGA GGC GCT CCT GCT GTG GGA GGA GGA CCT AAG GCT GCC GAC AAG GTC GAC ATC GG
Linker (KLGGGAPAVGGGPKAADK)-	CC GAT GTC GAC CTT GTC GGC AGC CTT AGG TCC TCC TCC CAC AGC AGG AGC GCC TCC GCC CAG CTT AGC TAG CAT GCC
Bfa linker +	GTA ATA CGA CTC ACT ATA GGG CTC CGC TTA AGG GAC
Bfa linker -	p-TAG TCC CTT AAG CGG AG-Amino
Sonic TA link(+)	GTA ATA CGA CTC ACT ATA GGG CTC CGC TTA AGG GAC CAT ACG AGC TCT TCC GAT CT
Sonic Link(-)amino	GAT CGG AAG AGC TCG TAT G-Amino

2.2. Methods

Standard molecular biology methods were executed as described in Sambrook and Russel (Sambrook and Russel 2001).

2.2.1. Cloning of plasmid constructs

All PCR reactions amplifying DNA used for cloning of vector constructs were done with PfuUltra Fusion HS DNA polymerase (Stratagene). All other PCR reactions were conducted using InviTaq DNA polymerase (Invitek).
All PCR-amplified DNA sequences used for cloning were verified by sequencing.

2.2.1.1. Cloning of E2C ZF/SB transposase fusion constructs

The vector pFV4aE2C was created by removing the sequence coding for 10xGly/SB by NheI and NotI digest from pFV4aE2C/10xGly/SB (kindly provided by Zoltan Ivics (Ivics et al. 2007)). The vector pF4a is described in (Caldovic and Hackett 1995). The hyperactive SB transposase mutant M3a was PCR-amplified using primers fw_NheI/SalI-SB and SB-Not-rv from CMV M3a (kindly provided by Zoltan Ivics), the PCR fragment was digested with NheI and NotI and cloned into pFV4aE2C resulting in pFV4aE2C/SB. Linker variants where constructed by partial digest of pFV4aE2C/SB using SalI cutting only at position 3323 bp between the E2C ZF and SB transposase, followed by NheI digest. Plus- and minus strand of oligonucleotides for linkers 1A, 2A, 2B, 3A and 3B (see 2.1.6) were annealed (see 2.2.7) producing SalI and NheI overhangs and ligated into NheI and partially SalI digested pFV4aE2C/SB. For practical reasons the ampicillin (amp) resistance gene originally present on pFV4a was removed from pFV4aE2C/SB by BspHI digest and the vector blunt-ended by Klenow fragment (Fermentas) treatment. The coding sequence of the chloramphenicol (cam) resistance gene was PCR amplified from pACYC184 (ATCC) using primers CAT fw and CAT rv and cloned into the prepared vector backbone. To clone pFV4aM3a SB transposase mutant M3a was cut out from pTRE2hygM3a (kindly provided by Andrea Schmitt from the laboratory of Zsusanna Izsvák) with XhoI and BamHI and blunt-ended through treatment with Klenow fragment (Fermentas). The vector pFV4a was cut by SmaI and the blunt-ended M3a coding region inserted. The plasmid pFV4aN57/E2C was kindly provided by Zoltan Ivics.

2.2.1.2. Creation of ZF proteins binding multi copy recognition sites in the human genome

LINE1.3 an element of the Ta subfamily of LINE1 elements was analyzed for suitable ZF binding sequences. The 3´-terminal 900 bp of *LINE1.3* containing the 3´-region of ORF2 and the 3´-UTR were subjected to a search for 18 bp sequences that seemed promising for ZF protein design using the "zinc finger tools" website (Mandell and Barbas 2006). Three auspicious sites were selected considering possible target site overlap problems, nucleotide triplet composition and predicted proximal *SB* insertion hot spots: ZF A recognition site: 5´-ACC AAC AGT GTA AAA GTG-3´; ZF B recognition site: 5´-GCC ATA AAA AAT GAT GAG-3´; ZF C recognition site: 5´-GGT GGG GTC GGG GGA GGG-3´. ZF proteins potentially binding to selected sites were designed also using the "zinc finger tools" website. The individual finger and corresponding binding site are indicated with corresponding ZF DNA recognition helices from amino acid -1 to +6 in brackets:

ZFA F1-GTG (RSDELVR), F2-AAA (QRANLRA), F3-GTA (QSSSLVR), F4-AGT (HRTTLTN), F5-AAC (DSQNLRV), F6-ACC (DKKDLTR)

ZFB F1-GAG (RSDNLVR), F2-GAT (TSGNLVR), F3-AAT (TTGNLTV), F4-AAA (QRANLRA), F5-ATA (QKSSLIA), F6-GCC (DCRDLAR)

ZFC F1-GGG (RSDKLVR), F2-GGA (QRAHLER), F3-GGG (RSDKLVR), F4-GTC (DPGALVR), F5-GGG (RSDKLVR), F6-GGT (TSGHLVR)

Codon-optimized synthetic genes encoding the designed ZF proteins were synthesized by GENEART (Regensburg). ZF genes were delivered in the backbone of pGA4. They are referred to as pGA4_ZFA, pGA4_ZFB and pGA4_ZFC.

2.2.1.3. Cloning of ZF activation domain (AD) fusion constructs

E2C/VP64 fusion plasmid was kindly provided by the laboratory of Carlos Barbas III. VP64 is an artificial tetrameric repeat of the viral VP16´s minimal activation domain which has been shown to produce five-fold stronger induction of transcription than VP16 (Beerli et al. 1998).

For LINE1 targeting Rep.TZ was removed from pcDNA.Rep.TZ.AD, which was kindly provided by Toni Cathomen (Cathomen et al. 2000), using EcoRI and NotI resulting in pcDNA_AD. The AD encoded in this plasmid was VP16. ZF proteins ZF A, B and C were cut out from pGA4_ZFA, pGA4_ZFB and pGA4_ZFC respectively using EcoRI and NotI and cloned into pcDNA_AD creating pcDNA_ZFA/AD, pcDNA_ZFB/AD and pcDNA_ZFC/AD respectively.

2.2.1.4. Cloning of LINE1 ZF transposase constructs

The E2C ZF domain was removed from pFV4aE2C/SBcam by SacII and NheI restriction digest. ZFB was cut out from pGA4_ZFB by SacII and NheI and cloned into the SB transposase gene containing vector backbone pFV4acam. This construct is referred to as pFV4aZFB/M3a. A linker (KLGGGAPAVGGGPKAADK) (Szuts and Bienz 2000) was introduced between ZFB and SB transposase. The vector pFV4aZFB/linker/M3a was cut with NheI and partial SalI and a ds oligonucleotide encoding the linker sequence was cloned in resulting in pFV4aZFB/linker/M3a. The transposase cloned into pFV4aE2C/SBcam was the hyperactive mutant M3a. To test whether a fusion protein containing SB100x instead of M3a showed increased transposition activity, constructs containing SB100x were also cloned. M3a was cut out of pFV4aZF4/M3acam with NheI and NotI digest. SB100x coding sequence was PCR-amplified from pCMV (CAT) T7 SB100x (Mates et al. 2009) using primers fw_NheI/SalI-SB and SB-NotI-rv, digested with NheI and NotI and cloned into the prepared vector backbone pFV4aZFB yielding pFV4aZFB/SB100xcam. To create pFV4aZFB/linker/SB100xcam the same procedure was performed on pFV4aZFB/linker/M3acam. To create pFV4aSB100xcam pFV4acam was digested with ApaI and SpeI. The E2C ZF coding sequence was removed from pFVN57/E2C (kindly provided by Zoltan Ivics) by ApaI digest followed by Klenow fragment fill and XhoI digest. ZF B was PCR amplified using primer XhoI ohne ATG ZF4 fw and ZF4 STOP ApaI rv, digested by XhoI, followed by insertion into prepared pFVN57. The SB100x coding region was removed from pCMV (CAT) T7 SB100x by ApaI and SpeI digest and cloned into the prepared pFV4acam backbone.

2.2.1.5. Cloning of luciferase reporter plasmids

The luciferase reporter plasmid erbB2 (Beerli et al. 1998) contained the genomic nucleotide sequence -758 to -1 relative to the *erb-B2* initiation codon including the E2C ZF binding site upstream a luciferase reporter gene. ErbB2 was kindly provided by the laboratory of Carlos F. Barbas III. For creation of the negative control plasmid the whole plasmid erbB2 was PCR-amplified excluding the E2C binding site using primers erbB2ΔE2C_fw and erbB2ΔE2C_rv. The PCR product was ligated together yielding the plasmid erbB2ohneE2C.

The luciferase reporter plasmid pGLtk.11.Luc containing an HSV-tk promotor upstream a firefly luciferase gene (kindly provided by Toni Cathomen) was cut with NheI and AgeI creating pGLtk.Luc. Oligonucleotides encoding plus and minus strand of ZFA, ZFB and ZFC

binding sites were annealed (see 2.2.7), cut with NheI and AgeI and cloned into pGLtk.Luc. Successful cloning was verified by sequencing.

2.2.2. Tissue culture and transfection

HeLa cells were cultured at 37°C and 5% CO_2 in Dulbecco´s modified Eagle medium (Gibco/Invitrogen) supplemented with 10% fetal calf serum (PAA) and Antibiotic-Antimycotic (Invitrogen). One day prior transfection $3x10^5$ or $1,5x10^5$ HeLa cells were seeded per 6-well or 12-well respectively. Transfections were done with QIAGEN-purified endotoxin-free plasmid DNA using jetPEI (Polyplus-transfection) according to manufacture´s protocol. Cells were typically harvested 48 hours post transfection. Transfection efficiency was monitored using control transfections with pEGFP.

2.2.3. Cell culture transposition assay

Transposition assays were done as described (Ivics et al. 2007). In this work 6-wells were transfected with 50-100 ng of transposase expression plasmid and 20-500 ng of the neomycin resistance gene carrying transposon plasmid pTneo. 48 hours post transfection, a fraction (1/2-1/5) of transfected cells was replated on 10 cm dishes and selected for transposon integration using 1,4 mg/ml G418 (Biochrom). Residual cells were pelleted and used for PCR-based transposon excision assay (Izsvak et al. 2004). After three weeks of selection cell colonies were fixed with 10% v/v formaldehyde in PBS, stained with methylene blue in PBS and counted.

2.2.4. PCR-based transposon excision assay

HeLa cells were transfected and processed as described in Ivics et al. (1997). Plasmids were extracted from cells using the QIAprep Spin Miniprep Kit (QIAGEN) following a modified manufacturer´s protocol. In step 2 of the QIAprep Spin Miniprep Kit Protocol P2 buffer was replaced by 1.2 % SDS with 1 µg/µl proteinase K. Samples were incubated for 30 min at 55 °C. After addition of N3 samples were incubated at 4 °C for 15 min. The first nested PCR was performed as follows: 2 µl plasmid extract, 1 µl 2.5 mM dNTP, 2 µl 10 pmol/µl primer pUC2, 2 µl 10 pmol/µl pUC5, 2.5 U Taq polymerase, 6 µl 25 mM MgCl2, 5 µl 10 x Taq buffer in a 50 µl reaction using the following PCR program: 94 °C 3 min followed by 30 cycles of 94 °C 30 sec, 58 °C 20 sec, 64 °C 10 sec ended by 72 °C 2 min. One microliter 1:100 diluted sample of the first PCR was used as template for the second nested PCR which was performed as follows: 1 µl 2.5 mM dNTP, 2 µl 10 pmol/µl primer 19-3F, 2 µl 10 pmol/µl

19-3R, 2.5 U Taq polymerase, 3 µl 25 mM MgCl2, 5 µl 10 x Taq buffer in a 50 µl reaction using the following PCR program: 94 °C 3min followed by 30 cycles of 94 °C 30 sec, 58 °C 20 sec, 64 °C 10 sec ended by 72 °C 2min.

2.2.5. Luciferase assay

HeLa cells were seeded in 12-well plates and transfected in duplicates with 250 ng luciferase reporter plasmid carrying a ZF binding site upstream of the luciferase gene, 125 ng of a plasmid expressing an AD/ZF fusion protein and 50 ng of a ß-galactosidase expression plasmid using jetPEI (Polyplus-transfection) following manufacturer´s protocol. 48 hours post transfection cells were lysed in 200 µl CCLR buffer (125 mM Tris-phosphate pH 7.8 H_3PO_4, 10 mM DTT, 10 mM 1,2-Diaminocyclohexan-N,N,N´,N´-tetraacetic acid, 50 % v/v glycerol, 5 % v/v Triton X-100) per well, vortexed for 15 sec, centrifuged and the supernatant was kept. For normalization of transfection efficiency a ß-galactosidase assay was performed. 15 µl of sample was incubated with 500 µl ß-galactosidase assay buffer (100 mM HEPES pH 7.3 KOH, 150 mM NaCl, 4.5 mM Aspartate (hemi-Mg salt), 1 % w/v bovine serum albumin, 0.05 % v/v Tween 20, 1.6 mM chlorphenol red- ß-D-galactopyranosid) until colour change. The reaction was then stopped by adding 250 µl 3 mM $ZnCl_2$ and absorbance was measured at 578 nm. To determine efficiency of cell lysis, protein content was measured using a Bradford assay. 1 ml 1:5 diluted Bio-Rad Protein assay solution (BioRad) was incubated with 10 µl sample. Following short incubation time absorbance was measured at 595 nm. For determination of luciferase activity 15 µl sample was added to 50 µl luciferase assay reagent (20 mM Tricine pH 7.8 NaOH, 1.07 mM $(MgCO_3)_4Mg(OH)_2$ x 5 H_2O, 2.67mM $MgSO_4$, 0.1 mM EDTA, 270 µM Coenzyme A, 470 µM luciferin, 530 µM adenosine 5´-triphosphate), vortexed and immediately measured with a 10-sec integration period in a luminometer (Lumat LB 9507 (Berthold)). ß-galactosidase values were calculated using the following formula: ß-gal units = 1000 x $OD_{578\ nm}$ / t x $OD_{595\ nm}$ where t is the time of incubation of the ß-galactosidase assay buffer with the sample before addition of $ZnCl_2$. Raw luciferase reads are normalized by using the following formula: normalized luciferase reads = raw luciferase reads / ß-gal units.

2.2.6. Competitive luciferase assay

HeLa cells in 12-well plates were transfected in duplicates with 50 ng luciferase reporter plasmid carrying a ZF binding site upstream of the luciferase gene, 15 ng of a plasmid expressing an AD/ZF fusion protein, 935 ng of a ZF transposase fusion protein and 50 ng of a

ß-galactosidase expression plasmid complexed with jetPEI (Polyplus-transfection) following manufacturers protocol. Cells were harvested and processed as described for the standard luciferase assay.

2.2.7. Generation of ds oligonucleotides/linkers

To generate ds oligonucleotides 50 µl of both, plus and minus strand of complementary oligonucleotides (each 100 pmol/µl) are boiled for five minutes in annealing buffer (50 mM NaCl, 10 mM TrisHCl, 10 mM $MgCl_2$, 1 mM DTT at pH 7.9). Annealing takes place during cool down of the water bath (1.5 l) overnight.

2.2.8. Inter-plasmid targeted transposition assay

6-well plates were triple-transfected with 200 ng transposon donor plasmid pTkan (kindly provided by Zoltan Ivics), 200 ng target plasmid erbB2 (kindly provided by the laboratory of Carlos F. Barbas III) and generally 50 ng of transposase helper plasmid. For the helper plasmids either E2C/SB, E2C/SB together with SB or SB alone were transfected. In case of transfections of E2C/SB together with SB 50 ng E2C/SB and 5 ng SB were applied. The target plasmid carries a promoter fragment encompassing nucleotides -758 to -1 relative to the ATG start codon of the erbB-2 gene cloned into pGL3basic (Beerli et al. 1998). As additional negative control a target plasmid (erbB2ohneE2C) identical to erbB2 except that it lacked the E2C binding site was transfected together with E2C/SB, unfused SB transposase and the donor plasmid pTkan. Forty eight hour post transfection cells were washed twice with PBS. Cells were lysed in 400µl lysis buffer (0.6 % SDS, 0.01 M EDTA). Through addition of 100 µl 5 M NaCl (1 M final concentration) high molecular weight genomic DNA precipitated. Low molecular weight DNA was subjected to phenol/chloroform and chloroform extraction and precipitated using 0.1 vol 2.5 M KAc pH 8, 10µg glycogen and 0.8 vol isopropanol. DNA pellets were washed twice with 70 % ethanol and resuspended in 10 µl ddH_2O. All 10 µl extracted plasmid DNA was electroporated into ElectroMAX DH10B cells according to manufacturer´s protocol. To avoid division and consequently amplification of individual transposition events bacteria were plated directly after electroporation on LB agar plates containing 100 µg/ml amp and 25 µg/ml kan. Plasmids from individual clones were extracted and checked for transposon insertions into the target plasmid by XmnI digest. Plasmids with promising restriction patterns were sequenced for transposon insertions.

2.2.9. Semi-nested locus-specific PCR

HeLa E2CTA cells were transfected and selected as described for the cell culture transposition assay (section 2.2.3). The HeLa E2CTA cell line contained at least one extra E2C recognition site flanked by 25x TA dinuleotides compared to the HeLa cell line used in our laboratory. After selection of transposon containing cells with G418, cells were either pooled or single clones picked and further cultivated in 12 or 24-well plates. Cells were washed twice in PBS and lysed adding PBS with 0.2 mg/ml Proteinase K at 50 °C for two hours. Before the first round of PCR Proteinase K was heat inactivated for 20 min at 80 °C. In the first PCR about 500 ng of genomic DNA, 20 mM Tris pH 8.4, 50 mM KCl, 3 mM $MgCl_2$, 0.4 mM dNTP, 0.2 pM primer erbB2/5 and BalRev3 or T-Jobb1, 2 U Taq-Polymerase were used. PCR program: 94 °C 4 min, 30 cycles of 94 °C 30 sec, ramp to 59 °C 1 °C/sec, 59 °C 30 sec, 72 °C 2.5 min. The second PCR was done on 1 µl of the first PCR , 20 mM Tris pH 8.4, 50 mM KCl, 2 mM $MgCl_2$, 0.2 mM dNTP, 0.2 pM primer erbB2/5 and BalRev or T-Jobb2, 2 U Taq-Polymerase. PCR program: 94 °C 4 min, 30 cycles of 94 °C 30 sec, ramp to 59 °C 1 °C/sec, 59 °C 30 sec, 72 °C 2.5 min.

2.2.10. LAM-PCR for Illumina sequencing

Linear amplification

A schematic for the LAM-PCR procedure is depicted in Fig. 7A. HeLa cells were transfected as described for a cell culture transposition assay. For transfections where E2C/SB, N57/E2C or N57/ZF B were supplemented with respective unfused transposase 1/20 of the amount of plasmid encoding the fusion protein was transfected from the plasmid expression the unfused transposase. About 10,000 HeLa cell clones carrying at least one transposon insertion were pooled. Genomic DNA was extracted from cells using the DNeasy blood and tissue kit (QIAGEN) according to manufacturer´s protocol. Genomic DNA was sheared with a Covaris S2 (Covaris) ultrasonicator using adaptive focused acoustics. Sheared genomic DNA fragments were of around 300 bp after ultrasonication. 500 ng of genomic DNA was applied to a linear amplification procedure. Linear PCR reactions were done in 20 mM Tris pH 8.4, 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM dNTP, 0.02 pM biotinylated primer LAM-SB/L-50Bio, 2.5 U Taq-Polymerase. PCR program: 94 °C 4 min, 50 cycles of 94 °C 40 sec, ramp to 54 °C 1 °C/sec, 54 °C 30 sec, 72 °C 1 min. After one round of linear PCR, an extra 0.5 µl of Taq Polymerase was added and samples were subjected to a second round of linear PCR.

Magnetic capture

Biotinylated linear PCR products were captured by streptavidin coupled beads (Dynabeads kilobaseBINDER Kit, Invitrogen). Upon exposure to a magnet DNA coupled to the beads was separated from supernatant. Magnetic capture was done according to manufacturer´s protocol.

Second strand synthesis

Linear PCR products were converted into dsDNA. The second strand synthesis reaction was performed using 2 µl 10 x Hexanucleotide mix (Roche), 0.25 mM dNTP, 2 U Klenow fragment (Fermentas), x µl H_2O in 20 µl on linear DNA coupled to magnetic beads.

End repair

Possible overhangs of dsDNA created by the second strand synthesis were blunt-ended and 5´-ends phosphorylated using the End It DNA End-Repair Kit (Epicentre Biotechnologies) according manufacturer´s instructions.

Adding A´-overhangs

A´-overhangs were added to the 3´-ends of DNA molecules. Magnetic beads were incubated in 5 µl NEB2 buffer, 0.2 mM dATP, 1 µl Klenow fragment Exo (-) (New England Biolabs) and 43 µl H_2O for 30 min at 37 °C following 20 min at 75 °C for heat inactivation.

Linker ligation

A double stranded sonic linker (see 2.2.7) with T´-overhangs was ligated to DNA coupled to beads using the Fast-Link DNA Ligation Kit (Epicenter Biotechnologies): 1 µl 10x ligation buffer, 1 µl 10 mM ATP, 1 µl Fast-Link ligase, 1 µl 50 pmol/µl sonic linker. Ligation reaction was carried out at 4 °C overnight.

Nested PCR

After ligation DNA plus linker coupled to beads was resuspended in 10 µl TE. 2 µl of bead suspension, 1 µl 10 mM dNTP, 1 µl 10 pmol/µl linker primer, 1 µl 10 pmol/µl LAM SBL-20 hmr, 2.5 U Taq polymerase, 3 µl 25 mM $MgCl_2$, 5 µl 10 x Taq buffer and H_2O to fill up the reaction to 50 µl were PCR amplified using the following PCR program: 94 °C 2 min followed by 35 cycles of 94 °C 30 sec, ramp to 55 °C 1 °C/sec, 55 °C 20 sec, 72 °C 30 sec ended by 72 °C 5min. A nested PCR was performed with 1 µl of PCR 1, 1 µl 10 mM dNTP, 1 µl 10 pmol/µl nested primer, 1 µl 10 pmol/µl primer SB III barcode_OVH, 2.5 U Taq polymerase, 3 µl 25 mM $MgCl_2$, 5 µl 10 x Taq buffer and H_2O to fill up the reaction to 50 µl were PCR amplified using the following PCR program: 94 °C 3 min followed by 35 cycles of 94 °C 30 sec, ramp to 51 °C 1 °C/sec, 51 °C 20 sec, 72 °C 30 sec ended by 72 °C 5 min. Barcoded primers annealed at the end of *SB* transposon. They were identical except for four bases within the primer. All PCR reactions ascribing to one transposase (fused, mixed or

unfused) were done with the same barcoded primer, so that transposon integrations mapped from PCR products could be traced back to corresponding transposases. A third PCR was performed to add overhangs necessary for recognition by the Genome Analyzer IIx (Illumina). 1 µl of PCR 2, 1 µl 10 mM dNTP, 1 µl 10 pmol/µl Illumina Pr1, 1 µl 10 pmol/µl Illumina Pr1, 0.5 µl Phusion polymerase, 10 µl Phusion buffer and H_2O to fill up the reaction to 50 µl were PCR-amplified using the following PCR program: 98 °C 30 sec followed by 14 cycles of 98 °C 10 sec, 65 °C 30 sec, 72 °C 30 sec ended by 72 °C 5 min.

PCR reactions were run on an agarose gel and verified for product sizes ranging from 100 bp to 500 bp. Equal amounts for all samples with differing barcodes were mixed together and loaded to a Genome Analyzer IIx (Illumina) analyzer

Sequencing with the Genome Analyzer IIx (Illumina)

With the Genome Analyzer IIx (Illumina) millions of DNA fragments can theoretically be sequenced from a single sample in a single flow cell.

DNA compatible to the primers of the Genome Analyzer are incorporated into sequencing sample DNA during PCR reactions. Sample DNA is attached to the surface of the flow cell channel and bridge amplified. Bridge amplification results in dense clusters of dsDNA. DNA sequence is determined using reversible terminator-base sequencing chemistry.

Bioinformatic analysis

DNA fragments read by the Genome Analyzer IIx (Illumina) were checked for the presents of intact *SB* transposon ends followed by a TA dinucleotide indicating SB transposase-mediated transposition opposed to random integration. DNA flanking the transposon end was aligned against the human genome using a basic local alignment tool (BLAST) search. The human reference sequence used for BLAST analysis was NCBI genome reference consortium human 37 (GRCh37) also referred to as hg19 published in February 2009. A transposon insertion at one particular site in the human genome was counted as a single insertion regardless of the number of reads obtained by the Genome Analyzer IIx (Illumina). Number of reads for single insertions varied between one to over one thousand. In some rare events up to 5,000 reads were detected for a single insertion. All transposon insertions are subjected to multiple PCR-amplification steps during LAM-PCR procedure. Detection of different numbers of reads for individual transposon insertion can result from PCR bias and different amplification rates as well as different amount of template at beginning of the LAM-PCR procedure due to multiple individual insertions at the same site. The LAM-PCR method unfortunately was not apt to detect targeting events at single TA dinucleotides in the human genome since multiple independant transposon insertions at the same TA dinucleotide will be read as one single

insertion. Insertions into DNA that map to multiple regions in the human genome were also discarded from the analysis. Genomic DNA around unique transposon insertions site was screened for full length ZF binding sites allowing zero to two mismatches. For ZF B also ZF binding sites of the type 5′-GCCNTANAANATGATGAG-3′ allowing zero to one mismatch were screened for since ZFs binding 5′-ANN-3′ triplets have loose specificity for the 5′-A. Genomic DNA around the transposon insertion site was also screened for ZF half sites. For comparison a randomTA data set was calculated which indicates the likelihood of a DNA insertion at any TA dinucleotide in the human genome.

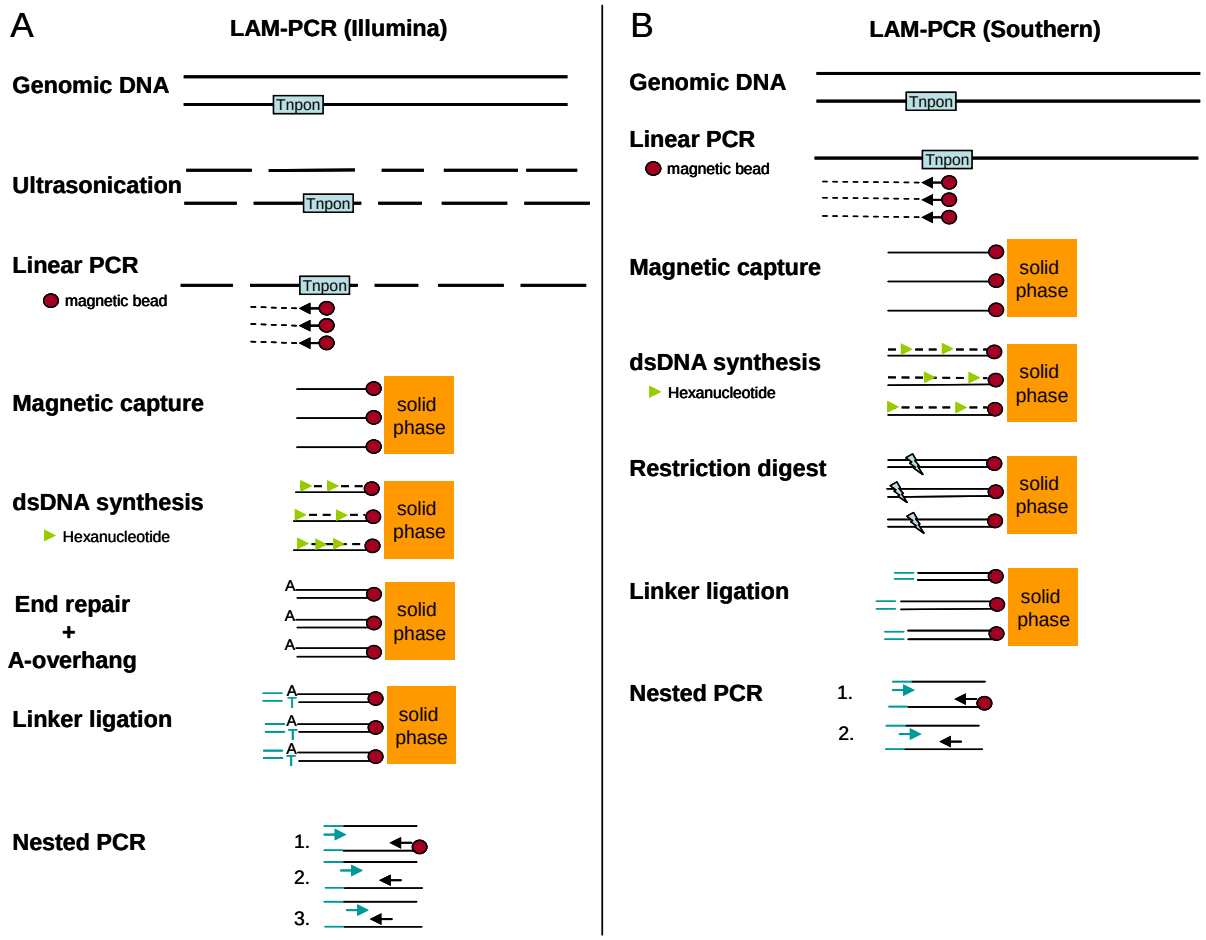


Figure 7. Schematic of the LAM-PCR procedure. (A) Schematic of the LAM-PCR protocol used for Illumina sequencing. **(B)** Schematic of the LAM-PCR protocol used for Southern Blot.

2.2.11. Southern Blot on LAM-PCR samples

A schematic for the LAM-PCR procedure is depicted in Fig. 7B.

Preparation of genomic DNA

HeLa cells seeded in 6-well plates, $3x10^5$ cells per well, were transfected with 100 ng transposase helper plasmid and 20 ng of transposon donor plasmid pTneo. For transfections

where E2C/SB, N57/E2C or N57/ZF B were supplemented with respective unfused transposase, 1/20 of the amount of plasmid encoding the fusion protein was transfected from the plasmid expression the unfused transposase. After 48 h cells were trypsinized, replated onto 10 cm petri dishes and selected for transposon insertions with 1,4mg/ml G418 (Biochrom). Selected cells were harvested and genomic DNA extracted following standard molecular biology methods.

Linear amplification

500 ng of genomic DNA was subjeted to a linear amplification procedure. Linear PCR reactions were done in 20 mM Tris pH 8.4, 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM dNTP, 0.02 pM biotinylated primer LAM-SBleft-Bio, 2.5 U Taq-Polymerase. PCR program: 94 °C 4 min, 50 cycles of 94 °C 40 sec, ramp to 59 °C 1 °C/sec, 59 °C 20 sec, 72 °C 1 min. After one round of linear PCR, an extra 0.5 µl of Taq Polymerase was added and samples were subjected to a second identical round of linear PCR.

Magnetic capture

Biotinylated linear PCR products were captured by streptavidin coupled beads (Dynabeads kilobaseBINDER Kit, Invitrogen). Upon exposure to a magnet DNA coupled to the beads was separated from supernatant. Magnetic capture was done according to manufacturer´s protocol.

Second strand synthesis

Linear PCR products were converted into dsDNA. The second strand synthesis reaction was performed using 2 µl 10 x Hexanucleotide mix (Roche), 0.25 mM dNTP, 2 U Klenow fragment (Fermentas), x µl H_2O in 20 µl on linear DNA coupled to magnetic beads.

Restriction digest

DsDNA coupled to beads was digested to create compatible ends for linker ligation. A restriction digest using FspBI (Fermentas) was prepared using proper buffer conditions and temperature. Samples were also digested with SacI which cut immediately outside the *SB* transposon in the pTneo vector backbone creating ends incompatible for linker ligation. This should eliminate propagation of non-SB transposase mediated pTneo vector insertions into the genome.

Linker ligation was done as described above (2.2.10) except that the double stranded Bfa linker was ligated instead of the sonic linker.

Nested PCR

After ligation DNA plus linker coupled to beads was resuspended in 10 µl TE. 2 µl of bead suspension, 1 µl 10 mM dNTP, 1 µl 10 pmol/µl linker primer, 1 µl 10 pmol/µl BalRev, 2.5 U Taq polymerase, 3 µl 25 mM $MgCl_2$, 5 µl 10 x Taq buffer and H_2O to fill up the reaction to

50 µl were PCR amplified using the following PCR program: 94 °C 3 min followed by 35 cycles of 94 °C 30 sec, ramp to 55 °C 1 °C/sec, 55 °C 20 sec, 72 °C 30 sec ended by 72 °C 5min. A nested PCR was performed with 1 µl of PCR 1, 1 µl 10 mM dNTP, 1 µl 10 pmol/µl nested primer, 1 µl 10 pmol/µl LAM SBleft II, 2.5 U Taq polymerase, 3 µl 25 mM $MgCl_2$, 5 µl 10 x Taq buffer and H_2O to fill up the reaction to 50 µl were PCR amplified using the following PCR program: 94 °C 3min followed by 35 cycles of 94 °C 30 sec, ramp to 51 °C 1 °C/sec, 51 °C 20 sec, 72 °C 30 sec ended by 72 °C 5min. Primers BalRev and LAM SBleft II annealed at the end of *SB* transposon. Linker primer and nested primer annealed on the added linker.

Southern Blot

After separation of 40 µl of the final nested PCR from the LAM-PCR protocol in a 2 % agarose gel, the gel was equilibrated in 0.4 NaOH for 30 min. A standard wet blotting Southern transfer stack was assembled with GB005 Whatman (Schleicher & Schuell) soaked in 0.4 NaOH and a nylon membrane (Hybond-XL, Amersham Biosciences) in 0.4 NaOH. Nucleic acids were transferred onto the nylon membrane over-night at room temperature. The blot was dried and DNA crosslinked using UV radiation (10000 µJ/cm^2). Radioactive probes that annealed to the genomic E2C locus from about 1.6 kb upstream of the E2C binding site to 1.1 kb downstream of the E2C binding site were prepared. Probe one covered base pairs 1882 to 1009, probe two covered base pairs 688 to 4 upstream the E2C recognition site. Probe three included the E2C recognition site and reached 449 bp downstream of it, probe four stretched from 449 bp to 1055 bp downstream the E2C recognition site. DNA fragments for preparation of probe one and two were cut out from the plasmid erbB2ΔE2C (Beerli et al. 1998) using restriction enzymes BsgI, AflII and NcoI. To provide DNA templates for probe three and four 1046 bp covering the E2C binding site and DNA downstream were amplified from genomic DNA using primer erb-B2/2 and E2C rec site fw and cloned into pGEM-T (Promega) resulting in pGEM-T 1046 bp E2C downstream. PGEM-T 1046 bp E2C downstream was restriction digested with SphI, HindIII and SpeI to isolate DNA fragments used for probe preparation. DNA used to prepare all four E2C probes was mixed together equimolar. A 1.2 kb DNA fragment covering the 3´-end of the element *LINE1.3* was used for making radioactive probes to detect LINE1 sequences in LAM-PCR reactions. The 1.2 kb *LINE1.3* fragment was cut out from the plasmid pJM101/L_{RP}Δneo (Wei et al. 2001) using BamHI. For labeling DNA α-^{32}P-dATP (Perkin Elmer; 3000 Ci/mmol, 10 mCi/ml) and the Prime-It II Random prime labelling kit (Stratagene) was used following manufacturer´s protocol. Unincorporated nucleotides were removed using illustra Microspin G-50 columns

(Amersham/GE Healthcare). After incubation of the membrane in 20x SSC, pH 7 for 5 min and 2x SSC, pH 7 for 5 min the blot was prehybridized at 68 °C for at least 30 min in Rapid-hyb (Amersham Biosciences) buffer. $1x10^6$ cpm/ml of both the E2C probe mix and the LINE1 probe were added to the prehybridisation buffer and blots hybridized at 65 °C over-night. Blots were washed with 2x SSC + 0.1 w/v % SDS and 0.1x SSC + 0.5 w/v % SDS at room temperature or 65 °C for several times. Radioactive signals were detected on phosphor imaging plates (MS Imaging Plates (Fujifilm)).

2.2.12. Western Blotting

Total protein was extracted from HeLa cells according to standard molecular biology methods. Protein amount was measured using the BCA Protein Assay Reagent Kit (BioRad) according to manufacturer´s instructions. Equal amounts of protein were separated on a 12.5 % polyacrylamide gel at 170 V. Separated proteins were transferred from the polyacryamide gel onto nitrocellulose membranes (Hybond-C Extra (Amersham/GE Healthcare)) at 0.4 Amp for one hour at 4 °C. Nitrocellulose membranes were blocked with 10 % nonfat dried milk in TBS-T (200 mM Tris Base, 0.05 % v/v Tween 20, 50 mM Tris HCl, pH 7.4) for one hour at room temperature. Incubation with the primary antibodies (polyclonal Anti-SB Transposase antibody 1:1,000; monoclonal Anti-Actin antibody 1:400) was done in 5 % nonfat dried milk in TBS-T over-night at 4 °C. Before addition of the secondary antibodies the nitrocellulose membranes were washed four times 10 min in TBS-T. The nitrocellulose membrane was incubated with secondary antibodies (Mouse Anti-Goat IgG; Goat Anti-Mouse IgG, both peroxidase conjugated, both 1:5,000) for 1.5 h at room temperature. After washing five times for five minutes in TBS-T, bound antibodies were visualized by chemiluminescence (ECL Plus Western blotting detection system (Amersham/GE Healthcare)) according to manufacturer´s protocol.

2.2.13. Statistical analysis

All experiments (except for Illumina sequencing) were conducted at least three times. Error bars represent the standard error of the mean (sem). The standard error of the mean represents the standard deviation divided by the square root of number of experiments done.

P-values were calculated for differences in distribution of transposon insertions within the human genome obtained with Illumina sequencing.

3. Results

The aim of my work was to target insertions of the *SB* transposon near selected sites in the human genome. My strategy to pursue this aim was to fuse the enzymatic component of the *SB* transposon system, the transposase, to a DBD: the attached DBD was expected to tether the transposon/transposase complex to its recognition site and the transposase part of the protein was expected to integrate the transposon into a TA dinucleotide at adjacent DNA sequences. A DBD eligible for a targeting a DBD/transposase fusion protein had to fulfill certain criteria: an endogenous unique recognition site in the human genome and specific binding of this recognition site. The recognition site of the DBD had to be accessible for binding and transgene insertion and provide the environment for sustained gene expression. The artificial six finger ZF protein E2C (Beerli et al. 1998) represented a good candidate as DBD fusion partner for the SB transposase binding a single-copy target in the human genome. Since no ZF proteins binding a multi-copy target in the human genome had been published so far, I designed and tested three ZF proteins that bound to targets in the 3´-region of LINE1 in the human genome.

3.1. E2C/SB fusion proteins exhibit reduced transposition activity

Bringing two individual proteins close together may force the individual fusion partners to misfold and thus influence their individual activity. Fusion of tags to the catalytically active C-terminus of SB transposase resulted in most of the cases in complete loss of transposition activity (Yant et al. 2007) (Ivics et al. 2007). Fusions to the N-terminus of the SB transposase which is responsible for transposon binding and protein-protein interaction seem to be tolerated to a better extent, even though N-terminal fusions also show a decrease in transpositional activity compared to unfused SB transposase (Wilson et al. 2005) (Yant et al. 2007) (Ivics et al. 2007). Looking at other published fusion proteins between a DBD and a transposase, the order DBD/transposase generally seems to be the preferred choice (Wu et al. 2006) (Maragathavally et al. 2006). This may be due to the fact that, like the SB transposase, most transposases have their catalytically active site at their C-terminus. For these reasons I fused the E2C ZF to the N-terminus of SB transposase in order to create a fusion protein with good transpositional activity. Including a linker between both fusion partners could help to reduce the influence that both fusion partners have upon each other. I inserted various linkers with different structural properties and lengths between both fusion partners and examined the efficiency in transposition of the fusion protein. Linker 1 (GSGGSGGSGGSGGSG) and

linkers 2 (2A: GTSSGGSGSSGSGGS; 2B: GTSSGGSGSSGSGSGGSGGS) were assumed to be flexible due to their composition of simple amino acids like glycine and serine allowing individual proteins to act less constraint. Stability and functionality of proteins fused together with various serine/glycine linkers were tested in Robinson and Sauer (1998). For biological activity of the composite protein a minimum length of eleven residues was determined for the linker. However, in other reports, fusion proteins with shorter linkers also showed biological activity (Yant et al. 2007). Robinson and Sauer (1998) further state that composition rather than sequence of the linker contributes to protein stability. Serine containing linkers were shown to slow down unfolding of the single chain protein examined in this work. Linkers 3 (3A: LAEAAAKEAAAKAAA; 3B: LAEAAAKEAAAKEAAAKAAA) were assumed to form a helical structure due to interaction between positively charged lysine and negatively charged glutamate residues giving it a rigid structure which would keep both individual proteins apart from each other. Linkers of this composition had shown the ability to control the distance of two proteins fused together and reduce interference between both domains (Arai et al. 2001). Linkers with the suffix A were 15aa, linker with suffix B were 20 aa in length.

No eminent differences of transposition activity between fusion constructs containing different linkers were observed using a cell culture transposition assay. In a cell culture transposition assay a transposon donor plasmid which carried an antibiotic resistance gene and a transposase expressing helper plasmid were transfected into HeLa cells. After excision from the donor plasmid the transposon was integrated into genomic DNA. Cells with transposon insertions survived selection with the corresponding antibiotic. After the selection process surviving cell colonies were fixed and stained (Fig. 8A). All constructs showed about two to three-fold reduced transposition activity compared to unfused SB10 transposase but still about double as much integrations than seen for the negative control (Fig. 8C).

Capability of E2C/SB fusion proteins to excise a *SB* transposon from a plasmid was shown using a PCR-based method called excision PCR (Fig. 8D). Excision PCR is performed with primers annealing at the transposon donor plasmid backbone that either amplify the full transposon or a shorter product in case of excision of the transposon from the plasmid (Fig. 8E).

Due to best performance in transposition as well as PCR-based excision assay, the E2C/SB protein with the minimal linker (LAAVD) was chosen for subsequent experiments (Fig. 8B).

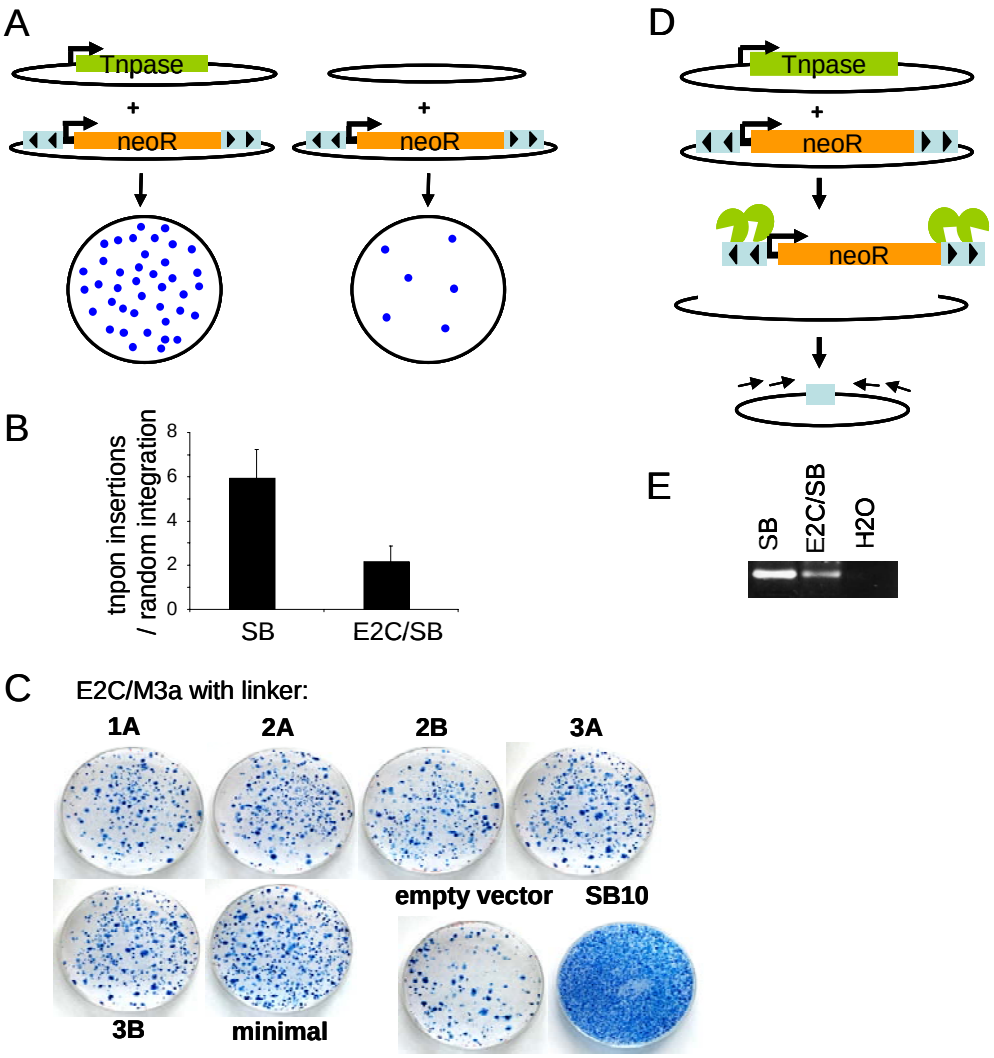


Figure 8. Transposition activity of E2C/SB fusion proteins. (A) Schematic of a cell culture transposition assay. For use of the transposon as DNA delivery tool the enzymatic component the transposase and the DNA component of the transposon were separated: a helper plasmid provided the transposase and a donor plasmid carried the transposon which encoded an antibiotic resistance gene. Both plasmids were cotransfected into HeLa cells. 48 hours post transfection cells were treated with an antibiotic selecting for cells with transposon insertions. Selected cells were fixed and stained. Black circle with blue dots represents a petri dish with stained cell colonies. The difference in cell colony number between cotransfections of transposase and transposon and a control transfection with no or transpositionally inactive transposase and transposon represents the transpositional activity of the transposase transfected. **(B)** Cell culture transposition assay. Bars represent fold transposition of SB (M3a) transposase and E2C/SB transposase (containing the minimal linker LAAVD) compared to negative controls. Error bars represent the standard error of the mean. **(C)** Cell culture transposition assay. Petri dishes with stained HeLa colonies produced by E2C/SB variants containing different linkers. **(D)** PCR-based transposon excision assay. After excision of the transposon from the donor plasmid, the plasmid backbone was ligated together by host cell proteins. PCR with primers annealing on the plasmid backbone pointing towards the transposon either amplified the transposon or a smaller fragment in case of excision. **(E)** PCR fragments corresponding to donor plasmid with excised transposon were detected for transfections of unfused and E2C/SB transposase.

3.2. The E2C/SB fusion protein binds to the E2C recognition site

The E2C ZF binds its recognition site with good affinity and specificity. ZFs are naturally found in fusions with other regulatory domains like in transcription factors, and may thus not react sensitive to fusion to other domains by losing their ability to bind DNA. However, in order to check whether a fusion protein consisting of the E2C ZF and the SB transposase still bound the E2C recognition site, I performed competition-based luciferase reporter assays. In this assay a luciferase reporter plasmid carrying an E2C binding site upstream a firefly luciferase gene and a plasmid expressing a VP64 activation domain (AD) fusion protein were transfected into HeLa cells. Upon binding of the E2C ZF to its recognition site luciferase expression was induced by the AD. Cotransfection of the E2C/SB fusion protein diminished induction of luciferase expression by competing with the E2C/AD fusion protein for binding to the E2C recognition site (Fig. 9A).

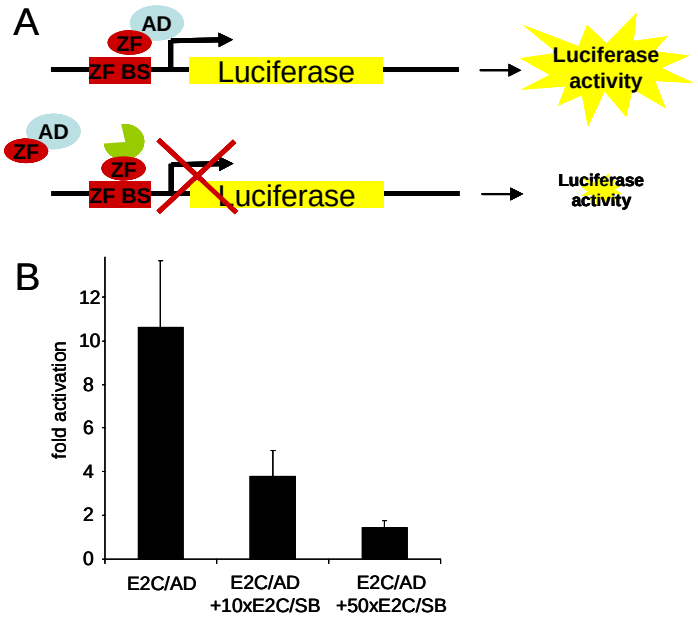


Figure 9. Competitive luciferase reporter assay. (A) Schematic of the competitive luciferase assay. Competitive luciferase assays were performed to detect DNA binding activity of ZF/transposase fusion proteins. HeLa cells were transfected with a luciferase reporter plasmid containing a ZF binding site (ZF BS) upstream a minimal promoter, activator protein (ZF/AD) and an excess of competitor protein (ZF/transposase fusions; red oval and green three quarter square). Binding of the competitor protein results in reduced luciferase *trans*-activation. **(B)** E2C recognition site-binding of the E2C/SB fusion protein was examined using a competitive luciferase assay transfecting E2C/AD fusion proteins with no competitor (first bar) and a tenfold and a fiftyfold molar excess of competitor (second and third bar). Luciferase reporter assays were done in HeLa cells. Bars represent normalized relative luminescence units (RLUs) for activator proteins (ZF/AD) divided by RLUs for a negative control (AD). Error bars represent the standard error of the mean.

The viral AD VP64 fused to the E2C ZF induced luciferase activity over ten-fold compared to AD only. When E2C/SB transposase fusion protein was added to this transfection in a ten-fold excess induction of luciferase activity decreased to fourfold compared to transfections of

the luciferase reporter plasmid and AD only. With a 50-fold excess of the E2C/SB transposase fusion construct luciferase induction further decreased to below twofold of control transfections (Fig. 9B).

3.3. E2C/SB fusions slightly alter transposon integration pattern in plasmid context

Before going into genomic context I first examined targeting ability of the E2C/SB fusion protein in an inter-plasmid transposition assay. In the human genome a single complete E2C recognition site exists for E2C ZF binding as opposed to approximately 180 million TA dinucleotides in the human genome sufficient for *SB* transposon insertion. Since the SB transposase in the E2C/SB fusion protein still contains its intrinsic ability to bind DNA, off-target transposon insertions are likely to occur. On a smaller DNA scale like a 5.5 kb plasmid I expected changes of transposon integration patterns due to ZF binding to be more noticeable and easier to detect than in the context of 3 billion bp of the human genome. Once changes in the transposon integration pattern on plasmid level indicated successful targeting, I examined targeting on genomic level. For the inter-plasmid transposition assay three plasmids were transfected into HeLa cells (*i*) a donor plasmid encoding a kan-marked transposon (*ii*) a cam resistant helper plasmid encoding E2C/SB or SB transposase alone (*iii*) an amp resistant target plasmid containing nucleotides -758 to -1 relative to the ATG initiation codon of the human gene *erbB-2* including the E2C binding site. The target plasmid was 5.55 kb in size and contained 280 TA dinucleotides each providing a possible *SB* integration site. In cells containing all three plasmids the transposon could get excised and in rare events inserted into the target plasmid. Plasmids extracted from HeLa cells were electroporated into *E.coli* cells and screened for amp/kan double resistance indicating target plasmids containing a transposon insertion. With an adaquate restriction digest target plasmids containing a transposon insertion were further validated (Fig. 10). Insertions were mapped onto the target plasmid and compared between different transposases. Insertions into the antibiotic resistance gene or the origin of replication (ori) were not expected to be recovered.

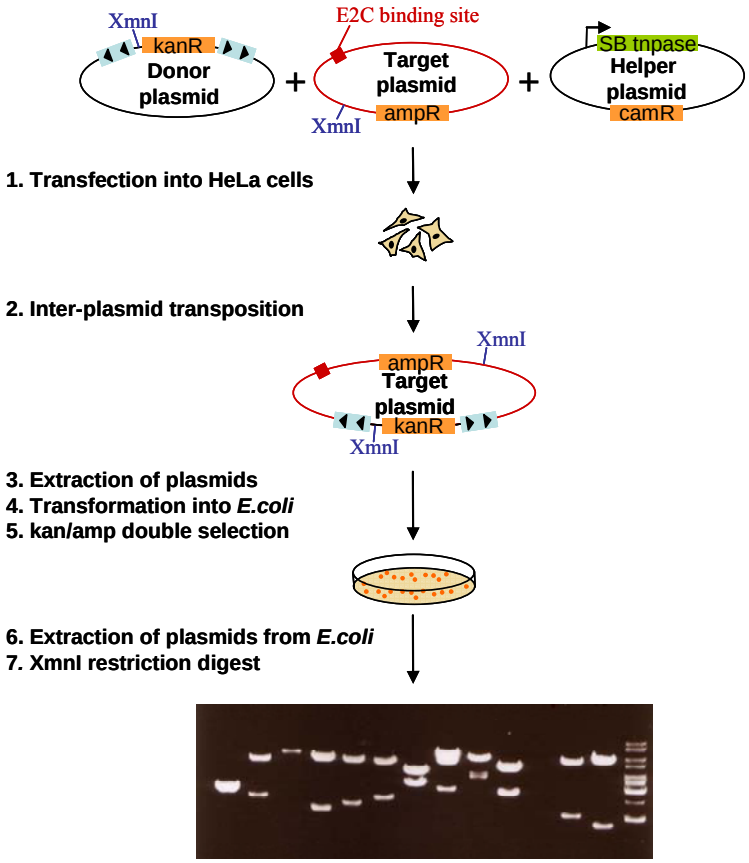


Figure 10. Schematic of the inter-plasmid transposition assay. A donor plasmid containing an *SB* transposon carrying a kan resistance gene, a target plasmid containing the E2C binding site and a helper plasmid expressing SB transposase were transfected into HeLa cells. Plasmids were recovered from HeLa cells 48 hours post transfection and electroporated into *E.coli*. Using kan/amp double selection only target plasmids containing a transposon insertion were able to grow. Plasmids from kan/amp resistant bacteria were extracted and validated for correct band sizes after XmnI digest. Exact position of transposon insertions sites were identified by sequencing.

E2C/SB fusion proteins showed decreased transposition activity compared to unfused SB transposase. During transposition the SB transposase is believed to form a tetramer (Izsvak et al. 2002). Heterotetramers of unfused and DBD/SB fusion transposase could theoretically target transposon insertions with improved transpositional activity. Such heterotetramers are expected to tether the transposition complex to the target site with the DBD of the fusion protein and efficient transposition will be mediated by unmodified full length transposase. Proof-of-principle for this strategy was published by Ivics et al. (2007). To test this strategy I transfected the E2C/SB fusion protein together with unfused SB transposase. Since unfused SB transposase can form homotetramers as well as heterotetramers the amount of unfused SB transposase in relation to E2C/SB should be limited. Unfused SB transposase was added to transfections with E2C/SB and transposon donor plasmid in different ratios. The lowest amount of unfused SB sufficient to increase transposition efficiency was used for further experiments. Maps of transposon insertions on the target plasmid were established for the

E2C/SB fusion protein (E2C/SB), cotransfections of E2C/SB and unfused SB transposase (E2C/SB + SB) and unfused transposase (SB) as a control. For E2C/SB transfections 84 transposon integrations (Fig. 11A), for E2C/SB + SB transfections 77 transposon integrations (Fig. 11B) and for SB alone 95 transposon integrations (Fig. 11C) were mapped on the target plasmid erbB2. For E2C/SB + SB transfections also 41 transposon integrations into the target plasmid erbB2ohneE2C were mapped (Fig. 11D). ErbB2ohneE2C was identical to erbB2 except that it lacked the E2C recognition site. Insertion patterns on the target plasmids looked overall similar. However, in a region about 1 kb upstream of the E2C binding site just outside the cloned *erbB-2* promoter sequence an enrichment of transposon insertions could be observed using E2C/SB and E2C/SB + SB (23 % and 25 % respectively) compared to 9.4 % for SB only. On the target plasmid erbB2ohneE2C transposon integrations into this region dropped back to 12 %.

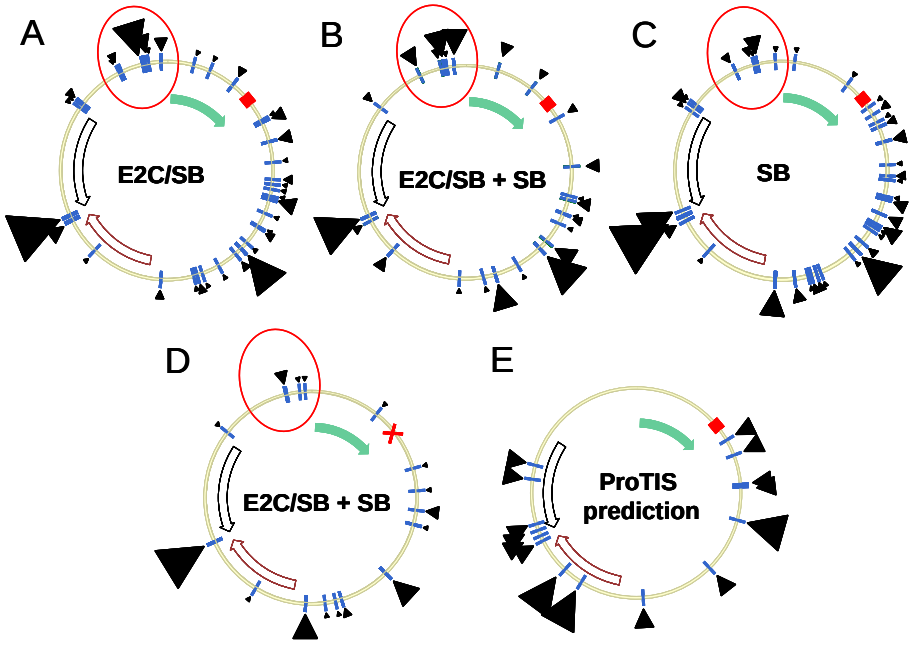


Figure 11. Transposon integration profiles on the target plasmid erbB2 for different SB transposases. An inter-plasmid transposition assay was done to compare transposon integration profiles for E2C/SB, E2C/SB mixed with unfused SB and unfused SB transposase. Each transposon integration (black arrowheads) recovered from the inter-plasmid transposition assay was mapped on the target plasmid erbB2 (5.5 kb). Transposon insertions into the same TA or close to each other were combined into one arrowhead. The bigger the arrowhead the more insertions at this particular region. Target plasmids contained the E2C recognition site (red square) and an adjacent 758 bp fragment from the *erbB2* gene (filled green arrow). No insertions into the antibiotic resistance gene (black unfilled arrow) and only few integrations into the ori (red unfilled arrow) could be recovered due to the experimental set up. The red circle indicates a DNA region of 1 kb upstream the E2C binding site. **(A)** Transposon integrations (n=84) into the target plasmid erbB2 for E2C/SB transposase. **(B)** Transposon integrations (n=77) for E2C/SB and unfused SB transfected in a 10:1 ratio. **(C)** Transposon integrations (n=95) into the target plasmid for SB transposase. **(D)** Transposon integrations (n=41) in a target plasmid identical to erbB2 which lacked the E2C binding site using E2C/SB and SB in a 10:1 ratio as transposase source. **(E)** Prediction of *SB* transposon insertion preference using the program ProTIS. Predicted hotspots for transposon insertions are depicted with black arrowheads, big arrows heads indicate 4-peak preferred TA dinuleotides, small arrowheads indicate 3.5-peak semi preferred TA dinucleotides.

3.4. Transposon insertions near the E2C binding site in human cells

After observing slightly shifted transposon insertion patterns in the inter-plasmid transposition assay I examined targeting ability of E2C/SB in genomic context.

For targeted transposon integration upon binding of the E2C ZF the SB transposase has to find DNA fulfilling *SB* insertion criteria to integrate the transposon. As mentioned before, a strict integration requirement for the *SB* transposon is a TA dinucleotide. On plasmid level *SB* was reported to prefer insertion into TA-rich DNA in general (Liu et al. 2005). This was confirmed by analysis of 138 unique *SB* insertions into the human genome where the consensus sequence for insertions was found to be tandem TA-repeats. However target site selection was found to be primarily dependent on physical properties of DNA not on the actual sequence (Vigdal et al. 2002). Later publications analyzing genomic *SB* integrations also suggest that structural characteristics of DNA like bendability or protein-induced deformation determines preferred integration sites (Yant et al. 2005) (Geurts et al. 2006). However, in general common insertion sites for *SB* were found to be palindromic AT sequences. The *erbB-2* locus in the human genome around the E2C binding site is rather GC-rich. GC-content of 1500 bp upstream of the E2C binding site is 51 %, GC-content for 1500 bp downstream the E2C binding site is 60 %. Analysis of the *erbB-2* promoter region with ProTIS (Geurts et al. 2006), a program that analyses DNA for potential *SB* integration hot spots, showed very few promising TA dinucleotides in the region 750 bp upstream the E2C binding site. To offer eligible *SB* insertion sites near the E2C binding site a HeLa cell line carrying at least one extra E2C recognition site flanked by 25x TA dinuleotides was created. This cell line is from now on referred to as HeLa E2CTA**.** HeLa E2CTA cells were transfected with E2C/SB or SB and a donor plasmid carrying a neoR marked transposon. Cells were selected for transposon insertions by neomycin treatment and 48 clones picked per transposase. Semi-nested E2C locus specific PCR was performed using one primer annealing in the *erbB-2* locus (711 bp upstream of the E2C binding site) and two nested primers annealing in the transposon IRs. A product would thus only be produced if a transposon insertion occurred around the E2C binding site (Fig. 12A). No product was seen for SB transposase only. For the E2C/SB fusion protein one out of these 48 clones produced a band. The product was cut out from an agarose gel and sequenced. BLAST analysis mapped a proper transposon insertion 179 bp upstream the E2C binding site (Fig. 12C).

In a different experiment HeLa E2CTA cells were transfected with a neoR marked transposon donor plasmid and E2C ZF fusions to N57 or full length SB transposase and unfused SB

transposase. Full length SB transposase was connected through a 10x glycine linker both C- and N-terminal to the E2C ZF. Both of these N- and C-terminal fusions had shown no transposon excision activity using PCR-based transposon excision assay.

As mentioned before, full length SB transposase is expected to perform the catalytical steps of the transposition process, whereas catalytically inactive E2C/10xglycine/SB, SB/10xglycine/E2C and N57/E2C are expected to lead the synaptic complex consisting of the transposase tetramer and excised transposon to the E2C binding site. Cells selected for transposon insertions were pooled. Semi-nested E2C locus-specific PCR was performed on pooled cells (approximately 500 individual cell clones) as described above. No products were observed for control transfections with unfused SB+transposon and transposon only. One product was seen for SB/E2C and unfused SB transposase using an *erbB-2* specific primer and a nested primer set annealing in the left IR of the *SB* transposon. A second product was detected for transfections of N57/E2C and unfused SB transposase using the *erbB-2* specific primer and a nested primer set annealing in the R-IR. Both band were excised from an agarose gel and sequenced (Fig. 12B). The first product showed to be an insertion into the *erbB-2* promoter region 730 bp upstream the *erbB-2* start codon. The second product could unfortunately not be identified despite repeated sequencing.

Semi-nested E2C locus specific PCR with transposon-specific primers and a primer annealing to the E2C binding site were performed to amplify transposon insertion at the artificially introduced E2C binding site. No products were obtained using these primer combinations nor combinations of transposon-specific primers and primers annealing at various other positions around the *erbB-2* locus.

After these encouraging results I decided to do an unbiased and extensive mapping of transposon insertions into the human genome.

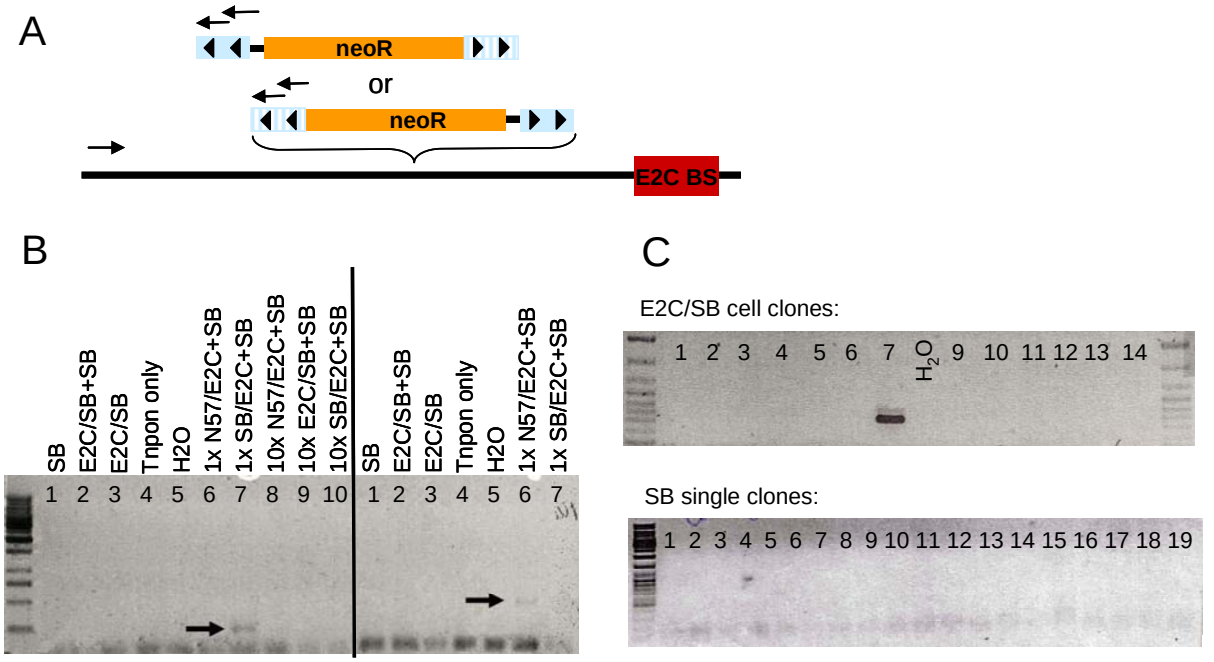


Figure 12. Semi-nested E2C locus-specific PCR for detection of targeted transposon insertions of E2C/SB transposase fusion proteins. (A) Schematic of the semi-nested E2C locus specific PCR. The E2C locus-specific primer annealed 750 bp upstream the E2C binding site at the human genomic *erbB-2* locus. Two nested pairs of primers annealed within the left or right IR of the *SB* transposon. **(B)** HeLa cells were transfected with different transposase helper plasmids and the transposon donor plasmid pTneo carrying a neomycin resistance gene. Cells selected for transposon insertions were pooled, genomic DNA extracted from cell pools and subjected to semi-nested E2C locus-specific PCR. 1x transposase fusion protein+SB indicates transfections of equal amounts of transposase proteins. 10x transposase fusion protein+SB indicates transfections with fusion transposase plasmids and unfused SB transposase in a ratio 10:1. **(C)** HeLa cells were transfected with E2C/SB transposase helper plasmids together with the transposon donor plasmid pTneo carrying a neomycin resistance gene. Cells were selected for transposon insertion, single cell clones picked and cultivated. Genomic DNA was extracted from single cell clones and subjected to semi-nested E2C locus-specific PCR.

3.5. Targeting LINE1 elements

Most transposon insertions mediated by E2C/SB or combinations of E2C fusion proteins with unfused SB occured off-target. Insertion at a single site in the human genome may be severely underrepresented considering that this unique site theoretically competes with all TA dinuclotides in the human genome for transposon insertion. Using a target site that existed at high copy number in the human genome could increase the chances of targeted transposition events over off-target transposon integration. The Ta subset of the L1H subfamily of LINE1 elements exists in approximately 500 copies in the human genome. Elements of this subset have shown recent transpositional activity (Boissinot et al. 2000), indicating that they reside in genomic context that could be accessible for proteins. LINE elements are often 5´-truncated but exhibit a 3´-region including ORF2 and 3´-UTR which is to some extent conserved throughout the LINE1 family (Boissinot et al. 2000). Sequences within this region were therefore expected at much higher copy number in the human genome than for Ta elements

only. Analysing the 3´-region of the Ta element for favourable *SB* insertion sites using the *SB* insertion site prediction program ProTIS (Geurts et al. 2006) showed that DNA composition in this region could provide an eligible target for *SB* transposition (Fig. 13). The 3´-region of this LINE1 element therefore seemed to be a good target choice for proof-of-principle targeting with *SB*. No ZF protein binding within this or a similar region had been published so far. However, some publications claimed that ZF proteins could be designed to bind any sequence in the human genome (Mandell and Barbas 2006) (Wei et al. 2008). For this reason I decided to create ZF proteins by rational design that bound within the 3´-region of the *LINE1.3* element which belongs to the Ta subset of LINE1 elements. As mentioned before, targeting transposon insertions near LINE1 elements serveed as proof-of-principle experiments. Transgene integration near retrotranspositional-active LINE1 elements bears the risk of mobilisation of such elements with the risk of insertional mutagenesis at the new LINE1 integration site.

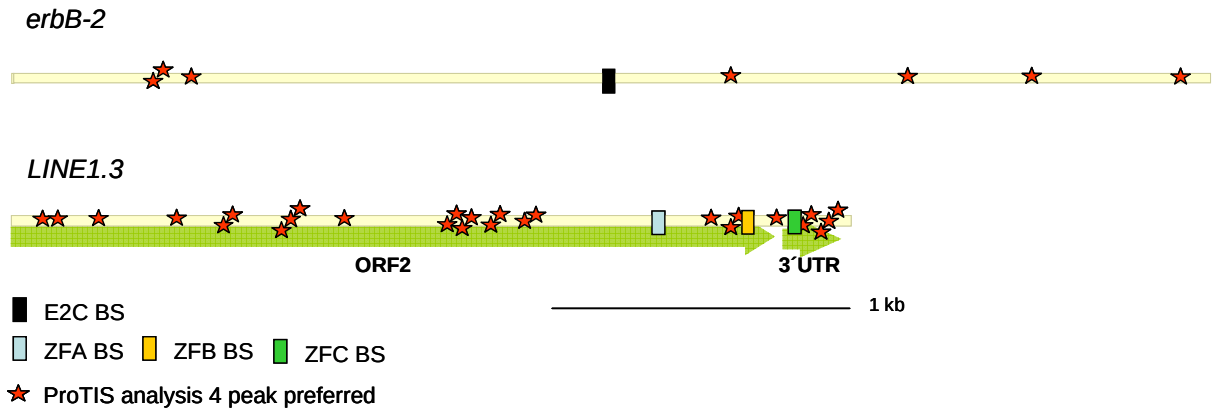


Figure 13. Analysis of genomic DNA around the ZF target sites for potential *SB* insertion site hot spots using the *SB* insertion site prediction tool ProTIS. Two kb up- and downstream of the E2C BS in the *erbB2* genomic region and the 3´-terminal 2814 bp of the *LINE1.3* element are depicted. Red stars indicate TA dinucleotides which are ranked as highest preferred for *SB* transposon integration corresponding to ProTIS analysis.

3.6. Two out of three ZFs designed by modular assembly bind their predicted recognition sequence

The ZF tools website (http://www.scripps.edu/mb/barbas/zfdesign/zfdesignhome.php) was used to select three 18 bp DNA sequences within the 3´-region of a Ta element of the LINE1 family of retrotransposons (Fig. 13 and Table 1).

Table 1. ZF recognition sites.

Name	ZF binding site 5´-3´
ZF A	ACC AAC AGT GTA AAA GTG
ZF B	GCC ATA AAA AAT GAT GAG
ZF C	GGT GGG GTC GGG GGA GGG

Using the same website amino acid sequences for polydactyl ZF proteins that potentially bound these DNA sequences were designed. ORFs encoding predicted ZF proteins were deduced, codon optimized for expression in human cells and synthesized by GENEART AG (Regensburg) through assembly from synthetic oligonucleotides. To test binding of ZF proteins to their predicted recognition sites, ZF ORFs were fused to a VP16 AD. Predicted ZF recognition sites were cloned upstream a luciferase gene, which was under the control of a HSV-tk promoter. Binding of synthetic ZF proteins to their predicted recognition sites in HeLa cells was examined using a luciferase reporter assay. Luciferase expression was induced by binding of the ZF/AD fusion protein to the predicted ZF recognition site (Fig. 9A). Two of the ZF proteins, ZF A and ZF B, showed fourfold stronger induction of luciferase expression compared to AD domain only control transfections (Fig. 14A). Unspecific binding of ZF B was tested by transfecting ZF B/AD together with a reporter plasmid lacking the ZF B binding site. On a control reporter plasmid lacking the ZF B recognition site ZF B/AD showed an approximate 1.5-fold increase in luciferase expression compared to transfections with AD only (Fig. 14B). As a reference the E2C ZF showed a ninefold increase in luciferase expression (Fig. 14A).

I decided to use ZF B for creation of targeting proteins and further experiments.

3.7. ZF B/SB fusion proteins compete with ZF B/AD fusion proteins for binding to the ZF B binding site

Fusion of the SB transposase to ZF B may alter protein structure and thus binding capacity of the ZF. I tested binding of different ZF/SB fusion proteins to the ZF B recognition site by a competition based luciferase reporter assay (Fig. 9A). Fusions of ZF B to the VP16 AD lead to an approximately fourfold increase of luciferase expression compared to VP16 activation domain only (Fig. 14C). Supplementing ZF B/SB to this experiment was expected to reduce activation of luciferase expression since ZF B/AD and ZF B/SB were expected to compete for binding to the ZF B binding site. Luciferase expression was reduced to 1.5- to 2.5-fold of ZF B/AD only transfections depending of ZF/SB construct transfected. Transfection of unfused transposase reduced ZF B/AD induced luciferase activation to 3.5-fold compared to

transfections with AD only (Fig. 14C). The ZF B/SB100x and the N57/ZF B construct reduced luciferase expression most extensively.

Off-target integration of the transgene is highly undesired in targeting approaches. For this reason binding to its recognition site was a key property of the ZF B/SB construct. For subsequent targeting experiments I chose to use the ZF B/SB fusion with best performance in the competitive luciferase reporter assay: the direct fusion ZF B/SB100x and N57/ZF B.

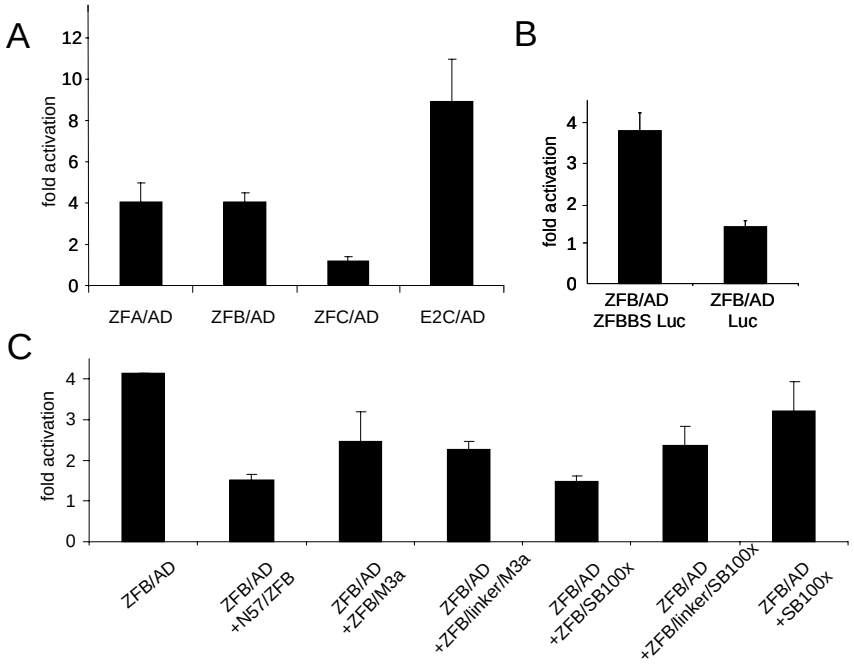


Figure 14. DNA-binding of designed ZF proteins and ZF/SB transposase fusion proteins. All figures show the results of luciferase reporter assays done in HeLa cells. Bars represent normalized relative luminescence units (RLUs) for activator proteins (ZF/AD) divided by RLUs for a negative control (AD). Error bars represent the standard error of the mean. Luciferase reporter plasmids carry ZF recognition sites corresponding to the ZF fused to the AD upstream of the minimal promoter. **(A)** Binding of different ZF/AD and the E2C/AD fusion proteins to their predicted ZF recognition sites. **(B)** Binding of ZF B to its recognition site and unspecific DNA-binding of ZF B is examined using luciferase reporter plasmids, one containing the ZF B recognition site (ZFBBS Luc) and one lacking the ZF B recognition site (Luc). **(C)** Binding of different ZF B/SB fusion proteins to the ZF B binding site was examined by competitive luciferase assays (see Fig. 9A) transfecting activator protein (ZFB/AD) with and without addition of fifty fold excess of competitor (different ZF B/SB fusion proteins).

3.8. ZF B/SB fusion proteins exhibit transposition activity

Fusion proteins between ZF B and the hyperactive SB transposase variants SB100x and M3a connected with and without a linker were created and examined for their transposition activity. The linker sequence introduced between both moieties consisted of amino acids KLGGGAPAVGGGPK. This linker was already successfully applied in transposase fusions used for targeting (Maragathavally et al. 2006).

All fusion constructs showed diminished performance in transposition compared to unfused transposase. The direct fusion of ZF B to M3a transposase was about twofold as efficient in transposition compared to background integration. Inclusion of the linker enhanced transposition activity to sevenfold compared to background integration. The direct fusion of ZF B to SB100x transposase, which was used in further targeting experiments, was about fourfold as efficient in transposition as background integration. Inclusion of the linker into ZF B/SB100x fusion proteins enhanced transposition activity to 13-fold compared to background integration (Fig. 15A). Results of the excision PCR and stained petri dishes are depicted in Fig. 15 C and B respectively. Further experiments were conducted with the ZF B/SB100x fusion protein for simplicity from the next chapter on termend ZF B/SB.

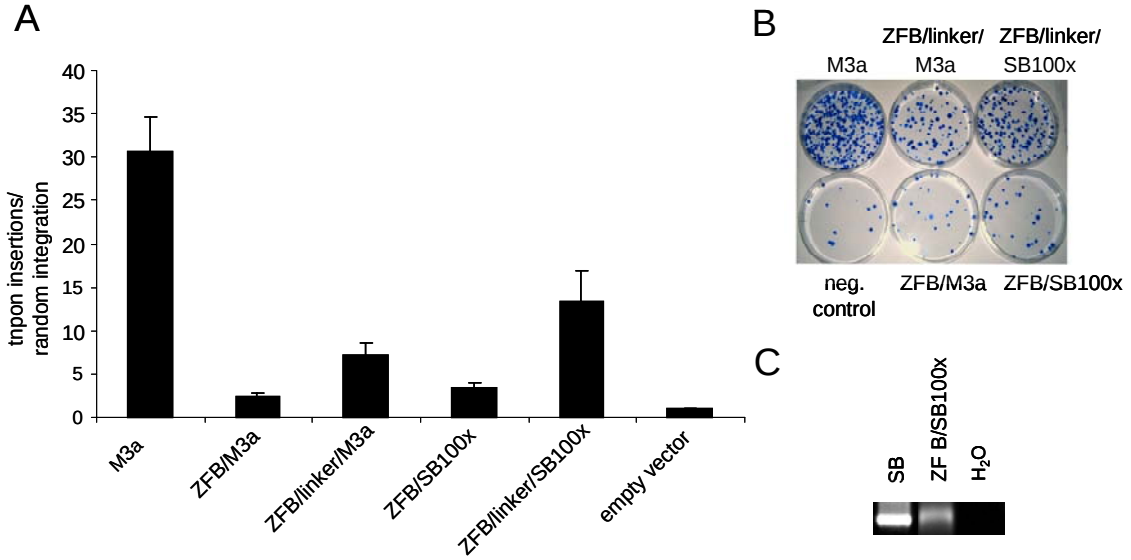


Figure 15. Transposition activity of different ZF B/SB fusion constructs. (A) A cell culture transposition assay was done in HeLa cells (Fig. 8A) to examine transposition activity of different ZF B/SB fusion proteins. Cells transfected with transposase helper plasmid and a transposon donor plasmid containing a neomycin resistance gene were selected for transposon insertions, stained and counted. Cell colony number produced by the hyperactive SB transposase mutant M3a and different ZF B/SB transposase constructs were divided by cell colony number of random integration. ZF B was connected with or without an intervening linker to either M3a or SB100x two hyperactive variants of the SB transposase. Error bars represent the standard error of the mean. **(B)** Petridishes of a characteristic cell culture transposition assay for different ZF B/SB constructs. **(C)** PCR-based transposon excision assay (Fig. 8D). Excision of the transposon from the donor plasmid pTneo results in PCR-fragment of characteristic size.

3.9. Southern blot of LAM-PCR-amplified *SB* transposon insertions

LAM-PCR offered a more unbiased method to detect transposon insertions than locus-specific PCR. Cells transfected with different hyperactive transposase mutants (M3a and SB100x) or different ZF/SB transposase helper plasmids together with transposon donor plasmids were selected for transposon insertions. Equal amount of genomic DNA isolated from selected cells was subjected to LAM-PCR procedure (Fig. 7B). In a first linear PCR step

biotinylated primer annealed at the end of the left IR of the transposon inserted into the genome, followed by elongation of the PCR product across the left transposon end into adjacent genomic DNA. This first linear PCR did not amplify DNA but rather made biotinylated single stranded transcripts of transposon integration sites into the genome. These transcripts were then captured by streptavidin coated magnetic beads (Dynabeads) followed by second strand synthesis using the Klenow fragment. Double stranded DNA was digested with the restriction enzyme FspBI, a four base cutter that is statistically expected to cut human DNA about every 256 bp. Linker with ends compatible to the FspBI-created overhangs were ligated to FspBI-digested sample DNA. Nested PCR was done with primers that annealed in the linker and the left transposon end to amplify transposon insertions in the genome. PCR products were separated on an agarose gel, blotted onto a nylon membrane and hybridized with either E2C locus specific probes or probes specific to the *LINE1.3* 3´-end. All lanes of the blot hybridized with probes specific to the E2C locus showed a faint smear with similar intensity indicating no targeting of the E2C containing transposases. Lanes hybridized with *LINE1.3*-specific probes showed a stronger signal for ZF B/SB and N57/ZFB+SB100x than for SB100x indicating an enrichment of transposon insertions near the ZF B recognition site (Fig. 16).

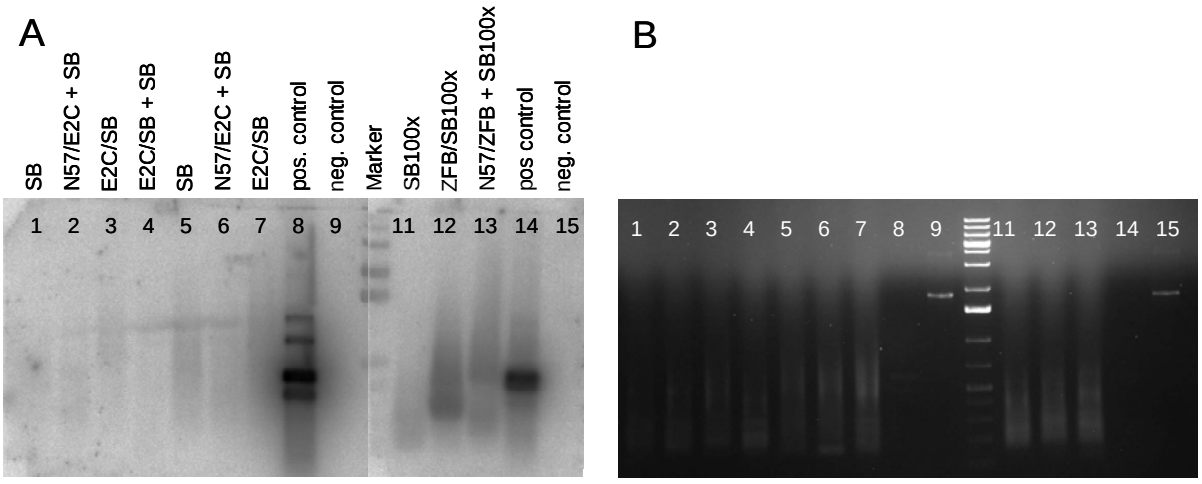


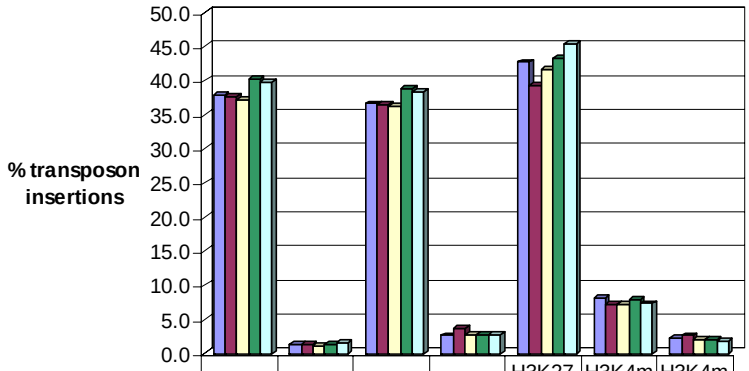
Figure 16. Southern blot of LAM-PCR samples. (A) Southern blot. HeLa cells were transfected with different SB transposase helper plasmids and a neomycin resistance gene carrying transposon donor plasmid. Cells were selected for transposon insertions. Genomic DNA was extracted from cells and subjected to LAM-PCR procedure. Equal amounts of LAM-PCR product was separated on an agarose gel. Lanes 1-9 were hybridized with probes annealing at the 3´-end of LINE1 elements. Lanes 11-15 were hybridized with four probes annealing at the *erbB2* region. Negative control was unrelated PCR product. Lanes 1-4 belong to one independent experiment, lanes 5-7 belong to another independant experiment. Positive control was DNA corresponding to the E2C or LINE1 probes. **(B)** Agarose gel before Southern blot loaded with equal amount of µl of LAM-PCR samples.

3.10. Illumina sequencing

Some changes were introduced to the LAM-PCR protocol for Illumina sequencing compared to the LAM-PCR protocol used in Southern experiments (section 3.9). Genomic DNA was ultrasonicated before performing linear PCR to avoid the use of restriction enzymes in later steps which might introduce a bias in the detection of transposon insertions. After second strand synthesis and end repair, A-overhangs were added to the DNA which was needed for ligation to a double stranded linker. Transposon insertions into the genome were amplified by three nested rounds of an exponential PCR using primer annealing at linker DNA and primer annealing at the left transposon end (Fig 7A). Primer annealing at linker DNA were barcoded differently for each transposase. The barcode was a four nucleotide long tag encluded in the second round PCR primer. Barcodes were designed in a way that different barcodes differed at least two positions so that a polymerase error or sequencing misread at one nucleotide did not directly result in a different barcode. DNA obtained from exponential PCRs using differently barcoded linker primers could thus be distinguished even if mixed together after the PCR reaction. This was important because samples from different transposases were pooled before submission to the Solexa sequencing reaction.

To examine the potential of E2C ZF fusion proteins to target *SB* transposon integrations E2C/SB transposase (E2C/SB), E2C/SB transposase mixed with unfused SB transposase (E2C/SB+SB) and N57/E2C mixed with unfused SB transposase (N57/E2C+SB) were tested. Between 740 transposon insertions (E2C/SB) and 8,314 transposon insertions (SB) were analysed, on average about 4,900 insertions per transposase. As controls, transposon integrations of unfused SB (SB) were analysed and a dataset was calculated for random integration at any TA dinucleotide in the human genome (randomTA). For all transposases tested no profound change in integration pattern was observed. For E2C/SB+SB a slight increase in transposon integrations into genes (40.2 % versus 37.6 % for other transposases) and here within introns (38.9 % versus 36.4 % for other transposases) was found. About 7 to 8 % of transposon insertions were found in active genes characterized by H3K4me1 for all transposases as well as for random calculations and about 2 % of transposon insertions occurred within transcription start sites characterized by H3K4me3. Most statistically relavant differences (p-value < 0.01) occurred between the randomTA dataset and SB mediated transposition. Slightly fewer insertions into genes (37.6 % (without E2C/SB+SB) versus 39.7 % for randomTA) including introns (36.4 % (without E2C/SB+SB) versus 38.3 % for randomTA) were found as well as fewer integrations into silent intergenic regions (41.7 %

versus 45.4 % for randomTA) characterised by H3K27me3 (Fig. 17). Surprisingly, not a single insertion into a 10 kb window around the E2C recognition site was detected for any of the transposases. Chances for a transposon integration to insert within this window by chance (for the randomTA data set) were calculated to be 0.816 x 10^{-3} or, in other words, of about 125,000 transposon integrations at any TA in the human genome one insertion could be expected to occur within a 5 kb window around the E2C binding site by chance. Allowing for 2 mismatches within the E2C recognition site three insertions (0.0495 %) are found for N57/E2C+SB with randomTA expectations at 0.0122 % and one insertion (0.012 %) for SB. For E2C/SB and E2C/SB+SB no insertions were found within a 5 kb window around the E2C recognition site allowing up to two mismatches. Improved targeting was observed with a partially mutated E2C recognition site which only allowed for binding of E2C ZF 1-3 by Yant et al. (2007). Therefore, presence of E2C half sites within a 5 kb window around the E2C binding site was also checked. No significant enrichment in transposon insertions were found for the 3´-half site (5´-GCC GCA GTG-3´). Only a slight enrichment of transposon insertions for E2C/SB compared to SB (4.86 % versus 3.61 %) for the 5´-half site (5´-GGG GCC GGA-3´) was seen (supplementary Fig. 2 and supplementary Table 1A).

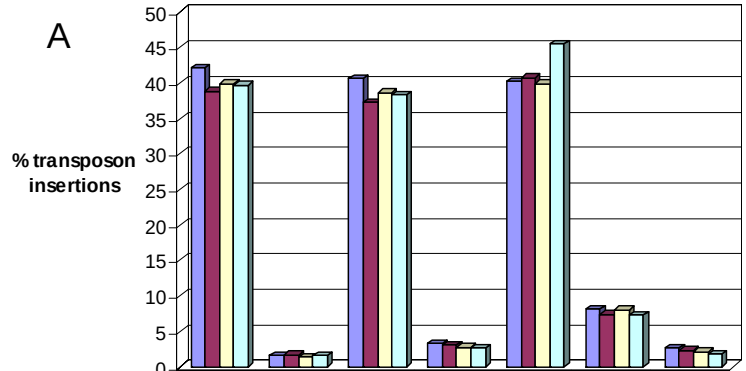


	gene	exon	intron	tss_5k	H3K27me3	H3K4me1	H3K4me3
SB	37.9	1.4	36.6	2.6	42.7	8.1	2.2
E2C/SB	37.7	1.4	36.5	3.6	39.2	7.2	2.6
N57/E2C+SB	37.2	1.0	36.2	2.7	41.7	7.1	2.1
E2C/SB+SB	40.2	1.4	38.9	2.8	43.2	7.8	2.0
randomTA	39.7	1.5	38.3	2.8	45.4	7.3	1.8

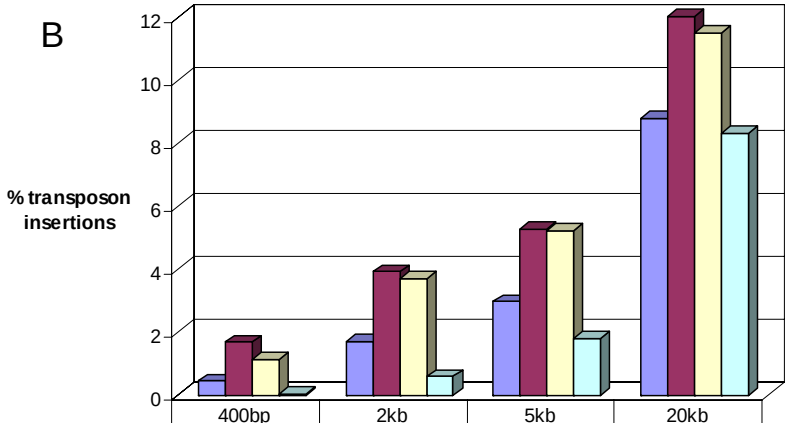
Figure 17. Distribution of *SB* transposon insertions in the human genome catalyzed by E2C/SB constructs. **(A)** Distribution of transposon insertions with respect to genes, introns, exons, transcription start sites (tss) and different histone methylations indicating intergenic and silent coding regions (H3K27me3), active genes (H3K4me1) and transcriptional start sites of expressed genes (H3K4me3). Color-coded bars represent the percentage of transposon insertions within a genomic region catalyzed by the indicated transposase. RandomTA represents bioinformatically calculated percentage of random insertion into any TA dinucleotide in the respective genomic region. Using E2C/SB+SB as transposase source a slight enrichment of transposon insertions into genes and here into introns can be observed.

All three classes of LINE elements make up about 21 % of the human genome. LINE1 elements contribute to about 17 % to the human genome. LINE1 elements are distributed stochastically over human chromosomes with the exception of the X-chromosome which contains a high load of LINE elements. However, they are found at much higher density (about fourfold enriched) in AT-rich regions (Lander et al. 2001). BLAST analysis for short sequences discovered 12,899 ZF B binding sites in the human genome. Since the human genome consists of about 3 billion bp, on average there should be one ZF B binding site approximately every 235,000 bp. However, some bias for ZF B binding site distribution is observed among different chromosomes. Chromosomes 4, 5 and especially the X chromosome contain more ZF B binding sites (< every 200,000 bp) whereas chromosomes 18, 19 and especially chromosome 22 contain fewer ZF B binding sites (> every 400,000 bp). Using a gene delivery vector with a random integration profile an insertion within 1 kb of a ZF B binding site is expected to occur by chance about once every 118 insertions. To examine the potential of ZF B to target *SB* transposon integrations to LINE1 sequences ZF B/SB transposase (ZFB/SB), and N57/ZF B mixed with unfused SB transposase (N57/ZFB+SB) were tested. Between 5,991 transposon insertions (ZFB/SB) and 14,178 transposon insertions (SB) were analysed by LAM-PCR followed by Illumina sequencing, on an average about 10,000 insertions per transposase. As controls transposon insertions of unfused SB transposase (SB) into the human genome were analysed and a dataset was calculated for random integration at any TA dinucleotide in the human genome (randomTA). Statistical relevant differences (p-value < 0.001) in genomic transposon distribution was found for insertions into genes: 39.4 % for ZF B containing transposase versus 42.1 % for unfused SB transposase and here within introns: 38 % versus 40.6 % respectively. Differences for transposon insertions into transcription start sites, silent or intergenic regions or active genes were not observed between different transposases. For ZFB/SB constructs an about fourfold enrichment of transposon insertions was found in a 400 bp window around ZF B binding sites. The wider the window around the ZF B binding site the less pronounced the enrichment of transposon insertions around the ZF B binding site (Fig 18B). Allowing one mismatch within the ZF B binding site 6.1 % (SB) versus 10.1 % (ZF B containing transposases) and allowing 2 mismatches 8.3 % (SB) versus 13 % (ZF B containing transposases) of transposon insertions occured within a 5 kb window around the ZF B recognition site. Since individual ZFs binding 5´-ANN-3´ triplets often do not make contact to the 5´-A, in an additional search all 5´-As of individual ZF recognition sites for ZF B were changed to any nucleotide (5´-N).

Targeting frequency was similar to that of the full length ZF B recognition site. The same was found for ZF B half sites (Fig. 18, supplementary Fig. 1 and supplementary Table 1B).
As described in the method section (2.2.10 LAM-PCR for Illumina sequencing; Bioinformatic analysis) insertions into DNA that map to multiple sites in the human genome were not considered in the data set shown in Fig. 18A and 18B. This could apply to insertions upstream of the ZF B binding site into the LINE1 element where DNA sequence can not be discriminated against other conserved LINE1 elements in the human genome. The bias towards insertion near ZF B binding sites seen for ZF B containing transposases could thus be even more pronounced when those insertions would be included into the analysis. Therefore transposon insertions into repetitive regions in the human genome that could not be mapped to a unique site in the human genome were assigned to the respective repetitive element they integrated into (a selection of analyzed repetitive elements can be seen in supplementary Table 1C). The SB transposon system seems to have a general tendency to integrate into LINE elements with > 28 % of transposon insertions occurring in LINE elements compared to 24.9 % calculated for the randomTA data set. This could be explained by the high TA-content of members of the LINE family. The tendency to insert into LINE elements was seen for LINE1 as well as LINE2 elements for unfused SB transposase compared to the randomTA data set. An even stronger bias towards insertion into LINE1 elements was seen for ZF B/transposase fusion proteins with 26.1 % (ZFB/SB) and 27 % (N57/ZFB+SB) compared to 23.6 % (SB) and 21 % (randomTA). Such a conform difference between ZF B fusion proteins and controls was only seen for LINE1 elements. For the majority of repetitive elements no or only small differences between the different data sets were observed.



	gene	exon	intron	tss_5k	H3K27 me3	H3K4m e1	H3K4m e3
SB	42.1	1.59	40.6	3.23	40.2	8.13	2.67
ZFB/SB	38.8	1.67	37.3	3.05	40.7	7.41	2.3
N57/ZFB+SB	39.9	1.36	38.6	2.69	39.9	8.01	2.08
randomTA	39.7	1.53	38.3	2.6	45.5	7.31	1.78



	400bp	2kb	5kb	20kb
SB	0.43	1.7	2.96	8.78
ZFB/SB	1.67	3.91	5.26	12
N57/ZFB+SB	1.13	3.68	5.23	11.5
randomTA	0.0375	0.585	1.8	8.3

Figure 18. Distribution of transposon insertions into the human genome catalyzed by ZF B/SB constructs. **(A)** Distribution of transposon insertions with respect to genes, introns. exons, transcription start sites (tss) and different histone methylations indicating intergenic and silent coding regions (H3K27me3), active genes (H3K4me1) and transcriptional start sites of expressed genes (H3K4me3). Colorcoded bars represent the percentage of transposon insertions within a genomic region catalyzed by the indicated transposase. RandomTA represents bioinformatically calculated random distribution of insertions into any TA dinucleotide in the respective genomic region. **(B)** Enrichment of *SB* transposon insertions near ZF B binding sites in the human genome using ZF B/SB fusion constructs. Depicted are percentages of transposon insertions within a defined window around a ZF B binding site.

3.11. Proteolytic cleavage products of SB transposase and ZF/SB transposase fusion proteins

Integrity and correct expression of the ZF B/SB fusion protein is a prerequisite for successful transposon targeting. Proteolytic cleavage products of the SB transposase in prokaryotic as well as eukaryotic cells have been observed earlier by members of our laboratory. Unfused SB transposase M3a, ZF B/SB and E2C/SB were transfected into HeLa cells and Western blots were performed to examine integrity of proteins. Apart from bands of expected size corresponding to SB transposase (37 kDa) and ZF/SB transposase fusion proteins (55 kDa) additional bands of smaller size were detected (about 28 kDa for SB transposase and about 37 kDa for ZF/SB fusion proteins) with a polyclonal *Anti-Sleeping Beauty Transposase Antibody* (R & D Systems) raised against full length SB transposase protein (Fig. 19).

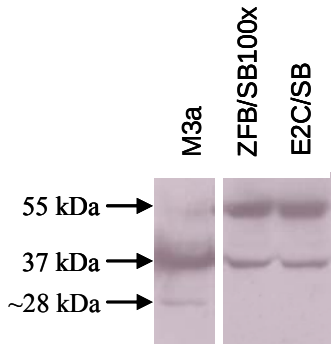


Figure 19. Expression of ZF B/SB100x and E2C/SB proteins in HeLa cells. Western Blot showing expression of the unfused SB transposase (37 kDa) and ZF B/SB100x and E2C/SB fusion proteins (55 kDa). For the ZF fusion proteins and the full length protein additional smaller protein fragments were detected.

4. Discussion

4.1. Targeting on genomic level

All approaches to target DNA integration to specific loci in the human genome pursued to date have their strengths and have their shortcomings. Possible explanations for the failure of ZF/SB fusion proteins to target transposon insertions to their respective ZF binding sites can reside within the general concept of targeting transposon insertions with a full length SB transposase fused to a ZF protein or within the particular fusion proteins used in this work themselves.

4.1.1. Targeting strategy

The E2C ZF fused to regulatory domains was shown to effect expression of the *erbB-2* gene which lies downstream of the E2C binding site in HeLa cells (Beerli et al. 1998). Targeting retroviral integrations to the *erbB-2* locus in the human genome has been shown for an E2C/HIV IN fusion protein (Tan et al. 2006). Fusion of E2C to the HIV IN resulted in an up to 10-fold higher preference to insert into the *erbB-2* locus than found for HIV IN alone. Direct fusions of transposases to DBDs have also been described in (Szabo et al. 2003; Wilson et al. 2005). Thus, targeting *SB* transposon insertions to the *erbB-2* locus in the human genome using direct fusions between the transposase and the E2C protein appeared to be a promising approach. The *erbB-2* gene encodes an epidermal growth factor receptor which stimulated cell proliferation (Lane et al. 2000) and inhibits apoptosis (Zhou et al. 2000) and has been shown to be overexpressed in breast cancer. So clearly insertion of a transgene into this region with the possible consequence of altered expression of the *erbB-2* gene is not of gene therapeutic interest. The aim of this project was to show the feasibility of targeting the SB transposase to a unique site in the human genome with a ZF/SB fusion protein in a proof-of-principle experiment. Even though site-specific PCRs could detect some *SB* transposon integrations near the *erbB-2* locus mediated by E2C containing SB transposases, analysis of hundreds and thousands of transposon insertions by LAM-PCR and Illumina sequencing did not confirm this finding. One insertion into the *erbB-2* locus was found using site-specific PCR on fifty individual cell clones using the E2C/SB transposase fusion protein. Number of transposon insertions for these cell clones was not determined. However estimating two transposon insertions per cell clone results in a targeting frequency of 1/100. Thus in a total of 740 transposon insertions (as analysed by Illumina sequencing for E2C/SB transposase) 7.4

insertions would be expected to occur within the *erbB-2* locus to confirm the results of the locus-specific PCR, but none was actually found. However with site-specifc PCR only those insertions within a defined window between the transposon insertion and the locus-specific primer can be detected and ideally no transposon insertions at other loci in the human genome will be amplified. During LAM-PCR hundreds or thousands of transposon insertion in the human genome are amplified maybe masking insertions into the *erbB-2* locus that could be amplified by site-specific PCR. This could be supported by the fact that the *erbB-2* locus is rather GC-rich and might amplify less efficient than TA-rich targets in the human genome. However if insertions into the *erbB-2* locus were frequent and made up a significant proportion of *SB* transposon insertions this would be expected to be evident in the data. Lack of genomic insertions near the E2C binding site is consistent with work from Yant et al. (2007) where no targeting of the genomic E2C recognition site by an E2C/SB fusion protein was observed.

Unfortunately both the catalytic and DNA-binding domains of the transposase are present in the ZF/SB fusion protein, in which the SB transposase retains its intrinsic ability to recognize any TA dinucleotide in the genome. Binding of SB transposase to DNA independent of the ZF domain results in off-target transposon integrations. In the SB transposase the catalytic domain and the domain responsible for target DNA binding and protein-protein interaction overlap, making it difficult to separate the target DNA binding domain from the rest of the transposase. Using a SB transposase mutant that retains the ability to mediate catalytic steps of transposition and lacks the ability to select an insertion site independent of an additional DBD would provide a solution to this problem. However no such SB mutant is published to date. Serine recombinases have spatially separated catalytic and DNA-binding domains. Fusion constructs between a foreign DNA-binding domain and solely the catalytic part of the recombinase would thus be expected to completely alter insertion pattern. This approach was successfully taken by Gordley et al. (2009) who fused the catalytic domain of a hyperactive version of the Gin invertase from bacteriophage Mu to a five-finger ZF domain. With this fusion protein targeted integration of a transgene in the human genome with > 98 % accuracy was achieved. For efficient integration the Gin invertase hyperactive mutant requires a 20 bp core sequence which is flanked by the two ZF binding sites. ZF proteins can theoretically be designed to bind any sequence in the human genome, in contrast the 20 bp internal sequence required by the recombinase for catalysis had to be inserted into the genome prior targeting experiments by (Gordley et al. 2009). Whether Gin mutants that insert at any given sequence in the human genome can be created remains to be further explored. Invertases with altered

specificities have already been created by directed evolution (Gordley et al. 2007) (Gersbach et al. 2010). Retroviral integrases also have spatially separated DBDs and catalytic domains reviewed in (Jaskolski et al. 2009). The catalytic domain is evolutionary conserved for INs of retroviruses (Khan et al. 1991) and shows limited activity even in the absence of the other domains found in full length IN for Rous sarcoma virus and HIV (Kulkosky et al. 1995). Fusing a heterologous DBD to the catalytic IN domain could therefore eliminate intrinsic target DNA binding of the integrase. Target site selection would then fully depend on the heterologous DBD. However in (Katz et al. 1996) full length IN fusions to LexA were most successful in targeting DNA integration whereas constructs with the LexA protein fused to the catalytic domain only showed reduced activity and targeting ability. Similarly Goulaouic and Chow (1996) and Tan et al. (2004) observed reduced integration performance for fusions with truncated IN. Their targeting experiments also focused on full length IN fusions to a heterologous DBD. The heterologous DBD thus seems to act dominant over the intrinsic DNA binding activity of the IN. This finding supports the hypothesis that a heterologous DBD could act dominantly over the intrinsic DNA-binding ability of SB transposase as well. Using N57/ZF B and ZF B fusions of full length SB transposase (ZF B/SB) to target LINE1 elements, no pronounced difference in targeting ability was found. Both approaches seemed to work equally well with slightly better targeting values for ZF B/SB. Different approaches for targeting transposon insertions in the human genome were taken in Ivics et al. (2007). A shift of *SB* transposon insertion pattern was achieved by tethering the transposon to S/MARs on chromosomes in human cell culture. S/MARs mediate structural organization of chromatin in the nucleus by providing anchor points of DNA for the chromatin scaffold. A *LexA* operator site was included in the *SB* transposon. The *E.coli* LexA protein was fused to a SAF-box domain which is known to interact with S/MARs (Kipp et al. 2000). Using this method the DNA component of the *SB* transposon system the transposon is tethered to a target site in the genome in contrast to approaches where the enzymatic component the transposase is tethered to the target site. One advantage of tethering the transposon to DNA lies within the fact that unfused SB transposase protein carries out transposition bypassing the problem associated with the use of a fusion protein like diminished transposition activity. In a similar approach a LexA/tetracyclin repressor (TetR) fusion protein was used to tether a *LexA* containing *SB* transposon to an artificially introduced TRE^2 site in the human genome. Two out of 400 transposon insertions mediated with this set up occurred downstream of the TRE^2. In a third approach N57 was fused to the TetR. Upon tethering of the SB transposase to an artificially introduced TRE^2 site in the human genome using N57/TetR > 10 % of transposon

insertions occurred in neighboring DNA sequences (Ivics et al. 2007). However in contrast to viral vector systems (Tan et al. 2006) until today no genomic targeting of a transposon system into a naturally occurring site in the human genome has been published to my knowledge. As mentioned earlier a fusion protein of the E2C ZF and a hyperactive SB transposase has been tested for targeting transposon insertions by Yant et al. (2007). While this publication could show some extent of targeting in a plasmid-based assay, no targeted transposon integration could be found in the context of the human genome. In inter-plasmid targeting experiments transposition from the donor plasmid to the target plasmid occurs in the nucleus. Theoretically, the plasmid target competes with genomic DNA for transposition like in genomic targeting experiments. However for inter-plasmid targeting experiments only those events occurring into the target plasmid are considered. Shifts in transposon integration pattern are easier to observe on plasmid of for example 5 kb than in the 3 billion bp human genome. In contrast to the transfected target plasmid the endogenous target locus might be CpG methylated. The E2C binding site contains two CpGs which could potentially be methylated. Methylation of DNA can affect protein-DNA interaction as seen for some restriction enzymes or transcription factors (Jones et al. 1992) (Tate and Bird 1993). Methylation sensitive binding for ZF proteins has been described (Hark et al. 2000) (Daniel et al. 2002). ZF proteins binding methylated cytosine or unmethylated cytosine exclusively as well as ZF proteins that do not discriminate between methylated and unmethylated cytosine have been created by (Choo 1998). A phage display library of the Zif268 three-finger ZF protein was screened with altered versions of the Zif268 binding sites. The recognition site of the second finger, originally 5′-TGG-3′, was replaced by 5′-GNG-3′ where N was either thymine, cytosine or 5-meC. Thymine is a pyrimidine base and carries a methyl group at its fifth carbon equal to 5-meC. ZFs with an alanine at position 3 of the α-helix of the second finger bound triplets containing thymine or 5-meC by interaction of the amino acid side chain with the 5-methyl side chain. ZFs with a valine at the same position bound all three triplets and ZFs with aspartate at the same position bound cytosine exclusively. The E2C ZF was already successfully applied in fusions to regulatory domains to regulate gene expression and in fusions to the HIV IN for targeting experiments. Regulatory studies were conducted on plasmids containing *erbB-2* promotor sequence transfected into HeLa cells. Thus these plasmids were not expected to contain CpG methylations. However targeting studies of the E2C/HIV IN were done in HeLa cells where the E2C binding site resides in its natural methylation state. Since targeting the genomic *erbB-2* locus with an E2C/HIV IN construct

had been successful (Tan et al. 2006) the E2C binding site either had to be accessible for E2C ZF binding or the E2C ZF is not sensitive to altered methylation status of its binding site.

Since targeting the *SB* transposon to a defined site in the human genome already worked using a fusion protein of two DBDs which tethered a transposon containing one DBD binding site to the second DBD binding site present in the genome, for further targeting experiments this method could also be exploited further. In this approach the transposon would be integrated by unmodified transposase. This strategy would circumvent the problems associated with creating a functional transposase fusion protein, but would still harbour the problem of potential off-target insertions. However this approach involves the use of three components rather than two. For gene therapy applications a simpler two component system might be more favourable. The same problem arises for a second alternative approach that is fusing a protein known to interact with the transposase to a DBD. This approach was applied in my work by including the N57/ZF fusion protein into experiments. As pointed out before one major problem of targeting transposon insertions with a ZF/SB fusion protein are off-target transposon insertions. Increasing the number of binding sites for the fusion partner in the human genome increases chances for the fusion protein to bind its target and thus increases chances for targeted transposon insertions. Apart from Ivics et al. (2007) who targeted S/MARs with a tethering protein, Gijsbers et al. (2010) targeted the HIV IN to pericentric heterochromatin and intergenic regions by replacing the chromatin interaction-domain of LEDGF/p75 a cellular factor that normally tethers HIV IN to transcription start sites with CBX1. CBX1 binds H3K9m2 or H3K9m3 which is associated with pericentric heterochromatin and intergenic regions. This strategy to target HIV IN has also been taken by Ferris et al. (2010). HIV DNA integration was targeted to the 5′-region of active genes by fusion of the homeodomain finger from inhibitor of growth protein 2 to LEGDF lacking its chromatin binding domain. HIV DNA integrations were also targeted to chromatin by fusion of the chromodomain of heterochromatin protein 1-α to LEDGF lacking its chromatin binding domain. All of these approaches use proteins known to interact with certain DNA regions or structures in contrast to a unique DNA binding site. Being able to target a defined DNA sequence might specify the insertion site to a more defined region. However alternatively to using a ZF (E2C) which recognized a unique site in the human genome for targeting, I fused a ZF (ZF B) that recognized a site that existed in multiple copies in the human genome to the SB transposase. An enrichment of transposon insertions near ZF B binding sites indicates the general feasibility of this approach. However, for gene therapeutic applications a much higher

rate of targeting is requiered. Enhancing performance of the ZF or fusion protein might enhance targeting ability.

4.1.2. Targeting protein

Apart from the general concept of targeting transposon insertions to specific sites in the human genome with ZF/SB transposase fusion proteins success or failure of this project greatly depends on the composition and design of the actual ZF/SB transposase fusion protein. The ZF/SB transposase protein used in targeting experiments is expected to bind its ZF recognition site with good affinity and specificity and catalyze efficient transposon integration. The E2C/SB fusion protein created in this work showed reduced transposition activity compared to unfused transposase. Whether this was advantageous or adverse for targeting transposition resides in the reason for the diminished transpositional performance. In this work the ZF domain is fused to the N-terminal part of the SB transposase. The N-terminal part carries a paired-like domain which among other things may also be responsible for target DNA binding. Since fusion of tags to the catalytical C-terminus of the transposase often completely abolish transposition capacity, fusion of tags to the N-terminus could thus result in restricted DNA binding. About 100,000,000 TA dinucleotides, the strict target site requirement for *SB* transposition, exist in the human genome compared to one single E2C binding site. Chances for SB transposase to encounter an adequate insertion site are thus much higher than for the E2C ZF to encounter its binding site. If the transposase moiety in a fusion protein, due to sterical hindrance, has only limited access to target DNA and binding of target DNA is diminished this could actually be beneficial for targeting transposition to a specific site in the human genome because less off-target insertions can be expected. However, for gene therapy applications efficient gene transfer is desired, and thus a transposase with high transposition efficiency would be desirable.

The E2C ZF is considered the gold standard of six-finger ZF proteins to date in terms of binding affinity and specificity. However, stable docking of the transposase may interfere with individual steps during transposition. Transposase proteins are believed to undergo several coordinated conformational changes during the transposition process and strict tethering may be debilitating (Yant et al. 2007). Yant et al. (2007) tested targeting of an E2C/SB fusion protein on a partially mutated E2C-binding site in an inter-plasmid transposition assay. Binding sites for finger 4-6 of the E2C ZF protein were mutated so that only fingers 1-3 of the E2C ZF protein were able to bind. The E2C/SB fusion protein showed better targeting on plasmids containing the partially mutated binding site than on plasmids

with the full length E2C binding site. Interaction of the E2C ZF protein with the partially mutated E2C binding site might be more transient and allow for more flexibility for the transposase. Yant et al. (2007) concluded that flexible transposase tethers may further improve performance of site-directed transposition. However three-finger ZF proteins bind 9 bp target sites which should occur by chance every ~11,500 bp in the human genome. Thus with a three-finger ZF protein it seems impossible to insert a transgene at a specific site or region in the human genome (Wilson et al. 2005). Naturally occurring C_2H_2 type ZFs can consist of one to as many as 29 or even more individual fingers (Dhanasekaran et al. 2007) (Tsai and Reed 1998). In ZF proteins containing multiple fingers not all have to contribute to DNA binding some might have different substrates like other proteins or RNA or provide other functions. A lot of artificial ZF proteins are three-finger ZF proteins used as DNA-binding fusion partner in ZFNs. Six-finger ZF proteins have also been created (Dreier et al. 2001) (Segal et al. 2003b). If and to what extent designed six finger ZFs bind partial target sites remains to be investigated. For ZF B no enhancement of targeting was found when screening for ZF B binding half sites in Illumina sequencing data of LAM PCR samples.

The ZF B protein showed reduced induction of reporter gene expression compared to the E2C ZF protein. Affinity of the ZF B protein for its target site was not determined. However reduced affinity of ZF B for its binding site compared to the E2C ZF might lead to more transient binding which will lead to more transient tethering of the SB transposase. Furthermore there are approximately 13,000 copies of the ZF B binding site in the human genome. Chances of ZF B to encounter its recognition site are thus much higher than for the E2C ZF.

Fusing two functionally distinct individual domains together may lead to compromised performance of each moiety due to sterical hindrance. Individual C_2H_2 type ZF domains are simply stichted together in multi-finger proteins and multi-finger ZF proteins are often found in fusion to regulatory domains in the human genome. Therefore ZF domains do not seem to be extremely sensitive to fusions to other protein domains. However the SB transposase is very sensitive to fusions to tags or other protein domains. Including a linker would allow for more flexibility for both fusion partners maybe resulting in enhanced freedom for the individual moeity. Linkers included in E2C/SB fusion proteins showed no improvement in transposition activity whereas the linker included in ZF B/M3a as well as ZF B/SB100x fusion proteins enhanced transposition activity of the fusion protein. Intervening linkers have been applied in numerous fusion proteins to provide sufficient space and flexibility for both proteins to function. An E2C/SB fusion protein containing a 10x glycine linker created by

Ivics et al. (2007) showed no transposition activity. For construction of functional fusion proteins between the E2C ZF and SB transposase Yant et al. (2007) included intervening flexible linkers consisting of variable numbers of repeating glycine-glycine-serine $(GGS)_n$ linkers. Fusions of the E2C ZF to the C-terminus of the SB transposase showed drastically reduced transposition activity ranging from 2-3 % compared to unfused transposase. Transposition activity of SB transposase with N-terminally attached E2C ZF ranged from 4-10 % compared to unfused transposase. The E2C/SB fusion protein without a linker showed 6 %, whereas the $(GGS)_4$ linker, which is partially identical to linker 1A used in my work, showed 8 % of transposition activity compared to unfused transposase. Yant et al. (2007) used an E2C/SB fusion protein with the intervening $(GGS)_5$ linker for their targeting experiments which had 10 % of activity of unfused SB transposase. Increasing length of the linker did not result in a general increase in transposition activity. Including a 15 aa linker showed highest transposition activity, including longer (18 aa and 21 aa) as well as shorter (no linker, 3 aa, 6 aa and 9 aa) linkers yielded fusion proteins less active in transposition (Yant et al. 2007). The linker used in ZF B/SB fusion proteins was published in (Szuts and Bienz 2000) and already included in Gal4/PB and Gal4/Mos1 fusion transposases which showed successful targeting on plasmid level (Maragathavally et al. 2006). Transposition activity of Gal4/PB and Gal4/Mos1 fusion proteins compared to their respective transposases were not directly evaluated but rather transposition frequency into a cotransfected target plasmid was compared for these proteins. The KLGGGAPAVGGGPKAADK linker was also tested for construction of a PB transposase fusion protein (Cadinanos and Bradley 2007). Among constructs with three different intervening linkers the construct containing the KLGGGAPAVGGGPKAADK linker showed highest activity in transposition.

In general linkers are designed to be flexible and hydrophilic not to disturb the function of two protein domains fused together. In Arai et al. (2001) linkers expected to form a rigid helical structure due to interaction of lysine and glutamic acid residues were tested for their ability to spatially separate two protein moieties. Linkers of different length of the amino acid sequence $A(EAAAK)_nA$ with n = 2-5 were introduced between two fluorescent proteins,enhanced green fluorescent protein (EGFP) and enhanced blue fluorescent protein (EBFP). EGFP and EBFP can be applied in fluorescence resonance transfer (FRET) studies. EGFP and EBFP fused together with an intervening helical linker showed reduced FRET compared to fusions with flexible intervening linkers. FRET also decreased with increasing length of the helical linker but not with increasing length of flexible linkers. Thus helical linkers seem to increase distance between both proteins keeping both domains spatially

separated. Arai et al. (2001) advice helical linkers for construction of bifunctional fusion proteins because they can efficiently separate functional domains and avoid interaction between the two units. However for the E2C/SB construct created in this work no increase in transposition activity was observed by including intervening helical linkers. Maybe the rigid structure of the fusion protein interferes with some of the conformational changes of the SB transposase during different steps in transposition. Design of linkers 2A and 2B was guided by experiments from Robinson and Sauer (1998). In this publication two Arc repressor proteins were fused by linkers varying in length and composition. Linkers containing 13 or more amino acid residues show wt activity of the Arc repressor proteins. Modeling studies suggest that linkers of shorter length lead to distortion of the proteins or that the linker crosses the DNA-binding domain of the proteins. Thermodynamic stability was claimed to be greatest for proteins containing a 19 aa linker and stability decreased with decreasing as well as increasing linker length. Folding and unfolding rates decreased with increasing linker length. After linker length the composition of the linker was examined, which consisted of glycine and non-glycine residues. Non-glycine residues were either serine or alanine residues. Composition rather than sequence seemed to be important for stability. Linkers with too many or too less glycine residues were less stable. Furthermore, the number and not the position of the residues in a linker was important for protein stability.

It might be surprising that inclusion of intervening linkers of different length as well as structure into the E2C/SB fusion protein did not have a stronger effect on transposition efficiency. However in Yant et al. (2007) different linker containing E2C/SB constructs have transposition efficiencies ranging from 4-10 % of that of unfused transposase. Five out of seven of the constructs do not show big differences in transposition efficiency (6-8 %). Overall selection of an adequate linker in a fusion protein seems to be rather empirical and findings for proteins other than transposases can not readily be transferred to fusions of SB transposase. However the linker with the amino acid sequence KLGGGAPAVGGGPKAADK seems to be tolerated well by a variety of transposases in fusion proteins. Modeling studies could maybe aid in design of adequate linkers. However, despite extensive effort the structure of SB transposase has yet not been solved.

Apart from exchanging the ZF in the ZF/transposase fusion protein also other transposases could be tested for their ability to target transposition. Fusion proteins of ZF B and E2C to the C-terminus to Tol2 transposase were all inactive in transposition (data not shown). A ZF B fusion to the PB transposase was active in transposition but did not show promising results in binding to the ZF B binding site (data not shown). The PB transposase requires a TTAA

sequence for insertion. The next TTAA sequences within the *erbB-2* promotor region lay approximately 500 bp up- and 1 kb downstream of the E2C binding. Even though targeted integration of the *PB* transposon by a PB transposase Gal4 DNA-binding domain fusion was observed at a TTAA about 1 kb upstream of the *UAS,* the *erbB-2* region may not provide enough *PB* target sites for successful transposition upon binding of the E2C ZF.

4.1.3. Target site selection

For gene therapy purposes selection of an adequate insertion site for the therapeutic transgene is fundamentally important. A suitable target site has to provide sustained transgene expression and a "safe" DNA context where expression of endogenous genes is not altered after insertion of the transgene cassette. In this work I wanted to show proof-of-principle for targeting the *SB* transposon to defined sites in the human genome using ZF/SB transposase fusion proteins. For proof-of-principle targeting experiments the ZF binding site and adjacent genomic sequences had to be accessible for the gene delivery complex, the ZF binding site had to provide a specific target for the ZF protein and adjacent genomic sequences had to provide DNA structure and sequence suitable for SB transposase integration. SB transposase absolutely requires a TA dinucleotide for transposon integration. Preferred *SB* insertion sites are generally found in TA-rich regions but target site preference rather relies on physical DNA structure than sequence. *SB* prefers integration into distorted DNA where the axis around the target site is off-center, rotation of the helix is non-uniform and the distance of the central base pairs are increased (Liu et al. 2005). Preferred *SB* integration sites can be predicted with a bioinformatics tool called ProTIS (Geurts et al. 2006). DNA context at the *erbB-2* locus is rather GC-rich which disagrees with *SB* integration preferences in general. Analysis of the *erbB-2* locus using the ProTIS prediction tool predicts only few TA dinucleotides semi-preferred for *SB* transposon insertions (Fig. 13). After binding of the E2C ZF to its target site transposon integration could be limited by the lack of adequate TA dinucleotides in the surrounding DNA. In contrast the 3´-region of the *LINE1.3* element is rather TA-rich and analysis with the ProTIS program predicts numerous *SB* insertion hot-spots. Successful transposon insertion within this region should not be limited by eligible TA dinucleotides (Fig. 13). However DNA sequence flanking the *LINE1.3* element varies from element to element.

For gene therapy applications long term expression of the transposon-encoded therapeutic transgene gene is needed. Silencing of transposons by chromatin is a phenomenon found in *Arabidopsis thaliana* (Lippman et al. 2004) (Llave et al. 2002), *Saccharomyces pombe*

(Martienssen et al. 2005), and mammals (Martens et al. 2005) (Bourc'his and Bestor 2004). Different ways of silencing expression of a transposon-encoded transgene after insertion into the human genome exist. Expression of the transposon-encoded transgene can be influenced by spreading of heterochromatin from the adjacent genomic regions into the transposon. Thus silencing of transposons inserted into heterochromatic regions would not be surprising. Transposons inserted into a highly expressed region can be silenced by formation of heterochromatin induced by an RNAi-mediated pathway (Garrison et al. 2007). This mechanism of silencing has been suggested to also silence expression of *SB* transposon-encoded transgenes. Garrison et al. (2007) report that the percentage of silenced transgenes raised to 71 % 42 weeks after transfection for transposons expressing the transgene from a viral promoter. Initial transposition efficiency in this publication was surprisingly high at 41-52 % using non-selective fluorescence-activated cell sorter-based method compared to 2-3 % previously reported for *SB*. However in other publications *SB* transposition events were selected for by antibiotic treatment. Low transposition efficiencies observed by the latter method could in part be explained by high rates of silencing of the transposon-encoded antibiotic resistance gene during the selection process. However, other studies report sustained *SB* transposon-encoded transgene expression followed up around 5 month in human cells as well as mice (Mikkelsen et al. 2003) (Yant et al. 2000) (Ohlfest et al. 2005) (Grabundzija et al.).

4.2. Problems of ZF design by modular assembly

Design of novel ZF proteins by modular assembly offers a fast and easy method to create a unique and customized DNA binding protein. Websites like “zinc finger tools” or websites created more recently like the zinc finger database (ZiFDB) (Fu et al. 2009) and zinc finger targeter (ZiFiT) (Sander et al. 2007) enable laboratories lacking expertise in ZF design to design ZFs for their specific needs. ZiFDB offers a database that lists engineered ZFs which derive from Sangamo Bioscience, the laboratory of Carlos Barbas, Toolgen and ZFs deriving from the laboratory of Keith Joung. ZiFiT offers a software for finding adequate ZF binding sites in a specified DNA sequence. While modular assembly is fast and effortless it is not always reliable (Ramirez et al. 2008). Other selection-based methods for ZF design like OPEN (Maeder et al. 2009) are more reliable but also more time consuming. In literature the use of ZF proteins designed by modular assembly is discussed controversially (Kim et al. 2010). A variety of novel multi finger proteins has been established using modular assembly

(Dreier et al. 2001) (Beerli and Barbas 2002) (Bae et al. 2001) (Lui et al. 2002) (Segal et al. 2003a). Early studies claim success rates for modular ZF assembly of 100 % (Segal et al. 2003a) and 60 % (Bae et al. 2001). The main field of application for artificially designed ZFs lies within the use in ZFNs. Since ZFNs are engineered to work as heterodimers, most ZF proteins tested and applied are three or four finger proteins. The success rate for novel ZFNs was claimed to be 24 %. Out of 315 ZFNs that were tested on 33 DNA targets in the human genome, 21 ZFNs successfully modified eight of the targeted DNA sequences (Kim et al. 2009). On the other hand, this implies that 93.3 % of the ZFN pairs tested failed to modify their targets. Other groups report failure rates of 94-76 % for modular assembled ZFN pairs (Ramirez et al. 2008). Whether failure of ZFN pairs was due to ZF design or ZFN-specific remains to be clarified. It is important to note however that three or four finger ZF proteins are more thoroughly studied and may be more specific and easier to engineer than six finger proteins. The use of a three finger protein in fusions with the SB transposase does not provide enough specificity to target a unique site in the human genome. Creating SB transposase fusions with two different three finger ZFs that need to heterodimerize in the process of transposition could take advantage of the already established ZF proteins. However creating such a SB mutant is maybe more laborious and difficult than finding an adequate DNA binding domain that binds a unique site in the human genome.

Three different ZF proteins were designed and tested by me using the “Zinc finger tools” website from the laboratory of Carlos Barbas. ZF A was designed to bind four 5′-ANN-3′ and two 5′-GNN-3′ triplets, ZF B bound three 5′-ANN-3′ and three 5′-GNN-3′ triplets and ZF C bound six 5′-GNN-3′. Few ZF domains binding 5′-ANN-3′ type base triplets exist in nature. Examples are finger 5 (5′-AAA-3′) of Gfi-1 (Zweidler-Mckay et al. 1996), finger 3 (5′-AAT-3′) of YY1 (Hyde-DeRuyscher et al. 1995), fingers 4 and 6 (5′-A/GTA-3′) of CF2II (Gogos et al. 1996) and finger 2 (5′-AAG-3′) of (Fairall et al. 1993). However experimental data for these ZFs indicate that the 5′ adenine of the 5′-ANN-3′ triplets is actually not bound by the corresponding amino acid in the ZF α helix (Fairall et al. 1993) (Gogos et al. 1996). In multi finger ZF proteins not every individual finger necessarily binds DNA and one individual finger does not have to contact all three bases within its 3 bp binding site (Pavletich and Pabo 1993) (Nolte et al. 1998). Different amino acids at positions -1 and 3 of the ZF α-helix that bind to bases at the 3′ and middle position of the 3 base pair ZF recognition site have been identified, for position 6 of the ZF α-helix however only arginine and lysine binding to 5′-guanine have been observed. This is why before systematic screens for synthetic ZF modules uncovered ZF units that were able to bind 5′-ANN-3 base pair triplets creating such ZFs by

rational design was difficult. Aart a synthetic six finger protein was originally designed to bind to 5′-ATG TAG AGA AAA ACC AGG-3′ using the same procedure as for the design of the ZF protein E2C. Three finger proteins were assembled from characterized ZF domains using Sp1C as framework followed by fusion of the two three-finger proteins (Dreier et al. 2001). Aart bound its predicted target site with an affinity of approximately 7.5 pM. It showed strong and specific induction of gene expression in a luciferase reporter assay. However from all six fingers of the Aart ZF only finger 1 was actually expected to make contact to the 5′-A of its recognition triplet. For the other five ZFs the ZF α-helix was expected to bind only the middle and 3′-base on the DNA recognition strand. In a subsequent study cyclic amplification and selection of targets (CAST) analysis found that Aart preferably bound to a 5′-ATG (G/T)AG (A/G)GA AAA GCC CNN-3′ consensus site (Segal et al. 2003a). Affinity for the original versus the consensus binding site was only slightly increased for the consensus (90 pM and 50 pM respectively). This suggests that even if a particular multi finger ZF binds its predicted target site with good affinity, off-target binding can not be excluded, especially for ZF modules binding triplets other than 5′-GNN-3′. On the other side, artificially designed multi finger ZF proteins like E2C (Beerli et al. 1998) have shown to alter integration pattern of HIV on plasmid as well as genomic level (Tan et al. 2004) (Tan et al. 2006). In the E2C ZF all individual fingers bind 5′-GNN-3′ triplets. A high content of 5′-GNN-3′ triplets in the ZF recognition site is believed to enhance chances that a designed ZF protein will actually bind its predicted target site. ZF C tested for binding to a target site in the *LINE1.3* element in my work also consisted of individual ZFs that all bound 5′-GNN-3′ triplets. This ZF protein however did not show the best performance of all three ZF tested in binding to its recognition site in a luciferase reporter assay. Another problem for the design of ZF proteins binding 5′-ANN-3′ triplets arises with the phenomenon called target site overlap. ZF proteins carrying an aspartic acid at position 2 of ZF α-helix bind either cytosine or adenosine at the 3′-position of the complementary DNA strand of the preceding finger. The preceeding finger could thus only bind 5′-TNN-3′or 5′-GNN-3′ triplets without creating target site overlap issues. This problem can be circumvented by selecting two fingers at a time; however, for true modular assembly individual units that can be stitched together autonomously are required. When selecting an appropriate ZF protein for a given binding site possible target site overlap issues should thus be excluded. For ZF A-ZF C tested in my work no target site overlap issues were found.

While designing novel ZF proteins via modular assembly seemed a promising approach at the start of my project, publications and reports involving the design of novel ZF by modular

assembly published during the last years strongly suggest the use of some kind of selection strategy like a bacterial two hybrid system like in the OPEN strategy (Maeder et al. 2009) for design of adaquate ZF proteins.

4.3. Detecting targeted transposon insertions with Southern blot

All lanes of the Southern blot hybridized with probes specific to the E2C locus showed a faint smear of similar intensity. Only two copies of the E2C locus exist in the genome of a normal diploid human cell transposon; thus, integrations near this site by chance are expected to be quite rare. Thus targeted insertions near the E2C binding site were expected to be represented as discrete bands in this Southern blot on LAM-PCR DNA. No bands neither for E2C/SB transposases nor for SB were observed. However targeted transposon insertions could have occurred outside of the 1.5 kb window around the E2C binding site that was covered by the E2C probes or amplification of transposon insertions within this window was too weak to be detected by Southern blot.

About 500,000 copies of LINE1 elements exist in the human genome (Mandal and Kazazian 2008) which belong to different families, subfamilies and subsets thereof. Even though families, subfamilies and subsets of LINE1 elements all differ in sequence to some extent the predicted 18 bp recognition site of ZF B in the 3´-region of the particular *LINE1.3* element is found about 13,000 times in the human genome. Hence, statistically, one in about 118 transposon insertions should insert within 1 kb up- or downstream of a ZF B recognition site by chance. Since multiple ZF B binding sites exist in the human genome, and all lie within different DNA context, no discrete bands but rather a smear is expected on a Southern blot. The intensity of signal obtained with a probe that hybridized to the 3´-end of *LINE1.3* elements was stronger when transposition was mediated by ZFB/SB or N57/ZFB+SB compared to unfused SB. This could indicate that more transposon insertions occurred within LINE1 elements using ZFB/SB or N57/ZFB+SB. Transposon insertions outside LINE1 elements and thus most of transposon insertions to the 3´-end of the ZF B recognition site were not detected using these probes.

Results of the Southern blot agree with results obtained by Illumina sequencing where no targeting was found for E2C fusion proteins, but some enrichment of transposon insertions near LINE1 elements using ZF B fusion proteins was found.

4.4. Targeting on plasmid level

I determined transposon integration pattern on target plasmids using the E2C/SB fusion protein, unfused SB transposase or a mixture of both. An enrichment of transposon insertions was found for E2C/SB fusion protein in a region about 1 kb upstream the E2C binding site. Around 25 % of transposon integrations occurred in this region for transfections with E2C/SB or E2C/SB together with unfused SB transposase compared to about 9 % of transposon integrations for unfused SB transposase alone. When a plasmid which was identical to the target plasmid except that it lacked the E2C binding site was transfected together with the transposon donor plasmid, E2C/SB and unfused SB transposase the percentage of transposon integration within this region dropped to 12 %. Upon binding to the target plasmid the E2C/SB transposase selects a suitable TA dinucleotide for transposon integration. Possible transposon insertion sites are limited to TA dinucleotides accessible to the SB transposase part of the fusion protein. Within accessible DNA some TA dinucleotides might fulfil the requirements of the SB transposoase for a particular DNA structure for successful transposon integration. Targeted transposon insertions will occur at one particular site or, if more than one suitable TA dinucleotide is accessible, at multiple sites within a certain DNA region. Depending on the flexibility of the fusion protein or target DNA, spatially separated regions could be targeted concomitantly like regions up- and downstream of the binding site. However if the two moieties of the DBD/transposase fusion protein restrict spatial movement of each other, it is more likely that targeted integration of the transposon occurs within one narrow range in target DNA. The conformation of the target plasmid could play an important role for target site choice. A supercoiled plasmid might offer different regions for transposon integration than a plasmid in open circular conformation. Plasmids used for experiments were all produced by the same routine. Predominantly the supercoiled form was extracted but also plasmids with open circular conformation could be detected on an agarose gel. To create transposon integration maps on the target plasmid for different transposase proteins, target plasmid preparations of similar conformation composition were used. Thus, differences of the transposon integration pattern on target plasmids achieved with E2C/SB compared to transposon insertions pattern on target plasmids achieved with unfused SB can be assumed to result from differences in the transposase protein and not from differences in plasmid conformation. Fusion of any tag or protein to the SB transposase might suffice to alter integration preferences. However in such case no increase in transposon insertions at a

specific region would be expected but rather a different distribution of hot spots around the whole target plasmid.

For two other transposase/DBD fusions, PB/Gal4 and Mos1/Gal4, 67 % and 96 % of transposon insertions, respectively, were targeted to the same site about 1 kb upstream of the *UAS* on a target plasmid which was about 2.8 kb in size (Maragathavally et al. 2006). All targeted transposon integrations occurred in this case not only exclusively upstream of the UAS but actually into the same TA or TTAA, respectively. SB fusions to Gal4 or E2C showed a 7 or 8-fold enrichment, respectively, of transposon insertions in a 443 bp window around the corresponding binding site of the DBD in a target plasmid, which was about 5 kb in size (Yant et al. 2007). This 443 bp window contained 27 potential target TA dinucleotides with 23/27 of the target sites located at the 5´-site of the E2C binding site. Consequently, most transposon insertions occurred at the 5´-site of the E2C binding site. No clusters of transposon insertions or preferred integration into a particular TA dinucleotide were observed by Yant et al. (2007). DBD fusions to different retroviral integrases ASV IN/LexA (Katz et al. 1996), HIV IN/λ-repressor (Bushman 1994), HIV IN/LexA (Goulaouic and Chow 1996) or HIV IN/E2C (Tan et al. 2004) target insertions directly adjacent to or within about 60 bp to the DBD binding sites on the respective target plasmid. All publications used plasmids as target DNA except for (Bushman 1994) who used λ-phage DNA as target. These results could lead to the conclusion that DBD/integrases target to the immediate adjacency of the DBD binding sequence, whereas DBD/transposases insert into a wider window around it at least in plasmid context. An explanation for this observation could be that transposases often have minimal sequence requirements on target DNA for insertion (e.g. TA dinucleotide for *SB*, TTAA for piggyBac) whereas integrases do not have such minimal sequence requirements. DBD/IN fusion proteins could thus insert DNA wherever it is accessible, whereas DBD/transposases have to encounter their appropriate target site DNA signature. For the E2C/HIV IN fusions targeted integrations occurred within 10 bp upstream (to the 5´-end) of the E2C binding site. No such hot spot was found downstream (to the 3´-end) of the E2C binding site. On the contrary preferred integration sites downstream the E2C binding site found for wt HIV IN disappeared using the E2C/IN fusions (Tan et al. 2004). The same bias for integration adjacent to the 5´-end of the binding site was observed for ASV IN/LexA fusion proteins (Katz et al. 1996). For other IN fusion proteins mentioned above (HIV IN/LexA and HIV IN/λ-repressor) targeted integration occurred adjacent to the respective binding site with no site apparently favored over the other. The preference to insert at one particular site of the

respective binding site seems to depend on the individual fusion protein or maybe on target DNA composition rather than follow a general rule.

Yant et al. (2007) reported an enrichment of transposon insertions in a 443 bp window around the E2C recognition site on a target plasmid in an inter-plasmid transposition assay. The target plasmids used by Yant et al. (2007) and my target plasmid erbB2 were in both cases identical to the luciferase reporter plasmids used for E2C ZF binding studies. Both plasmids were about 5 kb in size (the plasmid used in my work was 5.5 kb and the plasmid used by Yant et al. (2007) about 5 kb in size), contained a luciferase gene, an amp resistance gene and a ColE1 ori. However the target plasmid used by Yant et al. (2007) contained five tandem unidirectional copies of the E2C recognition site which could enhance binding of the E2C/SB fusion protein compared to a single E2C recognition site on the target plasmid erbB2. The target plasmid used by Maragathavally et al. (2006) also contained five copies of the *UAS* (the Gal4 binding site). Increasing the number of DNA binding sites on the reporter plasmid lead to an increase in reporter gene expression in reporter plasmid-based DNA binding studies for the Gal4 protein (Webster et al. 1988). If multicopy binding sites on the target plasmid also enhance transposon targeting remains to be investigated. Goulaouic and Chow (1996) and Tan et al. (2004) could target HIV IN/DBD fusions on target plasmids containing a single DBD binding site. Apart from one versus five E2C binding site, the erbB2 plasmid, used in my experiments, contained 750 bp of the *erbB-2* promotor region concluding with the E2C binding site upstream the luciferase reporter gene. Thus the E2C binding site lay to its 5´-site in DNA context identical to that found in the human genome mirroring the DNA environment found for the E2C binding site in the human genome. However this DNA context offered only few TA dinucleotides that meet *SB* transposon integration preferences. Differences in transposon integration pattern between the Yant et al. (2007) study and my results could vary due to DNA sequence or sequence-induced structural differences of the DNA resulting in altered transposon integration. Yant et al. (2007) and my work use different hyperactive versions of SB transposase and different linker sequences joining ZF and transposase. This could also lead to fusion proteins with altered enzymatic performance due to structural differences within the protein and may also alter integration patterns. Comparing genomic transposon insertion patterns for SB100x and M3a transposases analysed with Illumina sequencing in this work, some difference in insertion pattern can be observed. SB100x seems to prefer insertion into Ref Seq genes (42 %) with 2 % of insertions in exons compared to M3a (38 %) with 1 % of insertions in exons. Conversely 43 % of transposon insertion

mediated by M3a occurred in silent intergenic regions and only 40 % of SB100x were found within this region.

For targeting gene insertions with ZF proteins mainly three finger ZF proteins have been created for the use in ZFNs. Liu et al. (1997) created novel six finger ZF proteins by fusing together two established three finger ZF proteins. The constructed six finger proteins showed higher affinity to their 18 bp recognition site than the three finger protein alone to their respective 9 bp recognition site. Interestingly, the six finger protein did not appear to bind its 9 bp half site as indicated by luciferase reporter assays whereas binding of a partially mutated recognition site (binding sites for finger 2 and finger 5 were partially mutated) did show some extent of induction of luciferase expression. For the E2C ZF no systematic study of binding to partial binding sites has been upublished. However, controversially to Liu et al. (1997) Yant et al. (2007) report improved targeting on plasmids with partially mutated E2C binding sites. On the target plasmid used in my work the same E2C subsite reported to achieve better targeting was present within the amp resistance gene. The partial E2C binding site was about 1.1 kb downstream to the region where an increase of transposon insertions using the E2C/SB fusion protein could be detected. Elimination of this partial binding site would maybe alter transposon integration pattern into the target plasmid.

Using the *SB* transposon insertion prediction program ProTIS (Geurts et al. 2006), a map of the target plasmid was generated showing potential *SB* insertion hot spots. Only ProTIS-predicted hot spots with the highest rating (so-called 4-peak preferred) were mapped onto the erbB2 target plasmid. Hot-spots with a lower rank (so-called 3.5 to 2.5-peak semi preferred) were not indicated. One major integration hot spot predicted by ProTIS which lay between the *amp* gene and the ColE1 ori corresponded with an integration hot spot actually observed in the inter-plasmid transposition assay for all transposases (Fig. 11). Two hot spots upstream the ColE1 ori could also be verified *in vivo*, whereas hotspots predicted to be within the region of the ColE1 ori were not observed. This is probably due to the fact that insertions at some of these sites within the ColE1 ori as well as in the antibiotic resistance gene will be detrimental to plasmid propagation. Insertion hot spots downstream the *erbB-2* promotor fragment were also not confirmed by *in vivo* experiments. Instead of a limited number of transposon integration hot spots predicted by ProTIS the inter-plasmid transposition results show a uniform distribution with one insertion hotspot about 1.3 kb from the E2C binding site. No hot spots with the highest ranking were calculated for the *erbB-2* fragment and the region between the *amp* gene and the *erbB-2* promoter fragment. Within this region the increase in transposon insertions was seen for E2C ZF-containing transposases. However

some lower-ranked hot spots (so-called 3.5-peak semi preferred) can be found within this region. Worth mentioning is that within the first 500 bp upstream the E2C binding site within the *erbB-2* fragment not a single hot spot regardless to its ranking was observed. The ProTIS prediction program was developed on the basis of transposon insertion mediated by the original reconstructed first version of the SB transposase (Ivics and Izsvak 1997) if different hyperactive versions of this transposase have altered target site preferences in DNA sequence level remains to be investigated**.**

4.5. Truncated versions of the SB transposase and ZF/SB transposase fusion proteins

Western blots hybridized with polyclonal Anti-*Sleeping Beauty* Transposase Antibodies raised against full length transposase protein detected protein of expected size corresponding to SB transposase (37 kDa) for transfections with M3a and ZF/SB fusion protein (55 kDa) for transfections with the fusion protein but also identified bands of smaller size (about 28 kDa for transfections with M3a and about 37 kDa for transfections with fusion proteins). Proteins were extracted in the presence of a protease inhibitor cocktail to avoid degradation of proteins by proteases, but the same protein bands were detected in three independent experiments. Occurrence of smaller products by proteolytic cleavage during or after protein extraction from cells can not fully be excluded. However, truncated versions of the SB transposase protein have been observed in prokaryotic as well as eukaryotic cells before by members of our laboratory. No increase in the amount of the proteolytic cleavage product relative to full length transposase was observed with increase of expression. The amount of the proteolytic cleavage product rather increased with time after induction. Because of these observations Izsvak and coworkers came to the conclusion that truncated transposase versions were proteolytic cleavage products rather than products of early transcription or translation termination. Some SB transposase mutants with amino acid changes introduced at characteristic proteolytic cleavage sites showed somewhat enhanced transposition activity (Izsvak, Ivics, Plasterk unpublished data) which could argue for a regulatory role of proteolytic cleavage products in the transposition process.

If full length SB transposase was proteolytically cleaved resulting in a truncated fragment of about 28 kD at least one additional band of about 9 kD would be expected to be detected. Such fragment however was not seen. This may be due to insufficient sensitivity of the Western blot, further cleavage of the fragment or complete degradation of the small fragment

in the cell after cleavage. The size of the truncated version of ZF B/SB suggests that it is not product of the same cleavage reaction that resulted in the truncated version of unfused SB. If the truncated 28 kD protein detected for unfused SB transposase represents the C-terminus of the SB transposase the same protein fragment would be expected to show for ZF B/SB. An N-terminal 28 kD SB transposase fragment still fused to the 18 kD ZF would show at about 46 kD. ZF B and SB transposase brought close together in the ZF B/SB fusion protein may influence their individual folding. Altered accessibility of certain protein domains may account for the drop in transpositional activity seen for ZF B/SB compared to unfused SB transposase. This altered accessibility of domains in the SB transposase may also be the reason for the unexpected size of the truncated protein. The size of the truncated protein for ZF B/SB runs at the same height as the unfused SB transposase. This could suggest that ZF B got cleaved off from the ZF B/SB fusion protein. Truncated protein fragments were not further analyzed but even if the ZF got cleaved from the transposase it would be questionable if this cleaved transposase was still able to catalyze transposition. Furthermore the targeting strategy of mixing unfused SB transposase with a targeting molecule was already successfully applied (Ivics et al. 2007). Some amount of cleaved and thus unfused SB transposase could even be helpful for targeting.

Regulation of transposition by truncated versions of the transposase protein is known for the bacterial transposase Tn5. Apart from the full length Tn5 transposase an N-terminal truncated version called Inhibitor (Inh) is also expressed from the *Tn5* transposon using a alternative start codon. Inh lacks the N-terminal residues important for transposon end binding and is therefore inactive as a transposase. Inh however still contains the dimerisation interface and can lead to inactivation of otherwise functional full length Tn5 transposase proteins through heterodimerisation. Inh in fact enhances binding to the *Tn5* transposon but inhibits excision of the transposon (de la Cruz et al. 1993). Inh also shows enhanced ability to dimerize compared to full length Tn5 transposase (Mahnke Braam et al. 1999) and under steady-state conditions fourfold more Inh is found in cells compared to full length Tn5 (Johnson and Reznikoff 1984).

Truncated versions of the transposase protein were also found for the *Tc1* element from *Caenorhabditis elegans.* Expressed in *E.coli* a truncated transposase protein consisting of the N-terminal approximately 153 aa of the *Tc1* transposase (Tc1A) was accountable for the majority of DNA/protein complexes found in bandshift assays with DNA representing *Tc1* transposon ends. Full length Tc1A complexed with DNA was far less abundant even though full length Tc1A was expressed at much higher levels than the truncated version (Vos et al.

1993). Two truncated transposase protein consisting of the N-terminal 78 aa and about 153 aa were isolated from *E.coli* together with the full length Tc1A transposase. Both N-terminal transposase fragments were able to bind to the transposon ends (Vos and Plasterk 1994).

Regulation of transposition by truncated transposase versions has been described for the prokaryotic *IS911* element. *IS911* encodes two proteins: OrfAB, the transpositional active protein, results from a translational frameshift between two partially overlapping ORFs and OrfA which shares the N-terminal 86 aa of OrfAB but differs in its C-terminal 14 aa. OrfA stimulates OrfAB-mediated integration reactions (Ton-Hoang et al. 1998). Aside from these two proteins truncated versions of OrfAB are found in *E.coli* which have an inhibitory effect on OrfAB activity (Gueguen et al. 2006).

It remains to be clarified which part of the SB transposase is represented in the truncated versions detected on Western blots and if it serves any purpose like regulation of transposition.

5. References

NIH report 2002. Assessment of adenoviral vector safety and toxicity: report of the National Institutes of Health Recombinant DNA Advisory Committee. *Hum Gene Ther* **13**(1): 3-13.

Akopian, A., He, J., Boocock, M.R., and Stark, W.M. 2003. Chimeric recombinases with designed DNA sequence recognition. *Proc Natl Acad Sci U S A* **100**(15): 8688-8691.

Anastassiadis, K., Fu, J., Patsch, C., Hu, S., Weidlich, S., Duerschke, K., Buchholz, F., Edenhofer, F., and Stewart, A.F. 2009. Dre recombinase, like Cre, is a highly efficient site-specific recombinase in E. coli, mammalian cells and mice. *Dis Model Mech* **2**(9-10): 508-515.

Arai, R., Ueda, H., Kitayama, A., Kamiya, N., and Nagamune, T. 2001. Design of the linkers which effectively separate domains of a bifunctional fusion protein. *Protein Eng* **14**(8): 529-532.

Aye, M., Dildine, S.L., Claypool, J.A., Jourdain, S., and Sandmeyer, S.B. 2001. A truncation mutant of the 95-kilodalton subunit of transcription factor IIIC reveals asymmetry in Ty3 integration. *Mol Cell Biol* **21**(22): 7839-7851.

Bachman, N., Gelbart, M.E., Tsukiyama, T., and Boeke, J.D. 2005. TFIIIB subunit Bdp1p is required for periodic integration of the Ty1 retrotransposon and targeting of Isw2p to S. cerevisiae tDNAs. *Genes Dev* **19**(8): 955-964.

Bae, K.H., Kwon, Y.D., Shin, H.C., Hwang, M.S., Ryu, E.H., Park, K.S., Yang, H.Y., Lee, D.K., Lee, Y., Park, J., Kwon, H.S., Kim, H.W., Yeh, B.I., Lee, H.W., Sohn, S.H., Yoon, J., Seol, W., and Kim, J.S. 2003. Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nat Biotechnol* **21**(3): 275-280.

Bazopoulou, D. and Tavernarakis, N. 2009. The NemaGENETAG initiative: large scale transposon insertion gene-tagging in Caenorhabditis elegans. *Genetica* **137**(1): 39-46.

Beerli, R.R., Barbas, C.F., 3rd. 2002. Engineering polydactyl zinc-finger transcription factors. *Nat Biotechnol* **20**(2): 135-141

Beerli, R.R., Segal, D.J., Dreier, B., and Barbas, C.F., 3rd. 1998. Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. *Proc Natl Acad Sci U S A* **95**(25): 14628-14633.

Blancafort, P., Magnenat, L., and Barbas, C.F., 3rd. 2003. Scanning the human genome with combinatorial transcription factor libraries. *Nat Biotechnol* **21**(3): 269-274.

Boissinot, S., Chevret, P., and Furano, A.V. 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol Biol Evol* **17**(6): 915-928.

Bourc'his, D. and Bestor, T.H. 2004. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**(7004): 96-99.

Bowers, W.J., Mastrangelo, M.A., Howard, D.F., Southerland, H.A., Maguire-Zeiss, K.A., and Federoff, H.J. 2006. Neuronal precursor-restricted transduction via in utero CNS gene delivery of a novel bipartite HSV amplicon/transposase hybrid vector. *Mol Ther* **13**(3): 580-588.

Bryk, M., Banerjee, M., Murphy, M., Knudsen, K.E., Garfinkel, D.J., and Curcio, M.J. 1997. Transcriptional silencing of Ty1 elements in the RDN1 locus of yeast. *Genes Dev* **11**(2): 255-269.

Buchholz, F., Angrand, P.O., and Stewart, A.F. 1998. Improved properties of FLP recombinase evolved by cycling mutagenesis. *Nat Biotechnol* **16**(7): 657-662.

Buchholz, F. and Stewart, A.F. 2001. Alteration of Cre recombinase site specificity by substrate-linked protein evolution. *Nat Biotechnol* **19**(11): 1047-1052.

Bushman, F.D. 1994. Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. *Proc Natl Acad Sci U S A* **91**(20): 9233-9237.

Bushman, F.D. and Miller, M.D. 1997. Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. *J Virol* **71**(1): 458-464.

Cadinanos, J. and Bradley, A. 2007. Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Res* **35**(12): e87.

Caldovic, L. and Hackett, P.B., Jr. 1995. Development of position-independent expression vectors and their transfer into transgenic fish. *Mol Mar Biol Biotechnol* **4**(1): 51-61.

Cathomen, T., Collete, D., and Weitzman, M.D. 2000. A chimeric protein containing the N terminus of the adeno-associated virus Rep protein recognizes its target site in an in vivo assay. *J Virol* **74**(5): 2372-2382.

Chalberg, T.W., Portlock, J.L., Olivares, E.C., Thyagarajan, B., Kirby, P.J., Hillman, R.T., Hoelters, J., and Calos, M.P. 2006. Integration specificity of phage phiC31 integrase in the human genome. *J Mol Biol* **357**(1): 28-48.

Choo, Y. 1998. Recognition of DNA methylation by zinc fingers. *Nat Struct Biol* **5**(4): 264-265.

Choo, Y. and Klug, A. 1994. Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc Natl Acad Sci U S A* **91**(23): 11163-11167.

Chung, T., Siol, O., Dingermann, T., and Winckler, T. 2007. Protein interactions involved in tRNA gene-specific integration of Dictyostelium discoideum non-long terminal repeat retrotransposon TRE5-A. *Mol Cell Biol* **27**(24): 8492-8501.

Ciuffi, A., Diamond, T.L., Hwang, Y., Marshall, H.M., and Bushman, F.D. 2006. Modulating target site selection during human immunodeficiency virus DNA integration in vitro with an engineered tethering factor. *Hum Gene Ther* **17**(9): 960-967.

Ciuffi, A., Llano, M., Poeschla, E., Hoffmann, C., Leipzig, J., Shinn, P., Ecker, J.R., and Bushman, F. 2005. A role for LEDGF/p75 in targeting HIV DNA integration. *Nat Med* **11**(12): 1287-1289.

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

Collins, C.H., Yokobayashi, Y., Umeno, D., and Arnold, F.H. 2003. Engineering proteins that bind, move, make and break DNA. *Curr Opin Biotechnol* **14**(4): 371-378.

Corbi, N., Libri, V., Fanciulli, M., Tinsley, J.M., Davies, K.E., and Passananti, C. 2000. The artificial zinc finger coding gene 'Jazz' binds the utrophin promoter and activates transcription. *Gene Ther* **7**(12): 1076-1083.

Cortes, M.L., Oehmig, A., Saydam, O., Sanford, J.D., Perry, K.F., Fraefel, C., and Breakefield, X.O. 2008. Targeted Integration of Functional Human ATM cDNA Into Genome Mediated by HSV/AAV Hybrid Amplicon Vector. *Mol Ther* **16**(1): 81-88.

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

Daniel, J.M., Spring, C.M., Crawford, H.C., Reynolds, A.B., and Baig, A. 2002. The p120(ctn)-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. *Nucleic Acids Res* **30**(13): 2911-2919.

de la Cruz, N.B., Weinreich, M.D., Wiegand, T.W., Krebs, M.P., and Reznikoff, W.S. 1993. Characterization of the Tn5 transposase and inhibitor proteins: a model for the inhibition of transposition. *J Bacteriol* **175**(21): 6932-6938.

Desjarlais, J.R. and Berg, J.M. 1992. Towards rules relating zinc finger protein sequences and DNa binding proteins. *Proc Natl Acad Sci U S A* **89**(16): 7345-7349.

-. 1993. Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc Natl Acad Sci U S A* **90**(6): 2256-2260.

Dhanasekaran, M., Negi, S., Imanishi, M., and Sugiura, Y. 2007. DNA-Binding ability of GAGA zinc finger depends on the nature of amino acids present in the beta-hairpin. *Biochemistry* **46**(25): 7506-7513.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. 2005. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122**(3): 473-483.

Dreier, B., Beerli, R.P., Segal, D.J., Flippin, J.D., and Barbas, C.F., 3rd. 2001. Development of zinc finger domains for recognition of the 5′-ANN-3′ family of DNA sequences and their use in the construction of artificil transcription factors. *J Biol Chem* **276**(31): 29466-29478

Dreier, B., Fuller, R.P., Segal, D.J., Lund, C.V., Blancafort, P., Huber, A., Koksch, B., and Barbas, C.F., 3rd. 2005. Development of zinc finger domains for recognition of the 5′-CNN-3′ familiy DNA sequences and their use in the construction of artificial transcription factors. *J Biol Chem* **280**(42):35588-35597

Dreier, B., egal, D.J., and Barbas, C.F., 3rd. 2000. Insights into the molecular recognition of the 5′-GNN-3′ family of DNA sequences by zinc finger domains. *J Mol Biol* **303**(4):498-502

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

Fairall, L., Schwabe, J.W., Chapman, L., Finch, J.T., and Rhodes, D. 1993. The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature* **366**(6454): 483-487.

Ferris, A.L., Wu, X., Hughes, C.M., Stewart, C., Smith, S.J., Milne, T.A., Wang, G.G., Shun, M.C., Allis, C.D., Engelman, A., and Hughes, S.H. 2010. Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration. *Proc Natl Acad Sci U S A* **107**(7): 3135-3140.

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

Follenzi, A., Santambrogio, L., and Annoni, A. 2007. Immune responses to lentiviral vectors. *Curr Gene Ther* **7**(5): 306-315.

Forni, P.E., Scuoppo, C., Imayoshi, I., Taulli, R., Dastru, W., Sala, V., Betz, U.A., Muzzi, P., Martinuzzi, D., Vercelli, A.E., Kageyama, R., and Ponzetto, C. 2006. High levels of Cre expression in neuronal progenitors cause defects in brain development leading to microencephaly and hydrocephaly. *J Neurosci* **26**(37): 9593-9602.

Fu, F., Sander, J.D., Maeder, M., Thibodeau-Beganny, S., Joung, J.K., Dobbs, D., Miller, L., and Voytas, D.F. 2009. Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays. *Nucleic Acids Res* **37**(Database issue): D279-283

Garrison, B.S., Yant S.R., Mikkelsen J.G., Kay M.A. 2007. Postintegrative gene silencing within the Sleeping Beauty transposition system. *Mol Cell Biol* (24): 8824-33

Ge, H., Si, Y., and Roeder, R.G. 1998. Isolation of cDNAs encoding novel transcription coactivators p52 and p75 reveals an alternate regulatory mechanism of transcriptional activation. *Embo J* **17**(22): 6723-6729.

Gersbach, C.A., Gaj, T., Gordley, R.M. and Barbas, C.F., 3rd. 2010. Directed evolution of recombinase specificity by slit gene assembly. *Nucleic Acids Res* **34**(9): 2803-2811

Geurts, A.M., Hackett, C.S., Bell, J.B., Bergemann, T.L., Collier, L.S., Carlson, C.M., Largaespada, D.A., and Hackett, P.B. 2006. Structure-based prediction of insertion-site preferences of transposons into chromosomes. *Nucleic Acids Res* **34**(9): 2803-2811.

Geurts, A.M., Yang, Y., Clark, K.J., Liu, G., Cui, Z., Dupuy, A.J., Bell, J.B., Largaespada, D.A., and Hackett, P.B. 2003. Gene transfer into genomes of human cells by the sleeping beauty transposon system. *Mol Ther* **8**(1): 108-117.

Gijsbers, R., Ronen, K., Vets, S., Malani, N., De Rijck, J., McNeely, M., Bushman, F.D., and Debyser, Z. 2010. LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. *Mol Ther* **18**(3): 552-560.

Gogos, J.A., Jin, J., Wan, H., Kokkinidis, M., and Kafatos, F.C. 1996. Recognition of diverse sequences by class I zinc fingers: asymmetries and indirect effects on specificity in the interaction between CF2II and A+T-rich elements. *Proc Natl Acad Sci U S A* **93**(5): 2159-2164.

Gordley, R.M., Gersbach, C.A., and Barbas, C.F., 3rd. 2009. Synthesis of programmable integrases. *Proc Natl Acad Sci U S A* **106**(13): 5053-5058.

Gordley, R.M., Smith, J.D., Graslund, T., and Barbas, C.F., 3rd. 2007. Evolution of programmable zinc finger-recombinases with activity in human cells. *J Mol Biol* **367**(3): 802-813.

Goulaouic, H. and Chow, S.A. 1996. Directed integration of viral DNA mediated by fusion proteins consisting of human immunodeficiency virus type 1 integrase and Escherichia coli LexA protein. *J Virol* **70**(1): 37-46.

Grabundzija, I., Irgang, M., Mates, L., Belay, E., Matrai, J., Gogol-Doring, A., Kawakami, K., Chen, W., Ruiz, P., Chuah, M.K., Vandendriessche, T., Izsvak, Z., and Ivics, Z. 2010 Comparative Analysis of Transposable Element Vector Systems in Human Cells. *Mol Ther*.

Greisman, H.A. and Pabo, C.O. 1997. A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* **275**(5300): 657-661.

Gueguen, E., Rousseau, P., Duval-Valentin, G., and Chandler, M. 2006. Truncated forms of IS911 transposase downregulate transposition. *Mol Microbiol* **62**(4): 1102-1116.

Hacein-Bey-Abina, S., Le Deist, F., Carlier, F., Bouneaud, C., Hue, C., De Villartay, J.P., Thrasher, A.J., Wulffraat, N., Sorensen, R., Dupuis-Girod, S., Fischer, A., Davies, E.G., Kuis, W., Leiva, L., and Cavazzana-Calvo, M. 2002. Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *N Engl J Med* **346**(16): 1185-1193.

Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M.P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C.S., Pawliuk, R., Morillon, E., Sorensen, R., Forster, A., Fraser, P., Cohen, J.I., de Saint Basile, G., Alexander, I., Wintergerst, U., Frebourg, T., Aurias, A., Stoppa-Lyonnet, D., Romana, S., Radford-Weiss, I., Gross, F., Valensi, F., Delabesse, E., Macintyre, E., Sigaux, F., Soulier, J., Leiva, L.E., Wissler, M., Prinz, C., Rabbitts, T.H., Le Deist, F., Fischer, A., and Cavazzana-Calvo, M. 2003. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**(5644): 415-419.

Hamlet, M.R., Yergeau, D.A., Kuliyev, E., Takeda, M., Taira, M., Kawakami, K., and Mead, P.E. 2006. Tol2 transposon-mediated transgenesis in Xenopus tropicalis. *Genesis* **44**(9): 438-445.

Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M., and Tilghman, S.M. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**(6785): 486-489.

Harraghy, N., Gaussin, A., and Mermod, N. 2008. Sustained transgene expression using MAR elements. *Curr Gene Ther* **8**(5): 353-366.

Houbaviy, H.B., Usheva, A., Shenk, T., and Burley, S.K. 1996. Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proc Natl Acad Sci U S A* **93**(24): 13577-13582.

Hyde-DeRuyscher, R.P., Jennings, E., and Shenk, T. 1995. DNA binding sites for the transcriptional activator/repressor YY1. *Nucleic Acids Res* **23**(21): 4457-4465.

Isalan, M., Klug, A., and Choo, Y. 2001. A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nat Biotechnol* **19**(7): 656-660.

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

Ivics, Z. and Izsvak, Z. 1997. Family of plasmid vectors for the expression of beta-galactosidase fusion proteins in eukaryotic cells. *Biotechniques* **22**(2): 254-256, 258.

-. 2006. Transposons for gene therapy! *Curr Gene Ther* **6**(5): 593-607.

Ivics, Z., Katzer, A., Stuwe, E.E., Fiedler, D., Knespel, S., and Izsvak, Z. 2007. Targeted sleeping beauty transposition in human cells. *Mol Ther* **15**(6): 1137-1144.

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

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

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

Jamieson, A.C., Kim, S.H., and Wells, J.A. 1994. In vitro selection of zinc fingers with altered DNA-binding specificity. *Biochemistry* **33**(19): 5689-5695

Jamieson, A.C., Wang, H., and Kim, S.H., 1996. A zinc finger directory for high-affinity DNA recognition. *Proc Natl Acad Sci U S A* **93**(23):12834-12839

Jaskolski, M., Alexandratos, J.N., Bujacz, G., and Wlodawer, A. 2009. Piecing together the structure of retroviral integrase, an important target in AIDS therapy. *Febs J* **276**(11): 2926-2946.

Johnson , R.c. and Rezikoff, W.s. 1984. Role of the IS50 R proteins in the promotion and control of Tn5 transposition. *J Mol Biol* **177**(4):645-661

Jones, P.A., Rideaut, W.M., 3rd, Shen, J.C., Spruck, C.H., and Tsai, Y.c. 1992. Methylation, mutation and cancer. *Bioessays* **14**(1):33-36

Katz, R.A., Merkel, G., and Skalka, A.M. 1996. Targeting of retroviral integrase by fusion to a heterologous DNA binding domain: in vitro activities and incorporation of a fusion protein into viral particles. *Virology* **217**(1): 178-190.

Kawakami, K., Shima, A., and Kawakami, N. 2000. Identification of a functional transposase of the Tol2 element, an Ac-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage. *Proc Natl Acad Sci U S A* **97**(21): 11403-11408.

Kazazian, H.H., Jr. and Moran, J.V. 1998. The impact of L1 retrotransposons on the human genome. *Nat Genet* **19**(1): 19-24.

Kazazian, H.H., Jr., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., and Antonarakis, S.E. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**(6160): 164-166.

Khan, E., Mack, J.P., Katz, R.A., Kulkosky, J., and Skalka, A.M. 1991. Retroviral integrase domains: DNA binding and the recognition of LTR sequences. *Nucleic Acids Res* **19**(4): 851-860.

Kim, A., Terzian, C., Santamaria, P., Pelisson, A., Purd'homme, N., and Bucheton, A. 1994. Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster. *Proc Natl Acad Sci U S A* **91**(4): 1285-1289.

Kim, H.J., Lee, H.J., Kim, H., Cho, S.W., and Kim, J.S. 2009. Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. *Genome Res* **19**(7): 1279-1288.

Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., and Voytas, D.F. 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. *Genome Res* **8**(5): 464-478.

Kim, J.S., Lee, H.J., and Carroll, D. 2010. Genome editing with modularly assembled zinc-finger nucleases. *Nat Methods* **7**(2): 91; author reply 91-92.

Kipp, M., Gohring, F., Ostendorp, T., van Drunen, C.M., van Driel, R., Przybylski, M., and Fackelmayer, F.O. 2000. SAF-Box, a conserved protein domain that specifically recognizes scaffold attachment region DNA. *Mol Cell Biol* **20**(20): 7480-7489.

Kirchner, J., Connolly, C.M., and Sandmeyer, S.B. 1995. Requirement of RNA polymerase III transcription factors for in vitro position-specific integration of a retroviruslike element. *Science* **267**(5203): 1488-1491.

Klippel, A., Cloppenborg, K., and Kahmann, R. 1988. Isolation and characterization of unusual gin mutants. *Embo J* **7**(12): 3983-3989.

Koga, A., Suzuki, M., Inagaki, H., Bessho, Y., and Hori, H. 1996. Transposable element in fish. *Nature* **383**(6595): 30.

Kowolik, C.M., Topp, M.S., Gonzalez, S., Pfeiffer, T., Olivares, S., Gonzalez, N., Smith, D.D., Forman, S.J., Jensen, M.C., and Cooper, L.J. 2006. CD28 costimulation provided through a CD19-specific chimeric antigen receptor enhances in vivo persistence and antitumor efficacy of adoptively transferred T cells. *Cancer Res* **66**(22): 10995-11004.

Kren, B.T., Unger, G.M., Sjeklocha, L., Trossen, A.A., Korman, V., Diethelm-Okita, B.M., Reding, M.T., and Steer, C.J. 2009. Nanocapsule-delivered Sleeping Beauty mediates therapeutic Factor VIII expression in liver sinusoidal endothelial cells of hemophilia A mice. *J Clin Invest* **119**(7): 2086-2099.

Kulkosky, J., Katz, R.A., Merkel, G., and Skalka, A.M. 1995. Activities and substrate specificity of the evolutionarily conserved central domain of retroviral integrase. *Virology* **206**(1): 448-456.

Lander, E.S. Linton, L.M. Birren, B. Nusbaum, C. Zody, M.C. Baldwin, J. Devon, K. Dewar, K. Doyle, M. FitzHugh, W. Funke, R. Gage, D. Harris, K. Heaford, A. Howland, J. Kann, L. Lehoczky, J. LeVine, R. McEwan, P. McKernan, K. Meldrim, J. Mesirov, J.P. Miranda, C. Morris, W. Naylor, J. Raymond, C. Rosetti, M. Santos, R. Sheridan, A. Sougnez, C. Stange-Thomann, N. Stojanovic, N. Subramanian, A. Wyman, D. Rogers, J. Sulston, J. Ainscough, R. Beck, S. Bentley, D. Burton, J. Clee, C. Carter, N. Coulson, A. Deadman, R. Deloukas, P. Dunham, A. Dunham, I. Durbin, R. French, L. Grafham, D. Gregory, S. Hubbard, T. Humphray, S. Hunt, A. Jones, M. Lloyd, C. McMurray, A. Matthews, L. Mercer, S. Milne, S. Mullikin, J.C. Mungall, A. Plumb, R. Ross, M. Shownkeen, R. Sims, S. Waterston, R.H. Wilson, R.K. Hillier, L.W. McPherson, J.D. Marra, M.A. Mardis, E.R. Fulton, L.A. Chinwalla, A.T. Pepin, K.H. Gish, W.R. Chissoe, S.L. Wendl, M.C. Delehaunty, K.D. Miner, T.L. Delehaunty, A. Kramer, J.B. Cook, L.L. Fulton, R.S. Johnson, D.L. Minx, P.J. Clifton, S.W. Hawkins,

T. Branscomb, E. Predki, P. Richardson, P. Wenning, S. Slezak, T. Doggett, N. Cheng, J.F. Olsen, A. Lucas, S. Elkin, C. Uberbacher, E. Frazier, M. Gibbs, R.A. Muzny, D.M. Scherer, S.E. Bouck, J.B. Sodergren, E.J. Worley, K.C. Rives, C.M. Gorrell, J.H. Metzker, M.L. Naylor, S.L. Kucherlapati, R.S. Nelson, D.L. Weinstock, G.M. Sakaki, Y. Fujiyama, A. Hattori, M. Yada, T. Toyoda, A. Itoh, T. Kawagoe, C. Watanabe, H. Totoki, Y. Taylor, T. Weissenbach, J. Heilig, R. Saurin, W. Artiguenave, F. Brottier, P. Bruls, T. Pelletier, E. Robert, C. Wincker, P. Smith, D.R. Doucette-Stamm, L. Rubenfield, M. Weinstock, K. Lee, H.M. Dubois, J. Rosenthal, A. Platzer, M. Nyakatura, G. Taudien, S. Rump, A. Yang, H. Yu, J. Wang, J. Huang, G. Gu, J. Hood, L. Rowen, L. Madan, A. Qin, S. Davis, R.W. Federspiel, N.A. Abola, A.P. Proctor, M.J. Myers, R.M. Schmutz, J. Dickson, M. Grimwood, J. Cox, D.R. Olson, M.V. Kaul, R. Raymond, C. Shimizu, N. Kawasaki, K. Minoshima, S. Evans, G.A. Athanasiou, M. Schultz, R. Roe, B.A. Chen, F. Pan, H. Ramser, J. Lehrach, H. Reinhardt, R. McCombie, W.R. de la Bastide, M. Dedhia, N. Blocker, H. Hornischer, K. Nordsiek, G. Agarwala, R. Aravind, L. Bailey, J.A. Bateman, A. Batzoglou, S. Birney, E. Bork, P. Brown, D.G. Burge, C.B. Cerutti, L. Chen, H.C. Church, D. Clamp, M. Copley, R.R. Doerks, T. Eddy, S.R. Eichler, E.E. Furey, T.S. Galagan, J. Gilbert, J.G. Harmon, C. Hayashizaki, Y. Haussler, D. Hermjakob, H. Hokamp, K. Jang, W. Johnson, L.S. Jones, T.A. Kasif, S. Kaspryzk, A. Kennedy, S. Kent, W.J. Kitts, P. Koonin, E.V. Korf, I. Kulp, D. Lancet, D. Lowe, T.M. McLysaght, A. Mikkelsen, T. Moran, J.V. Mulder, N. Pollara, V.J. Ponting, C.P. Schuler, G. Schultz, J. Slater, G. Smit, A.F. Stupka, E. Szustakowski, J. Thierry-Mieg, D. Thierry-Mieg, J. Wagner, L. Wallis, J. Wheeler, R. Williams, A. Wolf, Y.I. Wolfe, K.H. Yang, S.P. Yeh, R.F. Collins, F. Guyer, M.S. Peterson, J. Felsenfeld, A. Wetterstrand, K.A. Patrinos, A. Morgan, M.J. de Jong, P. Catanese, J.J. Osoegawa, K. Shizuya, H. Choi, S. and Chen, Y.J. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lane, H.A., Beuvink, I., Motoyama, A.B., Daly, J.M., Neve, R.M., and Hynes, N.E. 2000. ErbB2 potentiates breast tumor proliferation through modulation of p27(Kip1)-Cdk2 complex formation: receptor overexpression does not determine growth dependency. *Mol Cell Biol* **20**(9): 3210-3223.

Lee, M.S., Gippert, G.P., Soman, K.V., Case, D.A., and Wright, P.E. 1989. Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science* **245**(4918): 635-637.

Lippman, Z., Gendrel, A.V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., Carrington, J.C., Doerge, R.W., Colot, V., and Martienssen, R. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**(6998): 471-476.

Liu, G., Geurts, A.M., Yae, K., Srinivasan, A.R., Fahrenkrug, S.C., Largaespada, D.A., Takeda, J., Horie, K., Olson, W.K., and Hackett, P.B. 2005. Target-site preferences of Sleeping Beauty transposons. *J Mol Biol* **346**(1): 161-173.

Liu, J., Jeppesen, I., Nielsen, K., and Jensen, T.G. 2006. Phi c31 integrase induces chromosomal aberrations in primary human fibroblasts. *Gene Ther* **13**(15): 1188-1190.

Liu, Q., Segal, D.J., Ghiara, J.B., and Barbas, C.F., 3rd. 1997. Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. *Proc Natl Acad Sci U S A* **94**(11): 5525-5530.

Liu, Q., Xia, Z., Zhong, X., and Case, C.C., 2002. Validated zinc finger protein designs for all 16 GNN DNA triplet targets. *J Biol Chem* **277**(6):3850-3856.

Llano, M., Vanegas, M., Fregoso, O., Saenz, D., Chung, S., Peretz, M., and Poeschla, E.M. 2004. LEDGF/p75 determines cellular trafficking of diverse lentiviral but not murine oncoretroviral integrase proteins and is a component of functional lentiviral preintegration complexes. *J Virol* **78**(17): 9524-9537.

Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. 2002. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14**(7): 1605-1619.

Lombardo, A., Genovese, P., Beausejour, C.M., Colleoni, S., Lee, Y.L., Kim, K.A., Ando, D., Urnov, F.D., Galli, C., Gregory, P.D., Holmes, M.C., and Naldini, L. 2007. Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nat Biotechnol* **25**(11): 1298-1306.

Loomis, W.F., Welker, D., Hughes, J., Maghakian, D., and Kuspa, A. 1995. Integrated maps of the chromosomes in Dictyostelium discoideum. *Genetics* **141**(1): 147-157.

Loonstra, A., Vooijs, M., Beverloo, H.B., Allak, B.A., van Drunen, E., Kanaar, R., Berns, A., and Jonkers, J. 2001. Growth inhibition and DNA damage induced by Cre recombinase in mammalian cells. *Proc Natl Acad Sci U S A* **98**(16): 9209-9214.

Lu, B., Geurts, A.M., Poirier, C., Petit, D.C., Harrison, W., Overbeek, P.A., and Bishop, C.E. 2007. Generation of rat mutants using a coat color-tagged Sleeping Beauty transposon system. *Mamm Genome* **18**(5): 338-346.

Maeder, M.L., Thibodeau-Beganny, S., Sander, J.D., Voytas, D.F., and Joung, J.K. 2009. Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat Protoc* **4**(10): 1471-1501.

Mahnke Braam, L.A., Goryshin, I.Y., and Reznikoff, W.S. 1999. A mechanism for Tn5 inhibition. carboxyl-terminal dimerization. *J Biol Chem* **274**(1): 86-92.

Mandal, P.K. and Kazazian, H.H., Jr. 2008. SnapShot: Vertebrate transposons. *Cell* **135**(1): 192-192 e191.

Mandell, J.G. and Barbas, C.F., 3rd. 2006. Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases. *Nucleic Acids Res* **34**(Web Server issue): W516-523.

Maragathavally, K.J., Kaminski, J.M., and Coates, C.J. 2006. Chimeric Mos1 and piggyBac transposases result in site-directed integration. *Faseb J* **20**(11): 1880-1882.

Martens, J.H., O'Sullivan, R.J., Braunschweig, U., Opravil, S., Radolf, M., Steinlein, P., and Jenuwein, T. 2005. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *Embo J* **24**(4): 800-812.

Martienssen, R.A., Zaratiegui, M., and Goto, D.B. 2005. RNA interference and heterochromatin in the fission yeast Schizosaccharomyces pombe. *Trends Genet* **21**(8): 450-456.

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

Mates, L., Izsvak, Z., and Ivics, Z. 2007. Technology transfer from worms and flies to vertebrates: transposition-based genome manipulations and their future perspectives. *Genome Biol* **8 Suppl 1**: S1.

McClintock, B. 1946. *Maize genetics*. Carnegie Inst. of Wash. Yearbook 45: 176-186.

-. 1947. *Cytogenetic studies of maize and Neurospora*. Carnegie Inst. of Wash. Year Book 46: 146-152.

-. 1948. *Mutable loci in maize*. Carnegie Inst. of Wash. Year Book 47: 155-169.

Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K.W., Vogelstein, B., and Nakamura, Y. 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res* **52**(3): 643-645.

Mikkelsen, J.G., Yant, S.R., Meuse, L., Huang, Z., Xu, H., and Kay, M.A. 2003. Helper-Independent Sleeping Beauty transposon-transposase vectors for efficient nonviral gene delivery and persistent gene expression in vivo. *Mol Ther* **8**(4): 654-665.

Miller, J., McLachlan, A.D., and Klug, A. 1985. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *Embo J* **4**(6): 1609-1614.

Miskey, C., Izsvak, Z., Kawakami, K., and Ivics, Z. 2005. DNA transposons in vertebrate functional genomics. *Cell Mol Life Sci* **62**(6): 629-641.

Mitchell, R.S., Beitzel, B.F., Schroder, A.R., Shinn, P., Chen, H., Berry, C.C., Ecker, J.R., and Bushman, F.D. 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol* **2**(8): E234.

Mizuguchi, H. and Hayakawa, T. 2002. Adenovirus vectors containing chimeric type 5 and type 35 fiber proteins exhibit altered and expanded tropism and increase the size limit of foreign genes. *Gene* **285**(1-2): 69-77.

Narezkina, A., Taganov, K.D., Litwin, S., Stoyanova, R., Hayashi, J., Seeger, C., Skalka, A.M., and Katz, R.A. 2004. Genome-wide analyses of avian sarcoma virus integration sites. *J Virol* **78**(21): 11656-11663.

Nekrutenko, A. and Li, W.H. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* **17**(11): 619-621.

Nolte, R.T., Conlin, R.M., Harrison, S.C., and Brown, R.S. 1998. Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex. *Proc Natl Acad Sci U S A* **95**(6): 2938-2943.

Ohlfest, J.R., Frandsen, J.L., Fritz, S., Lobitz, P.D., Perkinson, S.G., Clark, K.J., Nelsestuen, G., Key, N.S., McIvor, R.S., Hackett, P.B., and Largaespada, D.A. 2005. Phenotypic correction and long-term expression of factor VIII in hemophilic mice by immunotolerization and nonviral gene transfer using the Sleeping Beauty transposon system. *Blood* **105**(7): 2691-2698.

Ohno, S. 1972. So much "junk" DNA in our genome. *Brookhaven Symp Biol* **23**: 366-370.

Orgel, L.E. and Crick, F.H. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**(5757): 604-607.

Papworth, M., Kolasinska, P., and Minczuk, M. 2006. Designer zinc-finger proteins and their applications. *Gene* **366**(1): 27-38.

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

-. 1993. Crystal structure of a five-finger GLI-DNA complex:new perspectives on zinc fingers. *Science* **261**(5129):1701-1707

Perez, E.E., Wang, J., Miller, J.C., Jouvenot, Y., Kim, K.A., Liu, O., Wang, N., Lee, G., Bartsevich, V.V., Lee, Y.L., Guschin, D.Y., Rupniewski, I., Waite, A.J., Carpenito, C., Carroll, R.G., Orange, J.S., Urnov, F.D., Rebar, E.J., Ando, D., Gregory, P.D., Riley, J.L., Holmes, M.C., and June, C.H. 2008. Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat Biotechnol* **26**(7): 808-816.

Peters, J.E. and Craig, N.L. 2001. Tn7: smarter than we thought. *Nat Rev Mol Cell Biol* **2**(11): 806-814.

Plasterk, R.H., Van Luenen, H., Vos, C., Ivics, Z., Iszvak, Z., and Fischer, S. 1998. Transposable elements as tools from mutagenesis and transgenesis of vertebrates. *Pathol Biol (Paris)* **46**(9): 674-675.

Porteus, M.H. and Baltimore, D. 2003. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**(5620): 763.

Ramirez, C.L., Foley, J.E., Wright, D.A., Muller-Lerch, F., Rahman, S.H., Cornu, T.I., Winfrey, R.J., Sander, J.D., Fu, F., Townsend, J.A., Cathomen, T., Voytas, D.F., and Joung, J.K. 2008. Unexpected failure rates for modular assembly of engineered zinc fingers. *Nat Methods* **5**(5): 374-375.

Raymond, C.S. and Soriano, P. 2007. High-efficiency FLP and PhiC31 site-specific recombination in mammalian cells. *PLoS One* **2**(1): e162.

Rebar, E.J., and Pabo, C.O. 1994. Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science* **263**(5147):671-673

Recchia, A., Perani, L., Sartori, D., Olgiati, C., and Mavilio, F. 2004. Site-specific integration of functional transgenes into the human genome by adeno/AAV hybrid vectors. *Mol Ther* **10**(4): 660-670.

Roberts, A.P., Chandler, M., Courvalin, P., Guedon, G., Mullany, P., Pembroke, T., Rood, J.I., Smith, C.J., Summers, A.O., Tsuda, M., and Berg, D.E. 2008. Revised nomenclature for transposable genetic elements. *Plasmid* **60**(3): 167-173.

Robinson, C.R. and Sauer, R.T. 1998. Optimizing the stability of single-chain proteins by linker length and composition mutagenesis. *Proc Natl Acad Sci U S A* **95**(11): 5929-5934.

Sambrook, J. and Russel, D. 2001. *Molecular Cloning*. Cold Spring Harbor Laboratory Press.

Sander, J.D., Zaback, P., Joung, J.K., Voytas, D.F., and Dobbs, D. 2007. Zinc Finger Targeter (ZiFiT): an engineered zinc finger/target site design tool. *Nucleic Acids Res* **35**(Web Server issue): W599-605.

Sarkar, I., Hauber, I., Hauber, J., and Buchholz, F. 2007. HIV-1 proviral DNA excision using an evolved recombinase. *Science* **316**(5833): 1912-1915.

Sauer, B. and McDermott, J. 2004. DNA recombination with a heterospecific Cre homolog identified from comparison of the pac-c1 regions of P1-related phages. *Nucleic Acids Res* **32**(20): 6086-6095.

Schmidt, E.E., Taylor, D.S., Prigge, J.R., Barnett, S., and Capecchi, M.R. 2000. Illegitimate Cre-dependent chromosome rearrangements in transgenic mouse spermatids. *Proc Natl Acad Sci U S A* **97**(25): 13702-13707.

Schroder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, J.R., and Bushman, F. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**(4): 521-529.

Segal, D.J. and Barbas, C.F., 3rd. 2000. Design of novel sequence-specific DNA-binding proteins. *Curr Opin Chem Biol* **4**(1): 34-39.

Segal, D.J., Beerli, R.R., Blancafort, P., Dreier, B., Effertz, K., Huber, A., Koksch, B., Lund, C.V., Magnenat, L., Valente, D., and Barbas, C.F., 3rd. 2003a. Evaluation of a modular strategy for the construction of novel polydactyl zinc finger DNA-binding proteins. *Biochemistry* **42**(7): 2137-2148.

Segal, D.J., Dreier, B., Beerli, R.R., and Barbas, C.F., 3rd. 1999. Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. *Proc Natl Acad Sci U S A* **96**(6): 2758-2763.

Segal, D.J., Stege, J.T., and Barbas, C.F., 3rd. 2003b. Zinc fingers and a green thumb: manipulating gene expression in plants. *Curr Opin Plant Biol* **6**(2): 163-168.

Shi, Y. and Berg, J.M. 1995. A direct comparison of the properties of natural and designed zinc-finger proteins. *Chem Biol* **2**(2): 83-89.

Singh, H., Manuri, P.R., Olivares, S., Dara, N., Dawson, M.J., Huls, H., Hackett, P.B., Kohn, D.B., Shpall, E.J., Champlin, R.E., and Cooper, L.J. 2008. Redirecting specificity of

T-cell populations for CD19 using the Sleeping Beauty system. *Cancer Res* **68**(8): 2961-2971.

Sinzelle, L., Izsvak, Z., and Ivics, Z. 2009. Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell Mol Life Sci* **66**(6): 1073-1093.

Skowronski, J., Fanning, T.G., and Singer, M.F. 1988. Unit-length line-1 transcripts in human teratocarcinoma cells. *Mol Cell Biol* **8**(4): 1385-1397.

Song, S.U., Gerasimova, T., Kurkulos, M., Boeke, J.D., and Corces, V.G. 1994. An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus. *Genes Dev* **8**(17): 2046-2057.

Staunstrup, N.H., Moldt, B., Mates, L., Villesen, P., Jakobsen, M., Ivics, Z., Izsvak, Z., and Mikkelsen, J.G. 2009. Hybrid lentivirus-transposon vectors with a random integration profile in human cells. *Mol Ther* **17**(7): 1205-1214.

Sumiyoshi, T., Holt, N.G., Hollis, R.P., Ge, S., Cannon, P.M., Crooks, G.M., and Kohn, D.B. 2009. Stable transgene expression in primitive human CD34+ hematopoietic stem/progenitor cells, using the Sleeping Beauty transposon system. *Hum Gene Ther* **20**(12): 1607-1626.

Szabo, M., Muller, F., Kiss, J., Balduf, C., Strahle, U., and Olasz, F. 2003. Transposition and targeting of the prokaryotic mobile element IS30 in zebrafish. *FEBS Lett* **550**(1-3): 46-50.

Szczepek, M., Brondani, V., Buchel, J., Serrano, L., Segal, D.J., and Cathomen, T. 2007. Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nat Biotechnol* **25**(7): 786-793.

Szuts, D. and Bienz, M. 2000. LexA chimeras reveal the function of Drosophila Fos as a context-dependent transcriptional activator. *Proc Natl Acad Sci U S A* **97**(10): 5351-5356.

Tan, W., Dong, Z., Wilkinson, T.A., Barbas, C.F., 3rd, and Chow, S.A. 2006. Human immunodeficiency virus type 1 incorporated with fusion proteins consisting of integrase and the designed polydactyl zinc finger protein E2C can bias integration of viral DNA into a predetermined chromosomal region in human cells. *J Virol* **80**(4): 1939-1948.

Tan, W., Zhu, K., Segal, D.J., Barbas, C.F., 3rd, and Chow, S.A. 2004. Fusion proteins consisting of human immunodeficiency virus type 1 integrase and the designed polydactyl zinc finger protein E2C direct integration of viral DNA into specific sites. *J Virol* **78**(3): 1301-1313.

Tate, P.H. and Bird, A.P. 1993. Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr Opin Genet Dev* **3**(2): 226-231.

Thibault, S.T., Singer, M.A., Miyazaki, W.Y., Milash, B., Dompe, N.A., Singh, C.M., Buchholz, R., Demsky, M., Fawcett, R., Francis-Lang, H.L., Ryner, L., Cheung, L.M., Chong, A., Erickson, C., Fisher, W.W., Greer, K., Hartouni, S.R., Howie, E., Jakkula, L., Joo, D., Killpack, K., Laufer, A., Mazzotta, J., Smith, R.D., Stevens, L.M., Stuber, C., Tan, L.R., Ventura, R., Woo, A., Zakrajsek, I., Zhao, L., Chen, F., Swimmer, C., Kopczynski, C., Duyk, G., Winberg, M.L., and Margolis, J. 2004. A complementary transposon tool kit for Drosophila melanogaster using P and piggyBac. *Nat Genet* **36**(3): 283-287.

Thorpe, H.M. and Smith, M.C. 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.

Thukral, S.K., Morrison, M.L., and Young, E.T. 1992. Mutations in the zinc fingers of ADR1 that change the specificity of DNA binding and transactivation. *Mol Cell Biol* **12**(6): 2784-2792.

Thyagarajan, B., Guimaraes, M.J., Groth, A.C., and Calos, M.P. 2000. Mammalian genomes contain active recombinase recognition sites. *Gene* **244**(1-2): 47-54.

Thyagarajan, B., Olivares, E.C., Hollis, R.P., Ginsburg, D.S., and Calos, M.P. 2001. Site-specific genomic integration in mammalian cells mediated by phage phiC31 integrase. *Mol Cell Biol* **21**(12): 3926-3934.

Ton-Hoang, B., Polard, P., and Chandler, M. 1998. Efficient transposition of IS911 circles in vitro. *Embo J* **17**(4): 1169-1181.

Tsai, R.Y. and Reed, R.R. 1998. Identification of DNA recognition sequences and protein interaction domains of the multiple-Zn-finger protein Roaz. *Mol Cell Biol* **18**(11): 6447-6456.

Urnov, F.D., Miller, J.C., Lee, Y.L., Beausejour, C.M., Rock, J.M., Augustus, S., Jamieson, A.C., Porteus, M.H., Gregory, P.D., and Holmes, M.C. 2005. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**(7042): 646-651.

Vigdal, T.J., Kaufman, C.D., Izsvak, Z., Voytas, D.F., and Ivics, Z. 2002. Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements. *J Mol Biol* **323**(3): 441-452.

Vink, C.A., Gaspar, H.B., Gabriel, R., Schmidt, M., McIvor, R.S., Thrasher, A.J., and Qasim, W. 2009. Sleeping beauty transposition from nonintegrating lentivirus. *Mol Ther* **17**(7): 1197-1204.

Vos, J.C., and Plasterk, R.H. 1994. Tc1 transposase of Caenorhabditis elegans is an endonuclease with a bipartite DNa binding domain. *EMBO J* **13**(24):6125-6132

Vos, J.C., van Luenen, H.G., and Plasterk, R.H. 1993. Characterization of the Caenorhabditis elegans Tc1 transposase in vivo and in vitro. *Genes Dev* **7**(7A): 1244-1253

Walisko, O., Schorn, A., Rolfs, F., Devaraj, A., Miskey, C., Izsvak, Z., and Ivics, Z. 2008. Transcriptional activities of the Sleeping Beauty transposon and shielding its genetic cargo with insulators. *Mol Ther* **16**(2): 359-369.

Wang, X., Sarkar, D.P., Mani, P., Steer, C.J., Chen, Y., Guha, C., Chandrasekhar, V., Chaudhuri, A., Roy-Chowdhury, N., Kren, B.T., and Roy-Chowdhury, J. 2009. Long-term reduction of jaundice in Gunn rats by nonviral liver-targeted delivery of Sleeping Beauty transposon. *Hepatology* **50**(3): 815-824.

Webster, N., Jin, J.R., Green, S., Hollis, M., and Chambon, P. 1988. The yeast UASG is a transcriptional enhancer in human HeLa cells in the presence of the GAL4 trans-activator. *Cell* **52**(2): 169-178.

Wei, W., Gilbert, N., Ooi, S.L., Lawler, J.F., Ostertag, E.M., Kazazian, H.H., Boeke, J.D., and Moran, J.V. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21**(4):1429-1439

Wei, Y., Ying, D., Hou, C., Cui, X., and Zhu, C. 2008. Design of a zinc finger protein binding a sequence upstream of the A20 gene. *BMC Biotechnol* **8**: 28.

Williams, D.A. 2008. Sleeping beauty vector system moves toward human trials in the United States. *Mol Ther* **16**(9): 1515-1516.

Wilson, M.H., Kaminski, J.M., and George, A.L., Jr. 2005. Functional zinc finger/sleeping beauty transposase chimeras exhibit attenuated overproduction inhibition. *FEBS Lett* **579**(27): 6205-6209.

Winckler, T., Dingermann, T., and Glockner, G. 2002. Dictyostelium mobile elements: strategies to amplify in a compact genome. *Cell Mol Life Sci* **59**(12): 2097-2111.

Wu, H., Yang, W.P., and Barbas, C.F., 3rd. 1995. Building zinc fingers by selection: toward a therapeutic application. *Proc Natl Acad Sci U S A* **92**(2): 344-348.

Wu, S.C., Meir, Y.J., Coates, C.J., Handler, A.M., Pelczar, P., Moisyadi, S., and Kaminski, J.M. 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.

Wu, X., Li, Y., Crise, B., and Burgess, S.M. 2003. Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**(5626): 1749-1751.

Xie, W., Gai, X., Zhu, Y., Zappulla, D.C., Sternglanz, R., and Voytas, D.F. 2001. Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. *Mol Cell Biol* **21**(19): 6606-6614.

Xue, X., Huang, X., Nodland, S.E., Mates, L., Ma, L., Izsvak, Z., Ivics, Z., LeBien, T.W., McIvor, R.S., Wagner, J.E., and Zhou, X. 2009. Stable gene transfer and expression in cord blood-derived CD34+ hematopoietic stem and progenitor cells by a hyperactive Sleeping Beauty transposon system. *Blood* **114**(7): 1319-1330.

Yant, S.R., Ehrhardt, A., Mikkelsen, J.G., Meuse, L., Pham, T., and Kay, M.A. 2002. Transposition from a gutless adeno-transposon vector stabilizes transgene expression in vivo. *Nat Biotechnol* **20**(10): 999-1005.

Yant, S.R., Huang, Y., Akache, B., and Kay, M.A. 2007. Site-directed transposon integration in human cells. *Nucleic Acids Res* **35**(7): e50.

Yant, S.R., Meuse, L., Chiu, W., Ivics, Z., Izsvak, Z., and Kay, M.A. 2000. Somatic integration and long-term transgene expression in normal and haemophilic mice using a DNA transposon system. *Nat Genet* **25**(1): 35-41.

Yant, S.R., Park, J., Huang, Y., Mikkelsen, J.G., and Kay, M.A. 2004. Mutational analysis of the N-terminal DNA-binding domain of sleeping beauty transposase: critical residues for DNA binding and hyperactivity in mammalian cells. *Mol Cell Biol* **24**(20): 9239-9247.

Yant, S.R., Wu, X., Huang, Y., Garrison, B., Burgess, S.M., and Kay, M.A. 2005. High-resolution genome-wide mapping of transposon integration in mammals. *Mol Cell Biol* **25**(6): 2085-2094.

Yieh, L., Hatzis, H., Kassavetis, G., and Sandmeyer, S.B. 2002. Mutational analysis of the transcription factor IIIB-DNA target of Ty3 retroelement integration. *J Biol Chem* **277**(29): 25920-25928.

Young, S.M., Jr. and Samulski, R.J. 2001. Adeno-associated virus (AAV) site-specific recombination does not require a Rep-dependent origin of replication within the AAV terminal repeat. *Proc Natl Acad Sci U S A* **98**(24): 13525-13530.

Yu, Y. and Bradley, A. 2001. Engineering chromosomal rearrangements in mice. *Nat Rev Genet* **2**(10): 780-790.

Zayed, H., Izsvak, Z., Walisko, O., and Ivics, Z. 2004. Development of hyperactive sleeping beauty transposon vectors by mutational analysis. *Mol Ther* **9**(2): 292-304.

Zhou, B.P., Hu, M.C., Miller, S.A., Yu, Z., Xia, W., Lin, S.Y., and Hung, M.C. 2000. HER-2/neu blocks tumor necrosis factor-induced apoptosis via the Akt/NF-kappaB pathway. *J Biol Chem* **275**(11): 8027-8031

Zhu, Y., Dai, J., Fuerst, P.G., and Voytas, D.F. 2003. Controlling integration specificity of a yeast retrotransposon. *Proc Natl Acad Sci U S A* **100**(10): 5891-5895.

Zou, S., Ke, N., Kim, J.M., and Voytas, D.F. 1996. The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. *Genes Dev* **10**(5): 634-645.

Zweidler-Mckay, P.A., Grimes, H.L., Flubacher, M.M., and Tsichlis, P.N. 1996. Gfi-1 encodes a nuclear zinc finger protein that binds DNA and functions as a transcriptional repressor. *Mol Cell Biol* **16**(8): 4024-4034.

6. Abbreviations

aa	amino acid
AAV	adeno-associated virus
Ac	*Activator*
AD	activation domain
amp	ampicillin
APC	adenomatosis polyposis coli
ASLV	avian sarcoma-leucosis virus
ASV	avian sarcoma virus
BLAST	basic local alignment tool
bp	base pair
BS	binding site
cam	chloramphenicol
CAR	coxsackievirus adenovirus receptor
CAST	cyclic amplification and selection of targets
CD	cluster of differentiation
cDNA	complementary desoxyribonucleic acid
CENP-B	centromere protein B
DBD	DNA binding domain
DSB	double strand break
DNA	desoxyribonucleic acid
DR	direct repeat
Ds	*Dissociator*
ds	double strand
DSB	double strand break
DR	direct repeat
DTT	Dithiothreitol
EBFP	enhanced blue fluorescent protein
e.g.	exempli gratia
EGFP	enhanced green fluorescent protein
EN	endonuclease
env	envelope

FCS	fetal calf serum
FRET	fluorescence resonance transfer
gag	group-specific antigen
glmS	glutamate synthetase gene
GOI	gene of interest
GRCh37	genome reference consortium human 37
h	hour
hAT	hobo Ac Tam3
HeLa	Henrietta Lacks
HIV	human immunodeficiency virus
IL	interleukin
IN	integrase
Inh	inhibitor
IR	inverted repeat
IS	insertion sequence
ITAM	immunoreceptor tyrosin-based activation motif
kan	kanamycin
l	liter
LEDGF	lens epithelium-derived growth factor
LMO2	LIM domain only 2
LINE	long interspersed nuclear element
L-IR	left inverted repeat
LTR	long terminal repeat
min	minutes
MLV	murine leukemia virus
mRNA	messenger ribonucleic acid
NLS	nuclear localization signal
ORF	open reading frame
ori	origin of replication
pA	polyadenylation sequence
PB	piggyBac
PBMC	peripheric blood mononuclear cell
PEI	polyethylenimin
pol	polymerase

PR	protease
Prots	proteins
RAG	recombination-activating gene
rep	replication protein
R-IR	right inverted repeat
RLU	relative luminescence unit
RNA	ribonucleic acid
RT	reverse transcriptase
SAF	scaffold attachment factor
SB	Sleeping Beauty
SCID-X	X-linked severe combined immunodeficiency
sec	seconds
sem	standard error of the mean
SINE	short interspersed nuclear element
S/MAR	scaffold/matrix attachment regions
Ta	transcriptionally active
TCR	T-cell receptor
TE	transposable element
TetR	tetracycline repressor protein
TF	transcription factor
Tnpon	transposon
TRE^1	tRNA gene-targeting retrotransposable elements
TRE^2	tetracycline respone element
tss	transcription start sites
UAS	upstream activating sequence
UTR	untranslated region
wt	wild type
ZF	zinc finger
ZFN	zinc finger nuclease
ZiFDB	zinc finger database
ZiFiT	zinc finger targeter

7. Zusammenfassung

Gentherapie stellt eine vielversprechende Alternative zur Behandlung von genetischen und erworbenen Krankheiten dar. Ziel dieser Arbeit ist es, die Möglichkeit zu untersuchen, mit Hilfe von Transposons ein therapeutisches Transgen gezielt an einer „sicheren" Stelle im menschlichen Genom einzubauen, um so eine Grundlage für eine „sicherere" Gentherapie zu schaffen.

Als Transposon wurde das *Sleeping Beauty* (*SB)* Transposon System gewählt, da es das zur Zeit am intensivsten untersuchte DNA Transposon in Vertebraten ist. Das *SB* Transposon wurde bereits bei einer Vielfalt unterschiedlicher Krankheiten in unterschiedlichen menschlichen Zelltypen als Vektor-System erfolgreich getestet. Um sicherzustellen, dass das *SB* Transposon an einer „sicheren" Stelle im menschlichen Genom inseriert, ist ein gezielter Einbau des Transposons notwendig.

In der vorliegenden Arbeit wurde untersucht, ob durch eine Fusion zwischen der enzymatischen Komponente des *SB* Transposon Systems, der Transposase, mit einer DNA-Bindedomäne ein gezielter Einbau des Transposons herbeigeführt werden kann.

Hierzu wurden zwei alternative Ansätze verfolgt, zum einem wurde die SB Transposase mit dem artifiziellen sechs-Finger Zinkfingerprotein (ZFP) E2C, dessen Bindestelle ein einziges Mal im menschlichen Genom vorkommt, fusioniert und zum anderen neue sechs-Finger ZFPe erschaffen und getestet, deren Bindestellen mehrfach im menschlichen Genom zu finden sind. Fusionsproteine aus ZFP und SB Transposase wurden auf ihre katalytischen Fähigkeiten mit Hilfe von Transpositions- und Transposon Excisionsassay getestet. Ihr Bindevermögen wurde mit Hilfe von Luciferase Reporterassays überprüft. Ihre Eignung für den gezielten Einbau von Transgenen wurde zunächst auf Plasmidebene für E2C/SB Fusionsproteine unter Anwendung eines sogenannten Plasmid-zu-Plasmidassays untersucht. Gezielter Einbau im menschlichen Genom wurde mit Hilfe von locus-spezifischer PCR für E2C/SB Fusionsproteine und Southern Blot und LAM-PCR für beide Ansätze überpüft. Alle Untersuchungen wurden in HeLa Zellen durchgeführt.

Für die E2C/SB Fusionsproteine konnte eine veränderte Verteilung der Transposonsinsertionen auf Plasmidebene gezeigt werden, jedoch nicht im Kontext des menschlichen Genoms.

Die drei, in dem zweiten Ansatz getesteten, sechs-Finger ZFPe wurden mit Hilfe der „Zinc finger tools" Internetseite erschaffen. Ihre prognostizierten Bindedomänen lagen in der 3´-Region der Ta Unterfamilie von LINE1 Elementen im menschlichen Genom. Zwei der drei

ZFPe (A und B) banden ihre prognostizierte Bindedomäne in Luciferase Reporterassays vierfach besser als entsprechende Negativkontrollen. Mit Hilfe eines Fusionsproteins aus ZFP B und der SB Transposase konnte eine Anreicherung von *SB* Transposon Insertionen in derNähe der ZFP B Bindedomäne im menschlichen Genom nachgewiesen werden.

8. Summary

Gene therapy is a promising alternative for the treatment of genetic as well as acquired diseases. Goal of my work was to create a non-viral transposon-based gene delivery vector that could target transposon insertions to “safe” sites in the human which is a prerequisite for “safer” gene therapy.

As a transposon tool I chose to use the *Sleeping Beauty* (*SB*) System since it is the most profound studied vertebrate DNA transposon to date. The *SB* transposon system has been applied as delivery tool for therapeutic transgenes for a variety of diseases in a variety of human cell types. To ensure insertion of the transposon into a “safe” site in the human genome, some form of targeting would be desirable.

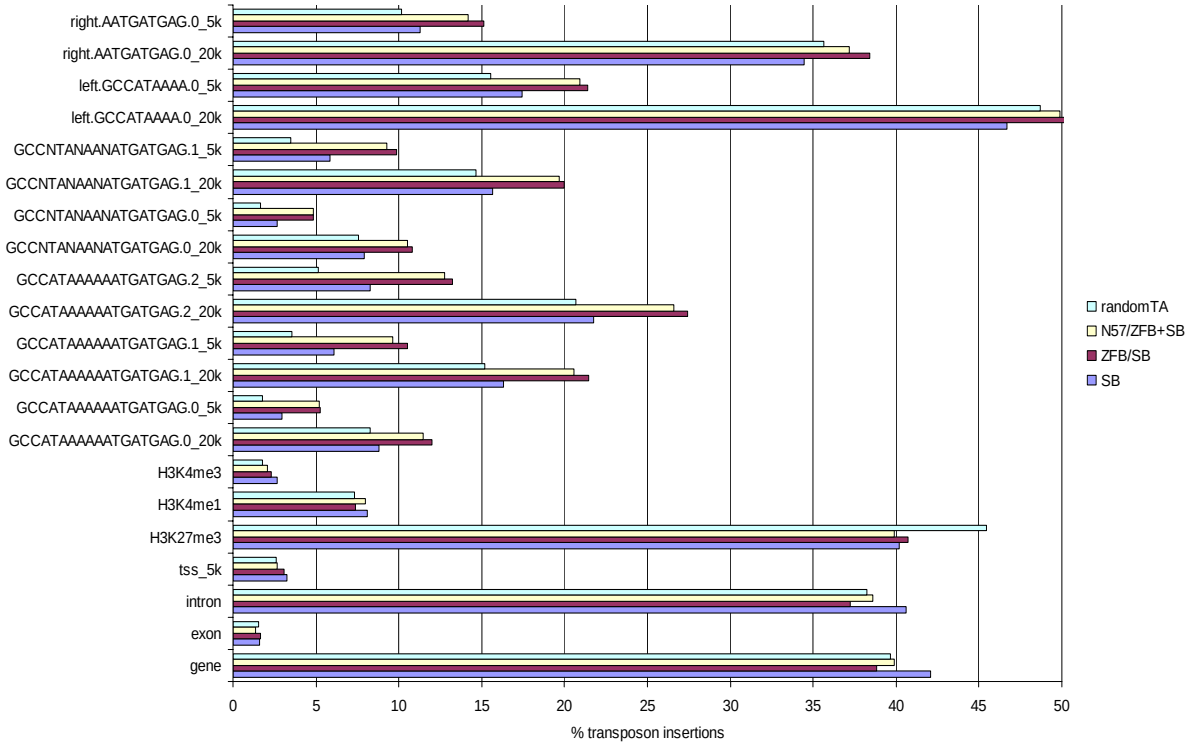
In this work I examined if targeted transposon insertions can be achieved using fusion proteins of a DNA binding domain and the catalytic component of the *SB* transposon, the transposase.

For this two alternative approaches were pursued. First the SB transposase was fused to the articifial six-finger zinc finger (ZF) protein E2C, which binds a single unique site in the human genome. Secondly new six-finger ZF proteins were designed who´s binding domain exist in multiple copies throughout the human genome. Fusion proteins of ZF proteins and SB transposase were tested for their catalytic activity in cell culture transposition and transposon excision PCR assays. Their ability to bind their predicted target sites was examined using luciferase reporter assays. Eligibility for targeted transposon insertion was tested on plasmid level for E2C/SB fusion proteins using a so-called plasmid-to-plasmid assay. Eligibility for targeted transposon insertion in the human genome was tested using site locus-specific PCR for the E2C/SB fusion proteins and Southern blot and LAM-PCR for both approaches. All experiments were done in HeLa cells.

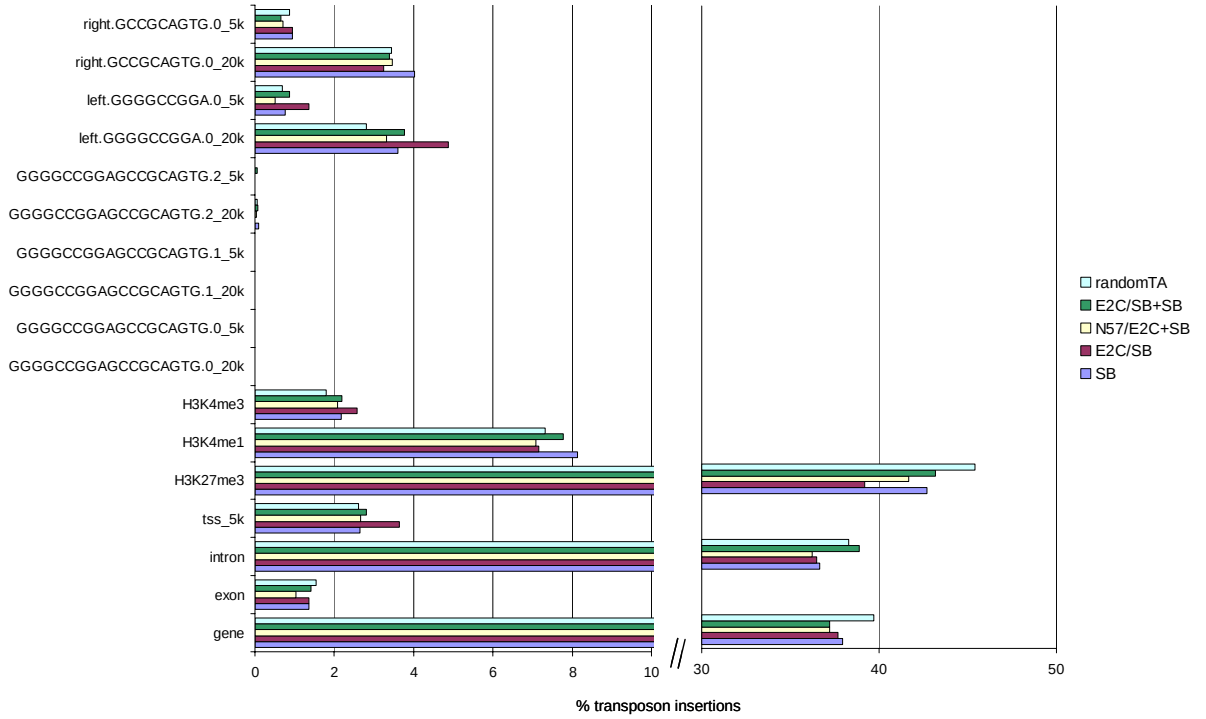
Targeting the E2C binding site with the E2C/SB transposase showed some success in plasmid context, however failed in genomic context.

In the second approach three six-finger ZF proteins were designed using the “Zinc finger tools” website. Their predicted ZF binding sites lay in the 3´-region of a member of the Ta-subfamily of LINE1 elements. Two of the three ZF proteins (ZF A and ZF B) bound their predicted target site fourfold better than negative controls respectively. Using ZF B/SB transposase fusion proteins an enrichment of transposon insertions near the ZF B binding site could be observed.

9. Supplementary data



supplementary Figure 1. Transposon insertions into the human genome mediated by ZF B/SB fusion proteins, unfused SB and calculations for random insertion (randomTA). Bars represent percentage of transposon insertions within a defined genomic region or within a defined window around a specified DNA sequence (y-axis). For example: GCCATAAAAAATGATGAG represents the predicted ZF B binding site, the appendage .0, .1 or .2 stands for the number of mismatched allowed within this binding site, the appendage _5k stand for a 5 kb window around the transposon insertion site that was screened for ZF B binding sites. Apart from the correct full length ZF B binding site DNA near transposon insertions was also screened for ZF B binding sites of the type GCCNTANAANATGATGAG because of loose 5´-A recognition of ZF proteins. Occurrence of ZF B binding half sites was also checked. Number of transposon insertion sites analyzed varied from 6,000 to 14,000 insertions per transposase.



supplementary Figure 2. Transposon insertions into the human genome mediated by E2C/SB fusion proteins, unfused SB and calculations for random insertion (randomTA). Bars represent percentage of transposon insertions within a defined genomic region or within a defined window around a specified DNA sequence (y-axis). For example: GGGGCCGGAGCAGTG represents the predicted E2C binding site the appendage .0, .1 or .2 stands for the number of mismatched allowed within this binding site. The appendage _5k stands for a 5 kb window around the transposon insertion site that was screened for E2C binding sites. Occurrence of E2C binding half sites was also checked. Number of transposon insertion sites analyzed varied from 750 to 8,000 insertions per transposase.

supplementary Table 1. Distribution of SB transposon insertions within the human genome. (A) Enlisted are the results for SB transposon insertion into the human genome using unfused SB transposase (SB), different E2C/SB fusion proteins (E2C/SB, N57/E2C+SB, E2C/SB+SB) and a calculated data set for random integration at any TA dinucleotide in the respective region of the human genome (randomTA). Only those transposon insertions giving a unique hit in the human genome were considered in this data set. The column named "total" comprises the total number of SB transposon insertions considered for analysis. the column "tss 5k" comprises transpson insertions within a 5 kb window around transcription start sites, the column GGGGCCGGAGCCGCAGTG.0_20k comprises transposon insertions within a 20 kb window around a GGGGCCGGAGCCGCAGTG motif allowing zero (other columns are labelled accordingly), the column left.GGGGCCGGA.0 20k comprises transposon insertions within 20 kb of the left E2C binding half site 5´-GGGGCCGGA -3´allowing no mismatch. Stars (*) represent p-Values for differences to the SB data set. * indicates a p-value of ≤ 0.05, ** a p-value of ≤ 0.01 and *** a p-value of ≤ 0.001. **(B)** Distribution of SB transposon insertions that map to unique sites within the human genome for ZF B containing SB transposase constructs. Columns are labelled accordingly to (A). Only those transposon insertions giving a unique hit in the human genome were considered in this data set. **(C)** Transposon insertions into repetitive regions of the human genome for ZF B/fusion proteins and controls. Transposon insertions of this data set could be mapped to multiple sites within the human genome and were not considered in the data set for (B).

A

	total	gene	exon	intron	tss_5k	H3K27me3	H3K4me1	H3K4me3
SB	8314 (100.0%)	3154 (37.9%)	113 (1.36%)	3047 (36.6%)	220 (2.65%)	3551 (42.7%)	677 (8.14%)	180 (2.17%)
E2C/SB	740 (100.0%)	279 (37.7%)	10 (1.35%)	270 (36.5%)	3187 (2.6%)	290 (39.2%)	53 (7.16%)	19 (2.57%)
N57/E2C+SB	4544 (100.0%)	1691 (37.2%)	47 (1.03%)	1646 (36.2%)	27 (3.65%)	1893 (41.7%)	322 (7.09%)*	95 (2.09%)
E2C/SB+SB	6057 (100.0%)	2433 (40.2%)**	86 (1.42%)	2354 (38.9%)**	121 (2.66%)	2616 (43.2%)	471 (7.78%)	133 (2.2%)
randomTA	122589 (100.0%)	48648 (39.7%)**	1879 (1.53%)	46906 (38.3%)**	170 (2.81%)	55685 (45.4%)***	8967 (7.31%)**	2188 (1.78%)*

	GGGGCCGGAGCCGCAGTG.0 20k	GGGGCCGGAGCCGCAGTG.0 5k	GGGGCCGGAGCCGCAGTG.1 20k	GGGGCCGGAGCCGCAGTG.1 5k
SB	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
E2C/SB	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
N57/E2C+SB	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
E2C/SB+SB	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
randomTA	2 (0.00163%)	1 (0.000816%)	3 (0.00245%)	1 (0.000816%)

	GGGGCCGGAGCCGCAGTG.2 20k	GGGGCCGGAGCCGCAGTG.2 5k	left.GGGGCCGGA.0 20k	left.GGGGCCGGA.0 5k
SB	8 (0.0962%)	1 (0.012%)	300 (3.61%)	64 (0.77%)
E2C/SB	0 (0.0%)	0 (0.0%)	36 (4.86%)	10 (1.35%)
N57/E2C+SB	2 (0.044%)	0 (0.0%)	151 (3.32%)	23 (0.506%)
E2C/SB+SB	4 (0.066%)	3 (0.0495%)	228 (3.76%)	53 (0.875%)
randomTA	64 (0.0522%)	15 (0.0122%)	3445 (2.81%)***	838 (0.684%)

	right.GCCGCAGTG.0 20k	right.GCCGCAGTG.0 5k
SB	335 (4.03%)	78 (0.938%)
E2C/SB	24 (3.24%)	7 (0.946%)
N57/E2C+SB	157 (3.46%)	32 (0.704%)
E2C/SB+SB	205 (3.38%)*	40 (0.66%)
randomTA	4229 (3.45%)**	1062 (0.866%)

B

	total	exon	gene	intron	tss_5k	H3K27me3	H3K4me1	H3K4me3
SB	14178 (100.0%)	225 (1.59%)	5967 (42.1%)	5758 (40.6%)	458 (3.23%)	5698 (40.2%)	1152 (8.13%)	378 (2.67%)
ZFB/SB	5991 (100.0%)	100 (1.67%)	2325 (38.8%)***	2232 (37.3%)***	183 (3.05%)	2441 (40.7%)	444 (7.41%)	138 (2.3%)
N57/ZFB+SB	10081 (100.0%)	137 (1.36%)	4019 (39.9%)***	3890 (38.6%)**	271 (2.69%)*	4023 (39.9%)	807 (8.01%)	210 (2.08%)**
randomTA	122589 (100.0%)	1879 (1.53%)	48648 (39.7%)***	46906 (38.3%)***	3187 (2.6%)***	55685 (45.4%)***	8967 (7.31%)***	2188 (1.78%)***

	GCCATAAAAAATGATGAG.0_20k	GCCATAAAAAATGATGAG.0_5k	GCCATAAAAAATGATGAG.1_20k	GCCATAAAAAATGATGAG.1_5k
SB	1245 (8.78%)	420 (2.96%)	2316 (16.3%)	862 (6.08%)
ZFB/SB	718 (12.0%)***	315 (5.26%)***	1286 (21.5%)***	630 (10.5%)***
N57/ZFB+SB	1155 (11.5%)***	527 (5.23%)***	2071 (20.5%)***	971 (9.63%)***
randomTA	10174 (8.3%)	2201 (1.8%)***	18656 (15.2%)***	4358 (3.55%)***

	GCCATAAAAAATGATGAG.2_20k	GCCATAAAAAATGATGAG.2_5k	GCCNTANAANATGATGAG.0_20k	GCCNTANAANATGATGAG.0_5k
SB	3087 (21.8%)	1174 (8.28%)	1289 (9.09%)	439 (3.1%)
ZFB/SB	1643 (27.4%)***	793 (13.2%)***	752 (12.6%)***	331 (5.52%)***
N57/ZFB+SB	2682 (26.6%)***	1288 (12.8%)***	1185 (11.8%)***	542 (5.38%)***
randomTA	25326 (20.7%)**	6307 (5.14%)***	10518 (8.58%)*	2260 (1.84%)***

	GCCNTANAANATGATGAG.1_20k	GCCNTANAANATGATGAG.1_5k	left.GCCATAAAA.0_20k	left.GCCATAAAA.0_5k
SB	2408 (17.0%)	910 (6.42%)	6622 (46.7%)	2468 (17.4%)
ZFB/SB	1319 (22.0%)***	644 (10.7%)***	3032 (50.6%)***	1281 (21.4%)***
N57/ZFB+SB	2147 (21.3%)***	1007 (9.99%)***	5031 (49.9%)***	2109 (20.9%)***
randomTA	19439 (15.9%)***	4559 (3.72%)***	59735 (48.7%)***	19069 (15.6%)***

	right.AATGATGAG.0_20k	right.AATGATGAG.0_5k
SB	4888 (34.5%)	1604 (11.3%)
ZFB/SB	2302 (38.4%)***	906 (15.1%)***
N57/ZFB+SB	3746 (37.2%)***	1430 (14.2%)***
randomTA	43677 (35.6%)**	12453 (10.2%)***

C

	DNA	DNA_hAT	DNA_PiggyBac	DNA_TcMar-Mariner	LINE_L1	LINE_L2	LINE	SINE
SB	433 (3.66%)	5 (0.0423%)	2 (0.0169%)	7 (0.0592%)	2793 (23.6%)	485 (4.1%)	3337 (28.2%)	926 (7.83%)
ZFB/SB	169 (3.32%)	2 (0.0393%)	0 (0.0%)	0 (0.0%)	1329 (26.1%)***	175 (3.44%)*	1533 (30.1%)*	395 (7.76%)
N57/ZFB+SB	302 (3.53%)	8 (0.0935%)	0 (0.0%)	5 (0.0585%)	2308 (27.0%)***	288 (3.37%)**	2632 (30.8%)***	754 (8.82%)*
randomTA	542 (4.14%)	18 (0.137%)*	2 (0.0153%)	12 (0.0916%)	2751 (21.0%)***	421 (3.21%)***	3262 (24.9%)***	1041 (7.94%)

	LTR_ERVK	LTR_Gypsy	LTR	RC_Helitron	rRNA_rRNA
SB	46 (0.389%)	1 (0.00846%)	856 (7.24%)	1 (0.00846%)	4 (0.0338%)
ZFB/SB	16 (0.314%)	1 (0.0197%)	445 (8.74%)***	0 (0.0%)	3 (0.059%)
N57/ZFB+SB	43 (0.503%)	1 (0.0117%)	661 (7.73%)	1 (0.0117%)	1 (0.0117%)
randomTA	43 (0.328%)	9 (0.0687%)*	1009 (7.7%)	1 (0.00763%)	0 (0.0%)

10. Acknowledgements

I would like to thank Dr. Zoltán Ivics for giving me the opportunity to work in his group on my favourite project. Zoltán was always available for motivating fruitful discussion and left room for my own ideas and self-reliant work.

I want to thank Csaba Miskey for sharing his abyss-deep transposon and lab knowledge that he generously shared with me especially for LAM PCR procedures.

I very much like to thank all the current and former lab members that made working so much enjoyable, especially Andrea Schorn, Andrea Schmitt and Tobias Jursch.

I want to thank Prof. Udo Heinemann and Prof. Reinhard Kunze very much for being my evaluators at the Freie Universität Berlin.

I also like to thank Andreas Gogol-Doehring for analysing gigabytes of Illumina sequencing data for me.