Advanced Genome Engineering Using CRISPR/Cas9 and Transposases

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

vorgelegt beim Fachbereich 14

der Johann Wolfgang Goethe-Universität

in Frankfurt am Main

von

Adrian Kovač

aus Frankfurt am Main

Frankfurt, 2020

Vom Fachbereich 14 der

Johann Wolfgang Goethe – Universität als Dissertation angenommen.

Dekan: Prof. Dr. Clemens Glaubitz

Gutachter: Prof. Dr. Rolf Marschalek

Datum der Disputation: 22.10.2020

Table of Contents

1 Introduction

1.1 Overview over gene therapy history and technology

Genetic disorders have affected the lives of humans for the entire span of human history and some of them are among the most debilitating diseases affecting our species. Additionally, heritable disorders are passed on across generations. The great challenge of treating genetic disorders is that, until recently, there was no way to correct the underlying defects that cause them. In contrast to, for example, infectious diseases, it was not possible to cure these disorders; rather, treating them meant combating their symptoms. This has only changed within the last decades with the advent of gene therapy. Gene therapy as a term covers different modifications of cells with nucleic acids, but in many cases it means directly modifying the genome of a target cell. This can either take the form of inserting genetic information into genomes – sometimes called augmentation gene therapy – or making other modifications to the genome, like gene knockouts or *in situ* correction of mutations[1].

The potential benefit of gene therapy was already outlined in 1972[2], however the authors urged to proceed with caution in this field of research. It took another 18 years for the first gene therapy trial to be approved and carried out in 1990[3], treating a patient with the immunodeficiency ADA-SCID. The first stable modification of a target cell followed three years later[4], but the field suffered a setback with the death of a patient from vector toxicity in 1999[5]. Since then successes in the field have restored some of the initial trust and enthusiasm and therapies for a wide range of diseases have been tested in pre-clinical or clinical studies so far, including Alzheimer's disease[6], blindness[7], cancers[8,9], cystic fibrosis[10], hemophilia[11], HIV infection[12,13], Huntington's disease[14], Muscular Dystrophy[15], Parkinson's disease[16] and immunodeficiencies[17]. In sum, more than 2500 gene therapy trials had been initiated by the end of 2018[1]. Methods used in these interventions include introduction of corrected copies of defective genes, introduction of engineered genes with novel functions (e.g. CAR-T cells) and other modifications. The first commercial gene therapy treatments have been approved in recent years; the first drug approved for the European market was Glybera[18] in 2012 and the first drug approved in the US was Luxturna, in 2017. As of September 2019, 14 gene therapy products were approved in the US, 8 in Europe and 4 in other countries[19].

In general, several strategies are available to introduce genetic material into target cells. Some methods simply allow introduction of DNA into cells or their nuclei, including physical methods like lipid-based transfection or electroporation. However, with these methods, DNA

is introduced into the target cells once and then lost over time. This would necessitate any gene therapy product to be administered repeatedly, especially if the target is a dividing cell population. Repetitive administration is problematic in a clinical context, both due to practical and financial reasons and due to the possibility of immune responses. This problem is avoided by a second group of vectors, called integrating vectors, which stably insert the genetic cargo into the genome of the target cells. The transgene is thus replicated with the genome of the cell during the cell cycle. This makes it possible for such gene therapy products to be effective after a single use. Yet another kind of gene therapy technology is based on re-writing genomic information in the target genome, rather than inserting a transgene. This approach, referred to as gene editing, has a unique set of advantages and drawbacks when compared to integrating vectors.

1.2 Integrating vectors

1.2.1 Viral vectors

The most widely used group of integrating vectors are viral vectors, in which viral particles are used to shuttle a transgene into the cell and viral enzymes are used to integrate it into the target genome. A viral vector, like a natural virus, consists of two components, a nucleic acid and proteins. In nature, the viral genome encodes both for the structural protein component of the virus and the enzymes required in the viral life cycle. Viral proteins insert the viral DNA into the host cell's genome and the viral genes are expressed after integration. Viral particles are assembled from viral genomes and the protein components and after release are able to infect other cells.

The ability to transfer DNA across cell membranes and to integrate it into host cell genomes makes integrating viruses very attractive tools for genome engineering and gene therapy. By replacing parts of the viral genome with transgenes, it is possible to transfer these transgenes into target cells in an efficient manner while also making it impossible for the virus to proliferate.

The two most widely used integrating viruses are lentiviruses like HIV and γ-retroviruses like MLV, both of which are ssRNA(+) viruses from the family *Retroviridae*. Both viruses are capable of efficiently integrating transgenes into target genomes and can carry cargoes of up to 8 kb. While MLV is only capable of infecting dividing cells, HIV can infect cells throughout the cell cycle.[1] Their efficient integration into the genome, while being their greatest asset, can also be a liability, as will be described in section 1.5.

Two other vectors commonly used for gene therapy applications are based on adenoviruses or adeno-associated virus (AAV). In contrast to lentiviral and γ-retroviral vectors, these vectors generally do not integrate their cargo into the genome but rather persist as episomes, although some integration can occur in the case of AAV[20] (see section 1.5.1). This means that they cannot be used for targeting rapidly dividing cell populations. On the other hand, they do not depend on active division in their target cells, like MLV, or like HIV which depends on target cells being in G1 phase[1]. The therapeutic use of Adenoviruses and AAV is also limited to some extent by the strong antiviral response they can cause[21–25], which in turn often results in loss of the transgene[26].

1.2.2 Transposable elements

The second major group of integrating vectors are based on transposable elements (TEs), also called transposons. Transposons are mobile genetic elements which share some of the characteristics of integrating viruses. In their natural state, they generally also encode an enzyme which catalyzes the reaction by which the transposon can be integrated into a target genome and are marked by specific sequences at the transposon ends. The main difference compared to viruses is that transposons do not form infectious particles. This means that transposons can move within a genome, but normally are not transferred between cells or organisms.

Transposons were originally discovered in the 1940s by Barbara McClintock[27,28], who was studying the maize genome. They occur naturally in organisms from all kingdoms of life and make up a significant fraction of some genomes (e.g. 45% of the human genome[29] and 90% of the maize genome[30]), which initially was somewhat puzzling due to the fact that transposons seemed to serve no obvious purpose and were considered "junk DNA". However, since then transposons have been shown to interact with their host genomes in a complex manner.

One the one hand, transposons shape evolution of the genome. These changes can be rather destructive, like the disruption of host genomic sequences or regulatory networks that will be described in more detail in section 1.5. The highly mutagenic nature of transposons can be illustrated by the fact that more than 50% of isolated phenotypic mutants of *Drosophila melanogaster*[31] and 10-15% of those of the mouse[32] are caused by TEs. However, transposons can also lead to constructive changes by processes like capture of host sequences by the transposon[33] and exon shuffling[34]. Intronic SINEs in particular are often captured as alternative exons[35,36] and retroelements can transport flanking DNA sequences with them when they transpose[37]. Another interesting phenomenon is the accidental integration of cellular RNAs by

transposon-encoded reverse transcriptases, which is termed "retroposition"[38–40]. Transposition reactions using transposon ends from different transposon copies in one genome can result in large-scale genomic rearrangements[41,42], potentially causing inversions, duplications and deletions of over 100 kb[43,44].

Apart from being passively influenced by TEs, there are several mechanisms by which organisms prevent transposons from jumping, thus limiting the destabilization of their genomes. Transposons can be silenced by a variety of mechanisms – depending on the host organism – including chromatin remodeling, methylation, or RNAi[45–47]. However, the fact that some epigenetic marks have to be erased at certain stages of development[48] can give TEs chances to escape repression[49]. There are also sequence-specific repressor proteins dedicated to limiting TE expression, like KRAB-ZFPs[50,51]. While host cells have several mechanisms to repress TEs, in some cases transposons themselves control their copy number in order to limit disruption of their host[52,53]. These opposing trends – TEs destabilizing their host genomes, but creating novel variants in the process and genomes limiting TE activity to maintain stability – and their balance have great influence over the evolution of genomes and consequently the diversification of species[54,55].

A very interesting case of constructive interaction between transposons and hosts is the co-opting of transposon-encoded genes, or parts of these genes, by the host cell, referred to as "domestication". It is often hard to conclusively determine whether a gene is the product of a domestication process, and, depending on the exact criteria used, the number of domesticated genes reported so far varies between a few and thousands[56,57]. A prominent example of this process are the proteins RAG1 and RAG2[58], which are involved in V(D)J recombination, the process by which the adaptive immune system generates the huge number of distinct antibodies it produces. RAG1 has been shown to be closely related to transposases from the *Transib* superfamily[59] and both RAG proteins seem to have originated from the same DNA transposon[60]. A more recent domestication event led to the generation of SETMAR, a fusion protein between a SET domain and a *mariner* transposase[61]. Apart from transposon-encoded genes being repurposed by the cell, TEs also contribute to the generation of non-coding RNAs, including lncRNAs[62], miRNAs[63] and piRNAs[64].

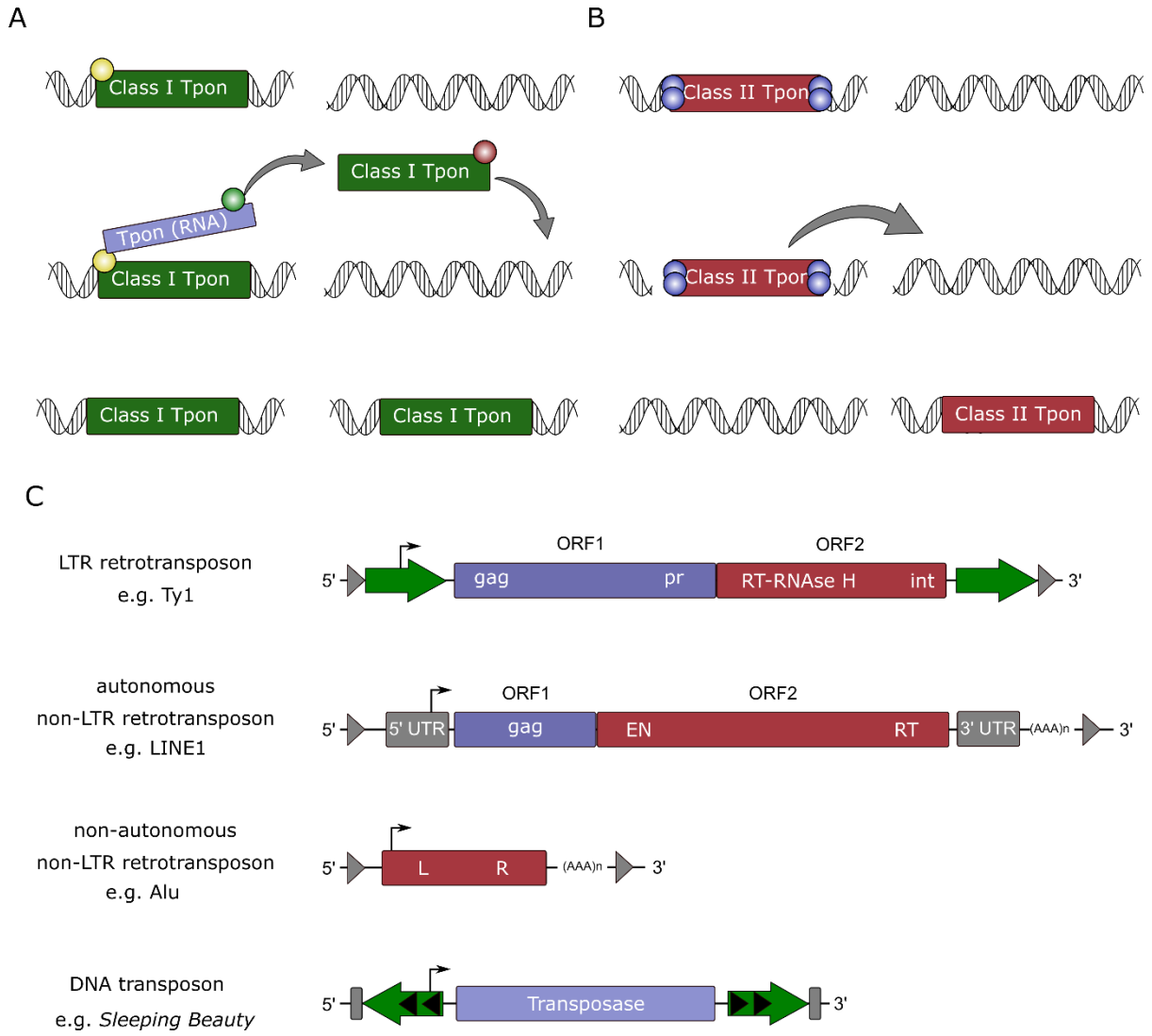


Figure 1.1 – Classification of transposable elements. A Simplified transposition mechanism of class I transposons**.** Class I transposons are transcribed into RNA by RNA polymerase (yellow sphere) and the RNA is reverse transcribed back into DNA by a reverse transcriptase (green sphere). The DNA copy of the transposon is then re-integrated into the genome at a different position by an integrase (red sphere). For non-LRT retrotransposons, a pathway called target-primed reverse transcription is utilized, meaning that the reverse transcription and integration steps are performed in conjunction at the target site. Depending on the element, more than one of these functions can be fulfilled by the same protein and in some cases the necessary proteins are provided by the host cell or by other TEs. As a result of the mechanism, the original copy of the transposon remains in the genome. **B** Simplified transposition mechanism of most class II transposons. The transposase enzymes (blue spheres) bind to the ITRs of the transposon and excise it from the genome. The enzymes remain bound to the transposon ends until they integrate the transposon into a different position in the genome in a concerted reaction. **C** Structure of the main classes of transposons. LTR retrotransposons are similar in structure to retroviruses and are flanked by long terminal repeats. Non-LTR retrotransposons lack LTRs, generally encode two ORFs and have a poly-A tail. Non-autonomous non-LTR retrotransposons lack the genes required for active transposition and need to be mobilized by the machinery from autonomous non-LTR retrotransposons. DNA transposons generally encode a single gene and are flanked by characteristic ITR structures. Gray triangles or rectangles indicate target site duplications.

There are two main classes of transposons: retrotransposons (or class I transposons, Figure 1.1A) and DNA transposons (or class II transposons, Figure 1.1B). Retrotransposons have a similar life cycle to retroviruses: they are transcribed into RNA, which is then converted back into DNA and inserted into a DNA target[65]. This means that the original copy of a

retrotransposon remains present in the genome ('copy-and-paste' mechanism or replicative transposition). Class I TEs are further divided into LTR (long terminal repeat) and non-LTR retrotransposons. LTR retrotransposons, like the Ty1, Ty3 and gypsy families, are very similar to retroviruses in structure and replicative mechanism. As is the case for retroviruses, the critical step of converting RNA into DNA is catalyzed by a reverse transcriptase[66]. Non-LTR retrotransposons comprise both LINEs and the non-autonomous SINEs (e.g. L1 and Alu elements) and replicate by a mechanism known as target-primed reverse transcription[67].

DNA transposons, on the other hand, are generally excised from the original DNA molecule and re-inserted elsewhere; the original transposon physically moves to a different location ('cut-and-paste' mechanism or conservative transposition). Both of these reactions are catalyzed by a single enzyme called a transposase, and autonomous class II transposons need an intact copy of the transposase gene to be mobile. Some DNA transposons also contain additional genes[68]. Both transposon ends of class II transposons are generally similar, but in opposite orientation, which is why they are referred to as inverted terminal repeats (ITRs). Prominent superfamilies of class II transposons include the Tc1/mariner, hAT and piggyBac families. While most class II transposons utilize the mechanism described above, there are some exceptions. Some DNA transposons (or integrated phages) are not excised from the genome at all, instead they replicate by strand-transfer reactions which are later resolved during DNA replication[69]. *Helitron* transposons use a rolling circle-based replicative mechanism[70] and *Polinton* transposons use a different mechanism also based on a single-stranded intermediate[71,72]. All of these mechanisms are replicative rather than conservative, but these transposons are still considered class II transposons because they do not transpose via an RNA intermediate. For the sake of simplicity, the terms "DNA transposon" and "class II transposon" will be used to refer to those DNA transposons which move by a conservative (i.e. cut-and-paste) mechanism, excluding replicative DNA transposons.

Due to the fact that their transposition mechanism is (in most cases) not replicative, DNA transposons cannot increase their copy number as easily. However, they may still be duplicated by transposing from an already replicated to a non-replicated site during S phase[73], and some transposons actively increase the likelihood of such an event[74]. Alternatively, DNA transposons can increase copy numbers by homologous recombination of the gap left behind by transposon excision using a homologous chromosome or sister chromatid containing a transposon copy[75].

The ratio between class I and class II transposons shows drastic variation between different organisms. While the genomes of the yeasts *S. cerevisiae* and *S. pombe* only contain

retrotransposons[68], and retrotransposons make up ca. 95% of the transposons in the human[76] and mouse genomes, other organisms like *Entamoeba histiolytica* and *Trichomonas vaginalis* almost exclusively have DNA transposons[68]. While there are still some DNA transposons present in the human genome, all of them have been inactivated by accumulated mutations, with the last transposition event of a class II transposon being estimated to have occurred around 40 million years ago[77]. The only transposons still active in the human genomes are LINE1 non-LTR retrotransposons and their associated SINEs like the Alu elements[78].

1.2.3 Transposon vectors

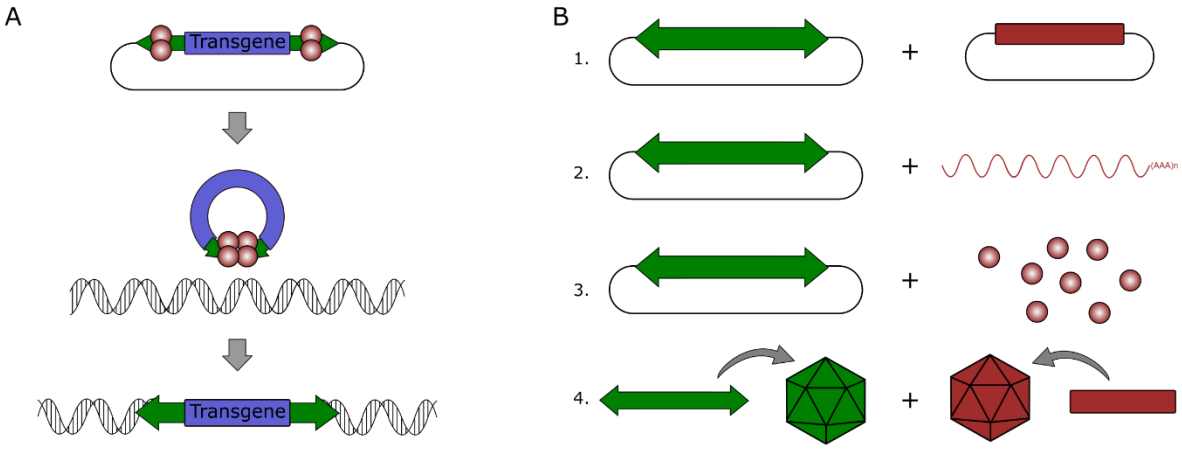


Figure 1.2 – DNA transposons as vectors. A Principle of DNA transposons as gene delivery tools. A transposon is supplied to the cell as a plasmid and transposase molecules bind to the ITRs (top). The transposon is excised and the pre-integration complex finds a suitable target site in the target genome (middle). The transposase integrates the transposon into the genome (bottom). **B** Methods of delivery of transposon (green double arrow) and transposase (red shapes). Both components can be supplied as plasmids (1), but the transposase can also be delivered as RNA (2) or protein (3). In hybrid vectors, the DNA of both components is packed into viral capsids to increase the efficiency of delivery (4).

Like viruses, transposons have been modified for use as genome engineering tools. This is achieved by replacing parts of the DNA sequence of the natural transposon with a transgene and supplying the required enzymes *in trans*. Because class I transposons move by a replicative mechanism and a few retrotransposons – LINE1 elements and dependent SINEs[79] - are still active in the human genome, class II transposons are generally preferred for gene therapy purposes. Attempts to utilize retrotransposons for gene therapy have been made[80], but in general retrotransposons are more likely to be used in non-therapeutic applications[81,82].

Transposons have several advantages compared to viruses, but also some significant drawbacks (Table 1-1). The main drawback is the lower efficiency of transposon vectors when compared to viral vectors[83]. The protein component of viruses is evolved to transfer genetic material across cell membranes, and is very efficient at this task. Transposon vectors lack this machinery and thus have to be transferred into target cells using other methods. Simple chemical methods

can be used for many cell lines *in vitro*[84,85], but are often inefficient for clinically relevant cells. Electroporation is a viable alternative for many of these cells, for example T cells[86] and iPSCs[87], but it usually comes with high cell mortality.

Table 1-1 – Comparison between transposon and viral vectors

	Transposon	**Viral**	**Naked DNA**
Efficiency	High	Very high	Low
Cost	Low	High	Low
Immunogenicity	Low	High	Low
Genotoxicity	Low	High	Low
Size limitation	Relaxed	Strict	Relaxed

The lack of a structural protein component, while limiting the efficiency of transposon vectors, is also their greatest advantage over viral vectors. The need to package viral vectors into infectious particles makes them more complicated and time-consuming in production, handling and storage when compared to transposon vectors[83,88], especially under GMP conditions. This means that transposon vectors will generally have lower costs and it will be easier for non-specialized labs to implement them[89]. The absence of viral proteins also reduces the danger of immune reactions, which can both be a safety concern and limit the expression of transgenes when viral vectors are used[90]. Additionally, the size of the viral capsid places a strict cap on the size of the genetic cargo. For transposons, integration efficiency generally decreases after a certain size, but there is no hard limit regarding the length of DNA that can be packed into them and over 100 kb of DNA have been successfully inserted[91–93]. This size limitation is extremely relevant for therapeutic addition of transgenes that are larger than the packing capacity of commonly used viral vectors.

While the transposon itself is by its nature a DNA molecule, the transposase can be delivered to target cells in several forms (Figure 1.2B). Delivering the transposase in the form of a plasmid is appropriate in some contexts and results in the simplest handling of the transposon system. However, control over transposition activity is limited as the transposase is expressed for an extended time[88] and can even integrate into the target genome[94], which can in turn result in genomic remobilization of the transposon[95]. One alternative is the delivery of transposase in the form of mRNA, which has been shown to reduce cytotoxicity, but is limited in scalable production and stability[96]. Transposase can also be delivered as protein molecules, which results in the most precise control over transposition activity[96].

It is possible to combine some of the characteristics of viral and transposon vectors into so-called hybrid vectors. In these vectors, the DNA transposons of viral vectors are packaged into a viral capsid to allow for easier entry into the target cells. This can increase overall efficiency of transposon vectors, but it negates the advantages of lower cost, reduced immunogenicity and relaxed size limitation that they normally have over viral vectors. The advantage of hybrid vectors over viral vectors is their more favorable integration pattern and thus reduced genotoxicity, which will be discussed in more detail in section 1.5. Hybrid vectors have been generated using a range of different viruses, including adenoviruses[97], HSV[98], AAV[97,99] and lentiviruses[100].

The challenges in the delivery of transposons to cells described above refer to the delivery into cells that were previously isolated from an organism, or *ex vivo* delivery. It is also possible to administer transposon-based therapeutics *in vivo*. This is most routinely done using a technology called hydrodynamic injection (HD)[101], where large amounts of DNA solution are delivered over a short amount of time. This has been mostly tested in mice (e.g. [102–104]), but also seems to be possible in larger animals[105–107]. As an alternative, transfection of cells with PEI *in vivo* has been tested, but was found to be significantly less efficient than HD[108]. Further methods for *in vivo* delivery of transposons include the use of nanoparticles[109,110], intramuscular electroporation[111] and the use of hybrid vectors[97,98,104].

Transposon vectors are also used for other purposes than gene therapy. For example, the mutagenic nature of transposons can be utilized in functional genomics screens[112–114]. However, due to the focus of this project, transposons will be mostly discussed here as gene delivery tools. The two most commonly used transposons for gene therapy are the *Sleeping Beauty* (SB) and *piggyBac* (PB) vectors, but other transposons like *Tol2* are also used. The SB transposon will not be described here as it will be discussed in detail in section 1.3.

The PB transposon was isolated from the genome of the cabbage looper moth (*Trichoplusia ni*)[115,116]. The transposon has a length of ~2.4 kb and the transposase consists of 594 aa[117]. Long believed to be an isolated element, *PB* has since been shown to be part of a family of transposon present in a range of organisms[118–120], with active copies even found in the genomes of some mammals[121]. The transposons of the *PB* family integrate into TTAA tetranucleotides[122] and *PB* transposition does in general not leave a footprint at the site of excision[123]. Several optimizations have been performed on both the *PB* transposase and the transposon ([123–126], reviewed in [83]). The *PB* system is also capable of integrating very large DNA cargoes of up to 200 kb[92,93], making it an attractive tool for insertion of very large genes.

The *Tol2* transposon belongs to a DNA transposon superfamily called *hAT*; it was discovered in a fish genome (*Orizyas latipes*)[127]. The transposon in its natural form is ca. 4.7 kb in size and the most active transposase isoform consists of 649 aa. As in the case of the *PB* system, both the transposon and transposase components have been optimized for increased activity[128,129] and up to ~10 kb of DNA can be delivered without significant loss of activity[130]. The *Tol2* system is furthermore active in a range of organisms[131–133] and is routinely used to generate transgenic zebrafish models[134,135].

1.3 The SB transposon

The transposon system used in this project is a DNA transposon called *Sleeping Beauty* (SB). The SB transposon belongs to the Tc1/*mariner* superfamily of transposons and is a class II transposon that propagates by a cut-and-paste reaction which is mediated by the enzyme SB transposase. Like other Tc1/mariner transposons, SB integrates into TA dinucleotides which are duplicated during the insertion process. The SB transposase was reconstructed from a consensus sequence of several inactive transposases from different fish genomes[136] and since then both the transposase and the transposon have undergone several cycles of optimization. The SB system has been shown to be active in a range of organisms, including fish [137], mouse[138], rat[139], pig[140], cattle[141] and hamster[142], but mostly restricted to vertebrates[142], with the exception of the chordate *Ciona intestinalis*[143]. In addition, the SB system works in different human cells, including therapeutically relevant primary cells[144] and iPSCs[96].

1.3.1 *Sleeping Beauty* transposon and transposase structure

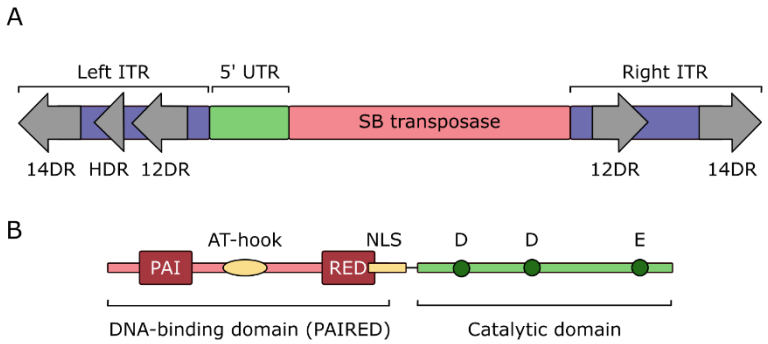


Figure 1.3 – SB transposon and transposase structure. A Structure of the SB transposon. Different transposase binding sites in the ITRs are indicated by gray arrows. The 5'-UTR of the natural SB transposon is indicated in green. **B** – Structure of the SB transposase. The N-terminal DNA-binding PAIRED domain (red) contains two HTH motifs called PAI and RED (dark red), as well as an AT-hook (yellow oval) and a NLS (yellow box). The C-terminal catalytic domain (green) contains a DDE motif as a catalytic center (residues marked with dark green circles).

The ITRs of the original SB transposon have the distinct structure consisting of two direct repeats within the inverted terminal repeats which is characteristic for the IR/DR subfamily of

Tc1/mariner class transposons[145] (Figure 1.3A). The length of the ITRs is around 230 bp and the imperfect direct repeats are around 32 bp long. The outer DRs are found at the very ends of the transposon ITRs while the inner DRs are located ca. 165 bp from the transposon ends. The outer DRs, referred to as 14DR, are 2 bp longer than the inner DRs, which are called 12DR. Each DR represents a single binding site for the SB transposase and both sites in both ITRs are required for efficient transposition[137]. While the DRs are similar, they are not perfectly identical, and it has been shown that the 12DRs are bound more tightly by the SB transposase DNA-binding domain than the 14DRs[146], where cleavage occurs. In addition to the two regular DRs, the left ITR of the transposon also contains a third site, called HDR between the two DRs. The 11-bp HDR site resembles the 3'-end of a DR and acts as a transpositional enhancer necessary for efficient transposition[147]. In the natural SB transposon configuration, the left ITR is separated from the transposase gene by an untranslated region (UTR) of ca. 160 bp. This region has moderate promoter activity[148], but does not seem to serve any function in transposition[149] and is thus usually omitted when the transposon is used as a genome engineering tool.

The SB transposase is a 340 aa and 39 kDa enzyme consisting of a N-terminal domain with DNA-binding activity and a C-terminal catalytic domain[136] (Figure 1.3B). The N-terminal domain contains two separate HTH motifs; the HTH structure that was first described in the eukaryotic Pax transcription factors[150], but which might originate from a transposase[151]. The two HTH motifs are referred to as PAI and RED and the domain as a whole is called the PAIRED domain. The PAI subdomain was found to mediate specific DNA interaction as well as multimerization of SB transposase molecules via a leucine zipper motif, while the RED subdomain interacts with DNA in a less specific manner[147,152]. During binding of the PAIRED domain to the DRs, PAI interacts with the 3'-end of the DR and RED interacts with the 5'-end[153]. Specifically, the RED subdomain has been shown to be responsible for the distinction between 12DRs and 14DRs, binding almost exclusively to 12DRs, while the PAI subdomain does not distinguish between the DRs[153]. In addition, the RED subdomain has also been implicated in multimerization[153]. A GRPR-like sequence (GRRR in the case of SB), also called AT-hook, due to its function in contacting DNA at AT base pairs[154], is found between the two HTH motifs. This motif is conserved across Tc1/*mariner* transposons[145] and has been found to mediate DNA-interaction, for example in the RAG1 recombinase[155]. The PAIRED domain is followed by a bipartite nuclear localization signal (NLS), the N-terminal part of which overlaps with the RED subdomain[147].

The DNA-binding domain (DBD) and the catalytic domain are connected by a flexible interdomain linker of 10 amino acids. The C-terminus of the SB transposase is made up from a catalytic domain with a RNAseH-like fold and containing a DDE motif. DDE recombinases are found in a wide range of organisms, including viruses, prokaryotes and eukaryotes[145,156]. The reaction catalyzed by the DDE domain requires the presence of two metal ions[157]. In SB, the three catalytically active residues are D153, D244 and E279[158]. In addition to the core RNAseH domain, the C-terminal domain contains a flexible clamp loop inserted between two β sheets of the RNAseH fold. This motif mediates protein-protein interaction between two catalytic domains, possibly during a pre-catalytic dimerized state[158].

1.3.2 *Sleeping Beauty* transposition mechanism and target selection

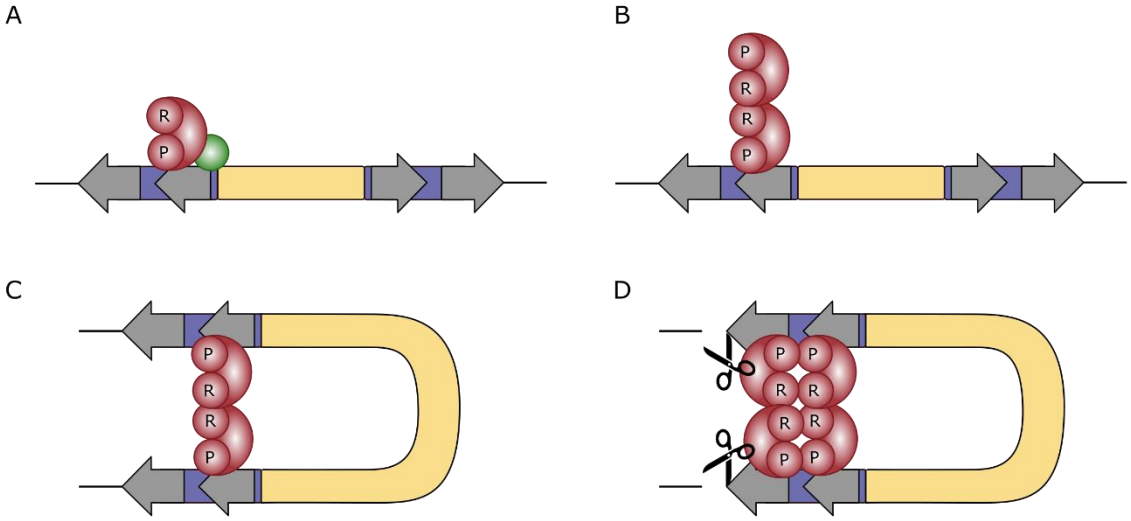


Figure 1.4 – Mechanism of SB excision. A Transposase molecules (red shapes) bind to the IR/DR via the PAI (P) subdomain. Initial binding preferentially occurs at the 12DR is mediated by the RED (R) subdomain and facilitated by HMGB1 (green sphere). **B** Dimerization of SB molecules occurs via RED-RED interaction. **C** A second 12DR is introduced to the complex, bringing the transposon ends into close proximity. No cleavage occurs at the 12DR sites. **D** Addition of further SB molecules via PAI-PAI interaction results in a cleavage-competent complex and cleavage occurs exclusively at the 14DRs.

The first steps in the transposition reaction are binding of transposase molecules to the transposon ends, followed by pairing of the ends to form a synaptic complex or paired-end complex (PEC)[153]. Formation of the PEC is followed by stepwise cleavage of both DNA strands on both transposon ends. Synaptic complex formation is a highly regulated process, which serves to suppress aberrant reactions, e.g. transposition events involving ITRs from separate transposons or from very short, internally deleted transposons (MITEs)[26,153]. The excised transposon can then re-integrate as soon as it encounters a suitable piece of target DNA (tDNA). Cleavage of the tDNA occurs at TA dinucleotides with a 2 bp stagger, resulting in short gaps

after the strand transfer reaction is completed. These short gaps are filled by the host cell's DNA repair machinery, resulting in 2 bp target site duplications (TSDs).

Primary binding to the IR/DR is mediated by PAI subdomains (Figure 1.4A). Preferential binding to the inner 12DR repeats is mediated by the RED subdomain and facilitated by HMGB1. A second transposase molecule is captured by RED-RED dimerization (Figure 1.4B) and a second 12DR (and not 14DR) is introduced (Figure 1.4C). Cleavage at the 12DRs is suppressed. Further transposase molecules bound to 14DRs are incorporated into the complex (PEC) via PAI-PAI interaction (Figure 1.4D). It is likely that DNA-free or single end-bound pre-catalytic auto-inhibited states are involved in the assembly of the PEC[158].

The cleavage reaction catalyzed by the DDE domain is initiated by nicking of a single strand[159], then the opposite strand is cleaved to generate a DSB. In the case of SB transposition, these two cleavage reaction occur sequentially and without a hairpin intermediate[160]. As the first and second cuts are offset by several nucleotides, DNA ends with 3' overhangs are generated[160,161]. The DSB generated by transposon excision is generally repaired by NHEJ, resulting in characteristic footprints at the excision sites[160,161]. The two 3'-OH ends liberated by these hydrolysis reactions are then available as nucleophiles for the strand transfer reaction at the target DNA molecule[162]. Their exact spatial orientation in the target capture complex (TCC) has some implications for target site selection, as described below.

The SB transposon integrates relatively randomly on a genomic scale[163–170], but exhibits some insertion site preferences on a local level. First of all, nearly all SB insertions occur into TA dinucleotides. The nucleotides flanking the insertion site also have some amount of influence over target site quality; the preferred consensus sequence of the SB transposase is ATA<u>TA</u>TAT, with the underlined TA as the actual insertion site. The first and last position of this octanucleotide show particularly strong preference for the presence of T or A nucleotides. However, only around 1.8% of insertions catalyzed by SB transposase occur into the consensus sequence (Kesselring et al., unpublished manuscript).

SB integration also shows preference regarding the physical properties of target DNA. In particular, SB preferentially integrates into DNA that is highly bendable, A-philic and displays a symmetrical pattern of hydrogen bonds in the major groove of the target DNA[163]. The preference for a consensus sequence that was mentioned above also seems to be caused by a unique deformation that is associated with this sequence[167]. The preference for bent DNA is likely due to the spacing of the 3'-OHs in the TCC[158,167], which is a feature shared by other

Tc1/*mariner* transposons[163] and other DDE recombinases[171,172]. It might serve to make the integration reaction irreversible, forcing the DNA to leave the active site after catalysis[171,172]. When the SB transposon is mobilized from a chromosome, an additional effect called 'local hopping' is observed, meaning that sites on the chromosome from which the transposon was mobilized are targeted with a higher frequency[161,173,174].

The efficiency of the SB transposition reaction is size-dependent[137,175]. This effect is already observed at the excision stage of the transposition reaction[142]. While early studies claimed that only residual transposition could be observed for transposons larger than 10 kb[176], the SB transposon has by now been shown to be capable of mobilizing bacterial artificial chromosome (BAC) constructs of ca. 150 kb[91]. In addition, SB transposition is subject to a phenomenon called 'overproduction inhibition' (OPI), in which excessive quantities of transposase reduce the efficiency of the transposition reaction[176]. This observation was also made for other mariner transposons[52,177] and, to a lesser extent, other DNA transposons[178–180]. OPI might be caused by 'quenching' of transposases bound to ITRs[176] or the generation of catalytically inactive multimeric complexes.

Several types of dysregulated transposition reactions can result in non-canonical transposition events. While SB transposition has generally been thought to exclusively occur at TA dinucleotides, it has recently been suggested that non-TA integrations may occur at a rate of ~1.4%[181]. SB transposition is also subject to an effect called 'suicidal autointegration', in which excised transposons re-integrate into their own genomes[142]. In the case of SB, integration can occur in the DNA flanked by the ITRs, but relatively rarely into the ITRs themselves or close to the ITRs[142] and autointegration occurs much more frequently *in vitro* for Hsmar1, another mariner transposon[182]. This process is not specific to mariner transposons, but has been previously observed in prokaryotic[183] and eukaryotic[182] transposons as well as retroviruses[184]. This is in contrast to some prokaryotic transposons like Tn7, which have mechanisms that prevent autointegration[185]. A host-encoded factor called BANF1 has been found to protect retroviruses[186] as well as transposons[142] against autointegration.

Several other cofactors are involved in SB transposition. Ku70 is an NHEJ factor involved in the repair of DSBs left behind by SB excision[160]. Miz-1 is a transcription factor and interaction partner of SB transposase, through which SB transposase downregulates cyclin D1. This slows the cells cycle[187], which results in a prolonged G(1) phase and increased transposition. HMGB1 is a non-histone chromatin factor involved in regulation of RAG1 recombination and is required for efficient differential binding of the 12DRs and 14DRs and thus for proper synaptic complex

formation [188–190]. HMGXB4 is another transcription factor that is involved in regulation of SB transposase expression[148].

1.3.3 *Sleeping Beauty* optimization and variants

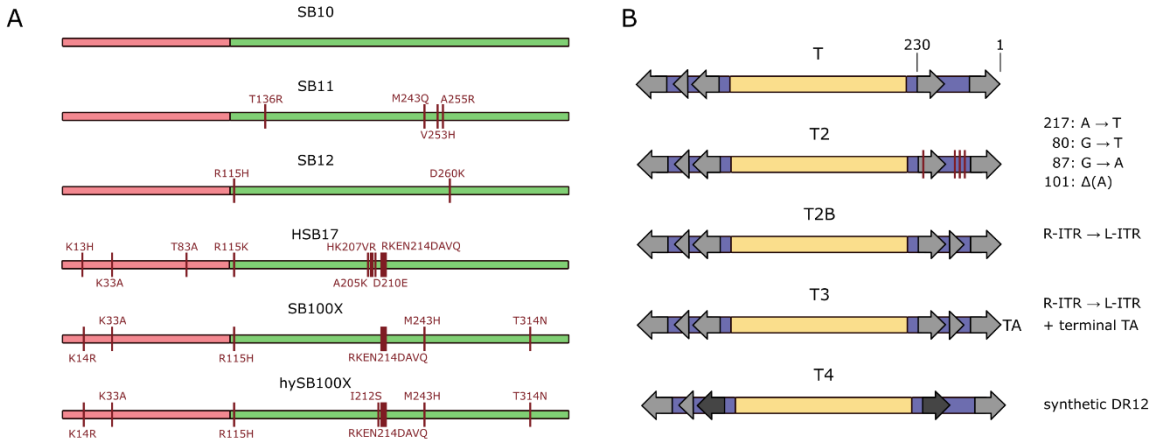


Figure 1.5 – Optimization of the SB system. A Optimized transposase variants. Each mutation (red lines) is indicated above or below the line representing the transposon (DNA-binding domain in red, catalytic domain in green). **B** Optimized transposon variants. The changes made in the transposons are indicated to the right of each schematic. Nucleotides are numbered from the right end of the transposon.

As previously mentioned, evolutionarily successful transposons have to strike a balance between activity and repression to prevent excessive disruption of the host genomes. This means that generally, the components of a transposon system are not evolved for maximal efficiency[149,191] and changes to both the SB transposon and the SB transposase can be made to make them into more efficient tools.

Several successive alterations of the SB transposase sequence have resulted in enzymes with increased activity (Figure 1.5A). The introduction of four point mutations, all in the catalytic domain of the transposase, to the first active transposase (SB10), resulted in an ~ 3-fold increase and the new transposase was termed SB11[176]. Two different point mutations, one in the interdomain linker and the other one in the catalytic domain, resulted in a ~ 4-fold increase compared to SB10[149]. However, the mutations of this new mutant, called SB12, did not result in a synergistic effect when combined with the mutations introduced into SB11[149]. The mutants HSB1-HSB4 were generated by combination of point mutations in the N-terminal PAIRED domain, achieving increases of up to ~ 10-fold compared to SB10[152]. Further combinations of mutations in both the N- and C-terminal domains resulted in the generation of the mutants HSB5-HSB17; the most active of these, HSB17, was ca. 17-fold as active as SB10 and had a total of 12 amino acid replacements distributed across both domains[192]. The transposon variant SB100X is up to 100-fold more active than SB10[191]. The SB100X transposase was generated

by first randomly, then manually combining mutations from a pool of 41 single mutants, which included both previously identified hyperactive mutations as well as phylogenetically conserved residues from related transposases. An additional mutation in the catalytic domain, I212S, was found to result in a further 30% increase of transposition activity compared to SB100X[158].

In addition to changes of the transposase, the transposon has also been modified to increase transposition activities (Figure 1.5B). Three point mutations and one deletion of 1 bp were introduced into the T transposon to generate the transposon T2, which is approximately four times as active as T[146]. A slightly more active variant of T2 is called T2B; in this version the transgene is flanked by two left inverted terminal repeats, which increases transposition efficiency approximately three-fold due to the presence of an additional HDR site[147]. Similarly, the transposon T3 also introduces an additional HDR site into the right ITR and adds an additional flanking TA dinucleotide, increasing transpositional efficiency ca. two-fold[152]. In a final variation, called T4, the 12DR sequences are replaced by synthetic sequences with stronger interaction with the transposase molecules, resulting in a 75% increase in transposition efficiency[153].

Apart from simply trying to increase the activity of the system by optimizing both the transposon and the transposase, many other modification can be made to either component for more specialized applications. For example, hsSB, a SB100X enzyme with the mutations I121S and C176S, has increased solubility[96]. This is relevant for applications in which the transposase should be delivered as a protein; for the SB system, this has so far been limited by a bottleneck in production and delivery[188].

The K248T and K248S mutants of the SB transposase are competent in excision of transposons, but deficient in integration (ex^+ int^- phenotype)[162]. The excision follows a different mechanism compared to wt SB transposase, following a path of nicking, transesterification and hairpin formation instead of double-strand cleavage, reminiscent of other recombinases like RAG1[153]. Excision-only transposons can be used to remove previously integrated sequences from target genomes without risk of unwanted re-integration[193], for example during the generation of iPSCs[162].

Other SB mutants with interesting phenotypes are the mutants H187P and K248R, which have similar changes in their integration profiles. As mentioned in section 1.3.2, the consensus sequence for SB integrations is ATATATAT[163], but only 1.8% of total integrations occur at this

sequence. In both of the mutants mentioned above, the fraction of insertions occurring at the consensus sequence is markedly increased, to 17.9% for H187P and to 33.3% for K248R (Kesselring et al., unpublished manuscript).

Apart from optimizing the transposase or the transposon itself, many variations of the system can serve to increase efficiency or safety. One of these variations is referred to as a 'sandwich' transposon. In this case, the transgene is flanked by two complete empty SB transposons and the 14DRs of the IR/DRs facing the cargo are inactivated, preventing mobilization of the empty SB transposons by themselves[149]. This allows the efficient transposition of significantly larger transgenes than with a simple SB transposon, which loses efficiency when a certain transposon size is exceeded[137,175]. In order to improve the safety of the SB system, insulators can be included in the transposon[148]. This reduces the chance of aberrant activation of cellular transcription by promoters contained in the transgene cassette.

1.3.4 Applications of the *Sleeping Beauty* system

The SB system has a range of possible uses, which can generally be categorized as either gene therapy and transgenesis or gene discovery and mutagenesis. In mutagenesis screens, the disruption of genes in the target cells is the intended outcome. The purpose of these screens is often the discovery of oncogenes[194–197], but they can also be performed to answer questions pertaining to the regulatory architecture of the genome[113].

In contrast to mutagenesis screens, in transgenesis and gene therapy, expression of the genetic cargo is the desired result. In terms of transgenesis, the SB system can be used both in tissue culture[198] and in a range of animals[199]. Apart from studying the effects of a gene, e.g. in the context of a disease model, gene insertion can be performed to generate a desired phenotype, e.g. for the generation of iPSCs. In these gene addition applications, the expression cassette can be varied in a number of ways depending on the exact application, e.g. by making expression conditional, cell- or tissue-specific or including additional components like selection markers or suicide switches. While these considerations can be vital to the success of an experiment or a therapy, they will not be discussed here, as they do not relate directly to the vector itself.

Often, addition of a gene is performed in order to treat a pathogenic phenotype, i.e. for augmentation gene therapy. The SB system has been widely used both in pre-clinical animal models as well as in clinical studies. On the pre-clinical side, SB has been successfully tested in a range of disease models in therapeutically relevant human cells, for example for Fanconi anemia[200], Huntington's disease[201], sickle cell disease[202] and epidermolysis bullosa[203].

Additionally, several studies have applied the SB transposon system to animal models, including mouse models of type I diabetes[102], hemophilia A[204] and B[104], tyrosinemia type I[205], pulmonary hypertension[206] and several malignancies[207,208].

Building on successes in preclinical studies, a dozen clinical trials have been started over the last decade. Most of these have been aimed at treating cancers by generation of CD19-specific CAR-T cells, but a clinical trial against the lysosomal storage disease Hurler Syndrome is also currently planned[209]. SB vectors have been tested for the generation of CD19-specific CAR T cells and a phase I clinical trial has been completed with promising results[210]. The SB transposon has also been used in a phase Ib study attempting to treat Alzheimer's disease[211].

One aspect of SB application that is particularly relevant for gene therapy is the delivery of the SB components, as both safety and efficacy need to meet exceptionally high standards in these contexts. For cell-based, *ex vivo* therapies, this question generally comes down to what format the transposase is delivered in and which protocol is used to deliver the components into the cells. The advantages and drawbacks of delivering the transposase as DNA, mRNA or protein have already been discussed in section 1.2.3. In addition to the benefits of tighter temporal control, nucleofection of mRNAs has been shown to be less toxic than nucleofection of DNA[144]. Instead of – or in addition to – switching from DNA to RNA, safety and efficacy can be improved by eliminating bacterial sequences from plasmid vectors. This reduces problems with gene silencing[212], immunogenicity[213] and potential horizontal transfer of resistance genes[144]. One such technology referred to as Minicirlce (MC) vectors, which almost exclusively contain the GOI, has been successfully used in combination with the SB system and a variety of cell types[214–217]. As an alternative to nucleofection-based protocols, nanoparticle-based vectors have been used in combination with SB, some of which can be combined with targeting molecules to make them target cell-specific (e.g. for MSCs[218], cancer cells[218,219] or hepatocytes[109,220]). Hybrid vectors (see section 1.2.3) of the SB system are also available (reviewed in [144]), including several using adenoviruses[99,104,221] or AAV[222].

For *in vivo* therapies, the question of delivery has the additional dimension of how to deliver the SB components into the target organism. In preclinical studies performed in mice, hydrodynamic injection has been successfully applied to deliver transposon and transposase – or a hybrid viral vector – to the liver[99,223], and to the lung[224] and trials using nanoparticles have achieved high rates of gene transfer *ex vivo*[109].

1.4 Site-specific genome engineering

In contrast to the integrating vectors described in the previous sections, gene editing technologies aim to modify the genome at a precisely defined position. Most of these technologies are based on the introduction of double-strand breaks (DSBs) at a target site, but some gene editing technologies work without DSBs and customizable DBDs can be utilized for other tasks as well.

1.4.1 Nuclease-based genome engineering

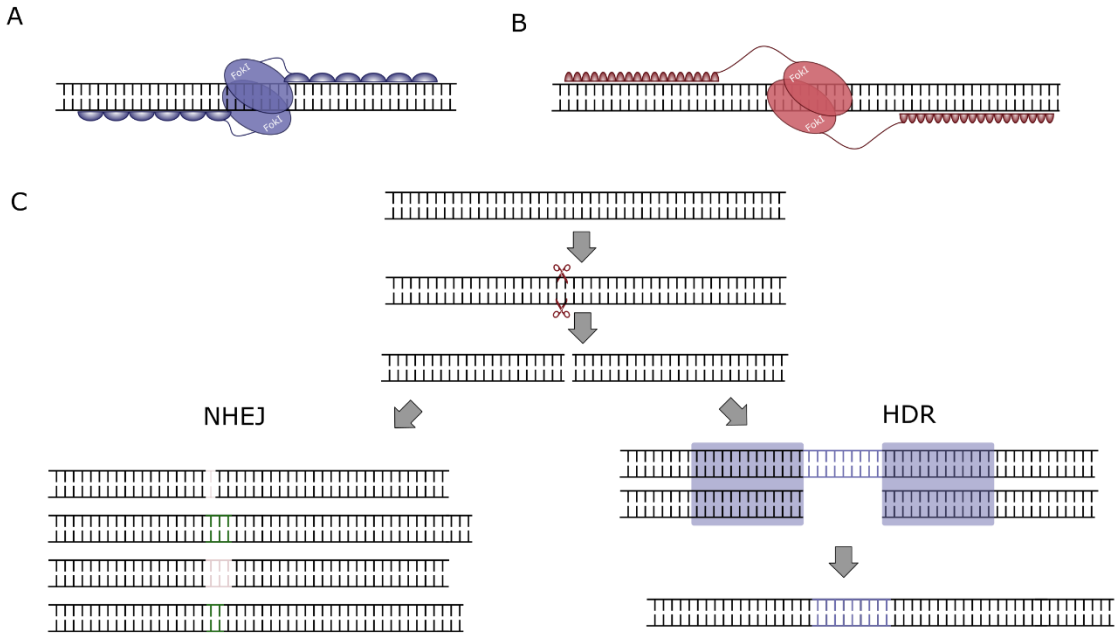


Figure 1.6 – Principles of nuclease-based genome engineering. A Zinc finger nucleases consist of modular DBDs and *Fok*I endonuclease domains. Each module recognizes a specific trinucleotide. **B** TALENs consist of a modular DBD and *Fok*I endonuclease domains, with each module recognizing a specific nucleotide. **C** DSBs can be resolved either via the NHEJ or HDR pathways. In the NHEJ pathway, the free ends are joined directly, but the error-prone nature of this process leads to the formation of indels. In the HDR pathway, a repair template with homology to the sequences flanking the DSB is used. Information contained in the repair template is copied into the DSB site.

The most widely used designer nucleases are zinc finger nucleases (ZFNs)[225], transcription activator-like endonucleases (TALENs)[226], and the CRISPR/Cas system.[227,228]. ZFNs and TALENs are specific classes of DNA-binding proteins than can be engineered to recognize a wide range of DNA sequences, while the CRISPR/Cas system is RNA-guided and will be discussed in detail in section 1.4.2. ZFNs consist of DBDs called zinc finger proteins (ZFPs) and *Fok*I endonuclease domains. ZFPs are a family of DNA-binding proteins abundant in many eukaryotes, including humans[229]. They are modular in structure, consisting of repeated elements (the so-called zinc fingers), that make more or less specific contacts with specific trinucleotides in the target DNA (Figure 1.6A). ZFPs with 3-6 individual fingers can be engineered,

recognizing target sites of 9-18 bp. Thus, a 6-finger ZFP can be sufficient to target a unique site in the human genome. However, when ZFPs are used as nuclease components, two ZFPs binding to target sequences separated by 5-7 bp are designed, each fused to a *Fok*I domain[229]. Simultaneous binding of both ZFPs allows dimerization of the *Fok*I domains and cleavage of the target DNA. This dimerization step additionally increases the specificity of the reaction. While ZFPs recognizing almost all possible trinucleotides have been developed, enabling the targeting of a wide range of sequences, ZFP design is not straightforward due to, among other things, potential interactions between the individual fingers[230].

TALEs (transcription activator-like effectors), like ZFPs, are modular DNA-binding proteins; in contrast to ZFPs, they are of bacterial origin[230]. While ZFPs recognize target nucleotides in groups of three, each TALE repeat contacts a single nucleotide (Figure 1.6B). This allows TALE domains recognizing nearly every possible sequence to be generated[231], the only limitation seeming to be a preference for a 5'-terminal T nucleotide[232]. Engineering of TALE domains is also easier than engineering of ZFPs because they can be linked together in a more straightforward manner[230]. One challenge specific to TALE construction is that the highly repetitive nature of the involved sequences necessitates the use of specialized cloning procedures[233,234]. Like ZFPs, two TALE domains recognizing adjacent DNA sequences and fused to *Fok*I domains can be used as highly specific nucleases for genome engineering, referred to as TALENs.

Both ZFNs and TALENs are used to introduce DSBs at specific positions in the target genome. The introduction of genomic DSBs has a range of problematic consequences. For one thing, DSBs induce a DNA damage response via p53[235,236], which, apart from the immediate toxic effect, can lead to unintended selection of cells with defects in this pathway. Treatment of cells with CRISPR/Cas9 has also been reported to result in complex, large-scale genomic rearrangements, which could be another source of malignant transformation[237]. Both ZFN-based and TALEN-based genome editing can result in off-target cleavage, for example when cleavage occurs after homodimerization[232]. Some obligate heterodimeric *Fok*I nuclease domains have been developed to reduce off-target effects[238,239].

All of these designer nucleases can be used both for knock-outs and knock-ins. This is due to the fact that the cell has two distinct pathways to repair DSBs (Figure 1.6C). The first pathway, called non-homologous end-joining (NHEJ) directly fuses the two DNA ends together. However, this process is imprecise and often leads to small insertions or deletions (indels) at the site of the former DSB. As such indels in an open reading frame (ORF) will often result in

frameshifts, targeting an ORF with a nuclease can be used to knock out a specific gene. The other pathway, called homology-directed repair (HDR), requires a repair template with homology arms flanking the site which is meant to be edited. While homologous recombination in principle does not rely on a DSB, its efficiency is highly increased by the presence of a DSB at the target site. HDR can be used to insert DNA sequences or make precise edits or deletions.

While generation of NHEJ-mediated knock-outs is a fairly efficient process, HDR is significantly less efficient[240] and HDR efficiency decreases with transgene size[241]. Thus, HDR efficiency represents a bottleneck in gene addition applications and recombination-based insertion of genes is generally less efficient than the use of integrating vectors. HDR-based editing generally also generates NHEJ-repaired byproducts[242], often at a higher rate than the desired modification. The ratio between HDR and NHEJ varies, depending on a range of different factors. For one thing, the HDR-NHEJ ratio is dependent on the site of the DSB within the genome[242], for another thing it depends on the cell type as well as the cell cycle phase the edited cell is in. HDR is only active in the late S and G2 phases of the cell cycle[243], making it impossible to edit post-mitotic cells, which make up most of the tissues in an organism, in a HDR-dependent manner[242,244].

The relative inefficiency of HDR when compared to NHEJ is a limiting factor when using nuclease-based genome engineering to insert sequences or make base-specific changes. Several attempts have been made to improve the HDR/NHEJ ratio. A typical approach is to inhibit the efficiency of the NHEJ pathway by either downregulation or inhibition NHEJ-related genes or enzymes[245,246], or stimulation of the HDR pathway[247,248]. An alternative strategy is to use enzymes that can result in initiation of HDR from nicks instead of DSBs[249], which prevents NHEJ. HDR efficiency can also be improved by introducing changes that prevent re-editing[250]. Finally, as the HDR pathway is only active in the S/G2/M phases of the cell cycle, HDR can be favored by arresting cells in these phases[251] or by coupling Cas9 expression to these periods[252]. One sub-project of this thesis will describe an attempt to increase the HDR/NHEJ ratio of the CRISPR/Cas9 system.

1.4.2 The CRISPR/Cas9 system

The CRISPR/Cas system is a recently discovered enzymatic system from prokaryotes which consists of enzyme and RNA components that together perform a nuclease function[228]. While a wide array of CRISPR/Cas systems have been described, the CRISPR/Cas9 system from *S. pyogenes* is most extensively studied and most widely used as a tool and thus only spCas9 will be discussed here. In this system, the nuclease Cas9 is directed to its target sequences by

short RNA molecules called guide RNAs (gRNAs). This is in marked contrast to previously described nucleases, where target specificity is determined by the amino acid sequence of the nuclease itself, making the CRISPR/Cas9 system significantly more flexible than previous nuclease-based genome engineering technologies. CRISPR/Cas technology is adapted from a bacterial immune system originally intended to defend cells against bacteriophage infections[253]. CRISPR/Cas systems can be found both in bacteria and archaea and generally consist of CRISPR loci and Cas (CRISPR-associated) enzymes. The CRISPR loci provide short RNA molecules that direct the Cas9 nuclease against viral DNA. Other Cas proteins are involved modifying the CRISPR loci and in processing these RNAs expressed from them.

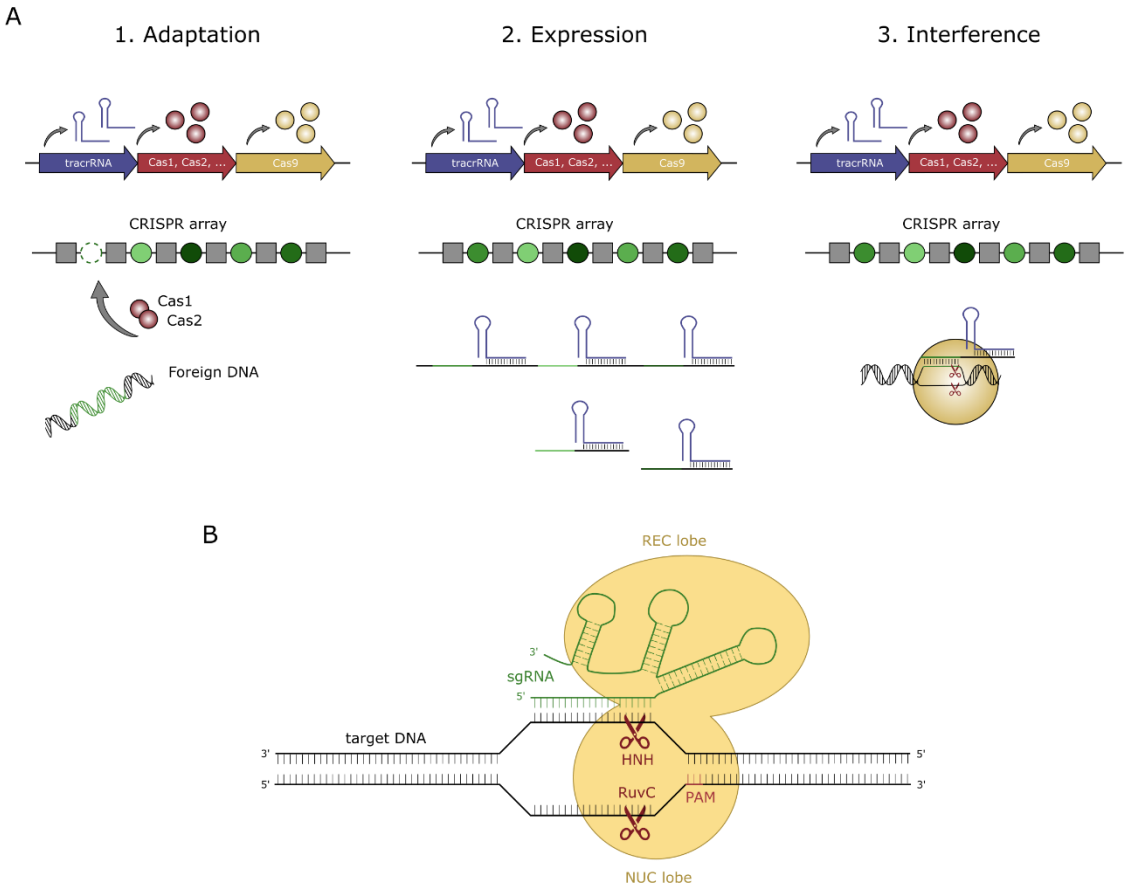


Figure 1.7 – The CRISPR/Cas9 system. A Schematic representation of the CRISPR/Cas9 mechanism as it occurs in nature. (1) Foreign DNA is incorporated into the CRISPR array as new spacers (green circles) between invariant repeats (gray boxes) by several enzymes including Cas1 and Cas2. (2) The CRISPR array is expressed into a pre-crRNA, which is processed into several gRNA complexes consisting of a crRNA and a tracrRNA. (3) Cas9 and gRNA form surveillance complexes which are capable of cleaving the sequences from which the respective spacer was derived. **B** Detailed schematic of cleavage catalyzed by Cas9 in conjunction with a sgRNA, which fulfills the function of crRNA and tracrRNA in one molecule. Cas9 complexes the RNA via its REC lobe, recognizing several hairpins in the region corresponding to the tracrRNA. The complex then scans target DNA for the presence of PAMs and, once a PAM is bound, unwinds the flanking DNA and checks for complementarity with the sgRNA recognition sequence. If complementarity is established, Cas9 cleaves the target strand with its HNH domain and the non-target strand with its RuvC domain, both of which are located in the NUC lobe.

In nature, the sequence information for guiding the nuclease is encoded in CRISPR arrays, genomic loci in which unique spacers alternate with short repeats (Figure 1.7A). The unique spacer sequences are derived from the DNA of invading bacteriophages or plasmids and thus represent a record of previous infections that the organism has encountered[254]. Accordingly, the first step in the defense process is integrating these infecting DNA sequences, called protospacers, as new spacers into the array, a process that involves the proteins Cas1 and Cas2 and is called adaptation[255]. The entire CRISPR array is then transcribed into a pre-crRNA, which is subsequently processed into individual crRNAs, each consisting of a spacer at the 5'-end and a part of the repeat at the 3'-end[256]. The repeat interacts with a separate RNA component of the CRISPR/Cas9 system called transactivating crRNA (tracrRNA)[257] and the Cas9 effector nuclease to generate the complex that cleaves DNA at the sequence defined by the spacer of the crRNA. An additional requirement for cleavage by the Cas9 complex is the presence of a protospacer-adjacent motif (PAM), NGG, next to the recognized target sequence[258]. This requirement prevents cleavage of the CRISPR array in the cell's genome.

In addition to assembly of the CRISPR arrays and processing of its transcripts, the mechanism of the interference step, in which Cas9 cleaves its target DNA, has been extensively studied (Figure 1.7B). The Cas9 enzyme consists of a recognition (REC) lobe, a nuclease (NUC) lobe and a C-terminal domain (CTD)[259]. By itself, the apoenzyme is inactive and binding of gRNAs is required to convert it into an active state[228]. Once bound to a gRNA, Cas9 begins to interrogate the DNA for complementary target sites[260], first scanning for PAMs via interaction with the CTD[259,261] and then checking the flanking DNA for complementarity[262]. If complementarity to the gRNA is established, the Cas9 enzyme is activated for cleavage. Cas9 contains two nuclease domains, a HNH-like domain that cleaves the target strand (complementary to the gRNA sequence) and a RuvC-like domain that cleaves the non-target strand[228]. Both of these cuts occur at a position 3 bp from the PAM[228].

For purposes of bioengineering, the CRISPR/Cas9 system can be significantly simplified. For one thing, if the gRNAs are supplied manually, all enzymes involved in processing the CRISPR loci and the guide RNAs can be omitted and only the nuclease (Cas9) is required. Additionally, tracrRNAs and crRNAs can be combined into a single molecule referred to as single guide RNA (sgRNA)[228]. Because the structure and function of Cas9 have been well studied, a number of Cas9 variants could be developed. As the HNH- and RuvC-like domains can be independently inactivated by point mutations (H840A and D10A, respectively), it was possible to generate Cas9 variants that only produce single-strand nicks instead of DSBs. These 'nickases' can be

used to stimulate HDR at a site without activating the NHEJ pathway[263]. Additionally, by combining both inactivating mutations into a single enzyme, a cleavage-deficient – but DNA binding proficient – Cas9 variant, dCas9 ('dead Cas9') was generated[228]. Another variant of Cas9 was recently developed to work with only a minimal PAM requirement (NYN instead of NGG)[264].

Like ZFNs and TALENs, Cas9 has the potential to generate off-target cleavage events and in the case of Cas9 off-target activity varies significantly between gRNAs. In the case of Cas9, the tolerance for mismatches is asymmetrically distributed across the length of the gRNA binding sequence. Bases in the so-called 'seed region', consisting of the 10-12 PAM-proximal nucleotides, generally need to be perfectly matched[262], while PAM-distal mismatches are better tolerated[265]. Mismatches in the non-seed region are particularly well tolerated during the DNA binding step[266] and sites with seed sequences as short as 5 bp have been found to be bound by Cas9[261,267]. Several Cas9 variants which have reduced off-target activity have been developed[268–272] and choices in sgRNA design[273] and spatiotemporal regulation of Cas9[274] can be used to minimize off-target effects.

Obviously, specific and simple genomic modification has many laboratory applications, for example in studying gene functions or generating disease models. The generation of gRNA libraries allows genome-wide screens to be performed with relative ease[275]. However, the most exciting prospect of this technology is its potential for treating genetic diseases and malignancies. The CRISPR/Cas system has been validated in preclinical studies on human cells or animal models for a wide range of diseases, including Duchenne muscular dystrophy[276,277], sickle cell disease[278], β-thalassemia[279], amyotrophic lateral sclerosis[280], Huntington's disease[281] and metabolic diseases[282]. Apart from success in preclinical studies, since 2016 first clinical trials have demonstrated the safety of CRISPR/Cas9 in patients, attempting to treat several cancers, sickle cell disease and thalassemia[283–285].

1.4.3 Nuclease-free editing and other site-specific methods

As described in the previous two sections, the ability to generate highly specific DSBs allows both knock-outs and knock-ins via NHEJ and HDR pathways, respectively. However, DSBs are associated with a range of negative consequences for the target cell and some parameters, like gene regulation, cannot be controlled in this manner. To address this, the specific DNA-binding abilities of ZFNs, TALENs and the Cas9 system can be uncoupled from their nuclease activity to direct other types of enzymes to specific target sites.

In the case of ZFPs and TALEs, this can be achieved by simply replacing the *Fok*I nuclease domains by other functional domains or enzymes. It should be noted that in this case a layer of regulation is lost because dimerization is no longer required for activity at the target sites. This means that DBDs with shorter recognition sequences will often not be sufficient to specify unique sites in a target genome. In contrast to ZFNs or TALENs, the DNA-binding and nuclease activities cannot be clearly separated in the CRISPR/Cas9 system. However, the dCas9 mutant, which is cleavage-deficient but active in DNA-binding can be used as a DBD and fused to effector proteins in its entirety[286,287].

Several types of effector proteins have been fused to DBDs. Fusion of transcriptional activators and repressors[288–290] as well as epigenetic modifiers[291–293] to DBDs allows specific and reversible up- or downregulation of genes or epigenetic remodeling. Addition of polymerases to DBDs can be used to diversify sequences at a desired locus[294]. By combining DBDs with base editors, all possible transition mutations can be introduced into the DNA without the generation of DSBs[295–298]. An interesting variant of this approach is prime editing, in which a Cas9 domain (in this case a nickase instead of dCas9) is fused to a reverse transcriptase domain to flexibly write changes directly into the genome from a provided template[299]. Fluorescent proteins fused to dCas9 can be utilized to visualize genomic loci[300] or entire chromosomes[301]. Finally, some recombinases have been successfully directed to specific sequences with custom DBDs. A special case of this, combinations of DBDs and transposases, will be discussed in a later section.

1.5 Natural and artificial targeted insertion systems

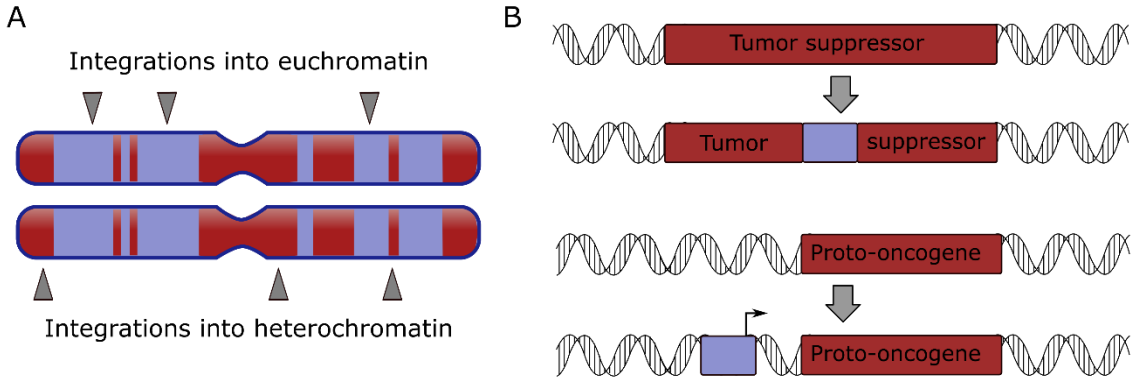


Figure 1.8 – Problems caused by random insertion of integrating vectors. A Position effects: A transgene cassette can either integrate into regions accessible by the transcription machinery (e.g. euchromatin, blue) or less accessible regions (e.g. heterochromatin, red). Due to effects of the genomic surroundings on the transgene, expression levels can be hard to predict. **B** Genotoxicity: Dysregulation of the host genome by the integrated cassette (blue) can have serious negative effects if either a tumor suppressor gene is inactivated (top) or a proto-oncogene is activated by an element contained in the cassette (bottom). In both cases, transformation of the target cell is a potential outcome.

The site-specific genome modification technologies described in section 1.4, as described above, all suffer from some drawbacks, including the problems associated with generating DSBs, and the relatively low efficiency of HDR-based integration of transgenes, especially large ones. Thus, for many applications integrating vectors seem to be the better tool. However, all of the integrating vectors described so far integrate in a semi-random fashion. While some of them have requirements on the sequence level – e.g. the SB transposon only integrating in TA dinucleotides – their integration is, to some degree, random on a genome-wide level. While some vectors have a biased integration patterns, ultimately the site of integration is determined by chance. This type of non-specific integration is in contrast to the gene editing technologies described in section 1.4, which are always aimed at a specific sequence.

Random integration results in two major problems, position effects and genotoxicity, both of which are particularly problematic for gene therapy. Position effects describe the influence on genomic context on the transgene (Figure 1.8A). The genomic location can influence the level of expression of a transgene and whether the transgene is even expressed at all. Overexpression of a transgene may have deleterious effects on the cell while low level or complete lack of expression would prevent any therapeutic effect. Genotoxicity, on the other hand, is caused by effects transgene integration can have on the host cell (Figure 1.8B). Insertion of transgenes can disrupt genomic homeostasis by inserting directly into coding or regulatory sequences and regulatory elements contained in the vectors can result in upregulation of genes, as in the case described below. This is particularly problematic if cellular homeostasis is disturbed by knockout of a tumor suppressor gene or by upregulation of a proto-oncogene, which can result in transformation of the target cell and – in a gene therapy context – tumor formation in the patient. Both transposon[302,303] and viral[304] insertions have previously been shown to be the underlying cause of genetic diseases. In sum, more than 100 naturally occurring pathogenic mutations have been associated with transposition events[32,305] and hundreds of natural viral insertions have been linked to diseases[304].

The issue of genotoxicity most prominently became apparent after 5 out of 20 patients in a trial using γ-retroviral vectors to treat X-SCID developed leukemia[306]. This was later shown to be related to insertional mutagenesis[307,308]. Specifically, this seems to be caused by insertion of the transgene near the *LMO2* gene and subsequent overexpression of LMO2 induced by the LTRs of the retroviral genome. However, insertion-unrelated events seem to have also been involved in the development of leukemia in the patients[309]. The development of leukemia was also an observed outcome in a separate γ-retrovirus based gene therapy trial attempting to treat Wiskott-

Aldrich syndrome[310] and genes associated with leukemia were found to be activated in another trial for X-CGD[311]. Self-inactivating (SIN) vectors have since been developed to increase the safety of γ-retroviral therapy[312,313]. Lentiviruses are generally considered safer than γ-retroviruses because of their integration patterns (see below) and the absence of viral regulatory elements in SIN lentiviral vectors[314], but clonal expansion has also been observed after the use of lentiviral vectors[315]. In the end, both viral systems have unfavorable integration profiles and even random integration comes with a risk of insertional mutagenesis.

While, assuming a random integration profile, the chance of activating a proto-oncogene has been estimated to be less than one in 10 million[316], the observed frequency of such events is significantly higher, depending both on the used vector and the target cell[317]. All of these observations combined lead to the conclusion that it is of high interest to develop novel types of integrating vectors with safer insertion patterns, or, ideally, vectors which can be targeted to specific elements of genomic positions at will.

1.5.1 Natural targeted insertion systems

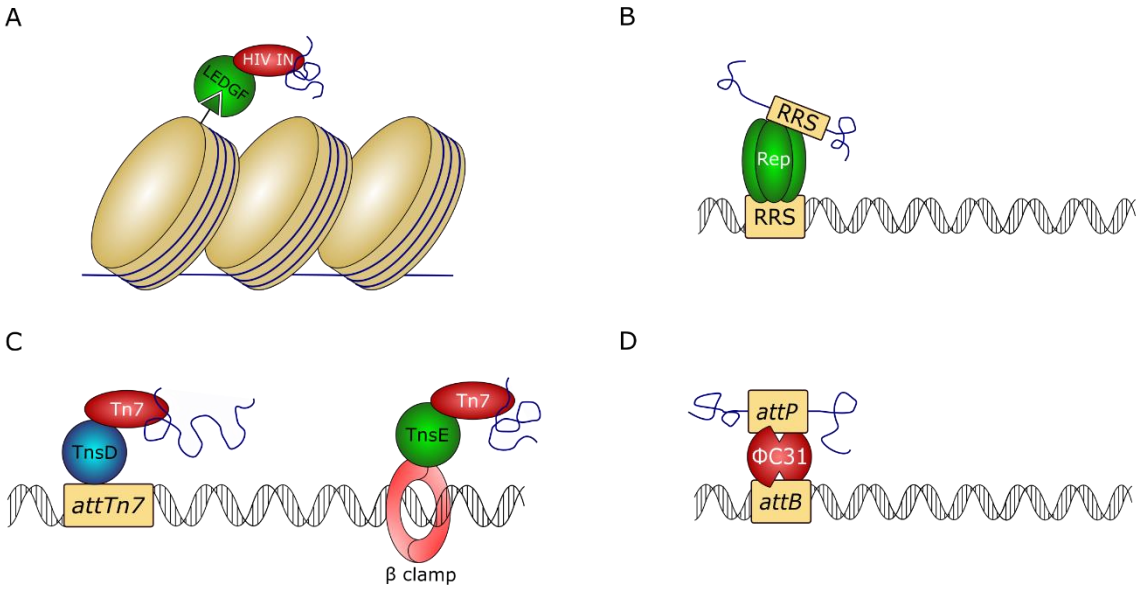


Figure 1.9 – Naturally occurring targeted integration. A HIV preferentially integrates into active transcription units, marked by the histone modification H3K36me3 (teal triangle). The modification is bound by the host protein LEDGF, which in turn interacts with the HIV IN. **B** AAV specifically integrates into the *AAVS1* site on chromosome 19. The virus-encoded Rep protein bind to Rep recognition sequences (RRS) present both in the human and in the viral genome. **C** The Tn7 transposon has two distinct natural targeting mechanisms. The transposon-encoded factor TnsD recognizes the *attTn7* site in the target genome in a sequence-dependent manner, while the protein TnsE binds to the β clamp and steers integrations to actively replicating DNA. **D** The ΦC31 integrase catalyzes site-specific recombination between two distinct sites, one present in the host genome (*attB*) and one in the phage genome (*attP*).

Many naturally occurring integrating viruses or transposons have unique biases in their target site selection and these biases can be the result of a range of different mechanisms (reviewed in

[318]). Studying these naturally targeted systems can be of great use for artificial retargeting attempts.

The two viral integrating vectors mentioned above have distinct integration patterns, but the underlying mechanism is similar for both. The γ-retrovirus MLV has a tendency to integrate near transcription start sites (TSS), CpG islands and DNAse I-hypersensitive sites[319–321]. This seems to be due to an interaction between the MLV integrase (IN) and the host factors of the bromodomain and extraterminal domain (BET) family[322–324], a group of chromatin reader proteins[325]. Indeed, disruption of the interaction between the MLV IN and BET proteins results in a loss of the characteristic integration pattern of MLV[326]. Lentiviruses like HIV, on the other hand, tend to insert into active transcription units, with no preference for the TSS[327]. Again, the effect is due to an interaction between the viral IN and a host protein, in this case lens epithelium-derived growth factor (LEDGF)[328–331], which binds in the vicinity of actively transcribed genes[332] (Figure 1.9A). The interaction between the viral and the host protein tethers the integration complex to LEDGF binding sites[333] and as a result the HIV integration profile shows remarkable similarity to the binding profile of LEDGF[332,334].

The biased integration of wild-type AAV differs from HIV and MLV in that it is targeted to a specific site rather than to a general genomic feature. Specifically, AAV integrates at a site on chromosome 19 called *AAVS1*[335,336] with an efficiency of approximately 0.1%-0.5% of all infecting genomes[337]. In contrast to HIV and MLV, the specificity of AAV integration does not rely on any host factor but is entirely mediated by the virus-encoded Rep protein (Figure 1.9B). Rep is a DNA-binding protein that recognizes sequences present in the ITRs of the viral genome as well as the human genome (at the *AAVS1* locus)[338–340]. While this specific integration would be desirable in many contexts, especially since the *AAVS1* locus satisfies genomic safe harbor (GSH) criteria[341], the Rep protein is generally removed from recombinant AAV (rAAV) vectors in order to free up space for the transgene and because of the toxicity associated with the Rep protein, precluding specific integration[342].

Biased integration near genomic features or sequences is not only observed in viruses, but also in transposons. The *Saccharomyces cerevisiae* retrotransposons Ty1 and Ty3 have a similar integration profile, both integrate upstream of genes transcribed by Pol III[343,344], a group which mostly comprises tRNA, 5S rRNA and other short non-coding RNA genes[345]. While the targeted region is the same for both TEs, Ty1 has a relatively wide targeting window of several hundred base pairs upstream of the TSS[346–348], Ty3 integrates in a very narrow window of only one or two base pairs[349]. The targeting mechanisms are similar for both TEs, relying on an

interaction between the respective integrase (Ty1 IN or Ty3 IN) and components of the Pol III complex (TFIIIB[350,351] and TFIIIB/TFIIIC[352–354], respectively). The Ty5 retrotransposon, also from *S. cerevisiae*, has a markedly different integration pattern from Ty1 and Ty3, preferentially integrating into heterochromatin[189,355,356]. Again, interaction between the IN and a host factor is responsible for the effect, the host factor in this case being Sir4p[357–359], a protein component of heterochromatin in yeast[360,361].

One interesting tendency that has been noted before[349] is that many viruses tend to target actively transcribed genomic regions, while transposons often prefer less actively transcribed loci. A possible explanation for this effect could be the differences in the life cycles of viruses and transposons. In order to generate new viral particles, viral genomes need to be transcribed and insertion into loci accessible to the transcription machinery may grant viruses an evolutionary advantage[317]. Transposons, on the other hand, may need to avoid severe disruption of the host cell as they lack an extracellular phase in their life cycle[179] and death of the host cell would mean loss of the transposon. The Ty1 and Ty3 retrotransposons mentioned above tend to integrate near Pol III promoters, but the regions upstream of these promoters are often gene-poor[362]. In fact, the regions upstream of Pol III promoters seem to generally be good targets for integration in dense genomes, balancing transcriptional accessibility with minimal disruption of the host genome. Apart from the Ty1 and Ty3 retrotransposons mentioned above, the same target preference seems to have developed independently at least six times[363] However, the observations concerning target preferences of transposons versus viruses describe general tendencies and not strict rules, as evidenced, for example, by the *Schizosaccharomyces pombe* retrotransposon Tf1, which integrates into Pol II promoters[364,365].

All naturally targeted TEs described so far have been retrotransposons, but DNA transposons also show biases in their target site selection. A particularly interesting case it the bacterial DNA transposon Tn7, which can target both specific genomic structures and DNA sequences, depending on the target factor that is utilized (Figure 1.9C). One of these targeting factors, TnsE, interacts with the β clamp[366], a component of the DNA replication machinery, directing Tn7 into actively replicating DNA[367]. Alternatively, the targeting factor TnsD can be used to direct Tn7 integration to specific sequences called *attTn7* sites[368,369]. Interestingly, three sequences with high similarity to *attTn7* (pseudo *attTn7* sites) are present in the human genome, but while preferential integration into two of these sites could be shown *in vitro*, no targeted insertion into the human genome could be demonstrated *in vivo*[370].

Another interesting system displaying high specificity is the IN of the ΦC31 bacteriophage[371] (Figure 1.9D). In contrast to most mechanisms described so far, the ΦC31 IN does not rely on interaction with another protein to tether it to a target site, but rather recognizes the target sequences (the *attB* site in the bacterial genome and the *attP* site in the phage genome) itself. ΦC31 IN is also active in human cells[372], and some directed integration into pseudo *attP* sites in the human genome could be demonstrated[373]. While ΦC31 has been tested in preclinical models[374–379], it seems to induce a relatively strong DNA damage response, making its use in a therapeutic setting problematic[380,381]. The system is, however, useful for other applications like transgenesis in model organisms like *Drosophila*[382–384].

The two most widely used transposon vectors, PB and SB, also show some integration bias, although the SB system only diverges very slightly from a random distribution. The PB transposon displays an integration pattern remarkably similar to that described above for MLV[170]. Indeed, it was shown that the distribution is based on the same mechanism, as like the MLV IN the PB transposase interacts with BET proteins, specifically with BRD4[170]. An interesting bias can be observed with the SB system; in contrast to the other systems described above, the SB transposase is not directed to specific sites by interaction with host factors. Rather, in this case, the tethering factor seems to be other SB transposase molecules. Because they interact with one another via their N-terminal domains, SB transposase molecules can bring other SB integration complexes into the vicinity of pseudo SB sites[170], i.e. sites in the genome with high sequence similarity to the sites recognized by the SB transposase in the transposon ITRs. This causes some enrichment near these pseudo SB sites in the human genome[170].

1.5.2 Artificial retargeting of integrating vectors

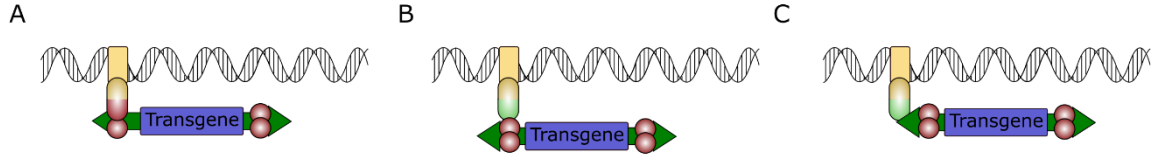


Figure 1.10 – Artificial vector retargeting. A Vectors can be retargeted by generating a direct fusion between the recombinase (red half-sphere) and a DNA-binding domain (yellow-half sphere). The DBD recognizes its target site (yellow box) in the genome and tethers the integration complex to the genomic location specified by the DBD. **B** Alternatively, DNA-protein adapters can be generated by fusing a DBD to a protein-interacting domain (green half-sphere) that noncovalently interacts with the recombinase. **C** DNA-DNA adapters can be generated by fusing two DBDs, one recognizing the desired target and one binding to a sequence in the viral or transposon genome.

As is often the case, nature can be used as an inspiration when designing artificially targeted integrating vectors. Two main types of mechanism were described in the previous section:

1) the integrating enzyme (IN/transposase/recombinase) directly interacts with the target genome (ΦC31 IN and AAV Rep), or 2) the integrating enzyme interacts with another protein which has a DNA- or chromatin-binding activity (e.g. PB, Ty1, Ty3, Ty5, Tn7, MLV and HIV IN). Both of these mechanisms can potentially be mimicked when attempting to target integrating vectors, whether they are viral or transposon vectors. Artificially targeted vectors are useful mainly for avoiding disruption of the target genome and making the behavior of transgenic cells more predictable, particularly in gene therapy settings. However, fusions of integrating vectors to DBDs of unknown specificity can also be used to gain information about the binding profile of the DBD[385–387]. Retargeting of the SB transposon system will be discussed in detail in section 1.5.3.

The first targeting mechanism described above, where the integrating enzyme directly interacts with the target genome, can often be mimicked by directly fusing a DBD to the integrating enzyme (Figure 1.10A). An advantage of this approach is its simplicity, as only a single targeting enzyme is needed. However, in many cases integrating activity is reduced after fusion to a DBD. While many viral INs tolerate fusions relatively well, capsid packaging can suffer from the addition of a new domain and *in vivo* activity can be negatively affected[388,389].

The HIV IN has been the subject of several retargeting attempts by direct fusion with DBDs. DBDs used in these experiments include LexA[390], λR[388], Zif268[391] and E2C[392]. For all of these fusions, some targeting activity could be observed *in vitro*. The HIV-IN/E2C was also shown to increase integration near the genomic *erbB-2* site ten-fold, but *in vivo* activity was significantly reduced[389]. The MLV IN has also been shown to target genomic binding sites after fusion to a Sp1 zinc finger domain[393] and ASV IN has been retargeted *in vitro* by fusion to LexA[394].

Like viral IN fusions, transposase fusions have been shown to exhibit some targeting effects. A fusion of PB transposase to a ZF DBD targeting the *CHK2* gene, has been shown to be effective in a plasmid-to-plasmid targeting assay, but did not successfully direct insertion to the endogenous genomic target[395]. Fusions of PB to Gal4[396] and to a TAL domain recognizing the *CCR5* gene[397] have demonstrated the capability of targeting exogenous and endogenous genomic sites in cell culture assays. Other transposons that have been re-targeted by direct fusion with a DBD include the bacterial transposons IS30[398] and ISY100[399] and the *C. elegans* mariner transposon Mos1[400].

As described in the previous section, many naturally targeted insertion systems rely on adapter proteins rather than on binding of the target DNA by the recombinase itself (Figure 1.10B,C). This approach has the advantage of enabling the use of an unmodified recombinase and thus avoids the issue of activity loss in fusion proteins. Adapter-based targeting is exemplified by HIV IN, whose interaction with LEDGF biases its integration pattern towards active transcription units. This interaction has been successfully used to re-target HIV insertions *in vitro* with an adapter protein consisting of the LEDGF IN-binding domain and λR[401]. HIV insertions have also been directed towards intergenic regions *in vivo*, using a modified LEDGF protein in which the chromatin-binding domain is replaced with CBX1, which recognizes H3K9me3 chromatin marks[402] and this system has been validated in an X-CGD model[403]. Equivalent constructs using ING2[404] and HP1α[405] DBDs have been shown to bias HIV integrations in cell culture-based assays. The safety profile of HIV insertions can also be improved by deletion of the LEDGF chromatin-interaction domain, which essentially results in a random integration profile[406]. However, for this approach to work, endogenous LEDGF has to be knocked down in the target cell, so it cannot be utilized in many contexts.

Adapter-based targeting has also been demonstrated for transposons, for example the yeast retrotransposon Ty5[359]. While the PB transposon has been re-targeted by direct DBD fusion several times, generating adapter proteins for this system is not straightforward because no interaction domains are known for the PB transposase. However, a fusion of a TALE domain targeting the *CCR5* locus and Gal4 has been shown to bias PB insertions when a Gal4 binding site is included in the transposon itself[397].

Some of the DBDs mentioned in this section are occurring naturally and thus only offer very limited flexibility in target site selection. Others, like ZFP and TALE domains, can be custom made to target a wide range of sequences, but have to be specifically engineered for each of these sites. The CRISPR/Cas9 system, due to being RNA-guided, is significantly more flexible than the other systems. Catalytically inactivated Cas9 can be used as a DBD to flexibly target other effector enzymes to sites defined by a gRNA (see section 1.4.2). Generating a targeted integrating vector based on this technology would therefore be a significant step forward, combining the easy target site selection of CRISPR/Cas9 with the efficient DNA insertion of integrating vectors.

Strikingly, an attempt to target PB integrations to the *HPRT* gene with a fusion of dCas9 and PB transposase failed to generate any targeted insertions, and even seemed to prevent insertions at the targeted locus, even though analogous constructs using ZFP or TALE DBDs were able

to produce a targeting effect[407]. A fusion between dCas9 and the *mariner* transposon *Hsmar1* has been shown to efficiently retarget transposition in an *in vitro* plasmid-based assay, but has failed to produce a measureable targeting effect in *E. coli*[408]. A more recent attempt to retarget the PB transposon by directly fusing it to dCas9 has demonstrated targeted transposition into the *CCR5* locus with an estimated efficiency of around 0.06%[409].

Two recent studies have demonstrated highly efficient RNA-guided transposition into bacterial genomes. In one study, the Tn7-like transposon Tn6677 from *Vibrio cholerae*, which is naturally associated to a type I-F CRISPR/Cas-system, was targeted to several genomic sites in *E. coli* [410]. The cas transposon of Tn6677 encodes for the protein TniQ, a homolog of TnsD, which mediates interaction between the cleavage-deficient Cas system and the transposase complex. By supplying alternative gRNAs, Tn6677 can be re-targeted and insertions occurred mainly ~ 50 bp downstream of the targeted sites, with targeting efficiencies of over 90%[410]. Two other Tn7-like transposons from cyanobacteria, which are associated with type V-K CRISPR systems via TniQ, were similarly shown to be targetable by providing appropriate gRNAs[411]. Again, insertions occurred ~ 60 bp downstream of sites bound by the gRNAs and at rates of 50% or higher[411]. It is striking that both of these highly efficient systems are based on naturally evolved RNA-mediated targeted transposition, rather than being *de novo* creations. However, a major limitation of both of these systems is that these transposons are only active in bacterial and not in eukaryotic cells. It might be possible to adapt them for use in mammalian cells, but even if this is possible, it remains to be seen if similarly high targeting efficiencies could be achieved in the significantly larger genomes of these cells.

1.5.3 *Sleeping Beauty* retargeting

The *Sleeping Beauty* transposon system has been re-targeted several times using both direct fusion and adapter-based approaches. While direct fusions of DBDs to the C-terminus of the SB transposase have been shown to completely abolish transpositional activity[165,412,413], N-terminal additions are tolerated to some extent. Direct fusions of SB transposase to several DBDs have been generated.

Proteins generated by C-terminal additions of Gal4 or the ZFP E2C (which targets the *erbB-2* locus) to HSB5 transposase have been shown to increase integration near target sites up to 11- and 8-fold in inter-plasmid assays, respectively[412]. However, the genomic *erbB-2* locus, could not be efficiently targeted by the E2C-HSB5 fusion. A similar construct using the SB transposase M3a instead of HSB5 was shown to direct some integrations towards the endogenous *erbB-2* locus, but this result could not be confirmed by analysis on a whole-genome

level[165]. While the E2C ZFP recognizes a single-copy target, it is also possible to design custom DBDs that recognize a high number of sites in the target genome. This approach was utilized with a fusion of SB to a ZFP called ZF-B, recognizing a sequence near the 3'-end of the human L1 element. With this fusion protein, the fraction of insertions within 400 bp of ZF-B binding sequences increased four-fold and 40% of total insertion were found to occur in L1 elements, up from 32%[165]. An additional DBD that has been used to generate SB fusions is the AAV Rep protein. While the fusion failed to direct any insertions towards the *AAVS1* locus, some insertions were found to occur close to Rep recognition sequence (RRSs), sites with a few mismatches to the sequence recognized at the *AAVS1* locus[164].

In contrast to the PB transposon system, for which no transposase-interacting domains are known, adapter proteins for the SB system can be easily constructed. This can be achieved by fusing DBDs to the peptides known as N57 and N123, both of which are N-terminal fragments of the SB transposase. N57 consists of the N-terminal PAI subdomain of SB and N123 consists of the entire PAIRED domain, including the NLS[147]. The N-terminal domain of SB is involved in multimerization and these fragments consequently interact with SB transposase molecules. Additionally, the N-terminal domain is also involved in recognition of the transposon DNA[147], enabling the fragments to also tether the transposon itself to potential target sites. A fusion of N57 to TetR was able to bias integration towards a chromosomally integrated tetracycline response element (TRE)[413]. The three DBDs mentioned above for direct fusions, E2C, ZF-B and Rep, were also used to generate adapter proteins by fusing them to N57. N57-E2C, like E2C-SB, was able to direct integrations towards the *erbB-2* locus at a low frequency (<1%) and a slight increase in insertions near ZF-B binding sites could be observed with N57-ZF-B[165]. As with the Rep-SB direct fusion, some enrichment near RRSs could be detected with Rep-N57 adapters, but no integration at *AAVS1* itself[164].

While the availability of peptides like N57 makes it easy to generate target-transposase adapters, the SB system can also be tethered to target sites by generating adapters with dual DNA-binding activities and including the relevant sequences in the transposon itself. Retargeting has been achieved by introducing a LexA site in the transposon and using adapters comprising a DBD and LexA. This approach has been successfully used to target insertions to a chromosomally integrated TRE with a LexA-TetR fusion and towards endogenous matrix attachment regions with a LexA-SAF fusion[413].

All of the SB retargeting attempts undertaken so far have used DBDs with fixed specificities, like ZFPs. As described for other vector systems at the end of section 1.5.2, designing a

targetable SB system using an RNA-guided DBD like dCas9 would greatly increase the flexibility of these systems and would create a bridge between programmable nuclease and integrating vector technologies. The creation of such a system will be the main focus of the work presented here.

1.5.4 Targeting of ribosomal DNA

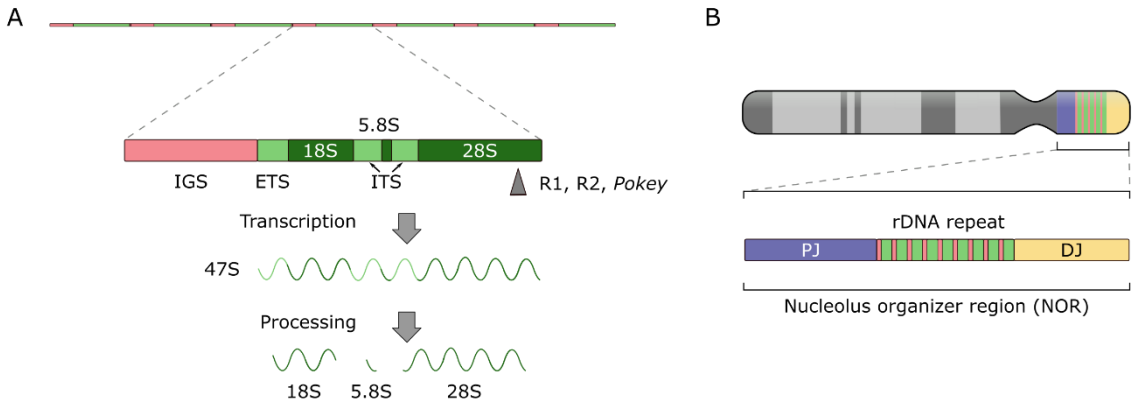


Figure 1.11 – Structure of eukaryotic rDNA and rRNA processing. **A** Schematic structure of rDNA loci. Repeats are tandemly arrayed, with the transcription units separated by intergenic spacers (IGS). The 18S, 5.8S and 28S rRNAs are separated by post-transcriptional processing which removes the external and internal transcribed spacers (ETS, ITS) from the transcript. Approximate insertion locations of R1, R2 and *Pokey* elements are indicated by a gray arrow. **B** Schematic representation of the organization of the NORs on the short arms of the acrocentric chromosomes. Proximal junction (PJ) and distal junction (DJ) regions make up the non-rDNA component of the NORs.

In the previous sections, a range of artificially selected or natural targets has been described, ranging from unique DNA sequences to repetitive elements or specific chromatin structures. However, one appealing target has not been mentioned so far: ribosomal DNA (rDNA). Ribosomal DNA consists of the DNA coding for ribosomal RNAs (rRNAs), which are a structural component of ribosomes. On a sequence level, rDNA consists of tandemly arrayed repeats coding for 18S, 5.8S and 28S rRNAs (the 5S rRNA is transcribed by Pol III from a cluster on chromosome 1[414,415]) and copy numbers of these genes display significant intra-species variation[416] (Figure 1.11A). These repeats are separated by intergenic spacers (IGSs) and the individual rRNA coding regions are separated by internal transcribed spacers (ITSs)[417]. In human cells, these repeats are present in ~ 400 copies on five different chromosomes (13, 14, 15, 21 and 22[417,418]). The rDNA loci are all on the short arms of acrocentric chromosomes and are thus flanked by centromeric and telomeric heterochromatin. In the interphase nucleus, rDNA loci, with the help of over 700 different proteins, organize into structures called nucleoli[419], the number of which varies from cell to cell, with different averages for different cell types[420]. Nucleoli are not enclosed by a membrane, but surrounded by a shell of heterochromatic DNA[421] from so called nucleolus-associated domains (NADs). There, among

a range of other processes, transcription of rDNA into pre-rRNAs and processing into rRNAs take place in the context of ribosome biogenesis[422]. The rDNA repeats are flanked on both sides by additional sequences, which, together with the rDNA are involved in the formation of the nucleolus; together with the rDNA repeats, these sequences are called nucleolar organizer regions (NORs)[423,424] (Figure 1.11B).

The genomic rDNA loci are attractive targets for integration of TEs due to two features. One the one hand, they are actively transcribed, which would likely aid the further propagation of any elements integrated there as well as making it difficult for the host cell to silence the element[425]. Additionally, rRNA loci are highly redundant, thus their disruption is unlikely to have negative consequences for the host cell[416], especially given the fact that only a fraction of rDNA genes are transcribed at any time[426]. Indeed, several transposons have been found to preferentially integrate into rDNA. A whole superfamily of non-LTR retrotransposons, called R elements, display some preference for rDNA as a target. Both the R1 and R2 subfamilies target positions in the 28S region of the rDNA and in both cases, this target site preference seems to be mediated by the fact that these elements encode highly specific endonucleases[417]. The *Pokey* DNA transposons of the water flea *Daphnia pulex*, which is a member of the *piggyBac* superfamily, also preferentially insert into a specific site in the 28S gene region of rDNA, close to the insertion sites of R1 and R2 elements[427,428]. The mechanism for this preference is still unknown and an equilibrium between rDNA and non-rDNA copies of the transposons seems to be involved in the maintenance of the transposon over evolutionary timeframes, as the high recombination rate at rDNA loci tends to contribute to the elimination of TEs from these sites[429].

While elimination via concerted evolution represents a problem for naturally occurring TEs, it is obviously not a problem for applications like somatic gene therapy. However, the advantages of redundancy of the target sequence as well as high transcriptional activity still remain. Thus, rDNA might represent an attractive target for directed integrations. Interestingly, integrations of rAAV vectors have been shown to occur at rDNA loci at a higher-than-random rate[430,431], possibly due to the fact that the vector is processed at nucleoli[432]. In order to enhance this tendency of rAAV to integrate at rDNA, a rAAV-rDNA vector was developed. This rAAV variant contained rDNA sequences flanking the expression cassette, which should favor homologous recombination into rDNA loci[433]. Indeed, the rAAV-rDNA vector integrated into rDNA with a ~ 10-fold increased frequency or an estimated overall frequency of ca. 30%[433].

One major drawback of this approach is that the already low packing capacity of AAV vectors is further reduced by the introduction of homology regions.

Given that rAAV vectors possess an intrinsic preference for integration into rDNA due to their processing in nucleoli, it might be possible to introduce a similar bias to other integrating vectors by physically localizing it in these regions of the nucleus. Several nucleolar localization signals and nucleolar proteins that could perform this task have been described[434]. Thus, one aspect of the work presented here will be the targeting of SB transposition to rDNA by nucleolar localization.

2 Material and methods

2.1 Material

2.1.1 Chemicals

Table 2-1 - Chemicals

Chemical	Supplier
6-thioguanine	Sigma
Acrylamide/Bisacrylamide solution	National Diagnostics
APS	AppliChem
β-mercaptoethanol	Sigma
Bromophenol blue	Sigma
Ethanol	In-house
Glycerol	Sigma
Glycine	Roth
Milk powder	Roth
Neomycin	Invitrogen
Polyethyleneimine	Sigma
Puromycin	InvivoGen
Sodium chloride	Sigma
Sodium deoxycholate	Sigma
Sodium dodecyl sulfate	Roth
TEMED	Roth
Tris base	Roth
Tris-HCl	Roth
Triton X-100	Sigma

2.1.2 Media, buffers and solutions

Table 2-2 – Media, buffers and solutions

Solution	Supplier
CutSmart buffer	New England Biolabs
Dulbecco’s Modified Eagle Medium	Gibco
Fetal calf serum	Pan Biotech

L-Glutamine solution	Sigma
NEB2 buffer	New England Biolabs
NEB3 buffer	New England Biolabs
PBS	In-house
PBS-T	In-house
Penicillin/Streptomycin	In-house
Orange G loading dye	Thermo Fisher
Quick Ligation Buffer	New England Biolabs
TBS	In-house
TBS-T	In-house
Trypsin-EDTA (0.05% Trypsin)	In-house

2.1.3 Kits

Table 2-3 - Kits

Kit	Supplier
DNA Clean and Concentrator ™	Zymo Research
DNeasy® Blood & Tissue Kit	Qiagen
ECL™ Prime Western Blotting Reagents	Thermo Fisher
LightShift™ Chemiluminescent EMSA Kit	Thermo Fisher
Lipofectamine® 3000 Transfection Kit	Invitrogen
Lipofectamine ® LTX and Plus™ Reagents	Invitrogen
MEGAshortscript™ T7 High Yield Transcription Kit	Ambion
NE-PER™ Nuclear and Cytoplasmic Extraction Reagents	Thermo Fisher
QIAprep® Spin Miniprep Kit	Qiagen
QIAquick® Gel Extraction Kit	Quiagen
Qubit™ dsDNA BR Assay Kit	Invitrogen
Qubit™ dsDNA HS Assay Kit	Invitrogen
Quick-DNA Miniprep Kit	Zymo Research
Zymoclean™ Gel DNA Recovery Kit	Zymo Research
Zymoclean™ Large Fragment DNA Recovery Kit	Zymo Research
ZymoPure™ Plasmid Maxiprep Kit	Zymo Research

ZymoPure™ II Plasmid Midiprep Kit	Zymo Research

2.1.4 Other consumables

Table 2-4 – Other consumables

Consumable	Supplier
Biodyne™ Nylon Membranes	Thermo Fisher
Magnetic beads	Beckman Coulter
MultiScreen® Nylon Filter Plates	Merck
Nitrocellulose membrane	GE Healthcare Life Sciences
Pierce™ Protease Inhibitor Mini Tablets	Thermo Fisher

2.1.5 Equipment

Table 2-5 - Equipment

Equipment	Supplier
BD™ High Throughput Sampler	BD Biosciences
BD™ LSRII	BD Biosciences
BS™ LSRFortessa	BD Biosciences
C1000 Touch™ Thermal Cycler	Biorad
ECL Chemocam Imager	Intas
Gel iX20 Imager	Intas
M220 Focused Ultrasonicator	Covaris
MSC-Advantage™ Class II Biological Safety Cabinet	Thermo Fisher
NanoDrop™ 2000 Spectrophotometer	Peqlab
peqSTAR Thermocycler	Peqlab
Qubit® 2.0 Fluorometer	Invitrogen
T*i* Eclipse Inverted Microscope	Nikon

2.1.6 Antibodies and enzymes

Table 2-6 - Antibodies and enzymes

Enzyme	Supplier

Antibodies	
α-Sleeping beauty (goat)	R&D Systems, RRID: AB_622119
α-actin (mouse)	Thermo Fisher, RRID: AB_2223496
α-Cas9 (mouse)	Thermo Fisher, RRID:AB_2610639
α-goat-HRP (rabbit)	Sigma, RRID: AB_258425
α-goat-Alexa488	Thermo Fisher, RRID:AB_2534102
α-mouse-HRP (goat)	Thermo Fisher, RRID: AB_228313
Restriction enzymes	
AgeI	New England Biolabs
BamHI-HF	New England Biolabs
BbsI-HF	New England Biolabs
DpnI	New England Biolabs
EcoRI-HF	New England Biolabs
FseI	New England Biolabs
HindIII	New England Biolabs
NotI	New England Biolabs
SmaI	New England Biolabs
Other enzymes	
Antarctic phosphatase	New England Biolabs
Blunt/TA Master Mix	New England Biolabs
spCas9	PNA Bio
End Repair Enzyme Mix	New England Biolabs
Klenow exo-	New England Biolabs
NEBNext 2x PCR Master Mix	New England Biolabs
Q5 DNA polymerase	New England Biolabs
RNAse A	Thermo Fisher
T4 DNA ligase	New England Biolabs
T4 Polynucleotide kinase	New England Biolabs
T7 Endonuclease	New England Biolabs
Taq DNA polymerase	Thermo Fisher

2.1.7 Bacterial strains and eukaryotic cell lines

Table 2-7 - Bacterial strains and eukaryotic cell lines

Line	Source
Bacterial strains	
E. coli DH5α	Invitrogen
E. coli TOP10	Invitrogen
Eukaryotic cell lines	
HCT116	Lab stock
HeLa	Lab stock
HEK293T	Lab stock
HEK293T-TLR	Dr. Ralf Kühn, [435]

2.1.8 Plasmids

Plasmids with a name containing "sg*" contain a sgRNA cassette into which different guide sequences can be cloned and plasmids containing several sgRNAs were constructed, depending on the experiments. Plasmids that only differ in their guide sequence are not listed independently in this table. In the results section, plasmids containing guide sequences are marked as such, i.e. dCas9 refers to a plasmid containing dCas9 with no sgRNA and dCas9-sgAluY-1 refers to a plasmid expressing both dCas9 and the sgRNA sgAluY-1 The procedure for cloning of sgRNAs into these plasmids is detailed in section 2.2.1.4 and the sequences of the oligos used for cloning are listed in Table 2-11.

Table 2-8 - Plasmids

Name	Notes
pSpCas9(sg*)-2A-GFP (PX458)	Prof. Feng Zhang, Addgene #48138, [436]
pSpCas9(sg*)-2A-Puro (PX459)	Prof. Feng Zhang, Addgene #62988, [436]
pAC2-dual-dCas9VP48-sgExpression	Prof. Rudolf Jaenisch, Addgene #48236, [437]
pT7-SB100X	Dr. Zoltán Ivics
pT7-SB100X(K248R)	Dr. Zoltán Ivics
pDsRed-B23	[438]
pFv-SB10	Dr. Zoltán Ivics
pT2B/puro	Dr. Zoltán Ivics

pT2/puroDR3	Dr. Zoltán Ivics
pmaxGFP	Lonza
pU6-sg*	George Church Lab
pTALE_HPRT_L	Dr. Claudio Mussolino
pTALE_HPRT_R	Dr. Claudio Mussolino
pU6-sgRosa26_CBh_Cas9-T2A-BFP	Dr. Ralf Kühn, [435]
pTLR_repair_vector	Dr. Ralf Kühn, [435]
Constructed plasmids	
pCBh-Cas9-SB100X-sg*	See section 2.2.1.3
pCBh-Cas9-N57-sg*	See section 2.2.1.3
pCBh-Cas9-N123-sg*	See section 2.2.1.3
pCBh-dCas9-SB100X-sg*	See section 2.2.1.2
pCBh-dCas9-SB(K248R)-sg*	See section 2.2.1.2
pCBh-dCas9-N57-sg*	See section 2.2.1.2
pCBh-dCas9-N123-sg*	See section 2.2.1.2
pCBh-N57-dCas9-sg*	See section 2.2.1.2
pCBh-dCas9-sg*	See section 2.2.1.2
pT7-SB(K120A)	See section 2.2.12
pT7-SB(Q124A)	See section 2.2.12
pT7-SB(H127A)	See section 2.2.12
pT7-SB(H128A)	See section 2.2.12
pT7-SB(K129A)	See section 2.2.12
pT7-SB(R131A)	See section 2.2.12
pT7-SB(K156A)	See section 2.2.12
pT7-SB(R166A)	See section 2.2.12
pT7-SB(K186A)	See section 2.2.12
pT7-SB(H187A)	See section 2.2.12
pT7-SB(K208A)	See section 2.2.12
pT7-SB(N245A)	See section 2.2.12
pT7-SB(P247A)	See section 2.2.12
pT7-SB(K256A)	See section 2.2.12
pT7-SB(K259A)	See section 2.2.12
pT7-SB(Q271A)	See section 2.2.12
pT7-SB(R293A)	See section 2.2.12

pT7-SB(N296A)	See section 2.2.12
pT7-SB(K339A)	See section 2.2.12
pT7-Tat-SB100X	See section 2.2.1.5
pT7-Rev-SB100X	See section 2.2.1.5
pT7-p120-SB100X	See section 2.2.1.5
pT7-Rex-SB100X	See section 2.2.1.5
pT7-B23-SB100X	See section 2.2.1.6
dCas9-SB(R131A)-sg*	See section 2.2.1.2
dCas9-SB(R166A)-sg*	See section 2.2.1.2
dCas9-SB(K186A)-sg*	See section 2.2.1.2
dCas9-SB(N245A)-sg*	See section 2.2.1.2
dCas9-SB(Q271A)-sg*	See section 2.2.1.2
dCas9-SB(C42)-sg*	See section 2.2.1.2
pU6-sgRosa26_CBh_Cas9-N57-T2A-BFP	See section 2.2.1.7
pU6-sgRosa26_CBh_Cas9-N123-T2A-BFP	See section 2.2.1.7

2.1.9 Primers

Overhangs (non-binding regions) are underlined, restriction sites are **bold**, artificially introduced start or stop codons are doubly underlined, [Phos] indicates 5'-phosphorylation, filler sequences are in lowercase, replaced codons are *italic* in mutagenesis primers.

The primers T-bal_long_*BC** and SB20hmr_*BC** contain hexanucleotide barcodes (represented by NNNNNN in the sequence). These primers were used in 20 different variations to allow multiplexing during integration library analysis.

Table 2-9 - Primers

Primer	Sequence (5'→3')
	1. RNA-guided retargeting
L-SB100X_fwd1	ACCTGCTGTGGGCGGAGGCCCTAAGATGGGAAAATCAAAAGAAATCAGCCAAGAC
FseI-L-SB100X_fwd2	aatc**GGCCGGCC**AAACTGGGCGGAGGCGCACCTGCTGTGGGCGGAG
EcoRI-SB100X_rev	aatc**GAATTC**TAGTATTTGGTAGCATTGCCTTTAAATTGTTTAACTT
EcoRI-N57_rev	aatc**GAATTC**GCGGTATGACGGCTGCGTG

EcoRI-N123_rev	aatc**GAATTC**GAGCAGTGGCTTCTTCCTTGCTGAGTG
dCas9_FseI/EcoRI_seq_fwd	CGGATCGACCTGTCTCAGC
dCas9_FseI/EcoRI_seq_rev	GAGGGGCAAACAACAGATGG
AgeI-N57_fwd	aatc**ACCGGT**accATGGGAAAATCAAAAGAAATCAGCCAAGAC
AgeI-L-N57_rev1	CAGGTGCGCCTCCGCCCAGTTTGCGGTATGACGGCTGCG
AgeI-L-N57_rev2	aatc**ACCGGT**CTTAGGGCCTCCGCCCACAGCAGGTGCGCCTCCGC
dCas9_AgeI_seq_fwd	GGTATTAATGTTTAATTACCTGGAGCACCTG
dCas9_AgeI_seq_rev	CGCTTCGACCTTGCGCTTTTT
EcoRI-L-SB100X_fwd2	aatc**GAATTC**AAACTGGGCGGAGGCGCACCTGCTGTGGGCGGAG
TRC-F	CAAGGCTGTTAGAGAGATAATTGGA
HPRT_fwd	GTAGTCAGGGTGCAGGTCTC
HPRT_rev	AGAAGTGTCACCCTAGCCTG
T7-sgL1-1_fwd	TTAATACGACTCACTATAGGGGCGCATATTCTCACTCATAGG
T7-sgL1-2_fwd	TTAATACGACTCACTATAGGGGGGATTCCTTAGCGGTGTGACTGA
T7-sgL1-3_fwd	TTAATACGACTCACTATAGGGGGTATATACCCAGTAATGGGA
T7-sgTA1_fwd	TTAATACGACTCACTATAGGGGGTCTCCCTAATTCTAATTCA
T7-sgTA2_fwd	TTAATACGACTCACTATAGGGGGCTTCAGGATTATGCTGCAT
T7-sgTA3_fwd	TTAATACGACTCACTATAGGGGGATCCAGGGAGGTTAAGTAG
T7-sgAAVS1_fwd	TTAATACGACTCACTATAGGGGGCCACTAGGGACAGGAT
T7_rev	AAAAGCACCGACTCGGTGCC
HS4.1_fwd	TTGTAAGCCTTGTGGCAACC
HS4.1_rev	GCTTCCCTCTTACCTCTGCT
HS4.2_fwd	GTAGCAAACCTGCCCATCCT

HS4.2_rev	TCCGTACTAGGCATCAGGGG
HS8.1_fwd	TCACCCGCACTCATGGTCT
HS8.1_rev	GCCATCATATGGTAGACGGGG
HS8.2_fwd	GAAACAATGCCCGCCTCTTG
HS8.2_rev	GCAAAGACTGGCACTAGGGA
HS10.1_fwd	TGCCTATGCACTGAGAACAGC
HS10.1_rev	CACAAAACCATTCGTGAGGGG
HS10.2_fwd	CCCACCGAGAGATCAGGC
HS10.2_rev	TACTTGTTTGTCACAGCCCGT
AAVS1_fwd	TGCCCAAGGATGCTCTTTCC
AAVS1_rev	AGCACCAGGATCAGTGAAACG
SB20hmr	ACTTAAGTGTATGTAAACTTCCGACT
frag1_fwd	ATGGGAAAATCAAAAGAAATCAGCCAAGAC
frag1_rev	ACTCGTTTTACTGTGGATATAGATACTTTTGTACC
frag2_rev	TTGTGAAGATGCTGGAGGAAACAGG
frag2_rev	TGCACCAGTCCCTCCTGC
frag3_rev	GGGTGGCAGCATCATGTTGT
frag3_rev	CGCACACGCTTTTTCAGTTCT
frag4_rev	CTGACCTCAATCCTATAGAAAATTTGTGGG
frag4_rev	CTAGTATTTGGTAGCATTGCCTTTAAATTGTTTAACT
	2. SB mutagenesis primers
K120_fwd	AGCAAGGAAG*GCT*CCACTGCTCC
K120_rev	GAGTGGCCTTTCAGGTTATG
Q124_fwd	GCCACTGCTC*GCT*AACCGACATAAG
Q124_rev	TTCTTCCTTGCTGAGTGG
H127_fwd	CCAAAACCGA*GCT*AAGAAAGCCAGACTAC
H127_rev	AGCAGTGGCTTCTTCCTT
K128_fwd	AAACCGACAT*GCT*AAAGCCAGACTACG
K128_rev	TGGAGCAGTGGCTTCTTC
K129_fwd	CCGACATAAG*GCT*GCCAGACTACG
K129_rev	TTTTGGAGCAGTGGCTTC
R131_fwd	TAAGAAAGCC*GCT*CTACGGTTTGCAACTG
R131_rev	TGTCGGTTTTGGAGCAGT
K156_fwd	TGATGAAACA*GCT*ATAGAACTGTTTGGC

K156_rev	GACCAGAGGACATTTCTC
R166_fwd	TAATGACCAT*GCT*TATGTTTGGAGGAAGAAG
R166_rev	TGGCCAAACAGTTCTATTTTTG
K186_fwd	CCCAACCGTG*GCT*CACGGGGGTG
K186_rev	ATGGTGTTCTTCGGCTTG
H187_fwd	AACCGTGAAG*GCT*GGGGGTGGCAG
H187_rev	GGGATGGTGTTCTTCGGC
K208_fwd	TGCACTTCAC*GCT*ATAGATGGCATCATGGACGC
K208_rev	CCAGTCCCTCCTGCAGCA
N245_fwd	CCAACACGAC*GCT*GACCCCAAGC
N245_rev	AAGACCCATTTGCGACCA
P247_fwd	CGACAATGAC*GCT*AAGCATACTTC
P247_rev	TGTTGGAAGACCCATTTG
K256_fwd	AGTTGTGGCA*GCT*TGGCTTAAGGACAAC
K256_rev	TTGGAAGTATGCTTGGGG
K259_fwd	AAAATGGCTT*GCT*GACAACAAAGTCAAGG
K259_rev	GCCACAACTTTGGAAGTATG
Q271_fwd	GTGGCCATCA*GCT*AGCCCTGACC
Q271_rev	TCCAATACCTTGACTTTG
R293_fwd	GCGAGCAAGG*GCT*CCTACAAACC
R293_rev	ACACGCTTTTTCAGTTCTG
N296_fwd	GAGGCCTACA*GCT*CTGACTCAGTTACACC
N296_rev	CTTGCTCGCACACGCTTT
K339_fwd	CAATGCTACC*GCT*TACTAGGGGCC
K339_rev	CCTTTAAATTGTTTAACTTGG
	3. Ribosomal targeting
Tat-SB100X_fwd	AGAGAAGAAGAGCCCACCAGAACGGAAAATCAAAAGA AATCAGCCAAGACCTC
Tat-SB100X_rev	GTCTTCTCTTCTTTCTGCCCATGGTACTAGTCCCTATAGT GAGTCGTAT
Rev-SB100X_fwd	AGAAGAAGATGGAGAGAGAGACAGAGACAGGGAAAA TCAAAAGAAATCAGCCAAGACC
Rev-SB100X_rev	TCTGTTTCTTCTGGCCTGTCTGCCCATGGTACTAGTCCC TATAGTGAGTCGTAT

p120-SB100X_fwd	AGAAAGAGAGCCGCCAAGAGAAGACTGGGAAAATCAAAAGAAATCAGCCAAGACCTC
p120-SB100X_rev	GGCTCTGCTGCTCAGTCTCTTGCTGCCCATGGTACTAGTCCCTATAGTGAGTCGTAT
Rex-SB100X_fwd	AGAAGCCAGAGAAAGAGACCTCCCACCCCTGGAAAATCAAAAGAAATCAGCCAAGACCTC
Rex-SB100X_rev	TCTGGGTCTTCTTCTGGTCTTGGGGCCCATGGTACTAGTCCCTATAGTGAGTCGTAT
L-SB100X_fwd1_alt	ACCTGCTGTGGGCGGAGGCCCTAA*G*GGAAAATCAAAAGAAATCAGCCAAG
NotI-SBT7_rev	aatc**GCGGCCGC**AGTCCCTATAGTGAGTCGTATTAATTCCTTCCG
NotI-S-B23_fwd	aatc**GCGGCCGC**ATGTATGGAAGATTCGATGGACATGGACATGA
FseI-B23_rev	aatc**GGCCGGCC**tAAGAGACTTCCTCCACTGCCAGA

4. HDR enhancement

HindIII-L-SB100X_fwd2	aatc**AAGCTT**AAACTGGGCGGAGGCGCACCTGCTGTGGGCGGAG
HindIII-N57_rev	aatc**AAGCTT**GCGGTATGACGGCTGCGTG
HindIII-N123_rev	aatc**AAGCTT**GAGCAGTGGCTTCTTCCTTGCTGAGTG
TLR_donor_fwd	AGCTTGATTAAGCCGCCACC
TLR_donor_rev	CCCCAGCTGGTTCTTTCCG
SBBS_TLR_donor_fwd	aatcTACAGTTGAAGTCGGAAGTTTACATACACTTAA*G*AGCTTGATTAAGCCGCCACC
SBBS_TLR_donor_rev	aatcCTTAAGTGTATGTAAACTTCCGACTTCAACTGTACCCCAGCTGGTTCTTTCCG

5. Other primers

PE_nest_*BC**	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAG
pUC3	CGATTAAGTTGGGTAACGCCAGGG
pUC5	TCTTTCCTGCGTTATCCCCTGATTC

SB20hmr_*BC**	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCT ACACGACGCTCTTCCGATCTNNNNNNACTTAAGTGTAT GTAAACTTCCGACT
T-bal_long	CTTGTGTCATGCACAAAGTAGATGTCCTAACTGACT
TS_linker	GTAATACGACTCACTATAGGGC

2.1.10 Other oligonucleotides

For oligonucleotides used for sgRNA cloning (oligos with names beginning with sg), *italic* nucleotides are cloning overhangs and the remaining 20 nt are the actual guide sequence. The top oligo defines the actual RNA sequence (identical to the non-target strand). The first (PAM-distal) nucleotide of the target sequence is always replaced by a G in the sgRNA sequence to improve expression from the U6 promoter. Note that the term 'targeting sequence' of a sgRNA refers to the original sequence targeted in the genome and not the sequence with the replaced first nucleotide.

Table 2-10 - Oligonucleotides

Oligonucleotide	**Sequence (5'→3')**
FseI-STOP-EcoRI_top	[Phos]CCTGAG
FseI-STOP-EcoRI_btm	[Phos]AATTCTCAGGCCGG
N57_EMSA_top	TACAGTTGAAGTCGGAAGTTTACATACACTTAAG
N57_EMSA_btm	CTTAAGTGTATGTAAACTTCCGACTTCAACTGTA
sgAluY-1_top	*CACC*GTCCCAAAGTGCTGGGATTAC
sgAluY-1_btm	*AAAC*GTAATCCCAGCACTTTGGGAC
sgAluY-2_top	*CACC*GCCTGTAATCCCAGCACTTT
sgAluY-2_btm	*AAAC*AAAGTGCTGGGATTACAGGC
sgAluY-3_top	*CACC*GTTTGTATTTTTAGTAGAGA
sgAluY-3_btm	*AAAC*TCTCTACTAAAAATACAAAC
sgL1-1_top	*CACC*CGCATATTCTCACTCATAGG
sgL1-1_btm	*AAAC*CCTATGAGTGAGAATATGCG
sgL1-2_top	*CACC*GGATTCCTTAGCGGTGTGACTGA
sgL1-2_btm	*AAAC*TCAGTCACACCGCTAAGGAATCC
sgL1-3_top	*CACC*GTATATACCCAGTAATGGGA
sgL1-3_btm	*AAAC*TCCCATTACTGGGTATATAC
sgHPRT-0_top	*CACC*GAAGTAATTCACTTACAGTC
sgHPRT-0_btm	*AAAC*GACTGTAAGTGAATTACTTC

sgHPRT-1_top	*CACC*GCTTGCTCGAGATGTGATGA
sgHPRT-1_btm	*AAAC*TCATCACATCTCGAGCAAGC
sgHPRT-2_top	*CACC*GAAATTCTTTGCTGACCTGC
sgHPRT-2_btm	*AAAC*GCAGGTCAGCAAAGAATTTC
sgHPRT-3_top	*CACC*GTGATAAAATCTACAGTCAT
sgHPRT-3_btm	*AAAC*ATGACTGTAGATTTTATCAC
sgHS4.1_top	*CACC*GCTTTTTACAGTTTTGGTC
sgHS4.1_btm	*AAAC*GACCAAAACTGTAAAAAGC
sgHS4.2_top	*CACC*GATCAATCTAAGTGTACGT
sgHS4.2_btm	*AAAC*ACGTACACTTAGATTGATC
sgHS8.1_top	*CACC*GCCTTTCGACCATGCAAAG
sgHS8.1_btm	*AAAC*CTTTGCATGGTCGAAAGGC
sgHS8.2_top	*CACC*GACCACAGTGCGAATCCTG
sgHS8.2_btm	*AAAC*CAGGATTCGCACTGTGGTC
sgHS10.1_top	*CACC*GCTGCTCTACAACTAGCAG
sgHS10.1_btm	*AAAC*CTGCTAGTTGTAGAGCAGC
sgHS10.2_top	*CACC*GTGGGCTAATAAACACTAT
sgHS10.2_btm	*AAAC*ATAGTGTTTATTAGCCCAC
sgTA1_top	*CACC*GTCTCCCTAATTCTAATTCA
sgTA1_btm	*AAAC*TGAATTAGAATTAGGGAGAC
sgTA2_top	*CACC*CCTTCAGGATTATGCTGCAT
sgTA2_btm	*AAAC*ATGCAGCATAATCCTGAAGG
sgTA3_top	*CACC*GATCCAGGGAGGTTAAGTAG
sgTA3_btm	*AAAC*CTACTTAACCTCCCTGGATC
sgAAVS1_top	*CACC*GGGGCCACTAGGGACAGGAT
sgAAVS1_btm	*AAAC*ATCCTGTCCCTAGTGGCCCC

2.1.11 Software

- ImageJ was used for processing of IF images and the ImageJ plugin Colony Counter was used for counting cell colonies and processing of IF images.
- SnapGene was used for *in silico* cloning, management of DNA and protein sequences as well as generation of alignments.
- Inkscape was used for generation of vector graphics and GIMP was used for generation of raster graphics.
- R, Microsoft Excel and GraphPad QuickCalcs were used for statistical testing.

- The online tool CCTop CRISPR/Cas9 target online predictor[439] was used for design of sgRNAs.
- The online tool Primer3[440] was used for the design of some PCR primers.

2.1.12 Services

- Sanger sequencing was performed by GATC / Eurofins Genomics.
- Integration library deep sequencing was performed by GeneWiz, or in-house.
- Primers and other DNA oligos were supplied from Eurofins Genomics.

2.2 Methods

2.2.1 Plasmid construction

2.2.1.1 Molecular cloning methods

PCRs for cloning substrates were performed with Q5 DNA polymerase; other PCRs were performed with Q5 or Taq polymerase. Restriction digestions were performed in the buffers and at temperatures specified by the suppliers. Ligations were performed using T4 DNA ligase over night at 16°C or using T4 DNA ligase with Quick Ligation Buffer for 15 min at room temperature. Column purifications were done with Zymo DNA Clean and Concentrator. Gel extractions were done with Zymo Gel DNA Extraction Kit, Zymo Large Fragment DNA Recovery Kit or Quiagen QIAquick Gel Extraction Kit. For amplification, plasmids were transformed into *E. coli* DH5α or TOP10. DNA concentration measurements were done using a NanoDrop ND1000 or a Qubit 2.0 using dsDNA BR or HS kits.

2.2.1.2 dCas9-SB100X, dCas9-SB10 dCas9-N57, dCas9-N123, N57-dCas9, dCas9

Expression plasmids for fusion proteins of dCas9 with SB100X, N57 and N123 were generated by replacing the VP48 domain in the vector pAC2-dual-dCas9VP48-sgExpression with SB100X, N57 or N123. In order to achieve the necessary flexibility of the domains, a 14 amino acid linker with the sequence KLGGGAPAVGGGPK had to be introduced between dCas9 and the fusion partners. This linker was added to the inserts via overhangs on PCR primers. Due to the length of the linker sequence (42 bp) it had to be added in two consecutive PCRs. The full-length SB100X insert containing the linker sequence and restriction sites was generated by amplifying the SB100X sequence from the plasmid pT7-SB100X with the primers L-SB100X_fwd1 and EcoRI-SB100X_rev, then re-amplifying the PCR products with primers FseI-L-SB100X_fwd2 and EcoRI-SB100X_rev. To generate inserts for cloning of N57 and N123, analogous PCRs were performed, replacing the primer EcoRI-SB100X_rev with EcoRI-N57_rev and EcoRI-N123_rev, respectively. These inserts were used to replace the VP48

sequence in pAC2-dual-dCas9VP48-sgExpression by FseI/EcoRI digestion and ligation. The generated plasmids pCBh-dCas9-SB100X, pCBh-dCas9-N57 and pCBh-dCas9-N123 were verified by Sanger sequencing using the primers dCas9_FseI/EcoRI_seq_fwd and dCas9_FseI/EcoRI_seq_rev.

A dCas9 expression plasmid was generated by replacing the VP48 sequence of pAC2-dual-dCas9VP48-sgExpression with a stop codon. To this end, the two phosphorylated oligonucleotides FseI-STOP-EcoRI_top and FseI-STOP-EcoRI_btm were annealed by heating 50 µl of a 200 nmol/ml solution to 40°C and cooling to 4°C over 1 h. Annealed oligos were ligated into the FseI/EcoRI-digested vector as described above. The ligation product was sequenced with the primer dCas9_FseI/EcoRI_seq_fwd.

In order to generate a N57-dCas9 fusion, an N57 insert was generated by performing two consecutive PCRs on pT7-SB100X, as described above. The primers used were AgeI-N57_fwd and AgeI-L-N57_rev1 for the first PCR and AgeI-N57_fwd and AgeI-L-N57_rev2 for the second PCR. The generated insert was ligated into the AgeI site of the newly generated dCas9 expression plasmid by the methods described above. Correct construction of the N57-dCas9 fusion plasmid was verified by sequencing with dCas9_AgeI_seq_fwd and dCas9_AgeI_seq_rev.

A dCas9-SB10 fusion was generated using the same procedure as described above for dCas9-SB100X, using pFv-SB10 as a template for the insert PCRs instead of pT7-SB100X.

2.2.1.3 Cas9-N57 and Cas9-N123

Expression plasmids for Cas9-N57 and Cas9-N123 fusion proteins were generated by replacing the GFP sequence in the plasmid PX458 with N57 or N123. As described in section 2.2.1.2., a linker had to be introduced by two consecutive PCRs. The primer pairs used for the N57 insert were L-SB100X_fwd and EcoRI-N57_rev, followed by EcoRI-L-SB100X_fwd2 and EcoRI-N57_rev. For the N123 insert, the primer EcoRI-N123_rev was used instead of EcoRI-N57_rev. The inserts were used to replace the GFP sequence by EcoRI digestion and ligation and constructed plasmids were verified by Sanger sequencing with dCas9_FseI/EcoRI_seq_fwd and dCas9_FseI/EcoRI_seq_rev.

2.2.1.4 Cloning of guide sequences into vectors containing sgRNA scaffolds

All Cas9-based vectors (e.g. PX458, pAC2-dual-dCas9VP48-sgExpression and derived vectors) contain a common sgRNA-expression cassette that features a human U6 promoter and two BbsI sites between which the 20 nt sgRNA binding sequence can be cloned upstream of

the sgRNA scaffold. To generate the inserts, two ssDNA oligos were diluted to 10µM in NEB T4 DNA buffer and phosphorylated with T4 Polynucleotide Kinase for 30 min at 37°C. Annealing was subsequently by heating to 95°C for five minutes, then ramping down to 25°C at -1°C/min. The annealed oligos were then directly ligated into BbsI-digested Cas9-based vectors. Insertion of the complete sgRNA sequence was verified by Sanger sequencing with the primer TRC-F.

2.2.1.5 NoLS-SB100X fusions

Four different NoLS sequences were introduced into pT7-SB100X at the N-terminus of the SB sequence: Tat (MGRKKRRQRRRAHQ)[441], Rev (MGRQARRNRRRRWRERQRQ)[442], p120 (MGSKRLSSRARKRAAKRRLG)[443], Rex (MGPKTRRRPRRSQRKRPPTP)[444]. This was achieved by amplifying the entire pT7-SB100X plasmid with primers that contained one half of the NoLS sequence as an overhang on each primer. The primer pairs were Tat-SB100X_fwd and Tat-SB100X_rev, Rev-SB100X_fwd and Rev-SB100X_rev, p120-SB100X_fwd and p120-SB100X_rev and Rex-SB100X_fwd and Rex-SB100X_rev, respectively. After PCR amplification, the product was digested with DpnI to remove the original templates from the samples. PCR products were column purified and circularized with T4 DNA ligase (500 ng of DNA in 100 µl). Insertion of the NoLS sequences was verified by Sanger sequencing using the primer EcoRI-N57_rev.

2.2.1.6 B23-SB100X

A B23-SB100X fusion was generated by amplifying the plasmid pT7-SB100X in two consecutive PCRs, first with the primers L-SB100X_fwd1_alt and NotI-SBT7_rev, then with FseI-L-SB100X_fwd2 and NotI-SBT7_rev. This process removes the SB100X start codon and inserts a linker sequence at the new N-terminus of the SB100X sequence. For the insert, the B23 sequence was amplified from the plasmid pDsRed-B23 using NotI-S-B23_fwd and FseI-B23_rev, which also introduces a start codon at the 5'-end of the B23 sequence. The insert is then introduced into the vector using FseI/NotI digestion and ligation. Correct insertion was verified by Sanger sequencing using the primer EcoRI-S-N57_rev.

2.2.1.7 Cas9-N57 and Cas9-N123 TLR expression plasmids

Plasmids expressing Cas9-N57/Cas9-N123 together with BFP and a sgRNA targeting the TLR target were generated by amplifying the N57/N123 inserts with a two-round PCR as described above (L-SB100X_fwd1 + HindIII-N57_rev, then HindIII-L-SB100X_fwd2 + HindIII-N57 for Cas9-N57; L-SB100X_fwd1 + HindIII-N123_rev, then HindIII-L-SB100X_fwd2 + HindIII-N123_rev for Cas9-N123), then digesting the generated inserts as well as the plasmid pU6-

sgRosa26_CBh_Cas9-T2A-BFP with HindIII and cloning the SB subdomains into the digested plasmid. Correct construction was verified by Sanger sequencing using the primer dCas9_FseI/EcoRI_seq_fwd.

2.2.2 Cell culture and transfection

Unless otherwise specified, all eukaryotic cell lines were cultured in Dulbecco's modified Eagle medium (DMEM), supplemented with 10% FCS, Penicillin/Streptomycin and 2 mM L-Glutamine. Cells were cultured at 37°C and 5% CO_2.

For antibiotic selection, antibiotics were added to the medium described above at the following concentrations: 1 µg/ml puromycin, 500 mg/ml neomycin and 50 mg/ml 6-TG. An untransfected control was included in every selection experiment to determine the time point at which selection is completed.

Transfections were performed in DMEM supplemented with 5% FCS (without antibiotics or L-Glutamine). As transfection reagents, Lipofectamine 3000, Lipofectamine LTX or PEI (PEI) were used. Lipofectamine reagents were used according to manufacturer's instructions. For transfections with PEI, DNA was diluted in Opti-MEM and 2 µg of PEI were added per µg of DNA. After 30 min at RT, PEI-DNA complexes were added to the cells. Unless otherwise specified, transfections were done in a 6-, 12- or 24-well format and total DNA amounts per well were 2 µg, 1 µg or 500 ng, respectively.

2.2.3 Selection-based transposition assays

In order to test the transpositional activity of a transposase variant or compare transpositional activity under different conditions, expression plasmids encoding the variant were co-transfected with a compatible transposase plasmid containing an antibiotic resistance marker. Unless otherwise specified, pT2/puro was used as a transposon plasmid. Cells were transfected in 6-, 12- or 24-well format and plated onto 10 cm dishes 48 h or 72 h after transfection. Cells were grown in selection media for 10 to 14 days, then fixed with 4% PFA in PBS for two hours. Cells were then stained in methylene blue solution overnight. Plates were scanned and colonies were counted using ImageJ's Colony Counter plugin at the following settings: size > 50 px, circularity > 0.75.

2.2.4 Western Blot

For Western Blots, HEK293T cells were transfected in 10 cm dishes and cultured for 48 h. Cells were then washed with PBS and lysed with 0.6 ml of RIPA buffer (150 mM NaCl, 0.1% Triton X-100, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris-HCl pH 8.0 and one PI tablet per

10 ml) for 15 min at 4°C. Lysates were sheared by passing through a 21-gauge needle, incubated 30 min on ice and centrifuged at 10,000 g for 10 min at 4°C to remove cellular debris. Protein concentrations were determined using Coomassie Protein Assay Reagent (Thermo). Lysate containing 50 µg of total protein was mixed 1:1 with loading buffer (125 mM Tris-HCl, pH 8.0, 4% SDS. 20% glycerol, 200 ng/ml bromophenol blue, 1% β-mercaptoethanol) and heated to 99°C for 2 min. Proteins were separated on discontinuous acrylamide gels (acrylamide percentages of the separating gels varied between 8% and 12.5%, depending on the size of the protein of interest) at 200V. Nitrocellulose membranes were activated by soaking in water for 10 min, then in transfer buffer (25 mM Tris, 200 mM glycine, 20% methanol) for another 10 min. Proteins were transferred from the gel to the membrane in a blotting tank at 100V for 1h at 4°C. Membranes were blocked by incubation in 5% milk in TBS-T overnight at 4°C. Membranes were incubated with primary antibodies in 5% milk in TBS-T for 2-4h. Used antibody dilutions are listed in Table 2-11. Membranes were then washed with TBS-T and incubated with secondary antibodies for 1h. After secondary antibody incubation, membranes were washed with TBS-T, developed with ECL™ Prime Western Blotting reagents and imaged.

Table 2-11 – Concentrations for WB antibodies

Antibody	Source organism	Dilution for WB
α-Sleeping Beauty	Goat	1:500
α-actin	Mouse	1:5000
α-goat	Rabbit	1:10000
α-mouse	Goat	1:10000

2.2.5 Electrophoretic mobility shift assay (EMSA)

An EMSA was performed to test the ability of fusion proteins containing N57 domains to bind to their recognition DNA sequence from the SB ITR. The EMSA was performed using a DNA oligonucleotide corresponding to the 14DR from either transposon end with the sequence TACAGTTGAAGTCGGAAGTTTACATACACTTAAG. To generate protein extracts, HeLa cells were transfected with the targeting factors and nuclear extracts were generated two days after transfection using NE-PER™ Nuclear and Cytoplasmic Extraction Reagents (Thermo Fisher). Approximate concentrations of the targeting factors were determined by a dot blot experiment using an α-Cas9 antibody and concentrations were adjusted to be similar between samples. A bacterial extract of N57 was used as a positive control. The EMSA was performed using a LightShift™ Chemiluminescent EMSA Kit (Thermo Fisher) according to the

manufacturer's instructions, with 10 µg of total protein input for the nuclear extracts and 2.5 µg of total protein input for the positive control.

2.2.6 Generation of integration libraries

In order to generate SB insertions, HeLa cells were transfected in 10 cm dishes with combinations of transposon plasmids, transposase expression plasmids and, where applicable, adapter protein expression plasmids. After 48 h, cells were re-seeded in 15 cm dishes and selected with the appropriate antibiotic (puromycin or neomycin) for 2-3 weeks. Genomic DNA was prepared from the cells after the end of selection. Generation of integration libraries is a multi-step process detailed in Figure 3.4.

Sonication

10-50 µg of genomic DNA were isolated from an agarose gel using a Zymo Large Fragment DNA Recovery Kit. Recovered DNA was sheared to an average length of 600 bp using a Covaris M220 Focused Ultrasonicator at the following settings: Peak Power 50.0, Duty Factor 5.0, Cycles/Burst 200 and Duration 80 s. Sonicated DNA was captured with magnetic beads, washed with EtOH and eluted in H_2O.

End repair, dA-tailing and ligation

End repair, dA-tailing and ligation were performed in three separate reactions. After each reaction, DNA was captured with magnetic beads, washed with EtOH and eluted in H_2O. End repair was performed using NEB End Repair Enzyme Mix in for 30 min at RT. The dA-tailing reaction was performed using Klenow ex- in dA-tailing buffer (NEB) for 30 min at RT. Linkers were generated by annealing the oligos TruSeq_linker_top and TruSeq_linker_btm. Ligation was performed by adding linkers to DNA at a concentration of 500 nM in Blunt/TA Master Mix (NEB) and incubating for 10 min at RT.

Nested PCR and size selection

Genome-transposon junctions were amplified by a nested PCR using the following primer pairs: T-bal_long and TS_linker, then PE_nest_*BC* and SB20hmr_*BC*. The second set of primer contains short *BC* (barcode) sequences. These 6 nt sequences allow the later identification of individual amplicons, enabling several samples to be processed in parallel during sequencing. The thermocycler program used for the first PCR was: 98°C for 30 s; 10 cycles of 98°C for 10 s, 72°C for 40 s; 20 cycles of 98°C for 10 s, 62°C for 30 s, 72°C for 30 s; 72°C for 5 min. The thermocycler program used for the second PCR was: 98°C for 30 s, then 20 cycles of 64°C for 30 s, 72°C for 30 s, then 72°C for 5 min. Products of the second PCR were run on an ultrapure

agarose gel and fragments with sizes between 200 and 500 bp were excised and purified from the gel.

2.2.7 Integration site sequencing and analysis

Integration libraries were sequenced using the MiSeq platform. The raw Illumina reads were processed in the R environment [445] as follows: the transposon-specific primer sequences were searched and removed and PCR specificity was controlled by checking for the presence of transposon end sequences downstream of the primer. The resulting reads were subjected to adapter-, quality-, and minimum-length-trimming by the *fastp* algorithm [446] using these settings: *adapter_sequence =AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC --cut_right --cut_window_size 4 --cut_mean_quality 20 --length_required 28*. The reads were then mapped to the hg38 human genome assembly using Bowtie2[447] with the *--very-fast* parameter in *--local* mode. The 'unambiguity' of the mapped insertion site positions were controlled by filtering the sam files using SAMtools[448] with the *samtools view –q* 10 setting. Since the mapping allowed for mismatches the insertion sites within 5 nucleotide windows were reduced to the one supported by the highest number of reads. Any genomic insertion position was considered valid if supported by at least five independent reads. Insertion site logos were calculated and plotted with the SeqLogo package. The frequencies of insertions around the sgRNA target sequences were displayed by the genomation package[449]. Probability values for nucleosome occupancy in the vicinity of *Alu*Y targets and non-targeted insertion sites were calculated with a previously published algorithm[450].

2.2.8 PCR-based insertion site analysis

In some targeting experiments, integrations were analyzed via site-specific PCRs rather than by generating a genome-wide integration site library. In these experiments, cells were transfected with transposase fusion (including sgRNAs) and transposon plasmids and cultured for >1 week, and gDNA was isolated from these cells, as described in section 2.2.6. Isolated gRNA was then used as template in PCR reactions using a primer that binds both transposon ends, facing outwards (SB20hmr) and a primer binding in the genome either upstream or downstream of the target site. This configuration should allow amplification of transposon-genome junctions for insertions that occurred close to the target site (the distance being defined by the extension time of the PCR reaction). The primers used for each of the tested target sites are listed in Table 2-11.

Table 2-12 – Primers for PCR-based insertion screening

Target site	Fwd primer	Rev primer
HS4	HS4.1_fwd	HS4.1_rev
HS8	HS8.2_fwd	HS8.2_rev
HS10	HS10.2_fwd	HS10.2_rev
HPRT	HPRT_fwd	HPRT_rev
AAVS1	AAVS1_fwd	AAVS1_rev

Because no positive controls (cell lines with SB integrations at the targeted loci) were available, two different gDNA (75 ng or 750 ng) template amounts and two different annealing temperatures (58°C and 66°C) were tested. PCRs were performed with Q5 DNA polymerase and for 25 cycles, with an extension time of 2.5 min.

2.2.9 *In vitro* digestion with Cas9

Some sgRNAs were validated by *in vitro* digestion of target DNA with purified Cas9 and *in vitro* transcribed sgRNAs. For *in vitro* transcription, sgRNA sequences were amplified from sgRNA expression plasmids using a primer pair that introduced a T7 promoter upstream of the sgRNA sequence (T7-sg*_fwd and T7_rev). The amplified DNA was column purified and *in vitro* transcribed using an Ambion® MEGAshortscript™ T7 Kit. Digestion reactions were performed in NEB3 buffer with 1x BSA and equal amounts of *in vitro* transcribed sgRNA and purified Cas9 protein (between 300 ng and 1 µg of each) in reaction volumes of 10 or 20 µl. The amount of target DNA varied between 100 ng (for short DNA fragments) and 1 µg (for gDNA). Samples were incubated for 1.5 h (plasmids or short DNA) or overnight (gDNA). After digestion, 4 µg of RNAse A were added and samples were incubated for an additional 15 min to remove input RNA.

2.2.10 T7 Endonuclease assay

T7 assays were performed in order to test the activity of sgRNAs *in vivo*. HEK293T were transfected with a PX459-based plasmid containing the sgRNA in question. After 36 h, cells were selected with puromycin for 36 h, or until all cells in an untransfected control had died. After selection, gDNA was prepared from remaining cells and the relevant target sites were amplified by PCR (with primers HS4.1_fwd and HS4.1_rev, HS4.2_fwd and HS4.2_rev, HS8.1_fwd and HS8.1_rev, HS8.2_fwd, HS8.2_rev, HS10.1_fwd and HS10.1_rev, HS10.2_fwd and HS10.2_rev). PCR products were column purified and eluted in NEB2 buffer. Heteroduplexes were formed using the following thermocycler program: 95°C for 10 min; 95°C to 85°C at -2°C/s; 85°C for 1 min; 85°C to 75°C at -0.3°C/s; 75°C for 1 min; 75°C to 65°C at -

0.3°C/s; 65°C for 1 min; 65°C to 55°C at -0.3°C/s; 55°C for 1 min; 55°C to 45°C for -0.3°C/s; 45°C for 1 min; 45°C to 35°C at -0.3°C/s; 35°C for 1 min; 35°C to 25°C at -0.3°C/s; 25°C for 1 min. 300 ng of each re-annealed amplicon were digested in 12 µl of NEB2 buffer with T7 endonuclease at 37°C for 20 min. The reaction was stopped by adding 2.4 µl of 10x Orange G Loading Dye and put on ice. Bands were then visualized by agarose gel electrophoresis. Efficiencies of the sgRNAs were estimated by measuring the signal intensity for each lane around the main band produced by the PCR and in a region comprising the cleaved products and subtracting the background intensity for each measurement. Cleavage efficiency was defined as the intensity value of the lower region of the lane divided by the total intensity (upper and lower part).

2.2.11 TIDE assay

A TIDE assay was performed in order to quantify the efficiency of the sgRNA sgHPRT-0. To generate editing events, HeLa cells were transfected with PX459-sgHPRT-0 or PX459 as a reference. Puro selection was started 36 h after transfection and genomic DNA was isolated 36 h after starting the selection. The target locus was amplified from both samples using the primers HPRT_fwd and HPRT_rev. The PCR products were column purified and Sanger sequenced. Sequences were uploaded to the TIDE webtool[451] and analyzed using standard parameters.

2.2.12 *Sleeping Beauty* mutagenesis

To mutagenize the SB transposase, the plasmid pT7-SB100X was amplified with the primers listed in Table 2-9, section 2 using Q5 DNA polymerase. 1 µl of DpnI was added to each PCR reaction and the samples were incubated for 1 h at 37°C. Samples were column purified and 50 ng of each amplicon were diluted in 50 µl of T4 DNA ligase buffer. Amplicons were phosphorylated for 1 h at 37°C after addition of 0.5 µl of T4 PNK. For circularization, 1 µl of T4 DNA ligase was added to each sample and they were incubated overnight at 16°C. Circularized amplicons were transformed into *E.coli* DH5α and the SB100X sequences of the plasmids were Sanger sequenced to verify successful mutagenesis.

2.2.13 Assembly of the SB mutant library

Four fragments of the SB sequence were PCR amplified from the SB mutants generated in section 2.2.12. Each fragment was amplified from all SB mutants containing the mutation in the respective fragment (Table 2-13). PCR amplicons of each fragment were pooled in equal amounts and sonicated to an average size of < 100 bp using a Covaris M220 ultrasonicator at the following settings peak power: 50, duty factor: 20, cycles/burst: 200, time 2500 s. For

assembly, pooled fragments were mixed in a ratio based on their sizes – 1:1.12:0.96:0.64 for fragments 1-4, respectively. Total DNA concentration of the mixture was adjusted to 20 ng/µl and the SB sequence was reassembled by a primerless PCR program: 1 min at 94°C; 45 cycles of 30 s at 94°C, 30 s at 50°C, 30 s at 72°C; 5 min at 72°C. Reassembly was verified by agarose gel electrophoresis. 1 µl of the reassembly product was used as a template for a PCR with primers L-SB100X_fwd1 and EcoRI-SB100X_rev. The PCR product was gel extracted and amplified with SB100X_fwd2 and EcoRI-SB100X_rev. The second PCR product was gel extracted and cloned into an EcoRI/FseI digested pCBh-dCas9 vector. Ligated plasmids were transformed into *E.coli* DH5α and a few plasmids of each library were picked and sequenced using primer dCas9_FseI/EcoRI_seq_rev in order to estimate the average number of mutations in the library and overall integrity of the sequence.

Table 2-13 – List of fragment PCR primers and templates for SB mutant library generation

	Fragment 1	**Fragment 2**	**Fragment 3**	**Fragment 4**
Primers	frag1_fwd frag1_rev	frag2_fwd frag2_rev	frag3_fwd frag3_rev	frag4_fwd frag4_rev
Templates	pT7-SB100X	pT7-SB(K120A) pT7-SB(Q124A) pT7-SB(H127A) pT7-SB(K128A) pT7-SB(K129A) pT7-SB(R131A) pT7-SB(K156A) pT7-SB(R166A) pT7-SB(K186A) pT7-SB(H187A)	pT7-SB(K208A) pT7-SB(N245A) pT7-SB(P247A) pT7-SB(K256A) pT7-SB(K259A) pT7-SB(Q271A)	pT7-SB(R293A) pT7-SB(N296A) pT7-SB(K339A)

2.2.14 Screening of the SB mutant library

In order to test whether some of the generated SB mutants displayed the desired phenotype – i.e. reduced activity which can be (partially) rescued by a fused DNA domain - 864 individual plasmids were transfected into HeLa cells in combination with a GFP-tagged transposon and a plasmid expressing either the sgRNA sgAluY-1 or no sgRNA. Transfections were done in 96-well plates (10^4 cells/well) with Lipofectamine LTX. Cells were transfected with 25 ng of each transposase plasmid, 75 ng of the transposon plasmid pT2HB/GFP and 125 ng of either pU6-

sg- or pU6-sgAluY-1. After transfection, cells were split 1:4 every other day until background fluorescence was not detectable in a sample transfected with only the transposon plasmid (ca. 2 weeks). Cells were then prepared for HT FACS analysis.

2.2.15 High-throughput FACS analysis

In order to prepare cells for FACS analysis in a 96-well format, cells were trypsinized by addition of 20 µl of Trypsin-EDTA and the reaction was stopped by the addition of 80 µl of DMEM after ca. 10 min. Cells were passed through a 40 µm nylon filter plate (Merck MultiScreen®) and the receiver plate was centrifuged for 5 min at 1200 g. After removal of the supernatant, cells were fixed for 10 min with 2% PFA in PBS (50 µl) and the reaction was stopped by addition of 100 µl of FACS buffer (5% FCS and 1% penicillin/streptomycin in PBS). Cells were centrifuged for 5 min, supernatant was removed and cells were subsequently resuspended in 150 µl of FACS buffer. Plates were put on ice until measurement. Measurement was performed using a BD LSRFortessa™ flow cytometer and a BD High Throughput Sampler.

2.2.16 Immunofluorescence microscopy

For IF analysis of subcellular localization of SB proteins, HeLa cells were seeded in a 6-well format on cover glasses and transfected with 100 ng of expression plasmids for the relevant proteins. After 48 h, the medium was removed and cells were rinsed with PBS, then fixed with 4% PFA in PBS for 30 minutes. Cells were washed six times with PBS and incubated with 100 mM glycine in PBS for 10 min. Subsequently, cells were washed two times with 100 mM glycine in PBS and three times with PBS. Cell membranes were permeabilized by 30 min of incubation with 1% PSA and 0.1% Triton X-100 in PBS. Cells were then incubated for 1 h with SB antibody (1:100 in 1% BSA in PBS), washed three times with PBS and incubated for 1h in darkness with α-goat-Alexa488 (1:1000) and DAPI (1:10000) in 1% BSA in PBS. Cells were washed three more times with PBS and mounted on microscope slides.

3 Results

3.1 Retargeting of SB with dCas9

3.1.1 Generation of fusions between dCas9 and transposase components

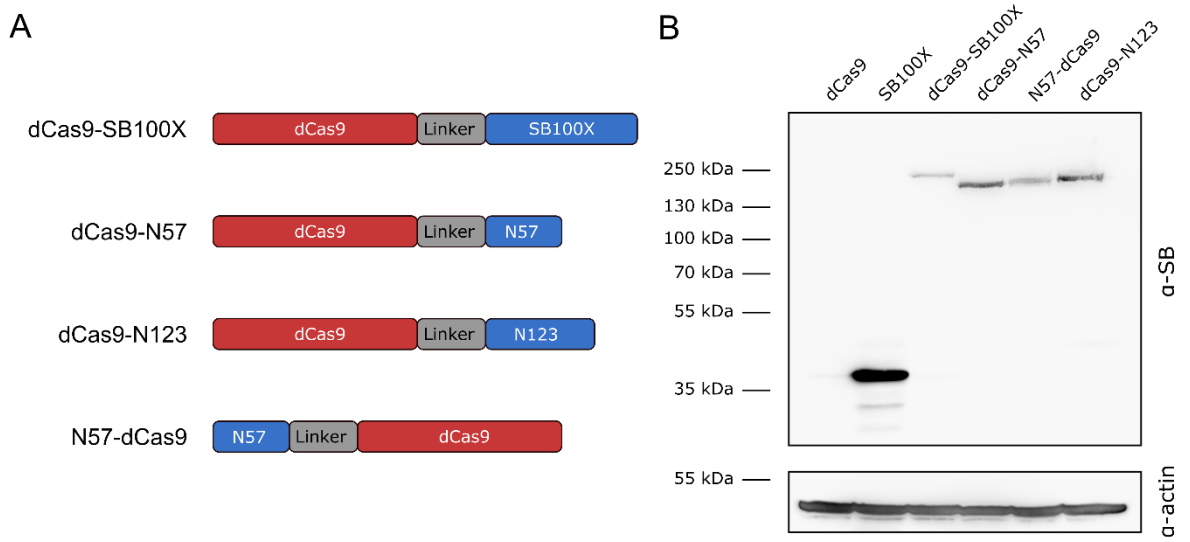


Figure 3.1 – Fusion constructs of dCas9 and SB components. A Schematic representation of the generated fusion constructs dCas9-SB100X, dCas9-N57, dCas9-N123 and N57-dCas9. **B** Western Blot of the generated constructs, developed with an α-SB antibody (top half) and an α-actin antibody (bottom half, loading control). Expected sizes were 202.5 kDa for dCas9-SB100X, 177.3 kDa for dCas9-N123, 169.7 kDa for dCas9-N57 and N57-dCas9, 39.3 kDa for SB100X and 42 kDa for actin. Unfused dCas9 is included as a negative control.

In order to target SB insertions towards specific sites in the genome, several strategies are available, as described in section 1.5.2 and a range of different constructs was generated to test these approaches (Figure 3.1A). The first option was generating a direct fusion between the DBD (dCas9) and the full-length, hyperactive transposase SB100X. Because previous studies showed that the SB transposase does not tolerate additions to its C-terminus[165,412,413], dCas9 was only added to the N-terminus of SB100X. To allow the necessary flexibility between the domains, a 14 aa linker with the sequence KLGGGAPAVGGGPK was inserted between dCas9 and SB100X. This linker has previously been used in a retargeting study using the SB transposon[164].

The alternative to generating a direct fusion was to generate a tethering factor that would localize the transposition complex to the target sites via non-covalent interaction. For this purpose, the SB fragments N57 and N123, consisting of the PAI subdomain or the entire PAIRED DNA-binding domain, respectively, were used. These relatively short peptides non-covalently interact both with full-length SB transposase molecules as well as with the ITRs of the SB transposon, which should allow them to attract both components of the pre-integration

complex to the target site. N123 binds ITR DNA more tightly than N57[147]. As no transpositional activity is required of N57 and N123, these peptides were fused to both the N- and the C-terminus of dCas9.

The generated constructs were tested for correct sizes and expression in a Western Blot (Figure 3.1B). All tested constructs were expressed and had the calculated sizes. All constructs were expressed at lower levels than unfused SB100X, which might be due to the different promoter strengths of the T7 and CBh promoters or due to altered protein degradation in the cells. Some differences in expression levels could also be observed between the targeting constructs, with the direct transposase fusion expressed at a lower level than the adapter proteins. While this difference cannot be attributed to the promoter, it might reflect differences in blotting efficiencies of the differently sized proteins or, again, altered protein dynamics in the cells.

For each targeting experiment, sgRNAs were cloned into a site on the backbone of the Cas9 plasmids, so sgRNAs and targeting constructs could be expressed from the same plasmid. The cloning procedure is described in section 2.2.1.4. The sgRNA expressed from each plasmid is specified for each experiment. If no target is specified, the plasmids do not express an active sgRNA, although the sgRNA scaffold is still expressed. Constructs with a sgRNA will be referred to by the name of the construct and the name of the sgRNA, separated by a slash. Thus dCas9-SB100X/sgHPRT-0 refers to a construct expressing the fusion protein dCas9-SB100X and the sgRNA sgHPRT-0 from the same plasmid.

3.1.2 Transpositional activity of dCas9-SB100X

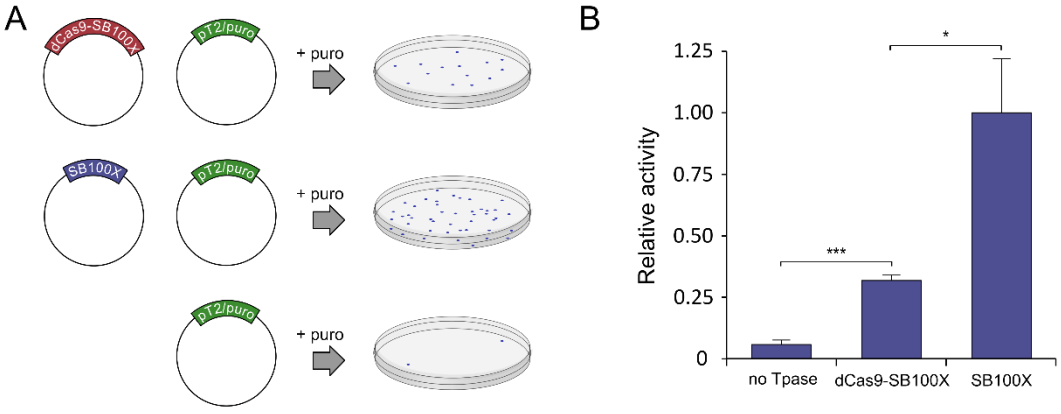


Figure 3.2 – Transpositional activity of dCas9-SB100X. A The general principle of the selection-based transposition assay. An antibiotic-tagged transposon is co-transfected into cells with the transposases to be tested (or no transposase as a negative control). After selection with the appropriate antibiotic, only cells containing stable integrations in the genome form colonies. **B** Relative transpositional activity of dCas9-SB100X compared to SB100X. The activity of dCas9-SB100X is significantly reduced to ca. 30% of the activity of SB100X. * $p \leq 0.05$, *** $p \leq 0.001$

In order to work as a tool for targeted transposition, the dCas9-SB100X fusion protein needs to retain a reasonable amount of transpositional activity. It has previously been shown that fusions of DBDs to the SB transposase often have a severe negative effect on the enzymatic activity of the transposase. Thus, the transpositional activity of dCas9-SB100X was tested by comparing the number of puromycin-resistant colonies formed after co-transfection of dCas9-SB100X with a puromycin-tagged transposon to the number formed by co-transfection of the transposon with SB100X (Figure 3.2A). Transfection of the transposon without transposase served as a negative control.

The assay revealed that the dCas9-SB100X retained the ability to catalyze the transposition reaction ($p<0.001$). However, activity of dCas9-SB100X was significantly reduced in comparison to SB100X ($p<0.05$), to a level of ca. 30% (Figure 3.2B).

3.1.3 DNA-binding activities of dCas9 fusions

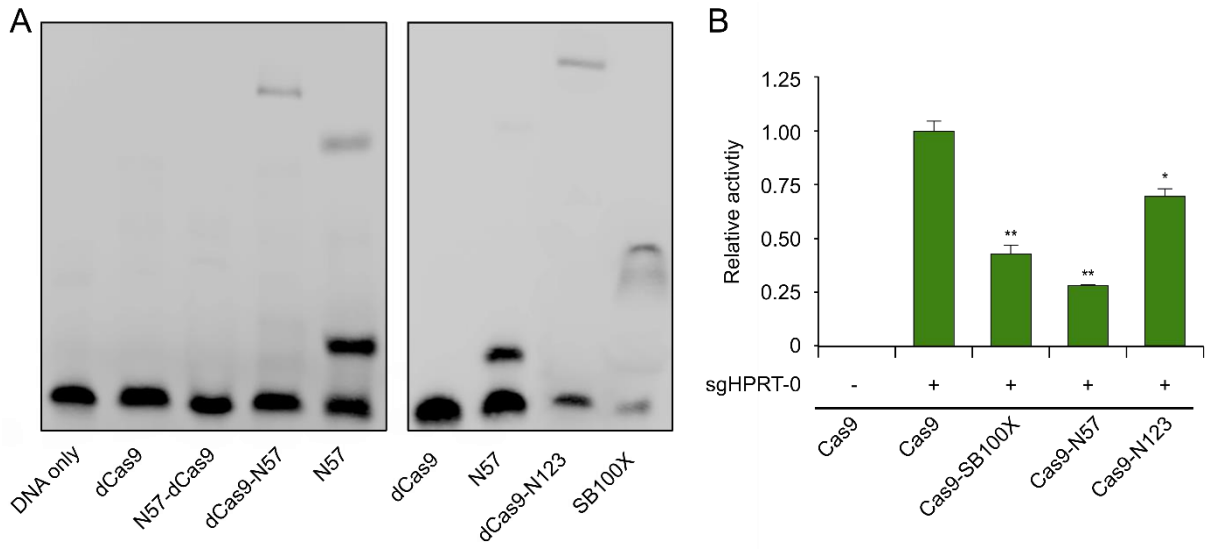


Figure 3.3 – DNA binding activities of fusion domains. A EMSAs of N57-dCas9, dCas9-N57 and dCas9-N123. N57 is included as a positive control and dCas9 is included as a negative control. dCas9-N57 still binds to the target oligonucleotide, but no interaction can be detected for N57-dCas9. N57 produces two bands, likely corresponding to monomeric and multimeric binding to the DNA. The adapter protein dCas9-N123 is analyzed with a separate EMSA, using the same positive and negative controls, with SB100X as an additional control. **B** Cleavage activity of Cas9 fusions, measured as a proxy for DNA binding activity of dCas9 domains in analogous dCas9 fusions. The *HPRT* disruption assay was performed as detailed in section 3.1.8. Cas9-SB100X, Cas9-N57 and Cas9-N123 were capable of catalyzing the cleavage reaction, although cleavage efficiency was reduced compared to unfused Cas9. Successful cleavage implies that sgRNA and target DNA interaction were possible. Indicated significance is relative to Cas9 + sgHPRT-0. * $p \leq 0.05$, ** $p \leq 0.01$

In order to work in targeting, the generated targeting proteins need two distinct enzymatic activities. While the direct fusions needs to be able to recognize their targets via their DBD and to catalyze the transposition reaction, the adapter proteins need to have two different binding activities (the actual transposition reaction is performed by separately supplied transposase

molecules). This secondary binding activity is needed to recruit either the transposon or the unfused transposase to the target site. In the constructs described here, the N57 or N123 domains should theoretically be able to bind to both the SB transposon and the SB transposase. DNA-binding activity was demonstrated in an EMSA with an oligo comprising the SB recognition sequence from the 14DR used as a binding partner (Figure 3.3A). Unfused N57 was used as a positive control in this assay.

It could be shown that in dCas9-N57, the N57 domain retains some DNA-binding activity. However, for N57-dCas9, no DNA-binding could be detected. Due to the lack of DNA-binding activity of N57-dCas9, this construct was excluded from all subsequent experiments. A separate EMSA was performed for dCas9-N123, using the same positive and negative controls, with the addition of unfused SB100X as a reference for binding strength. Like dCas9-N57, dCas9-N123 retained detectable DNA-binding activity. Both dCas9-N57 and dCas9-N123 bind to their target DNA significantly more weakly than unfused N57. However, it should be noted that unfused N57 is not a component of the naturally occurring transposition reaction and its high binding strength is not representative of a physiological situation. Indeed, the binding strength of unfused SB100X, which is high enough to enable the transposition reaction to take place, is lower than that of unfused N57 and only somewhat higher than the binding activity of dCas9-N123. The capability of N57 in these constructs to bind other transposase molecules was not separately tested, but if N57 retains its DNA-binding activity, its conformation and accessibility cannot have changed drastically and it is likely that it retains some protein-binding activity as well.

For either the direct transposase fusion or the adapter proteins to work in retargeting integrations, the dCas9 domain needs to be able to bind its sgRNA and recognize its target site in the context of these fusion proteins. In order to test this, cleavage-competent fusion proteins were generated (see section 2.2.1.3), replacing the dCas9 domain with an active Cas9 domain. Cas9 and dCas9 only differ in two point mutations (D10A and H840A)[228] and dCas9 is routinely used to target effector domains in a sgRNA-dependent manner, so the overall structure of Cas9 and dCas9 is the same. Because DNA binding is a necessary step in the cleavage reaction, cleavage of the Cas9 fusions can be used as a proxy measurement for binding of the dCas9 fusions. It should be noted that the addition of fusion partners may not influence DNA binding and overall cleavage to the same extent, so no quantitative inference can be made regarding DNA binding efficiency of the dCas9 fusions.

The Cas9 cleavage assay was tested by measuring disruption of the *HPRT* gene, as described in section 3.1.8. For this, each Cas9 fusion was transfected on a plasmid also expressing sgHPRT-0. All three tested fusions, Cas9-SB100X, Cas9-N57 and Cas9-N123 retained a significant fraction of cleavage activity (Figure 3.3B). However, the cleavage efficiency was reduced for all constructs: to ~30% for Cas9-SB100X ($p \leq 0.01$), to ~30% for Cas9-N57 ($p \leq 0.01$) and to ~70% for Cas9-N123 ($p \leq 0.05$).

3.1.4 Integration library generation

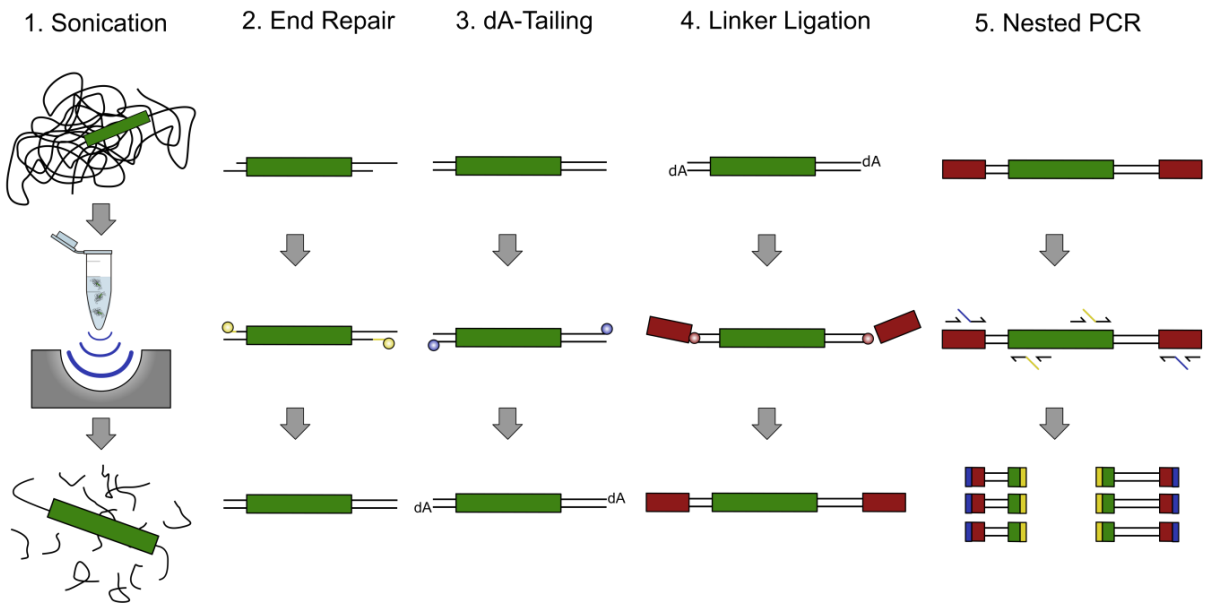


Figure 3.4 – Overview of integration library generation. Genomic DNA is sonicated to fragments with an average length of 600 bp, some of which will contain transposon sequences (green rectangle) (1). Staggered DNA ends are repaired (2) and dAs are added to the 3' ends (3) in separate enzymatic reactions, allowing short linkers (red rectangles) to be ligated to the DNA ends (4). A nested PCR is performed with primer pairs binding the transposon end and the linker, thus specifically amplifying genome-transposon-junctions from the fragmented gDNA (5). The inner set of primer contains overhangs that introduce sequences required for further processing as well as 6 nt barcodes that allow multiplexing (indicated by blue and yellow rectangles).

In order to test whether the generated direct transposase fusions or adapter proteins target insertions to their respective target sites, integration libraries were generated using different combinations of targeting constructs and sgRNAs. The general procedure of integration library generation is described in this section, while validation of the targeting sgRNAs and the results of individual libraries are discussed in the following sections.

HeLa cells were transfected with each targeting factor, the transposon pT2/puroDR3 and – except for the samples containing direct transposase fusions – an SB transposase (either SB100X or SB10). The transposon pT2/puroDR3 contains an additional SB binding site in the right ITR, which has been shown to positively influence targeting[164]. After 2-3 weeks of puromycin selection, gDNA was isolated from the cells and integration libraries were

generated. An overview of the principle of integration library generation can be found in Figure 3.4, and the process is described in detail in section 2.2.6.

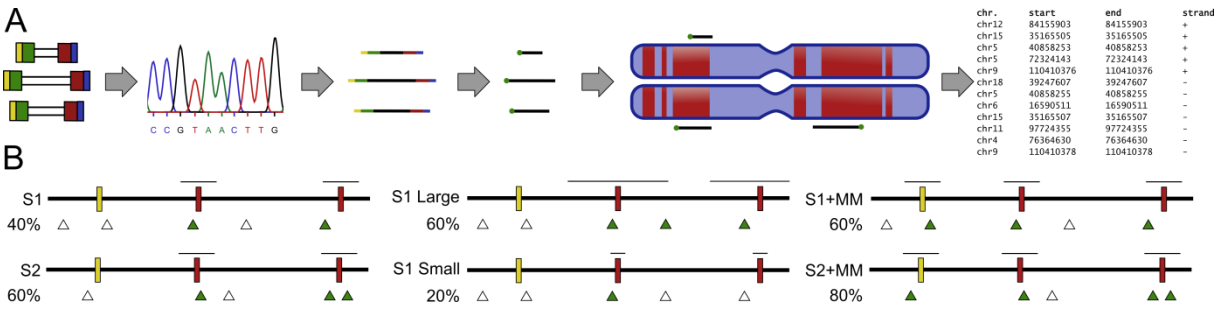


Figure 3.5 – Integration library analysis. A Generation of insertion site lists. DNA samples generated as described in Figure 3.4 are Illumina sequenced. Sequencing reads are checked for the presence of both tails (which were added by the last PCR reaction) and a transposon-genome junction. Demultiplexing occurs at this step. Non-genomic sequences are trimmed and the remaining genomic sequences are mapped to the human genome, resulting in a list of positions where integrations have occurred. **B** Window-based analysis of enrichment near target sites. Targeting windows are defined around target sites defined by the used sgRNAs (red rectangle). Insertions that occurred within the specified distance to a target site are defined as "hit" insertions (green triangles), all others are classified as "miss" insertions (white triangles). This analysis is applied to two insertion sets (S1 and S2) in this example (left), percentages indicate fraction of "hit" insertions. The analysis can be varied by increasing or decreasing the size of the targeting windows (center) or by allowing mismatches in the target sequence (yellow rectangle; right).

After library generation and quality control, libraries were deep sequenced. Reads were trimmed and mapped to the human genome (assembly hg38, Figure 3.5A). For some libraries, enrichment of integrations in windows around target sites was calculated as a measure for a targeting effect. For each target, every site in the human genome with 100% identity to the targeting sequence of the sgRNA was considered a target site. Targeting windows were then defined as extending to both sides of each target site. For example, a targeting window size of 1 kb (total) means considering every insertion that is less than 500 bp from a target site a "hit" insertion and all other insertions "miss" insertions (Figure 3.5B, left). For single-copy targets, only a single targeting window is present in the genome, while for multi-copy targets many – potentially overlapping – target windows are defined. The sizes of these windows can be varied to probe the distribution of insertions around target sites (Figure 3.5B, center). In addition to analyzing insertion frequencies around perfectly matched sites, in some cases mismatches to the sgRNA targeting sequence were allowed in order to test whether integrations were enriched around other sites with high similarity to the sgRNA-defined targets (Figure 3.5B, right). The final measure of targeting efficiency was enrichment achieved by use of targeting factors, i.e. the frequency of "hit" insertions was compared between the targeted sample and a reference (generally the same factor with an unrelated sgRNA). If, for example, the number of insertions – as a fraction of total insertions recovered from the respective library – was 50% higher in the

targeted sample than in the untargeted sample, it would be considered a 1.5-fold enrichment into the tested targeting window.

Total numbers of insertions recovered from each integration library are listed in Supplementary Figure 1A and fractions of insertions in windows around multicopy target sgRNA binding sites are listed in Supplementary Figure 1B.

3.1.5 Validation of L1-directed sgRNAs

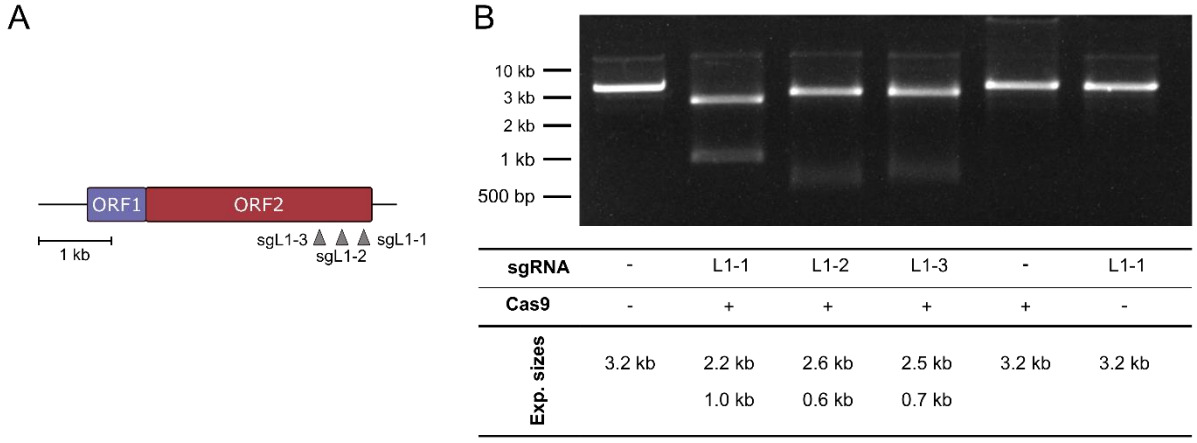


Figure 3.6 – *In vitro* validation L1-directed sgRNAs. A Schematic representation of the binding sites of the three L1-targeted sgRNAs within the L1 sequence. **B** *In vitro* digestion of a 3.2 kb fragment containing target sites for the sgRNAs sgL1-1, sgL1-2 and sgL1-3 with Cas9 and the respective sgRNAs. All three sgRNAs are effective at cleaving their target sites *in vitro*.

In order to get a first indication whether it is possible to influence the integration pattern of SB with dCas9-based adapter proteins or direct transposase fusions, we chose to use sgRNAs targeting repetitive elements in the genome. Using targets which are present in many copies should increase the chances of the targeting construct encountering their target sequences and allow a more detailed analysis of distributions around the target sites.

The first of these targets was the L1Hs (human LINE1) element, the human version of a ubiquitous family of non-LTR retrotransposons. L1Hs are ~ 6 kb in length and comprise ~ 17% of the human genome[452]. They might be attractive targets for SB transposition due to their relatively low GC content and consequently high number of TA sites available for SB transposition. Additionally, L1Hs elements have previously been successfully targeted using a ZFP binding near the 3'-end of the element[165].

For L1Hs elements, three sgRNAs called sgL1-1, sgL1-2 and sgL1-3 were tested, all of which targeted the 3'-end of L1Hs (Figure 3.6A). It should be noted that these sgRNAs target only a subset of L1 elements, as their target sequence is not present in every single L1 element in the

genome. These sgRNAs were tested in an *in vitro* cleavage reaction using a fragment of the plasmid JJ101-L1.3 generated by BamHI digestion as a target. The ~3 kb fragment generated by this digestion contains all three L1-directed sgRNA target sequences. Incubation of the target DNA with Cas9 and any of the three sgRNAs resulted in complete digestion of the target DNA (Figure 3.6B), indicating that the sgRNAs are capable of efficiently directing cleavage *in vitro*.

3.1.6 Validation of AluY-directed sgRNAs

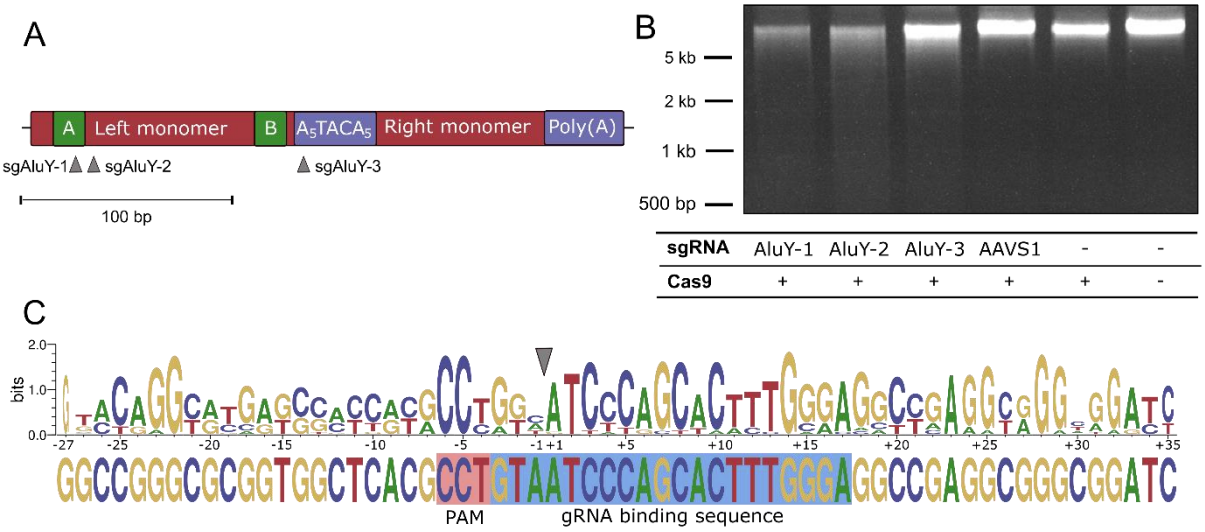


Figure 3.7 – Validation of AluY-directed sgRNAs. A Schematic representation of the position of sgRNA target sites within the Alu consensus sequence. sgAluY-1 binds in the A-box, sgAluY-2 binds close to the A-box in the left monomer and sgAluY-3 binds in the A-rich stretch between the two monomers. **B** *In vitro* digestion of HeLa gDNA with Cas9 and the AluY-directed sgRNAs. A sgRNA targeting AAVS1 (a single copy locus) is used as a negative control. Distinct fragmentation of gDNA can be observed in samples containing sgRNAs sgAluY-1 and sgAluY-2. **C** Sequence logo generated from the ends of DNA digested by Cas9 with sgAluY-1 (top). Digested and purified DNA was cloned into a plasmid and plasmid-genome junctions were Sanger sequenced. Recovered sequences were aligned to the AluY sequence, either immediately downstream or upstream of the cleavage site. The recovered sequences exhibit clear similarity to the AluY consensus sequence (bottom), demonstrating that the fragmentation of the genomic DNA was indeed the result of cleavage mediated by Cas9 and sgAluY-1.

The second multicopy target tested was the Alu element, a primate-specific SINE that comprises ~11% of the genome in humans. It was chosen because it is present in even higher numbers than the L1Hs element. The consensus sequence of the AluY element was chosen as the basis for sgRNA design due to the high copy number and conservation of this Alu subclass. However, Alu elements may be less attractive targets than L1Hs due to their small size (~250 bp) and their relatively high GC content (~63%)[453], which results in a lower number of available TA dinucleotides for SB transposition.

As for L1, three sgRNAs were designed for AluY, two of which bind in the left monomer of the Alu element and one binding in the A-rich stretch between the two monomers (Figure 3.7A). The sgRNAs sgAluY-1, sgAluY-2 and sgAluY-3 were tested by *in vitro* digestion of gDNA. Digestion of gDNA with sgAluY-1 and sgAluY-2 resulted in clear fragmentation of the DNA,

while digestion with sgAluY-3 resulted in less pronounced fragmentation (Figure 3.7B). In order to verify that the observed fragmentation was indeed the result of specific cleavage of AluY sequences, DNA digested with Cas9 and sgAluY-1 was purified and fragments were cloned into the SmaI site in the lacZα sequence of the plasmid pUC19. After transfection into *E. coli* DH5α and cultivation on a substrate containing X-gal, 16 white colonies were picked and both plasmid-insert junctions were sequenced from both directions. Of 32 plasmid-genome junctions recovered, 31 could be aligned to the AluY consensus sequence and started at the predicted cleavage site. 12 of these sequences matched the short left end of the AluY sequence and 19 matched the longer right end (Supplementary Figure 2A). By aligning both pools of sequences to the canonical AluY sequence and generating a sequence logo, it could be demonstrated that the DNA ends generated by digestion with Cas9/sgAluY-1 show clear similarity to the expected sequence in the AluY element (Figure 3.7C). However, it is also apparent that all recovered sequences contain a number of mismatches to the canonical AluY sequence, even in the region bound by the sgRNA (Supplementary Figure 2B and C). The two to G nucleotides in the spCas9 PAM NGG (represented by the complementary C nucleotides in the logo) are invariant, as their presence is required for cleavage by Cas9. It should also be noted that the higher variability upstream of the cut site can be attributed to the lower number of sequences making up this part of the logo than the part downstream of the cut site (12 vs. 19 sequences).

3.1.7 AluY- and L1-targeted integration libraries

While some trends in the distribution of insertions around sgL1-1 sites could be observed, the total number of insertions there was too low to make any statistical conclusions on whether a targeting effect had occurred. To increase the number of target sites available for analysis, up to one mismatch was allowed compared to the original sgL1-1 binding sequence. This increased the number of sgL1-1 sites approximately threefold (from 5438 to 14264).

Comparing the pattern of insertions around the sgL1-1 binding sites obtained with dCas9-SB100X/sgL1-1 to dCas9-SB100X/sgAluY-1 shows no striking enrichment near the target sites (Figure 3.8A). While the comparison is made somewhat difficult due to the comparatively low overall number of insertions in the sgAluY-1 dataset, it seems clear that for both datasets the number of insertions seems to be mostly influenced by the TA frequency around the target sites, which is relatively high upstream of the target sites (i.e. in the L1 element) and lower downstream of the target sites. Sequences from -100 to +100 bp around the target sites are particularly TA-poor and peaks in TA frequency can be observed at around -200 and -100 bp.

A clear insertion hotspot can be seen in the dCas9-SB100X/sgL1-1 dataset at around -175 bp, where the insertion frequency at one position is approximately 10-fold higher than at surrounding TAs. While insertions are also concentrated near this position in the sgAluY-1 dataset, the enrichment at this position seems to be less pronounced. In general, the dCas9-SB100X/sgL1-1 dataset has a higher number of insertions in the TA-depleted regions immediately downstream of the binding site, where no insertions were recovered from the dCas9-SB100X/sgAluY-1 dataset. While this effect is similar to the observation made for targeting of sgAluY-1 binding sites (see below), i.e. increased integration rates in a generally disfavored region immediately downstream of the binding sites, it could also be attributed to the low overall number of insertions in the sgAluY-1 dataset. While, comparing the dCas9-SB100X/sgAluY-1 and dCas9-SB100X/sgL1-1 datasets, a ~25% enrichment can be observed for a symmetrical 500 bp region around the target sites, it is not statistically significant due to the overall relatively low number of insertions in this region.

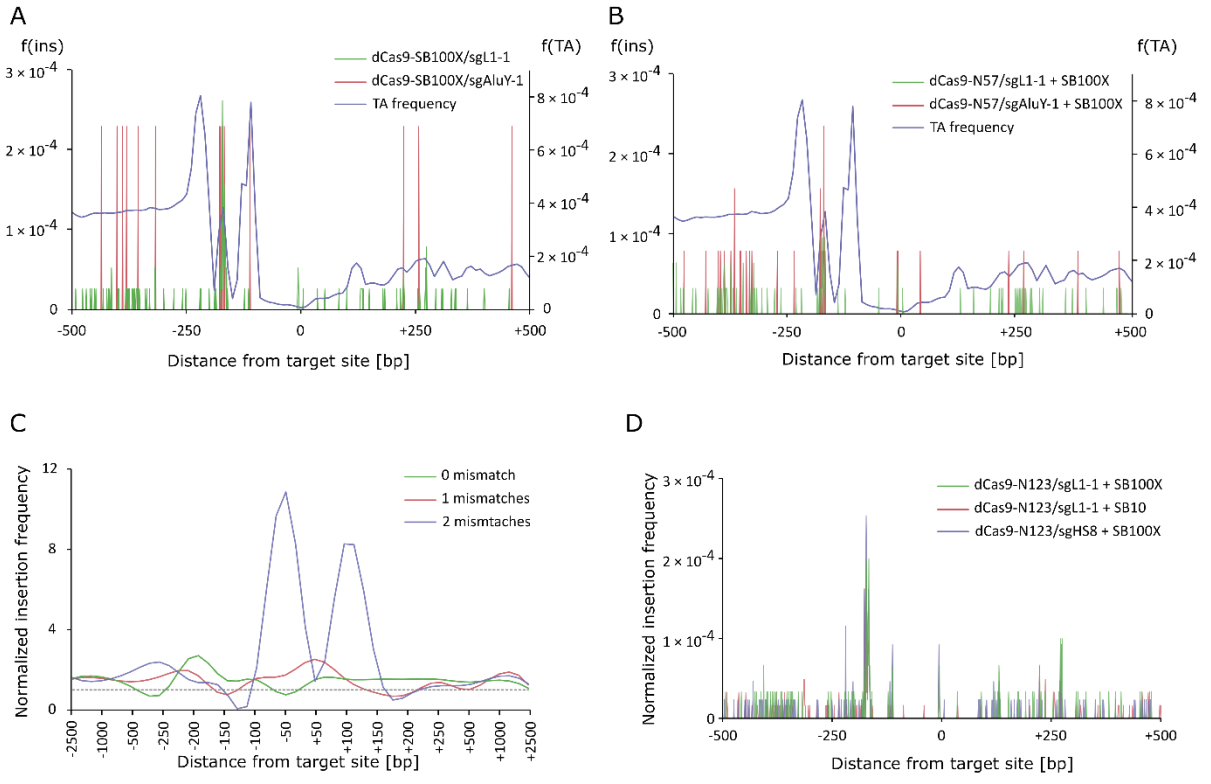


Figure 3.8 – Targeting of the L1 element. A Normalized insertion frequencies around sgL1-1 target sites, including sites with a single mismatch, of insertion libraries catalyzed with dCas9-SB100X/sgL1-1 or dCas9-SB00X/sgAluY-1 as a control. Note that the library of insertions with sgAluY-1 contains fewer total insertions than the sgL1-1 library, resulting in higher peaks. Left y-axis: f(ins) = normalized insertion frequency; right y-axis: f(TA) = average TA frequency. **B** Normalized insertion frequencies around sgL1-1 target sites (up to 1 mismatch) obtained with SB100X and dCas9-N57/sgL1-1 and dCas9-N57/sgAluY-1 as a control. **C** Insertion frequencies obtained with dCas9-N57/sgL1-1, normalized to dCas9-SB100X/sgAluY-1, for sgL1-1 target sites with different numbers of mismatches. Pseudocouts were added to prevent zero values and division by zero. **D** Normalized insertion frequencies around sgL1-1 target sites (up to 1 mismatch), catalyzed by SB100X or SB10 with dCas9-N123/sgL1-1. As no library with dCas9-N123/sgAluY-1 was available as a reference, a library of dCas9-N123/HS8 (see section 3.1.11) was used instead.

Analyzing insertions catalyzed with SB100X and dCas9-N57/sgL1-1, compared to dCas9-N57/sgAluY-1, yields a similar picture (Figure 3.8B). No clear targeting of the sgL1-1 binding sites can be observed and the tendency for insertions to occur in the disfavored region immediately downstream of the binding site seems less pronounced than in the dCas9-SB100X/sgL1 -1 dataset. Still, some TA sites that are not utilized in the sgAluY-1 sample can be recovered from the sgL1-1 set, although it is unclear if this can be attributed to any targeting effect or is simply due to the higher number of insertions in the dCas9-N57/sgL1-1 library. Otherwise, integration frequency again closely mirrors TA content of the DNA and the previously observed insertion hotspot is seen again, this time also in the control dataset.

Interestingly, an increase in insertion frequencies can be observed with dCas9-N57/sgL1-1, compared to dCas9-N57/sgAluY-1, when looking at sites with two mismatches compared to the sgL1-1 target sequence (Figure 3.8C). It is unclear why no similar peaks are visible at sites with 0 or 1 mismatch, although it is possible that the composition of DNA around the sites with two mismatches is different from the composition of DNA around sites with no mismatches or one mismatch. Again, due to the low number of insertions involved, the change is not statistically significant.

Targeting with the adapter protein dCas9-N123 was done with the hyperactive transposase SB100X as well as the less active SB10 transposase in order to test whether transposase activity would influence targeting specificity. Specifically, this was meant to test the hypothesis that lower transposase activity might improve targeting by reducing the number of untargeted background insertions. As no dataset using dCas9-N123/sgAluY-1 was available as a reference, a dataset obtained with dCas9-N123 and the single-copy sgRNA sgHS8 (see section 3.1.11) was used instead. Comparing the three datasets showed no enrichment with dCas9-N123/sgL1-1 compared to dCas9-N123/HS8 using either transposase. While insertion frequencies near the target sites with SB10 were somewhat higher than with SB100X, as would be expected under the assumption that lower transposase activity might favor targeting, frequencies with dCas9-N123/sgHS8 were higher than either sgL1-1 dataset (Figure 3.8D). As before, insertion frequency seems mainly determined by TA frequency, with insertions depleted in the TA-poor stretches around the target site and more frequent in the TA-rich L1 element upstream of the target site. The previously observed hotspot at around -175 bp is seen again, independently of the sgRNA used.

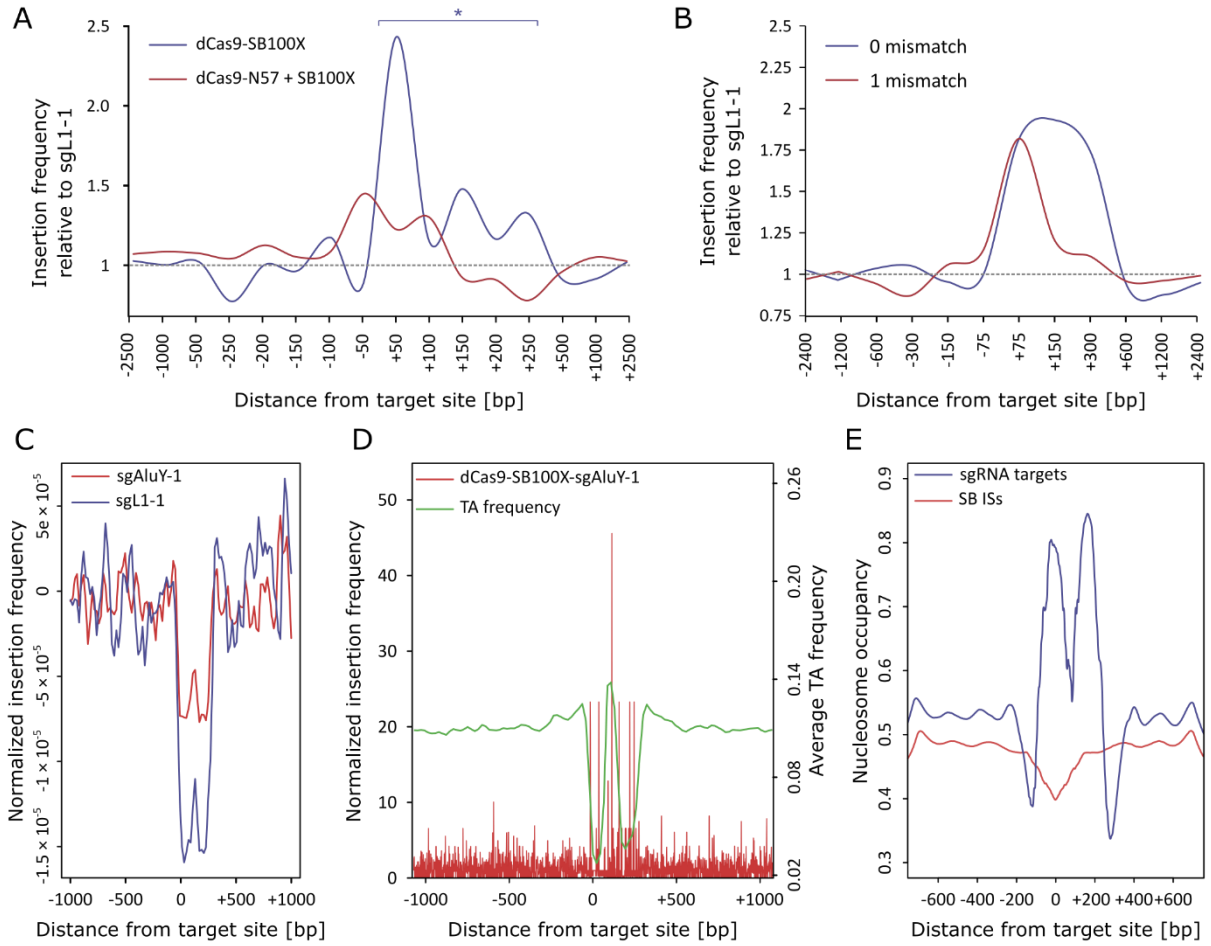


Figure 3.9 – Targeting of the AluY element. A Relative enrichment of insertions into small targeting windows upstream and downstream of sgAluY-1 target sites with dCas9-SB100X/sgAluY-1 and dCas9-N57/sgAluY1. Enrichment is relative to the same targeting construct in combination with sgL1-1. Enrichment is statistically significant for dCas9-SB100X and a targeting window from 0 to +300 bp. **B** Enrichment into targeting windows upstream or downstream of sites with zero or one mismatches to the sgAluY-1 target sequence, generated with dCas9-SB100X. **C** Insertion frequencies around sgAluY-1 target sites obtained with dCas9-SB100X, relative to the mean. A clear drop in insertion frequencies can be observed for both samples, but it is less pronounced for the sample expressing sgAluY-1. **D** Insertion frequency of dCas9-SB100X/sgAluY-1 around sgAluY-1 target sites, normalized to insertion frequencies with dCas9-SB100X/sgL1-1. Pseudocouts were added to prevent zero values and division by zero. Insertion frequency is superimposed with average TA frequency. **E** Average nucleosome occupancy around the target sites (blue) as well as around insertion sites from an untargeted SB dataset (red). * $p \leq 0.05$

While the changes in integration patterns around sgL1-1 target sites were generally inconclusive, targeting around sgAluY-1 target sites could be analyzed with a higher level of detail due to the significantly higher number of target sites in the genome (~300,000). Samples containing the unrelated sgRNA sgL1-1 were used as a reference and fractions of insertions were calculated for windows of varying sizes around sAluY-1 target sites (Figure 3.9A). These windows are non-cumulative, i.e. counts from the shorter windows were not included in the larger windows around them. Insertions were independently counted for the regions upstream and downstream of the target sites. This analysis revealed a statistically significant enrichment of insertions catalyzed by dCas9-SB00X/sgAluY-1 when compared to dCas9-SB100X/sgL1-1.

The enrichment occurred exclusively downstream of the target sites in a window of around 300 bp (p=0.019) and enrichment was ca. 2-fold for this window. The highest enrichment of ca. 2.5-fold occurred in a 50 bp window immediately downstream of the target site. For dCas9-N57, a slight enrichment in a window of ca. 200 bp around the target sites could be observed. In contrast to dCas9-SB100X, the enrichment observed with dCas9-N57 occurred symmetrically around the target sites, with no preference for downstream insertions. The slight enrichment with dCas9-N57 was also not statistically significant, even for this 200 bp window.

Previous targeting studies have shown that enrichment of insertions sometimes occurs around sites that have a similar, but not identical sequence to the sequence defined by the DBD[164]. Additionally, the relatively high variation in sequences recovered from the *in vitro* cleavage of gDNA with Cas9 and sgAluY-1 (see section 3.1.6) suggests that this sgRNA might have significant off-target activity. Thus, enrichment around sites with mismatches to the sgAluY-1 target sequence was analyzed. Some enrichment occurred downstream of sites with a single mismatch (Figure 3.9B), although it was weaker than the enrichment around perfectly matched sites and consequently lost its statistical significance. No enrichment could be observed around sites with more than one mismatch.

In order to probe the reason for the asymmetrical distribution of enrichment observed with dCas9-SB100X/sgAluY-1, the insertion frequencies around the sgAluY-1 target sites were analyzed for direct fusion transposase insertion datasets (Figure 3.9C). This revealed that the ~300 bp region downstream of the target sites is generally disfavored for SB insertions. A drop in insertion frequency occurs here for both datasets, but the drop is less pronounced when sgAluY-1 is used. Analyzing the average TA frequency around the target sites revealed that TA frequency is much lower in this window, likely the main reason for the drop in insertion frequency (Figure 3.9D). Superimposition of TA frequency with relative enrichment for individual positions also showed that the region of enrichment clearly coincides with the region of reduced TA frequency.

In addition to TA frequency, nucleosome occupancy around the target sites was evaluated (Figure 3.9E). This showed that the region downstream of the target sites has relatively high nucleosome occupancy. As can be seen from the average nucleosome occupancy around untargeted SB integration sites, the SB transposase preferentially integrates its cargo into nucleosome-free DNA. This might be a second reason for the reduced integration frequency in the region downstream of sgAluY-1 target sites, in addition to the low TA frequency observed there.

3.1.8 Validation of HPRT-directed sgRNAs

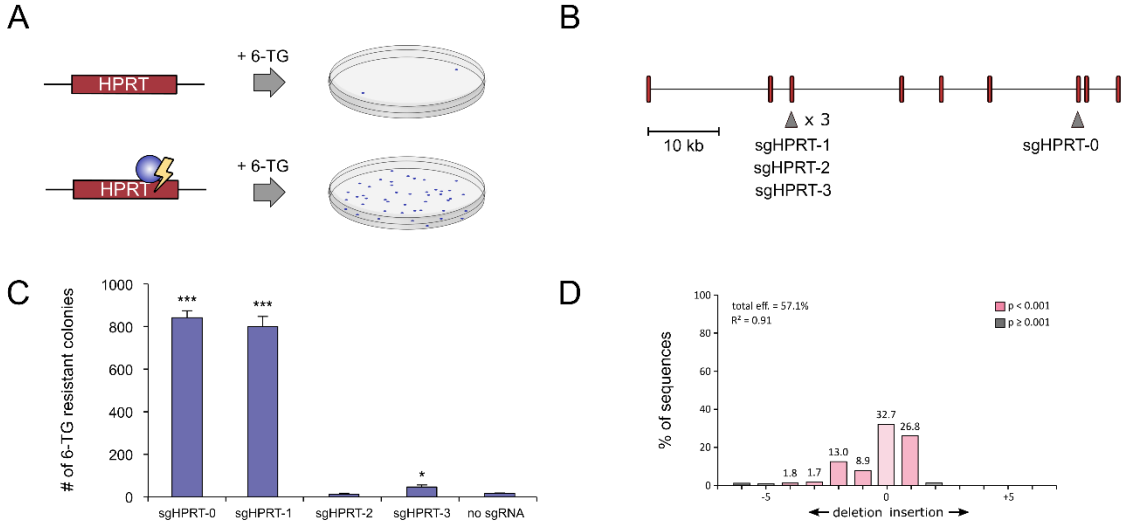


Figure 3.10 – Validation of *HPRT*-directed sgRNAs. A Principle of the *HPRT* disruption assay. The HPRT enzyme converts the non-toxic compound 6-TG into a toxic metabolite, making 6-TG lethal to cells with an intact *HPRT* gene. Disruption of *HPRT* by Cas9-mediated cleavage results in cells becoming tolerant to 6-TG. Tolerant cells are able to form colonies under 6-TG selection. **B** Schematic representation of the position of the sgRNA target sites. sgHPRT-0 recognizes a sequence in exon 7, the other sgRNAs bind in exon 3. **C** Numbers of colonies formed after transfection with Cas9 and four different *HPRT*-targeted sgRNAs. Statistical significance is calculated in comparison to the sample containing no sgRNA. * $p \leq 0.05$, *** $p \leq 0.001$ **D** Results of TIDE analysis using sgHPRT-0. Overall efficiency of the sgRNA was determined to be 57%.

The first single-copy target chosen for targeting experiments was the *HPRT* gene, a large (~30kb) gene on the X chromosome. This target was chosen because a counterselection procedure makes it possible to easily screen for cells in which the gene is disrupted. The HPRT enzyme is involved in the purine salvage pathway, converting guanine to guanosine monophosphate[454]. However, when *HPRT*$^{+}$ cells are supplied with 6-TG, this compound is converted into nucleotide form and integrated into the cell's DNA, resulting in toxicity. For cells in which the *HPRT* gene is disrupted, 6-TG is non-toxic, allowing for selection of *HPRT*$^{-}$ cells[455] (Figure 3.10A).

Four different sgRNAs were tested: sgHPRT-0, which binds in exon 7 of the *HPRT* gene and sgRNAs sgHPRT-1, sgHPRT-2 and sgHPRT-3, which bind in exon 3 (Figure 3.10B). All sgRNAs were tested by co-transfecting plasmids expressing the sgRNA with a Cas9 expression plasmid (Figure 3.10C). Cells transfected with Cas9 but no sgRNA served as a negative control. This analysis revealed that the sgRNAs sgHPRT-0 and sgHPRT-1 are highly active, both inducing a ca. 50-fold increase in the number of 6-TG resistant colonies ($p<0.001$ for both). The activity of sgHPRT-3 was comparably low with a 3-fold increase in the number of colonies ($p<0.05$) and sgHPRT-2 had no detectable activity.

Due to the highest activity determined by *HPRT* disruption, sgHPRT-0 was chosen for further analysis. The activity of this sgRNA was quantified with a TIDE assay (Figure 3.10), in which the generation of indels at the target site is measured by comparing the sequences from pools of edited and non-edited cells. The overall editing efficiency of Cas9 with sgHPRT-0 was determined to be ~57%.

3.1.9 *HPRT*-targeted integration libraries

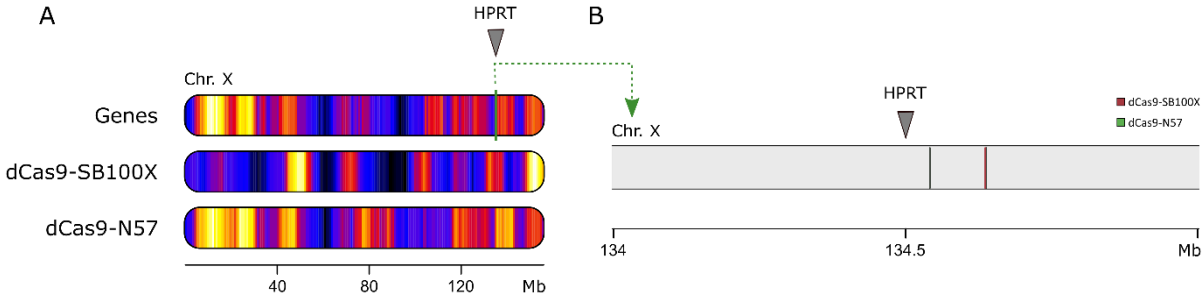


Figure 3.11 Targeting of the *HPRT* gene. A Distribution of insertions obtained with the different targeting constructs along the X chromosome, as well as distribution of genes. **B** Insertions in a 1 Mb window around the sgHPRT-0 target site.

Targeting of the *HPRT* locus was tested with dCas9-SB100X as well as with dCas9-N57 with SB100X. The sgRNA sgHPRT-0 was co-expressed from the plasmids dCas9-SB100X or dCas9-N57. Integration libraries were generated using the previously described protocol.

Analysis of the integration libraries showed that, although sgHPRT-0 has been shown to efficiently induce Cas9-mediated cleavage at the target site, no insertions could be recovered from the targeted locus in either library (Figure 3.11B). The closest insertion occurred at 42.2 kb from the target site with dCas9-N57 and the closest insertion catalyzed by dCas9-SB100X occurred at a distance of 137.9 kb. Overall, only three total insertions occurred in a region of 1 Mb around the target site.

Interestingly, the overall distribution of insertions catalyzed by dCas9-SB100X was clearly different from the distribution of insertions obtained with unfused SB100X and an adapter protein (Figure 3.11A). While the distribution of insertions catalyzed by unfused transposase roughly correlated with gene density, dCas9-SB100X showed a clear preference for inserting into two specific genomic regions, one of them near the end of the chromosome.

A single insertion recovered from the dCas9-N57 dataset was found in close proximity (<250 bp) to a site with three mismatches to the sgHPRT-0 binding sequence. However, based

on a single insertion it is not possible to determine whether this was the result of a targeting effect or whether the insertion simply occurred near a mismatched target site by chance.

3.1.10 Design and validation of GSH-targeted sgRNAs

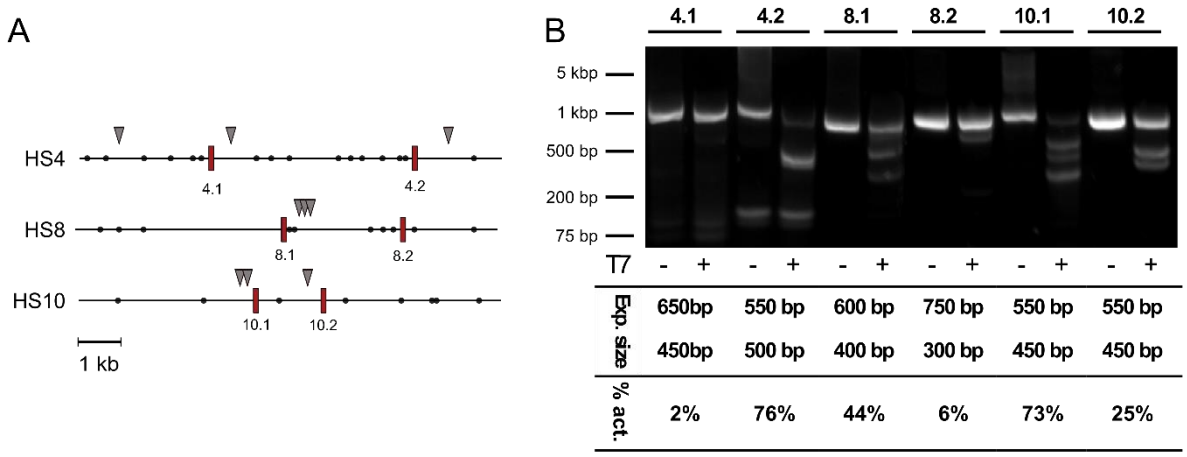


Figure 3.12 – Validation of GSH-directed sgRNAs A Schematic representation of the targeted loci. Grey triangles represent the position of previously identified untargeted SB insertions. Black circles represent the presence of pseudo SB sites in the genome. Red boxes represent the position of the designed sgRNAs. The entire lengths of the lines represent 10 kb. **B** Results of the T7 assay. Each set of two lanes represents DNA amplified from cells transfected with Cas9 and the sgRNA indicated at the top. The left lane of each set is a negative control not treated with T7 Endonuclease. The expected sizes of the two digestion fragments are indicated below. The ~150 bp band in the sample with sgHS4.2, which is also present in the negative control lane, is an unspecific PCR product from amplification of the locus. Activity values represent the approximate fraction of DNA that was cleaved and are only calculated for the +T7 lanes.

After the unsuccessful attempt to direct insertions with sgHPRT-0, we reasoned that the *HPRT* locus might simply not be a good target for SB insertions, especially since previous attempts to target it with TAL-based constructs had failed as well (unpublished data). In order to avoid this problem, we decided to attempt to target positions where several independent SB integrations had been previously identified in close proximity without any targeting. The rationale behind this was that loci that were already hit repeatedly without targeting likely represent favorable targets for SB insertion. In addition to being 'hotspots' for SB integration, we looked for loci that also fulfilled GSH criteria[341]. Using these two conditions, we identified three target sites which are located on chromosomes 4, 8 and 10 (Figure 3.12A). Interestingly, a high number of pseudo SB sites (sequences that are similar to the sequence of the transposon ITR recognized by the transposase) were found close to these sites, which might explain why they represent favorable targets for SB transposition.

Two sgRNAs were designed for each of the loci (sgHS4.1, sgHS4.2, sgHS8.1, sgHS8.2, sgHS10.1 and sgHS10.2) using the online tool CCTop[439], in each case choosing a target site that is close to the previously reported SB insertions, and predicted to result in a sgRNA with high efficiency and without any major predicted off-target sites. All sgRNAs were tested using

a T7 Endonuclease assay (Figure 3.12B). The assay revealed detectable cleavage activity for all sgRNAs except sgHS4.1; sgHS4.2, sgHS8.1 and sgHS10.1 were found to cleave DNA with high efficiencies of ca. 76%, 44% and 73%, respectively. Due to their high activities, these sgRNAs were chosen for the subsequent targeting experiment.

3.1.11 GSH-targeted integration libraries

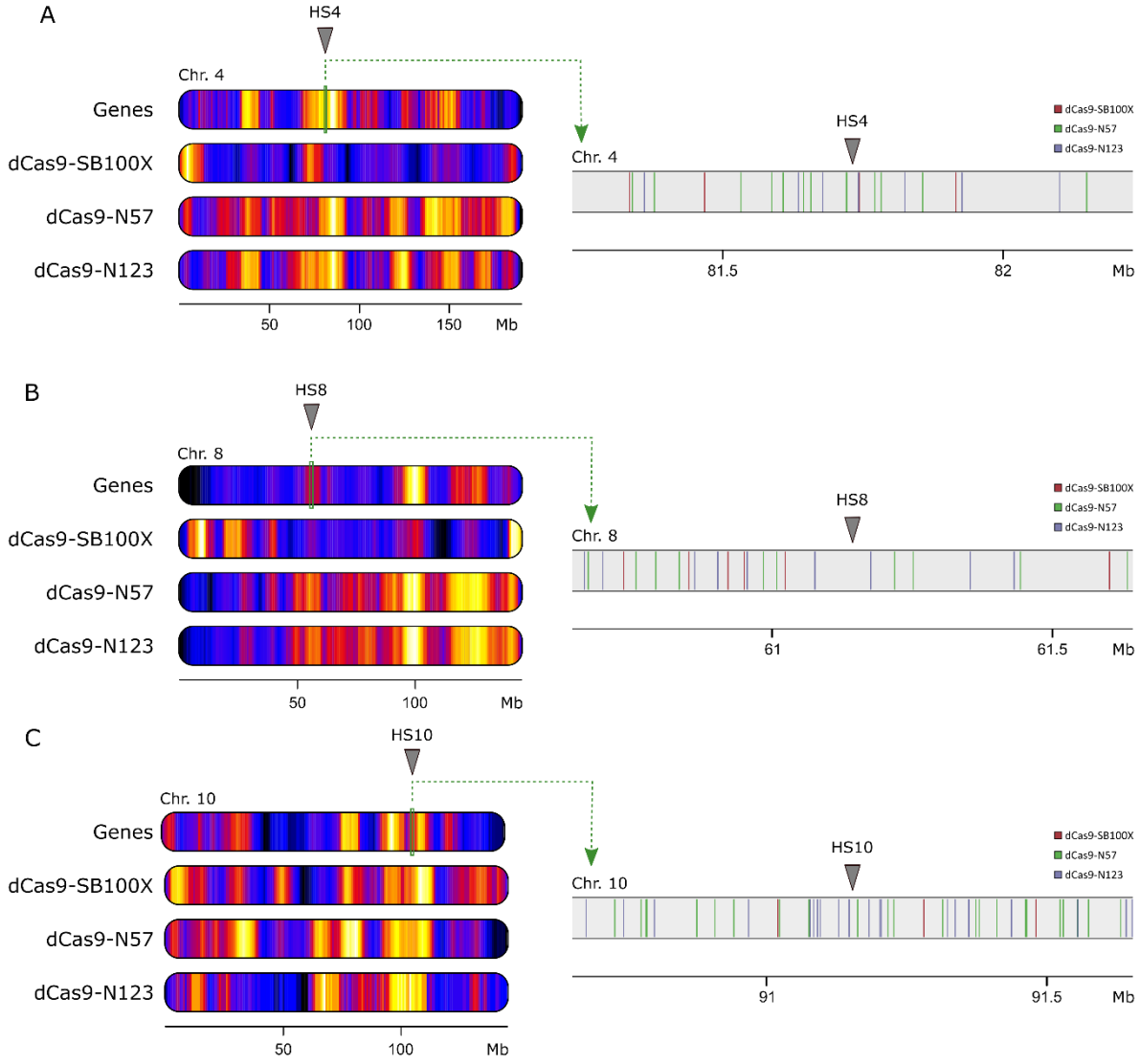


Figure 3.13 – Targeting of GSH sites. A Targeting of HS4. The left side shows the overall distribution of insertions on the chromosome obtained with each targeting construct and the overall distribution of genes. However, zooming in on a 1 Mb region around the target site shows that no insertions occurred at the targeted locus. **B** Targeting of HS8. **C** Targeting of HS10.

Integration libraries for the analysis of GSH/'hotspot' targeting were generated using the previously described protocol. Targeting was attempted using the direct fusion transposase dCas9-SB100X as well as the adapter proteins dCas9-N57 and dCas9-N123.

Like the targeting attempt of the *HPRT* gene before, targeting of the three selected GSH/'hotspot' sites did not produce any insertions at the intended loci. Even though the sites should theoretically be receptive to SB insertions and many insertions were recovered from a wider region around them, no insertions occurred in close proximity to the target sites, independently of which targeting construct was used.

While the target sites should be receptive to SB integrations, analysis of the libraries showed that only HS4 is located in a chromosomal region with a relatively high density of SB insertions (Figure 3.13A). HS8 and HS10, on the other hand, are in chromosomal areas of average SB insertion density (Figure 3.13B and C). The closest insertion to HS4 was recovered from the set using Cas9-N57/sgHS4 and was found at 10.4 kb from the target site. The closest insertion with dCas9-N123/sgHS4 occurred at a distance of 10.7 kb and the closest insertion with dCas9-SB100X/sgHS4 was at 11.9 kb from the target. The closest insertions around HS8 occurred at significantly higher distances, with the closest being 40.7 kb away, in the dCas9-N123/sgHS8 dataset. The closest integrations to HS8 for the other two datasets were 53.6 kb away with dCas9-N57/sgHS8 and 65.6 kb away for dCas9-SB100X/sgHS8. For HS10, the closest insertion occurred 16.9 kb away and was found in the dCas9-N123/sgHS10 dataset, the closest insertions for dCas9-SB100X/sgHS10 and dCas9-N57/sgHS10 were 46.9 kb and 67.3 kb, respectively. Enrichment was also tested for target sites with up to three mismatches, but, again, no targeting effect could be observed.

As has been observed for the *HPRT*-targeted libraries, the distribution of insertions catalyzed by the direct transposase fusion seems to be clearly distinct from the pattern observed for the adapter proteins. Again, insertions seem to preferentially occur in certain regions, often near the ends of the chromosomes, while insertions catalyzed by the unfused transposases are more evenly distributed along the length of the chromosomes.

3.1.12 Design and validation of TA_n-targeted sgRNAs

Another strategy to improve the low efficiency of targeting was the use of a SB mutant which has a more specific target preference than SB100X. The only requirement for SB insertion is a TA dinucleotide, and the most common 8-nt sequence for insertions is ATATATAT, the underlined TA dinucleotide being the actual insertion site. However, for SB100X, only 1.8% of insertions occur in this 8-nt sequence. The mutant SB(K248R) has a much more pronounced preference for the ATATATAT sequence, ~33% of insertions catalyzed by this mutant occur there. We reasoned that combining the inherent sequence preference of this mutant with targeting by dCas9 could result in a synergistic effect. We thus looked for unique, targetable

sites embedded in simple TA_n repeats to target with SB(K248R) instead of SB100X and identified three such sites (Figure 3.14A). Two of these sites, which we called TA1 and TA2, are located on chromosome 21. TA1 has eight potential target sequences embedded in a highly repetitive TA-stretch (one of which has a single mismatch to the others), which we speculated could further favor targeting by promoting multimerization of the transposase at the target site. A third target meeting the criteria was identified on chromosome 7 and called TA3. For detailed descriptions of the target sites, see Supplementary figure 3.

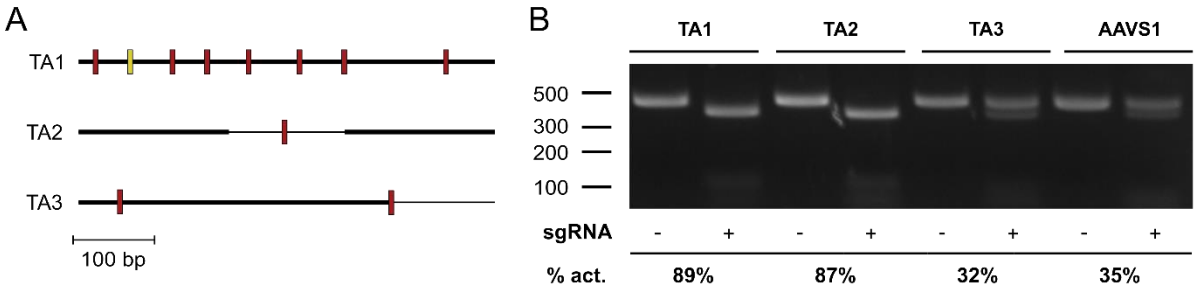


Figure 3.14 – *In vitro* validation of TA_n sgRNAs. A Schematic representation of the three target sites. Red boxes indicate sgRNA binding sequences, the yellow box indicates a binding sequence with a single mismatch. Thick lines indicate repetitive TA_n DNA, thin lines indicate non-TA_n DNA. The entire length of the lines represents ca. 500 bp of DNA. **B** DNA fragments of ~500 bp were digested using Cas9 and three TA-site specific sgRNAs, AAVS1 was included as a positive control. The expected sizes of the fragments were ~400 bp and ~100 bp; the 100 bp fragments are almost undetectable due to their small size and low overall DNA concentration. The sgRNAs directed against TA1 and TA2 mediate almost complete digestion of the target DNA, while the sgRNA against TA3 has a lower, but still clearly detectable activity.

For validation of sgRNA efficiency, it was not possible to use a T7 assay as for the GSH/'hotspot'-targeted sgRNAs because the repetitive nature of the targets made it difficult to amplify them by PCR. For this reason, the sgRNA target sequences were cloned into plasmid vectors and the sgRNAs were tested *in vitro*. For this purpose, a ~500 bp stretch of the plasmid containing the target sites was amplified by PCR and digested with purified Cas9 and *in vitro* transcribed sgRNAs. A previously verified sgRNA targeting the AAVS1 locus and its target site were used as a positive control. All sgRNAs were capable of inducing cleavage of their target sites. The sgRNAs sgTA1 and sgTA2 resulted in almost complete digestion of the target DNA, while the sgTA3 mediated cleavage at a level comparable to the positive control sgAAVS1 (Figure 3.14B). Due to the reasonable activity detected in an *in vitro* assay, all three sgRNAs were used for targeting assays in combination with SB(K248R) plus adapter proteins or the direct fusion transposase dCas9-SB(K248R).

3.1.13 TA_n -targeted integration libraries

In order to test whether the combination of the mutant transposase SB(K248R) and sgRNAs targeting sites embedded in TA_n repeats would have a synergistic targeting effect, new

integration libraries were generated as described before. Targeting was tested with dCas9-SB(K248R) and with the adapter proteins dCas9-N57 and dCas9-N123 in combination with SB(K248R).

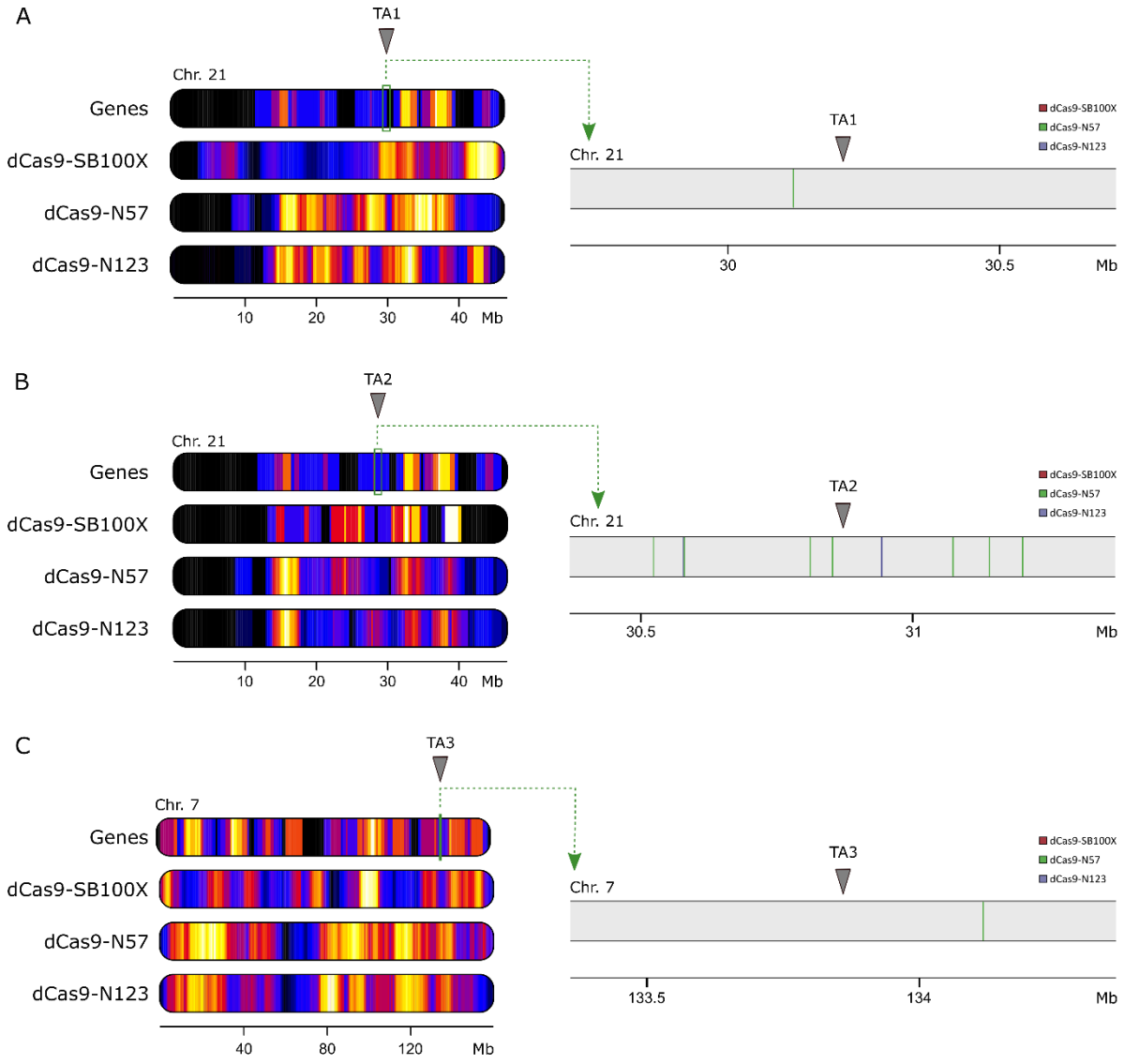


Figure 3.15 – Targeting of TA$_n$ repeats with SB(K248R) A Targeting of TA1. Distribution of genes as well as of insertions generated with the different targeting constructs along chromosome 21 is shown on the left, distribution of insertions around the target site on the right. T **B** Targeting of TA2. **C** Targeting of TA3.

Like in the previous targeting attempts, no insertions could be recovered from the target sites. For TA1, the closest insertion was 90.1 kb away, with dCas9-N57/sgTA1, the closest insertion with dCas9-N123/sgTA1 occurred at a distance of 903 kb and the closest insertion with dCas9-SB(K248R)/sgTA1 was 1037 kb away. The closest insertion at TA2 occurred at 18.2 kb and was found in the dCas9-N57/sgTA2 dataset; for dCas9-N123/sgTA2 the closest insertion was 72.4 kb away and the closest dCas9-SB(K248R)/sgTA2 insertion occurred at 831 kb. The insertion closest to TA3 on chromosome 8 occurred 259 kb away and is from the dCas9-

N57/sgTA3 dataset, the closest insertions for the other datasets were 560 kb away for dCas9-N123/sgTA3 and 2728 kb away for dCas9-SB(K248R)/sgTA3.

The large distances to the targeted sites clearly show that no targeting effect was achieved for the TA_n-embedded targets. Like in previous experiments, the integration pattern of the direct fusion transposase was distinct from the patterns generated with an adapter protein and unfused transposase. For targeting of TA1, the direct fusion again seemed to preferentially integrate into a region near the end of the chromosome. While the insertion patterns for both adapter proteins are similar for both targeting of TA1 and TA2, it should be noted that the patterns differ between those two targeting experiments, despite the fact that both targets are located on the same chromosome. This might suggest that the sgRNA used in the experiment might have some influence on target site selection, even if it does not direct insertions to the selected target sites. The large region near the start of chromosome 21, into which almost no insertions are mapped and where no genes are annotated corresponds to the short arm of the chromosome that contains the NOR, which contains mostly repetitive DNA and is not part of the standard genome assembly.

3.1.14 Generation of reduced-affinity SB mutants

Table 3-1 – List of residues selected for the SB mutagenesis screen. Residues in the first group are positively charged and are located close to the tDNA binding site of the SB transposase. Residues in the second group are positively charged and exposed on the surface of the enzyme. Residues in the third group are not positively charged, but have previously been identified as potentially being involved in tDNA interaction. Note that some residues are both surface exposed and close to the tDNA binding site and are consequently present in both categories.

Positively charged **Close to tDNA binding site**	**Positively charged** **Surface exposed**	**Not positively charged** **Likely tDNA interaction**
K156	K128	H127
K186	K129	N245
K248	R131	P247
K339	R166	Q271
	K186	N296
	K248	
	K259	
	R293	
	K339	

Due to the very limited efficiency of retargeting observed in experiments so far, we decided to look into modification of the transposase as a way to improve targeting. A severe limitation on

SB retargeting is that the number of integrations that can be targeted is very low compared to the overall number of insertions due to the high background activity of the transposase. Thus, we decided to look for a transposase variant that is less efficient at transposition by itself, but recovers some of the lost activity when it is directed to a target site by a fused DNA-binding domain. A similar mutant has been previously described for the PB transposon system[193] and in one study the use of this mutant seemed to positively influence the targeting outcome[409]. An analogous strategy, based on reducing non-specific DNA contact, has also been employed in generating a Cas9 variant with reduced off-target activity[268].

In order to decrease the DNA affinity of the transposase, we decided to reduce the overall positive charge of the enzyme by replacing positively charged residues that are close to the tDNA binding site or surface exposed. In addition, we planned to replace other amino acid residues known to play a role in tDNA interaction. A list of these residues can be found in Table 3-1. The residues are located along the length of the SB catalytic domain (Figure 3.16B)

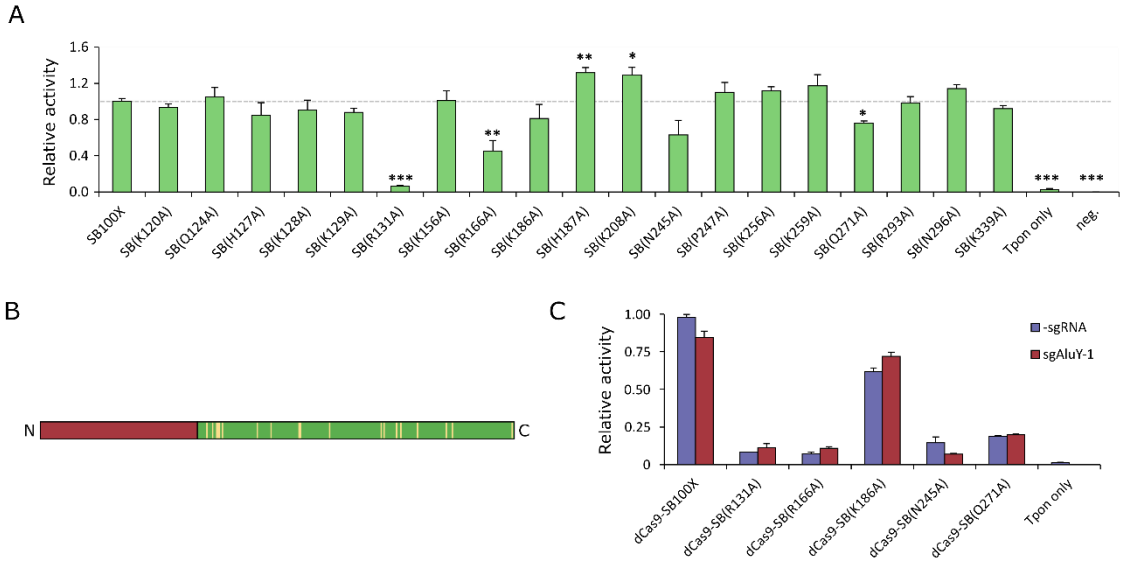


Figure 3.16 – Transposition assays with SB single residue mutants. A Activities of all single residue mutants, relative to SB100X. * p≤0.05, ** p ≤ 0.01, *** p ≤ 0.001 **B** Distribution of the replaced residues (yellow lines) along the length of the SB catalytic domain (green box) **C** Activities of five single residue mutants fused to dCas9, with and without sgAluY-1.

Each of the residues in the list was replaced with alanine and the resulting mutants were tested in a transposition assay (Figure 3.16A). A strong reduction of activity could be observed for SB(R131A) and some reduction was observed for SB(R166A), SB(K186A), SB(N245A) and SB(Q271A). For all other residues, only marginal reduction in activity or increases in activity were observed. In order to test whether the observed activity loss of any of these five mutants could be rescued by fusion with a DNA-binding domain, they were fused to the C-terminus of

dCas9 in a manner analogous to the generation of dCas9-SB100X as described in sections 2.2.1.2 and 3.1.1). The transpositional activity of those fusions was compared when co-expressed with sgAluY-1 or no sgRNA (Figure 3.16C). The sgRNA sgAluY-1 was selected due to the assumption that a high number of recognition sites would lead to a more pronounced effect. One observation that could be made is that the reduction in activity caused by the mutations seems to be more pronounced in the context of fusions with dCas9, all fusions except for dCas9-SB(K186A) showed significant drops in activity when compared to dCas9-SB100X. However, the addition of sgAluY-1 caused no significant differences in transpositional activity for any of the mutant fusions.

3.1.15 Generation of a random SB mutant library

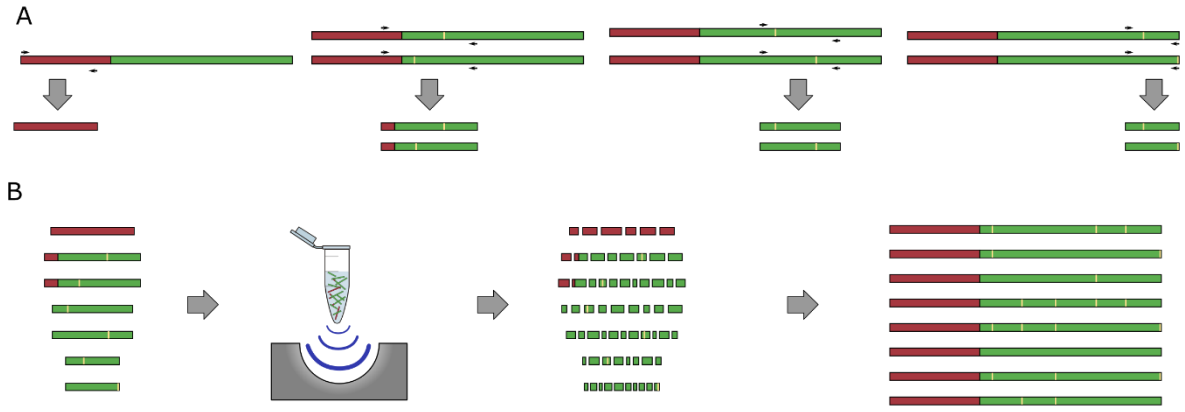


Figure 3.17 – Generation of a SB mutant library. A Four segments of the SB transposase sequence are amplified by PCR reactions. For each single nucleotide replacement mutant, the segment containing the mutation (yellow line) is amplified. The first segment contains no mutations and is amplified from the wt SB100X sequence. **B** Corresponding amplicons of the same segment from each mutant are pooled and pooled DNA is fragmented by sonication. Sonicated DNA of all four segments is pooled in a ratio corresponding to the sizes of the amplicons and then reassembled using a DNA polymerase, resulting in a library of randomly combined sequences containing multiple mutations.

Since none of the single nucleotide replacement mutants tested so far exhibited the desired phenotype, the mutations in the SB100X sequence were recombined in a random manner. In order to make sure that the recombined sequences would have, on average, more than one mutation, the SB sequence was subdivided into four segments (Figure 3.17A). The first segment, which does not contain any mutations, was PCR amplified from the wt SB100X sequence. For each single nucleotide SB mutant, the segment containing the mutation was PCR amplified and the PCR products of each segment were pooled. The pooled DNA was fragmented by sonication and mixed together in amounts corresponding to the length of each segment (Figure 3.17B). Finally, the fragmented DNA was reassembled in a primerless reaction using a polymerase.

The mixture of randomly assembled SB sequences was cloned into a dCas9 backbone to generate constructs analogous to dCas9-SB100X. Some constructs from this mixed pool were sampled to verify that each sequence indeed contained several mutations (Supplementary Figure 4A). The mutants from this library were found to have an average of 3.1 mutations (Supplementary Figure 4C), although a minority of mutations seems to have been caused by errors in the reassembly procedure and were not part of the set of intended amino acid replacements.

A separate mutant library was generated in an analogous fashion, using only mutants in which a positively charged residue (lysine or arginine) was altered as input. This was intended to generate some mutants with a more pronounced loss of positive charge when compared to those in which neutral residues have been replaced. This K&R mutant library was analyzed in the same manner as the first library (Supplementary Figure 4B) and was found to have a similar distribution of mutation numbers with an average of 2.8 mutations (Supplementary Figure 4C).

3.1.16 Screening of the mutant libraries

To test whether any of the randomly generated mutants exhibited the desired phenotype, i.e. increased transposition efficiency when combined with an active DBD, the dCas9 fusions of the library containing all 19 amino acid replacements (SB(ML)) and the library containing only the 13 lysine or arginine replacements (SB(KR)) were tested for transpositional activity in the presence or absence of sgAluY-1.

A total of 384 dCas9-SB(ML) and 480 dCas9-SB(KR) mutants were isolated from the libraries and co-transfected with the transposon pT2HB/GFP and either a plasmid expressing sgAluY-1 or an empty sgRNA scaffold. Due to the scale of the experiment, this was not done in duplicates. Fluorescence values were measured after two weeks of cell culture by high-throughput FACS analysis. Each mutant was scored according to the relative fluorescence in the reactions with and without sgAluY-1. Based on these scores, a list of 29 high priority and 49 low priority candidates was generated to be tested in replicates.

The screening was repeated as described above, in triplicates for the high priority candidates and in duplicates for the low priority candidates. From this analysis, two candidates, C5 and C42, were identified to have a significantly increased activity when co-transfected with sgAluY-1 (1.8-fold increase with $p=0.03$ and 14-fold increase with $p=0.01$, respectively). Due to the higher fold-change ratio obtained with SB(C42), this mutant was selected for further analysis.

3.1.17 Transposition with dCas9-SB(C42)

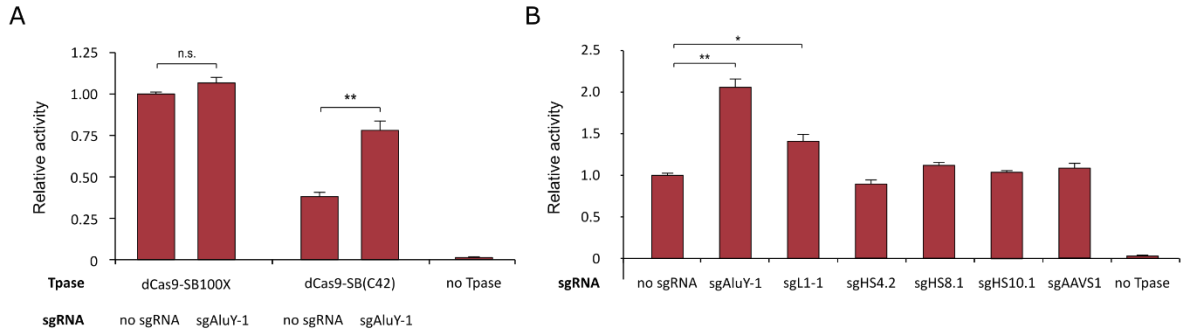


Figure 3.18 – Transpositional activity of dCas9-SB(C42). A Transpositional activity of dCas9-SB(C42) with sgAluY-1 or no sgRNA, relative to dCas9-SB100X with no sgRNA. Transpositional activity of dCas9-SB(C42) is significantly reduced compared to dCas9-SB100X ($p \leq 0.001$), but can be increased approximately two-fold by addition of sgAluY-1. ** $p \leq 0.01$ **B** Transpositional activity of dCas9-SB(C42) with different sgRNAs, relative to dCas9-SB(C42) with no sgRNA. Presence of either multicopy sgRNA increases transpositional activity, ca. 2-fold for sgAluY-1 and ca. 1.5-fold for sgL1-1. Presence of single-copy sgRNAs does not result a significant increase in transpositional activity for any of the sgRNAs tested. * $p \leq 0.05$, ** $p \leq 0.01$

The transpositional activity of dCas9-SB(C42) was subsequently tested in a large-scale transposition assay with a puro-tagged transposon, to see if the effect observed in the high throughput screen could be replicated. The sgRNA sgAluY-1 was cloned into a dCas9-SB(C42) vector backbone and the transposition rate with this construct was compared to dCas9-SB(C42) without a sgRNA, with dCas9-SB100X serving as a reference (Figure 3.18A). The activity of dCas9-SB(C42) was determined to be significantly reduced to ca. 40% of the activity of dCas9-SB100X ($p<0.001$). However, addition of sgAluY-1 increased the activity ~2-fold to around 80% of the activity of dCas9-SB100X ($p<0.01$). The addition of sgAluY-1 to dCas9-SB100X resulted in a non-significant ($p=0.17$) increase of 6%.

Although the increase observed in this assay was significantly lower than the value observed in the high-throughput assay, this experiment demonstrated that an increase can also be seen in transposition on a larger scale and with antibiotic resistance as a readout. Subsequently, the effect of different sgRNAs when combined with dCas9-SB(C42) was tested. The sgRNA sgL1-1 has a significantly lower number of binding sites than sgAluY-1 (~5500 vs. ~300000, or ca. 2%). When comparing the sgRNAs, sgAluY-1 was again found to induce a ca. 2-fold increase in transpositional activity ($p<0.01$) and sgL1-1 increased activity around 1.5-fold ($p<0.05$). In addition to the multicopy sgRNAs, the effect of four single-copy sgRNAs which were previously verified in cell culture-based assays, was tested. However, none of the single-copy sgRNAs effected a significant increase in transposition activity.

Sequencing revealed that SB(C42) contained three mutations: K129A, K259A and R293A, all of them positively charged and surface exposed residues. K129 is located near the N-terminus of the catalytic domain, immediately downstream of the interdomain linker. The other two residues are found closer to the C-terminus of SB, K259 20 residues upstream of the second glutamic acid of the DDE triad and R293 14 residues downstream of it. Interestingly, none of these three mutations caused a significant drop in transpositional activity by themselves and all of them were positively charged, suggesting that the mechanism by which they reduce transpositional activity in combination might indeed be reduced overall DNA affinity of the transposase.

3.1.18 Targeting of single-copy loci with dCas9-SB(C42)

Although no significant increase in transpositional activity was observed with single-copy sgRNAs, targeting of single-copy loci was tested with sgRNAs sgHS4.2, sgHS8.1, sgHS10.1 and sgAAVS1 with dCas9-SB(C42). Instead of analyzing the genome-wide integration pattern by generating integration libraries, insertions were screened with a PCR-based procedure. Each PCR used one primer binding in the transposon ITR (facing outward) and a primer binding in the genome close to the targeted sequence. Each PCR was performed with either a downstream or upstream genomic primer. Due to the extension time of the PCRs, integrations within ca. 4 kb either side of the target site should be recoverable. As no positive controls (i.e., cell lines with SB insertions close to the target sites) were available, the PCRs were done using four different combinations of annealing temperatures and template amounts for each reaction. Each primer combination was also done on a sample of gDNA isolated from cells transfected with the same targeting construct but no sgRNA in order to test whether any observed amplification was specific. However, no primer combination resulted in specific amplification (Supplementary Figure 5A), indicating that specific targeting of single-copy loci remained impossible using dCas9-SB(C42) or that targeted integration events were so severely underrepresented that they were outcompeted by the wild-type loci in the pooled DNA.

3.1.19 Staged targeting with dCas9-SB100X and dCas9-SB(C42)

One possible explanation for the failure of single-copy targeting is that, due to the high activity of the SB100X domain, transposon integrations mainly occur at random TAs throughout the genome before the dCas9 domain has time to encounter and bind its target sequence. In order to test this hypothesis, targeting experiments of single-copy loci with dCas9-SB100X and dCas9-SB(C42) were repeated in a staged manner, supplying the transposon only after the transposase expression plasmids. This was meant to give the targeting constructs time to be

expressed and find their target sites. We reasoned that the relatively long dwell time of dCas9 its the target sites[456] might ensure that when the transposon plasmids were transfected, the target sites would already be occupied by dCas9-based targeting constructs. In this case, the SB PEC might form at the target site and thus increase the chance for integration there.

This targeting approach was tested with dCas9-SB100X in combination with the sgRNAs sgHS4.2, sgHS8.2, sgHS10.2 and sgHPRT and with dCas9-SB(C42) with the sgRNAs sgHS4.2, sgHS8.2, sgHS10.2 and sgAAVS1. Supplying the transposon plasmids 36 h after the targeting factors resulted in a significantly decreased transpositional activity (<5% of the same combination of targeting construct and sgRNA when co-delivered). The integrations generated this way were analyzed with a PCR-based assay as described in section 3.1.18. However, no integrations near the target sites could be recovered from samples with either targeting construct (Supplementary Figure 5B and C), indicating that the staged delivery of components did not increase targeting specificity to a level where targeting of single-copy loci becomes detectable by PCR performed on gDNA isolated from pooled cells.

3.2 Retargeting of SB by ribosomal localization

3.2.1 Characterization of NoLS-SB100X fusions

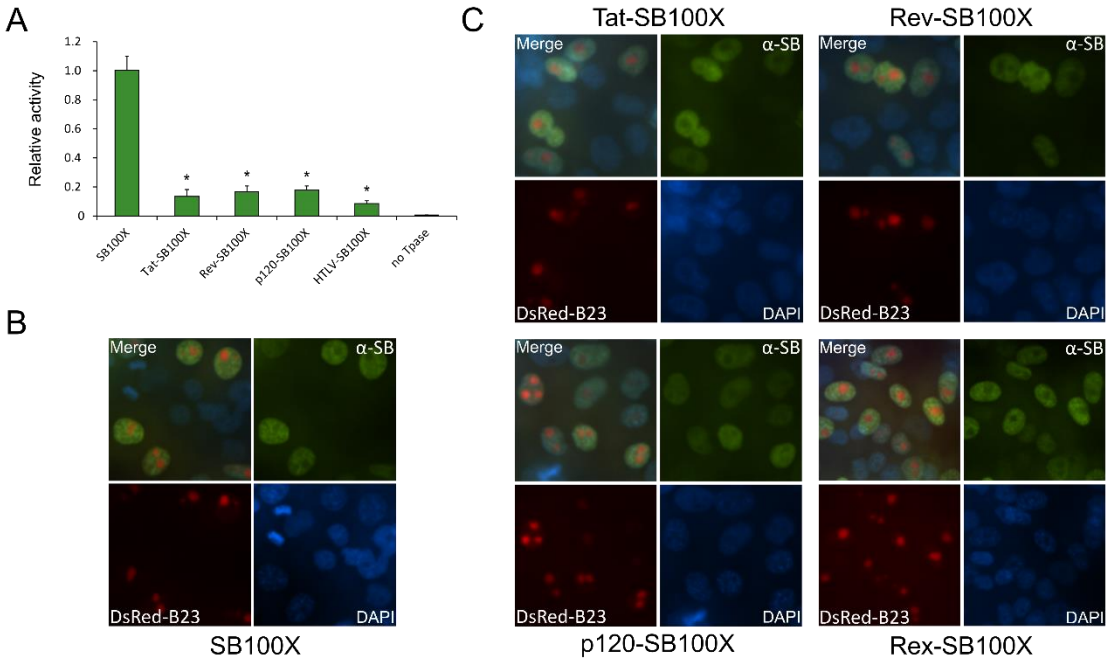


Figure 3.19 – Characterization of NoLS-SB100X fusions. A Transposition efficiencies of NoLS-SB fusions relative to SB100X. Activity is significantly reduced for all fusions. * p≤0.05 **B** Cellular localization of SB100X, with DsRed-tagged B23 as a nucleolar marker. SB100X is distributed across the nucleus and depleted in the nucleolus. **C** Cellular localization of the four NoLS-SB100X fusions. All fusions have the same localization pattern as SB100X.

Four different NoLS (Tat and Rev from HIV, p120 from human and Rex from HTLV) were fused to the N-terminus of SB100X in order to generate transposases that localize to the nucleolus (see section 2.2.1.5). The fusion proteins were tested in a transposition assay, with a puro-tagged transposon (Figure 3.19A). A significant reduction ($p<0.05$) in transpositional activity could be observed four all four NoLS-SB100X constructs, but activity was still detectable for all of them.

The subcellular localization of the NoLS-SB100X fusions was determined by IF staining of cells co-transfected with the NoLS-SB100X fusions and DsRed-B23, a nucleolar marker fused to a red fluorescent protein. No colocalization of the fusions with the nucleolar marker was observed for any of the NoLS-SB100X fusions; in fact, their localization was equivalent to that of unfused SB100X.

3.2.2 Characterization of B23-SB100X

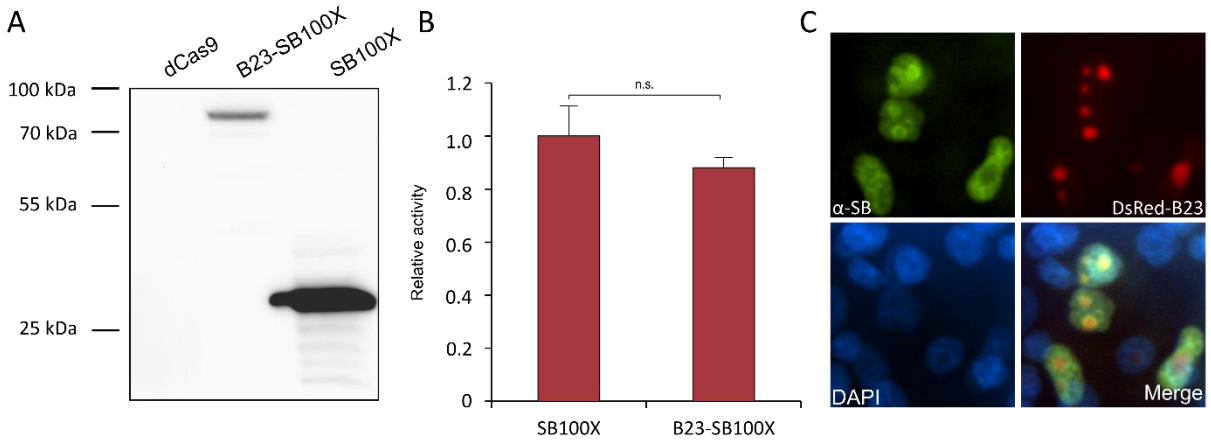


Figure 3.20 – Characterization of B23-SB100X A Western Blot of B23-SB100X using an α-SB antibody. dCas9 is included as a negative control and SB100X is included as a positive control. **B** Transposition assay with B23-SB100X, with SB100X as a reference. **C** IF images of cells co-transfected with B23-SB100X and pDsRed-B23 expression plasmids. B23-SB100X colocalizes with DsRed-B23 in some cells (top left cells), but is distributed like unfused SB100X in others (bottom cells).

After the failure of localizing SB100X to the nucleolus by fusing it to short NoLS sequences, a fusion of SB100X and the nucleolar protein B23/nucleophosmin, which is localized to the granular component of the nucleolus[457] by interaction with G-quadruplex structures of rDNA[458], was generated (see section 2.2.1.6). Western Blot analysis revealed that the fusion protein has the expected size and is expressed in HeLa cells (Figure 3.20A), but had a lower signal than unfused SB100X. A transposition assay was performed to test whether the fusion can still catalyze the entire transposition reaction; SB100X was used as a reference. The transpositional activity of B23-SB100X was not significantly reduced when compared to unfused SB100X (Figure 3.20B).

The subcellular localization of B23-SB100X was then tested by IF microscopy, in the same manner as described for the NoLS-SB100X fusions (Figure 3.20C). In contrast to the previous fusions, B23-SB00X could be shown to co-localize with DsRed-B23, but did not exclusively localize to the nucleolus. An SB signal could be detected from the entire nucleus, but intensity was highest in the nucleolus, whereas for unfused SB100X and the previous NoLS-SB100X fusions, the signal was depleted there. Nucleolar enrichment of B23-SB100X was also not observed in all transfected cells. B23-SB100X exhibited the same localization pattern when transfected by itself (rather than co-transfected with DsRed-B23), suggesting that the nucleolar localization in the co-transfected samples was not caused by an interaction with DsRed-B23 (data not shown).

3.2.3 B23-SB100X insertion libraries

SB insertion libraries were generated in the same manner as described in sections 2.2.6 and 3.1.4. Four different libraries were generated. HeLa cells were transfected with the B23-SB100X fusion or with SB100X. Each transposase was combined with either the transposon pT/SV40-neo or the transposon pT/HENA-neo. In the former, the neomycin marker is driven by a SV40 Pol II promoter, while in the latter the SV40 promoter is replaced with a 45S rRNA promoter sequence, which drives transcription by DNA polymerase I, and an IRES. The reason for this replacement was the assumption that if the transposon is successfully directed to the nucleolus, transcription by Pol I might be more efficient, as this polymerase is more abundant in the nucleolus. Additionally, interaction between the Pol I molecules localized in the nucleolus and the promoter may result in an enhancement of the targeting effect.

In order to test whether a preference for integration into nucleolar DNA was achieved, the fractions of insertions from each library that occurred into several sequence sets were compared. The first parameter that was analyzed was insertion into the nucleolar organizer regions (NORs)[424]. NORs consist of rDNA repeats and the flanking sequences on the short arms of the acrocentric chromosomes 13, 14, 15, 21 and 22. Comparing the fraction of insertions occurring into NORs between the samples shows a 1.6-fold increase with B23-SB100X compared to SB100X with the SV40 transposon ($p=0.005$) and a 1.4-fold increase using the HENA transposon ($p=0.024$) (Figure 3.21A, Supplementary Figure 6A). Interestingly, a very high fraction of NOR insertions recovered from all datasets were mapped to the NOR of chromosome 22 (Supplementary Figure 6B), independently of whether they were from a targeted or untargeted dataset.

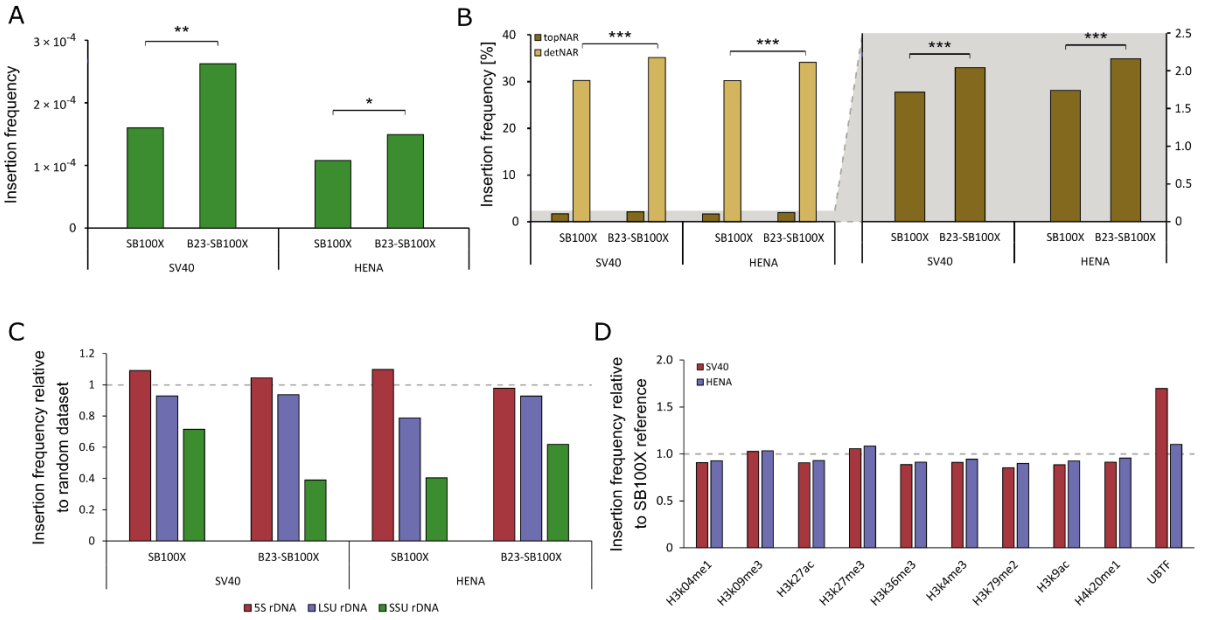


Figure 3.21 – Targeting of nucleolar DNA. A Insertion frequencies into NORs. **B** Insertion frequencies into two NAD datasets, topNAR and detNAR. Insertion into topNAR is shown again on the right with a rescaled y-axis. **C** Insertion frequency into rDNA genes as annotated by RepeatMasker, compared to a randomly generated control. **D** Insertion frequencies of B23-SB100X into significant peaks of ChIP-seq datasets from several chromatin marks as well as UBTF, compared to unfused SB100X. * p≤0.05, ** p≤0.01, *** p≤0.001

While NORs make up the main body of DNA in the nucleolus, a range of DNA sequences from other chromosomes has been shown to be associated with the nucleoli and form the chromatin shell of this compartment. These regions have been described as nucleolus-associated domains (NADs) and different datasets have been described as nucleolus-associated. Here, insertion frequencies into two of these datasets, topNAR[459] and detNAR[460], were compared (Figure 3.21B, Supplementary Figure 6C). Enrichment of ca. 25% was observed with the SV40 transposon and enrichment of ca. 19% was seen with the HENA transposon for the smaller topNAR dataset. For the larger sequence set, detNAR, enrichments of 16% (SV40) and 13% (HENA) were observed; all changes were statistically significant at p≤0.001.

The tool RepeatMasker contains an option to mask rDNA sequences dispersed throughout the genome, i.e. non-NOR copies of rDNA genes. Using these sequences as a target region allows to see whether enrichment into rDNA genes would occur independently of their position in NORs. RepeatMasker contains three separate tracks for 5S-rRNA, which is always located outside of the NORs, and LSU- and SSU-rRNAs (large and small subunit, respectively), which are normally located inside the NORs. The LSU-rRNAs comprise the 5.8S and 28S rRNAs while the SSU-rRNA is equivalent to the 18S rRNA.

Comparison of insertions frequencies between the SB100X and B23-SB100X datasets to a randomly generated control dataset show that insertions into 5S rDNA occur at close-to-random

frequency, while SB insertions seem to be slightly depleted in LSU rRNA and strongly depleted in SSU rDNA (Figure 3.21C and Supplementary Figure 6D). For 5S and LSU, the fraction of insertions obtained with B23-SB100X and unfused SB100X is similar. Comparing the fractions of insertion into SSU-rRNA results in a depletion of B23-SB100X insertions with the SV40 promoter and an enrichment of insertions with B23-SB100X and the HENA promoter, relative to the same transposon with unfused SB100X. However, it should be noted the enrichment in B23-SB100X with HENA still represents a lower fraction of insertions than both the random dataset and unfused SB100X with the SV40 transposon. Additionally, one has to consider the fact that the SSU-rDNA target region is significantly smaller than 5S-rRNA or LSU-rDNA target regions, thus higher variation is to be expected. This points to the fact that while enrichment with B23-SB100X occurs into NORs, non-NOR rDNA is not a preferred target.

Enrichment of insertions was also tested for significant peaks from a ChIP-seq dataset using the protein UBTF, a transcription factor of DNA polymerase I, which also localizes to the nucleolus. Integration into sequences recovered from ChIP-seq with other chromatin marks was also analyzed. This analysis revealed that SB insertions seem to be generally enriched in the UBTF dataset when compared to a randomly generated control (Figure 3.21D, Supplementary Figure 6E). Comparing the insertion frequencies of B23-SB100X and unfused SB100X showed only minor changes for other chromatin marks, but a 1.7-fold increase of insertions for the UBTF dataset when using the SV40 transposon, although the enrichment was not quite significant at $p \leq 0.05$ level ($p=0.055$). Interestingly, no enrichment could be seen with the HENA transposon, where only a minimal increase (1.1-fold) was observed.

3.3 HDR enhancement with Cas9 fusions

3.3.1 Characterization of Cas9-N57 and Cas9-N123 fusions

As described in section 1.4.1, one of the challenges of gene insertion with nuclease-based methods is the low efficiency of HDR compared to NHEJ. One potential approach to increase the HDR/NHEJ ratio is to increase the local concentration of the HDR donor at the edited site, rather than inducing any global changes in the cell. Similarly to the way adapter proteins work in retargeting of transposons, it should be possible to generate nuclease variants that have additional DNA-binding domains which recruit the HDR donor to the cut site by non-covalent interaction. In order to test this approach with Cas9 and components of the SB system, fusions between Cas9 and the SB subdomains N57 and N123 were generated using the procedure described in section 2.2.1.7. These constructs are functionally equivalent to the Cas9 fusions used to estimate dCas9 DNA-binding activity (see section 3.1.3), with the exception of

containing a BFP sequence in the same ORF as the Cas9 constructs, separated by a P2A sequence (Figure 3.22A).

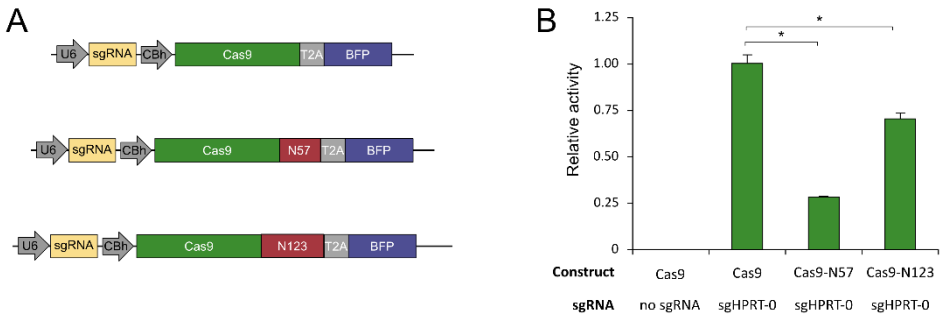


Figure 3.22 – Characterization of Cas9 fusions. A All three constructs generated for use with the TLR system contain a sgRNA targeting the Rosa26 sequence inserted into the Venus ORF in the TLR target. Cas9 expression is coupled to expression of a blue fluorescent protein via a P2A self-cleaving peptide in order to allow pre-gating of cells expressing Cas9 during FACS analysis. **B** Cas9-N57 and Cas9-N123 were tested in an *HPRT* disruption assay. Both constructs suffer a significant loss of cleavage activity compared to unfused Cas9, but remain active. ***** $p \leq 0.05$

Fusions were tested for expression and size in a Western Blot and the cleavage activity of the Cas9 domain was tested by a 6-TG disruption assay as described in section 3.1.8. Both Cas9 fusions had significantly reduced cleavage activity when compared to unfused Cas9, but still retained measureable activity (Figure 3.22B). Cas9-N57 had a slightly more pronounced reduction in cleavage activity to ca. 30% ($p<0.05$) while the activity of Cas9-N123 was slightly higher at ca. 70% ($p<0.05$).

3.3.2 Test of HDR enhancement using the TLR system

In order to test the ratio of HDR to NHEJ induced at DSBs catalyzed by the Cas9 fusion proteins, the traffic light reporter (TLR) system[435] was utilized (Figure 3.23). The TLR system consists of a reporter cell line with an editing target stably integrated into the AAVS1 locus and a compatible HDR donor. The target consists of two fluorescent protein genes. The first is a CAG-driven green fluorescent protein (Venus) sequence, which was inactivated by replacement of codons 117-152 with a sequence from the mouse *Rosa26* locus; it is followed by a P2A sequence and a red fluorescent protein (tagRFP) which are offset by 2 bp against the reading frame of Venus. An HDR repair donor containing the corrected Venus sequence is available, as is a sgRNA targeting Cas9 to the *Rosa26*-derived sequence in embedded in the Venus ORF. After Cas9 cleaves the target site, the DSB can be repaired either by NHEJ or by using the HDR donor. If the HDR donor is used, the wt Venus sequence is restored and the cell will emit green fluorescence. If the imprecise NHEJ pathway is used, small indels will be generated. In a third of these events, the tagRFP sequence will be shifted to be in-frame, resulting in red fluorescence of the cell. This allows the relative frequency at which the two repair pathways are utilized to

be analyzed by flow cytometry. In addition to the green and red output, the expression of the Cas9 constructs is coupled to expression of a blue fluorescent protein, allowing pre-gating for cells expressing the Cas9 enzyme.

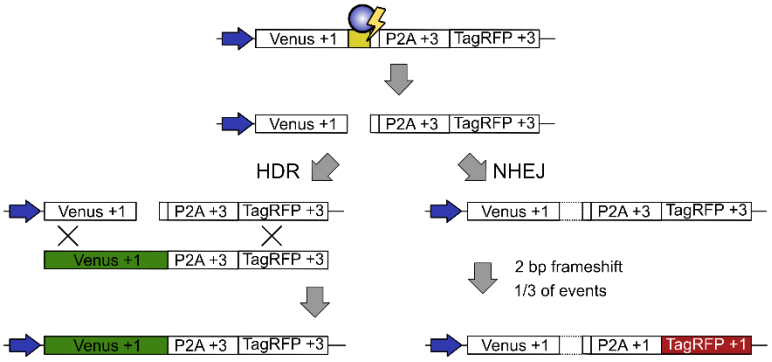


Figure 3.23 Principle of the TLR system – The TLR reporter construct consists of a Venus ORF, which is disrupted by replacement of codons 117-152 with a *Rosa26* sequence, as well as P2A and TagRFP sequences, which are shifted against the reading frame by two bp. Cas9 can be targeted to the *Rosa26*-derived sequence using an appropriate sgRNA, resulting in a DSB in the insert in the Venus sequence. The DSB can be repaired using the HDR pathway and the supplied donor construct (left), restoring the fully functional Venus sequence and resulting in green fluorescence. Alternatively, the DSB can be repaired using the error-prone NHEJ pathway, which results in small indels. In a third of NHEJ events, the P2A and TagRFP sequences are shifted into the reading frame, resulting in red fluorescence. Figure adapted from Chu et al. 2015[245].

The HDR donor used for these experiments was generated by amplifying the donor sequence from the plasmid pTLR_repair_vector using primers TLR_donor_fwd and TLR_donor_rev. HDR donors containing SB binding sites for interaction with N57/N123 at the left, right or both ends were made by an analogous PCR replacing the forward and/or reverse primer with SBBS_TLR_donor_fwd and SBBS_TLR_donor_rev, respectively (Figure 3.24A).

Cells transfected with the Cas9 constructs and the donors were analyzed by flow cytometry after 2 days. Cells were gated for blue fluorescence indicating successful transfection with the Cas9 construct and the eBFP positive cell population was analyzed for red and green fluorescence values (Figure 3.24B). The fraction of Venus positive cells was assumed to be identical to fraction of HDR-edited cells and the fraction of NHEJ-edited cells was calculated by multiplying the fraction of tagRed positive cells by three (Figure 3.24C). The ratios of HDR to NHEJ were calculated for each combination of Cas9 construct and donor (Figure 3.24D).

The HDR/NHEJ ratios were found to be significantly increased in samples using Cas9-N57 compared to samples using unfused Cas9 ($p<0.01$). This increased ratio can be attributed to both the significantly decreased efficiency of NHEJ ($p<0.01$) in these samples as well as a significant increase in HDR ($p<0.01$). However, this increase in HDR was not dependent on the presence of a SB binding site and in fact the overall efficiency was highest with the donor

containing no SB binding sites. The fact that the HDR/NHEJ ratio increased with the presence of 14DR sequences in the donor is due to the lowered NHEJ efficiency in these samples.

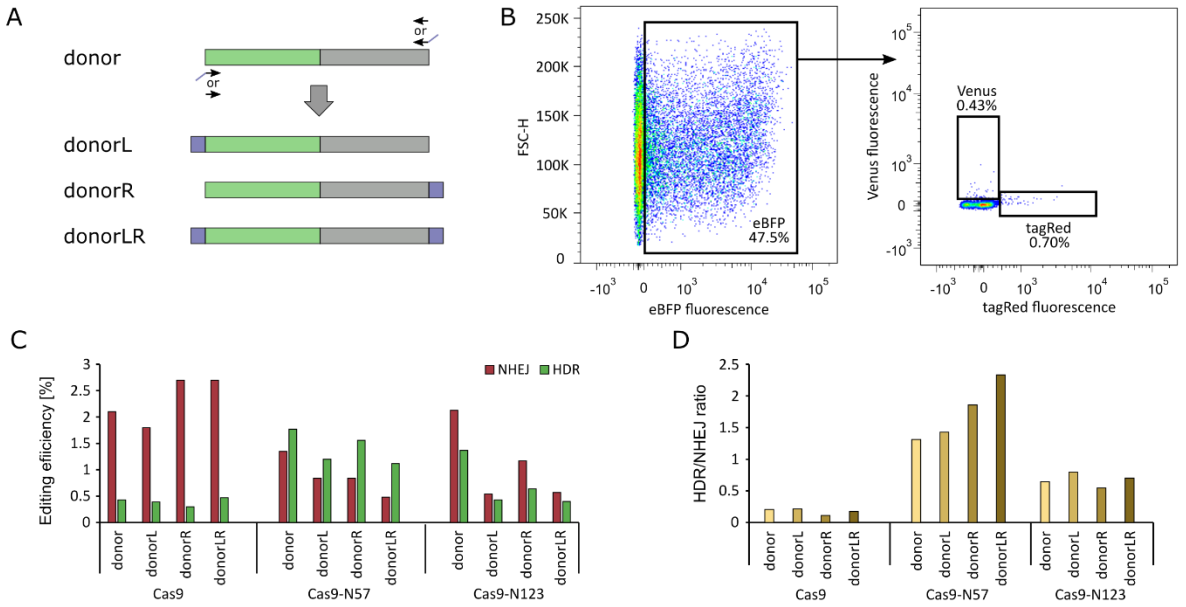


Figure 3.24 – Analysis of HDR enhancement with the TLR system. A Generation of the different HDR donor constructs. SB binding sites (blue rectangles) are added to the repair donor, which consists of the corrected venus sequence (green rectangle) and the reading frame-shifted tagRed sequence (gray rectangle). **B** Schematic of the gating strategy used, using the sample Cas9 + donor as an example. Cells are gated for blue fluorescence and then analyzed for red and green fluorescence. **C** Editing efficiency as percentage of eBFP positive cells. NHEJ efficiency is calculated as fraction of tagRed positive cells times three. **C** HDR/NHEJ ratios for all Cas9 constructs and donors.

A slight increase in HDR efficiency could also be observed with Cas9-N123 ($p<0.05$), but is it much less pronounced than for Cas9-N57. While the HDR/NHEJ ratios are significantly higher ($p<0.05$), this is mostly due to the drop in NHEJ efficiency compared to unfused Cas9 and the ratios are more similar to reactions mediated by unfused Cas9 than those using Cas9-N57. Again, the presence of SB binding sites in the donor was not a prerequisite for increased HDR/NHEJ ratios.

The fact that no SB binding site-dependent increase in HDR activity was seen either with Cas9-N57 or Cas9-N123 sites indicates that the strategy of increasing HDR efficiency by bringing donor templates into close proximity of the edited site by noncovalent interaction with the Cas9 fusion protein was not successful.

4 Discussion

4.1 RNA-guided transposition

4.1.1 Significance of targeted transposition

Given the plethora of different genome engineering systems available, it might be appropriate to ask whether generating an RNA-guided SB transposon system would in fact merit the effort that needs to be put into it. However, upon closer inspection, it becomes clear that such a system would fill a niche that is currently not covered by any other technology.

While the most commonly used integrating vectors are based on viruses, and they generally are the most efficient option, the necessity to package transgenes into viral particles makes them costly and difficult to handle when compared to transposon vectors and puts a strict size limitation on them. In addition, their often unfavorable insertion patterns result in an elevated risk of insertional mutagenesis and genotoxicity. This means that transposon vectors like the SB system, which have close-to-random insertion patterns, have a general safety advantage over viral vectors. However, even random integration can result in disruption of genes or regulatory elements, implying the risk of cancer formation in a clinical setting. This means that it is of great importance to develop systems that can deliver a genetic cargo to a clearly defined location.

As mentioned previously, several such systems already exist. Designer nucleases like ZFNs, TALENs and the CRISPR/Cas9 system are highly specific and enable editing at clearly defined genomic loci. In addition to this, the use of a HDR template makes it possible to introduce precise changes at a target site, including insertions. Thus, it might seem that HDR-directed repair of specifically introduced DSBs could make the use of integrating vectors obsolete. However, HDR-based insertion of sequences has several important limitations.

For one thing, the overall efficiencies of HDR-based editing are generally lower than those achieved with integrating vectors. The efficiency of HDR is also highly size-dependent[241]; although transposon integration efficiency also falls off with larger transgenes, the effect is generally less pronounced[137,461] and transposon vectors can integrate very large transgenes[91]. HDR is also limited to the S and G2 phases of the cell cycle[243], meaning that post-mitotic cells cannot be edited in this manner[242,244]. The SB transposon, on the other hand, is active in a wide range of cells, including non-dividing ones[187].

Additionally, the fact that nuclease-based editing will also produce a range of side products by NHEJ repair can be problematic in some contexts. The generation of DSBs in the genome can also have additional negative side effects, like p53 pathway activation[235,236] and the potential generation of large-scale genomic rearrangements[237]. Some DSB-free methods like prime editing can be used to generate insertions, but in this case the size limitation is even more restrictive than for HDR, making it impossible to insert entire genes.

Taken together, these considerations clearly show that there is a great need for a system that can integrate transgenes into clearly defined loci, without generating unwanted side products, particularly if the transgenes are large and if the target cell is non-dividing or does not have an active HDR pathway.

4.1.2 Construction of the targeting constructs

Two types of targeting constructs were generated, representing the two targeting strategies described in Figure 1.10: a direct fusion between dCas9 and SB100X and adapter proteins consisting of dCas9 and either N57 or N123, the subdomains of the SB transposase responsible for interaction with the transposon DNA and other transposase molecules. As previous studies have demonstrated that additions to the C-terminus of the SB transposase completely abolish transpositional activity[165,412,413], the dCas9 domain was only added to the N-terminus. The transpositional activity of the fusion transposase was reduced compared to unfused SB100X, but the fusion retained a reasonable amount of transpositional activity when compared with SB fusions to other DBDs[165,412,413].

Interestingly, the activity levels of fusions generated in this study vary significantly with the type of sequence added to the transposase. The addition of several small peptides caused a drastic reduction of transpositional activity (Figure 3.19), while the addition of the B23 sequence resulted in a fusion construct with no significant loss of activity (Figure 3.20), with the dCas9 fusion having an activity level between these two extremes. This demonstrates that the effect of N-terminal addition of domains to the SB sequence cannot be easily predicted and has to be determined experimentally for each construct generated.

An additional consideration that has to be taken into account for targeting of transposition with direct DBD-transposase fusions is whether a high activity of the transposase is even desirable. While high efficiencies are generally aimed for in genome engineering procedures, there might exist a tradeoff between high activity and targeting efficiency. The fact that the transposase in the fusion protein has a very high number of target sites available while the DNA-binding

domain that is meant to specify the target has a significantly lower number of target sites might mean that integration generally occurs at a random TA site before the targeting domain has encountered its target. Thus, a lower activity might be advantageous for targeting. If, on the other hand, the efficiency of integration at the target site is also reduced, lower activity might fail to have a positive effect on targeting. This consideration was the basis of testing targeting with an adapter protein and SB10 (instead of SB100X) and of the attempt to generate a transposase mutant with reduced untargeted activity.

Like SB100X, the dCas9 domain also lost some activity in the transposase fusion construct, as measured by the cleavage activity of Cas9-SB100X (Figure 3.3). However, the reduction to ca. 50% activity in the cleavage assay might not perfectly represent the difference in binding activity between dCas9-SB100X and dCas9, as it is possible that cleavage itself, rather than DNA binding is inhibited by addition of SB100X to the C-terminus of Cas9.

While the activities of both domains in the direct transposase fusion remained measurable, a loss of DNA-binding activity of N57 occurred in one of the fusion proteins; for N57-dCas9, no DNA-binding activity could be detected. The construct dCas9-N57 retained some DNA-binding activity, although at a much lower level than unfused N57. While this loss of activity may seem drastic, it should be noted that the DNA-binding activity of the isolated N57 is significantly higher than the DNA-binding activity of the domain in the context of a SB transposase molecule. Thus, the extremely high binding affinity of unfused N57 does not necessarily need to be matched for targeting to occur. Like dCas9-N57, dCas9-N123 retained a measureable level of tDNA interaction via its N123 domain.

Like in the direct transposase fusion, the Cas9 domain retained measurable cleavage activity after the addition of N57 or N123, with the addition of N57 causing a slightly greater loss of Cas9 cleavage activity than SB100X and the N123 fusion retaining a slightly higher activity level. Again, the relative binding activities might be higher than the relative cleavage activities measured in this assay if it is cleavage, rather than binding, which is inhibited in these fusions.

Overall, the functional testing revealed that, with the exception of the DNA-binding activity of N57 in the N57-dCas9 fusion, the activities of all domains remained measureable and were judged to be sufficient to exert some effect on target site selection.

4.1.3 Targeting of SB insertions in the human genome

4.1.3.1 Targeting to single-copy sites

Over the course of this study, seven single-copy sites were tested as targets: the *HPRT* gene, three GSH/'hotspot' sites and three sites embedded in TA_n repeats.

The *HPRT* gene was chosen for the availability of a simple counterselection procedure rather than for any properties related to SB transposition and no targeting effect could be observed for the target site. Because the sgRNA used in these experiments, sgHPRT-0, mediated highly efficient Cas9 cleavage *in vivo*, the failure to generate targeted insertions cannot be attributed to the sgRNA. One possible explanation for the failure to target any insertions to *HPRT* might be that the locus does not represent a good transposition target, even though it has a relatively high TA content at around 60%. No insertions near the target site could be detected in a dataset of untargeted SB insertions and a previous attempt to target *HPRT* with a fusion between a TAL domain and N57 also failed to generate insertions there (unpublished data). Additionally, a separate study attempting to target the *HPRT* gene with a dCas9-PB fusion also could not generate any targeted insertions there (although it was successful with ZFP- and TALE-based targeting)[407]. On the other hand, analysis of the distribution of insertions along the X chromosome in this study showed that the wider region around the target site is not particularly disfavored.

The uncertainty of whether some sites may, for unknown reasons, be disfavored for SB integration and would thus not be efficiently targetable, led us to address this question by targeting sites that are known to be receptive to SB insertions, picking three genomic loci at which three independent SB insertions had occurred without any targeting. In order to make sure that the sites, should they be found to be easily targetable, would also be candidates for targeting in a gene therapy context, we selected sites that fulfill GSH criteria[341]. However, none of the three sites, termed HS4, HS8 and HS10 (after the chromosomes they are located on) could be targeted. The targeting factors dCas9-SB100X, dCas9-N57 and dCas9-N123 were used in combination with a sgRNA for each of the three loci and no insertions were recovered from either the targeted loci or any mismatched sites. As for sgHPRT-0, the sgRNAs sgHS4.2, sgHS8.2, sgHS10.1 were shown to have high *in vivo* activities in combination with Cas9, indicating that they were not the weak point in the system. Also excluding the possibility that the selected sites are poor targets for the SB system in general, this suggests that the targeting system itself lacks the required specificity to direct insertions to single-copy target sites at a detectable level. However, it should, be noted that although the sites were selected on the basis

of several insertions occurring there in an untargeted dataset, in this experiment they did not seem to be particularly favored. This might simply be a statistical effect, the sites may still have been accessible but simply did not receive any insertions by chance. Another explanation might be that the dataset from which the sites were identified as favored targets was based on T cells while the targeting experiment was performed in HeLa cells. Chromatin differences between these cell types might explain why the GSH sites were less preferred this time.

One approach chosen to attempt to increase the specificity of the system sufficiently to allow targeting of single-copy loci was to use a mutant that already has increased specificity compared to SB100X. While the presence of a TA is the only strict requirement for SB insertion, there is some preference for a consensus sequence of ATATATAT, with integration occurring at the underlined TA dinucleotide. While there is almost no nucleotide preference for positions 2 and 5, an A at position 1 and a T at position 8 are clearly favored. However, only 1.8% of total SB insertions occur at this consensus sequence. The mutant transposase SB(K248R), on the other hand, has a more pronounced preference for this octanucleotide, inserting there in 33.3% of cases (Kesselring et al, unpublished manuscript). We attempted to use this to our advantage by targeting insertions catalyzed by SB(K248R) (as a fusion with dCas9 or in combination with an adapter protein) to unique sites embedded in TA_n repeats, where the consensus octanucleotide would occur with a high frequency.

However, none of the three loci TA1, TA2, or TA3 could be targeted at a detectable level. This might reflect a general failure of the strategy of combining two systems with distinct target site preferences. Sequence logos show that the preference for the consensus sequence is conserved in the targeted datasets (data not shown); this might indicate that the dCas9-based components did not exert a strong effect on target site selection. On the other hand, a clear difference in the insertion patterns of the direct fusion transposase and unfused transposase with adapter proteins could be observed, although it occurred in a sgRNA-independent manner. This suggests that the use of a fusion protein influenced the target site selection while still retaining the target site preference of SB(K248R). The fact that the distribution of insertions along the same chromosome was different for datasets containing sgTA1 and sgTA2 might also indicate that the distribution is to some extent sgRNA-dependent, even if the intended target sites do not receive any insertions.

As for the *HPRT* gene, no previous SB insertions could be found at these loci, so it is also possible that some property of the loci themselves makes them unfavorable for SB insertion (although, they are naturally very AT-rich and should present ample possible positions for SB

integrations). Additionally, as the target loci could not be PCR amplified due to their repetitive nature, no cell culture-based assays like the T7 endonuclease assay of the TIDE assay could be performed to verify the efficiency of the sgRNAs. All three sgRNAs were found to be reasonably efficient at digesting synthetically generated target DNA *in vitro*, but it is possible that their *in vivo* efficiency is too low to be used for targeting. Apart from being problematic for *in vivo* testing of the sgRNAs, the repetitive nature of these target sites could also be problematic for mapping of insertions there. Even if an insertion is recovered from these sites, it might be problematic to map it to the genome, especially if the reads are relatively short. Finally, it needs to be mentioned that while ATATATAT sequences are obviously a lot less common in the genome than TA dinucleotides, they are still present in a significantly higher number than the unique sgRNA target sequences. Thus, it is still possible that the mutant transposase simply integrates at any of these abundant targets without ever encountering the dCas9-defined target.

One observation that could be made for several of the single-copy targeted insertion libraries is that the dCas9-SB100X fusion transposase does seem to have a distinct insertion pattern compared to SB100X in combination with adapter proteins, more or less independently of the sgRNA used. While the unfused transposase insertions are distributed relatively evenly along the chromosomes, with a slight correlation to gene density, in many cases the fusion transposase had clearly preferred genomic regions. In many cases, at least one of these regions of high integration density occurred near the end of the chromosome, although many libraries also had hotspots near the center of the chromosomes. The fact that this occurs independently of the sgRNA used suggests that it might be some form of unspecific binding of the dCas9 domain. Indeed, dCas9 has been shown to exhibit some gRNA-independent binding in ChIP-seq experiments[266]. In this study, dCas9 was found to associate with GC-rich and GC-skewed regions in accessible chromatin. It might be interesting to verify whether the regions where dCas9-SB100X integrated with increased frequency share some of these characteristics.

An additional manner of increasing the targeting efficiency at single-copy loci that was tested was the staged delivery of components. In these experiments, direct fusion transposases were delivered to the cells first and the transposon was only supplied after some time. The idea behind this was to allow the transposase fusions to be expressed and to occupy their target sites before the transposon becomes available. This was theorized to allow SB PECs to assemble at the target sites and favor insertion there. However, no specific targeting of any of the tested single-copy loci could be achieved this way. It should be noted that this method significantly reduced

the overall efficiency of the system, so only a relatively low number of insertions could be screened. It might be necessary to adjust the parameters of this type of reaction to allow a higher number of insertions to be catalyzed, e.g. by fine-tuning the time between delivery of the components.

4.1.3.2 Targeting to multicopy sites

While the attempted targeting of seven different single-copy sites failed to have any measurable effects, some changes in the integration pattern could be observed when sgRNAs with multiple binding sites in the human genome were utilized.

Changes in the integration pattern could be analyzed in particularly high detail for targeting of AluY, due to the high number of available target sites (ca. 300000). This high number made it possible to analyze even small effects statistically and to test changes in integration into small windows. Targeting with dCas9-N57 and sgAluY-1 resulted in a small, symmetric enrichment around target sites when compared to the same targeting construct with the sgRNA sgL1-1, but the change was not statistically significant even for the window in which enrichment was observed. However, samples using the targeting factor dCas9-SB100X and sgAluY-1 contained a significantly higher fraction of insertions into a 300 bp window downstream of the sgAluY-1 biding sites than samples containing dCas9-SB100X and sgL1-1. Asymmetrical integration around target sites is a phenomenon that has been observed in a range of other targeting studies, including studies that used similar dCas9-based targeting constructs like dCas9-Hsmar1[408] or dCas9-PB[409].

In general, an asymmetric integration pattern can be caused by the targeting constructs themselves, i.e. by the architecture of the hybrid proteins that are used to direct integrations to the target sites, or by the sequence or structure of the DNA around the targeting sites, e.g. the availability of potential integration sites on either side of the target. In the AluY targeting analysis, downstream is defined as the direction of Alu transcription. Using this definition, the target sequence of sgAluY-1 is on the top strand, the bottom strand has the same sequence as the sgRNA with the PAM facing to the left, or upstream. Taking into consideration what is known about the structure of Cas9 and the position of its termini, a fusion to the C-terminus of Cas9 is expected to be closer to the 5'-end of the target strand[462]. This would suggest that the architecture of the targeting factor should favor enrichment upstream of the target site rather than downstream, where enrichment was observed.

Given that the architecture of the transposase fusion itself seemed to be an unlikely explanation for the asymmetric enrichment around the target sites, we speculated that the nature of the flanking DNA sequences might be the cause. However, closer analysis showed that the window in which insertions were enriched was actually a disfavored target for SB insertion rather than a favored one. The low frequency of SB insertions into this window can be explained by lower TA frequency than in the surrounding DNA, which is in accordance with the fact that the Alu consensus sequence is relatively GC-rich. In addition to the low number of potential target sites, the region also has a high nucleosome occupancy and the presence of nucleosomes in defined positions of the Alu sequence has been previously reported[463]. The presence of a nucleosome at this site also disfavors SB insertion, as untargeted SB insertions have a tendency to occur into nucleosome-free DNA. It remains unclear why insertions generated with the active targeting factor are enriched in a region that is overall a poor target for SB. One potential explanation might be that the asymmetric distribution is caused by the fact that the enrichment is masked by the background events on the other side of the target and only becomes visible due to the lower background in the 300 bp region downstream of the sgAluY-1 binding sequence.

Due to the significantly lower number of sgL1-1 target sites, distribution around these sites could not be analyzed with the same level of detail as integrations around sgAluY-1 sites. However, while no statistically significant enrichment could be reported due to the relatively low insertion counts around these sites, some trends can be seen. As for sgAluY-1 target sites, the region immediately downstream of sgL1-1 sites is TA-poor; additionally there is a region of high TA density ca. 200 bp upstream of the target sites. While no clear statistically significant enrichment could be observed for any targeting condition, there seemed to be a slightly higher occurrence of integrations into the disfavored region of 0 to +250 bp in samples where the sgRNA sgL1-1 was provided. However, this effect can also be attributed to the higher number of insertions in the sgL1-1 datasets compared to the sgAluY-1 datasets. Indeed, for the targeting factor dCas9-N123, where a single-copy sgRNA was used as a control due to the lack of an equivalent sgAluY-1 dataset, the difference in numbers between the datasets is lowest and the observed effect is the weakest.

The comparison between the datasets generated with dCas9-N123 in combination with SB100X and SB10 showed no clear difference in distribution of insertions around the target sites. While a higher fraction of insertions occurred close to the target site with SB10 than with SB100X, the change is not significant and both datasets are depleted in the region compared to the control.

This demonstrates that simply reducing the overall activity of the transposase does not automatically improve the ratio of targeted to untargeted insertions. To positively influence this value, more specific changes to the transposase are likely required, as will be discussed in the next section.

Taken together, the results of the attempts to target multicopy targets in the human genome were better than targeting of single-copy loci, but were still clearly lacking in specificity. A statistically significant enrichment could be only observed for dCas9-SB100X with the sgRNA sgAluY-1, which has a very high number of target sites. Even reducing the number of target sites to several thousand by using sgL1-1 made it impossible to observe a targeting effect with any certainty. This clearly indicates that while the use of RNA-guided components to retarget the SB transposon is theoretically feasible, the specificity of the system in this form is much too low to be of any practical utility.

4.1.3.3 Engineering of a reduced-affinity transposase

The previously described lack of specificity of the targeting system is likely caused by the general, unspecific integration of SB100X. An ideal solution to this problem would be to remove the parts of SB transposase that are required for tDNA interaction and replace them with a target-specific DBD. Unfortunately, this approach is not applicable to the SB system because the tDNA-binding and catalytic functions are performed by the same C-terminal domain; simple removal of the intrinsic tDNA-binding activity of SB while still maintaining transpositional activity is not possible.

Thus, the approach taken in this study was to try to disrupt tDNA binding of the SB transposase by replacing selected residues in the C-terminal domain, leaving the residues required for catalysis unaffected. Apart from removing individual residues implicated in tDNA binding, positively charged residues were removed from the surface of the protein to reduce the overall positive charge, and thus the unspecific DNA affinity, of the enzyme. Single replacement of some of these residues resulted in reduced activities (which would be a likely consequence of lowered DNA affinity). However, these mutations did not experience an increase in activity when coupled to dCas9 with an active sgRNA, indicating that the transposition mechanism itself was likely disrupted by these mutations.

In order to avoid replacing residues that are required for efficient transposition, we settled on the strategy of randomly recombining mutations. While a single amino acid replacement may not significantly alter the DNA affinity of the enzyme (and the respective mutant transposase

may experience no drop or only a minimal drop in activity), replacing several residues might significantly reduce the DNA affinity of the enzyme without disrupting the transposition mechanism itself. Our assumption was that it would be possible to identify such a mutant by comparing the transpositional activity of the mutant in a dCas9 fusion with and without a sgRNA.

After screening several hundred SB mutants, a mutant with this phenotype was discovered and termed SB(C42); the fusion protein dCas9-SB(C42) experienced a 2-fold increase in transpositional activity when co-expressed with sgAluY-1 and a 1.5-fold increase with sgL1-1; no significant increase was observed with single-copy targeting sgRNAs. The transposase fusion dCas9-SB(C42) was subsequently tested for targeting of both multicopy and single-copy loci.

Targeting of single-copy loci with SB(C42) was tested with the sgRNAs sgHS4, sgHS8, sgHS10 and sgAAVS1 using a PCR-based procedure. The PCR conditions should have theoretically allowed to recover integrations at distances of up to 4 kb from the sgRNA target sites in either direction, however, no specific amplification was observed for any locus. While the absence of a proper control for these reactions means that one cannot conclusively say that no integrations occurred at the targeted loci, it seems likely that the specificity of the system remained too low to efficiently target single-copy loci, which was not entirely unexpected as the addition of single-copy sgRNAs to dCas9-SB(C42) did not result in a measurable increase in transpositional efficiency. Targeting of single-copy loci was also attempted using the staged approach mentioned in the previous sections, but no targeting could be observed. Again, this might indicate either a general failure of the strategy or be related to the drastic reduction of efficiency after staged delivery of the components.

Due to time constraints, analysis of targeting with dCas9-SB(C42) together with the multicopy sgRNAs sgAluY-1 and sgL1-1 could not be completed in time to be included in this thesis. The fact that higher transposition rates could be observed with dCas9-SB(C42) after the addition of these sgRNAs suggests that the addition of these sgRNAs might have also had an effect on the target site selection. However, whether the use of dCas9-SB(C42) improves targeting rates to the multicopy targets compared to dCas9-SB100X remains to be seen.

4.1.4 Potential improvements to the targeting system

Some modifications to the targeting constructs could help to increase the limited efficiency of targeting observed in this study. One possible modification to the structure of the transposase

fusion (and adapter protein) constructs would be the use of alternative linkers between the two parts of the fusion protein. While the 14 aa linker used here has been previously used in a fusion between Rep and SB transposase[164], and a slightly extended 18 aa linker worked well in a fusion between a ZFP and SB transposase[165], both of these DBDs are significantly smaller than dCas9. It is possible that a longer linker sequence would allow the SB transposase better access to the target DNA while the dCas9 domain is bound to its target site. Indeed, a retargeting study using a direct fusion between dCas9 and the PB transposon reported a minimal length of 23 aa to be necessary to allow for genomic targeting[409]. As the structures of the PB and SB transposases are unrelated, it is unlikely that such an observation made for a dCas9-PB fusion would be directly applicable to a dCas9-SB fusion. However, it might indicate that targeting with SB might also benefit from a longer linker sequence when fused to a very large DBD like dCas9.

Another potential area where the targeting components could be optimized is in the use of alternative dCas9-based DBDs. All the constructs generated in this study use a standard dCas9 domain, but a range of optimized Cas9 variants are available and might help to increase targeting efficiency. One simple modification might be using a dCas9 domain derived from a high fidelity Cas9 like spCas9-HF1[268]. This approach has been used in a retargeting study using dCas9-PB[409], although no construct with a standard dCas9 domain was included in this study, so it is not evident whether the use of dCas9-HF1 contributed to the successful targeting reported there.

A particularly interesting idea may be the use of Cas9 circular permutants[464]. These Cas9 variants have the same overall structure as Cas9, but have alternative N- and C-termini, allowing the addition of other domains to different parts of the Cas9 (or dCas9) protein. Some of these variants have termini that are closer to the bound DNA molecule than those of wt Cas9. Bringing the transposase into closer proximity to the DNA while dCas9 is bound might improve the efficiency of transposition at the target sites. The use of circular permutated dCas9 might be particularly beneficial for SB transposase mutants with reduced unspecific DNA affinity like SB(C42). Many of the mutations tested in the high-throughput screen in this study resulted in a significant drop in transpositional activity that could not be rescued by the addition of dCas9 and a sgRNA. While it is possible that for these mutants the transpositional mechanism is simply disrupted in a way that is not caused by reduced DNA affinity, it could also be the case that fusion with dCas9 did not bring the transposase molecule close enough to the target DNA to overcome the reduced DNA affinity. Thus, it might be interesting to test whether mutated

SB transposases with very low or no transpositional activity would regain some activity after fusion to a circularly permutated dCas9.

A strategy that has been successfully applied in another dCas9-based transposon retargeting study was the parallel use of several sgRNAs[409]. In this case, using 8 separate sgRNAs targeting the same locus resulted in successful retargeting, while using 4 sgRNAs did not. The increase of targeting efficiency with a higher number of sgRNAs could be explained by statistical reasons, by improved multimerization of transposase molecules at the target site or by the increased likelihood of a suitable integration sequence being available at favorable distance from one of the target sites. Multiple binding sites were present at the target loci of the sgRNAs sgTA1 (eight binding sites) and sgTA3 (two binding sites), but no targeting could be observed for either locus. However, these sgRNAs were only verified *in vitro* and it is not known whether the loci might be poor targets for SB integration overall. Thus, it might be interesting to repeat targeting to the GSH/'hotspot' sites by adding multiple sgRNAs that bind there.

Modifications to the existing targeting system, like changes in targeting factor architecture (e.g. alternative dCas9 domains or alternative linkers) could be implemented in the short term and they might result in increased targeting efficiencies. It is possible that such increases would be sufficient to raise the specificity of the system to a point where detectable targeted insertion into single-copy loci can be observed. However, it seems unlikely that these types of modifications would result in a system where a sizeable fraction or even a majority of insertions occur into a specific locus. A targeting system with an overall low specificity might still be useful for applications where it is possible to isolate, expand and genotype individual clones. However, directing a few percent of insertions into a site will not be sufficient for situations where the targeted cell population is meant to be used in bulk. To generate a highly specific system, more general changes in the targeting strategy are likely required. Such long-term alternative strategies will be discussed in the next section.

4.1.5 Outlook and conclusion

Overall, the results presented here show that, while it is possible to introduce a minor bias to the integration profile of SB by fusing the transposase with dCas9, the overall efficiency of this process is very low. The effects could only be observed by using sgRNAs with high numbers of target sites in the human genome, which would limit its use in most applications, where targeting to a specifically defined site like a GSH is desirable.

The low efficiency of retargeting observed here was, however, not completely surprising. Across the range of studies using the targeting mechanisms described here with many different transposases and DBDs, a low rate of targeting compared to background insertions was a commonly observed effect (reviewed in [318]). This can be explained by the nature of the components used: unlike catalytically active Cas9, which only adopts a cleavage-competent conformation when bound to target, or designer nucleases using FokI domains that become active upon dimerization at the target sites, the transposases used in these retargeting attempts do not require binding of the associated DBDs and will thus catalyze high numbers of untargeted insertions. The observations made in this and previous studies suggest that in order to achieve the required level of specificity for targeted transposition, transposase fusions or adapter proteins have to be combined with additional mechanisms.

Some recent studies using alternative systems have reported very high rates of RNA-guided transposition targeting with a low level of background insertions[410,411]. Strikingly, these rates have not been achieved by generating hybrid proteins consisting of a transposase and a DBD. Rather, they are based on naturally evolved systems that already have the desired targeting properties; they were discovered rather than invented. The clear difference in efficiency between most purely artificially retargeted transposase systems and these naturally evolved systems suggests that the key to a high targeted insertion ratio is the use of specialized components, rather than the mechanistically simple transposase fusions and adapter proteins presented in this study. Unfortunately, the highly efficient RNA-guided transposases mentioned above are only active in bacteria, and if they can be adapted for use in mammalian cells, it remains to be seen whether similarly high targeting ratios could be achieved in a significantly larger eukaryotic genome. While it cannot be strictly excluded that equivalent systems might exist in some eukaryotic organisms, the more promising approach would be to attempt to generate a system with similar properties from components already active in mammalian cells.

The generation of a reduced-affinity mutant was the main approach taken in this study to address the problem of high background insertions. While the mutant fusion transposase dCas9-SB(C42) apparently still did not exhibit the required level of specificity to target single-copy loci, it remains to be seen whether the observed increase in transpositional activity after addition of multicopy sgRNAs will result in increased targeting of these sites. However, it is clear that while transpositional activity of this mutant without a sgRNA was reduced, it was still significant. This means that even if targeting is improved, background insertions will remain an issue. Further engineering of the SB transposase might ultimately result in a transposase that

can be targeted in a conditional manner, i.e. which is only active when an associated DBD is bound to its target site. It is important to note that while a pool of residues that were selected based on structural data was randomly combined and almost a thousand mutants were tested, this number is still low compared to the numbers of variants that might arise during natural evolution. In order to generate a DBD-dependent transposase, it might be necessary to use methods that allow the testing of much higher numbers of variants.

One project currently under development attempts to use such an approach to screen large numbers of SB mutants for hyperactivity. It is based on a reporter cell line in which both excision and integration activity of SB transposase variants can be screened for via a double selection procedure. The aim is to generate a high number (~10^6) of random SB mutants by error-prone PCR and packing these into γ-retroviral vectors. To find hyperactive SB mutants, the reporter cell line would be transduced with the γ-retroviral mutant library, and cells in which transposition occurred would be isolated by the selection procedure. The SB sequences incorporated into these cells could be determined by next generation sequencing and comparison with the general representation of mutants in the mutant library should show which mutants have a hyperactive phenotype. It would be conceivable to use an alternative cell line expressing a sgRNA in addition to the reporter cassette and generating a viral library of the dCas9-SB(mut) mutant libraries presented here. This library could be transduced into the reporter cell line expressing the sgRNA and into the reporter cell line lacking a sgRNA and after selection and sequencing, some SB mutants may be enriched in, or exclusively recovered from, the cell line that expresses the sgRNA. Using this approach, it should be possible to screen a very large number of randomly recombined mutants in a way that was not feasible with the methodology utilized in this project.

However, it is also possible that the construction of such a mutant is not possible for the SB system. For this reason it might be interesting to look into other transposase systems that have more favorable properties than SB. As mentioned above, many highly specific systems like ZFNs and TALENs are based on the multimerization of components at the target sites. Providing two targeting factors with distinct target sequences greatly increases specificity, especially if the sequences by themselves are already relatively long, as for dCas9. This approach has been shown to be workable with highly specific dCas9-FokI fusions[465], which combine the easy RNA-guided targeting of the Cas9 system with the specificity of FokI-based designer nucleases. Unfortunately, a split SB transposase is not available to emulate this approach. Designing a split SB system that becomes active upon multimerization at the target

site would be challenging, as all the parts of the SB transposase are already required for excision of the transposon and, as mentioned above, the tDNA binding activity cannot be easily separated from the rest of the enzyme.

While such an approach seems hard to implement for the SB system, other transposons have properties that make them attractive for such a strategy. The Harbinger transposase, for example, requires a second, transposon-encoded Myb-like cofactor to function[466]. It might be possible to direct both the transposase and the Myb-like protein towards a target site by fusing each component to a dCas9 domain and supplying two sgRNAs that bind near the intended integration site. Specifying two target sequences, in combination with the fact that the Harbinger transposon already has a high degree of specificity (it preferentially integrates into a 15 nt sequence), might result in highly specific targeting. While the Harbinger transposon system is already a lot less active than the SB system, and fusing dCas9 domains to both its components will likely further reduce this activity, low activity might be an acceptable tradeoff for high specificity.

4.2 Targeting by ribosomal localization

4.2.1 Nucleolar localization constructs

Several constructs were generated in an attempt to produce a SB transposase that localizes to the nucleolus in order to promote integration into rDNA. These loci are highly expressed, which should ensure that transgenes integrated there are also expressed at relatively high levels, something that may not be the case if integration occurs into heterochromatin. Additionally, the fact that these genes are present in a high copy number in the human genome means that disrupting them should not negatively affect the target cells, even though rRNAs themselves are essential to the functioning of the cell.

Surprisingly, the addition of four different nucleolar localization signals to the SB transposase failed to exert any measurable effect on the localization of the protein. In fact, like for unfused SB100X transposase, the NoLS-SB100X fusions were depleted in the nucleolus rather than enriched. It is striking that this was observed for all four NoLS, which have distinct phylogenetic backgrounds: two of the nucleolar localization signals were taken from the HIV-1 proteins Tat and Rev, one from the human protein p120 and one from the Rex protein of HTLV.

HIV-1 Tat has been shown to exhibit either nucleolar[467,468] or nuclear[469] localization and it has been shown that the localization is concentration-dependent[470], with high concentrations

favoring nucleolar localization. This suggests that a low concentration level in the cell might contribute to a failure of Tat, and possibly proteins tagged with the Tat NLS/NoLS to localize to the nucleolus. However, expression from a transfected plasmid, which is the method that was used in the localization experiments, should generally result in a high level of expression compared to physiological levels. A previous study fusing the Tat NLS/NoLS to a large protein also reported a nucleoplasmic pattern[471], while other studies reported distinct nucleolar localization[441].

In contrast to Tat, HIV-1 Rev has been consistently shown to accumulate in the nucleolus, and fusion proteins containing the Rev-NoLS generally replicate this pattern[442]. The same can be said of the human protein p120; the NoLS of p120 has been shown to be sufficient to localize proteins to the nucleolus[443] if the protein also contains an NLS. Just like the previous two NoLSs, the sequence of HTLV Rex has previously been shown to efficiently localize hybrid proteins to nucleoli[444].

The fact that all four NoLS failed to exert any effect on the localization of the fusion proteins suggests that a general problem in the way these proteins are constructed, rather than a specific failure of the NoLS themselves, is the underlying reason for the lack of re-localization. In general, nucleolar localization of proteins occurs via strong interaction with nucleolar components like rDNA, rRNA transcripts or other nucleolar proteins[472], rather than by recognition by a dedicated import machinery, like for nuclear import. This is exemplified by the mechanism by which B23/nucleophosmin localizes to the granular component of the nucleolus[457] via interaction with G quadruplex DNA structures which are formed by rDNA[458]. This suggests that disruption or weakening of these interactions might impede nucleolar localization. If the position at the N-terminus of the SB transposase is not favorable for accessibility of these short peptides, they might lose the interaction strength required to achieve nucleolar localization for the fusion protein. One possible way to improve the localization of these proteins might be the addition of a longer peptide linker between the NoLS and the transposase to improve the interaction between the NoLS and its nucleolar target. Additionally, a range of other NoLS sequences that could be tested have been described[434].

In contrast to the SB fusions to short NoLS, the fusion of the full-length B23/nucleophosmin sequence to the N-terminus of SB100X exerted a measureable effect on the localization of the transposase; the fusion protein clearly accumulated in the nucleolus. The limited resolution of the IF images made it impossible to make any conclusions about the localization of B23 within the nucleolus, but B23 has been previously reported to be spread throughout the granular

component (GC) of the nucleolus and be enriched near the nucleolar periphery[473], which is in accordance with its suggested role of tethering other genomic components to the perinucleolar chromatin[474–476].

The observed enrichment of the fusion protein in the nucleolus is in marked contrast to the nucleolar depletion of unfused SB100X or the NoLS-SB100X fusions. This seems to indicate that the interaction of B23 with other its target DNA structures is less affected by fusion with SB100X than the interaction of the short NoLS peptides. This might be due to the fact that while the NoLS peptides are very short (14-20 amino acids), B23 is a larger protein which is similar in size to SB100X (32.8 kDa vs. 39.1 kDa). This could prevent B23 interaction from easily being masked by fusion to the SB100X N-terminus. In addition, it should be noted that B23 was fused to SB100X via a flexible linker of 14 amino acids, which might additionally help to preserve the interaction of B23 to nucleolar components.

However, a direct comparison between the localization pattern of B23-SB100X and fluorescently tagged B23 showed that while B23 seems to be found exclusively in the nucleolus, the fusion protein was concentrated in the nucleolus, but also present in the nucleoplasm. This distribution might be caused by the general DNA affinity of SB100X. It is possible that this high affinity causes the fusion protein to ‘stick’ to the DNA distributed in across the nucleus and counteract the nucleolar localization caused by B23.

Interestingly, only around 50% of all observed cells showed clear nucleolar concentration, while the rest displayed only weak or no enrichment of the protein in the nucleolus. It is not clear what causes these distinct patterns in different cells. As has been previously mentioned for the HIV-1 Tat protein, concentration of cellular factors can influence the localization, but the overall signal strength between cells with and without nucleolar localization was found to be similar. It is also possible that the localization pattern of the fusion is influenced by the cells position in the cell cycle. B23 shows distinct localization during mitosis[438], but the cells that lack nucleolar enrichment are phenotypically clearly non-mitotic. Additionally, B23 localized to nucleoli with high specificity in all cells, even those cells where B23-SB100X was found in the nucleoplasm, indicating that the fusion with SB100X itself was the cause for partial failure of nucleolar localization. B23 has been shown to shuttle between the nucleolus and the nucleoplasm[477] depending on its phosphorylation state[478]. However, even if the phosphorylation state was somehow influenced by fusion with SB100X, this would not explain why the pattern should be distinct between cells.

4.2.2 Targeting of rDNA with B23-SB100X

Overall, analysis of the integration libraries generated with B23-SB100X showed indications of increased nucleolar integration, albeit at relatively low levels (independently of which exact parameter was used for defining nucleolar integrations, insertion frequencies remained below 1% of total recovered insertions). It should be taken into account that unambiguous mapping of insertions to repetitive rDNA can be difficult, thus the frequencies of insertions into NORs reported here might underrepresent the actual fractions of insertions at these loci. However, this should apply to targeted and untargeted samples to the same extent. Additionally, a clear preference for nucleolus-associated chromatin could be observed.

The clearest indication for successful targeting into nucleolar DNA was found by comparing integration frequencies into NOR sequences as defined by Floutsakou et al.[424]. Defining NOR sequences as targets resulted an increased integration rate of B23-SB100X compared to SB100X for transposons containing either a SV40 or a HENA promoter, although the enrichment was stronger (1.8-fold vs. 1.4-fold) and more statistically significant ($p=0.005$ vs. $p=0.024$) when comparing the datasets with the SV40 transposons.

In addition to the NORs, some enrichment was also seen for NADs, i.e. non-NOR chromatin that is associated with the nucleoli. Which DNA sequences are considered nucleolus-associated depends on the criteria used and consequently NADs of different sizes can be defined. However, significant increases of insertion frequencies were seen for both NAD datasets tested and for all targeting conditions. Given the spatial enrichment of the fusion transposase at the nucleoli, it is not surprising that increased insertion can also be observed in other DNA that is close to the nucleolus. The enrichment into NADs could also indicate that a significant fraction of B23-SB100X might be located close to, but not in, the nucleoli.

Comparing integration rates into rDNA repeats that are not part of the NORs, as annotated by the RepeatMasker tool, did not result in statistically significant enrichment. The overall frequencies of insertion into these genes were generally lower than for the random dataset, suggesting they are generally disfavored as SB targets. While 5S rDNA genes, like other Pol III-transcribed genes[479,480], are often found in NADs[459], they are located near, not in, the nucleolus. Even considering that integration into NADs is enriched with B23-SB100X, the 5S rDNA genes might be poor targets within these domains. Thus, lack of enrichment into these genes is not surprising. The failure of enrichment into non-NOR SSU- and LSU-rDNA is somewhat more ambiguous. If the non-NOR rDNA genes are incorporated into nucleoli, one might assume that enrichment should occur in these sequences as well as into the NOR

sequences. However, it is possible that the rDNA outside of NORs are not actually incorporated into nucleoli, which would explain the lack of enrichment into these loci.

Another indication for increased insertion into nucleoli was found when comparing the rates of insertion into sequences recovered from a ChIP-seq experiment with UBTF. In this analysis, the insertion frequency is increased (ca. 1.7-fold) with B23-SB100X when compared to SB100X, but only in combination with the SV40 transposon. No drastic changes could be observed for any of the tested chromatin marks, although it has to be noted that the total number of insertions into these sequences is significantly higher than for UBTF, meaning that higher fluctuations were expected for the UBTF dataset.

Interestingly, and contrary to initial expectation, targeting with B23-SB100X in combination with the transposon containing a SV40 promoter seemed to produce better results than the same targeting construct with the HENA-containing transposon. This can be seen in the comparison of integrations into NOR sequences and NADs, where the SV40 dataset showed clearer enrichment, and in the comparison with the UBTF ChIP-seq dataset, where no enrichment was observed for the HENA transposon. Different integration patterns between the same constructs using transposons with different promoters can be explained either by altered target site selection or by changes in expression of the antibiotic marker after integrations. Interaction between the HENA promoter and components of the Pol I complex were theorized to promote integrations at locations where the concentration of Pol I is high, i.e. in the inner part of the nucleolus. However, the fact that insertion into rDNA genes was generally lower for transposons containing the HENA promoter and the lack of enrichment near binding sites of UBTF, which is a transcription factor of Pol I, seem to contradict this theory. Similarly, the idea that transcription of the antibiotic marker would be improved by using a Pol I promoter when insertions occur into ribosomal DNA is not confirmed by the experimental results. In fact, higher fractions of insertions into NORs and rDNA genes were found with the transposon containing SV40, indicating that even when integration occurs in nucleolar DNA, expression by Pol II remains possible. This was somewhat expected as the NOR regions flanking the rDNA repeats contain Pol II-transcribed genes[424] and low levels of Poll II-mediated transcription have been reported for the rRNA genes themselves[415].

Overall, the fact that general enrichment into NORs was stronger with the SV40 transposon and that enrichment with this transposon and B23-SB100X was enriched in the UBTF ChIP-seq dataset suggests that there might be two distinct regions where enrichment occurs. There might be one region where enrichment occurs independently of which transposon is used, and a

second region where enrichment occurs only with the SV40 promoter-containing transposon. This might be attributed to nucleolar architecture; the different DNA components are not evenly distributed throughout the nucleolus. Some of the flanking, non-rDNA sequences of the NORs are generally anchored to the surface of the nucleolus while the rDNA genes are found in the center of the nucleolus when actively transcribed by Pol I, but on the surface when transcriptionally inactive[415,424]. It is tempting to speculate that enrichment occurs mostly in the non-rDNA parts of NORs, and additionally in the inner part when the SV40 transposon is used. While this would explain the enrichment in the UBTF dataset that can only be seen in the dataset using the SV40 transposon, it contradicts the theory that the inner part of the nucleolus would be more favored by the HENA-containing transposon due to being the location of Pol-I dependent transcription. The fact that B23 was generally found in the GC and enriched near the nucleolar surface suggests that both inactive rDNA repeats as well as other non-rDNA components of the NOR should be possible targets for B23-SB100X.

Taken together, the results generated with the B23-SB100X fusion, like the results obtained with dCas9-based targeting component, show some promise and serve as a proof-of-concept for this strategy, but are far away from practical applications. The rates of integration into nucleolar DNA are currently much too low to be useful for research or clinical applications, although, as mentioned previously, they might underrepresent the true insertion frequencies at these loci. One potential approach to increase the specificity might be to address the localization pattern of B23-SB100X, which is enriched in, but not exclusive to, nucleoli. If, as has been suggested in the previous section, the general DNA affinity of the SB transposase causes the fusion to be dispersed through the nucleus, the use of a transposase with reduced DNA affinity might be the solution. The previously described SB(C42) mutant might come in handy here, although its reduced transpositional activity has not been conclusively shown to be the result of lowered DNA affinity. Other changes to the architecture of the B23-SB100X fusions, like alternative linker domains, could also be explored as means to improve the nucleolar localization of the protein. Alternatively, other NoLS could be used, either directly fused to SB100X, or added to B23-SB100X as a second NoLS. It is also tempting to speculate that the specificity could be improved by combining it with other targeting approaches, e.g. sequence based targeting. It might be possible to use a sequence-specific adapter protein, for example dCas9-N57 with an appropriate sgRNA, to target a rDNA-specific sequence, while also using the B23-SB100X fusion to preferentially localize the transposase to the nucleolus.

While limited in its efficiency, the nucleolar targeting strategy is unique in that it uses subcellular localization, rather than specific recognition of a DNA sequence or DNA-associated protein, to generate its targeting effect. To our knowledge, such an approach has not been previously applied to target a transposon or any other integrating vector. The fact that a change in the integration pattern could be effected this way indicates that it might be worthwhile to further pursue this novel approach.

4.3 HDR enhancement with Cas9 fusions

The attempt to increase the rate of HDR to NHEJ by noncovalently tethering the repair template to Cas9 failed to produce the desired result. The fact that measurable interaction between dCas9-N57 as well as dCas9-N123 and an oligo comprising the 14DR could be seen in EMSA experiments suggests that both fusion proteins should be able to recruit at least some donor DNA containing 14DR sequences to the site of editing. However, it should be noted that the interaction of the fusion proteins with the 14DR DNA is relatively weak, particularly compared to the isolated N57 subdomain, but also somewhat weaker than unfused SB100X transposase. As has been mentioned in the discussion of targeting attempts using N57 and N123 fusions to dCas9, it is possible that the interaction, while observable in an EMSA experiment, might be too weak to effectively change the localization of the desired interaction partner, whether this is the SB PEC or a HDR donor.

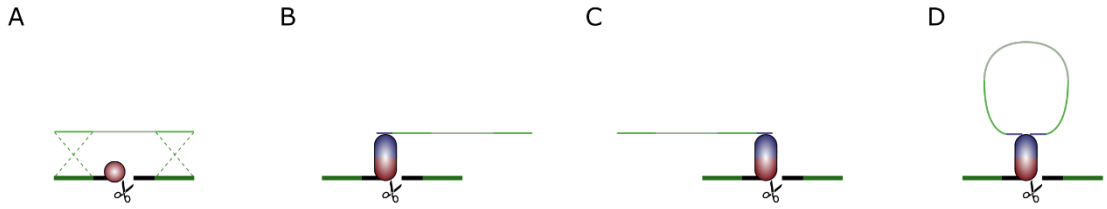


Figure 4.1 – Model of the conformation of the HDR machinery in combination with Cas9 fusions. A The homology arms (green lines) of the targeted site extend to either side of the cut made by Cas9 (red sphere) and the HDR donor is positioned accordingly. **B** Cas9-N57 or Cas9-N123 recruit the HDR template to the target site via interaction between the N57/N123 domain (blue semicircle) and the 14DR SB binding site in the donor (blue line), positioning the left end of the donor close to the cut site. **C** Same as B, with the SB binding site on the right end of the donor. **D** Presence of SB binding sites on both ends of the donor might result in an unfavorable loop formation.

An alternative explanation for the failure of Cas9-N57 and Cas9-N123 may be found in the conformation of the complex: in order to prevent the 14DR SB binding sites from being integrated into the genome, they were positioned at the ends of the HDR donor, rather than in the middle of it. However, the site of Cas9 cleavage is in the middle of the two homology arms (Figure 4.1A). Thus, positioning one of the ends of the HDR donor close to the cleavage site may not result in a conformation that is conducive to HDR (Figure 4.1B,C). In a construct with

SB binding sites on both ends, there is an additional possibility of formation of a loop structure in the donor DNA, which may also prevent efficient HDR in cases in which the donor is bound to the Cas9 fusion protein (Figure 4.1D). Such unfavorable conformations in with donors containing SB binding sites could explain the drop in overall HDR frequency that was observed with these donors in combination with Cas9-N57 and Cas9-N123, but not with unfused Cas9 (Figure 3.24).

An interesting observation that could be made is the overall increase in HDR efficiency and HDR/NHEJ ratio with the fusion protein Cas9-N57. The fact that this effect was seen with donor constructs with or without 14DR sequences indicates that it is not caused by tethering of the donor construct to the edit site by specific interaction between N57 and the 14DR sequence. It is possible that some unspecific interaction between the donor DNA and the fusion protein is responsible for this effect. However, no sites with high similarity to 14DR are present in the donor sequence and no equivalent increase in HDR efficiency was observed with Cas9-N123.

Finally, it should be noted that a recent study recently reported a significant increase (up to ~30-fold) in HDR efficiency utilizing a similar strategy to the one attempted here[481]. However, some key differences may give some indication of why the attempt to increase the HDR/NHEJ ratio with Cas9-N57/Cas9-N123 was unsuccessful. One key difference is that in the cited study, a single-stranded oligodeoxynucleotide (ssODN) was used rather than a large double-stranded HDR donor. The fact that the donor was only 200 nt long, compared to the 1 kb donor used in the experiment described here, may have played a significant role. While the site that interacts with the Cas9 fusion is still situated at the end of the donor molecule, the fact that the donor is much shorter might result in a more favorable overall conformation of the complex. In addition, the ssODN was coupled covalently to the Cas9 fusion protein, rather than bound by noncovalent interaction. This significantly higher interaction strength might have resulted in a more defined increase in local concentration of the HDR donor near the edit site.

One drawback of the method described in the cited study is that the covalent linkage between the Cas9 ribonucleoprotein (RNP) and the ssODN donor has to be established *in vitro*, thus the method is only compatible with the use of Cas9 in RNP form; it cannot be applied when Cas9 is meant to be supplied as a plasmid and expressed in the target cell. For this reason, it might still be interesting to look into improvements of the system presented here. One simple approach might be to test whether the HDR efficiency of dCas9-N57 or dCas9-N123 fusion could be improved by using a shorter HDR donor.

5 References

1. Anguela, X. M. & High, K. A. Entering the Modern Era of Gene Therapy. *Annual review of medicine* **70,** 273–288; 10.1146/annurev-med-012017-043332 (2019).

2. Neufeld, E. F., Sweeley, C. C., Rogers, S., Friedmann, T. & Roblin, R. Gene Therapy for Human Genetic Disease? *Science* **178,** 648–649; 10.1126/science.178.4061.648 (1972).

3. Sheridan, C. Gene therapy finds its niche. *Nature biotechnology* **29,** 121–128; 10.1038/nbt.1769 (2011).

4. Oldfield, E. H. *et al.* Gene therapy for the treatment of brain tumors using intra-tumoral transduction with the thymidine kinase gene and intravenous ganciclovir. *Human gene therapy* **4,** 39–69; 10.1089/hum.1993.4.1-39 (1993).

5. Sibbald, B. Death but one unintended consequence of gene-therapy trial. *CMAJ: Canadian Medical Association Journal* **164,** 1612 (2001).

6. Rafii, M. S. *et al.* Adeno-Associated Viral Vector (Serotype 2)-Nerve Growth Factor for Patients With Alzheimer Disease: A Randomized Clinical Trial. *JAMA neurology* **75,** 834–841; 10.1001/jamaneurol.2018.0233 (2018).

7. MacLaren, R. E. *et al.* Retinal gene therapy in patients with choroideremia: initial findings from a phase 1/2 clinical trial. *The Lancet* **383,** 1129–1137; 10.1016/S0140-6736(13)62117-0 (2014).

8. Miliotou, A. N. & Papadopoulou, L. C. CAR T-cell Therapy: A New Era in Cancer Immunotherapy. *Current pharmaceutical biotechnology* **19,** 5–18; 10.2174/1389201019666180418095526 (2018).

9. Chavez, J. C., Bachmeier, C. & Kharfan-Dabaja, M. A. CAR T-cell therapy for B-cell lymphomas: clinical trial results of available products. *Therapeutic advances in hematology* **10,** 2040620719841581; 10.1177/2040620719841581 (2019).

10. Cooney, A. L., McCray, P. B. & Sinn, P. L. Cystic Fibrosis Gene Therapy: Looking Back, Looking Forward. *Genes* **9**; 10.3390/genes9110538 (2018).

11. Doshi, B. S. & Arruda, V. R. Gene therapy for hemophilia: what does the future hold? *Therapeutic advances in hematology* **9,** 273–293; 10.1177/2040620718791933 (2018).

12. Peterson, C. W. & Kiem, H.-P. Cell and Gene Therapy for HIV Cure. *Current topics in microbiology and immunology* **417,** 211–248; 10.1007/82_2017_71 (2018).

13. Tebas, P. *et al.* Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV. *The New England journal of medicine* **370,** 901–910; 10.1056/NEJMoa1300662 (2014).

14. Southwell, A. L., Ko, J. & Patterson, P. H. Intrabody gene therapy ameliorates motor, cognitive, and neuropathological symptoms in multiple mouse models of Huntington's disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **29,** 13589–13602; 10.1523/JNEUROSCI.4286-09.2009 (2009).

15. Crudele, J. M. & Chamberlain, J. S. AAV-based gene therapies for the muscular dystrophies. *Human molecular genetics* **28,** R102-R107; 10.1093/hmg/ddz128 (2019).

16. Kaplitt, M. G. *et al.* Safety and tolerability of gene therapy with an adeno-associated virus (AAV) borne GAD gene for Parkinson's disease: an open label, phase I trial. *The Lancet* **369,** 2097–2105; 10.1016/S0140-6736(07)60982-9 (2007).

17. Candotti, F. *et al.* Gene therapy for adenosine deaminase–deficient severe combined immune deficiency: clinical comparison of retroviral vectors and treatment plans. *Blood* **120,** 3635–3646; 10.1182/blood-2012-02-400937 (2012).

18. Gaudet, D. *et al.* Long-Term Retrospective Analysis of Gene Therapy with Alipogene Tiparvovec and Its Effect on Lipoprotein Lipase Deficiency-Induced Pancreatitis. *Human gene therapy* **27,** 916–925; 10.1089/hum.2015.158 (2016).

19. Shahryari, A. *et al.* Development and Clinical Translation of Approved Gene Therapy Products for Genetic Disorders. *Frontiers in genetics* **10,** 868; 10.3389/fgene.2019.00868 (2019).

20. Deyle, D. R. & Russell, D. W. Adeno-associated virus vector integration. *Current opinion in molecular therapeutics* **11,** 442–447 (2009).

21. Muruve, D. A., Barnes, M. J., Stillman, I. E. & Libermann, T. A. Adenoviral gene therapy leads to rapid induction of multiple chemokines and acute neutrophil-dependent hepatic injury in vivo. *Human gene therapy* **10,** 965–976; 10.1089/10430349950018364 (1999).

22. Zhang, Y. *et al.* Acute cytokine response to systemic adenoviral vectors in mice is mediated by dendritic cells and macrophages. *Molecular Therapy* **3,** 697–707; 10.1006/mthe.2001.0329 (2001).

23. Hareendran, S. *et al.* Adeno-associated virus (AAV) vectors in gene therapy: immune challenges and strategies to circumvent them. *Reviews in medical virology* **23,** 399–413; 10.1002/rmv.1762 (2013).

24. Basner-Tschakarjan, E. & Mingozzi, F. Cell-Mediated Immunity to AAV Vectors, Evolving Concepts and Potential Solutions. *Frontiers in immunology* **5,** 350; 10.3389/fimmu.2014.00350 (2014).

25. Rogers, G. L. *et al.* Innate Immune Responses to AAV Vectors. *Frontiers in microbiology* **2,** 194; 10.3389/fmicb.2011.00194 (2011).

26. Worgall, S., Wolff, G., Falck-Pedersen, E. & Crystal, R. G. Innate immune mechanisms dominate elimination of adenoviral vectors following in vivo administration. *Human gene therapy* **8,** 37–44; 10.1089/hum.1997.8.1-37 (1997).

27. McClintock, B. Maize genetics. *Carnegie Institute of Washington Yearbook* **45,** 176–186 (1946).

28. MCCLINTOCK, B. Controlling elements and the gene. *Cold Spring Harbor symposia on quantitative biology* **21,** 197–216; 10.1101/sqb.1956.021.01.017 (1956).

29. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921; 10.1038/35057062 (2001).

30. SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274,** 765–768; 10.1126/science.274.5288.765 (1996).

31. Eickbush, T. Fruit flies and humans respond differently to retrotransposons. *Current opinion in genetics & development* **12,** 669–674; 10.1016/S0959-437X(02)00359-3 (2002).

32. Maksakova, I. A. *et al.* Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. *PLoS genetics* **2,** e2; 10.1371/journal.pgen.0020002 (2006).

33. Jiang, N., Bao, Z., Zhang, X., Eddy, S. R. & Wessler, S. R. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431,** 569–573; 10.1038/nature02953 (2004).

34. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Exon shuffling by L1 retrotransposition. *Science* **283,** 1530–1534; 10.1126/science.283.5407.1530 (1999).

35. Lev-Maor, G. *et al.* Intronic Alus influence alternative splicing. *PLoS genetics* **4,** e1000204; 10.1371/journal.pgen.1000204 (2008).

36. Schmitz, J. & Brosius, J. Exonization of transposed elements: A challenge and opportunity for evolution. *Biochimie* **93,** 1928–1934; 10.1016/j.biochi.2011.07.014 (2011).

37. Xing, J. *et al.* Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 17608–17613; 10.1073/pnas.0603224103 (2006).

38. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature genetics* **24,** 363–367; 10.1038/74184 (2000).

39. Carelli, F. N. *et al.* The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Research* **26,** 301–314; 10.1101/gr.198473.115 (2016).

40. Kubiak, M. R. & Makałowska, I. Protein-Coding Genes' Retrocopies and Their Functions. *Viruses* **9**; 10.3390/v9040080 (2017).

41. Lim, J. K. & Simmons, M. J. Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster. *BioEssays : news and reviews in molecular, cellular and developmental biology* **16,** 269–275; 10.1002/bies.950160410 (1994).

42. Gray, Y. H.M. It takes two transposons to tango:transposable-element-mediated chromosomal rearrangements. *Trends in Genetics* **16,** 461–468; 10.1016/S0168-9525(00)02104-1 (2000).

43. Preston, C. R., Sved, J. A. & Engels, W. R. Flanking Duplications and Deletions Associated with P-Induced Male Recombination in Drosophila. *Genetics* **144,** 1623–1638 (1996).

44. Zhang, J. & Peterson, T. Transposition of reversed Ac element ends generates chromosome rearrangements in maize. *Genetics* **167,** 1929–1937; 10.1534/genetics.103.026229 (2004).

45. Sijen, T. & Plasterk, R. H. A. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. *Nature* **426,** 310–314; 10.1038/nature02107 (2003).

46. Goodier, J. L. Restricting retrotransposons: a review. *Mobile DNA* **7,** 16; 10.1186/s13100-016-0070-z (2016).

47. Berrens, R. V. *et al.* An endosiRNA-Based Repression Mechanism Counteracts Transposon Activation during Global DNA Demethylation in Embryonic Stem Cells. *Cell stem cell* **21,** 694-703.e7; 10.1016/j.stem.2017.10.004 (2017).

48. Miyoshi, N. *et al.* Erasure of DNA methylation, genomic imprints, and epimutations in a primordial germ-cell model derived from mouse pluripotent stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **113,** 9545–9550; 10.1073/pnas.1610259113 (2016).

49. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome biology* **19,** 199; 10.1186/s13059-018-1577-z (2018).

50. Imbeault, M. & Trono, D. As time goes by: KRABs evolve to KAP endogenous retroelements. *Developmental cell* **31,** 257–258; 10.1016/j.devcel.2014.10.019 (2014).

51. Yang, P., Wang, Y. & Macfarlan, T. S. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends in genetics : TIG* **33,** 871–881; 10.1016/j.tig.2017.08.006 (2017).

52. Lohe, A. R. & Hartl, D. L. Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Molecular biology and evolution* **13,** 549–555; 10.1093/oxfordjournals.molbev.a025615 (1996).

53. Saha, A. *et al.* A trans-dominant form of Gag restricts Ty1 retrotransposition and mediates copy number control. *Journal of virology* **89,** 3922–3938; 10.1128/JVI.03060-14 (2015).

54. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116; 10.1126/science.1090005 (1975).

55. Carroll, S. B. Evolution at Two Levels: On Genes and Form. *PLoS Biology* **3**; 10.1371/journal.pbio.0030245 (2005).

56. Zdobnov, E. M., Campillos, M., Harrington, E. D., Torrents, D. & Bork, P. Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Research* **33,** 946–954; 10.1093/nar/gki236 (2005).

57. Britten, R. Transposable elements have contributed to thousands of human proteins. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 1798–1803; 10.1073/pnas.0510007103 (2006).

58. Huang, S. *et al.* Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. *Cell* **166,** 102–114; 10.1016/j.cell.2016.05.032 (2016).

59. Kapitonov, V. V. & Jurka, J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biology* **3,** e181; 10.1371/journal.pbio.0030181 (2005).

60. Kapitonov, V. V. & Koonin, E. V. Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biology direct* **10,** 20; 10.1186/s13062-015-0055-8 (2015).

61. Shaheen, M., Williamson, E., Nickoloff, J., Lee, S.-H. & Hromas, R. Metnase/SETMAR: a domesticated primate transposase that enhances DNA repair, replication, and decatenation. *Genetica* **138,** 559–566; 10.1007/s10709-010-9452-1 (2010).

62. Kapusta, A. *et al.* Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS genetics* **9,** e1003470; 10.1371/journal.pgen.1003470 (2013).

63. Piriyapongsa, J., Mariño-Ramírez, L. & Jordan, I. K. Origin and evolution of human microRNAs from transposable elements. *Genetics* **176,** 1323–1337; 10.1534/genetics.107.072553 (2007).

64. McCue, A. D. & Slotkin, R. K. Transposable element small RNAs as regulators of gene expression. *Trends in genetics : TIG* **28,** 616–623; 10.1016/j.tig.2012.09.001 (2012).

65. Boeke, J. D., Garfinkel, D. J., Styles, C. A. & Fink, G. R. Ty elements transpose through an RNA intermediate. *Cell* **40,** 491–500; 10.1016/0092-8674(85)90197-7 (1985).

66. Brown, P. O., Bowerman, B., Varmus, H. E. & Bishop, J.M. Correct integration of retroviral DNA in vitro. *Cell* **49,** 347–356; 10.1016/0092-8674(87)90287-X (1987).

67. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell* **72,** 595–605; 10.1016/0092-8674(93)90078-5 (1993).

68. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annual review of genetics* **41,** 331–368; 10.1146/annurev.genet.40.110405.090448 (2007).

69. Hickman, A. B. & Dyda, F. Mechanisms of DNA Transposition. *Microbiology spectrum* **3,** MDNA3-0034-2014; 10.1128/microbiolspec.MDNA3-0034-2014 (2015).

70. Kapitonov, V. V. & Jurka, J. Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 8714–8719; 10.1073/pnas.151269298 (2001).

71. Kapitonov, V. V. & Jurka, J. Self-synthesizing DNA transposons in eukaryotes. *Cytogenetic and genome research* **103,** 4540–4545; 10.1073/pnas.0600833103 (2006).

72. Pritham, E. J., Putliwala, T. & Feschotte, C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* **390,** 3–17; 10.1016/j.gene.2006.08.008 (2007).

73. Craig, N. L. *Mobile DNA II* (ASM Press, Washington, D.C, 2002).

74. Ros, F. & Kunze, R. Regulation of activator/dissociation transposition by replication and DNA methylation. *Genetics* **157,** 1723–1733 (2001).

75. Engels, W. R., Johnson-Schlitz, D. M., Eggleston, W. B. & Sved, J. High-frequency P element loss in Drosophila is homolog dependent. *Cell* **62,** 515–525; 10.1016/0092-8674(90)90016-8 (1990).

76. Kazazian, H. H. & Moran, J. V. The impact of L1 retrotransposons on the human genome. *Nature genetics* **19,** 19–24; 10.1038/ng0598-19 (1998).

77. Pace, J. K. & Feschotte, C. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Research* **17,** 422–432; 10.1101/gr.5826307 (2007).

78. Deininger, P. Alu elements: know the SINEs. *Genome biology* **12,** 236; 10.1186/gb-2011-12-12-236 (2011).

79. Mills, R. E., Bennett, E. A., Iskow, R. C. & Devine, S. E. Which transposable elements are active in the human genome? *Trends in genetics : TIG* **23,** 183–191; 10.1016/j.tig.2007.02.006 (2007).

80. Soifer, H. *et al.* Stable integration of transgenes delivered by a retrotransposon-adenovirus hybrid vector. *Human gene therapy* **12,** 1417–1428; 10.1089/104303401750298571 (2001).

81. Yang, N., Zhang, L. & Kazazian, H. H. L1 retrotransposon-mediated stable gene silencing. *Nucleic Acids Research* **33,** e57; 10.1093/nar/gni056 (2005).

82. Hou, Y., Rajagopal, J., Irwin, P. A. & Voytas, D. F. Retrotransposon vectors for gene delivery in plants. *Mobile DNA* **1,** 19; 10.1186/1759-8753-1-19 (2010).

83. Tipanee, J., VandenDriessche, T. & Chuah, M. K. Transposons: Moving Forward from Preclinical Studies to Clinical Trials. *Human gene therapy* **28,** 1087–1104; 10.1089/hum.2017.128 (2017).

84. Converse, A. D. *et al.* Counterselection and co-delivery of transposon and transposase functions for Sleeping Beauty-mediated transposition in cultured mammalian cells. *Bioscience Reports* **24,** 577–594; 10.1007/s10540-005-2793-9 (2004).

85. Li, Z., Michael, I. P., Zhou, D., Nagy, A. & Rini, J. M. Simple piggyBac transposon-based mammalian cell expression system for inducible protein production. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 5004–5009; 10.1073/pnas.1218620110 (2013).

86. Tsukahara, T. *et al.* The Tol2 transposon system mediates the genetic engineering of T-cells with CD19-specific chimeric antigen receptors for B-cell malignancies. *Gene therapy* **22,** 209–215; 10.1038/gt.2014.104 (2014).

87. Belay, E. *et al.* Novel hyperactive transposons for genetic modification of induced pluripotent and adult stem cells: a nonviral paradigm for coaxed differentiation. *Stem cells (Dayton, Ohio)* **28,** 1760–1771; 10.1002/stem.501 (2010).

88. Kebriaei, P., Izsvák, Z., Narayanavari, S. A., Singh, H. & Ivics, Z. Gene Therapy with the Sleeping Beauty Transposon System. *Trends in Genetics* **33,** 852–870; 10.1016/j.tig.2017.08.008 (2017).

89. Tipanee, J., Chai, Y. C., VandenDriessche, T. & Chuah, M. K. Preclinical and clinical advances in transposon-based gene therapy. *Bioscience Reports* **37**; 10.1042/BSR20160614 (2017).

90. Manno, C. S. *et al.* Successful transduction of liver in hemophilia by AAV-Factor IX and limitations imposed by the host immune response. *Nature medicine* **12,** 342–347; 10.1038/nm1358 (2006).

91. Rostovskaya, M. *et al.* Transposon-mediated BAC transgenesis in human ES cells. *Nucleic Acids Research* **40,** e150; 10.1093/nar/gks643 (2012).

92. Li, M. A. *et al.* Mobilization of giant piggyBac transposons in the mouse genome. *Nucleic Acids Research* **39,** e148; 10.1093/nar/gkr764 (2011).

93. Katter, K. *et al.* Transposon-mediated transgenesis, transgenic rescue, and tissue-specific gene expression in rodents and rabbits. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **27,** 930–941; 10.1096/fj.12-205526 (2013).

94. Wang, Z. *et al.* Detection of integration of plasmid DNA into host genomic DNA following intramuscular injection and electroporation. *Gene therapy* **11,** 711–721; 10.1038/sj.gt.3302213 (2004).

95. Liang, Q., Kong, J., Stalker, J. & Bradley, A. Chromosomal mobilization and reintegration of Sleeping Beauty and PiggyBac transposons. *Genesis (New York, N.Y. : 2000)* **47,** 404–408; 10.1002/dvg.20508 (2009).

96. Querques, I. *et al.* A highly soluble Sleeping Beauty transposase improves control of gene insertion. *Nature biotechnology*; 10.1038/s41587-019-0291-z (2019).

97. Cooney, A. L., Singh, B. K. & Sinn, P. L. Hybrid nonviral/viral vector systems for improved piggyBac DNA transposon in vivo delivery. *Molecular therapy : the journal of the American Society of Gene Therapy* **23,** 667–674; 10.1038/mt.2014.254 (2015).

98. Silva, S. de *et al.* Herpes simplex virus/Sleeping Beauty vector-based embryonic gene transfer using the HSB5 mutant: loss of apparent transposition hyperactivity in vivo. *Human gene therapy* **21,** 1603–1613; 10.1089/hum.2010.062 (2010).

99. Yant, S. R. *et al.* Transposition from a gutless adeno-transposon vector stabilizes transgene expression in vivo. *Nature biotechnology* **20,** 999–1005; 10.1038/nbt738 (2002).

100. Vink, C. A. *et al.* Sleeping Beauty Transposition From Nonintegrating Lentivirus. *Molecular Therapy* **17,** 1197–1204; 10.1038/mt.2009.94 (2009).

101. Bonamassa, B., Hai, L. & Liu, D. Hydrodynamic gene delivery and its applications in pharmaceutical research. *Pharmaceutical research* **28,** 694–701; 10.1007/s11095-010-0338-9 (2011).

102. He, C.-X. *et al.* Insulin expression in livers of diabetic mice mediated by hydrodynamics-based administration. *World journal of gastroenterology* **10,** 567–572; 10.3748/wjg.v10.i4.567 (2004).

103. Liu, L., Mah, C. & Fletcher, B. S. Sustained FVIII expression and phenotypic correction of hemophilia A in neonatal mice using an endothelial-targeted sleeping beauty transposon. *Molecular Therapy* **13,** 1006–1015; 10.1016/j.ymthe.2005.11.021 (2006).

104. Hausl, M. A. *et al.* Hyperactive sleeping beauty transposase enables persistent phenotypic correction in mice and a canine model for hemophilia B. *Molecular therapy : the journal of the American Society of Gene Therapy* **18,** 1896–1906; 10.1038/mt.2010.169 (2010).

105. Yazawa, H. *et al.* Hydrodynamics-based gene delivery of naked DNA encoding fetal liver kinase-1 gene effectively suppresses the growth of pre-existing tumors. *Cancer gene therapy* **13,** 993–1001; 10.1038/sj.cgt.7700970 (2006).

106. Aliño, S. F., Herrero, M. J., Noguera, I., Dasí, F. & Sánchez, M. Pig liver gene therapy by noninvasive interventionist catheterism. *Gene therapy* **14,** 334–343; 10.1038/sj.gt.3302873 (2007).

107. Hyland, K. A. *et al.* Transgene Expression in Dogs After Liver-Directed Hydrodynamic Delivery of Sleeping Beauty Transposons Using Balloon Catheters. *Human gene therapy* **28,** 541–550; 10.1089/hum.2017.003 (2017).

108. Podetz-Pedersen, K. M. *et al.* Gene expression in lung and liver after intravenous infusion of polyethylenimine complexes of Sleeping Beauty transposons. *Human gene therapy* **21,** 210–220; 10.1089/hum.2009.128 (2010).

109. Kren, B. T. *et al.* Nanocapsule-delivered Sleeping Beauty mediates therapeutic Factor VIII expression in liver sinusoidal endothelial cells of hemophilia A mice. *The Journal of clinical investigation* **119,** 2086–2099; 10.1172/JCI34332 (2009).

110. Smith, T. T. *et al.* In situ programming of leukaemia-specific T cells using synthetic DNA nanocarriers. *Nature nanotechnology* **12,** 813–820; 10.1038/nnano.2017.57 (2017).

111. Ley, D. *et al.* A PiggyBac-mediated approach for muscle gene transfer or cell therapy. *Stem cell research* **13,** 390–403; 10.1016/j.scr.2014.08.007 (2014).

112. Dupuy, A. J., Jenkins, N. A. & Copeland, N. G. Sleeping beauty: a novel cancer gene discovery tool. *Human molecular genetics* **15 Spec No 1,** R75-9; 10.1093/hmg/ddl061 (2006).

113. Ruf, S. *et al.* Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nature genetics* **43,** 379–386; 10.1038/ng.790 (2011).

114. Ivics, Z. *et al.* Transposon-mediated genome manipulation in vertebrates. *Nature methods* **6,** 415–422; 10.1038/nmeth.1332 (2009).

115. Fraser, M. J., Smith, G. E. & Summers, M. D. Acquisition of Host Cell DNA Sequences by Baculoviruses: Relationship Between Host DNA Insertions and FP Mutants of Autographa californica and Galleria mellonella Nuclear Polyhedrosis Viruses. *Journal of virology* **47,** 287–300 (1983).

116. Cary, L. C. *et al.* Transposon mutagenesis of baculoviruses: Analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology* **172,** 156–169; 10.1016/0042-6822(89)90117-7 (1989).

117. Yusa, K. piggyBac Transposon. *Microbiology spectrum* **3,** MDNA3-0028-2014; 10.1128/microbiolspec.MDNA3-0028-2014 (2015).

118. Xu, H.-F. *et al.* Identification and characterization of piggyBac-like elements in the genome of domesticated silkworm, Bombyx mori. *Molecular genetics and genomics : MGG* **276,** 31–40; 10.1007/s00438-006-0124-x (2006).

119. Hikosaka, A., Kobayashi, T., Saito, Y. & Kawahara, A. Evolution of the Xenopus piggyBac transposon family TxpB: domesticated and untamed strategies of transposon subfamilies. *Molecular biology and evolution* **24,** 2648–2656; 10.1093/molbev/msm191 (2007).

120. Bonasio, R. *et al.* Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. *Science (New York, N.Y.)* **329,** 1068–1071; 10.1126/science.1192428 (2010).

121. Ray, D. A. *et al.* Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. *Genome Research* **18,** 717–728; 10.1101/gr.071886.107 (2008).

122. Fraser, M. J., Brusca, J. S., Smith, G. E. & Summers, M. D. Transposon-mediated mutagenesis of a baculovirus. *Virology* **145,** 356–361; 10.1016/0042-6822(85)90172-2 (1985).

123. Yusa, K., Zhou, L., Li, M. A., Bradley, A. & Craig, N. L. A hyperactive piggyBac transposase for mammalian applications. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 1531–1536; 10.1073/pnas.1008322108 (2011).

124. Lacoste, A., Berenshteyn, F. & Brivanlou, A. H. An efficient and reversible transposable system for gene delivery and lineage-specific differentiation in human embryonic stem cells. *Cell stem cell* **5,** 332–342; 10.1016/j.stem.2009.07.011 (2009).

125. Cadiñanos, J. & Bradley, A. Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Research* **35,** e87; 10.1093/nar/gkm446 (2007).

126. Meir, Y.-J. J. *et al.* Genome-wide target profiling of piggyBac and Tol2 in HEK 293: pros and cons for gene discovery and gene therapy. *BMC biotechnology* **11,** 28; 10.1186/1472-6750-11-28 (2011).

127. Koga, A., Suzuki, M., Inagaki, H., Bessho, Y. & Hori, H. Transposable element in fish. *Nature* **383,** 30; 10.1038/383030a0 (1996).

128. Ni, J. *et al.* Active recombinant Tol2 transposase for gene transfer and gene discovery applications. *Mobile DNA* **7**; 10.1186/s13100-016-0062-z (2016).

129. Keng, V. W. *et al.* Efficient transposition of Tol2 in the mouse germline. *Genetics* **183,** 1565–1573; 10.1534/genetics.109.100768 (2009).

130. Balciunas, D. *et al.* Harnessing a high cargo-capacity transposon for genetic applications in vertebrates. *PLoS genetics* **2,** e169; 10.1371/journal.pgen.0020169 (2006).

131. Hamlet, M. R. J. *et al.* Tol2 transposon-mediated transgenesis in Xenopus tropicalis. *Genesis (New York, N.Y. : 2000)* **44,** 438–445; 10.1002/dvg.20234 (2006).

132. Sato, Y. *et al.* Stable integration and conditional expression of electroporated transgenes in chicken embryos. *Developmental biology* **305,** 616–624; 10.1016/j.ydbio.2007.01.043 (2007).

133. Yang, Y., Wang, W., Huang, T., Ruan, W. & Cao, G. Transgenesis of Tol2-mediated seamlessly constructed BAC mammary gland expression vectors in Mus musculus. *Journal of biotechnology* **218,** 66–72; 10.1016/j.jbiotec.2015.11.024 (2016).

134. Urasaki, A., Asakawa, K. & Kawakami, K. Efficient transposition of the Tol2 transposable element from a single-copy donor in zebrafish. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 19827–19832; 10.1073/pnas.0810380105 (2008).

135. Suster, M. L., Abe, G., Schouw, A. & Kawakami, K. Transposon-mediated BAC transgenesis in zebrafish. *Nature protocols* **6,** 1998–2021; 10.1038/nprot.2011.416 (2011).

136. Ivics, Z., Hackett, P. B., Plasterk, R. H. & Izsvák, Z. Molecular Reconstruction of Sleeping Beauty, a Tc1-like Transposon from Fish, and Its Transposition in Human Cells. *Cell* **91,** 501–510; 10.1016/S0092-8674(00)80436-5 (1997).

137. Izsvák, Z., Ivics, Z. & Plasterk, R. H. Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *Journal of Molecular Biology* **302,** 93–102; 10.1006/jmbi.2000.4047 (2000).

138. Yant, S. R. *et al.* Somatic integration and long-term transgene expression in normal and haemophilic mice using a DNA transposon system. *Nature genetics* **25,** 35–41; 10.1038/75568 (2000).

139. Ivics, Z., Izsvák, Z., Chapman, K. M. & Hamra, F. K. Sleeping Beauty transposon mutagenesis of the rat genome in spermatogonial stem cells. *Methods (San Diego, Calif.)* **53,** 356–365; 10.1016/j.ymeth.2010.12.014 (2011).

140. Garrels, W. *et al.* Germline transgenic pigs by Sleeping Beauty transposition in porcine zygotes and targeted integration in the pig genome. *PloS one* **6,** e23573; 10.1371/journal.pone.0023573 (2011).

141. Yum, S.-Y. *et al.* Efficient generation of transgenic cattle using the DNA transposon and their analysis by next-generation sequencing. *Scientific reports* **6,** 27185; 10.1038/srep27185 (2016).

142. Wang, Y. *et al.* Suicidal autointegration of sleeping beauty and piggyBac transposons in eukaryotic cells. *PLoS genetics* **10,** e1004103; 10.1371/journal.pgen.1004103 (2014).

143. Hozumi, A. *et al.* Germline transgenesis of the chordate Ciona intestinalis with hyperactive variants of sleeping beauty transposable element. *Developmental dynamics : an official publication of the American Association of Anatomists* **242,** 30–43; 10.1002/dvdy.23891 (2013).

144. Narayanavari, S. A. & Izsvák, Z. Sleeping Beauty transposon vectors for therapeutic applications: advances and challenges. *Cell Gene Therapy Insights* **3,** 131–158; 10.18609/cgti.2017.014 (2017).

145. Plasterk, R. H.A., Izsvák, Z. & Ivics, Z. Resident aliens: the Tc1/ mariner superfamily of transposable elements. *Trends in Genetics* **15,** 326–332; 10.1016/S0168-9525(99)01777-1 (1999).

146. Cui, Z., Geurts, A. M., Liu, G., Kaufman, C. D. & Hackett, P. B. Structure–Function Analysis of the Inverted Terminal Repeats of the Sleeping Beauty Transposon. *Journal of Molecular Biology* **318,** 1221–1235; 10.1016/S0022-2836(02)00237-1 (2002).

147. Izsvak, Z. *et al.* Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in Sleeping Beauty transposition. *The Journal of biological chemistry* **277,** 34581–34588; 10.1074/jbc.M204001200 (2002).

148. Walisko, O. *et al.* Transcriptional activities of the Sleeping Beauty transposon and shielding its genetic cargo with insulators. *Molecular therapy : the journal of the American Society of Gene Therapy* **16,** 359–369; 10.1038/sj.mt.6300366 (2008).

149. Zayed, H., Izsvák, Z., Walisko, O. & Ivics, Z. Development of hyperactive sleeping beauty transposon vectors by mutational analysis. *Molecular Therapy* **9,** 292–304; 10.1016/j.ymthe.2003.11.024 (2004).

150. Czerny, T., Schaffner, G. & Busslinger, M. DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site. *Genes & development* **7,** 2048–2061; 10.1101/gad.7.10.2048 (1993).

151. Breitling, R. & Gerber, J. K. Origin of the paired domain. *Development genes and evolution* **210,** 644–650; 10.1007/s004270000106 (2000).

152. Yant, S. R., Park, J., Huang, Y., Mikkelsen, J. G. & Kay, M. A. Mutational Analysis of the N-Terminal DNA-Binding Domain of Sleeping Beauty Transposase: Critical Residues for DNA Binding and Hyperactivity in Mammalian Cells. *Molecular and cellular biology* **24,** 9239–9247; 10.1128/MCB.24.20.9239-9247.2004 (2004).

153. Wang, Y. *et al.* Regulated complex assembly safeguards the fidelity of Sleeping Beauty transposition. *Nucleic Acids Research* **45,** 311–326; 10.1093/nar/gkw1164 (2016).

154. Aravind, L. & Landsman, D. AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Research* **26,** 4413–4421; 10.1093/nar/26.19.4413 (1998).

155. Spanopoulou, E. *et al.* The Homeodomain Region of Rag-1 Reveals the Parallel Mechanisms of Bacterial and V(D)J Recombination. *Cell* **87,** 263–276; 10.1016/S0092-8674(00)81344-6 (1996).

156. Craig, N. L. Unity in transposition reactions. *Science* **270,** 253–254; 10.1126/science.270.5234.253 (1995).

157. Montaño, S. P. & Rice, P. A. Moving DNA around: DNA transposition and retroviral integration. *Current opinion in structural biology* **21,** 370–378; 10.1016/j.sbi.2011.03.004 (2011).

158. Voigt, F. *et al.* Sleeping Beauty transposase structure allows rational design of hyperactive variants for genetic engineering. *ncomms* **7,** 11126; 10.1038/ncomms11126.

159. Mizuuchi, K. Polynucleotidyl transfer reactions in transpositional DNA recombination. *The Journal of biological chemistry* **267,** 21273–21276 (1992).

160. Izsvák, Z. *et al.* Healing the Wounds Inflicted by Sleeping Beauty Transposition by Double-Strand Break Repair in Mammalian Somatic Cells. *Molecular cell* **13,** 279–290; 10.1016/S1097-2765(03)00524-0 (2004).

161. Luo, G., Ivics, Z., Izsvák, Z. & Bradley, A. Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 10769–10773; 10.1073/pnas.95.18.10769 (1998).

162. Kesselring, L. *et al.* A single amino acid switch converts the Sleeping Beauty transposase into an efficient unidirectional excisionase with utility in stem cell reprogramming. *Nucleic Acids Research*; 10.1093/nar/gkz1119 (2019).

163. Vigdal, T. J., Kaufman, C. D., Izsvák, Z., Voytas, D. F. & Ivics, Z. Common Physical Properties of DNA Affecting Target Site Selection of Sleeping Beauty and other Tc1/mariner Transposable Elements. *Journal of Molecular Biology* **323,** 441–452; 10.1016/S0022-2836(02)00991-9 (2002).

164. Ammar, I. *et al.* Retargeting transposon insertions by the adeno-associated virus Rep protein. *Nucleic Acids Research* **40,** 6693–6712; 10.1093/nar/gks317 (2012).

165. Voigt, K. *et al.* Retargeting sleeping beauty transposon insertions by engineered zinc finger DNA-binding domains. *Molecular therapy : the journal of the American Society of Gene Therapy* **20,** 1852–1862; 10.1038/mt.2012.126 (2012).

166. Yant, S. R. *et al.* High-resolution genome-wide mapping of transposon integration in mammals. *Molecular and cellular biology* **25,** 2085–2094; 10.1128/MCB.25.6.2085-2094.2005 (2005).

167. Liu, G. *et al.* Target-site preferences of Sleeping Beauty transposons. *Journal of Molecular Biology* **346,** 161–173; 10.1016/j.jmb.2004.09.086 (2005).

168. Geurts, A. M. *et al.* Structure-based prediction of insertion-site preferences of transposons into chromosomes. *Nucleic Acids Research* **34,** 2803–2811; 10.1093/nar/gkl301 (2006).

169. Huang, X. *et al.* Gene Transfer Efficiency and Genome-Wide Integration Profiling of Sleeping Beauty, Tol2, and PiggyBac Transposons in Human Primary T Cells. *Molecular Therapy* **18,** 1803–1813; 10.1038/mt.2010.141 (2010).

170. Gogol-Doring, A. *et al.* Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac

Transposon in Primary Human CD4(+) T Cells. *Molecular Therapy* **24,** 592–606; 10.1038/mt.2016.11 (2016).

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

172. Montaño, S. P., Pigli, Y. Z. & Rice, P. A. The μ transpososome structure sheds light on DDE recombinase evolution. *Nature* **491,** 413–417; 10.1038/nature11602 (2012).

173. Fischer, S. E., Wienholds, E. & Plasterk, R. H. Regulated transposition of a fish transposon in the mouse germ line. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 6759–6764; 10.1073/pnas.121569298 (2001).

174. Horie, K. *et al.* Characterization of Sleeping Beauty transposition and its application to genetic screening in mice. *Molecular and cellular biology* **23,** 9189–9207; 10.1128/mcb.23.24.9189-9207.2003 (2003).

175. Karsi, A., Moav, B., Hackett, P. & Liu, Z. Effects of insert size on transposition efficiency of the sleeping beauty transposon in mouse cells. *Marine biotechnology (New York, N.Y.)* **3,** 241–245; 10.1007/s101260000072 (2001).

176. Geurts, A. M. *et al.* Gene transfer into genomes of human cells by the sleeping beauty transposon system. *Molecular Therapy* **8,** 108–117; 10.1016/S1525-0016(03)00099-6 (2003).

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

178. Wu, S. C.-Y. *et al.* piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 15008–15013; 10.1073/pnas.0606979103 (2006).

179. Skipper, K. A., Andersen, P. R., Sharma, N. & Mikkelsen, J. G. DNA transposon-based gene vehicles - scenes from an evolutionary drive. *Journal of biomedical science* **20,** 92; 10.1186/1423-0127-20-92 (2013).

180. Nakanishi, H., Higuchi, Y., Kawakami, S., Yamashita, F. & Hashida, M. piggyBac transposon-mediated long-term gene expression in mice. *Molecular therapy : the journal of the American Society of Gene Therapy* **18,** 707–714; 10.1038/mt.2009.302 (2010).

181. Guo, Y., Zhang, Y. & Hu, K. Sleeping Beauty transposon integrates into non-TA dinucleotides. *Mobile DNA* **9,** 8; 10.1186/s13100-018-0113-8 (2018).

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

183. Chiang, S. J., Jordan, E. & Clowes, R. C. Intermolecular and intramolecular transposition and transposition immunity in Tn3 and Tn2660. *Molecular & general genetics : MGG* **187,** 187–194; 10.1007/bf00331116 (1982).

184. Garfinkel, D. J. *et al.* Retrotransposon suicide: formation of Ty1 circles and autointegration via a central DNA flap. *Journal of virology* **80,** 11920–11934; 10.1128/JVI.01483-06 (2006).

185. Stellwagen, A. E. & Craig, N. L. Avoiding self: two Tn7-encoded proteins mediate target immunity in Tn7 transposition. *The EMBO Journal* **16,** 6823–6834; 10.1093/emboj/16.22.6823 (1997).

186. Lee, M. S. & Craigie, R. A previously unidentified host protein protects retroviral DNA from autointegration. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 1528–1533; 10.1073/pnas.95.4.1528 (1998).

187. Walisko, O. *et al.* Sleeping Beauty transposase modulates cell-cycle progression through interaction with Miz-1. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 4062–4067; 10.1073/pnas.0507683103 (2006).

188. Zayed, H., Izsvák, Z., Khare, D., Heinemann, U. & Ivics, Z. The DNA-bending protein HMGB1 is a cellular cofactor of Sleeping Beauty transposition. *Nucleic Acids Research* **31,** 2313–2322 (2003).

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

190. Aiuti, A. *et al.* Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science* **341,** 1233151; 10.1126/science.1233151 (2013).

191. Mátés, L. *et al.* Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nature genetics* **41,** 753–761; 10.1038/ng.343 (2009).

192. Baus, J., Liu, L., Heggestad, A. D., Sanz, S. & Fletcher, B. S. Hyperactive transposase mutants of the Sleeping Beauty transposon. *Molecular Therapy* **12,** 1148–1156; 10.1016/j.ymthe.2005.06.484 (2005).

193. Li, X. *et al.* piggyBac transposase tools for genome engineering. *Proceedings of the National Academy of Sciences of the United States of America* **110,** E2279-87; 10.1073/pnas.1305987110 (2013).

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

195. Copeland, N. G. & Jenkins, N. A. Harnessing transposons for cancer gene discovery. *Nature reviews. Cancer* **10,** 696–706; 10.1038/nrc2916 (2010).

196. Moriarity, B. S. & Largaespada, D. A. Sleeping Beauty transposon insertional mutagenesis based mouse models for cancer gene discovery. *Current opinion in genetics & development* **30,** 66–72; 10.1016/j.gde.2015.04.007 (2015).

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

198. Ammar, I., Izsvák, Z. & Ivics, Z. The Sleeping Beauty transposon toolbox. *Methods in molecular biology (Clifton, N.J.)* **859,** 229–240; 10.1007/978-1-61779-603-6_13 (2012).

199. Narayanavari, S. A., Chilkunda, S. S., Ivics, Z. & Izsvák, Z. Sleeping Beauty transposition: from biology to applications. *Critical reviews in biochemistry and molecular biology* **52,** 18–44; 10.1080/10409238.2016.1237935 (2017).

200. Hyland, K. A. *et al.* Sleeping Beauty-mediated correction of Fanconi anemia type C. *The journal of gene medicine* **13,** 462–469; 10.1002/jgm.1589 (2011).

201. Chen, Z. J., Kren, B. T., Wong, P. Y.-P., Low, W. C. & Steer, C. J. Sleeping Beauty-mediated down-regulation of huntingtin expression by RNA interference. *Biochemical and biophysical research communications* **329,** 646–652; 10.1016/j.bbrc.2005.02.024 (2005).

202. Sjeklocha, L. M., Wong, P. Y.-P., Belcher, J. D., Vercellotti, G. M. & Steer, C. J. β-Globin sleeping beauty transposon reduces red blood cell sickling in a patient-derived CD34(+)-based in vitro model. *PloS one* **8,** e80403; 10.1371/journal.pone.0080403 (2013).

203. Latella, M. C. *et al.* Correction of Recessive Dystrophic Epidermolysis Bullosa by Transposon-Mediated Integration of COL7A1 in Transplantable Patient-Derived Primary Keratinocytes. *The Journal of investigative dermatology* **137,** 836–844; 10.1016/j.jid.2016.11.038 (2017).

204. Ohlfest, J. R. *et al.* 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,** 2691–2698; 10.1182/blood-2004-09-3496 (2005).

205. Montini, E. *et al.* In vivo correction of murine tyrosinemia type I by DNA-mediated transposition. *Molecular Therapy* **6,** 759–769; 10.1006/mthe.2002.0812 (2002).

206. Xiao, Y. *et al.* Endothelial Indoleamine 2,3-Dioxygenase Protects against Development of Pulmonary Hypertension. *American Journal of Respiratory and Critical Care Medicine* **188,** 482–491; 10.1164/rccm.201304-0700OC (2013).

207. Wu, A. *et al.* Transposon-based interferon gamma gene transfer overcomes limitations of episomal plasmid for immunogene therapy of glioblastoma. *Cancer gene therapy* **14,** 550–560; 10.1038/sj.cgt.7701045 (2007).

208. Belur, L. R. *et al.* Inhibition of angiogenesis and suppression of colorectal cancer metastatic to the liver using the Sleeping Beauty Transposon System. *Molecular Cancer* **10,** 14; 10.1186/1476-4598-10-14 (2011).

209. ClinicalTrials.gov (Internet). MT2018-18: Sleeping Beauty Transposon-Engineered Plasmablasts for Hurler Syndrome Post Allo HSCT. NCT04284254. Available at https://clinicaltrials.gov/ct2/show/NCT04284254.

210. Kebriaei, P. *et al.* Phase I trials using Sleeping Beauty to generate CD19-specific CAR T cells. *The Journal of clinical investigation* **126,** 3363–3376; 10.1172/JCI86721 (2016).

211. Eyjolfsdottir, H. *et al.* Targeted delivery of nerve growth factor to the cholinergic basal forebrain of Alzheimer's disease patients: application of a second-generation encapsulated cell biodelivery device. *Alzheimer's research & therapy* **8,** 30; 10.1186/s13195-016-0195-9 (2016).

212. Šimčíková, M., Prather, K. L. J., Prazeres, D. M. F. & Monteiro, G. A. Towards effective non-viral gene delivery vector. *Biotechnology & genetic engineering reviews* **31,** 82–107; 10.1080/02648725.2016.1178011 (2015).

213. Tolmachov, O. E. Building mosaics of therapeutic plasmid gene vectors. *Current gene therapy* **11,** 466–478; 10.2174/156652311798192798 (2011).

214. Monjezi, R. *et al.* Enhanced CAR T-cell engineering using non-viral Sleeping Beauty transposition from minicircle vectors. *Leukemia* **31,** 186–194; 10.1038/leu.2016.180 (2017).

215. Holstein, M. *et al.* Efficient Non-viral Gene Delivery into Human Hematopoietic Stem Cells by Minicircle Sleeping Beauty Transposon Vectors. *Molecular therapy : the journal of the American Society of Gene Therapy* **26,** 1137–1153; 10.1016/j.ymthe.2018.01.012 (2018).

216. Garrels, W. *et al.* Cytoplasmic injection of murine zygotes with Sleeping Beauty transposon plasmids and minicircles results in the efficient generation of germline transgenic mice. *Biotechnology journal* **11,** 178–184; 10.1002/biot.201500218 (2016).

217. Sharma, N. *et al.* Efficient sleeping beauty DNA transposition from DNA minicircles. *Molecular therapy. Nucleic acids* **2,** e74; 10.1038/mtna.2013.1 (2013).

218. Wang, D.-D., Yang, M., Zhu, Y. & Mao, C. Reiterated Targeting Peptides on the Nanoparticle Surface Significantly Promote Targeted Vascular Endothelial Growth Factor Gene Delivery to Stem Cells. *Biomacromolecules* **16,** 3897–3903; 10.1021/acs.biomac.5b01226 (2015).

219. Ma, K. *et al.* Targeted delivery of in situ PCR-amplified Sleeping Beauty transposon genes to cancer cells with lipid-based nanoparticle-like protocells. *Biomaterials* **121,** 55–63; 10.1016/j.biomaterials.2016.12.033 (2017).

220. Wang, X., Mani, P., Sarkar, D. P., Roy-Chowdhury, N. & Roy-Chowdhury, J. Ex vivo gene transfer into hepatocytes. *Methods in molecular biology (Clifton, N.J.)* **481,** 117–140; 10.1007/978-1-59745-201-4_11 (2009).

221. Boehme, P., Zhang, W., Solanki, M., Ehrke-Schulz, E. & Ehrhardt, A. A High-Capacity Adenoviral Hybrid Vector System Utilizing the Hyperactive Sleeping Beauty Transposase SB100X for Enhanced Integration. *Molecular therapy. Nucleic acids* **5,** e337; 10.1038/mtna.2016.44 (2016).

222. Zhang, W. *et al.* Hybrid adeno-associated viral vectors utilizing transposase-mediated somatic integration for stable transgene expression in human cells. *PloS one* **8,** e76771; 10.1371/journal.pone.0076771 (2013).

223. Mikkelsen, J. G. *et al.* Helper-Independent sleeping beauty Transposon–Transposase vectors for efficient nonviral gene delivery and persistent gene expression in vivo. *Molecular Therapy* **8,** 654–665; 10.1016/S1525-0016(03)00216-8 (2003).

224. Belur, L. R. *et al.* Gene insertion and long-term expression in lung mediated by the sleeping beauty transposon system. *Molecular Therapy* **8,** 501–507; 10.1016/S1525-0016(03)00211-9 (2003).

225. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nature reviews. Genetics* **11,** 636–646; 10.1038/nrg2842 (2010).

226. Ousterout, D. G. & Gersbach, C. A. The Development of TALE Nucleases for Biotechnology. *Methods in molecular biology (Clifton, N.J.)* **1338,** 27–42; 10.1007/978-1-4939-2932-0_3 (2016).

227. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science (New York, N.Y.)* **346,** 1258096; 10.1126/science.1258096 (2014).

228. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)* **337,** 816–821; 10.1126/science.1225829 (2012).

229. Beerli, R. R. & Barbas, C. F. Engineering polydactyl zinc-finger transcription factors. *Nature biotechnology* **20,** 135–141; 10.1038/nbt0202-135 (2002).

230. Gaj, T., Gersbach, C. A. & Barbas, C. F. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in biotechnology* **31,** 397–405; 10.1016/j.tibtech.2013.04.004 (2013).

231. Reyon, D. *et al.* FLASH Assembly of TALENs Enables High-Throughput Genome Editing. *Nature biotechnology* **30,** 460–465; 10.1038/nbt.2170 (2012).

232. Kim, H. & Kim, J.-S. A guide to genome engineering with programmable nucleases. *Nature reviews. Genetics* **15,** 321–334; 10.1038/nrg3686 (2014).

233. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Research* **39,** e82; 10.1093/nar/gkr218 (2011).

234. Schmid-Burgk, J. L., Schmidt, T., Kaiser, V., Höning, K. & Hornung, V. A ligation-independent cloning technique for high-throughput assembly of transcription activator–like effector genes. *Nature biotechnology* **31,** 76–81; 10.1038/nbt.2460 (2013).

235. Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nature medicine* **24,** 927–930; 10.1038/s41591-018-0049-z (2018).

236. Ihry, R. J. *et al.* p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. *Nature medicine* **24,** 939–946; 10.1038/s41591-018-0050-6 (2018).

237. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nature biotechnology* **36,** 765–771; 10.1038/nbt.4192 (2018).

238. Miller, J. C. *et al.* An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature biotechnology* **25,** 778–785; 10.1038/nbt1319 (2007).

239. Szczepek, M. *et al.* Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nature biotechnology* **25,** 786–793; 10.1038/nbt1317 (2007).

240. Lieber, M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual review of biochemistry* **79,** 181–211; 10.1146/annurev.biochem.052308.093131 (2010).

241. Li, K., Wang, G., Andersen, T., Zhou, P. & Pu, W. T. Optimization of genome engineering approaches with the CRISPR/Cas9 system. *PloS one* **9,** e105779; 10.1371/journal.pone.0105779 (2014).

242. Fung, H. & Weinstock, D. M. Repair at single targeted DNA double-strand breaks in pluripotent and differentiated human cells. *PloS one* **6,** e20514; 10.1371/journal.pone.0020514 (2011).

243. Takata, M. *et al.* Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *The EMBO Journal* **17,** 5497–5508; 10.1093/emboj/17.18.5497 (1998).

244. Orthwein, A. *et al.* A mechanism for the suppression of homologous recombination in G1 cells. *Nature* **528,** 422–426; 10.1038/nature16142 (2015).

245. van Chu, T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nature biotechnology* **33,** 543–548; 10.1038/nbt.3198 (2015).

246. Maruyama, T. *et al.* Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nature biotechnology* **33,** 538–542; 10.1038/nbt.3190 (2015).

247. Pinder, J., Salsman, J. & Dellaire, G. Nuclear domain 'knock-in' screen for the evaluation and identification of small molecule enhancers of CRISPR-based genome editing. *Nucleic Acids Research* **43,** 9379–9392; 10.1093/nar/gkv993 (2015).

248. Song, J. *et al.* RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. *Nature communications* **7,** 10548; 10.1038/ncomms10548 (2016).

249. Rees, H. A., Yeh, W.-H. & Liu, D. R. Development of hRad51-Cas9 nickase fusions that mediate HDR without double-stranded breaks. *Nature communications* **10,** 2212; 10.1038/s41467-019-09983-4 (2019).

250. Paquet, D. *et al.* Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533,** 125–129; 10.1038/nature17664 (2016).

251. Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* **3,** e04766; 10.7554/eLife.04766 (2014).

252. Gutschner, T., Haemmerle, M., Genovese, G., Draetta, G. F. & Chin, L. Post-translational Regulation of Cas9 during G1 Enhances Homology-Directed Repair. *Cell reports* **14,** 1555–1566; 10.1016/j.celrep.2016.01.019 (2016).

253. Mojica, F. J. M. & Rodriguez-Valera, F. The discovery of CRISPR in archaea and bacteria. *The FEBS journal* **283,** 3162–3169; 10.1111/febs.13766 (2016).

254. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)* **315,** 1709–1712; 10.1126/science.1138140 (2007).

255. Knott, G. J. & Doudna, J. A. CRISPR-Cas guides the future of genetic engineering. *Science (New York, N.Y.)* **361,** 866–869; 10.1126/science.aat5011 (2018).

256. Brouns, S. J. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science (New York, N.Y.)* **321,** 960–964; 10.1126/science.1159689 (2008).

257. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–607; 10.1038/nature09886 (2011).

258. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology (Reading, England)* **155,** 733–740; 10.1099/mic.0.023960-0 (2009).

259. Jiang, F. & Doudna, J. A. CRISPR-Cas9 Structures and Mechanisms. *Annual review of biophysics* **46,** 505–529; 10.1146/annurev-biophys-062215-010822 (2017).

260. Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science (New York, N.Y.)* **348,** 1477–1481; 10.1126/science.aab1452 (2015).

261. Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature biotechnology* **32,** 670–676; 10.1038/nbt.2889 (2014).

262. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67; 10.1038/nature13011 (2014).

263. Davis, L. & Maizels, N. DNA Nicks Promote Efficient and Safe Targeted Gene Correction. *PloS one* **6**; 10.1371/journal.pone.0023981 (2011).

264. Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science (New York, N.Y.)*; 10.1126/science.aba8853 (2020).

265. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature biotechnology* **31,** 839–843; 10.1038/nbt.2673 (2013).

266. O'Geen, H., Henry, I. M., Bhakta, M. S., Meckler, J. F. & Segal, D. J. A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. *Nucleic Acids Research* **43,** 3389–3404; 10.1093/nar/gkv137 (2015).

267. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nature biotechnology* **32,** 677–683; 10.1038/nbt.2916 (2014).

268. Kleinstiver, B. P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529,** 490–495; 10.1038/nature16526 (2016).

269. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science (New York, N.Y.)* **351,** 84–88; 10.1126/science.aad5227 (2016).

270. Chen, J. S. *et al.* Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550,** 407–410; 10.1038/nature24268 (2017).

271. Casini, A. *et al.* A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nature biotechnology* **36,** 265–271; 10.1038/nbt.4066 (2018).

272. Vakulskas, C. A. *et al.* A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nature medicine* **24,** 1216–1224; 10.1038/s41591-018-0137-0 (2018).

273. Yin, H. *et al.* Partial DNA-guided Cas9 enables genome editing with reduced off-target activity. *Nature chemical biology* **14,** 311–316; 10.1038/nchembio.2559 (2018).

274. Richter, F. *et al.* Switchable Cas9. *Current opinion in biotechnology* **48,** 119–126; 10.1016/j.copbio.2017.03.025 (2017).

275. Pickar-Oliver, A. & Gersbach, C. A. The next generation of CRISPR-Cas technologies and applications. *Nature reviews. Molecular cell biology* **20,** 490–507; 10.1038/s41580-019-0131-5 (2019).

276. Long, C. *et al.* Postnatal genome editing partially restores dystrophin expression in a mouse model of muscular dystrophy. *Science (New York, N.Y.)* **351,** 400–403; 10.1126/science.aad5725 (2016).

277. Amoasii, L. *et al.* Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy. *Science (New York, N.Y.)* **362,** 86–91; 10.1126/science.aau1549 (2018).

278. Young, C. S. *et al.* A Single CRISPR-Cas9 Deletion Strategy that Targets the Majority of DMD Patients Restores Dystrophin Function in hiPSC-Derived Muscle Cells. *Cell stem cell* **18,** 533–540; 10.1016/j.stem.2016.01.021 (2016).

279. Xie, F. *et al.* Seamless gene correction of β-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac. *Genome Research* **24,** 1526–1533; 10.1101/gr.173427.114 (2014).

280. Gaj, T. *et al.* In vivo genome editing improves motor function and extends survival in a mouse model of ALS. *Science advances* **3,** eaar3952; 10.1126/sciadv.aar3952 (2017).

281. Staahl, B. T. *et al.* Efficient genome editing in the mouse brain by local delivery of engineered Cas9 ribonucleoprotein complexes. *Nature biotechnology* **35,** 431–434; 10.1038/nbt.3806 (2017).

282. Yin, H. *et al.* Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype. *Nature biotechnology* **32,** 551–553; 10.1038/nbt.2884 (2014).

283. Cyranoski, D. CRISPR gene-editing tested in a person for the first time. *Nature* **539,** 479; 10.1038/nature.2016.20988 (2016).

284. You, L. *et al.* Advancements and Obstacles of CRISPR-Cas9 Technology in Translational Research. *Molecular therapy. Methods & clinical development* **13,** 359–370; 10.1016/j.omtm.2019.02.008 (2019).

285. Stadtmauer, E. A. *et al.* CRISPR-engineered T cells in patients with refractory cancer. *Science (New York, N.Y.)* **367**; 10.1126/science.aba7365 (2020).

286. Komor, A. C., Badran, A. H. & Liu, D. R. CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. *Cell* **168,** 20–36; 10.1016/j.cell.2016.10.044 (2017).

287. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152,** 1173–1183; 10.1016/j.cell.2013.02.022 (2013).

288. Beerli, R. R., Segal, D. J., Dreier, B. & Barbas, C. F. 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. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 14628–14633; 10.1073/pnas.95.25.14628 (1998).

289. Maeder, M. L. *et al.* Robust, synergistic regulation of human gene expression using TALE activators. *Nature methods* **10,** 243–245; 10.1038/nmeth.2366 (2013).

290. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154,** 442–451; 10.1016/j.cell.2013.06.044 (2013).

291. Konermann, S. *et al.* Optical control of mammalian endogenous transcription and epigenetic states. *Nature* **500,** 472–476; 10.1038/nature12466 (2013).

292. Maeder, M. L. *et al.* Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nature biotechnology* **31,** 1137–1142; 10.1038/nbt.2726 (2013).

293. Mendenhall, E. M. *et al.* Locus-specific editing of histone modifications at endogenous enhancers. *Nature biotechnology* **31,** 1133–1136; 10.1038/nbt.2701 (2013).

294. Halperin, S. O. *et al.* CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. *Nature* **560,** 248–252; 10.1038/s41586-018-0384-8 (2018).

295. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533,** 420–424; 10.1038/nature17946 (2016).

296. Nishida, K. *et al.* Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science (New York, N.Y.)* **353**; 10.1126/science.aaf8729 (2016).

297. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551,** 464–471; 10.1038/nature24644 (2017).

298. Komor, A. C. *et al.* Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Science advances* **3,** eaao4774; 10.1126/sciadv.aao4774 (2017).

299. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*; 10.1038/s41586-019-1711-4 (2019).

300. Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155,** 1479–1491; 10.1016/j.cell.2013.12.001 (2013).

301. Zhou, Y. *et al.* Painting a specific chromosome with CRISPR/Cas9 for live-cell imaging. *Cell research* **27,** 298–301; 10.1038/cr.2017.9 (2017).

302. Kazazian, H. H. *et al.* Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332,** 164–166; 10.1038/332164a0 (1988).

303. Miki, Y. *et al.* Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer research* **52,** 643–645 (1992).

304. Zhao, X. *et al.* Dr.VIS: a database of human disease-related viral integration sites. *Nucleic Acids Research* **40,** D1041-6; 10.1093/nar/gkr1142 (2011).

305. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mobile DNA* **7,** 9; 10.1186/s13100-016-0065-9 (2016).

306. Cavazzana-Calvo, M. *et al.* Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science* **288,** 669–672; 10.1126/science.288.5466.669 (2000).

307. Hacein-Bey-Abina, S. *et al.* Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *The Journal of clinical investigation* **118,** 3132–3142; 10.1172/JCI35700 (2008).

308. Hacein-Bey-Abina, S. *et al.* LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science (New York, N.Y.)* **302,** 415–419; 10.1126/science.1088547 (2003).

309. Howe, S. J. *et al.* Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *The Journal of clinical investigation* **118,** 3143–3150; 10.1172/JCI35798 (2008).

310. Braun, C. J. *et al.* Gene therapy for Wiskott-Aldrich syndrome--long-term efficacy and genotoxicity. *Science translational medicine* **6,** 227ra33; 10.1126/scitranslmed.3007280 (2014).

311. Ott, M. G. *et al.* Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nature medicine* **12,** 401–409; 10.1038/nm1393 (2006).

312. Hacein-Bey-Abina, S. *et al.* A modified gamma-retrovirus vector for X-linked severe combined immunodeficiency. *The New England journal of medicine* **371,** 1407–1417; 10.1056/NEJMoa1404588 (2014).

313. Thornhill, S. I. *et al.* Self-inactivating gammaretroviral vectors for gene therapy of X-linked severe combined immunodeficiency. *Molecular therapy : the journal of the American Society of Gene Therapy* **16,** 590–598; 10.1038/sj.mt.6300393 (2008).

314. Montini, E. *et al.* Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. *Nature biotechnology* **24,** 687–696; 10.1038/nbt1216 (2006).

315. Cavazzana-Calvo, M. *et al.* Transfusion independence and HMGA2 activation after gene therapy of human β-thalassaemia. *Nature* **467,** 318–322; 10.1038/nature09328 (2010).

316. Stocking, C. *et al.* Distinct classes of factor-independent mutants can be isolated after retroviral mutagenesis of a human myeloid stem cell line. *Growth factors (Chur, Switzerland)* **8,** 197–209 (1993).

317. Cavazza, A., Moiani, A. & Mavilio, F. Mechanisms of retroviral integration and mutagenesis. *Human gene therapy* **24,** 119–131; 10.1089/hum.2012.203 (2013).

318. Kovač, A. & Ivics, Z. Specifically integrating vectors for targeted gene delivery: progress and prospects. *Cell Gene Therapy Insights* **3,** 103–123; 10.18609/cgti.2017.013 (2017).

319. Wu, X., Li, Y., Crise, B. & Burgess, S. M. Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300,** 1749–1751; 10.1126/science.1083413 (2003).

320. Mitchell, R. S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biology* **2,** E234; 10.1371/journal.pbio.0020234 (2004).

321. Cattoglio, C. *et al.* High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. *Blood* **116,** 5507–5517; 10.1182/blood-2010-05-283523 (2010).

322. Gupta, S. S. *et al.* Bromo- and extraterminal domain chromatin regulators serve as cofactors for murine leukemia virus integration. *Journal of virology* **87,** 12721–12736; 10.1128/JVI.01942-13 (2013).

323. Sharma, A. *et al.* BET proteins promote efficient murine leukemia virus integration at transcription start sites. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 12036–12041; 10.1073/pnas.1307157110 (2013).

324. Rijck, J. de *et al.* The BET family of proteins targets moloney murine leukemia virus integration near transcription start sites. *Cell reports* **5,** 886–894; 10.1016/j.celrep.2013.09.040 (2013).

325. Debyser, Z., Christ, F., Rijck, J. de & Gijsbers, R. Host factors for retroviral integration site selection. *Trends in biochemical sciences* **40,** 108–116; 10.1016/j.tibs.2014.12.001 (2015).

326. El Ashkar, S. *et al.* BET-independent MLV-based Vectors Target Away From Promoters and Regulatory Elements. *Molecular therapy. Nucleic acids* **3,** e179; 10.1038/mtna.2014.33 (2014).

327. Schroder, A. R. W. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110,** 521–529 (2002).

328. Cherepanov, P. *et al.* Solution structure of the HIV-1 integrase-binding domain in LEDGF/p75. *Nature structural & molecular biology* **12,** 526–532; 10.1038/nsmb937 (2005).

329. Ciuffi, A. *et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nature medicine* **11,** 1287–1289; 10.1038/nm1329 (2005).

330. Busschots, K. *et al.* The interaction of LEDGF/p75 with integrase is lentivirus-specific and promotes DNA binding. *The Journal of biological chemistry* **280,** 17841–17847; 10.1074/jbc.M411681200 (2005).

331. Llano, M. *et al.* An essential role for LEDGF/p75 in HIV integration. *Science* **314,** 461–464; 10.1126/science.1132319 (2006).

332. Rijck, J. de, Bartholomeeusen, K., Ceulemans, H., Debyser, Z. & Gijsbers, R. High-resolution profiling of the LEDGF/p75 chromatin interaction in the ENCODE region. *Nucleic Acids Research* **38,** 6135–6147; 10.1093/nar/gkq410 (2010).

333. Hombrouck, A. *et al.* Virus evolution reveals an exclusive role for LEDGF/p75 in chromosomal tethering of HIV. *PLoS pathogens* **3,** e47; 10.1371/journal.ppat.0030047 (2007).

334. Singh, P. K. *et al.* LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes & development* **29,** 2287–2297; 10.1101/gad.267609.115 (2015).

335. Kotin, R. M. *et al.* Site-specific integration by adeno-associated virus. *Proceedings of the National Academy of Sciences of the United States of America* **87,** 2211–2215; 10.1073/pnas.87.6.2211 (1990).

336. Samulski, R. J. *et al.* Targeted integration of adeno-associated virus (AAV) into human chromosome 19. *The EMBO Journal* **10,** 3941–3950 (1991).

337. McCarty, D. M., Young, S. M. & Samulski, R. J. Integration of adeno-associated virus (AAV) and recombinant AAV vectors. *Annual review of genetics* **38,** 819–845; 10.1146/annurev.genet.37.110801.143717 (2004).

338. Giraud, C., Winocour, E. & Berns, K. I. Site-specific integration by adeno-associated virus is directed by a cellular DNA sequence. *Proceedings of the National Academy of Sciences of the United States of America* **91,** 10039–10043 (1994).

339. McCarty, D. M., Ryan, J. H., Zolotukhin, S., Zhou, X. & Muzyczka, N. Interaction of the adeno-associated virus Rep protein with a sequence within the A palindrome of the viral terminal repeat. *Journal of virology* **68,** 4998–5006 (1994).

340. Henckaerts, E. *et al.* Site-specific integration of adeno-associated virus involves partial duplication of the target locus. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 7571–7576; 10.1073/pnas.0806821106 (2009).

341. Papapetrou, E. P. & Schambach, A. Gene Insertion Into Genomic Safe Harbors for Human Gene Therapy. *Molecular Therapy* **24,** 678–684; 10.1038/mt.2016.38 (2016).

342. Goncalves, Manuel A F V. Adeno-associated virus: from defective virus to effective vector. *Virology journal* **2,** 43; 10.1186/1743-422X-2-43 (2005).

343. Chalker, D. L. & Sandmeyer, S. B. Ty3 integrates within the region of RNA polymerase III transcription initiation. *Genes & development* **6,** 117–128 (1992).

344. Devine, S. E. & Boeke, J. D. Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III. *Genes & development* **10,** 620–633 (1996).

345. Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M. & Pagano, A. The expanding RNA polymerase III transcriptome. *Trends in genetics : TIG* **23,** 614–622; 10.1016/j.tig.2007.09.001 (2007).

346. Ji, H. *et al.* Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. *Cell* **73,** 1007–1018 (1993).

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

348. Bryk, M. *et al.* Transcriptional silencing of Ty1 elements in the RDN1 locus of yeast. *Genes & development* **11,** 255–269 (1997).

349. Voigt, K., Izsvak, Z. & Ivics, Z. Targeted gene insertion for molecular medicine. *Journal of molecular medicine (Berlin, Germany)* **86,** 1205–1219; 10.1007/s00109-008-0381-8 (2008).

350. Bachman, N., Gelbart, M. E., Tsukiyama, T. & Boeke, J. D. TFIIIB subunit Bdp1p is required for periodic integration of the Ty1 retrotransposon and targeting of Isw2p to S. cerevisiae tDNAs. *Genes & development* **19,** 955–964; 10.1101/gad.1299105 (2005).

351. Curcio, M. J., Lutz, S. & Lesage, P. The Ty1 LTR-Retrotransposon of Budding Yeast, Saccharomyces cerevisiae. *Microbiology spectrum* **3,** MDNA3-0053-2014; 10.1128/microbiolspec.MDNA3-0053-2014 (2015).

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

353. Aye, M., Dildine, S. L., Claypool, J. A., Jourdain, S. & Sandmeyer, S. B. A truncation mutant of the 95-kilodalton subunit of transcription factor IIIC reveals asymmetry in Ty3 integration. *Molecular and cellular biology* **21,** 7839–7851; 10.1128/MCB.21.22.7839-7851.2001 (2001).

354. Yieh, L., Hatzis, H., Kassavetis, G. & Sandmeyer, S. B. Mutational analysis of the transcription factor IIIB-DNA target of Ty3 retroelement integration. *The Journal of biological chemistry* **277,** 25920–25928; 10.1074/jbc.M202729200 (2002).

355. Zou, S., Kim, J. M. & Voytas, D. F. The Saccharomyces retrotransposon Ty5 influences the organization of chromosome ends. *Nucleic Acids Research* **24,** 4825–4831 (1996).

356. Zou, S. & Voytas, D. F. Silent chromatin determines target preference of the Saccharomyces retrotransposon Ty5. *Proceedings of the National Academy of Sciences of the United States of America* **94,** 7412–7416 (1997).

357. Gai, X. & Voytas, D. F. A single amino acid change in the yeast retrotransposon Ty5 abolishes targeting to silent chromatin. *Molecular cell* **1,** 1051–1055 (1998).

358. Xie, W. *et al.* Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. *Molecular and cellular biology* **21,** 6606–6614 (2001).

359. Zhu, Y., Dai, J., Fuerst, P. G. & Voytas, D. F. Controlling integration specificity of a yeast retrotransposon. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 5891–5895; 10.1073/pnas.1036705100 (2003).

360. Palladino, F. *et al.* SIR3 and SIR4 proteins are required for the positioning and integrity of yeast telomeres. *Cell* **75,** 543–555; 10.1016/0092-8674(93)90388-7 (1993).

361. Cockell, M. *et al.* The carboxy termini of Sir4 and Rap1 affect Sir3 localization: evidence for a multicomponent complex required for yeast telomeric silencing. *The Journal of cell biology* **129,** 909–924 (1995).

362. Boeke, J. D. & Devine, S. E. Yeast retrotransposons: finding a nice quiet neighborhood. *Cell* **93,** 1087–1089 (1998).

363. Spaller, T., Kling, E., Glöckner, G., Hillmann, F. & Winckler, T. Convergent evolution of tRNA gene targeting preferences in compact genomes. *Mobile DNA* **7,** 17; 10.1186/s13100-016-0073-9 (2016).

364. Guo, Y. & Levin, H. L. High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in Schizosaccharomyces pombe. *Genome Research* **20,** 239–248; 10.1101/gr.099648.109 (2010).

365. Chatterjee, A. G. *et al.* Serial number tagging reveals a prominent sequence preference of retrotransposon integration. *Nucleic Acids Research* **42,** 8449–8460; 10.1093/nar/gku534 (2014).

366. Parks, A. R. *et al.* Transposition into replicating DNA occurs through interaction with the processivity factor. *Cell* **138,** 685–695; 10.1016/j.cell.2009.06.011 (2009).

367. Peters, J. E. & Craig, N. L. Tn7 recognizes transposition target structures associated with DNA replication using the DNA-binding protein TnsE. *Genes & development* **15,** 737–747; 10.1101/gad.870201 (2001).

368. Gringauz, E., Orle, K. A., Waddell, C. S. & Craig, N. L. Recognition of Escherichia coli attTn7 by transposon Tn7: lack of specific sequence requirements at the point of Tn7 insertion. *Journal of bacteriology* **170,** 2832–2840 (1988).

369. Bainton, R. J., Kubo, K. M., Feng, J. N. & Craig, N. L. Tn7 transposition: target DNA recognition is mediated by multiple Tn7-encoded proteins in a purified in vitro system. *Cell* **72,** 931–943 (1993).

370. Kuduvalli, P. N., Mitra, R. & Craig, N. L. Site-specific Tn7 transposition into the human genome. *Nucleic Acids Research* **33,** 857–863; 10.1093/nar/gki227 (2005).

371. Thorpe, H. M. & Smith, M. C. In vitro site-specific integration of bacteriophage DNA catalyzed by a recombinase of the resolvase/invertase family. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 5505–5510 (1998).

372. Groth, A. C., Olivares, E. C., Thyagarajan, B. & Calos, M. P. A phage integrase directs efficient site-specific integration in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 5995–6000; 10.1073/pnas.090527097 (2000).

373. Thyagarajan, B., Olivares, E. C., Hollis, R. P., Ginsburg, D. S. & Calos, M. P. Site-Specific Genomic Integration in Mammalian Cells Mediated by Phage C31 Integrase. *Molecular and cellular biology* **21,** 3926–3934; 10.1128/MCB.21.12.3926-3934.2001 (2001).

374. Olivares, E. C. *et al.* Site-specific genomic integration produces therapeutic Factor IX levels in mice. *Nature biotechnology* **20,** 1124–1128; 10.1038/nbt753 (2002).

375. Ortiz-Urda, S. *et al.* Stable nonviral genetic correction of inherited human skin disease. *Nature medicine* **8,** 1166–1170; 10.1038/nm766 (2002).

376. Held, P. K. *et al.* In vivo correction of murine hereditary tyrosinemia type I by phiC31 integrase-mediated gene delivery. *Molecular Therapy* **11,** 399–408; 10.1016/j.ymthe.2004.11.001 (2005).

377. Ehrhardt, A., Xu, H., Huang, Z., Engler, J. A. & Kay, M. A. A direct comparison of two nonviral gene therapy vectors for somatic integration: in vivo evaluation of the bacteriophage integrase phiC31 and the Sleeping Beauty transposase. *Molecular Therapy* **11,** 695–706; 10.1016/j.ymthe.2005.01.010 (2005).

378. Chalberg, T. W., Genise, H. L., Vollrath, D. & Calos, M. P. phiC31 integrase confers genomic integration and long-term transgene expression in rat retina. *Investigative ophthalmology & visual science* **46,** 2140–2146; 10.1167/iovs.04-1252 (2005).

379. Chavez, C. L. & Calos, M. P. Therapeutic applications of the PhiC31 integrase system. *Current gene therapy* **11,** 375–381 (2011).

380. Liu, J., Jeppesen, I., Nielsen, K. & Jensen, T. G. Phi c31 integrase induces chromosomal aberrations in primary human fibroblasts. *Gene therapy* **13,** 1188–1190; 10.1038/sj.gt.3302789 (2006).

381. Liu, J., Skjorringe, T., Gjetting, T. & Jensen, T. G. PhiC31 integrase induces a DNA damage response and chromosomal rearrangements in human adult fibroblasts. *BMC biotechnology* **9,** 31; 10.1186/1472-6750-9-31 (2009).

382. Groth, A. C., Fish, M., Nusse, R. & Calos, M. P. Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31. *Genetics* **166,** 1775–1782; 10.1534/genetics.166.4.1775 (2004).

383. Bischof, J., Maeda, R. K., Hediger, M., Karch, F. & Basler, K. An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 3312–3317; 10.1073/pnas.0611511104 (2007).

384. Voutev, R. & Mann, R. S. Robust ΦC31-Mediated Genome Engineering in Drosophila melanogaster Using Minimal attP/attB Phage Sites. *G3 (Bethesda, Md.)* **8,** 1399–1402; 10.1534/g3.118.200051 (2018).

385. Brady, T. L., Schmidt, C. L. & Voytas, D. F. Targeting integration of the Saccharomyces Ty5 retrotransposon. *Methods in molecular biology (Clifton, N.J.)* **435,** 153–163; 10.1007/978-1-59745-232-8_11 (2008).

386. Wang, H., Mayhew, D., Chen, X., Johnston, M. & Mitra, R. D. Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. *Genome Research* **21,** 748–755; 10.1101/gr.114850.110 (2011).

387. Wang, H., Mayhew, D., Chen, X., Johnston, M. & Mitra, R. D. "Calling cards" for DNA-binding proteins in mammalian cells. *Genetics* **190,** 941–949; 10.1534/genetics.111.137315 (2012).

388. Bushman, F. D. Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. *Proceedings of the National Academy of Sciences of the United States of America* **91,** 9233–9237 (1994).

389. Tan, W., Dong, Z., Wilkinson, T. A., Barbas, C. F. 3. & Chow, S. A. 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. *Journal of virology* **80,** 1939–1948; 10.1128/JVI.80.4.1939-1948.2006 (2006).

390. Goulaouic, H. & Chow, S. A. Directed integration of viral DNA mediated by fusion proteins consisting of human immunodeficiency virus type 1 integrase and Escherichia coli LexA protein. *Journal of virology* **70,** 37–46 (1996).

391. Bushman, F. D. & Miller, M. D. Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. *Journal of virology* **71,** 458–464 (1997).

392. Tan, W., Zhu, K., Segal, D. J., Barbas, C. F. & Chow, S. A. 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. *Journal of virology* **78,** 1301–1313; 10.1128/JVI.78.3.1301-1313.2004 (2004).

393. Peng, W.-J., Chang, C.-M. & Lin, T.-H. Target integration by a chimeric Sp1 zinc finger domain-Moloney murine leukemia virus integrase in vivo. *Journal of biomedical science* **9,** 171–184; 10.1007/bf02256029 (2002).

394. Katz, R. A., Merkel, G. & Skalka, A. M. 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,** 178–190; 10.1006/viro.1996.0105 (1996).

395. Kettlun, C., Galvan, D. L., George, A. L., Kaja, A. & Wilson, M. H. Manipulating piggyBac transposon chromosomal integration site selection in human cells. *Molecular Therapy* **19,** 1636–1644; 10.1038/mt.2011.129 (2011).

396. Owens, J. B. *et al.* Chimeric piggyBac transposases for genomic targeting in human cells. *Nucleic Acids Research* **40,** 6978–6991; 10.1093/nar/gks309 (2012).

397. Owens, J. B. *et al.* Transcription activator like effector (TALE)-directed piggyBac transposition in human cells. *Nucleic Acids Research* **41,** 9197–9207; 10.1093/nar/gkt677 (2013).

398. Szabó, M. *et al.* Transposition and targeting of the prokaryotic mobile element IS 30 in zebrafish. *FEBS letters* **550,** 46–50; 10.1016/S0014-5793(03)00814-7 (2003).

399. Feng, X., Bednarz, A. L. & Colloms, S. D. Precise targeted integration by a chimaeric transposase zinc-finger fusion protein. *Nucleic Acids Research* **38,** 1204–1216; 10.1093/nar/gkp1068 (2010).

400. Maragathavally, K. J., Kaminski, J. M. & Coates, C. J. Chimeric Mos1 and piggyBac transposases result in site-directed integration. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **20,** 1880–1882; 10.1096/fj.05-5485fje (2006).

401. Ciuffi, A., Diamond, T. L., Hwang, Y., Marshall, H. M. & Bushman, F. D. Modulating target site selection during human immunodeficiency virus DNA integration in vitro with an engineered tethering factor. *Human gene therapy* **17,** 960–967; 10.1089/hum.2006.17.960 (2006).

402. Gijsbers, R. *et al.* LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. *Molecular Therapy* **18,** 552–560; 10.1038/mt.2010.36 (2010).

403. Vets, S. *et al.* Transient Expression of an LEDGF/p75 Chimera Retargets Lentivector Integration and Functionally Rescues in a Model for X-CGD. *Molecular therapy. Nucleic acids* **2,** e77; 10.1038/mtna.2013.4 (2013).

404. Silvers, R. M. *et al.* Modification of integration site preferences of an HIV-1-based vector by expression of a novel synthetic protein. *Human gene therapy* **21,** 337–349; 10.1089/hum.2009.134 (2010).

405. Ferris, A. L. *et al.* Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 3135–3140; 10.1073/pnas.0914142107 (2010).

406. Vranckx, L. S., Demeulemeester, J., Debyser, Z. & Gijsbers, R. Towards a Safer, More Randomized Lentiviral Vector Integration Profile Exploring Artificial LEDGF Chimeras. *PloS one* **11,** e0164167; 10.1371/journal.pone.0164167 (2016).

407. Luo, W. *et al.* Comparative analysis of chimeric ZFP-, TALE- and Cas9-piggyBac transposases for integration into a single locus in human cells. *Nucleic Acids Research* **45,** 8411–8422; 10.1093/nar/gkx572 (2017).

408. Bhatt, S. & Chalmers, R. Targeted DNA transposition in vitro using a dCas9-transposase fusion protein. *Nucleic Acids Research* **47,** 8126–8135; 10.1093/nar/gkz552 (2019).

409. Hew, B. E., Sato, R., Mauro, D., Stoytchev, I. & Owens, J. B. RNA-guided piggyBac transposition in human cells. *Synthetic biology (Oxford, England)* **4,** ysz018; 10.1093/synbio/ysz018 (2019).

410. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* **571,** 219–225; 10.1038/s41586-019-1323-z (2019).

411. Strecker, J. *et al.* RNA-guided DNA insertion with CRISPR-associated transposases. *Science (New York, N.Y.)* **365,** 48–53; 10.1126/science.aax9181 (2019).

412. Yant, S. R., Huang, Y., Akache, B. & Kay, M. A. Site-directed transposon integration in human cells. *Nucleic Acids Research* **35,** e50; 10.1093/nar/gkm089 (2007).

413. Ivics, Z. *et al.* Targeted Sleeping Beauty transposition in human cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **15,** 1137–1144; 10.1038/sj.mt.6300169 (2007).

414. Ciganda, M. & Williams, N. Eukaryotic 5S rRNA biogenesis. *Wiley interdisciplinary reviews. RNA* **2,** 523–533; 10.1002/wrna.74 (2011).

415. Schöfer, C. & Weipoltshammer, K. Nucleolus and chromatin. *Histochemistry and cell biology* **150,** 209–225; 10.1007/s00418-018-1696-3 (2018).

416. Zhang, X., Eickbush, M. T. & Eickbush, T. H. Role of recombination in the long-term retention of transposable elements in rRNA gene loci. *Genetics* **180,** 1617–1626; 10.1534/genetics.108.093716 (2008).

417. Eickbush, T. H. & Eickbush, D. G. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* **175,** 477–485; 10.1534/genetics.107.071399 (2007).

418. Sakai, K. *et al.* Human ribosomal RNA gene cluster: identification of the proximal end containing a novel tandem repeat sequence. *Genomics* **26,** 521–526; 10.1016/0888-7543(95)80170-q (1995).

419. Andersen, J. S. *et al.* Nucleolar proteome dynamics. *Nature* **433,** 77–83; 10.1038/nature03207 (2005).

420. Farley, K. I., Surovtseva, Y., Merkel, J. & Baserga, S. J. Determinants of mammalian nucleolar architecture. *Chromosoma* **124,** 323–331; 10.1007/s00412-015-0507-z (2015).

421. Németh, A. & Längst, G. Genome organization in and around the nucleolus. *Trends in genetics : TIG* **27,** 149–156; 10.1016/j.tig.2011.01.002 (2011).

422. O'Sullivan, J. M., Pai, D. A., Cridge, A. G., Engelke, D. R. & Ganley, A. R. D. The nucleolus: a raft adrift in the nuclear sea or the keystone in nuclear structure? *Biomolecular concepts* **4,** 277–286; 10.1515/bmc-2012-0043 (2013).

423. McStay, B. Nucleolar organizer regions: genomic 'dark matter' requiring illumination. *Genes & development* **30,** 1598–1610; 10.1101/gad.283838.116 (2016).

424. Floutsakou, I. *et al.* The shared genomic architecture of human nucleolar organizer regions. *Genome Research* **23,** 2003–2012; 10.1101/gr.157941.113 (2013).

425. Zhou, J., Eickbush, M. T. & Eickbush, T. H. A population genetic model for the maintenance of R2 retrotransposons in rRNA gene loci. *PLoS genetics* **9,** e1003179; 10.1371/journal.pgen.1003179 (2013).

426. Haaf, T., Hayman, D. L. & Schmid, M. Quantitative determination of rDNA transcription units in vertebrate cells. *Experimental cell research* **193,** 78–86; 10.1016/0014-4827(91)90540-b (1991).

427. Penton, E. H., Sullender, B. W. & Crease, T. J. Pokey, a new DNA transposon in Daphnia (cladocera: crustacea). *Journal of molecular evolution* **55,** 664–673; 10.1007/s00239-002-2362-9 (2002).

428. Elliott, T. A., Stage, D. E., Crease, T. J. & Eickbush, T. H. In and out of the rRNA genes: characterization of Pokey elements in the sequenced Daphnia genome. *Mobile DNA* **4,** 20; 10.1186/1759-8753-4-20 (2013).

429. Eagle, S. H. C. & Crease, T. J. Distribution of the DNA transposon family, Pokey in the Daphnia pulex species complex. *Mobile DNA* **7,** 11; 10.1186/s13100-016-0067-7 (2016).

430. Nakai, H. *et al.* Large-scale molecular characterization of adeno-associated virus vector integration in mouse liver. *Journal of virology* **79,** 3606–3614; 10.1128/JVI.79.6.3606-3614.2005 (2005).

431. Miller, D. G. *et al.* Large-scale analysis of adeno-associated virus vector integration sites in normal human cells. *Journal of virology* **79,** 11434–11442; 10.1128/JVI.79.17.11434-11442.2005 (2005).

432. Johnson, J. S. & Samulski, R. J. Enhancement of adeno-associated virus infection by mobilizing capsids into and out of the nucleolus. *Journal of virology* **83,** 2632–2644; 10.1128/JVI.02309-08 (2009).

433. Lisowski, L. *et al.* Ribosomal DNA integrating rAAV-rDNA vectors allow for stable transgene expression. *Molecular therapy : the journal of the American Society of Gene Therapy* **20,** 1912–1923; 10.1038/mt.2012.164 (2012).

434. Emmott, E. & Hiscox, J. A. Nucleolar targeting: the hub of the matter. *EMBO reports* **10,** 231–238; 10.1038/embor.2009.14 (2009).

435. Certo, M. T. *et al.* Tracking genome engineering outcome at individual DNA breakpoints. *Nature methods* **8,** 671–676; 10.1038/nmeth.1648 (2011).

436. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature protocols* **8,** 2281–2308; 10.1038/nprot.2013.143 (2013).

437. Cheng, A. W. *et al.* Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell research* **23,** 1163–1171; 10.1038/cr.2013.122 (2013).

438. Lechertier, T., Sirri, V., Hernandez-Verdun, D. & Roussel, P. A B23-interacting sequence as a tool to visualize protein interactions in a cellular context. *Journal of cell science* **120,** 265–275; 10.1242/jcs.03345 (2007).

439. Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J. & Mateo, J. L. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PloS one* **10,** e0124633; 10.1371/journal.pone.0124633 (2015).

440. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Research* **40,** e115; 10.1093/nar/gks596 (2012).

441. Siomi, H., Shida, H., Maki, M. & Hatanaka, M. Effects of a highly basic region of human immunodeficiency virus Tat protein on nucleolar localization. *Journal of virology* **64,** 1803–1807 (1990).

442. Cochrane, A. W., Perkins, A. & Rosen, C. A. Identification of sequences important in the nucleolar localization of human immunodeficiency virus Rev: relevance of nucleolar localization to function. *Journal of virology* **64,** 881–885 (1990).

443. Valdez, B. C. *et al.* Identification of the nuclear and nucleolar localization signals of the protein p120. Interaction with translocation protein B23. *The Journal of biological chemistry* **269,** 23776–23783 (1994).

444. Siomi, H. *et al.* Sequence requirements for nucleolar localization of human T cell leukemia virus type I pX protein, which regulates viral RNA processing. *Cell* **55,** 197–209; 10.1016/0092-8674(88)90043-8 (1988).

445. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, Vienna, Austria, 2017).

446. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)* **34,** i884-i890; 10.1093/bioinformatics/bty560 (2018).

447. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9,** 357–359; 10.1038/nmeth.1923 (2012).

448. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078–2079; 10.1093/bioinformatics/btp352 (2009).

449. Akalin, A., Franke, V., Vlahoviček, K., Mason, C. E. & Schübeler, D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics (Oxford, England)* **31,** 1127–1129; 10.1093/bioinformatics/btu775 (2015).

450. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442,** 772–778; 10.1038/nature04979 (2006).

451. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Research* **42,** e168; 10.1093/nar/gku936 (2014).

452. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921; 10.1038/35057062 (2001).

453. Price, A. L., Eskin, E. & Pevzner, P. A. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Research* **14,** 2245–2252; 10.1101/gr.2693004 (2004).

454. Silverman, L. J., Kelley, W. N. & Palella, T. D. Genetic analysis of human hypoxanthine-guanine phosphoribosyltransferase deficiency. *Enzyme* **38,** 36–44 (1987).

455. Liao, S., Tammaro, M. & Yan, H. Enriching CRISPR-Cas9 targeted cells by co-targeting the HPRT gene. *Nucleic Acids Research* **43,** e134; 10.1093/nar/gkv675 (2015).

456. Ma, H. *et al.* CRISPR-Cas9 nuclear dynamics and target recognition in living cells. *The Journal of cell biology* **214,** 529–537; 10.1083/jcb.201604115 (2016).

457. Louvet, E., Junéra, H. R., Berthuy, I. & Hernandez-Verdun, D. Compartmentation of the nucleolar processing proteins in the granular component is a CK2-driven process. *Molecular biology of the cell* **17,** 2537–2546; 10.1091/mbc.e05-10-0923 (2006).

458. Chiarella, S. *et al.* Nucleophosmin mutations alter its nucleolar localization by impairing G-quadruplex binding at ribosomal DNA. *Nucleic Acids Research* **41,** 3228–3239; 10.1093/nar/gkt001 (2013).

459. Németh, A. *et al.* Initial genomics of the human nucleolus. *PLoS genetics* **6,** e1000889; 10.1371/journal.pgen.1000889 (2010).

460. Dillinger, S., Straub, T. & Németh, A. Nucleolus association of chromosomal domains is largely maintained in cellular senescence despite massive nuclear reorganisation. *PloS one* **12,** e0178821; 10.1371/journal.pone.0178821 (2017).

461. Ding, S. *et al.* Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122,** 473–483; 10.1016/j.cell.2005.07.013 (2005).

462. Oakes, B. L., Nadler, D. C. & Savage, D. F. Protein engineering of Cas9 for enhanced function. *Methods in enzymology* **546,** 491–511; 10.1016/B978-0-12-801185-0.00024-6 (2014).

463. Tanaka, Y., Yamashita, R., Suzuki, Y. & Nakai, K. Effects of Alu elements on global nucleosome positioning in the human genome. *BMC Genomics* **11,** 309; 10.1186/1471-2164-11-309 (2010).

464. Oakes, B. L. *et al.* CRISPR-Cas9 Circular Permutants as Programmable Scaffolds for Genome Modification. *Cell* **176,** 254-267.e16; 10.1016/j.cell.2018.11.052 (2019).

465. Guilinger, J. P., Thompson, D. B. & Liu, D. R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nature biotechnology* **32,** 577–582; 10.1038/nbt.2909 (2014).

466. Sinzelle, L. *et al.* Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 4715–4720; 10.1073/pnas.0707746105 (2008).

467. Kuppuswamy, M., Subramanian, T., Srinivasan, A. & Chinnadurai, G. Multiple functional domains of Tat, the trans-activator of HIV-1, defined by mutational analysis. *Nucleic Acids Research* **17,** 3551–3561; 10.1093/nar/17.9.3551 (1989).

468. Ruben, S. *et al.* Structural and functional characterization of human immunodeficiency virus tat protein. *Journal of virology* **63,** 1–8 (1989).

469. Coiras, M. *et al.* Modifications in the human T cell proteome induced by intracellular HIV-1 Tat protein expression. *Proteomics* **6 Suppl 1,** S63-73; 10.1002/pmic.200500437 (2006).

470. Stauber, R. H. & Pavlakis, G. N. Intracellular trafficking and interactions of the HIV-1 Tat protein. *Virology* **252,** 126–136; 10.1006/viro.1998.9400 (1998).

471. Efthymiadis, A., Briggs, L. J. & Jans, D. A. The HIV-1 Tat nuclear localization sequence confers novel nuclear import properties. *The Journal of biological chemistry* **273,** 1623–1628; 10.1074/jbc.273.3.1623 (1998).

472. Carmo-Fonseca, M., Mendes-Soares, L. & Campos, I. To be or not to be in the nucleolus. *Nature cell biology* **2,** E107-12; 10.1038/35014078 (2000).

473. Holmberg Olausson, K., Nistér, M. & Lindström, M. S. Loss of nucleolar histone chaperone NPM1 triggers rearrangement of heterochromatin and synergizes with a deficiency in DNA methyltransferase DNMT3A to drive ribosomal DNA transcription. *The Journal of biological chemistry* **289,** 34601–34619; 10.1074/jbc.M114.569244 (2014).

474. Yusufzai, T. M., Tagami, H., Nakatani, Y. & Felsenfeld, G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Molecular cell* **13,** 291–298; 10.1016/s1097-2765(04)00029-2 (2004).

475. Foltz, D. R. *et al.* The human CENP-A centromeric nucleosome-associated complex. *Nature cell biology* **8,** 458–469; 10.1038/ncb1397 (2006).

476. Padeken, J. & Heun, P. Nucleolus and nuclear periphery: velcro for heterochromatin. *Current opinion in cell biology* **28,** 54–60; 10.1016/j.ceb.2014.03.001 (2014).

477. Fankhauser, C., Izaurralde, E., Adachi, Y., Wingfield, P. & Laemmli, U. K. Specific complex of human immunodeficiency virus type 1 rev and nucleolar B23 proteins:

dissociation by the Rev response element. *Molecular and cellular biology* **11,** 2567–2575; 10.1128/mcb.11.5.2567 (1991).

478. Okuwaki, M., Tsujimoto, M. & Nagata, K. The RNA binding activity of a ribosome biogenesis factor, nucleophosmin/B23, is modulated by phosphorylation with a cell cycle-dependent kinase and by association with its subtype. *Molecular biology of the cell* **13,** 2016–2030; 10.1091/mbc.02-03-0036 (2002).

479. Thompson, M., Haeusler, R. A., Good, P. D. & Engelke, D. R. Nucleolar clustering of dispersed tRNA genes. *Science (New York, N.Y.)* **302,** 1399–1401; 10.1126/science.1089814 (2003).

480. Haeusler, R. A. & Engelke, D. R. Spatial organization of transcription by RNA polymerase III. *Nucleic Acids Research* **34,** 4826–4836; 10.1093/nar/gkl656 (2006).

481. Aird, E. J., Lovendahl, K. N., St. Martin, A., Harris, R. S. & Gordon, W. R. Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template. *Communications Biology* **1,** 54; 10.1038/s42003-018-0054-2.

482. Kovač, A. *et al.* RNA-guided retargeting of Sleeping Beauty transposition in human cells. *eLife* **9**; 10.7554/eLife.53868 (2020).

6 Abbreviations

6-TG	6-thioguanine
AAV	Adeno-associated virus
APS	Ammonium persulfate
BAC	Bacterial artificial chromosome
bp	Base pair(s)
btm	Bottom (strand)
CAR-T	Chimeric antigen receptor T cell
crRNA	CRISPR RNA
DBD	DNA-binding domain
DR	Direct repeat
DSB	Double strand break
EDTA	Ethylenediaminetetraacetic acid
EMSA	Electrophoretic mobility shift assay
EtOH	Ethanol
ETS	External transcribed spacer
FACS	Fluorescence-activated cell sorting
FCS	Fetal calf serum
fwd	Forward (primer)
gDNA	Genomic DNA
GOI	Gene of interest
gRNA	Guide RNA
GSH	Genomic safe harbor
HD	Hydrodynamic injection
HDR	Homology directed repair / half direct repeat
HIV	Human immunodeficiency virus
HPRT	Hypoxanthine-guanine phosphoribosyltransferase
HR	Homologous recombination
HSV	Herpes simplex virus
HTH	Helix-turn-helix
HTLV	Human T-lymphotropic virus
IF	Immunofluorescence
IGS	Intergenic spacer
IN	Integrase

iPSC	Induced pluripotent stem cell
IR	Inverted repeat
IR/DR	Inverted repeat / direct repeat
IRES	Internal ribosomal entry site
ITR	Inverted terminal repeat
ITS	Internal transcribed spacer
LEDGF	Lens epithelium derived growth factor
LINE	Long interspersed nuclear element
LTR	Long terminal repeat
LSU	Large subunit
MLV	Murine leukemia virus
NAD	Nucleolus-associated domain
NHEJ	Non-homologous end-joining
NLS	Nuclear localization signal
NoLS	Nucleolar localization signal
NOR	Nucleolar organizer region
nt	Nucleotide(s)
Oligo	Oligonucleotide
OPI	Overproduction inhibition
ORF	Open reading frame
PB	PiggyBac
PBS	Phosphate buffered saline
PBS-T	PBS + Tween
PFA	para-Formaldehyde
PEI	Polyethyleneimine
PEC	Paired-end complex
rAAV	Recombinant Adeno-associated virus
rev	Reverse (primer)
rDNA	Ribosomal DNA
RNP	Ribonucleoprotein
rRNA	Ribosomal RNA
RRS	Rep recognition sequence
RT	Room temperature
SB	Sleeping Beauty

SDS	Sodium dodecyl sulfate
sgRNA	Single guide RNA
SIN	Self inactivating
SINE	Short interspersed nuclear element
ssDNA	Single stranded DNA
SSU	Small subunit
TALE	Transcription activator-like effector
TALEN	Transcription activator-like effector nuclease
TBS	Tris-buffered saline
TBS-T	TBS + Tween
tDNA	Target DNA
TCC	Target capture complex
TE	Transposable element
TLR	Traffic light reporter
Tpase	Transposase
Tpon	Transposon
tracrRNA	Transactivating crRNA
tRNA	Transfer RNA
TSD	Target site duplication
TSS	Transcription start site
wt	Wild type
ZFN	Zinc finger nuclease
ZFP	Zinc finger protein

7 List of figures

8 List of tables

9 Summary

Gene therapy is an emerging field which encompasses different modifications to the genomes of target cells and organisms with the aim of treating a range of diseases. In general, two major types of genome editing tools are used: integrating vectors and nuclease-based technologies.

Nuclease-based genome editing is particularly useful due to its high specificity; while some off-target effects exist, editing with nucleases generally occurs at precisely defined positions in the target genome. Designer nucleases like ZFNs and TALENs can be designed to target a wide range of sites and the CRISPR/Cas9 system offers even greater flexibility by recognizing target sites through a provided short RNA molecule called gRNA.

Nuclease-based genome engineering can be used to either disrupt sequences or make precise edits (sequence replacements, insertions or deletions), depending on the repair pathway used by the target cell. If only a nuclease is provided, introduced double strand breaks (DSBs) are repaired by the non-homologous end-joining (NHEJ) pathway, which directly joins the two free DNA ends. Because this process is imprecise, small insertions and deletions (indels) are often produced at the target site. If the target is within an exon, this will often result in a frameshift mutation and inactivation of the target gene. Supplying a homology template with the nuclease allows the cell to repair DSBs via homology-directed repair (HDR), which allows precise edits to be made but is less efficient than the NHEJ pathway. The efficiency of insertion of sequences is inversely correlated with the size of the insert, thus it is often difficult to insert long sequences like entire genes by HDR.

For this reason, integrating vectors are often used for gene addition. Integrating vectors comprise both viral and transposon-based vectors, two types of systems that allow a genetic cargo to be stably integrated into the genome. While viral vectors generally have higher efficiencies, transposon vectors are cheaper and easier to use and do not have a fixed size limit for their cargo. The transposons used for gene therapy are generally DNA transposons that move by a cut-and-paste mechanism, i.e. the transposase molecule excises the DNA sequence flanked by the transposon inverted terminal repeats (ITRs) and integrates it elsewhere. In a biotechnological context, the transposon is usually excised from a plasmid molecule and integrated into the genome of a target cell.

The *Sleeping Beauty* transposon is one of the most widely used transposon vectors and combines high activity with a favorable insertion profile compared to viral vectors or other transposons like *piggyBac*. While integrating vectors are efficient at inserting DNA, they lack

the specificity of nuclease-based genome engineering. Some vectors have preferences regarding their target site selection, but the choice of integration site is ultimately random. This is problematic in a gene therapy context due to position effects and especially due to genotoxicity, which can result in edited cells forming tumors.

To address this problem, several attempts have been made to re-target integrating vectors to improve their safety profile. For the SB system, several DNA-binding domains (DBDs) have been fused to the transposase or to components that noncovalently interact with either the transposase or the transposon. This is done in order to physically tether paired-end complexes (PECs) to the DBD-defined target sites. Several of these constructs have been shown to cause some enrichment of SB integrations near their target sites, although generally at low frequencies when compared to background insertions.

The main aim of this project was to generate a new targeted SB system that uses catalytically inactive Cas9 (dCas9) as a DBD, which should allow flexible retargeting of SB insertions by providing different gRNAs. To this end, both the direct fusion and adapter protein approach were tested. For the direct fusion approach, the dCas9 domain was fused to the N-terminus of the hyperactive SB transposase SB100X (no C-terminal fusion was generated because C-terminal additions to the SB transposase have been previously shown to completely abolish transpositional activity). To test targeting via adapter proteins, dCas9 was fused both to the N-terminus and the C-terminus of the N57 peptide. N57 comprises the N-terminal PAI subdomain of the SB transposase, which interact with the SB transposon as well as with other SB transposase molecules. Additionally, dCas9 was added to the N-terminus of N123, a peptide that spans the entire C-terminal domain of SB transposase and has a stronger DNA-binding activity than N57.

The individual domains in all generated constructs were tested for activity in a range of different assays. The SB domain of dCas9-SB100X was tested in a transposition assay and was shown to have reduced, but detectable transpositional activity. The N57 and N123 domains in their respective fusions were tested for DNA binding in an EMSA. While dCas9-N57 and dCas9-N123 retained detectable binding to their recognition sequences in the SB transposon, the N57-dCas9 fusion did not and was excluded from further experiments. Finally, the DNA-binding activity of dCas9 in all fusions was tested by proxy by analyzing the cleavage activity of analogous fusions containing Cas9 instead of dCas9. All fusion proteins retained detectable cleavage activity in a *HPRT* disruption assay.

As all domains in the fusions dCas9-SB100X, dCas9-N57 and dCas9-N123 retained detectable activities, they were used in an attempt to direct integrations to various target sites. Three targets were chosen initially: the AluY and L1 retrotransposon as multicopy targets and the *HPRT* gene as a single-copy target. Several sgRNAs were tested for each target and the most efficient one was chosen for subsequent targeting experiments. For AluY-1 the chosen sgRNA (sgAluY-1) had around 300,000 target sites and was shown to cause clear fragmentation of gDNA *in vitro*. For L1, a sgRNA with around 5000 target sites was chosen (sgL1-1), based on efficient *in vitro* digestion of a plasmid fragment containing the target site. For *HPRT*, effectivity of the sgRNA (sgHPRT-0) was verified in a 6-TG selection assay and via TIDE assay.

Targeting against AluY showed the best results, although the overall efficiency remained low. Using the direct fusion transposase dCas9-SB100X and sgAluY-1 resulted in a significant enrichment of insertions into a 300 bp window downstream of the target site when compared to the same targeting construct with the sgRNA sgL1-1. With dCas9-N57, slight symmetrical enrichment around the target site could be observed, but it was not statistically significant. The enrichment observed with dCas9-SB100X was shown to occur into a TA-poor region that is normally disfavored for SB integration.

For targeting of the L1 element, no statistically significant enrichment could be observed for any targeting construct. While an increased tendency to use target sites in a TA-poor stretch downstream of the target sites was observed, the low number of total insertions made statistical testing impossible. For the *HPRT* gene, no targeting effect could be seen. No insertions occurred near the targeted site on the X chromosome and only a single insertion was found close to a site with three mismatches to the intended target site.

The fact that no insertions could be recovered from the vicinity of the *HPRT* target site led us the theory that the *HPRT* gene might be a poor target for SB insertions. Consequently, we identified three sites in the human genome where untargeted SB insertions had occurred close to one another, suggesting that these sites represented good targets for the SB transposon. The sites were also chosen on the basis of fulfilling genomic safe harbor (GSH) criteria. The sgRNAs designed for targeting these sites were tested using a T7 endonuclease assay.

After identifying a highly active sgRNA for each of the three sites, targeting of these sites was tested with the constructs dCas9-SB100X, dCas9-N57 and dCas9-N123. However, no insertions close to the targeted sites or any mismatched target sites were observed.

In another attempt to increase the specificity of the system, the SB100X components of the targeting system were replaced by the SB mutant SB(K248R), which has a high propensity to integrate into TATATATA octanucleotides. This was coupled with specifically targetable sites embedded in simple TA repeats, which should present a high number of potential targets for SB(K248R). Three such sites were identified and sgRNAs for these sites were designed and tested in an *in vitro* assay. Targeting for all sites was tested with the constructs dCas9-SB(K248R), dCas9-N57 and dCas9-N123. However, no insertions at the targeted sites could be recovered using any of the targeting constructs.

One observation that was made for all single-copy targeting libraries was that, independently of the sgRNA used, using dCas9 fusions resulted in a characteristic distribution of insertions along the chromosomes, usually with preferred regions near the chromosome ends. This pattern was not observed with unfused transposase in combination with adapter proteins. It remains to be seen what causes this distribution and why these regions seem to be preferred over the sgRNA-defined target sites.

The failure to target any of 7 chosen single-copy sites led to the conclusion that the specificity of the system in its current form is too low due to the high number of background insertions. In order to address this, an attempt was made to generate SB transposase mutants with a reduced DNA affinity and reduced transpositional activity which could be partially rescued by the addition of an active DBD. To this end, 19 residues that are either positively charged or implicated in binding of the target DNA were mutated to alanine and the mutants were fused to dCas9. The transpositional activity of these fusions was tested both with and without the presence of the multicopy sgRNA sgAluY-1 in order to test whether DNA binding of dCas9 would have a positive impact on transpositional activity. However, no such increase was observed for any of the mutants.

Consequently, the 19 mutations were randomly recombined in a SB mutant library and almost 1000 individual mutants were tested for increased transpositional activity when combined with dCas9 and sgAluY-1. One of the mutants, called SB(C42), when fused to dCas9, was found to have a 2-fold increased transpositional activity after the addition of sgAluY-1 and a 1.5-fold increased activity after the addition of sgL1-1. No significant increase could be seen after the addition of single-copy sgRNAs.

The mutant construct dCas9-SB(C42) was tested for targeting of GSH/'hotspot' loci and the *AAVS1* locus, but no targeting of these single-copy loci could be observed. Testing of targeted

transposition with dCas9-SB(C42) was started, but data analysis could not be completed in time for inclusion in this thesis.

Targeting with both dCas9-SB100X and dCas9-SB(C42) was also tested in an alternative setup where the transposon plasmids were delivered 36 h after the transposase fusions. Targeting in this manner was tested for several single-copy loci, but no targeting could be observed, which might have been caused by the low overall transpositional efficiency observed with this setup.

In a side project, targeting of SB insertions to ribosomal DNA was attempted by fusing nucleolar localization signals (NoLSs) to the SB transposase in order to localize it to the nucleolus. While fusions with four different NoLS peptides failed to localize the transposase to nucleoli, addition of the protein B23 had the desired effect. However, while an enrichment of the construct in the nucleolus was observed, it was also found to be distributed throughout the nucleus and nucleolar localization could not be observed for all cells. Analysis of the distribution of insertions generated with B23-SB100X showed statistically significant enrichment into nucleolar organizer region (NOR) sequences and into nucleolus-associated domains (NADs) and some association with a ChIP-seq dataset of UBTF, a transcription factor of DNA polymerase I, which is present in the nucleolus. However, overall insertion rates into NORs remained low (<1%), possibly owing to the fact that the construct was not exclusively found in the nucleoli of transfected cells.

In a second side project, an attempt was made to increase the ratio of HDR to NHEJ after Cas9 editing by using fusions of Cas9 to N57 and N123 and including the recognition sequences of N57/N123 in the HDR donors. However, no binding-site dependent increase in the HDR/NHEJ could be observed. While the fusion constructs, especially Cas9-N57, had a higher HDR/NHEJ ratio than unfused Cas9, this effect was also seen for donor constructs with no recognition sites and apparently was caused by the constructs themselves rather than by specific recruitment of donor constructs to the target sites.

10 Zusammenfassung

Der Begriff Gentherapie beschreibt eine Reihe von neuartigen Therapieansätzen, bei denen Zellen auf genomischer Ebene modifiziert werden, um verschiedene Krankheiten zu behandeln. Genomengineering-Werkzeuge können grob in zwei Klassen unterteilt werden: integrierende Vektoren und Nuklease-basierte Methoden.

Nuklease-basierte Technologien zeichnen sich vor allem durch ihre hohe Spezifizität aus. Off-Target-Effekte sind zwar vorhanden, aber in der Regel wird das Genom an genau vorgegebenen Stellen modifiziert. Es ist möglich für fast jede Sequenz Designer-Nukleasen wie ZFNs und TALENs zu entwickeln und das CRISPR/Cas9-System ist besonders flexibel, da es durch kurze RNA-Moleküle, gRNAs genannt, gesteuert wird.

Nukleasen können verwendet werden, um Sequenzen zu inaktivieren oder um präzise Veränderungen zu bewirken (Ersetzungen, Insertionen oder Deletionen), je nachdem, welcher Reparaturmechanismus von der Zielzelle verwendet wird. Wenn nur die Nuklease vorhanden ist, werden Doppelstrangbrüche (DSBs) über *non-homologous end joining* (NHEJ) repariert. Dabei werden die freien DNA-Enden direkt zusammengefügt. Da dieser Vorgang ungenau ist, entstehen oft kleine Insertionen und Deletionen (indels), was, wenn der DSB in einem Exon liegt, zu Frameshift-Mutationen und Inaktivierung des Zielgens führt. Wenn ein homologes DNA-Molekül als Vorlage zugegeben wird, kann der DSB über *homology-directed repair* (HDR) repariert werden. So können präzise Veränderungen vorgenommen werden, allerdings ist HDR generell weniger effizient als NHEJ. Besonders die Insertion langer Sequenzen ist ineffizient, was es schwierig macht, ganze Gene über HDR ins Genom einzubauen.

Aufgrund der relativen Ineffizienz von HDR für Insertionen werden für Genaddition oft integrierende Vektoren verwendet. Integrierende Vektoren können in virale und Transposon-basierte Vektoren aufgeteilt werden; beide Typen können DNA stabil ins Genom integrieren. Virale Vektoren sind generell effizienter als nichtvirale, dafür sind Transposon-basierte Vektoren einfacher und günstiger in der Benutzung und haben keine feste Obergrenze bezüglich der Größe der integrierten DNA. Für Gentherapie werden generell DNA-Transposons verwendet, die sich über einen *cut-and-paste* Mechanismus mobilisieren. Hierbei schneidet die Transposase eine DNA-Sequenz aus, die von zwei invertierten terminalen Repeats (ITRs) umgeben ist, und integriert sie an anderer Stelle in die Ziel-DNA. In biotechnologischen Anwendungen wird das Transposon in der Regel von einem Plasmid mobilisiert und in das Genom der Zielzelle integriert.

Das *Sleeping Beauty* Transposon-System gehört zu den am häufigsten verwendeten Transposonvektoren und kombiniert hohe Aktivität mit einem günstigen Insertionsprofil (im Vergleich zu viralen Vektoren und anderen Transposons wie *piggyBac*). Integrierende Vektoren bauen DNA zwar effizient in Zielgenome ein, ihnen fehlt allerdings die Spezifizität der zuvor beschriebenen Nukleasen. Manche Vektoren integrieren bevorzugt in der Nähe bestimmter genomischer Elemente, aber letztendlich ist die Auswahl der Integrationstelle auf genomischer Ebene zufällig. Im Kontext der Gentherapie ist dies problematisch, sowohl wegen Positionseffekten als auch wegen Genotoxizität, die zu Tumorentwicklung führen kann.

Als Antwort auf diese Probleme wurden mehrere Versuche unternommen um integrierende Vektoren umzusteuern um ihre Sicherheit zu verbessern. Für das SB-System wurden mehrere DNA-bindende Domänen (DBDs) entweder direkt mit der Transposase oder mit Domänen, die nicht-kovalent mit der Transposase oder dem Transposon interagieren, fusioniert. Dies hat den Zweck, Transposon *paired-end*-Komplexe (PECs) in die Nähe der Zielstellen der DBDs zu bringen und Integration an diesen genomischen Loci zu begünstigen. Einige dieser Konstrukte konnten Anreicherungen von Integrationen in der Nähe der DBD-definierten Ziele bewirken, allerdings war die Anzahl dieser Integrationen generell niedrig im Vergleich zu ungezielten Hintergrundintegrationen.

Das Hauptziel dieses Projekts war, ein neues zielbares SB-System zu entwickeln, das katalytisch inaktives Cas9 (dCas9) als DBD benutzt, was es ermöglichen sollte, SB-Integrationen flexibel durch verschiedene gRNAs zu steuern. Hierfür wurden sowohl eine Fusionstransposase als auch Adapterproteine verwendet. Für die Fusionstransposase wurde dCas9 am N-Terminus der hyperaktiven Transposase SB100X hinzugefügt (C-terminale Additionen an SB100X zerstören die Transpositionsaktivität des Enzyms). Als Adapterproteine wurde dCas9 sowohl am N- als auch am C-Terminus von N57 hinzugefügt. N57 besteht aus der C-terminalen PAI Subdomäne, die sowohl mit dem Transposon als auch mit SB-Transposase-Molekülen interagiert. Außerdem wurde dCas9 am N-terminus von dCas9-N123 angebaut. N123 besteht aus der gesamten C-terminalen Domäne der SB-Transposase und interagiert stärker mit dem SB-Transposon als N57.

Die einzelnen Domänen im Kontext der Fusionskonstrukte wurden in verschiedenen Assays auf ihre Aktivität getestet. Die SB-Domäne in dCas9-SB100X wurde in einem Transpositionsassay getestet und es wurde reduzierte, aber nachweisbare Aktivität festgestellt. Die N57- und N123-Domänen wurden in einem EMSA getestet. Interaktion mit einer Sequenz aus dem SB-Transposon konnte für dCas9-N57 und dCas9-N123 festgestellt werden, jedoch

nicht für N57-dCas9, daher wurde das letzte Konstrukt von weiteren Experimenten ausgeschlossen. Um zu testen, ob die Fusionsproteine über ihre dCas9-Domäne und eine sgRNA an ihre Zielsequenzen binden können, wurden äquivalente Fusionen mit Cas9 auf ihre Nuklease-Aktivität getestet. Alle Fusionsproteine zeigten in einem *HPRT*-Disruptionsassay nachweisbare Aktivität der Cas9-Domäne.

Da alle Domänen in den Fusionsproteinen dCas9-SB100X, dCas9-N57 und dCas9-N123 nachweisbare Aktivitäten behielten, wurden die Konstrukte getestet, um eine Reihe verschiedener Ziele im Genom anzusteuern: die repetitiven AluY- und L1-Retrotransposons und das *HPRT*-Gen. Für jedes Ziel wurden mehrere sgRNAs entworfen und getestet, um hohe Aktivität zu gewährleisten. Die sgRNA für AluY, sgAluY-1, hat ca. 300.000 Zielsequenzen und bewirkte klare Fragmentierung von genomischer DNA *in vitro*. Die sgRNA für L1, sgL1-1, mit rund 5000 Zielstellen konnte ein Plasmidfragment *in vitro* mit hoher Effizienz schneiden. Die sgRNA sgHPRT-0 wurde über 6-TG-Selektion und einen TIDE-Assay getestet und zeigte ebenfalls hohe Aktivität.

Mit der sgRNA sgAluY-1 konnten die besten Ergebnisse erzielt werden, obwohl der Prozess insgesamt weiterhin ineffizient war. Verglichen mit dem gleichen Konstrukt und sgL1-1 konnte mit dCas9-SB100X/sgAluY-1 eine signifikante Anreicherung von Insertionen in einem Fenster von 300 bp hinter den Zielsequenzen, innerhalb des AluY-Elements, erreicht werden. Mit dCas9-N57 konnte eine leichte Anreicherung in einem symmetrischen 300-bp-Fenster festgestellt werden, die allerdings statistisch nicht signifikant war. Die Anreicherung mit dCas9-SB100X fand in einer TA-armen Region statt, die normalerweise kein bevorzugtes Ziel für SB-Integrationen darstellt.

Mit sgL1-1 wurde kein klarer Zieleffekt festgestellt. Es gab zwar eine leichte Anreicherung, wieder in einem TA-armen Gebiet hinter den Zielsequenzen, wegen der niedrigen Anzahl an Insertionen um die Zielstellen war dies jedoch nicht statistisch signifikant. Für das *HPRT*-Gen konnte ebenfalls kein Targeting-Effekt demonstriert werden. Keine Insertionen wurden in der Nähe der Zielstelle auf dem X-Chromosom nachgewiesen und nur eine einzige Insertion wurde in der Nähe einer Sequenz mit drei nichtübereinstimmenden Basen gefunden.

Da keine Insertionen in der Nähre des *HPRT*-Gens festgestellt wurden, entwickelten wir die Theorie, dass das *HPRT*-Gen womöglich ein ungünstiges Ziel für SB-Integrationen darstellt. Um dies zu testen, wurden drei gnomische Loci identifiziert, an denen mehrere ungezielte SB-Integrationen nachgewiesen wurden und die daher offenbar rezeptiv für SB-Insertionen sind.

Es wurde weiterhin darauf geachtet, dass die Loci Kriterien für *genomic safe harbors* (GSHs)-erfüllen. Die sgRNAs für diese Ziele wurden in einem T7-Endonuklease-Assay getestet. Nachdem für jedes Ziel eine aktive sgRNA identifiziert wurde, wurde versucht, diese Loci mit dCas9-SB100X, dCas9-N57 und dCas9-N123 anzusteuern, jedoch konnten keine Integrationen in der Nähe der Zielsequenzen nachgewiesen werden.

Ein weiterer Versuch, die Spezifizität des Systems zu erhöhen, bestand darin, SB100X durch SB(K248R) zu ersetzen. SB(K248R) ist eine Mutante die bevorzugt in die Sequenz TATATATA integriert. Dies wurde mit drei neuen Zielsequenzen kombiniert, die in einfache TA-Repeats eingebettet sind und daher eine hohe Anzahl an möglichen Zielsequenzen für SB(K248R) bieten sollten. Drei solcher Loci wurden identifiziert und neue sgRNAs wurden entworfen und in einem *in-vitro*-Assay getestet. Die sgRNAs wurden in Kombination mit den Konstrukten dCas9-SB(K248R), dCas9-N57 und dCas9-N123 verwendet, jedoch konnten keine Integrationen in der Nähe der Zielsequenzen gefunden werden.

In allen Targeting-Libraries mit single-copy sgRNAs wurde festgestellt, dass Integrationen mit dCas9-SB-Fusionen in der Regel klar in bestimmten Regionen der Zielchromosomen angereichert sind, oft in der Nähe der Chromosom-Enden. Dieses Muster war unabhängig von der benutzten sgRNA und wurde nicht in den äquivalenten Libraries mit Adapterproteinen beobachtet. Es ist noch unklar, warum diese chromosomalen Regionen bevorzugt werden.

Die Tatsache, dass keins der 7 getesteten single-copy Ziele angesteuert werden konnte, führte zu dem Schluss, dass das dCas9-SB100X-System in dieser Form nicht spezifisch genug ist, um die große Menge an Hintergrundinsertionen zu überwinden. Daher wurde ein Versuch unternommen, eine SB-Transposase-Mutante zu entwickeln, die weniger aktiv ist als SB100X, deren Aktivität jedoch durch die Fusion mit einer aktiven DBD erhöht werden kann. Um dies zu erreichen, wurden 19 Aminosäuren, die entweder positiv geladen sins oder vermutlich mit der Ziel-DNA interagieren, durch Alanin ersetzt. Die mutierten Transposasen wurden mit dCas9 fusioniert und die Aktivität der Fusionen wurde mit der sgRNA sgAluY-1 getestet und mit dem gleichen Konstrukt ohne sgRNA verglichen. Keine der 19 getesteten Mutanten zeigte den gewünschten Phänotyp, d.h. erhöhte Transposition in Kombination mit sgAluY-1.

Da keine Einzelmutante den gewünschten Phänotyp hatte, wurden die Mutationen zufällig in eine Mutations-Library kombiniert und fast 1000 einzelne Mutanten wurden auf denselben Effekt getestet. Tatsächlich wurde mit einer Mutante, SB(C42), sgRNA-abhängig erhöhte Transposition festgestellt. Mit der sgRNA sgAluY-1 erhöhte sich die Aktivität auf das

Zweifache, mit sgL1-1 auf das Eineinhalbfache. Single-copy sgRNAs konnten hingegen keine nachweisbare Erhöhung der Aktivität bewirken.

Daraufhin wurde ein Versuch unternommen, mit der Fusionstransposase dCas9-SB(C42) die GSH/“hotspot“-Ziele anzusteuern, es konnten aber keine Insertionen an den ausgewählten Loci nachgewiesen werden. Das Anzielen der repetitiven Ziele AluY und L1 wurde ebenfalls mit dieser Mutante wiederholt, allerdings konnte die Analyse der Daten nicht rechtzeitig für diese Arbeit fertiggestellt werden.

Die Konstrukte dCas9-SB100X und dCas9-SB(C42) wurden weiterhin mit einer alternativen Methode getestet, in der das Transposon 36 h nach den Fusionstransposasen zugegeben wird. Dies wurde für mehrere single-copy-Ziele getestet, doch es konnten keine gezielten Integrationen konnten werden.

In einem Nebenprojekt wurde versucht, SB-Integrationen in ribosomale DNA zu steuern, indem NoLS-Sequenzen mit der Transposase fusioniert wurden, um Fusionsproteine in Nukleloli zu lokalisieren. Während mehrere kurze Peptidsequenzen diesen Effekt nicht erzeugen konnten, wurde die gewünschte Lokalisation durch Fusion mit dem Protein B23 erreicht. Integrationen mit der Fusionstransposase B23-SB100X waren klar in Nukleolusorganisatorregionen (NORs) und Nukleolus-assoziierten Domänen (NADs) angereichert und assoziiert mit Sequenzen, die auch vom Pol-I-Transkriptionsfaktor UBTF gebunden werden und ebenfalls wahrscheinlich nukleolare DNA beinhalten.

In einem zweiten Nebenprojekt wurde versucht, das NHEJ/HDR-Verhältnis an durch Cas9 verursachten DSBs zu verbessen. Hierfür wurden die SB-Subdomänen N57 und N123 mit Cas9 fusioniert und die N57/N123-Bindesequenzen in die HDR-Donormoleküle eingebaut, um die Donormoleküle in die Nähe der DSBs zu bringen. Es konnte allerdings kein spezifisch erhöhtes HDR/NHEJ-Verhältnis erzielt werden. Das Fusionskonstrukt Cas9-N57 hatte zwar ein erhöhtes HDR/NHEJ-Verhältnis, dieses war jedoch nicht abhängig von der Anwesenheit einer entsprechenden Bindestelle im HDR-Donormolekül und scheint daher durch einen unspezifischen Effekt des Fusionsproteins entstanden zu sein.

11 Supplementary data

A

Target	Construct	# of insertions
AluY	dCas9-SB100X	4477
	dCas9-N57 + SB100X	12799
L1	dCas9-SB100X	39245
	dCas9-N57 + SB100X	31527
	dCas9-N123 + SB100X	30071
	dCas9-N123 + SB10	61572
HPRT	dCas9-SB100X	5612
	dCas9-N57 + SB100X	6648
HS4	dCas9-SB100X	5411
	dCas9-N57 + SB100X	18570
	dCas9-N123 + SB100X	26176
HS8	dCas9-SB100X	184
	dCas9-N57 + SB100X	35630
	dCas9-N123 + SB100X	12083
HS10	dCas9-SB100X	4769
	dCas9-N57 + SB100X	28255
	dCas9-N123 + SB100X	29122
TA1	dCas9-SB100X	1908
	dCas9-N57 + SB100X	7212
	dCas9-N123 + SB100X	6095
TA2	dCas9-SB100X	1574
	dCas9-N57 + SB100X	9860
	dCas9-N123 + SB100X	10539
TA3	dCas9-SB100X	886
	dCas9-N57 + SB100X	1908
	dCas9-N123 + SB100X	9297

B

Insertions around sgAluY-1 target sites

Condition	Insertions [%]			
	200 bp	500 bp	1500 bp	2000 bp
dCas9-SB100X/sgAluY-1	1.35	3.80	9.39	15.50
dCas9-SB100X/sgL1-1	1.23	3.21	9.53	15.99
dCas9-N57/sgAluY-1 + SB100X	1.34	3.87	10.15	16.10
dCas9-N57/sgL1-1 + SB100X	1.16	3.81	10.13	15.43

Insertions around sgL1-1 target sites (up to 1 mismatch)

Condition	Insertions [%]			
	200 bp	500 bp	1500 bp	2000 bp
dCas9-SB100X/sgL1-1	0.02	0.19	0.55	0.74
dCas9-SB100X/sgAluY-1	0.00	0.16	0.49	0.76
dCas9-N57/sgL1-1 + SB100X	0.05	0.22	0.87	1.35
dCas9-N57/sgAluY-1 + SB100X	0.04	0.24	1.06	1.62
dCas9-N123/sgL1-1 + SB100X	0.01	0.12	0.27	0.40
dCas9-N123/sgL1-1 + SB10	0.04	0.37	1.01	1.37
dCas9-N123/sgHS8 + SB100X	0.10	1.08	2.32	3.21

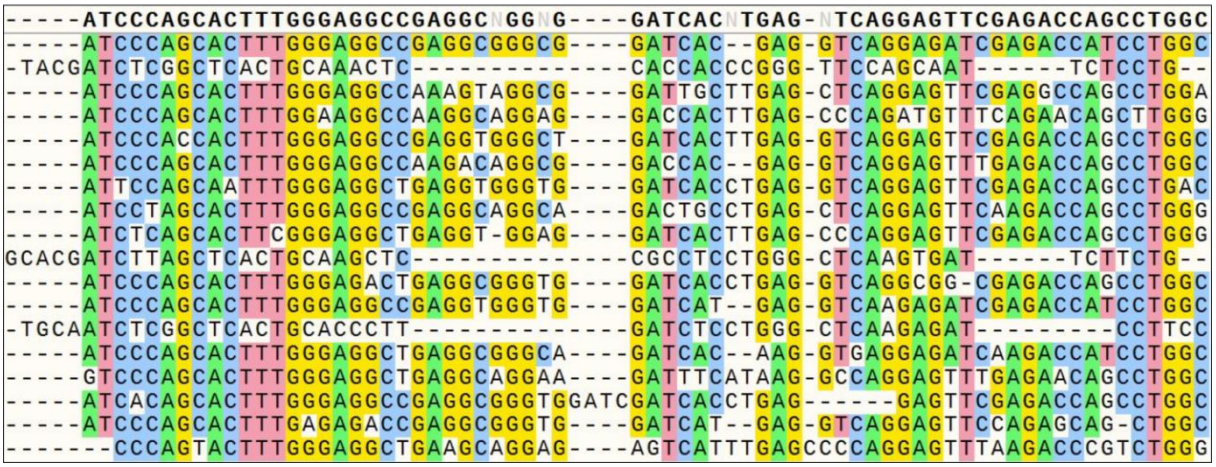
Supplementary Figure 1 – Insertion library data. A Total number of insertions in each dCas9-targeted insertion library. **B** Percentages of insertion into targeting windows around sgAluY-1 (top) and sgL1-1 (bottom) target sites.

A

	Forward	Reverse	N	L	R	Σ
1	left → read end	right → read end		1	1	2
2	no → right	same as fwd	1		1	2
3	left → read end	right → read end		1	1	2
4	left → read end	right? → read end		1	1	2
5	right → left	same as fwd		1	1	2
6	left → right	same as fwd		1	1	2
7	right → read end	right → read end			2	2
8	no seq	left? → right		1	1	2
9	left → right	same as fwd		1	1	2
10	left → left	same as fwd		2		2
11	left → read end	right → read end		1	1	2
12	left → right	same as fwd		1	1	2
13	right → read end	right → read end			2	2
14	right → read end	right → read end			2	2
15	left → left	same as fwd			2	2
16	right → left	no seq		1	1	2
		Σ	1	12	19	32

B

C



Supplementary Figure 2 – Alignment of DNA ends after sgAluY-1-mediated digestion. A Table containing all recovered plasmid-genome junctions. Each row represents a single picked plasmid, the columns represent junctions recovered by sequencing with the forward or reverse primer. For plasmids with large inserts (e.g. row 1) the sequencing read didn't cover the entire insert and one junction could be recovered with each sequencing reaction. For plasmids with smaller inserts (e.g. row 2), both sequencing reads covered the entire insert and both recovered the same two junctions. In these cases, the read generated with the reverse primer is disregarded. Grayed-oud reads failed to produce readable sequences. N = no match to AluY, L = match to AluY left end, R = match to AluY right end (in relation to the cleavage site). **B** Alignment of all reads that matched the AluY left end. Alignment was performed using SnapGene and the MUSCLE algorithm with a consensus threshold of >50%. Top row represents a consensus sequence generated from all other sequences, second row is the canonical AluY consensus sequence. **C** Alignment of all reads that matched the AluY right end, all parameters are the same as in B.

	Chr.	Targets	Closest gene 5'	Distance 5'	Closest gene 3'	Distance 3'
TA1	21	7(8)	claudin-17	44 kb	claudin-8	5 kb
TA2	21	1	keratin-associated proteins 7-1	40 kb	keratin-associated proteins 11-1	12 kb
TA3	7	2	In intron of exocyst complex component 4 gene			

>NC_000021.9 30210151 30210666
tatattatacacatattaatcacagagaatatatatatatatataGTCTCCCTAATTCTAATTCAAGGtaca
tatatatatatatatatatatatataGTCTCCCTAATTGTAATTCAAGGtacgtatgtgtatatatatat
atatatatatatatatatatatatatatatatatatatataGTCTCCCTAATTCTAATTCAAGGtacatata
tatatatatatatatatatatatataGTCTCCCTAATTCTAATTCAAGGtacatatatatatatatat
atatatatataGTCTCCCTAATTCTAATTCAAGGtacatatatatatatatatatatatatataGTCTCC
CTAATTCTAATTCAAGGtacatatatatatatatatataGTCTCCCTAATTCTAATTCAAGGtatatata
tatatatatatatatatatatataGTCTCCCTAATTCTAATTCAAGGtaCTtaATGTCAATGAGAtaata
taagatatgtgtgtgtatttatatat

>NC_000021.9 30869071 30869427
tatatatatatcccaaatatatatatttattatgcatatatatatgtacatatacacacacacatatata
tttgaatatatatacatatatgtgtatatatatatatatatatatatatatatttgaatcaGAGCAGAATTT
GATTAGCCAGCCTGTTTGACACTTTATATCACAGAATGTAGACGATAAAATAATGTAATGATCCTTCAGG
ATTATGCTGCATTGGCTCTTGCACAAGTCACATGGATATTTCCATCTTTCTtctgtatatatccatatat
acagggtctccctctgtcacttaggctgcaggatgtgtgtgtgtgtgtatatatatatatatatatatat
atatata

>NC_000007.14 133857097 133857366
tatatatatatatatgtgtatatatatatgtatatatatatacacatacacgtatatatatatatatata
tatatatatatatatatatatatatatatatatatttttaCCTCTACTTAACCTCCCTGGATGATTGTCTt
agtgtcacatatatatatatatatatatatatatatatacacgtatatatatatatatatatacacgtatat
atatatatatatatatacacgtatatatatatatatatatatatatatatatatatatatatattttaCCTCTA
CTTAACCTCCCTGGATGATTGTCTTAGTGTCACAGGATCCTTCGGGTGTCACTTCACCAGCCAGAACCCT
CTGTGGCTGGCGGCGCCTCTGCTTGAGTTTCACTCATGCCCGTTGGGCTTGTTCTGTCCC

Supplementary figure 3 - Detailed description of TA_n target sites. Information about location of the target sites in relation to adjacent genes and number of target sites can be found in the table at the top. The TA1 target site is present in 7 perfectly matched copies and a single copy with one mismatch. In the sequences below, target sites (including PAMs) are marked in yellow, repetitive TA_n DNA is in lowercase and non-TA_n DNA is in uppercase.

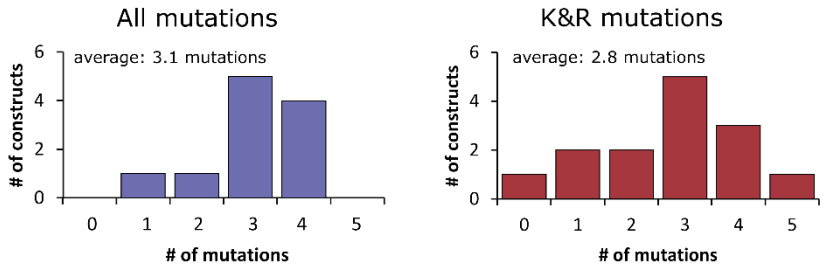
A

	K120	Q124	H127	K128	K129	R131	K156	R166	K186	H187	K208	N245	P247	K256	K259	Q271	R293	N296	K339	#	Additional mutations
1									x							x				2	
2																		x		1	
3		x												x			x			3	
4			x												x			x		3	
5				x								x								3	A255S
6			x							x					x			x		4	
7	x		x					x		x										4	
8					x								x					x		4	G59
9	x															x	x			4	R31S
10		x												x				x		3	
11		x		x														x		3	

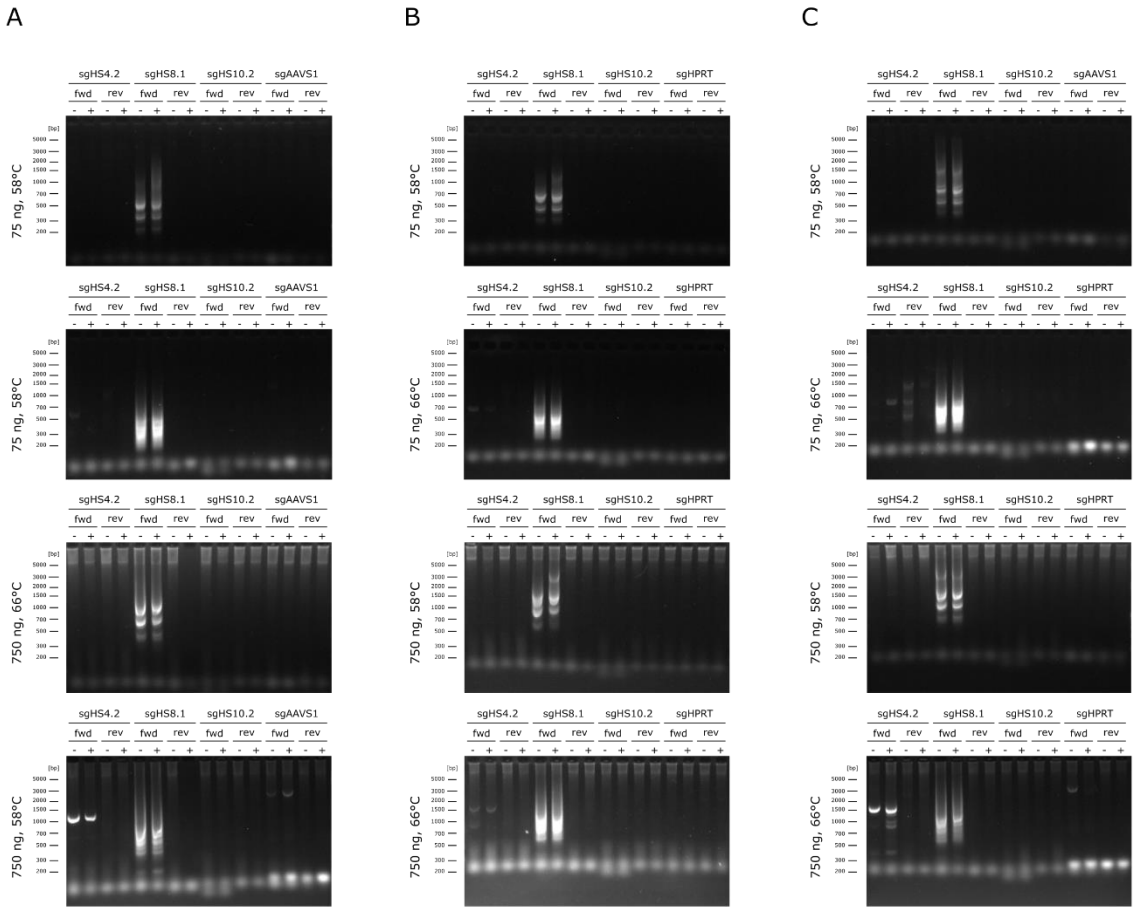
B

	K120	K128	K129	R131	K156	R166	K186	K208	K256	K259	R293	N245	K339	#	Additional mutations
1			x											2	Q9P
2	x		x					x		x				4	
3						x		x	x					3	
4				x		x		x			x			4	
5						x		x						2	
6							x	x			x			4	K287R
7														0	
8										x				1	
9			x		x				x					4	G50V
10				x			x	x			x			5	D214V
11						x			x					3	L72F
12						x		x	x					3	
13							x							1	A200T
14			x				x			x				3	
15				x		x		x						3	

C



Supplementary Figure 4 – Analysis of mutant libraries. A Mutations found in 11 sequences retrieved from the mutant library containing all 19 single amino acid replacements. **B** Mutations found in 15 sequences retrieved from the mutant library containing only lysine and arginine replacements. **C** Distribution and average of numbers of mutations in both mutant libraries.



Supplementary Figure 5 - Agarose electrophoresis of PCR-based insertion screening. A PCRs of gDNA containing insertions catalyzed with dCas9-SB(C42). Target sites, as defined by the sgRNA are indicated at the top. Which genomic primer (fwd or rev) is used is indicated directly below, the other primer is always SB20hmr. The individual primers for each target are listed in Table 2-12. Each primer pair is used on untargeted (-) and targeted (+) samples. The (-) sample always consists of the same targeting construct as the (+) sample, without a sgRNA. PCR conditions (template DNA amount and annealing temperature) are indicated to the left of each image. Note that strong unspecific amplification occurs with the primer HS8.1_fwd. **B** PCRs of gDNA containing insertions catalyzed by dCas9-SB100X, with the transposon delivered 36 h after the transposase. **C** PCRs of gDNA containing insertions catalyzed by dCas9-SB(C42), with the transposon delivered 36 h after the transposase.

A

		NOR insertions	Total Insertions	NORs
SV40	SB100X	68	750604	1.60E-04
	B23-SB100X	68	424511	2.62E-04
HENA	SB100X	81	259054	1.08E-04
	B23-SB100X	122	817286	1.49E-04

B

		NOR13	NOR14	NOR15	NOR21	NOR22
SV40	SB100X	5.88	0.00	0.00	2.94	91.18
	B23-SB100X	2.94	1.47	5.88	1.47	88.24
HENA	SB100X	7.41	1.23	1.23	9.88	80.25
	B23-SB100X	3.28	1.64	3.28	4.92	86.89

C

		topNAR	detNAR	Insertions
SV40	SB100X	12878	228866	750604
	B23-SB100X	8699	150483	424511
HENA	SB100X	4519	79155	259054
	B23-SB100X	17761	281232	817286

D

		5S rRNA	LSU rRNA	SSU rRNA	Insertions
SV40	SB100X	2.66E-02	6.12E-03	1.41E-03	750604
	B23-SB100X	2.54E-02	6.16E-03	7.71E-04	424511
HENA	SB100X	2.68E-02	5.19E-03	7.99E-04	259054
	B23-SB100X	2.38E-02	6.11E-03	1.22E-03	817286
	Random	2.44E-02	6.58E-03	1.97E-03	151923

E

		H3k04me1	H3k09me3	H3k27ac	H3k27me3	H3k36me3	H3k4me3	H3k79me2	H3k9ac	H4k20me1	UBTF
SV40	SB100X	1.67	1.07	1.62	0.99	1.49	1.63	1.67	1.67	1.53	5.00
	B23-SB100X	1.51	1.10	1.47	1.05	1.32	1.49	1.42	1.48	1.40	8.49
HENA	SB100X	1.75	1.05	1.73	0.97	1.56	1.71	1.79	1.72	1.55	5.56
	B23-SB100X	1.62	1.09	1.61	1.05	1.42	1.61	1.61	1.59	1.49	6.13
	Random	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Supplementary Figure 6 – Data on targeting of nucleolar DNA. A Total number of insertions and fraction of insertions into NORs. **B** Fraction of insertions into individual NOR sequences, as a fraction of total NOR insertions. **C** Numbers of insertions into two a smaller and a larger set of DNA defined as NADs (topNAR and detNAR, respectively). **D** Insertion frequencies into 5S rRNA, LSU rRNA and SSU rRNA. **E** Insertion frequency into significant peaks from ChIP-seq against several chromatin marks as well as UBTF, relative to a random dataset.

12 Publications

- Kovač, A. & Ivics, Z. Specifically integrating vectors for targeted gene delivery: progress and prospects. *Cell Gene Therapy Insights* **3,** 103–123; 10.18609/cgti.2017.013 (2017).[318]
- Kovač, A. *et al.* RNA-guided retargeting of Sleeping Beauty transposition in human cells. *eLife* **9**; 10.7554/eLife.53868 (2020). [482]

13 Acknowledgements

I would like to thank Dr. Zoltán Ivics for giving me the opportunity to perform my PhD work in his lab and for providing counseling and guidance throughout the last four years. Having a project so closely aligned with my area of interest has been a blessing and having a kind and understanding PI made things immeasurably easier.

I would like to thank Dr. Csaba Miskey for his great effort in sequencing and bioinformatic analysis of the data generated in this thesis. His efforts have been essential for both this thesis as well as the publication of my experimental data and this project would have been impossible without him.

I would like to thank all other members of the Ivics lab for providing the pleasant working atmosphere necessary to do four years of demanding work without severely impacting one's mental health.

I would like to thank Dr. Andreas Gogol-Döring and Michael Menzel for additional help with bioinformatic analyses and I would like to thank Dr. Claudio Mussolino and Dr. Ralf Kühn for providing material used in this project.

Finally, I would especially thank Delphina Hennig and Antonin Kovač, for always being tolerant of strange working hours and for allowing me to focus on my work when necessary.

Also, I want to thank them, as well as the rest of my family and friends, for providing the motivation needed to carry on when the obstacles in this project temporarily seemed overwhelming.

14 CV

Personal information	
Name	Adrian Kovač
Date of birth	13.09.1990
Place of birth	Frankfurt am Main, Germany

Education	
2010	Abitur Goethe-Gymnasium, Frankfurt am Main
2011-2014	BSc *Molecular Medicine*, Georg-August-Universität, Göttingen
2014-2016	MSc *Molecular Biology*, Max Planck International Research School, Göttingen
2016-2020	PhD studies Paul-Ehrlich-Institut, Langen Goethe-Universität, Frankfurt am Main