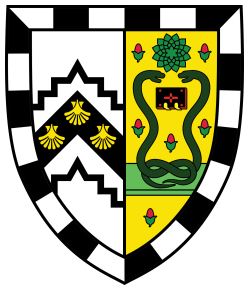
# Cas9-induced on-target genomic damage



**Michał Konrad Kosicki**

Supervisor: Prof. A. Bradley
Dr. S. Teichmann
Dr. M. Hemberg

Advisor: Prof. A. Ferguson-Smith

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
*Doctor of Philosophy*

Gonville & Caius College

September 2018

# Abstract

CRISPR/Cas9 is the gene editing tool of choice in basic research and poised to become one in clinical context. However, current studies on the topic suffer from a number of shortcomings. Mutagenesis is often assessed using bulk methods, which means rare events go undetected, unresolved or are discarded as potential sequencing errors. Many of the genotyping methods rely on short-range PCR, which excludes larger structural variants. Other methods, such as FISH, do not provide basepair resolution, making the genotype assessment imprecise. Furthermore, it is not well understood how Cas9 delivery format influences the dynamics of indel introduction. Finally, many studies of on-target activity were conducted in cancerous cell lines, which do not accurately model the mutagenesis of normal cells in the therapeutic context.

In my thesis, I have investigated on-target lesions induced by Cas9 complexed with single gRNAs and no exogenous template. I have followed the time dynamics of Cas9-induced small indels as a function of reagent delivery methods, established an assay for quantification of Cas9-induced genomic lesions that are not small indels ("complex lesions") and used this assay to isolate and genotype complex lesions, many of which would be missed by standard methods. I found that DNA breaks introduced by single guide RNAs frequently resolved into deletions extending over many kilobases. Furthermore, lesions distal to the cut site and cross-over events were identified. Frequent and extensive DNA damage in mitotically active cells caused by CRISPR/Cas9 editing may have pathogenic consequences.

# Declaration

I declare that his dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Acknowledgements and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation contains fewer than 60,000 words exclusive of tables, footnotes, bibliography, and appendices.

Michał Konrad Kosicki
September 2018

# Acknowledgements

Results presented in chapter 3 were obtained in collaboration with Rajan Sandeep, the first coauthor of the published book chapter. This work was supervised by Manos Metzakopian and Erik Bennett. Results related to culture of progenitor cells from murine bone marrow presented in chapter 5 were the work of Kärt Tomberg, second author on the published Nature Biotechnology article. Allan Bradley was the primary supervisor. Kärt, Allan and Jorge de la Rosa have read the thesis and offered many useful comments.

I would like to blame the members of Bradley group, in particular Mathias Friedrich and Kärt Tomberg, for making the lab a dangerously attractive place to spend my time in. Thanks to Ross Cook, Haydn Prosser, Alex Strong, Katta Hautaviita and Frances Law for their excellent technical support and putting up with millions of questions. I owe a lot to Manos Metzakopian for keeping my radar straight. Shout out to Lilliana Antunes, my 'thesis buddy', and Dimitris Garyfallos for always keeping my spirits up. Cheers to Rajan Sandeep for invigorating lab's social life outside of workplace. Thanks to everyone who participated in foosball matches, a surprisingly welcome distraction.

There is too many people to personally acknowledge at the Wellcome (Trust) Sanger Institute, who made my time here. Thank you to members of Teichmann lab (my second thesis supervisor) for socilizing and all the shady scientific projects we did together (hey, Johan!), to members of Hemberg lab (my third supervisor) and Trynka lab for lunch-time discussions and to all PhD students, in particular PhD14 cohort, for all the good times.

Two great mentors deserve a special mention: Jason Carroll, who inspired me to make the PhD leap and my External Advisor Anne Ferguson-Smith, who provided crucial advice in the time of need.

I am deeply grateful to all great people at Caius for reminding me that life is not all genetics, it is also linguistics, physics, neuroscience, law and port.

Last, but not least, thanks to Allan for being a great supervisor, for giving me a freedom to learn, for patience with my quasi-scientific ramblings and for supporting me even as I ignored his suggestions (mostly to my detriment).

# Table of contents

# Glossary

**BIR** break-induced replication.

**BL6** *Mus musculus*.

**CAST** *Mus musculus castaneus*.

**CRISPR** Clustered Regularly Interspaced Short Palindromic Repeats.

**DDR** DNA damage repair.

**DSB** double-stranded break.

**dsDNA** double-stranded DNA.

**ES** embryonic stem.

**gRNA** guide RNA.

**HR** homologous recombination.

**IDAA** Indel Detection by Amplicon Analysis.

**LOH** loss of heterozygosity.

**MMEJ** microhomology-mediated end-joining.

**NAHR** non-allelic homologous recombination.

**NGS** Next-Generation Sequencing.

**NHEJ** non-homologous end-joining.

**NMD** nonsense-mediated decay.

**PAM** protospacer adjacent motif.

**RNP** ribonucleoprotein.

**SNP** single-nucleotide polymorphism.

**SSA** single-strand annealing.

**ssDNA** single-stranded DNA.

**SSTR** single-strand template repair.

**TIDE** Tracking of Indels by DEcomposition.

# Chapter 1

# Introduction

## 1.1 Precise and efficient modification of DNA

The ability to modify DNA in mammalian cells at a chosen locus precisely and efficiently is highly desirable. In basic research, it allows to unambiguously establish the genetic causality. If introduction of a given DNA modification is accompanied by a change in phenotype, then this modification was sufficient for that phenotype to occur. Such modification has to be precise or else this clear conclusion may be confounded. If the process is not efficient enough, then the phenotype may be difficult to detect or not manifest at all. Precision and efficiency are also paramount in gene therapy. Inefficient DNA modification may fail to achieve the desired benefit. Imprecise modification may have negative consequences, for example excessive cell death, unintended loss of resistance or carcinogenesis. Since genetic modifications are mitotically heritable, even mild side-effects can accumulate over time and have to be avoided.

Many routinely used ways of modifying DNA are neither very precise, nor efficient. To make these methods useful in basic research and biotechnology, efficiency and precision have to be enforced by secondary means, like single cell cloning, breeding, positive and negative selection. Molecular cloning, transgene insertion, homologous recombination and Cre-Lox recombination may serve as examples. For the purpose of clarity, screens based on random mutagenesis will not be discussed here, although similar considerations apply.

**Molecular cloning** is a set of procedures that allow modification of about 3-350 kb DNA molecules in vitro. Typically, the DNA of interest is amplified using PCR or cut out of the donor DNA molecule using restriction enzymes. The resulting fragment is then ligated into a plasmid, a circular piece of DNA with the ability to propagate in bacterial hosts. The plasmid is transformed into bacteria for amplification (Cohen, 2013; Cohen et al., 1972). Restriction enzyme cutting, PCR amplification and ligation steps are usually reasonably precise, but they may not be 100% efficient, leaving behind unligated or uncut plasmids. Bacterial transformation is rarely 100% efficient either. Furthermore, as the number and size of fragments increase, the precision drops and incorrectly ligated plasmids are produced.

Without additional interventions, a cloning procedure will lead to the production of a mixture of correctly and incorrectly modified plasmids, with many bacteria harboring no plasmid at all. However, this outcome is routinely avoided by simply including a antibiotic resistance gene in the destination plasmid and removing non-transformed bacteria using that antibiotic. Efficiency can be further increased by placing a "suicide gene" (e.g. ccdB toxin, Bahassi et al., 1999) in the fragment to be replaced, which prevents undigested or religated backbone from being propagated. Finally, individual plasmids isolated by single cell cloning can be tested for precision by PCR, analytic restriction digest and sequencing. Thus, despite inherent inefficiency and imprecision of the method, a pure and correct product can often be obtained.

While very useful for modifying small DNA fragments, molecular cloning cannot be directly applied to genomic DNA (homologous recom-

bination being an exception, which is described in more detail below). The main obstacle is the short binding site of most restriction enzymes ($\leq$8 bp), which means even an average bacterial genome would be cut tens of times, making precise genomic modifications impossible. Furthermore, even though plasmid vectors can be maintained in bacteria and yeast (with proper origins of replication), they can only be expressed transiently in mammalian cells. This makes them impractical in gene therapeutic context, except when the expression only needs to be transient (notably, some solutions to this problem are being developed, e.g. Broll et al., 2010).

Naturally occuring mobile elements ("transposons") and genomically integrating viruses have been engineered to enable stable insertion of DNA of interest ("transgene") into the genome. Such **transgene insertion** is often efficient enough to affect the phenotype without need for selection. It makes possible the study of gene function by overexpression of wild-type or mutant product, genetic marking of cells for lineage tracing studies and therapeutic restoration of gene expression. Specific organs and even cell types can be modified at any time during development, given the availability of specific delivery methods and promoters.

Nevertheless, in most cases genomic integration is semi-random, which fails the "precision" criterion. Thus, no locus-specific editing is possible. Adeno-associated viruses are an exception, as they integrate at a defined genomic region. However, they are severely limited by the amount of exogenous DNA of interest they can carry (their "cargo capacity", Weitzman et al., 1994). Activation of oncogenes by viral elements posed a significant risk in the past, although newer generations of vectors reduced it by removing promiscuous promoter elements and adding insulators (Aiuti et al., 2013; Hacein-Bey-Abina et al., 2003; Schröder et al., 2002). Immune response to the viral capsid and silencing of viral repeat elements are also a concern (Chira et al., 2015). Finally, transgene insertion often cannot be used when the gene of interest needs to be under fine control from its local chromatin environment or when the pathogenic mutation is dominant negative (i.e. when it actively competes with the wild-type product).

Despite all these problems, the only three FDA-approved gene therapies are based on stable genomic integration of viral constructs. In two of these therapies, the virus deliveres a receptor (anti-CD19) to patient's T cells, which makes them attack B-cell lymphomas. In the third case, virus is used to directly deliver a missing gene (*RPE65*) into the retina, which prevents progressive vision loss in patients with Leber's congenital amaurosis. Many more therapies based on transgene insertion are under development.

Precise replacement or deletion of genomic DNA can be achieved by transfecting the cells with a linear double-stranded DNA (dsDNA) of interest flanked by long sequences identical ("homologous", in this context) to the target region (Smithies et al., 1985). This **homologous recombination** or "targeting" approach leads to precise, but inefficient target modification (1 in $10^5$-$10^8$ transfected cells). Furthermore, the rate of random insertion can be about 1000x higher than that of on-target editing leading to a risk of confounding off-target mutagenesis (Smithies et al., 1985; Thomas and Capecchi, 1987). Selection for correct insertion and against off-target mutagenesis made the process feasible by substantially enriching for the desired modification (5-80% correctly targeted cells among selected ones, Mansour et al., 1988; Yagi et al., 1993). The selection cassettes introduced into the genome during targeting may need to be removed in an additional step, e.g. using PiggyBac transposition, if "scarless" editing is desired (Lee et al., 2014; Yusa et al., 2011a). Because of these issues, targeting is only routinely applied to engineer embryonic stem (ES) cells, which can be single cell cloned and individually screened for correct insertion by PCR. Off-target insertions can be detected by Southern blotting or copy-number qPCR assays.

Since edited ES cells injected into a blastocyst can contribute to the germline, introduced mutations can be studied on an organismal level

(Bradley et al., 1984; Koller et al., 1989; Thompson et al., 1989). Animals obtained this way can be bred to homozygosity, yielding a congenic line with defined DNA modifications. While laborious, homologous recombination has been successfully employed to study the whole-organism phenotypes of many thousands of DNA modifications, among others through IMPC project (Austin et al., 2004). However, it is far too inefficient to create them directly in vivo, which is crucial when studying effects that are specific to a given tissue or developmental stage (notably, in vivo selection methods are being developed e.g. Nygaard et al., 2016).

Superior control over time and place of DNA modification can be achieved through the **Cre-lox recombination** system (or FLP-FRT; Broach, 1982; Golic and Lindquist, 1989; Schaft et al., 2001; Sternberg and Hamilton, 1981). The process uses Cre, a phage enzyme, which can be expressed in an inducible and tissue specific manner, to cause exchange of genetic material ("recombination") between specific DNA sequences called lox sites. Combining different positioning, orientation and sequence variants of these sites allows genomic inversion, deletion, translocation as well as insertion of exogenous DNA. Recombination is usually very efficient and precise. Some recombination lesions can even be engineered to be reversible, for example by using double-invertible splice acceptor constructs containing both lox and FRT sites (Andersson-Rolf et al., 2017; Elling et al., 2017). For all these reasons, Cre-lox system continues to contribute substantially to our understanding of basic biology, among others in mice models (Skarnes et al., 2011). However, recombination leaves behind a genomic scar, which may be a confounding factor in some experiments. This also makes it impossible to introduce single-nucleotide polymorphisms (SNPs) and indels in coding regions (unless a combined Cre-lox and PiggyBac strategy is employed, e.g. Lee et al., 2014). Finally, the lox sites need to be introduced by homologous recombination, which creates a substantial bottleneck in the procedure. Therefore, similarly to homologous recombination, Cre-lox cannot be directly applied to adult organisms, which precludes its use as a gene therapeutic tool.

Discovery of **precision nucleases** (e.g. HO, I-SceI, Zinc Finger and TAL Effector Nucleases) enabled targeted modification of genomes with precision and efficiency far higher than those offered by molecular cloning, transgene insertion, homologous recombination or Cre-lox recombination. While not completely replacing these methods, they complement some of them and open up new possibilities. In particular, they enable precise, genomic, on-target mutagenesis and vastly improve targeted homologous recombination efficiency. Their primary means of action is similar to restriction enzymes in that they introduce a double-stranded break (DSB) at their recognition site. In contrast to restriction enzymes, the binding site of precision nucleases is long enough (typically >15 bp), to enable precise genomic cutting at most loci. Understanding how the cell reacts to and repairs the nuclease induced DSB is crucial. The next section details how naturally occurring DSBs are resolved by cellular repair mechanisms.

## 1.2 DSB repair

DSBs are biologically important in many context, for example as a part of a systematic processes like V(D)J recombination, class switch recombination (both crucial to adaptive immunity), meiosis or transposition of mobile elements. Under these conditions, DSBs usually result in a well defined, localized mutagenic outcome. However, they are highly cytotoxic when induced outside of this context. Ionizing radiation, redox metabolism, nucleotide excision repair and replication fork collapse are some of the events, which cause pathogenic DSBs. Notably, so do precision nucleases. Mammalian cells have evolved a variety of ways to process DSBs, ranging from perfect repair to induction of programmed cell death. Failure to repair any DSB can prevent replication and correct assortment of the DNA. This could lead to activation of oncogenes or inactivation of tumor suppressors and thus cancer. Furthermore, inactivation of an essential gene would cause cell death.

### 1.2.1 Repair pathways

There is currently good genetic and functional evidence for at least four major DSB repair pathways (Fig. 1.1): non-homologous end-joining (NHEJ), microhomology-mediated end-joining (MMEJ), single-strand annealing (SSA) and homologous recombination (HR). The degree of end resection is the major mechanistic factor which determines the repair pathway usage. NHEJ (also known as classical-NHEJ) mediates direct rejoining of broken ends with no or little end-processing, resulting in either perfect repair or small indels <10 bp in vitro (Chang et al., 2017). MMEJ (also known as alternative NHEJ) rejoins mildly resected ends, often using microhomology of 1-16 bp, and is associated with inserts >10 bp and deletions >10 bp (Sfeir and Symington, 2015). SSA requires more extensive homology of >20 bp and always results in clean deletion between the homologous regions (Lin and Sternberg, 1984). HR involves long resection and strand invasion of the resected end into a double stranded template (usually the sister chromatid), which is guided by homology >50 bp, and results in near-perfect copying of genetic information (Jasin and Rothstein, 2013).

**NHEJ** is the default repair mechanism outside of replication, when extensive DSB resection is effectively blocked (Aylon et al., 2004; Escribano-Díaz et al., 2013; Ira et al., 2004). In NHEJ, exposed ends of the break are protected and brought together by Ku protein complex. If the ends are not cohesive, they can be resected in a limited fashion (<10 bp) as well as extended with templated and non-templated nucleotides (Chang et al., 2016). Cohesive ends are joined together by a ligase IV complex, even across a 1 bp gap.

**MMEJ** was originally discovered as the "salvage" pathway active in Ku knock-out cells (Boulton and Jackson, 1996), which requires limited resection for its activity. It also repairs mitochondrial DNA (which lack ligase IV crucial for NHEJ) and complex DSBs, such as those induced by ionizing radiation (Seol et al., 2018; Tadi et al., 2016). Removal of the protective Ku complex and limited resection of about 100 nt by the MRN (Mre11-Rad50-Nbs1) complex enables MMEJ and prevents NHEJ. Non-proofing polymerase Polϑ is central to MMEJ. Its main function is to add nucleotides to the ends of the break in three ways: non-templated, templated from the other end (in trans, resulting in duplications) or templated from the same end (in cis, resulting in inversions). Furthermore, Polϑ actively removes the single-strand binding protein RPA. This enables annealing of the small homologies between the ends of the break, whether natural or created by Polϑ action (Kent et al., 2016; Mateos-Gomez et al., 2017). At this stage, any non-matching terminal nucleotides ("flaps") are removed (Sharma et al., 2015), missing nucleotides are filled-in and the ends are ligated.

If binding of RPA to single-stranded DNA (ssDNA) prevails over Polϑ activity, the cell may instead proceed with the end resection (by Blm/Dna2/Exo1 complex), which enables SSA and HR. **SSA** is similar to MMEJ, as it involves annealing of homologies, flap removal, gap fill-in and ligation. However, homologies are longer (>20 bp) and no nucleotide addition is involved. Therefore, this pathway always results in a simple deletion.

**HR** is initiated by replacement of RPA by another ssDNA binding protein, Rad51 (Jensen et al., 2010; Taylor and Woodcock, 2015). This process also prevents SSA repair. The resected, Rad51-coated end invades into the dsDNA of the unbroken sister chromatid. It can progress through either synthesis-dependent strand-annealing (SDSA) or double-strand break repair (DSBR). In SDSA, the invading strand is extended by DNA copied from the sister chromatid and recaptured by the other side of the break (Nassif et al., 1994). SDSA always results in non-crossover (NCO), since both ends of the break remain on the same chromosome molecule. In DSBR, a so-called double Holliday junction (dHJ) is formed by both DSB ends of the break being captured in a tangled way with the invaded sister chromatid (Szostak et al., 1983). Depending on how this structure is resolved, DSBR can result in either NCO or crossover (CO).

SDSA appears to be the predominant HR pathway in mitosis, consistent with its exclusively non-crossover outcomes (Andersen and Sekelsky, 2010). Similarly, HR is limited to post-replicative cells, when a sister chromatid is present. Without sister chromatid, HR would have to use the homolog as the template, which would likely result in loss of heterozygosity (LOH) and thus loss of genetic information. Notably, even though HR usually results in perfect repair of the damaged locus and thus is a preferred pathway when a template is present, it is >1000 times more mutagenic than regular DNA replication due to lower fidelity of the involved polymerases (Deem et al., 2011; Hicks et al., 2010).

**Break-induced replication (BIR)** can serve as a backup to the other HR pathways, especially in collapsed replication forks, during telomere extension and in any other case, where the second end of the DSB is difficult to capture. The main feature of BIR is conservative replication of DNA from the site of the break till the end of the chromosome, primed by the invading ssDNA. BIR works even in non-replicative cells and requires Polα and a specialized Polδ polymerases (Sotiriou et al., 2016). Mechanistically, BIR can be placed at a similar level as SSA, since it requires RPA-coated ssDNA, but is inhibited by excessive end resection. However, BIR can also utilize Rad51, unlike SSA (Marrero and Symington, 2010; Ruff et al., 2016).

Rad51-independent single-strand template repair (SSTR) is a pathway that may have evolved to enable RNA-templated DNA repair. It has recently gained prominence as the mechanism for ssDNA templated genome editing (Gallagher and Haber, 2018). SSTR has been postulated to use proteins from Fanconi Anemia pathway, which is involved in repair of interstrand crosslinks (Richardson et al., 2018).

### 1.2.2 Cell-cycle arrest, apoptosis and controlled DSB induction

Even in a simple, unicellular organism such as yeast, a single DSB in a non-essential locus can trigger cell death (Bennett et al., 1993). In human cells, one unrepaired DSB can cause G1 arrest and 10-20 DSBs are enough for a G2 arrest (Deckbar et al., 2007; Huang et al., 1996). Excessive damage can result in activation of apoptotic pathways in a p53-dependent or independent manner (Blackford and Jackson, 2017; Ciccia and Elledge, 2010; Roos and Kaina, 2006).

Despite their high mutagenic and carcinogenic potential, DSBs are induced in many physiological processes. Separation of entangled daughter strands during replication, generation of immune diversity, meiotic recombination and transposition of mobile elements rely on them. These processes involve various specialized enzymes (for example topoisomerase II, Spo11, RAG1/2, PiggyBac transposase) that both catalyze the DSB and modulate the repair outcome. Whereas restriction nucleases leave a free terminal phosphate that can be easily religated by NHEJ, the mechanisms mentioned above often proceed through either a hairpin stage (V(D)J recombination and PiggyBac transposition, Mitra et al., 2008; van Gent et al., 1996) or a covalent linkage between DNA and the enzyme (meiotic recombination, Cre-lox recombination and disengagement of replicated strands by topoisomerases; Goto and Wang, 1982; Keeney and Kleckner, 1995). It is likely that these conditions reduce the oncogenic potential of induced lesions compared to spontaneous ones. While such mutations do occur, for example the translocation between *IgH* and *Myc* loci leading to Burkitt's lymphoma, they do so rarely (Alt et al., 2013).

### 1.2.3 Diversity in cellular DNA repair

While some pathway decision points are well-described (e.g. resection, strand invasion), a general, quantitative model for DNA repair is missing. In particular, differences in how cells utilize different repair pathways lack good explanation. For example, little is known about neural DSB repair. Neurons are post-replicative, which means that they do not suffer from replication-induced DSB and are at a lower risk of cancerous transformation. At the same time, they also cannot use sister chromatid to repair other spontaneous DSBs. Since they are largely irreplaceable due to limited adult

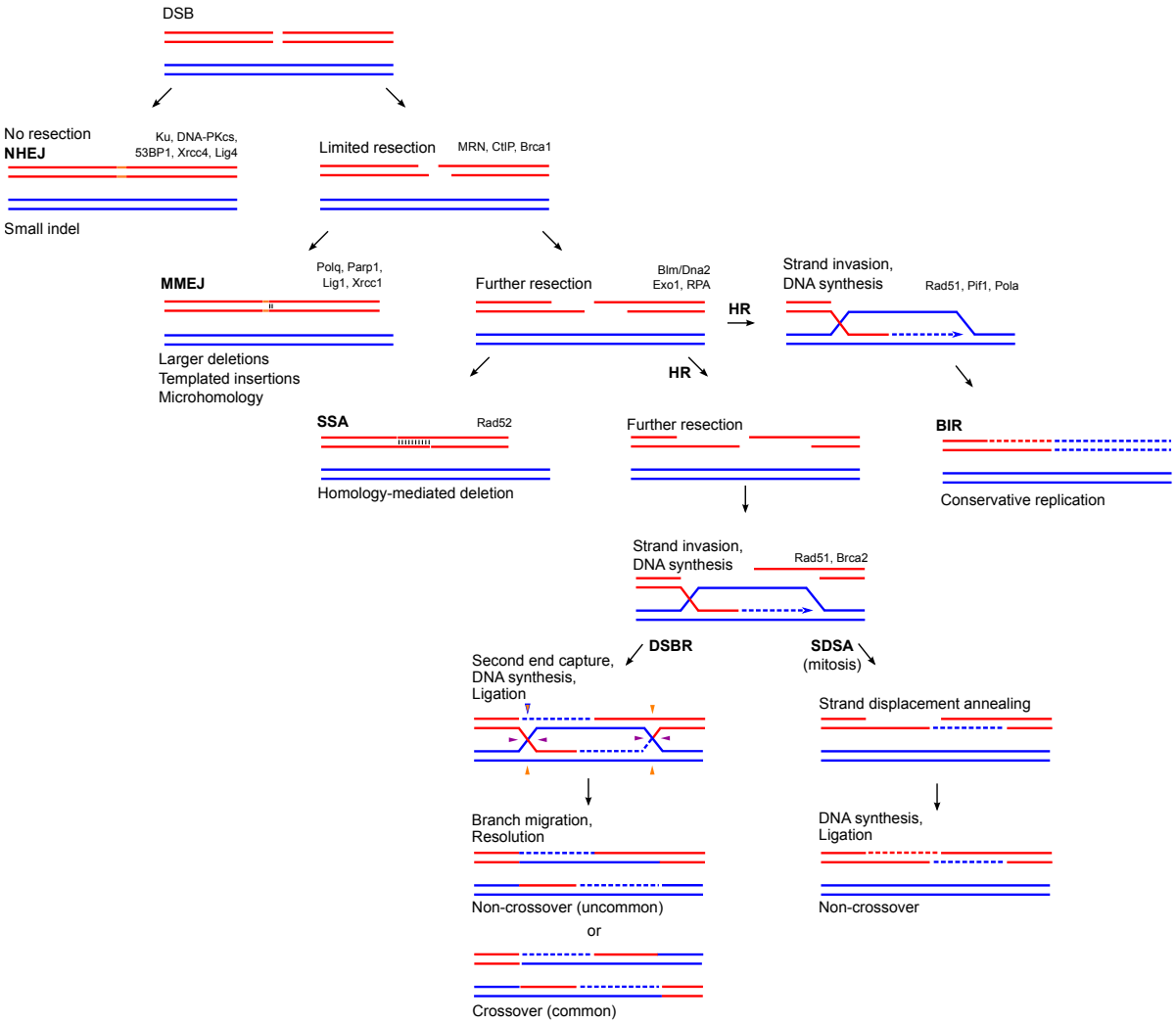



**Figure 1.1:** Pathways of DSB repair. Modified from Sung and Klein, 2006.

neurogenesis, they might be less likely to undergo apoptosis due to DNA damage. Collectively, these properties may explain why large structural variants are often found in mature neurons (Cai et al., 2014).

On the other end of the cellular spectrum, similar structural mutations and aneuploidies are seen in early embryos from IVF procedures (Voet et al., 2011). Consistently, mouse ES cells use less NHEJ and more mutagenic MMEJ and HR than the more differentiated mouse embryonic fibroblasts (MEF). ES cells also exhibit hallmarks of chronically unrepaired DNA damage, lack G1 checkpoint and only undergo apoptosis in a p53-independent manner (Ahuja et al., 2016; Aladjem et al., 1998; Hong and Stambrook, 2004; Tichy et al., 2010). It is currently not clear why mutagenic DNA repair seems to be associated with early embryos and ES cells and why the consequences of these events are rarely seen in adult organisms at similar frequencies, although a potential mechanisms involving immune and cellular elimination of affected cells have been proposed (Bolton et al., 2016; Daughtry et al., 2018; Santaguida et al., 2017). The cell-specific DNA repair may be related to balancing the risk of cancerous mutagenesis, need for timely cell division (for example during development) and broader consequences of cell death.

## 1.3 Precision nucleases

Precision nucleases substantially improved our ability to modify genomes. By generating a single DSB at their binding site, they can cause localized mutagenesis (mediated by NHEJ and MMEJ) and stimulate precise modification of the target using exogenous DNA templates (by HR or SSTR, (Richardson et al., 2018; Rouet et al., 1994)). If the nuclease is expressed constitutively, the reaction will only cease when mutagenesis or templated editing destroys the binding site. Targeted mutagenesis of exons is particularly useful in generating knock-out alleles by introduction of out-of-frame indels. Furthermore, larger deletions, inversions and translocation can also be created by two simultaneously induced DSBs (see subsection 1.3.3).

Some of the early precision nucleases discovered, such as HO, I-SceI and similar "meganucleases" (Plessis et al., 1992; Sugawara and Haber, 2012) could only bind one pre-defined sequence, which could not be easily modified by protein engineering (although some examples exist: Chevalier et al., 2002; Rosen et al., 2006; Seligman et al., 2002; Sussman et al., 2004). Their binding site would therefore often have to be introduced into the genome by traditional, low-efficiency homologous recombination approaches.

Programmable precision nucleases solved that problem by combining FokI nuclease with Zinc Finger proteins or TAL Effector domains, which can be engineered to bind specific DNA sequences. Since FokI introduces only a single stranded DNA break, two ZFNs (Zinc Finger Nucleases) or TALENs (TAL Effector Nucleases) need to bind in close proximity on opposite dsDNA strands to cause a DSB (Bibikova et al., 2003; Boch et al., 2009; Moscou and Bogdanove, 2009; Urnov et al., 2005). The ability to induce localized DSBs has allowed a more detailed dissection of the mechanisms involved in DSB repair (Mehta and Haber, 2014). Clinical trials using these tools to treat genetic diseases as well as to improve immune response to cancer or HIV by modifying T cells are under way (clinicaltrials.gov: NCT01044654, NCT02500849). Direct mutagenesis of integrated HPV virus using TALENs is also explored (clinicaltrials.gov: NCT03057912). A long, successful "track-record" of both ZFN and TALENs, and the unparalleled binding flexibility of new generation TALENs (which can be programmed to specifically bind sequences up to 30 bp with no composition constraints) make them tools of choice for many potential clinical applications. However, the complexity of design, which prevents many researchers from directly assembling their own nucleases and which drives up the cost of commercial solutions, have prevented their wide-spread use in basic science. This gap was largely filled by the discovery and development of a simpler, cheaper and more flexible CRISPR/Cas9 system.

### 1.3.1 CRISPR – biology and applications

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) are genomic DNA arrays found in most prokaryotes, which consist of repeat sequences interspersed with fragments of viruses "recorded" during viral invasion. Together with various Cas (CRISPR-associated system) proteins it acts as a prokaryotic immune system. Recorded viral fragments are used to direct Cas nucleases to an invading virus, causing its destruction. A fixed DNA sequence called protospacer adjacent motif (PAM), which is recognized by the nuclease, needs to be present next to the target site. This prevents the nuclease from digesting the host DNA, since PAM is not found in the repeat sequences of the genomic CRISPR array. Different classes of CRISPR system exist, many of which remain to be investigated (Wright et al., 2016).

The CRISPR/Cas9 system from *Streptococcus pyogenes* (SpCas9) was the first to be reprogrammed by the researchers to cut chosen sequences in vitro in plasmids and in human cell lines (Cong et al., 2013; Gasiunas et al., 2012; Jinek et al., 2012; Mali et al., 2013). In its natural form, it consists of the Cas9 nuclease loaded with two RNAs: a crRNA (CRISPR RNA) processed from the CRISPR array (which contains the sequence complementary to the target site and part

of the repeat sequence) and a universal trRNA (trans-activating crRNA), which mediates the interaction between the Cas9 protein and the crRNA. In biotechnological practice, the two RNAs are fused into a single guide RNA (gRNA) composed of 20 nt sequence complementary to the target site and a 76 nt scaffold. When introduced into cells, the gRNA-loaded nuclease finds the dsDNA target and cleaves both strands. The PAM requirement for SpCas9 is a simple 3' NGG sequence (Fig. 1.2). Unlike most transcription factors and many other Cas9 nucleases, SpCas9 can bind to and open heterochromatic regions, which broadens its targeting range (Barkal et al., 2016; Polstein et al., 2015). The modularity, simple targeting rule and wide genomic range have made SpCas9 the precision nuclease of choice, largely replacing ZFNs and TALENs in regular laboratory use. It is also the only CRISPR system so far to enter into clinical trials (e.g. clinicaltrials.gov: NCT03164135, NCT03166878, NCT03044743).

The targeting range of Cas9 is limited by the PAM requirement. Since Cas9-induced DSB only improves the efficiency of repair using exogenous DNA within 10 bp radius of the cut site (Paquet et al., 2016), the strict PAM requirement severely limits the number of sites that can be edited. This problem can be circumvented to some degree by using a CRISPR nuclease with a different PAM requirement, such as Cas12a (former Cpf1), C2c1 or Cas9 from other species (Yang et al., 2016b; Zetsche et al., 2015). Engineered Cas9 and Cas12a variants with altered PAM specificities are also available (Gao et al., 2017; Hirano et al., 2016; Kleinstiver et al., 2015). Notably, Cas9 from Neisseria meningitis can cleave ssDNA (but not dsDNA) without PAM limitation (Zhang et al., 2015). In principle, engineering of a Cas protein to cleave dsDNA without a PAM requirement should be feasible. However, such a protein would only work with synthetic gRNAs, as a dsDNA sequence producing the gRNA would be cut. Furthermore, its off-target activity will increase due to a shorter binding region.

Notably, various Cas proteins have been engineered to perform functions other than cleavage of DNA. Cas9 with an inactivating mutation in one of its two nuclease domains turns into a nickase that introduces single-stranded, rather than double-stranded breaks. Nickase coupled to a deaminating enzyme has been used as an efficient "base editor" capable of creating single basepair substitutions (CG to TA, and AT to GC, Gaudelli et al., 2017; Kim et al., 2017a; Komor et al., 2016). While normal activity of base editor Cas9 should suppress base-excision repair (instead proceeding through mismatch repair) and avoid creation of a DSB, indels are still observed at a frequency of 0.1-1%. These are likely caused by mutagenic intermediates of residually active base-excision process, a DSB caused by simultaneous base-excision and nicking or a DSB caused by a replication fork encountering a nick (Simonelli et al., 2005). Other uses of nickase enzymes are described in section 1.3.2. A "deactivated" Cas9 with both nuclease domains inactivated has been coupled to numerous effector domains to act as a "genomic delivery service", mediating among others transcriptional activation, inhibition or chromatin remodeling (Chavez et al., 2016; Gilbert et al., 2014; Kearns et al., 2015; Konermann et al., 2014; Liu et al., 2016; Thakore et al., 2015; Xu et al., 2016).

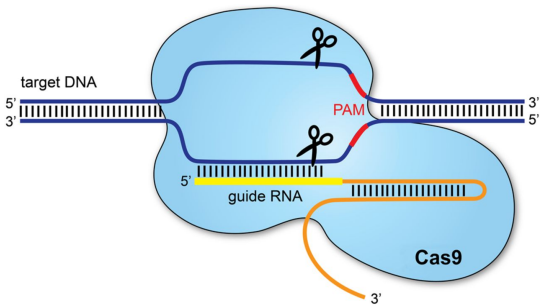



**Figure 1.2:** Schematic of Cas9 DNA cleavage mechanism. From Redman et al., 2016.

### 1.3.2 Cas9 off-target problem

The specificity of precision nucleases is limited by two factors. First, while many 23 bp Cas9 binding sites (including a 3 bp PAM) are unique, many are not due to repetitive nature of the genome. By definition, a site which is not unique is impossible to target specifically. SpCas9 gRNAs with a target-

ing segment longer than 20 nt can mediate binding and cutting, but do not confer increased specificity, possibly because they are trimmed down to 20 nt in vivo (Ran et al., 2013). The binding site of Cas12a is 24 bp long, which is the longest known among Cas enzymes and may underlie its higher specificity (Fonfara et al., 2016; Zetsche et al., 2015). No Cas enzyme has so far been engineered to have a longer binding site.

Second reason for limited specificity is that mismatches between the gRNA and the target DNA sequence do not always prevent activity. Such off-target mutagenesis has been detected in vivo at sequences mismatched at up to six positions (including the PAM sequence), as well as those with 1 bp indels (Akcakaya et al., 2018; Canela et al., 2016; Hsu et al., 2013b; Jiang et al., 2016; Lensing et al., 2016; Tsai et al., 2017, 2015). Frequencies of some of these off-target events are estimated to be around 0.01% and were obtained by either tagging of DSBs in vivo (e.g. GUIDE-seq, Tsai et al., 2015) or by selecting broken DNA upon in vitro Cas9 digestion (e.g. CIRCLE-seq, Tsai et al., 2017). Currently, indels resulting from such putative DSB events cannot be confirmed using direct amplicon sequencing, which has a resolution limit of around 0.1% due to inherent sequencing error rate of the Illumina platform. Systematic genome-wide studies have excluded the possibility that Cas9 may modify completely mismatched targets (Akcakaya et al., 2018; Iyer et al., 2018; Luo et al., 2018b). Notably, while Cas9 binding to the DNA is necessary for DSB induction, it is not sufficient. Therefore, the range of "off-target binding" is likely much larger than that of "off-target mutagenesis" and may potentially have consequences for nuclease-deactivated Cas9 enzymes engineered for their "genomic delivery" function. This may explain recent results questioning the specificity of CRISPR-interference approaches (Stojic et al., 2018).

A number of solutions to the off-target issue have been proposed. In practice, targets mismatched at more than two positions are cleaved very rarely. Therefore, choosing a target that differs on at least two positions from any other target in the genome is usually sufficient to maintain functional specificity. That choice can be improved by algorithms (Elevation, CFD, CCTop and MIT), which score off-targets based on empirical data and the likelihood of undesired modification of coding regions (Doench et al., 2016; Hsu et al., 2013b; Listgarten et al., 2018; Stemmer et al., 2015). In clinical setting, where the patient's genome is not be completely sequenced and where specificity is of paramount importance, empirical methods for detection of off-target mutagenesis may greatly improve gRNA selection prior to treatment. A number of such in vitro and in vivo methods are available (Tsai and Joung, 2016).

Since a modification at an on-target locus is usually more likely than at an off-target mismatched by a few nucleotides, the specificity of mutagenesis can be further increased at the cost of efficiency by reducing the effective concentration of the nuclease-gRNA complex. Shorter gRNAs (17-18nt match) as well as longer, 5' mismatched ones were reported to reduce the frequency of off-target mutagenesis, presumably by decreasing the affinity towards off-targets that are matched at the 5' end. These strategies occasionally came at a cost of creating new off-target sites and lower efficiency (Cho et al., 2013; Fu et al., 2014). A related strategy involves choosing gRNAs that are purposefully mismatched at the intended target site with the hope that further mismatches with off-target sites will increase specificity (Chavez et al., 2018). Furthermore, a number of SpCas9 variants with increased specificity have been engineered (Casini et al., 2018; Chen et al., 2017; Hu et al., 2018; Kleinstiver et al., 2016; Slaymaker et al., 2016), although some of them suffer from reduced efficiency (Chen et al., 2017). Some loss of efficiency has been linked to a 5' mismatch commonly introduced to enable expression of gRNAs from plasmid vectors (Kim et al., 2017b). While this suggests improved fidelity enzymes enforce a match with the target at the 5' end more stringently than wild-type enzymes, more research into the structural nature of these functional improvements is warranted.

Another way to reduce off-target mutagenesis is to use Cas9 nickase (or FokI-coupled deactivated Cas9), which creates single-stranded breaks. Analogously to ZFNs and TALENs, two nickase enzymes directed to two targets in close proximity of each other (10-30 bp) will induce a DSB. Conversely, a single off-target nick would normally be religated with no mutagenic effect (Guilinger et al., 2014; Mali et al., 2013; Ran et al., 2013). This strategy increases the specificity by sacrificing efficiency and targeting range, and by increasing the complexity of the system (as three components are needed). The off-target problem will likely continue to stimulate the development of new tools, detection techniques and computational methods. Notably, the specificity of the Cas9 is limited by the particular genetic and biochemical makeup of the target cell, which cannot always be known accurately (Lessard et al., 2017).

### 1.3.3 Cas9 on-target damage

The DSB induced by Cas9 and its resolution by DNA damage repair (DDR) mechanisms is the principal cause of the mutagenesis and templated editing. However, a Cas9-induced DSB differs from one caused by ionizing radiation or free radicals. In particular, a natural DSB is unlikely to occur *simultaneously* on all homologous sequences (homologs and sister chromatids), while highly active Cas9 may lead to such an outcome. Ionizing radiation often generates two single stranded breaks within 10 bp on opposite strands, which leads to a staggered DSB. Both staggered and blunt ionizing radiation-induced DSBs may also contain blocked ends and damaged nucleotides, which makes them difficult to repair using NHEJ (Mahaney et al., 2009). Conversely, DSBs caused by Cas9 are assumed to be predominantly blunt and clean, which makes them a good substrate for non-mutagenic NHEJ (Jinek et al., 2012). Occasionally, Cas9 induces a DSB with 1 nt 5' overhang, which has been linked to frequent occurrence of 1 bp and larger insertions templated from around the cut site (Lemos et al., 2018). In addition to an endonuclease activity, the nuclease domain which cleaves the strand non-complementary to gRNA may also have exonucleotic activity. This has been demonstrated in vitro, by resolving radioactively labelled dsDNA cleaved and digested by Cas9 over the course of about 10 min (Jinek et al., 2012; Stephenson et al., 2018). However, no in vivo proof has been presented so far. Finally, Cas9 remains bound to the DNA after cleavage (Sternberg et al., 2014). This could modulate the repair outcome by preventing proper assembly of the DSB repair machinery. Indeed, when Cas9 is bound to the transcribed strand of an active gene, its removal by the RNA polymerase activity mitigates the effect on DNA repair (Clarke et al., 2018).

Deletions smaller than 20 bp and insertions of 1-2 bp are the primary outcome of Cas9-induced DSB, when no template is provided. Each gRNA induces particular size indels at specific frequencies. This is often described as the "indel profile" of a given gRNA. These profiles are independent of the broader genomic context and generally stable across tested cell lines (Chakrabarti et al., 2018; Koike-Yusa et al., 2014; Tan et al., 2015; van Overbeek et al., 2016). However, small-molecule inhibition of NHEJ skews the profile towards larger indels, which indicates that differential expression of DNA repair pathways in normal or pathological settings may also influence the outcome of Cas9 cutting (van Overbeek et al., 2016). Other potential modifiers include the format of Cas9 delivery, which ranges from transient transfection of pure Cas9 protein and synthetic gRNAs, also called ribonucleoprotein (RNP), to stable lentiviral transduction of constructs expressing both. For example, RNP results in more rapid mutagenesis, because both components are preassembled and active as they enter the cells. Stable expression is associated with higher off-target rate, because both Cas9 and gRNA are present in the cell for a longer time (Kim et al., 2014; Lin et al., 2014; Liu et al., 2015; Ramakrishna et al., 2014; Zuris et al., 2015). In the presence of a template, both mutagenesis and templated editing can occur. The efficiency of editing is usually lower than that of mutagenesis, but varies widely between cell lines and loci. Efforts to increase it by modulating

DNA repair pathways and by modifying the Cas9 enzyme, gRNA and template itself, are very active areas of research (e.g. Chu et al., 2015; Maruyama et al., 2015; Riesenberg and Maricic, 2018).

Small indels are not the only documented outcomes of precision nuclease mutagenesis. Single gRNAs were shown to induce deletions of up to 600 bp in mouse zygotes (Shin et al., 2017). Deletions of up to 1.5kb in a haploid cancer cell line potentially induced by single gRNAs have been described, but since the guides were directed to a small part of the genome and provided as a pool, the possibility of rare double-cutting events could not be excluded (Gasperini et al., 2017). Although lesions non-contiguous with the cleavage site have been reported in yeast upon I-SceI nuclease cutting, no similar events were reported for Cas9 (Roberts et al., 2012; Sinha et al., 2017; Yang et al., 2008). Studies using paired gRNAs to induce localized deletions also reported generation of more complex genotypes, such as inversions, translocations, endogenous and exogenous DNA insertions and larger-than-expected deletions (Boroviak et al., 2016, 2017; Canver et al., 2014; Kraft et al., 2015; Parikh et al., 2015; Zuckermann et al., 2015). It is possible that even single gRNAs may generate such outcomes, for example due to DSB-proximal spontaneous damage or off-target DSB induction that is concomitant with on-target cutting.

## 1.4 Outstanding issues

Accurate characterization of genotypic and phenotypic consequences of on-target Cas9 mutagenesis is crucial to both basic research and therapeutic applications. However, current studies on the topic suffer from a number of shortcomings. Mutagenesis is often assessed using bulk methods, which means rare events go undetected, unresolved or are discarded as potential sequencing errors. Many of the genotyping methods rely on short-range PCR, which excludes larger structural variants. Other methods, such as FISH, do not provide basepair resolution, making the genotype assessment imprecise. Furthermore, it is not well understood how Cas9 delivery format influences the dynamics of indel introduction. Finally, many studies of on-target activity were conducted in cancerous cell lines, which do not accurately model the mutagenesis of normal cells in the therapeutic context.

In my thesis, I have investigated on-target lesions induced by Cas9 complexed with single gRNAs and no exogenous template. In chapter 3, I have followed the time dynamics of Cas9-induced small indels as a function of reagent delivery methods (published as Kosicki et al., 2017). In chapter 4, I established an assay for quantification of Cas9-induced genomic lesions that are not small indels ("complex lesions"). Finally, in chapter 5 I used this assay to isolate and genotype complex lesions, many of which would be missed by standard genotyping methods (most of the content of the last two chapters was published as Kosicki et al., 2018).

# Chapter 2

# Materials and Methods

## 2.1 Cell lines and cell culture

JM8A3 mouse ES cell line was derived from a C57BL/6N blastocyst (Pettitt et al., 2009). CB9, BC2 and BC8 mouse ES cell lines were derived from F1 cross between C57BL/6N and CAST/EiJ mice (Strogantsev et al., 2015), a gift from A.F. Smith. JBG7 and CBA9 are Cas9-expressing single cell clones derived from JM8 or CB9 cells, respectively. Cas9 was introduced by stable transduction using a Cas9-2A-Blast lentiviral construct (in pKLV2 backbone, see Vectors section) at a low titre to ensure single copy integration (<0.1% transduction rate). Human HEK293 cell line and its subclone expressing Cas9 from the same lentiviral Cas9-2A-Blast construct were single-cell cloned and their karyotype was verified (a gift from E. Metzakopian). hTERT RPE1, *trp*53-/- cell line expressing Cas9 was obtained from Steve Jackson's group. AB2.2 mCherry/GFP reporter cells were a gift from Dr. Xiufei Gao and Prof. Pentao Liu. 293T cells for lentivirus production were obtained from Ao Zhou. Virus was obtained by lipofectamine LTX mediated transfection of 293T cells with ViraPower Lentiviral Packaging Mix (Thermo Scientific) and the Cas9-2A-Blast construct, following manufacturers' instructions.

All ES cell lines were cultured in M15 media (High-Glucose DMEM, with 15% FSC, β-mercaptoethanol and L-Glutamate, Gibco) on sublethally irradiated feeder cells. Feeders were derived from SNL76/6 cell line (expressing neomycin resistance and LIF cassette, Ramírez-Solis et al., 1993) by transgenic insertion of a resistance cassette (blasticidin or puromycin). HEK cells were cultured in M15 and RPE1 cell lines were cultured in M10 (High-Glucose DMEM with 10% FSC).

## 2.2 Vectors

Vectors for expression of gRNAs contained a U6 promoter with a „F+E" scaffold (reported to mediate higher levels of mutagenesis than the standard scaffold, Hsu et al., 2013a) and a Puro-2A-BFP cassette driven by PGK promoter. Constructs were flanked by PiggyBac repeats (PBCV backbone from Mathias Friedrich), lentiviral repeats (pKLV1 backbone from K.Yusa, Koike-Yusa et al., 2014) or both PiggyBac and lentiviral repeat elements (pKLV2 backbone from E. Metzakopian, Metzakopian et al., 2017). Cas9-expression vectors contained a truncated EF1α (EFS) promoter driving a Cas9-2A-Blast cassette in a pKLV2 backbone. Hyperactive PiggyBac transposase was driven by CMV promoter (Yusa et al., 2011b). See vector schematics in Fig. 2.1. Vectors were amplified in NEB10β E.coli strain (Thermo Scientific) under Ampicillin selection and purified using Machelerey-Nagel plasmid extraction kits. gRNAs were cloned into BbsI digested backbones using DNA Ligation Kit V.1 (Takara). Subcloned plasmids were Sanger sequenced at Eurofins or GATC.

## 2.3 Transfections, flow cytometry and sequencing

Transfections took place in 24W plates coated with gelatin. About 300,000 wild-type mouse ES cells were "reverse" transfected with 2.5 µl

lipofectamine LTX, 0.5 µl plus reagent (Thermo Scientific), 200 ng hyperactive PiggyBac transposase (Yusa et al., 2011b), 100 ng of the pKLV2-PiggyBac Cas9-Blast plasmid and 50 ng of the PBCV-gRNA-Puro plasmid in 50 µl OptiMEM following manufacturer's instructions. For Cas9-expressing mouse ES cells, 50 ng hyperactive PiggyBac transposase and 150 ng of the PiggyBac gRNA-Puro plasmid were used. A similar setup was used for lipofection of 20 pmol of hybridized crRNA:trRNA (Sigma) and 20 pmol of EnGen Cas9 NLS (NEB), except plus reagent was omitted. Hybridization was performed by warming up mixed crRNA and trRNA to $95_oC$ and letting it cool down at room temperature. Neon Transfection System (Thermo Fisher Scientific; 1600 v / 10 ms / 3 pulses) was used for electroporation of 150,000 mouse ES cells in buffer R with 6 pmol each of crRNA:trRNA, electroporation enhancer (IDT) and Cas9 protein or 9 pmol each of crRNA:trRNA and Cas9 protein. Cells were cultured in M15 media supplemented with LIF to maintain pluripotence (Williams et al., 1988). Stable integration of Cas9 and gRNA-expressing constructs was selected for using blasticidin (10 µg/ml) and puromycin (3 µg/ml), respectively. The drugs were added on day 2 and cells were maintained in selective media for the duration of the experiment. Cells were split 1:1 or 1:2 on day 5 (depending on confluency) after 10-30' incubation with trypsin, 1:4 on day 7 and 1:6 from then on, any time they were nearing confluency.

For flow cytometric analysis, around 300,000 cells (1/6 of a near-confluent well) were collected by trypsinization, transfered to a U-bottomed 96W plate, washed once and stained for 15-60' in 50 µl buffer with 1 µg/ml FLAER reagent (Cedarlane) or 1:200 anti-Cd9-PE antibody (cat 124805, Biolegend). After staining, cells were washed three times and analysed using a Cytoflex flow cytometer. All procedures were performed at room temperature. PBS+0.1% BSA buffer was used throughout. All centrifugations were performed for 1 min at 500 G.

FACS sorting was performed on day 14 using MoFlow XDP (Beckman Coulter) or SH800 (Sony). Cells were plated at a limiting dilution of 500-2000 cells per 10 cm feeder plate (yielding around 100-400 colonies) in M15 supplied with Penicilin/Streptomycin and colonies were picked 7-10 days later in 96W feeder plates. Genomic DNA was extracted from grown colonies by overnight digestion at 56°C using a lysis buffer supplied with 1 µg/ml proteinase K (100 mM Tris pH 8.5, 5 mM EDTA pH 8.0, 0.2% SDS, 200 mM NaCl) followed by precipitation using 100% ethanol with 75 mM NaCl and three washes with 70% ethanol. DNA was resuspended in 200 µl T0.1E buffer (10 mM Tris-HCl with 0.1 mM EDTA). PCR amplification was performed using LongAMP or Q5 polymerase (NEB) following manufacturer's instructions. The products were resolved on an agarose gel (2% for primer pairs spanning <1.5 kb, 0.8% otherwise) and stained using ethidium bromide. If multiple

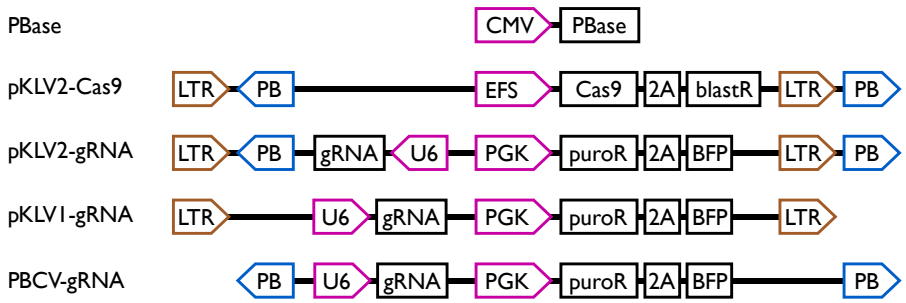



**Figure 2.1:** Vector schematics. LTR = Long Terminal Repeats, lentiviral elements; PB = PiggyBac repeats; blastR and puroR = blasticidin and puromycin resistance cassettes; LTRs, PBs and promoters (EFS, CMV and PGK PolII promoters and U6 PolIII promoter) are maked with colors.

products were present, they were individually cut out of the gel and purified using QIAquick Gel Extraction Kit (Qiagen). If only one product was obtained, the PCR reaction was purified using AMPure XP magnetic beads (1:1 ratio). Products were Sanger sequenced at Eurofins or GATC.

Similar procedures as for mouse ES cells were used with RPE1-Cas9 cell line, with following exceptions. For flow cytometry experiments, cells were transfected with 50 ng pKLV2-gRNA-Puro and 200 ng hyperactive PiggyBac transposase and selected with puromycin (3 µg/ml) from day 2 till the end of the experiment. For FACS, cells were transfected transiently with 500 ng pKLV1-gRNA-Puro plasmid and selected with puromycin on days 1-3. Instead of limiting dilution plating, RPE1 cells were single cell sorted on day 17 into a 96W plate with M10 media supplied with Penicilin/Streptomycin. Plating efficiency was around 10-20% on day 17 after sorting.

Bone marrow cells from a homozygous C57BL/6N CAS9-EGFP knock-in mouse (Platt et al., 2014) were isolated by flushing tibias and femurs in Hank's Balanced Salt Solution (Life Technologies) supplemented with 2% Fetal Bovine Serum (FBS) and 10mM HEPES (Sigma). Lineage negative cells were isolated using Direct Lineage Cell Depletion Kit Mouse (Miltenyl Biotec). After isolation and before sorting, cells were cultured in X-Vivo (Lonza) with 2% FBS, 50 ng/ml stem cell factor, 50 ng/ml thrombopoietin, 10 ng/ml IL-6 (Peprotech). Following a 3 h initial culture, 100'000 cells were electroporated (1550 v / 20 ms / 1 pulse) in buffer T with 44 pmols of crRNA:trRNA (IDT). On day 4 they were stained and sorted as described above. Single cell cloning was performed in Methocult M3434 media (6000 cells per 3 ml, StemCell Technologies) and colonies were picked 7-10 days later into 25 µl of direct PCR lysis buffer (Peqlab).

## 2.4 Comparison of delivery methods in HEK cells

For the RNP-electroporation condition, 150,000 HEK cells were electroporated (1700 V, 20 ms, 1 pulse) with 10 pmol Cas9 protein and 10 pmol hybridized crRNA:trRNA. For other methods, cells were seeded in 24W plates the day before the transfection so as to achieve 50%–70% confluency. For PiggyBac and transient plasmid conditions, cells were then transfected with 150 ng gRNA plasmid, 150 ng Cas9 plasmid, and 50 ng of either hyperactive PiggyBac transposase or carrier plasmid (pBluescript II SK+). For the protein + plasmid (P&P) and protein + plasmid + carrier (P&P-carrier) conditions, cells were transfected with two separately prepared mixes: (1) 3 pmol Cas9 protein, (2) 150 ng gRNA plasmid with or without 200 ng carrier plasmid (pBluescript II SK+). For RNP-lipofectamine conditions, cells were transfected with 3 pmol Cas9 protein, 3 pmol hybridized crRNA:trRNA (regular or stabilized). Plus reagent was added at 1 µL per 1 µg plasmid and Lipofectamine 3000 at double that volume. Cas9 protein was mixed with 1.5 µL Lipofectamine 3000. Cells were collected at indicated timepoints using trypsin, assessed for transfection effciency by flow cytometry and cell pellets were frozen for genomic DNA extraction.

Profiling of indels using IDAA procedure was performed according to the published protocol (Lonowski et al., 2017). In short, genomic DNA was extracted, a ~350 bp region around the cut site was amplified and tagged with a fluorescent dye using TEMPase Hot Start DNA polymerase (Ampliqon) and the products were resolved using a sequenator, yielding the indel profile.

## 2.5 PacBio sequencing and analysis

PCR amplification was performed using Q5 (NEB) or HiFi Hotstart ReadyMix (Kapa Biosciences). First 25 cycles used genomic primers with an adapter overhang (forward: GATGTACAGAGTGATATTATTGACACGCCC, reverse: CCAGGGGGATCACCATCCGTCGCCC or forward and reverse: CGACTCGCTACCAATGAAGACAGC). Products were purified using AMPure XP magnetic beads. One tenth of the eluate was used in a secondary 6 cycle PCR reaction to add recommended PacBio barcodes. For

*PigA*, these corresponded to different gRNAs and protein expression levels, whereas for *Cd9* they corresponded to single cell clones. Products were pooled equimolarly, prepared for sequencing by ligation of "SMRTbell" adapters by the Bespoke Sequencing Team (Wellcome Sanger Institute) and sequenced on the RSII instrument.

Analysis of PacBio data was performed using command line version of SMRT-Link software (pbtranscript 1.0.1.TAG-1470). For the purpose of calculating *PigA* locus coverage, a circular consensus sequences (CCS) were derived from multiple read-throughs of the same DNA molecule using "ccs --minPasses=1 --minPredictedAccuracy=0.9". Genome coverage was calculated with "bedtools genomecov –dz" (v 2.27.1) using CCS and visualized using ggplot2.

Individual *PigA* and *Cd9* alleles were reconstructed using Iso Seq workflow. In short, CCS were called using "ccs --minPasses=0 --minPredictedAccuracy=0.8" and classified into full length non chimeric ("FLNC", with both primer binding sites detected) and non full length ("NFL") reads using "classify" command. FL reads were also split by barcode, separating single cell clones (*Cd9*) or split into bins of 1 kb size using "separate_flnc" command (*PigA*). Iterative Clustering and Error correction (ICE) was peformed on each group (clone or size bin) individually using "cluster --targeted_isoseq" command. A custom script rebuilding the mapping index on each iteration of the clustering was used to fix a programming bug. Resulting "high quality" alleles (as classified by the clustering script) were mapped to the reference genome using "bwa mem" (v 0.7.17-r1188). Downstream analysis was performed using custom R (v 3.3.2) and bash scripts. For the *PigA* locus, reads were clustered furthered based on mapping and alleles with less than four FL reads support were filtered out. For the *Cd9* locus, additional filters based on FL to NFL ratio and within clone abundance were applied. Remaining *Cd9* alleles were visually inspected and ambiguities were resolved by Sanger sequencing. Additional alleles were discovered by custom PCRs (to detect larger deletions, large insertions and small indels) and Sanger sequenced. Lesions both smaller than 6 bp and farther than 20 bp from the cut site, as well as lesions in low complexity regions were removed from *Cd9* alleles.

## 2.6 Bioinformatics

Analysis of IDAA experiments was performed in R using binner package (https://github.com/plantarum/binner). Efficiency score was calculated as 1–(wild-type peak intensity/sum of wild-type and prominent peaks intensities). Spurious "–1 bp" signal present in wild-type samples was estimated to be around 10% of wild-type peak. This intensity was subtracted from "–1 bp" peak and added to the wild-type peak in all samples. Prominent peaks were defined using an arbitrary cutoff on the sum of intensities over many experiments.

Approximately 25 bp long primers with melting temperature of 60°C were designed using Primer3 or Primer3-BLAST. Guide RNAs were designed using Benchling and CRISPRscan (Moreno-Mateos et al., 2015), each guide being mismatched on at least two positions to any predicted off-target site. Flow cytometric data were processed with FlowJo (v 10.4.1). Mixed Sanger traces were resolved using the online tool PolyPeakParser (Hill et al., 2014). For visualization purposes, alignment of alleles (whether derived from PacBio or Sanger sequencing) was performed using BLAT (v 35, with settings -tileSize=6 -minScore=50 -minIdentity=90) and converted into BAM format using a customized script from Tobias Marschall (https://github.com/ALLBio/allbiotc2/tree/master/synthetic-benchmark). All visualizations were made in R using ggplot2 package.

# Chapter 3

# Dynamics of indel profiles induced by various Cas9 delivery methods

## 3.1 Introduction

### 3.1.1 Cas9 mutagenesis of genes

CRISPR/Cas9 has made site-specific mutagenesis highly efficient. Therefore, experimental and therapeutic edits can be performed on populations of cells, even without integration of an exogenous selection cassette. In many cases, good results can be achieved without extensive optimization, especially when cells constitutively express Cas9 or when selection for the desired phenotype is possible. In such circumstances, optimization may not be necessary. However, it may become paramount when phenotypic selection is impossible, when Cas9 has to be delivered into the cells, when the phenotypic effect is small or when cell numbers are limiting (e.g. patient material).

Ideally, the efficiency of a Cas9 experiment should be measured on the level of phenotype. However, it may be expensive, time consuming, the necessary reagents may not be available, the gene product may not be expressed in the edited cells or the phenotypic effect may not be detectable within a reasonable timeframe (e.g. if cells need to be quickly reintroduced into the patient). In this case, the genotype may be a good proxy for the phenotype. This is particularly the case, if the phenotype-genotype relationship has been established in a pilot experiment.

### 3.1.2 Genotyping of small indels in bulk cell populations

Routine genotyping of bulk cell populations can be performed by methods such as Enzyme Mismatch Cleavage (EMC), Indel Detection by Amplicon Analysis (IDAA), Tracking of Indels by DEcomposition (TIDE) and Next-Generation Sequencing (NGS) (Brinkman et al., 2014; Lonowski et al., 2017; Yang et al., 2015; Yeung et al., 2005). Each of these method requires short-range PCR amplification of the genomic region of interest (<600 bp), followed by various methods of allele separation and detection.

In EMC, colloquially known as the T7 or Surveyor assay, the pool of PCR products is denatured and rehybridized, resulting in formation of homo- and heteroduplexes. The latter are selectively digested by a heteroduplex sensitive endonucleases (e.g. T7EI, CELI or Surveyor) and separated from wild-type sized homoduplexes by agarose gel electrophoresis. Quantification of band intensities determines a simple efficiency score. In IDAA, fluorescently labelled PCR products are resolved at basepair resolution and quantified using a fragment analyzer machine (similar to those used in Sanger sequencing). This yields an indel profile, the frequency of different-sized indels (Fig. 3.1). In TIDE, PCR products from wild-type and mutagenized cells are Sanger sequenced and the resulting traces are computationally deconvoluted into an indel profile. Assessment of templated editing is also possible using a related TIDER procedure. In the NGS approach, the PCR products are sequenced using the Illumina platform. Both

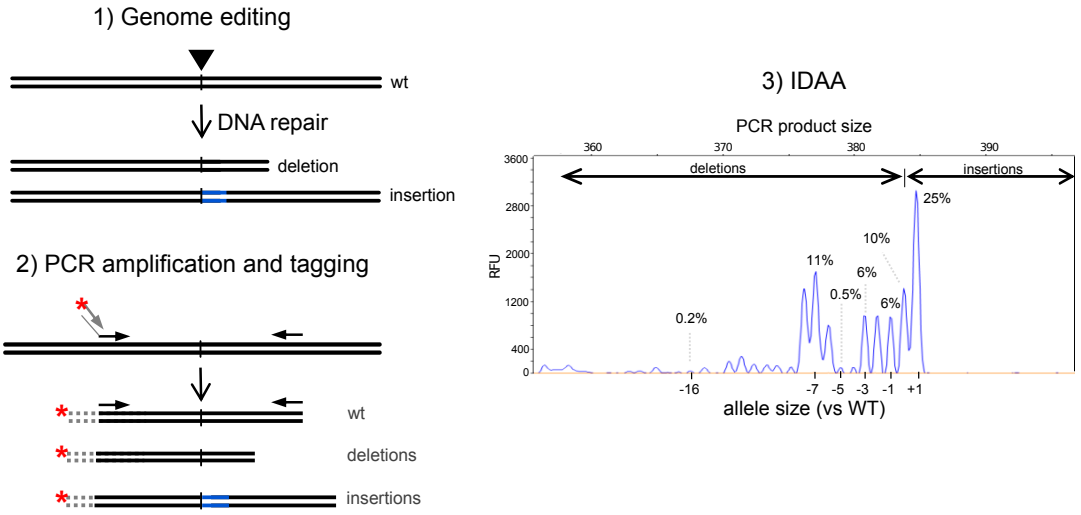



**Figure 3.1:** IDAA workflow. Cells are mutagenized in a pool, resulting in small deletions or insertions. The region of interest is amplified using a triprimer PCR reaction. One of the two genomic primers contains an overhang that allows annealing and amplification by a universal fluorescently labelled third primer. Fluorescent PCR products are separated at single basepair resolution and quantified using a fragment analysis machine (a capillary DNA sequencer). PCR product size (top x-axis) can be expressed as the allele size versus wild-type (wt, bottom x-axis) and the amount of given product/allele corresponds to the intensity of the peak expressed in relative fluorescent units (RFU, y-axis). Modified from Lonowski et al., 2017.

indel profiles and editing can be investigated using computational approaches (CRISPR Genome Analyzer, CRISPResso, CRISP-R).

EMC is the "quick and dirty" method. It takes less than a day and offers only a crude quantitative measure of efficiency, as it cannot distinguish between in-frame and out-of-frame mutations (both of which can form heteroduplexes that are cleaved by the endonuclease). Furthermore, endonucleases used in the assay are not sensitive to single base changes and heterozygosity at the wild-type locus may produce spurious signal due to formation of "natural" heteroduplexes in unedited samples. IDAA and TIDE resolve alleles by size at single basepair resolution with sensitivity similar to NGS (~0.1%, assuming the usual read quality cut-off). This makes them tools of choice for estimation of out-of-frame mutations, a proxy for functional knockout. When investigating many pools at once, NGS becomes an economically viable option, as many pools can be sequenced in one run for the same price. As the only method to produce sequence level data, it can resolve indels of same size, but different basepair composition. This property also allows it to reliably quantify templated editing. However, NGS is considerably more time-consuming than other methods to execute and to analyze.

### 3.1.3 Factors influencing Cas9 mutagenesis and genotyping

Genotyping gives an estimate of the overall efficiency of generating mutant alleles. This can be influenced by generic factors, such as transfection efficiency (proportion of cells receiving the reagents) and concentration of reagents within the cell over time. The latter is influenced by transfection multiplicity, activity of the gRNA and Cas9 promoters, whether expression constructs are stably integrated or transiently transfected, etc. If necessary, most of these factors can be optimized in a given cell system. Furthermore, the efficiency varies between different gRNAs, independently of the generic factors. The underlying reason is likely a combination of high and accurate expression of a particular gRNA sequence from a PolIII promoter (if used), the binding and cutting activity

of Cas9 at a given locus as well as how often the given locus reverts to wild-type upon repair.

How mutant genotype translates into gene knock-out will depend on both gene and gRNA-specific factors. Knock-out of non-coding genes is usually achieved by introducing a large deletion with paired gRNAs, as such genes are robust to small indels introduced by single gRNAs. Conversely, a small out-of-frame mutations in the first few exons of a protein-coding gene is often enough to abolish protein expression through nonsense-mediated decay (NMD) of its RNA transcript. However, alternative splicing and transcription start sites may lead to production of functional protein despite such mutations. Activity of many protein domains is also sensitive to in-frame mutations. Composition of mutant genotypes (the "indel profile") is specific to each gRNA (Chakrabarti et al., 2018; Taheri-Ghahfarokhi et al., 2018; van Overbeek et al., 2016). In consequence, two guides targeting the same protein domain with the same overall level of DNA mutation may differ substantially in the proportion of out-of-frame mutations and thus in the level of protein knock-out.

Little is known about the factors influencing the indel profile of a specific gRNA. Microhomologies around the cut site are speculated to contribute to it. Transcription has been reported to increase mutagenic efficiency of Cas9 in a strand-dependent manner, which implies it may also influence the indel profile. However, no general rules have yet been defined. The indel profile is therefore usually established empirically for each gRNA.

Knowing how different factors influence the "indel profile" may help predict the level of phenotypic knock-out. I set out to study two such previously unexplored factors - time of genotyping and Cas9/gRNA delivery method. If genotyping is performed before the indel profile becomes stable, the efficiency may be assessed incorrectly. On the other hand, delaying genotyping may be inconvenient, especially when cells need to be reinjected into the model organism or patient as soon as possible after the procedure. Similarly, if different delivery methods result in similar outcomes, they could be interchanged at convenience. If not, the differences could be exploited to achieve higher phenotypic knock-out. Finally, different Cas9/gRNA delivery methods may have different dynamics of mutagenesis. It may be beneficial to know whether genotyping time needs to be adjusted depending on the method.

## 3.2 Results

The Cas9 protein and guide RNA may be delivered into cells in a variety of forms (e.g. plasmid DNA, mRNA, protein) and using a variety of methods (e.g. electroporation, lipofection, transduction). Methods used in this study are summarized in Fig. 3.2. I collected cells at multiple timepoints post-delivery and analyzed the indels using IDAA. I chose IDAA, because it offers rapid results and high resolution at a low cost. I performed the experiment in a commonly used HEK293 cell line, known to be amenable to many delivery methods and previously shown to achieve high Cas9-induced mutagenesis rates. A validated, highly efficient gRNA against the *ST6GALNAC1* gene (Hansen and O'Shea, 2015), which is silent in HEK cells, was picked in hope of avoiding knockout-specific proliferative effects. To minimize differences between methods I utilized a plasmid backbone containing both lentiviral and PiggyBac functional elements (Metzakopian et al., 2017).

### 3.2.1 Transfection and integration dynamics

Methods compared in this study result in either stable (lentivirus, PiggyBac) or transient (RNP, transient plasmid, P&P - protein & plasmid) expression of Cas9 and gRNA. As neither component of the RNP was fluorescent, I did not monitor the transfection efficiency of this method. As the Cas9 plasmid contains no fluorescent marker, I monitored the BFP expression from BFP-gRNA cassette as a proxy for the overall transfection and integration efficiency (Fig. 3.3a). Cas9 was not selected for and therefore the percentage of

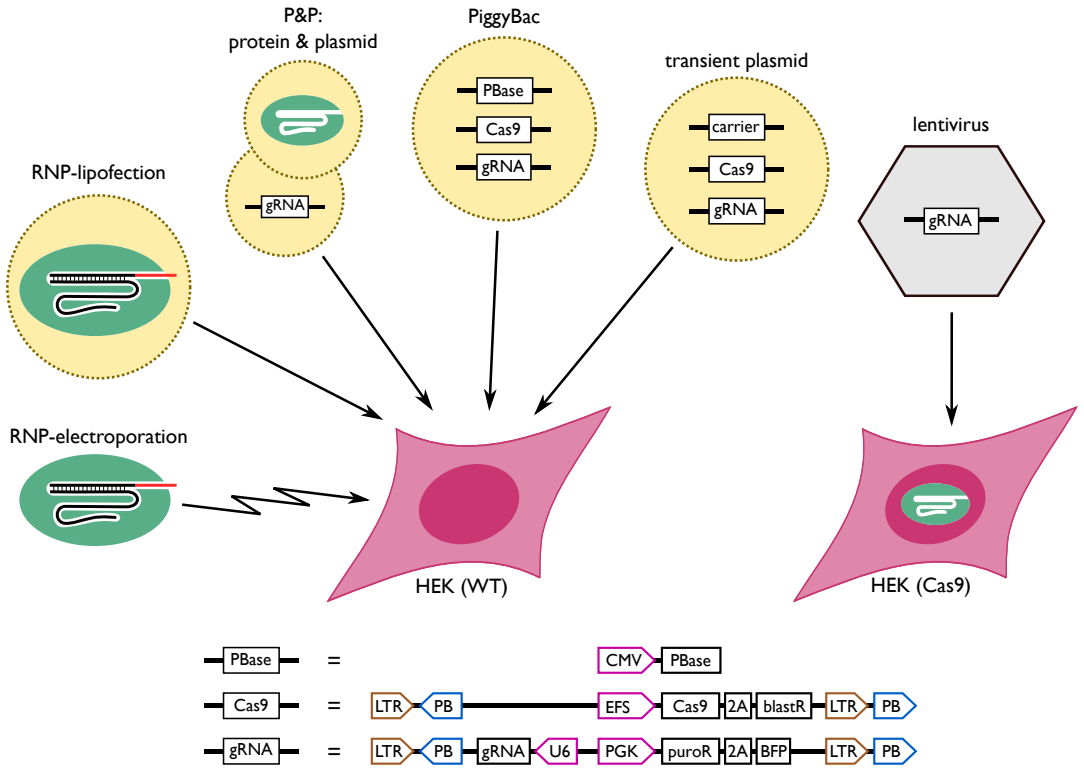


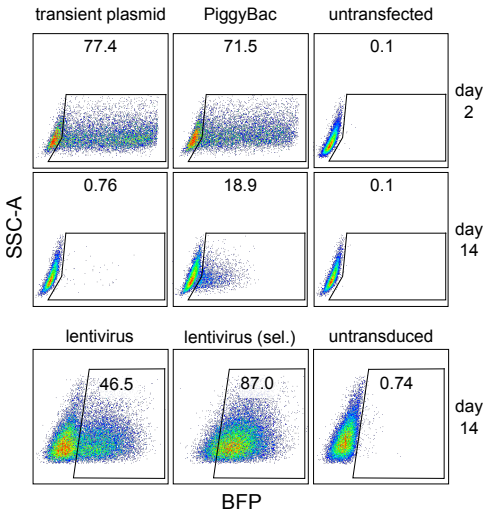

**Figure 3.2:** Cas9 and gRNA delivery methods. The same gRNA and Cas9 plasmids were used for all experiments. Yellow circle indicates transfection with Lipofectamine 3000. RNP-electro: electroporation of Cas9 protein and synthetic two-part gRNA (crRNA+trRNA) using Neon Transfection System. RNP-lipo: as RNP-electro, using lipofectamine. P&P - protein & plasmid: transfection of Cas9 protein and gRNA-encoding plasmid. A version with addition of carrier plasmid (pBluescript II SK+) was also used (P&P-carrier). PiggyBac: transfection of plasmids encoding Cas9 and gRNA together with PiggyBac transposase (resulting in stable cellular integration). transient plasmid: as PiggyBac, with transposase replaced by the carrier plasmid (pBluescript II SK+). lentivirus: transduction of HEK cells stably expressing Cas9 with gRNA lentivirus. Vector schematics in the bottom part of the figure, LTR = Long Terminal Repeats, lentiviral elements; PB = PiggyBac repeats; blastR and puroR = blasticidin and puromycin resistance cassettes; LTRs, PBs and promoters are marked with colors.

BFP positive cells is an overestimation of overall number of cells containing both Cas9 and gRNA-expressing constructs.

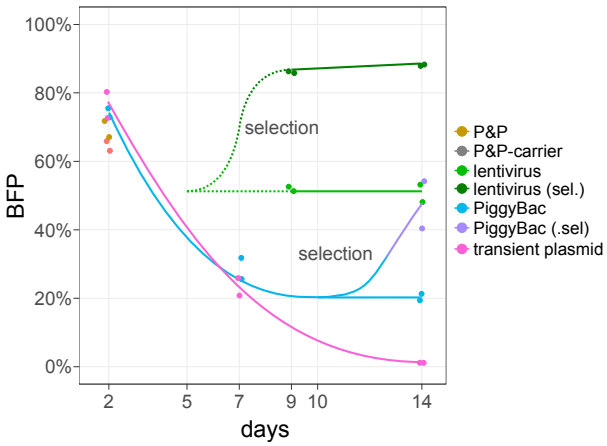
PiggyBac and transient methods resulted in the highest transfection efficiencies on day 2 (>70%), followed by P&P-carrier and P&P (55% and 40%; Fig. 3.3b). As expected, the BFP expression was all but extinguished in the transient plasmid, P&P and P&P-carrier conditions by day 14. The few remaining BFP positive cells may indicate rare cases of stable integration of the plasmids.

In the PiggyBac condition, 20% of the cells (about 1/4 of all transfected cells on day 2) remained BFP positive on day 14, indicating stable transposition of the PiggyBac transposon from the donor DNA into the genome. Short-term selection using puromycin (days 10–14) for the genomically integrated gRNA-BFP construct increased the percentage of positive cells to 44%.

Data from early timepoints in the lentiviral transductions was not collected, but the percentage of BFP positive cells was maintained at approximately 50% between days 9 and 14 post-transduction, indicating stable integration. Selection for the integrated gRNA-BFP pro-virus between days 5 and 14 increased this proportion to 87%. As puromycin normally kills all wild-




**(a) Examples of flow cytometry plots.** Samples were assessed using Cytoflex machine, except for the lentiviral method, which was assessed using BD Fortessa.




**(b) Comparison of transfection levels across methods and timepoints.** *Solid line* indicates the most likely path between collected timepoints. *Dotted line* connects day 7, the time when selection of lentiviral condition was started (no data available, efficiency inferred) to day 9 for selected and unselected samples.

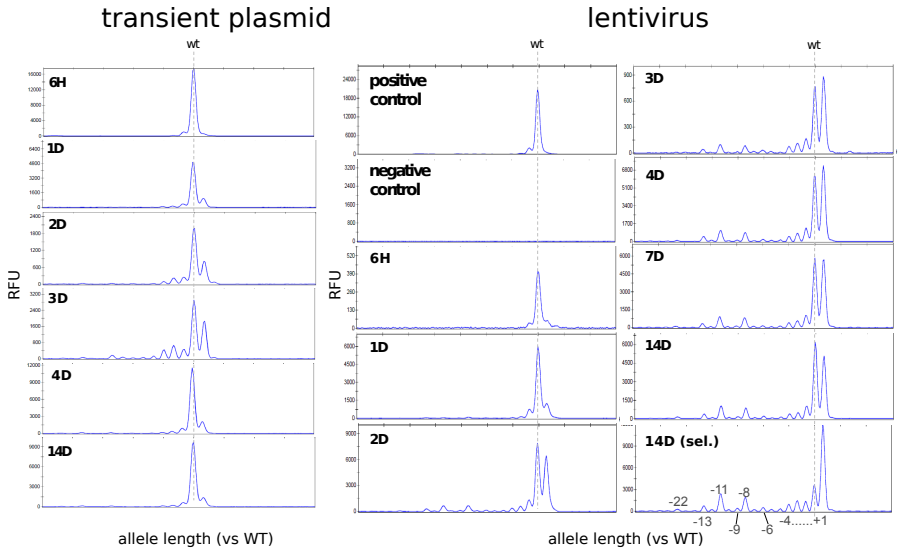
**Figure 3.3:** Transfection efficiency over time. HEK cells transfected with Cas9 and gRNA were analyzed by flow cytometry at various time points. BFP fluorescence comes from gRNA expressing plasmids and thus likely overestimates the actual percentage of Cas9/gRNA double-positive cells. The exception is lentiviral transduction, where cells constitutively express Cas9. As RNP has no fluorescent component, the transfection efficiency of this methods is unknown. sel.: selected for gRNA construct using puromycin.

type HEK cells (data not shown), I speculate that remaining 13% of cells were resistant, but expressed no or little BFP. Since BFP follows the puromycin resistance gene in the expression cassette (Fig. 3.2), it is possible for a knock-out mutation to occur within BFP without affecting the puromycin resistance (personal communication, Konstantinos Tzelepis).

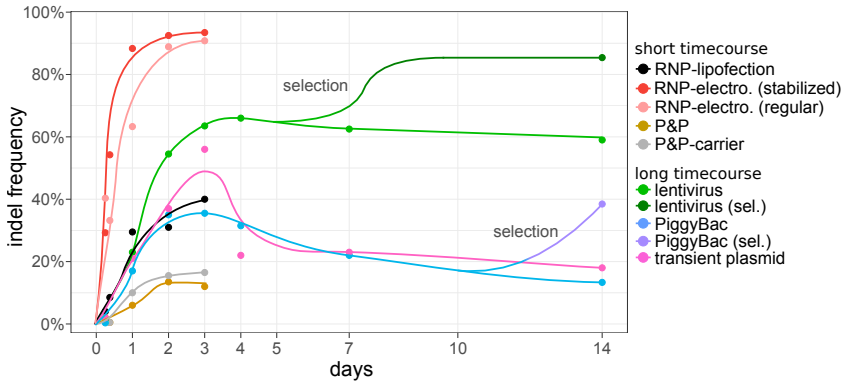
### 3.2.2 Mutagenesis efficiency over time

I studied the dynamics of indel generation using IDAA. For RNP and P&P methods I only collected samples up to day 3 post-delivery, on the assumption that Cas9 protein is degraded by that time. For other methods, I continued collecting samples until day 14 (examples of IDAA indel profiles in Fig. 3.4a). The mutagenesis efficiency was calculated as a ratio of intensity of the prominent, non wild-type peaks to all the peaks (Fig. 3.4a). As the wild-type size peak may represent rare SNPs or balanced indels in addition to the wild-type allele, this efficiency may be slightly underestimated.

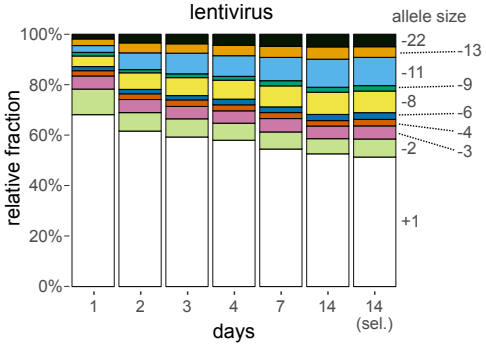
On day 3 post-delivery, RNP-electroporation was the most efficient method (91–93%), followed by lentivirus (64%), transient plasmid (56%), RNP-lipofection, PiggyBac (36%–40%) and P&P/P&P-carrier (12–17%; Fig. 3.4b). The transfection efficiency broadly correlated with the mutagenic efficiency, except for P&P and P&P-carrier, which induced relatively lower levels of indels. For transient plasmid transfection, no smooth curve fit could be found for experimental data, indicating some timepoints were outliers. However, selection for stable PiggyBac integrants between days 10 and 14 using puromycin more than doubled the final percentage of mutagenized alleles (from 17% to 38%). Similarly, selection for provirus integration from day 5 to day 14 increased that percentage from 59% to 86%. Despite stable Cas9 expression and long-term selection for the integration of the gRNA construct 100% mutagenic efficiency was not achieved. This may reflect the




**(a)** Examples of IDAA plots for lentivirus and transient plasmid conditions. Prominent peaks used for efficiency calculation are indicated in the selected sample. Asterisk indicates truncated product (see Discussion). Representative panels from duplicate experiments are shown. sel.: selected for gRNA construct using puromycin.




**(b)** Indel frequency over time. Solid line indicates most likely path between collected timepoints. Samples involving transfection of Cas9 protein were only collected up to day 3. N = 1-2.




**(c)** Indel profiles over time in the lentiviral sample. Prominent peaks, as indicated in Fig. 3.4a, are colored. N = 1-2.

**Figure 3.4:** Indel frequency and profiles overtime. HEK cells transfected with Cas9 and gRNA were analyzed by IDAA at various timepoints.

inherent limitation of the method, such as inability to distinguish substitutions and "balanced" indels from wild-type and occasional misclassification of the 1 bp insertion as wild-type due to Taq polymerase action (see Discussion). Furthermore, some gRNA-expressing cassettes may have been silenced independent of the puro-BFP cassette before mutagenesis could occur.

Both RNP-electroporation and RNP-lipofection show earliest indels formation at 6h and 9h post-transfection. Most of the methods reached their maximum efficiency on day 2 and 3. Interestingly, transient plasmid and PiggyBac exhibited a significant decrease in efficiency after day 3, while lentivirus plateaued.

To reduce technical variation, I repeated the experiment on a single day using transient plasmid, RNP-lipofection, PiggyBac, P&P and P&P-carrier methods (data not shown). On day 3 post-delivery, I analyzed the cells by flow cytometry and TIDE. All methods using protein Cas9 (RNP, P&P, P&P-carrier) achieved the same mutagenic efficiency of around 12%. Despite P&P-carrier transfection efficiency being slightly higher than with transient plasmid method (82% vs 76%), the mutagenic efficiency was much lower (12% vs 27%). These results indicate that Cas9 protein was equally likely to co-transfect with synthetic gRNA as with gRNA plasmid and that Cas9 protein was the limiting factor for mutagenic efficiency in the experiment.

Both transfection and mutagenic efficiency were higher in transient plasmid than PiggyBac condition on day 3, consistent with a similar difference on day 14 in the experiment presented in Fig. 3.4b. Therefore, higher transfection efficiency likely explains the higher mutagenic efficiency observed at day 14. It is unclear why there was a difference in transfection efficiency in the first place. It may be that the carrier plasmid in transient method increases the plasmid entry rate compared to the transposase plasmid, which it replaces, possibly due to its smaller size and hence higher copy number. Alternatively, the transposase may have effectively reduced the concentration of the gRNA expressed from the transposon, either by direct transcriptional interference or by exposing the transposon plasmid to degradation.

### 3.2.3 Indel profiles over time

To investigate whether the allelic composition stays stable or fluctuates over time and across methods, I quantified the relative abundances of IDAA indel peaks. As larger deletion indels (e.g. 22, 13, 11 and 8 bp) may correspond to microhomologies that I found around the target site (data not shown), I wondered if they would appear later than the smaller ones. This would be consistent with the observation that NHEJ which usually creates small indels is a faster repair pathways than MMEJ (Mladenov and Iliakis, 2011). In lentiviral condition, which yielded most reliable data, larger indels appear later in the timecourse (Fig. 3.4c). This suggests either some large indels take longer to form or that small indels (such as 1 bp insertion) and substitutions are susceptible to recutting, yielding larger indels upon mutagenic repair.

### 3.2.4 Dynamic effect of gRNA stability on indel efficiency

It has been proposed that gRNAs protected from cellular exonucleases by chemical modifications may increase the overall indel generation (Hendel et al., 2015). Therefore, I studied the dynamics of indel generation using IDAA in cells electroporated with Cas9 protein and either regular or stabilized gRNAs. Both gRNAs showed similar, rapid indel formation by 6h post-electroporation and demonstrate comparable indel profile and maximum efficiency at the end of the experiment on day 3 (Figs. 3.5). However, the stabilized gRNA reached its maximum efficiency on day 1, faster than the regular gRNA by about 24h. It is known, that gRNA loading is a key regulator of Cas9 enzyme function (Jiang et al., 2016) and I speculate that stabilized gRNA may improve its binding affinity and/or nuclease activity.

### 3.2.5 Comparison of IDAA and TIDE methods

As both IDAA and TIDE generate indel profiles, I decided to compare the results for selected samples. While the overall profiles were similar, the sensitivity of TIDE differed depending on which primer was used for the Sanger sequencing reaction (Fig. 3.6). As IDAA does not involve a sequencing step, it is not susceptible to this problem.

## 3.3 Discussion

I have investigated the time dynamics of indel profile generation by different methods of Cas9 and gRNA delivery. I showed that the mutagenesis and the initial transfection efficiency were roughly correlated, except when low protein Cas9 transfection was the likely limiting factor. Explicit measurement of the efficiency of Cas9 delivery, e.g. by using a Cas9-GFP fusion, would help solidify this finding. I have confirmed RNP delivery is characterized by rapid indel induction and found an indication that stabilized gRNAs may further speed up the process, which could potentially be useful in therapeutic setting.

### 3.3.1 Causes and consequences of mutagenesis efficiency fluctuation

Unexpectedly, when using plasmid transfection methods (but not lentivirus), the proportion of mutagenized cells did not plateau at the maximum level (usually around day 3 or 4), but peaked temporarily around day 3-4 and then decreased. One explanation for this effect may be that alleles, which end up being perfectly repaired by the slower process of homologous recombination remain undetectable for a longer time than the ones repaired mutagenically. Therefore, I observe a temporary enrichment for mutagenic genotypes. Discrepancy between lentivirus and plasmid methods may be explained by higher frequency of stable lentiviral integration. Most cells that are transduced continue expressing the gRNA from an integrated lentiviral cassette and are eventually mutagenized, leading to a plateau. In contrast, many cells in transient plasmid and PiggyBac conditions express the gRNA temporarily from the unintegrated plasmid, lose the construct due to cell division or plasmid degradation, repair the break perfectly and remain wild-type.

Another potential explanation is that cells mutagenized using the plasmid grow slower and end up being outgrown by wild-type cells. This could be due to toxicity associated with the DNA (lentivirus typically infects at low multiplicity, but many copies are delivered by transfection) or with the carrier (lipofectamine). One way to clarify this issue would be to monitor frequency of alleles "in repair" by qPCR and to compare growth rates. Comparing mutagenic efficiency of RNP-lipofectamine and RNP-electroporation beyond day 3 would indicate whether lipofectamine itself causes the effect.

As a consequence of this dynamic, correct estimation of the phenotypic effect from a snapshot genotype becomes difficult. In the "plateau" model, premature genotyping leads to underestimation of the effect, but the window for correct genotyping after reaching the plateau is large. In the "peak" model, premature genotyping may paradoxically better mirror the ultimate phenotypic effect, if only by chance. The decrease, at least in the PiggyBac condition, seems to take a long time and is quite substantial (from around 35% on day 3 to around 15% on day 14). One potential solution to this problem is to genotype at the peak and include the expected decline in phenotypic calculations.

### 3.3.2 Delayed large indel formation

The indel profile in lentiviral condition showed that larger indels form later (Fig. 3.4c). A recent paper found a similar effect and attributed it to the slower repair kinetics of MMEJ (Brinkman et al., 2018). While this is almost certainly the case early on in the timecourse, completion of MMEJ does not generally take many days, as observed in here (biochemical studies suggest $t_{1/2}$ of 2-20h, Iliakis, 2009; Perrault et al., 2004). Therefore, late formation of larger indels remains unexplained. Recut-

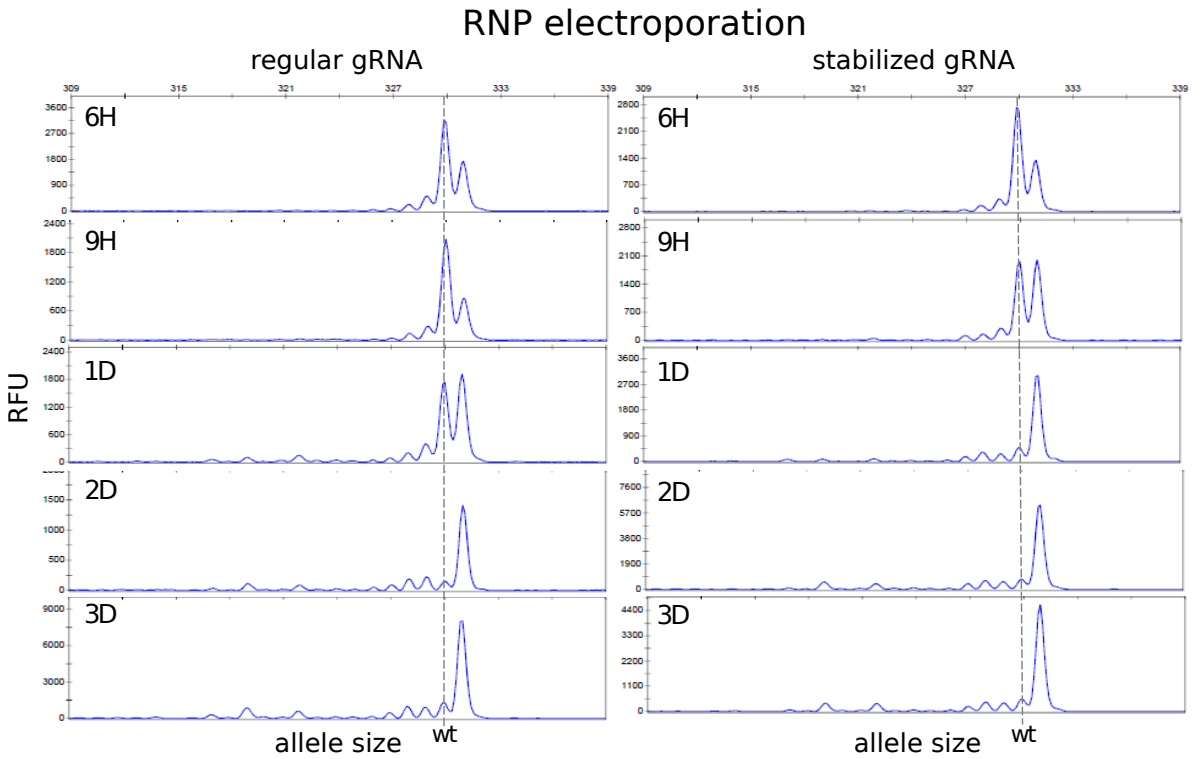



**Figure 3.5:** Comparison of normal and stabilized synthetic gRNAs. HEK cells were electroporated using Neon transfection system with protein Cas9 and regular or stabilized synthetic gRNAs. Representative panels from duplicate experiments are shown.

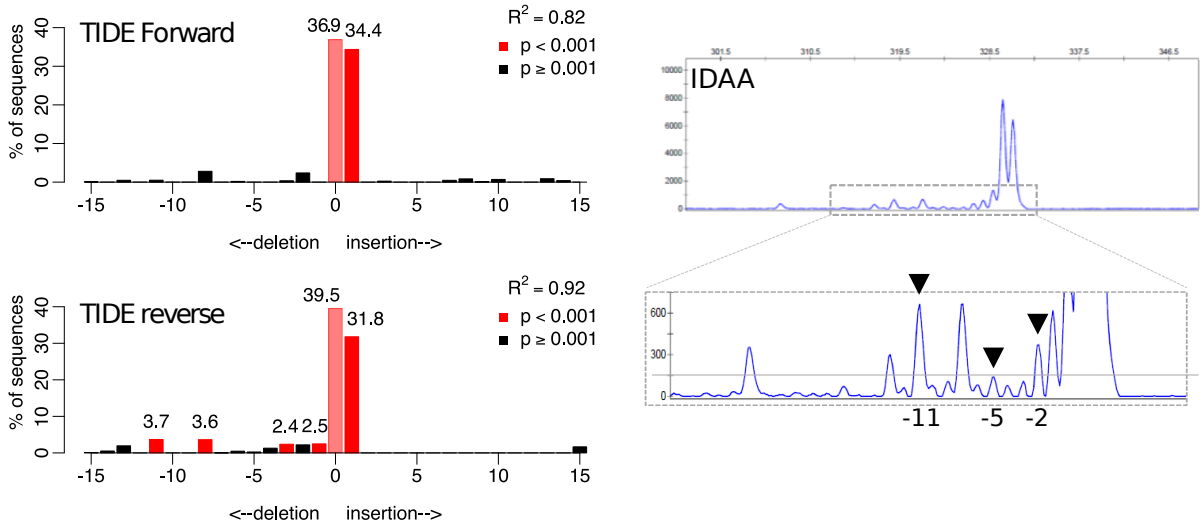



**Figure 3.6:** Comparison of IDAA and TIDE. The sample shown in Fig. 3.4a, lentivirus 3D, was indel profiled by TIDE analysis using both the forward or reverse Sanger sequencing raw data files. The results were compared to IDAA generated profiles, where a zoom in of the indel peaks is provided as inset for clarity. Recommended detection threshold (150RFU) is indicated by the horizontal line. Selected indels are indicated with black triangles. N = 1.

ting of small indels and substitutions may be the cause. If large indels continue to accumulate after two weeks, it would validate this hypothesis. Conversely, lack of additional accumulation would not falsify this hypothesis as recutting and mutagenic repair of such indels may be slow enough to reach an equilibrium with production of new wild-type alleles by DNA replication. Therefore, investigation of mutagenic dynamic using gRNAs with single mismatches could provide a more direct answer.

### 3.3.3 Caveats of indel profiling

Genotyping of the other delivery methods did not yield data of sufficient quality to perform a time-course analysis or to compare indel profiles across methods. Two issues caused this. First, the low efficiency of some methods, especially at the early timepoints, made signal detection and quantification difficult. Excess material had to be loaded to detect weak signal from indel alleles, which led to a strong wild-type signal overwhelming adjacent –1 and +1 peaks. Furthermore, I assumed co-linearity between peak intensity and abundance of an allele. However, for low intensity peaks this assumption is likely incorrect. Additional experiments to establish a standard curve and quantify the magnitude of the signal saturation could rectify these issues. The second issue was the presence of a spurious –1 peak. Since it was detected in all control samples (with an intensity of between 5-15% of wild-type peak), I assume it has been created by the use of Taq polymerase, which adds a single 3' adenine to all PCR products with less than 100% efficiency. This precluded accurate quantification of the –1 indel and likely affected quantification of other peaks as well. Usage of high-fidelity non-Taq polymerase would likely remedy this issue.

Finally, this study suffers from lack of replication. Only a single guide was used and experiments were repeated at most twice. More replication would increase confidence in the presented results.

# Chapter 4

# Detection and quantification of complex lesions

## 4.1 Introduction

### 4.1.1 Mutation reporters

Basic research in genetics, genome engineering and DNA damage repair uses simple assays, which allow isolation and quantification of cells with particular mutagenic events. The Ames test used to measure the mutagenicity of chemicals is one example (Ames, 1979). In this assay, bacteria harboring a mutation that makes them histidine dependent are exposed to potentially mutagenic compounds, which causes some of them to become histidine independent. The frequency of this reversion can be used as a measure of the mutagenicity of the compound. Furthermore, surviving bacteria can be analysed to establish the exact genetic cause of the reversion, which can be the restoration of the original genotype or a compensating mutation elsewhere in the gene (e.g. restoration of the reading frame). Variations on this assay continue to be used in genetics and toxicology research.

Guided by a similar, selective principle, the first systematic investigation of gene targeting used the endogenous *Hprt* gene and an exogenous neomycin resistance cassette (neoR) to isolate targeted cells (Thomas and Capecchi, 1987). *Hprt* plays a central role in the purine salvage pathway and is dispensable for viability of cultured cells under normal conditions. Hprt-proficient cells can be isolated by using hypoxanthine-aminopterin-thymidine (HAT) medium, which kills cells unable to salvage the necessary nucleotides. Conversely, 6-thioguanine (6-TG) kills Hprt-proficient cells, which convert it to a toxic product. The researchers transfected cells with an Hrpt-targeting construct containing a neoR cassette and isolated correctly targeted cells by selecting for loss of Hprt expression and gain of neomycin resistance. Varying the concentration of the reagents, the transfection conditions and the parameters of the vectors itself (like the length of homology arms) allowed optimization of the experimental protocol. In addition to basic research on gene targeting, *Hprt* is commonly used as a safe locus for insertion of transgenes and as cassette for positive and negative selection (Conway et al., 2014; van der Lugt et al., 1991). NeoR split into two parts, which can recombine following I-SceI induced DSB to form a functional unit, has been used to study using HR activity, isolate intra and interchromosomal repair events as well as to quantify the relative length of repair tracts (Brenneman et al., 2002, 2000; Johnson and Jasin, 2000).

Methods outlined above rely on drug resistance and colony formation for isolation and quantification of mutagenic outcomes. This is potentially problematic, if drug selection interferes with the repair processes, if non-mutagenized cells need to be analyzed or if cells of interest do not form colonies. Furthermore, the need for colony formation limits the throughput of the procedure, due to time needed for colony outgrowth and low density at which cells need to be plated in order to recover pure clones. Discovery and development of fluorescent proteins eliminated these problems, allowing a simple flow cytometric efficiency readout and FACS isolation of both posi-

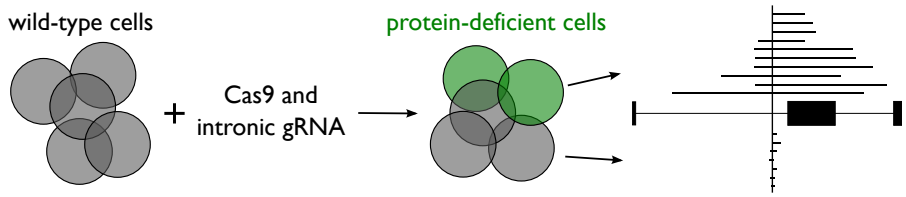



**Figure 4.1:** Assay design. Targeting introns allows quantification and isolation of complex lesions – i.e. lesions that are not small indels. The position of the gRNA is shown as a vertical line intersecting with the gene structure. Horizontal lines indicate indels and large deletions.

tive and negative cells (or intermediate states, if present; Julius et al., 1972; Prasher et al., 1992; Shimomura et al., 1962). Assays combining I-SceI induced DSB and fluorescent protein readout contributed substantially to research on DDR, allowing study of genetic requirements of HR (Pierce et al., 1999), translocations (Richardson and Jasin, 2000), SSA (Stark et al., 2004) and MMEJ (Bennardo et al., 2008). A principle similar to that in the split-neomycin assay was used, with split fluorescent proteins being placed at different loci and with different amounts of shared homology. Constructs combining multiple fluorescent proteins were designed to simultaneously quantify relative contributions of HR and NHEJ (Certo et al., 2011). An assay in which repair of Cas9-induced DSB using a ssDNA donor converts BFP to GFP, and mutagenic repair abolishes fluorescence altogether, was used to define optimal conditions for ssDNA donor integration (Richardson et al., 2016).

Most of the described DDR assays were designed to capture a specific type of mutation using a positive selection paradigm and often ignoring the negative population. I speculated that an assay based on negative selection against small indels, the most common lesion caused by Cas9 mutagenesis, will reveal repair outcomes that have been overlooked so far. Here, I describe the development of this assay.

## 4.2 Results

### 4.2.1 Assay design

I sought to establish an assay to detect and quantify cells with Cas9-induced lesions that are not small indels ("complex lesions"). I reasoned such an assay could be based on targeting intronic sites close to an exon (within 500 bp). Small intronic lesions are normally not expected to affect gene expression. Conversely, any other large intronic lesion, such as translocation, inversion or large deletions may affect gene expression (Fig. 4.1).

Mouse ES cells and hTERT immortalized, p53-deficient human retinoid pigment epithelial cells (RPE1) were used to establish the assay. In contrast to cancer-derived cell lines, both cell lines have a normal karyotype and intact DNA repair mechanisms, which makes them more representative of a normal somatic cell. Although mouse ES cells and embryonic fibroblasts differ in their use of DNA repair pathways, it is not known how they compare to other somatic cells (Tichy et al., 2010). P53 deficiency in the RPE1 cell line enabled easy characterization, as most Cas9-mutagenized p53-proficient RPE1 cells undergo apoptosis (Haapaniemi et al., 2018). Both ES and RPE1 cell lines can be single cell cloned, which allows creation of pure, Cas9 expressing lines as well as clonal genotypic analysis following mutagenesis.

Following criteria were used to pick targets for the assay:

- High surface expression or easily detectable function of the gene, which can be used as a readout and means of selection.
- Availability of flow cytometric reagents for detection of gene expression or function.
- "Isolated" exons flanked by more than 2 kb of intronic sequence in both directions, so that genotyping can be focused on one exon

only. Such exons could also be targeted on both flanks as a control.

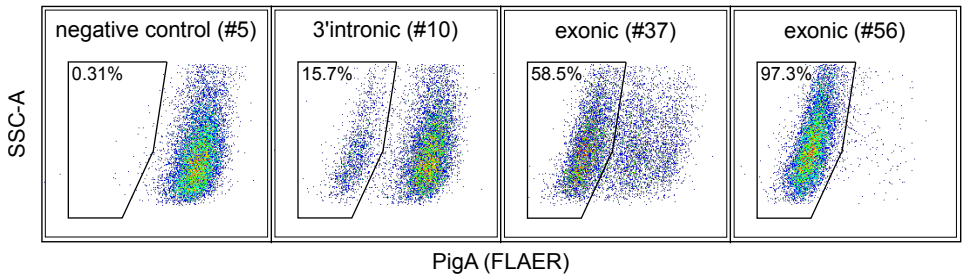
- Exons whose complete loss would change the reading frame of the transcript, so that no fully functional protein could be produced without them.
- Exons close to the 5' end of the transcript, as frameshifting mutations in these exons are more likely to trigger NMD.
- Genes which produce a single protein isoform (i.e. no alternative splicing and transcription start sites), as multiple ones could confound the readout.

For the initial study, I considered X-linked genes, which are present in only one copy in male mouse ES cells and are functionally hemizygous in female RPE1 cells, due to X inactivation. This makes the phenotypic readout stronger and easier to interpret, since only one copy of the gene needs to be inactivated to ablate the protein function. Additionally, in male ES cells I expected to detect exactly one allele per single cell clone, which would substantially simplify the genotyping strategy. Loss of chromosome X is lethal in male ES cells and they rarely maintain two X chromosomes. Therefore, detection of a single allele on chromosome X in male cells cannot be mistaken for detection of two identical alleles or monosomy, as is the case in female cells or at an autosomal locus.

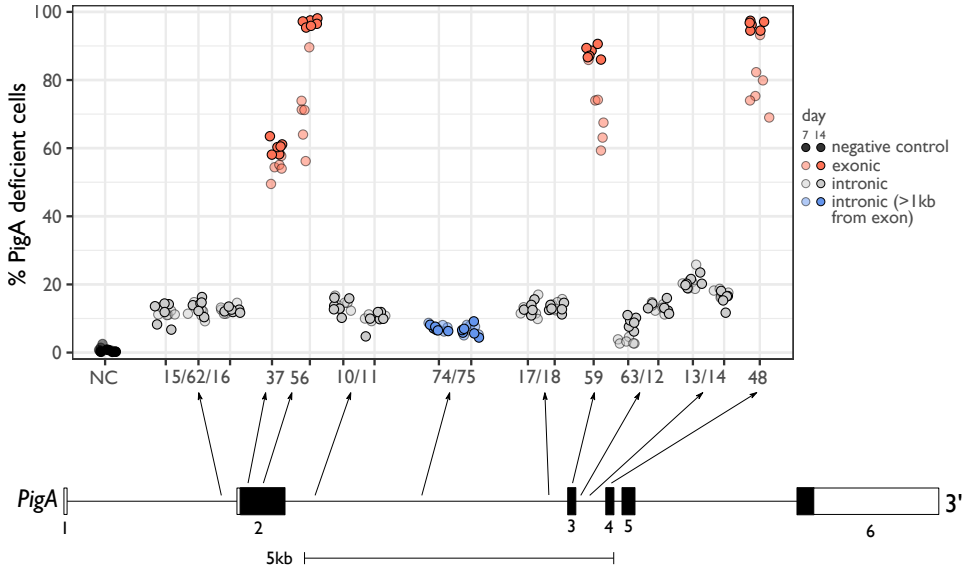
I considered *Hprt*, *Lamp2* and *PigA* as potential X-linked targets in my assay. **Hprt** is commonly used to enable gene targeting. However, my previous research indicated that the repeat rich regions around exons 2 and 3 make genotyping and Sanger sequencing particularly problematic. Furthermore, *Hprt* mutants can only be detected by a colony counting assay under 6-TG selection, which is time consuming. It may also be unreliable, if cells are plated too densely. Under such conditions some Hprt-deficient cells may be killed due to high local concentration of toxic products of 6-TG metabolism created by Hprt-proficient cells. **Lamp2** is a glycoprotein present at the lysosomal membrane that can also be found on the surface of mouse ES cells (unpublished data). However, exons 2-5 of *Lamp2* are not completely isolated, with both intron 2 and 4 being shorter than 1 kb. Moreover, available data implied Lamp2 may not be expressed highly enough to allow clear separation between positive and negative cells. **PigA** is one of the first elements of a biochemical pathway, which produces glycosylphosphatidylinositol (GPI) anchors necessary for attachment of some proteins to the surface of the cell (Miyata et al., 1993). The activity of the pathway can be assayed by flow cytometry using a fluorescent reagent, which binds to N-glycan on GPI-anchored proteins. This reagent (FLAER) is a fusion of the FITC molecule (FLuorescin isocyanate) and pro-aerolysin (AER), which is an inactive form of a bacterial toxin aerolysin (Sutherland et al., 2007). FLAER is routinely used in clinical practice to diagnose patients suspected to have paroxysmal nocturnal hemoglobinuria (PNH), a disease caused by deficient GPI-anchor production (Takeda et al., 1993). Genetic inactivation of *PigA* by CRISPR/Cas9 leads to complete loss of FLAER staining. It may result in slower cell growth, but the effect is modest (Koike-Yusa et al., 2014). Exon 2 is more than 2 kb away from exons 1 and 3, and its loss is an frameshifting mutation. Loss of exon 3 is also an out-of-frame mutation, but its proximity to exon 4 (500 bp) makes it less useful as a target. I chose to develop my assay based on the *PigA* gene.

### 4.2.2 Mouse *PigA* and human *PIGA* loci

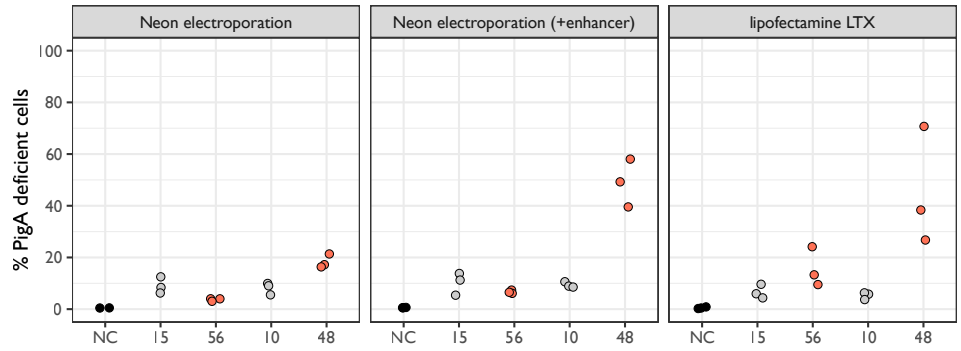
Cas9 and single gRNA constructs targeting intronic or exonic sites in chromosome X linked *PigA* gene were delivered into male JM8 mouse ES cells by PiggyBac transposition. Stable integration of both the Cas9 and gRNA expressing constructs was selected for using blasticidin and puromycin, respectively. This system allowed saturation mutagenesis of targeted loci, because even perfectly repaired targets would be recut until the site was destroyed. Staining with FLAER reagent was used to quantify the proportion of PigA-deficient cells 14 days post-delivery. Initial




**(a)** Examples of PigA mutagenesis revealed by FLAER staining, PiggyBac method.

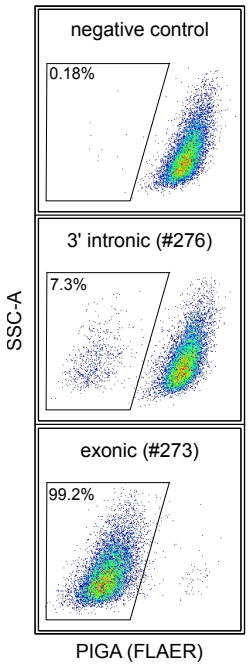



**(b)** Frequency of PigA loss, PiggyBac method. Each circle represents one independent cell culture (N = 6). Thick bars represent exons, hollow ones indicate UTRs. Dot transparency indicates time of sampling. NC: negative control, guide #5 targeting Cd9.

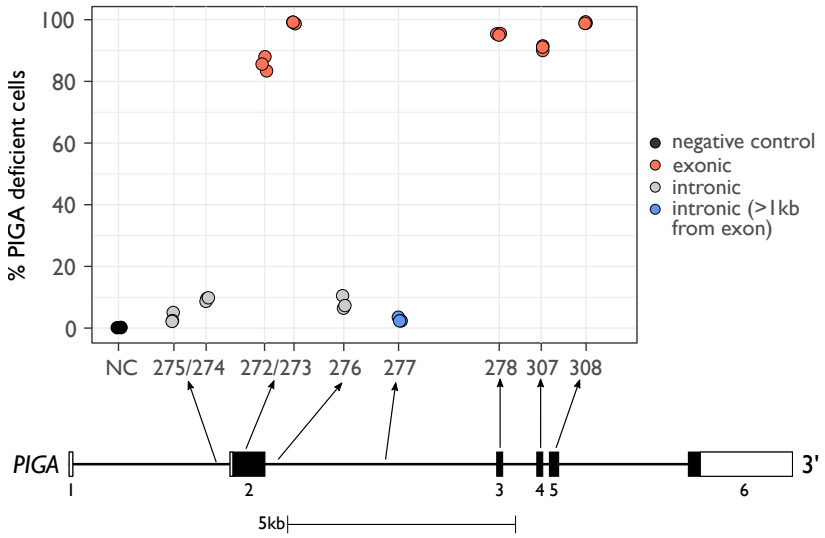



**(c)** Frequency of PigA loss, RNP method. Each circle represents one independent cell culture (N = 3).

**Figure 4.2:** Frequency of PigA loss upon mutagenesis with exonic and intronic guides in mouse ES cells. Individual guides are identified by numbers (Table 4.2).




**(a)** Examples of PIGA mutagenesis.




**(b)** Frequency of PIGA loss. Each circle represents one independent cell culture (N = 3). NC: negative control. Thick bars represent exons, hollow ones indicate UTRs.

**Figure 4.3:** Frequency of PIGA loss upon mutagenesis with exonic and intronic guides in human RPE1 cells using PiggyBac vectors. Individual guides are identified by numbers (Table 4.2).

experiments indicated that cells become FLAER negative gradually starting at around day 5 and plateauing at day 10. This is likely because it takes time for all GPI-anchored proteins to be recycled from the cell surface. Furthermore, some guides reached their plateau faster than others, likely due to guide-specific cutting rate or target-specific mutagenic repair (compare day 7 and day 14, Fig. 4.2b).

At 14 days post-transfection, three individual guides targeting exons 2 to 4 yielded very high rates of PigA loss (80–97%; Fig. 4.2a and 4.2b, red dots), consistent with frequent out-of-frame indels. Guide #37 targeting the 5' end of exon 2 yielded only 59%, which may be due to creation of hypomorphic PigA forms with in-frame mutations, as evidenced by intermediate FLAER intensity in some of the transfected cells (Fig. 4.2a).

Notably, guides targeting intronic sites also yielded PigA-deficient cells at significant frequencies. Ten different guides located 263–520 bp from the nearest exon caused 8–20% PigA loss, whereas two guides greater than 2 kb away induced 5–7% loss (Fig. 4.2b, gray and blue dots; Table 4.2), consistent with the mutagenic effect being distance-dependent.

I obtained similar results with transient expression using electroporation or lipofection of ribonucleoprotein complexes (RNP), proving that these observations were not a consequence of PiggyBac transposition, delivery method, antibiotic selection or cellular response to transfected plasmid DNA (Fig. 4.2c). While rates of PigA loss induced by intronic gRNAs #10 and #15 were nearly identical to those obtained by PiggyBac method, exonic gRNAs #48 and #56 were much less efficient. The difference was likely caused by slower cutting and mutagenic repair dynamics of the chosen exonic gRNAs combined with the fact their time of action is limited when using RNP. Consistently, in PiggyBac experiments the fraction of PigA-deficient cells plateaued earlier (on day 7) when using most intronic compared to exonic gRNAs (Fig. 4.2b).

To investigate whether loss of PigA expression upon intronic mutagenesis is an intrinsic property of undifferentiated mouse ES cells, I repeated the experiments in a human female differentiated cell line, RPE1 (Fig. 4.3a). I expected similar results, since RPE1 is functionally hemizygous at the *PIGA* locus, due to X inactivation. Complete ablation of FLAER staining was observed only by day 17 in RPE1 cells, later than in mouse ES cells (data not shown). This may be due to slower proliferation or increased stability of GPI-anchored proteins in this cell line. On day 17, mutagenesis of *PIGA* with all exonic and two intronic gRNAs #274 and #276 (<400 bp away from nearest exon) delivered with PiggyBac vectors resulted in a loss of PIGA at frequencies comparable to those observed in mouse ES cells (86-99% and 8.1-9.4%, respectively; Fig. 4.3b). An intronic guide #277 (>2 kb away from nearest exon) and another intronic guide #275 (<400 bp away) were much less efficient (2.7-3.2%). The exon-proximal gRNA #275 might have been exceptionally inefficient at inducing on-target damage.

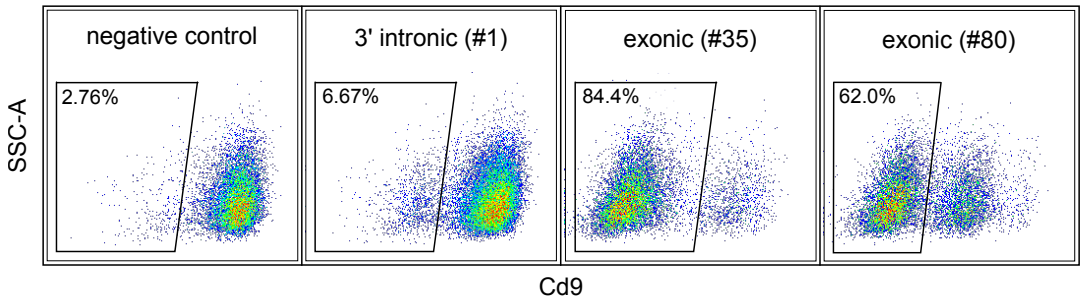
### 4.2.3 Autosomal *Cd9* locus

Given that only one copy of PigA is present in the male mouse ES cells I wished to exclude the possibility that the observations reflect some peculiarity of the lack of a homolog. I considered a number of autosomal genes, which are highly expressed on the surface of mouse ES cells (unpublished data) and whose exonic structure conforms to the conditions outlined above. To be able to distinguish the homologous chromosomes at the genotyping stage, I performed the experiments in mouse ES cells derived from an F1 cross between *Mus musculus* (BL6) and *Mus musculus castaneus* (CAST) mouse strains. Therefore, I was also looking for genes with high degree of divergence between the two mouse strains.

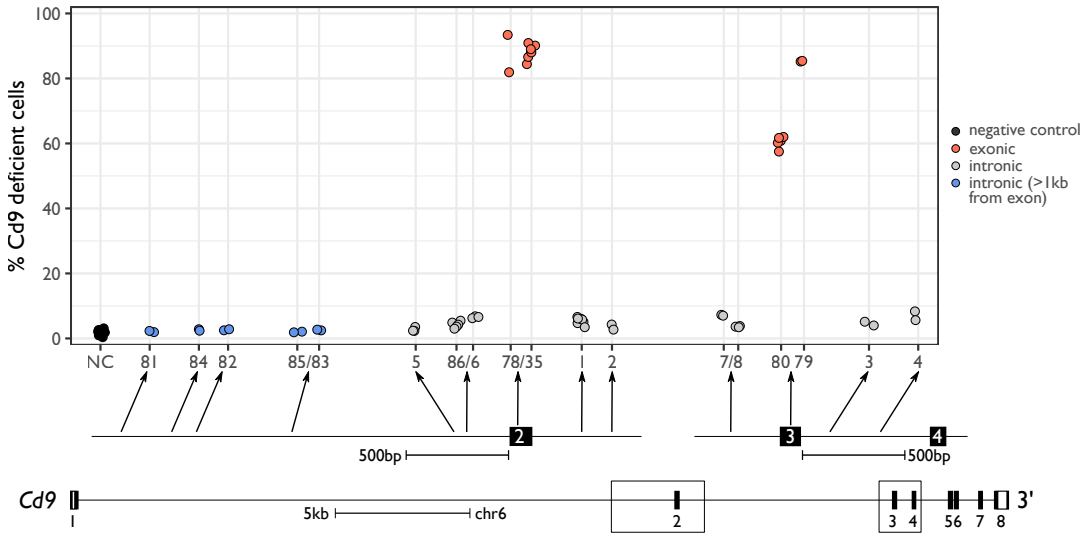
Genes fulfilling these criteria included *Cd9*, *Cd81*, *Itga6* and *Tfrc*. Initial flow cytometric tests confirmed high expression, but revealed sensitivity of Cd9 and Tfrc to differentiating conditions (plating on gelatin without LIF supplementation or dense plating on feeders). Furthermore, Cd81 and to some degree Tfrc proteins were sensitive to trypsinization, when compared to a milder Accutase treatment. Exonic guides abolished Itga6 and Cd9 expression, while Cd81 retained subpopulations with unaffected and intermediate expression. No viability phenotype was observed with these knock-outs, consistent with previous reports (Georges-Labouesse et al., 1996; Le Naour and Boucheix, 2000). Mutagenesis of *Tfrc* led to massive cell death, indicating it is an essential gene for cellular viability of mouse ES cells. This is consistent with evidence of depletion in knock-out CRISPR screens and essential role in mouse development (Blomen et al., 2015; Levy et al., 1999; Wang et al., 2015). I selected *Cd9* instead of *Itga6* for the assay, because of its higher expression and because I interpreted intermediate levels of staining with some guides as evidence that hemizygous populations can be isolated. This has proven to be misleading (see Discussion), but has not substantially influenced the results.

Most exonic guides against *Cd9* delivered with a PiggyBac vector yielded over 80% protein loss. Intronic guides 140-1900 bp away from the nearest exon generated 2.1-7.1% Cd9 loss (Fig. 4.4b; Table 4.2). Taking into account a 1.6% background of Cd9-deficient cells in the untransfected condition, I estimate the true proportion of Cd9 loss due to intronic cutting to be between 0.5–5.5%. This is consistent with results at the *PigA* locus, assuming both *Cd9* alleles have to be destroyed to prevent Cd9 expression. I confirmed that these results were not an artifact of a specific mouse ES cell line by using guides against *Cd9* locus in multiple independently derived lines (Fig. 4.5). Notably, different guides induced different levels of Cd9 loss (Fig. 4.4c and Discussion).

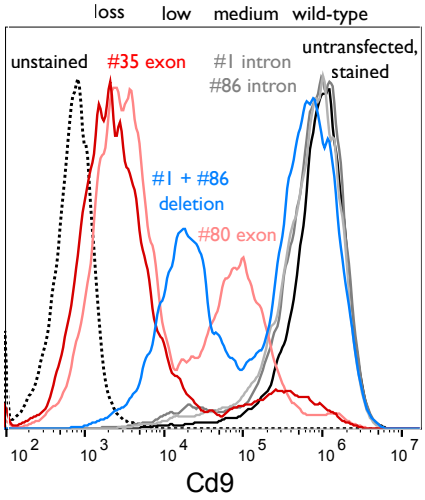
To understand the phenotypic outcomes of Cd9 editing, I isolated single-cell clones mutagenized with different gRNAs and ascertained their expression status by flow cytometry. Most clones retained the Cd9 expression status for which they were sorted. A few clones exhibited bimodal expression pattern (at 9-45% frequency), which may be the result of a mixed clone or



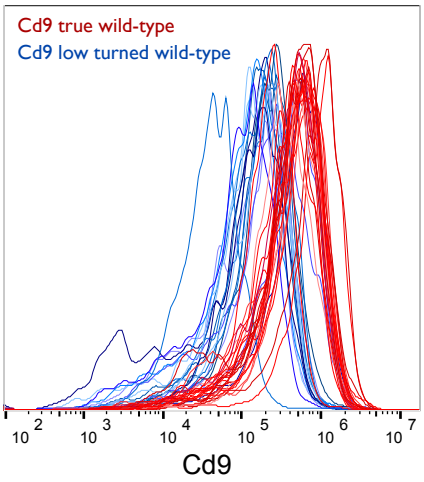

**(a)** Examples of Cd9 mutagenesis. A different gate was used for gRNA #80 (see Discussion).




**(b)** Frequency of Cd9 loss. Circles represent independent cell cultures (N = 2-8). NC: negative control. Boxed exons in the bottom diagram are magnified above.




**(c)** Histogram comparing possible Cd9 mutagenic outcomes. Main outcomes are named on top.




**(d)** Comparison of "wild-type" Cd9 expression between clones sorted for low and wild-type expression.

**Figure 4.4:** Frequency of Cd9 loss upon mutagenesis with exonic and intronic guides in mouse ES cells using PiggyBac vectors. Individual guides are identified by numbers (Table 4.2).

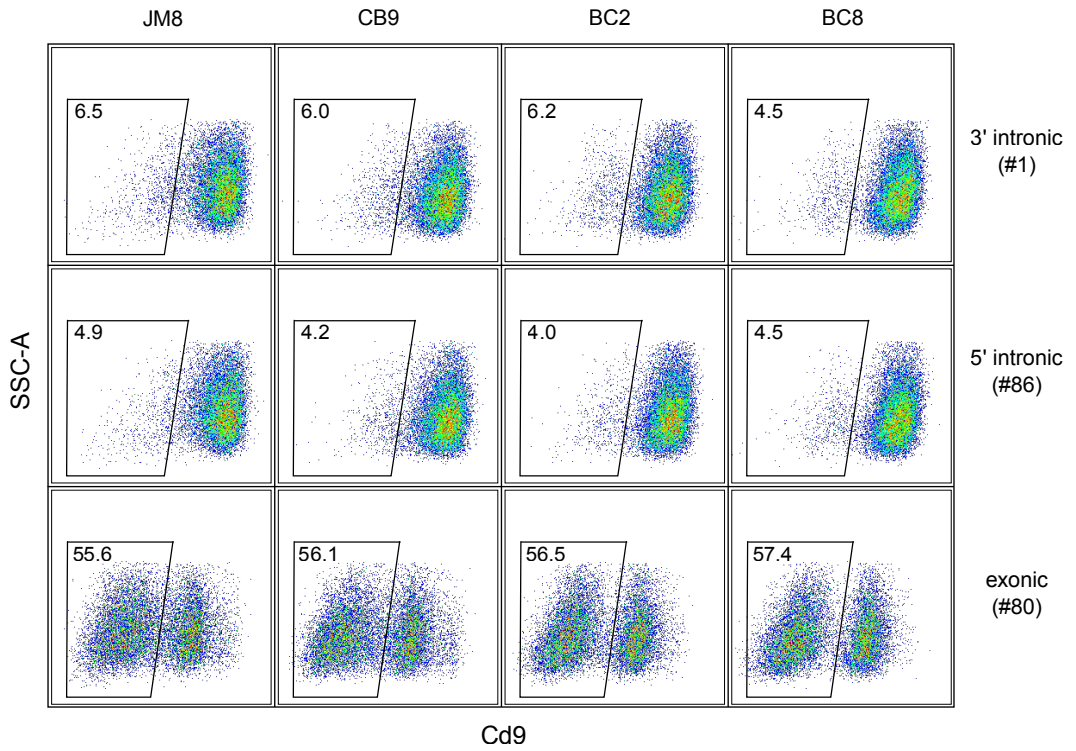



**Figure 4.5:** Mutagenesis in independently derived mouse ES cells lines. Name of the line indicated on top. Individual guides are identified by numbers (Table 4.2). Results shown are representative of three biological replicates. A different gate was used for gRNA #80.

mis-segregation of mutagenized chromatids during clone outgrowth (i.e. Cd9-deficient and Cd9-proficient chromatids segregating into separate cells). Notably, some of the clones derived from the Cd9-deficient population induced by intronic guides were later found to retain, on average, around 50% of wild-type levels of expression (Fig. 4.4d). They likely represent a distinct population found at the high end of the Cd9-deficient sorting gate (see Discussion). No cells mutagenized with the exonic gRNA #35 and sorted for "medium" expression of Cd9 retained that status, ending up as either "loss" or "wild-type" clones. This confirms that "medium" status is a unique population induced only by specific gRNAs (e.g. #80; data not shown).

## 4.3 Discussion

I have set up a simple flow cytometric assay for detection and quantification of complex Cas9-induced genomic lesions. It detected substantial levels of mutagenesis when targeting intronic sites at a hemizygous *PigA* and *PIGA* loci and an autosomal *Cd9* locus. This could be caused by either lesions destroying the nearby exon or ubiquitous presence of strong intronic regulatory elements at all intronic loci tested. The latter seems unlikely, as enhancers are neither ubiquitous, nor do they often have strong phenotypic effects. I investigate these hypotheses by directly genotyping PigA and Cd9-deficient cells in chapter 5.

*PigA*, *PIGA* and *Cd9* were actively transcribed. Outcomes could be different at inactive loci, if transcription or chromatin structure interferes with Cas9 activity or DNA repair. Low chromatin accessibility has been shown to impede Cas9 binding and lower editing efficiency (Horlbeck et al., 2016; Uusi-Mäkelä et al., 2018). Since in my assay both Cas9 and gRNA are constitutively active and since SpCas9 can functionally open the chromatin (Barkal et al., 2016; Polstein

**Table 4.1:** Non wild-type Cd9 expression levels.

| Level | gRNAs | Putative cause |
|---|---|---|
| loss | all exonic | out-of-frame mutation and NMD |
| low | single and paired intronic | alternative splicing due to exon skipping |
| medium | exonic #80 and #53 | in-frame mutation of exon 3 |
| l-wt | all intronic | monoallelic mutation |

**l-wt:** low turned wild-type. #53 is a Cpf1 guide targeting exon 3 (data not shown).

et al., 2015), its structure at inactive loci should not make much difference. When Cas9 is bound to the non-transcribed strand, it blocks DNA damage repair proteins from accessing the break and prevents formation of indels at 48 hours post-delivery (Clarke et al., 2018). It is not clear how the damage in those cells is eventually resolved. There does not seem to be any clear difference between intronic gRNAs targeting the template strand and non-template strand in term of frequency of PigA loss, but the dataset is not well balanced with respect to strandedness (Table 4.2). A direct experiment at a non-transcribed or temporarily silenced locus may be the only way to resolve this issue.

Lesions at the *PigA* locus resulted in either complete ablation of PigA expression or left the PigA expression unaffected (except those induced by gRNA #37, as Discussed in results section). In contrast, mutagenesis of *Cd9* locus had three distinct, non wild-type phenotypic outcomes separated by at least an order of magnitude fluorescent intensity, termed "loss", "low" and "medium" in Fig. 4.4c and Table 4.1. This variation in expression level suggests different underlying genotypes. A "negative" population was induced by three different gRNAs against exons 2 and 3 and may have resulted from out-of-frame indels triggering NMD and complete loss of protein expression. "Low" was only seen with single intronic gRNAs and deletions induced by a pair of intronic gRNAs flanking an exon. It may represent an alternative TSS or splice form, which "buffers" against complete loss of an "out-of-frame" exon 2 or 3. "Medium" expression was only observed with two specific gRNAs targeting exon 3 (incl. one Cpf1 gRNA, data not shown), which also induced a "negative" population. This "medium" state may result from an in-frame mutations that decreases the protein affinity for the antibody. This set of hypotheses can be tested by profiling local indels and RNA transcripts in cell populations sorted for their Cd9 expression level. If it is true, targeting different parts of *Cd9* would allow quantification and isolation of specific classes of genomic lesions.

Some cells edited with intronic gRNAs and sorted for low Cd9 expression were found to express near wild-type levels of Cd9 after clone outgrowth. These "low turned wild-type" clones could stem from the "background" $Cd9^{low}$ population observed in the negative control (Fig. 4.4b). Such population would be partially differentiated due to prolonged culture on gelatin with LIF supplementation. However, in a control experiment using gRNA against an irrelevant locus only about 3% of the expected number of such $Cd9^{low}$ cells formed colonies (all retaining wild-type expression; data not shown). Therefore, they would not have contributed substantially to the "low turned wild-type" population observed here. Observation that these clones express on average 50% less Cd9 than "true wild-type" clones indicates that they may represent a hemizygous population (see chapter 5) If this is the case, then exonic gRNAs should also induce a similar population.

**Table 4.2:** Flow cytometry results and gRNA sequences.

| Experiment | Guide | Sequence with PAM | Chr | Cutting position | Strand | Type | Distance from the nearest exon | %Expression deficient (mean) | %Expression deficient (sd) | N | %Expression deficient (adj.mean) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PigA | 5 | GCAGTGAAGATAAATCACAAGGG | 6 | 125472779 | T | NC | - | 0.4 | 0.3 | 6 | - |
| PigA | 15 | CGTTGTGTCACAGTGCATAATGG | X | 164422321 | NT | intronic | 260 | 12 | 3.3 | 6 | - |
| PigA | 62 | TGTGACACAACGTTTAAAAGTGG | X | 164422349 | T | intronic | 238 | 14 | 2.1 | 6 | - |
| PigA | 16 | GAACATCTACTTGCTTAGCAGGG | X | 164422416 | NT | intronic | 165 | 12 | 0.7 | 6 | - |
| PigA | 37 | AAGGTTTCCAGAGCTACCCGGGG | X | 164422701 | NT | exonic | - | 60 | 2.0 | 6 | - |
| PigA | 56 | GCAGAGAAAGAACTGTGGGAATGG | X | 164423023 | NT | exonic | - | 97 | 1.0 | 6 | - |
| PigA | 10 | AGGAAGCCATAAGATAGCCACGG | X | 164423864 | NT | intronic | 503 | 14 | 2.2 | 6 | - |
| PigA | 11 | GCATAAGAGTGGATAAAACCAGG | X | 164423884 | NT | intronic | 523 | 9.7 | 2.6 | 6 | - |
| PigA | 74 | TGAGGTACTGTACCATGCACAGG | X | 164425741 | NT | intronic | 2188 | 7.1 | 0.7 | 6 | - |
| PigA | 75 | GAGGGTAAGTAACTCGCCCAAGG | X | 164425844 | T | intronic | 2091 | 6.4 | 1.6 | 6 | - |
| PigA | 17 | ACTTGTTCATACAGCCTACGTGG | X | 164427667 | NT | intronic | 262 | 13 | 1.7 | 6 | - |
| PigA | 18 | GATATGGGTATGTGGCAGTAGCGG | X | 164427749 | T | intronic | 186 | 13 | 1.2 | 6 | - |
| PigA | 59 | GGGACCAAAGAGAATCATTTTGG | X | 164428028 | T | exonic | - | 88 | 1.8 | 6 | - |
| PigA | 63 | TGCCTCTTATAAATTGAAGCAGG | X | 164428148 | NT | intronic | 86 | 8.9 | 1.8 | 6 | - |
| PigA | 12 | ATAAGAGGCATGCAAATAGAAGG | X | 164428178 | T | intronic | 110 | 13 | 1.6 | 6 | - |
| PigA | 13 | CATACGAGCTGTGACACAACAGG | X | 164428347 | T | intronic | 196 | 21 | 1.6 | 6 | - |
| PigA | 14 | AAGTTGTGTCTATTACACTGGGG | X | 164428376 | T | intronic | 167 | 16 | 2.2 | 6 | - |
| PigA | 48 | ATGCAGAACGCTTCAGTGAGGG | X | 164428620 | NT | exonic | - | 96 | 1.3 | 6 | - |
| Cd9 | NC | [untransfected or edited at *PigA*] | - | - | - | NC | - | 1.6 | 1.0 | 8 | 0 |
| Cd9 | 81 | GTGCAGAGCACGCCTTCACGGGG | 6 | 125474376 | NT | intronic | 1865 | 2.1 | - | 2 | 0.5 |
| Cd9 | 84 | GCAGTGTCCTTATCTAAGAGGGG | 6 | 125474156 | T | intronic | 1645 | 2.6 | - | 2 | 1.0 |
| Cd9 | 82 | ATGTAAGCCCCCTTAGTCCCCGG | 6 | 125474028 | T | intronic | 1517 | 2.7 | - | 2 | 1.1 |
| Cd9 | 85 | CAGCCAGCCACTACACTGGAGGG | 6 | 125473582 | T | intronic | 1071 | 2.0 | - | 2 | 0.4 |
| Cd9 | 83 | ACCTCTTATCACTGGTACCAGGG | 6 | 125473547 | NT | intronic | 1036 | 2.6 | - | 2 | 1.0 |
| Cd9 | 5 | GCAGTGAAGATAAATCACAAGGG | 6 | 125472775 | NT | intronic | 264 | 2.7 | 0.7 | 3 | 1.1 |
| Cd9 | 86 | CAACTGCAGCACTTCCGGCAGGG | 6 | 125472720 | T | intronic | 209 | 4.2 | 1.0 | 5 | 2.6 |
| Cd9 | 6 | GATTCACACACAGTTCCCTGCCGG | 6 | 125472717 | NT | intronic | 206 | 6.5 | 0.3 | 3 | 4.9 |
| Cd9 | 78 | CAGTGCTTGCTATTGGACTATGG | 6 | 125472481 | T | exonic | - | 88 | - | 2 | 86 |
| Cd9 | 35 | TCTTGGTCTGAGAGTCGAATCGG | 6 | 125472467 | NT | exonic | - | 88 | 2.2 | 7 | 87 |
| Cd9 | 1 | AAGGATGCCACCACTCCTGAGGG | 6 | 125472162 | T | intronic | 246 | 5.4 | 1.1 | 7 | 3.8 |
| Cd9 | 2 | ATTCAGGAAGCCGGTCTGGAGGG | 6 | 125472028 | NT | intronic | 380 | 3.5 | - | 2 | 1.9 |
| Cd9 | 7 | GGTTGTCCCCTAAGCATCAAGGG | 6 | 125464747 | T | intronic | 260 | 7.1 | - | 2 | 5.5 |
| Cd9 | 8 | TCAACACTCTACCTCATCCTCGG | 6 | 125464703 | NT | intronic | 216 | 3.6 | 0.2 | 3 | 2.0 |
| Cd9 | 80 | AGCCGGGGCCCTCATGATGCTGG | 6 | 125464449 | T | exonic | - | 60 | 1.8 | 5 | 59 |
| Cd9 | 79 | GTACAGCTCCACAGCAGCCCAGG | 6 | 125464432 | NT | exonic | - | 85 | - | 2 | 84 |
| Cd9 | 3 | GCCTGAAGTAAGGATGGTGAGAGG | 6 | 125464252 | NT | intronic | 137 | 4.6 | - | 2 | 3.0 |
| Cd9 | 4 | CTTTGTTTCCCCGATCTCGGTGG | 6 | 125464003 | T | intronic | 386 | 7.0 | - | 2 | 5.4 |
| progenitor | 311 | GGGCGAGGAGCTGTTCACCGGGG | - | - | T | exonic | - | - | - | - | - |
| cherry/gfp | 33 | GAAGTTCGAGGGCGACACCCTGG | - | - | T | exonic | - | - | - | - | - |
| cherry/gfp | 34 | GGAACAGTACGAACGCGCCGAGG | - | - | T | exonic | - | - | - | - | - |
| RPE1 | 231 | AGGCTTCCCGCATTCAAAATCGG | 3 | 46371987 | NT | NC | - | 0.2 | 0.0 | 3 | - |
| RPE1 | 308 | GTTGTAAGTACCAGAGTTGGTGG | X | 15324855 | T | exonic | - | 99 | 0.3 | 3 | - |
| RPE1 | 307 | TTTGGGAGCTTTAGAACACAAGG | X | 15325127 | T | exonic | - | 91 | 0.8 | 3 | - |
| RPE1 | 278 | GGATATTTCTGACAGAGTTCAGG | X | 15326014 | NT | exonic | - | 95 | 0.3 | 3 | - |
| RPE1 | 277 | GAATGTCTTAAGTGAGAGAGAGG | X | 15328488 | T | intronic | 2278 | 2.7 | 0.7 | 3 | - |
| RPE1 | 276 | AGAGGGCAGGCCGTGTACGGTGG | X | 15330896 | T | intronic | 323 | 8.1 | 2.1 | 3 | - |
| RPE1 | 273 | TGCTCAGGTACATATTTGTTCGG | X | 15331580 | T | exonic | - | 99 | 0.3 | 3 | - |
| RPE1 | 272 | GTAATAGACTTTGAGGCCACTGG | X | 15331674 | NT | exonic | - | 86 | 2.3 | 3 | - |
| RPE1 | 274 | TGGTAAACCATGATATGCTGTGG | X | 15332254 | T | intronic | 261 | 9.4 | 0.6 | 3 | - |
| RPE1 | 275 | GGTAAAGTATAAGAGTAAAGGGG | X | 15332346 | T | intronic | 353 | 3.2 | 1.6 | 3 | - |

Genomic position is given with respect to the GRCm38 or GRCh38 (RPE1 experiment) reference genome. Last column contains negative control subtracted mean (Cd9 experiment). SD = standard deviation. Strand: T = transcribed, NT = non-transcribed.

# Chapter 5

# Genotyping of complex lesions

## 5.1 Introduction

Most PCR-based genotyping of Cas9-induced lesions has so far focused on the region immediately adjacent to the cut site (<1000 bp) in bulk cell populations. This biases the assessment by excluding lesions that destroy the primer binding sites (large deletions), disconnect them (translocation and large inversions) or prevent amplification by increasing the distance between them (large insertions). Cas9-induced lesions that are non-contiguous with the cut site and outside of the amplified region are also missed. Failure to recover such complex alleles is not apparent, when genotyping in bulk cell populations. While some specialized, PCR-based methods for detection of such lesions in bulk populations exist, they are not broadly used. PCR employed by these methods also has to be anchored in one flank of of the break, which biases the output (Cain-Hom et al., 2017; de Vree et al., 2014; Giannoukos et al., 2018; Zheng et al., 2014).

Here, I addressed some of these issues by combining long-range PCR with Sanger and PacBio sequencing to detect and describe complex Cas9-induced lesions in an unbiased way.

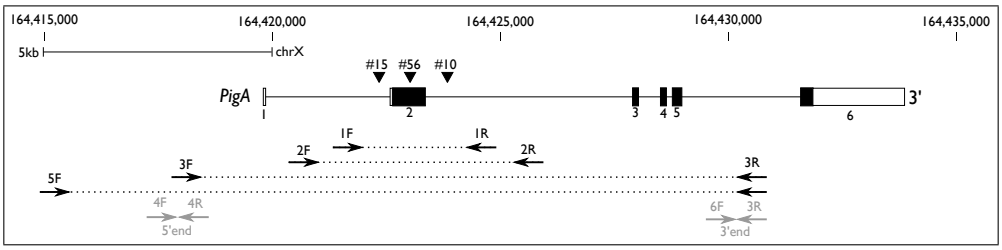
## 5.2 Results

### 5.2.1 Deletions underlying loss of gene expression caused by intronic gRNAs

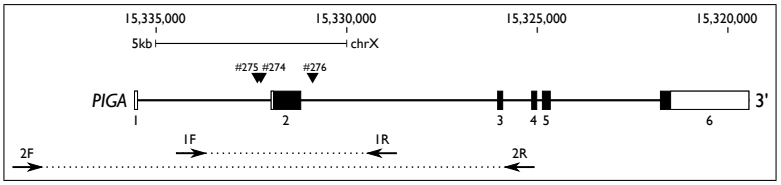
In chapter 4, I showed that individual, intronic gRNAs can cause loss of expression of the chromosome X linked PigA in about 12% of transfected male mouse ES cells. To understand what genetic changes underlie this phenotype, I amplified a 5.7 kb region around exon 2 from pools of cells mutagenized with three selected gRNAs introduced by PiggyBac transposition and sequenced the PCR products using the PacBio platform. I observed a depletion in read coverage on a kilobase-scale around the cut sites, consistent with the presence of large deletions (Fig. 5.2). Cells mutagenized with intronic guides and sorted for loss of PigA generally exhibited loss of the adjacent exon 2 (Fig. 5.1a). If intronic regulatory sequences were present around the exon, the DNA of cells sorted for retention of PigA expression would be wild type or contain only small indels around the cut site. However, the most frequent lesions in these cells were kilobase-scale deletions extending away from the exon. I conclude that, in most cases, loss of PigA expression was likely caused by loss of the exon, rather than damage to intronic regulatory elements.

PacBio sequencing of pooled edited DNA is biased towards detection of large deletions. PCR is more likely to amplify shorter amplicons, favoring deletions. Capture of short fragments is also more efficient during PacBio sequencing. Finally, individual, shorter DNA molecules are usually read more times during sequencing than longer ones. As a consequence, they have higher quality scores and are more likely to pass quality filters. Another disadvantage of the PacBio approach is the need to choose the amplicon size beforehand, which means some alleles with larger lesions may be missed. Finally, translocations cannot be amplified by a pair of fixed primers at all.

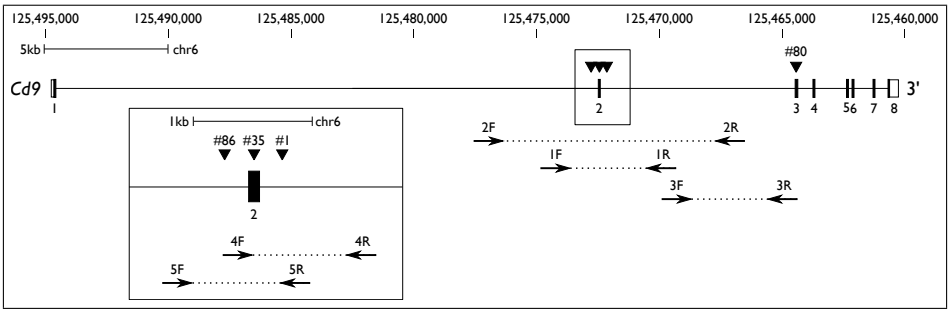
Therefore, to fully characterize the variety of mutagenized alleles, I isolated single cell clones. The loci around the gRNA target site were ampli-



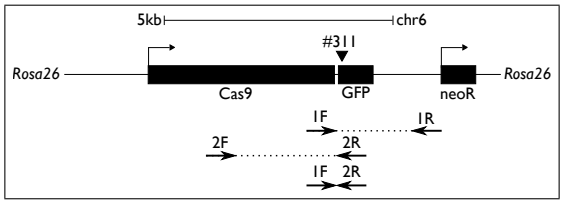

**(a)** *PigA* locus.




**(b)** *PIGA* locus.




**(c)** *Cd9* locus.




**(d)** *Cas9-GFP* locus, bone marrow progenitors experiment.


**(e)** Alleles recovered by Sanger sequencing from mouse bone marrow progenitor cells mutagenized at the *Cas9-GFP* locus. The position of the gRNA #311 is shown as a vertical line intersecting with the gene structure. Horizontal line indicates deletion.

**Figure 5.1:** Positions of primer pairs and gRNAs used for genotyping experiments (Table 5.1 and 5.3). Genomic position is given with respect to the GRCm38 or GRCh38 reference genome. In grey, primers used for diagnostic PCRs on alleles that could not be recovered.

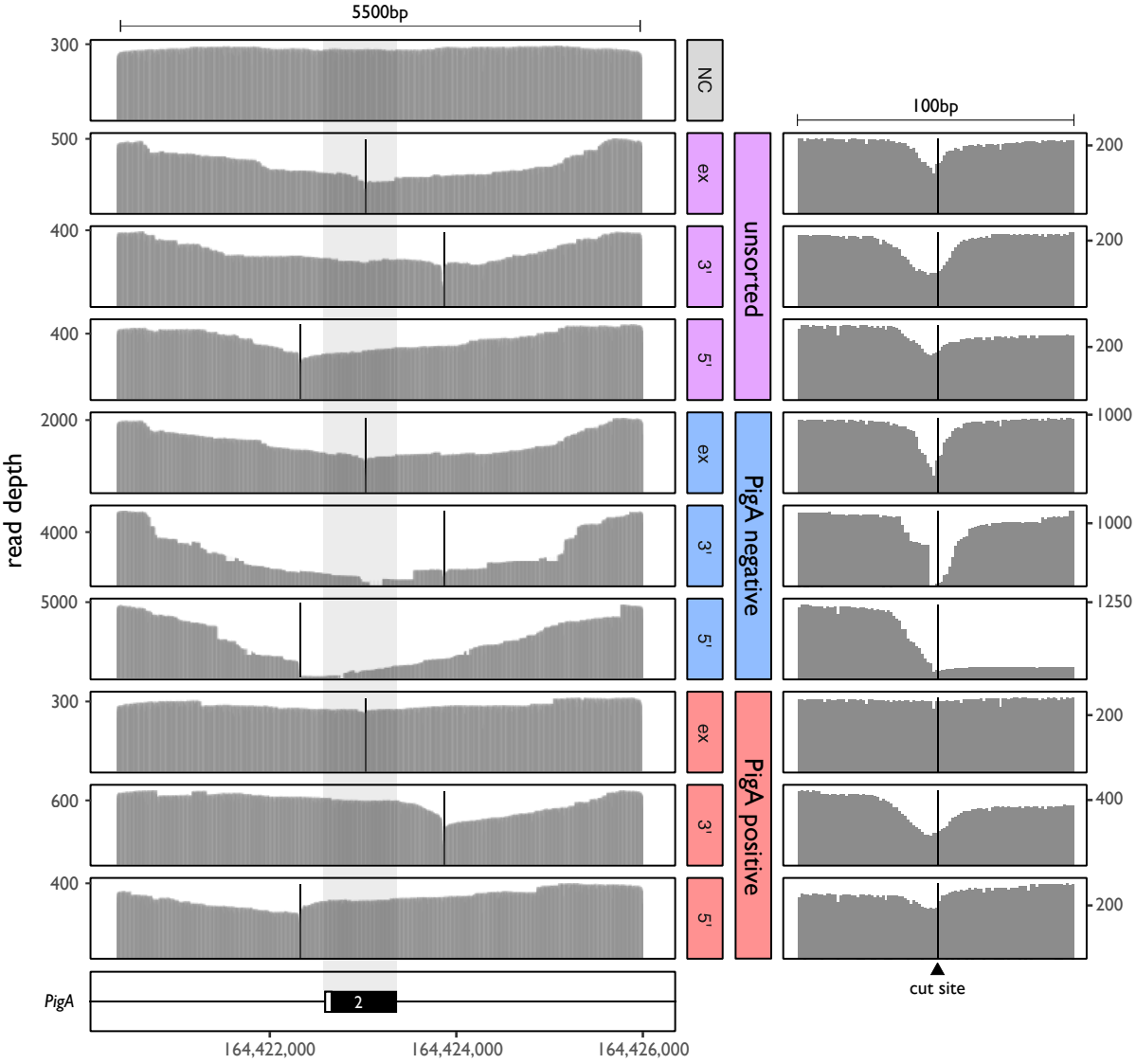
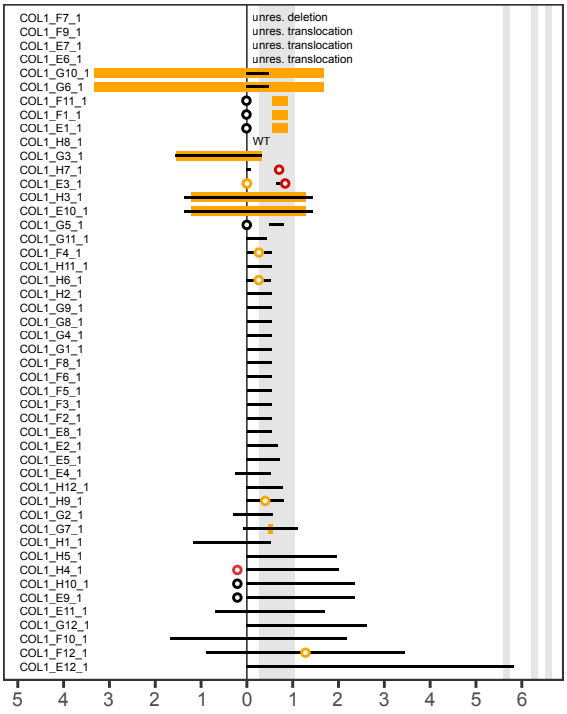
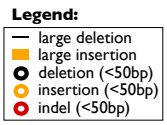
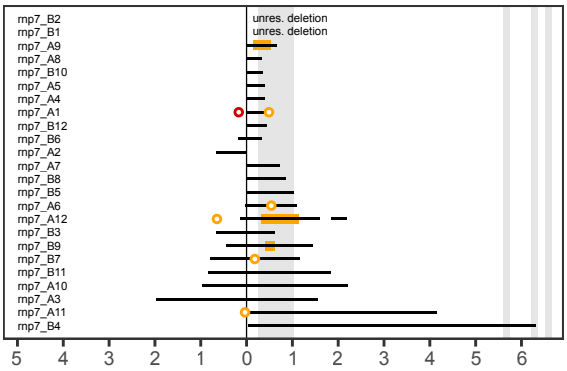



**Figure 5.2:** Analysis of the *PigA* locus mutagenized with selected gRNAs. Coverage of PacBio reads at the *PigA* locus. The locus was PCR-amplified from a pool of cells sorted for PigA expression (or from the unsorted population), and the resulting products were sequenced using the PacBio platform. The right panel depicts a 100 bp region centered at the cut site. NC: negative-control gRNA, ex: exonic gRNA (#56), 5' : 5' intronic gRNA (#15), 3' : 3' intronic gRNA (#10). The cut site of the gRNA is indicated with a vertical black bar. Genomic position is given with respect to the GRCm38 reference genome. N = 1.

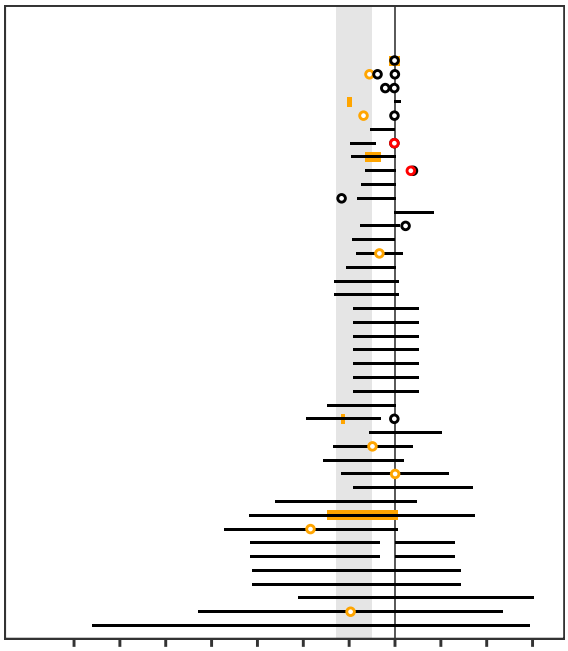



**(a)** 5' intronic guide #15, PiggyBac method.

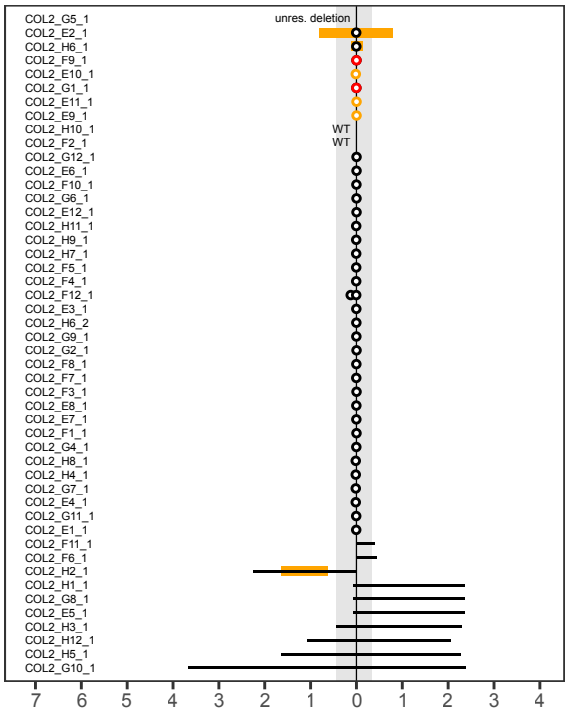



**(b)** 5' intronic guide #15, RNP method.

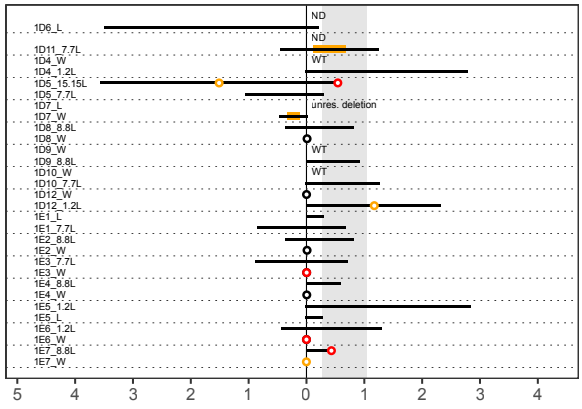


**(c)** 3' intronic guide #10, PiggyBac method.

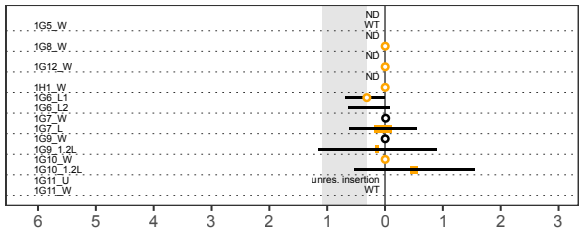



**(d)** Exonic guide #56, PiggyBac method.

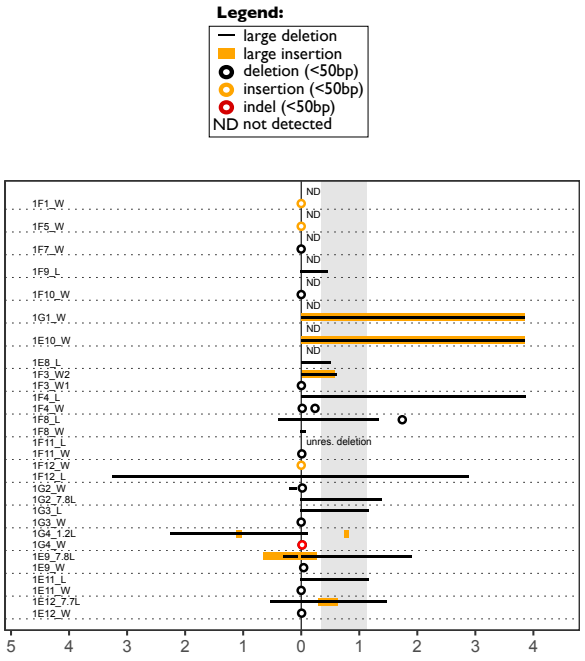
**Figure 5.3:** Alleles recovered by Sanger sequencing from Cas9-edited, PigA-deficient mouse ES cell clones. The position of the gRNA is shown as a vertical line. Pure insertions and deletions of <50 bp are indicated with orange and black circles, respectively. Combined insertion/deletion events of <50 bp and SNPs ("indel (<50 bp)" in the legend) are indicated with a red circle. Black lines represent deletions >50 bp. Orange bars indicate size of the >50 bp insertions (but not their map position). They are centered on the insertion locus or on the associated deletion. Gray shades represent exons 2 (large one), 3, 4 and 5. X-axis represents distance from the gRNA position in kilobases. Alleles are sorted by total length. Their names are indicated on the left.




**(a)** 5’ intronic guide #274.




**(b)** 3’ intronic guide #276.




**(c)** 5’ intronic guide #275.

**Figure 5.4:** Alleles recovered by Sanger sequencing from PIGA-deficient, human female RPE1 cell clones. Cas9-expressing cells were transiently transfected with gRNA-expressing plasmids. Grey shade represents exon 2. Names of individual alleles are indicated in the column on the left. Dotted horizontal line separates clones. Other display conventions as in Fig. 5.3.

fied using PCR primer pairs positioned progressively further apart (up to 12-16 kb), until amplicons were generated (see Fig. 5.1 for primers and gRNAs positions). Long elongation times were used to identify large insertions. Since mouse ES cells were grown on feeder cells (which help maintain their pluripotency), primer pairs which specifically exclude feeder cell DNA were used to avoid spurious wild-type alleles at the *PigA* locus. This could not be achieved at the *Cd9* locus due to low divergence between BL6 and feeder genomes. Amplicons were Sanger (all loci) or PacBio (only *Cd9* locus) sequenced. Since no wild-type CAST alleles were detected in any of the clones edited at the *Cd9* locus, I assumed all wild-type BL6 alleles in these clones were feeder derived.

Consistent with the results from PacBio sequencing, large deletions of >50 bp were detected in almost 85% (79/93) of PigA-deficient single cell clones generated by single, intronic gRNAs #10 and #15 (Fig. 5.3a and Fig. 5.3c). Most of them overlapped both the cut site and the nearest exon. The deletions varied in size, the largest spanning 9.5 kb. Identical results were obtained using electroporation of intronic gRNA #15 as RNP (Fig. 5.3b), as expected due to consistent rates of PigA-deficient cells between PiggyBac and RNP methods (chapter 4).

To assess the frequency of large deletions without strong selection for that outcome, I used the exonic guide #56 causing 97% PigA loss. Although two-thirds of alleles (32/48) from PigA-deficient cells had indels <50 bp, as expected, 20% (10/48) had deletions >50 bp, extending up to 6 kb. Some of the deletions exhibited clear directionality. Assuming this is also the case for deletions

induced by intronic guides, this explains why the observed rate of PigA loss with those guides is only ~12% (chapter 4), not 20%. It is also consistent with the depletion of read coverage distal to the exon in PacBio analysis (Fig. 5.2).

To replicate these results in another mouse ES cell line, I genotyped AB2.2 ES cell clones mutagenized at a hemizygous *GFP* or *mCherry* targeted transgene using the PiggyBac method. The lower frequency of deletions in these cells (7-12% vs the expected 20%, Table 5.1, "cherry/gfp" experiment)) is likely due to relatively shorter range of the PCR (<3 kb vs 16 kb). Consistently, no amplicon at all could be obtained in 15-26% of clones.

In chapter 4, I have shown that intronic gRNAs cause similar levels of PigA loss in both mouse ES cells and in human female differentiated RPE1 cell line. Consequently, I expected the rate of deletions in these cells to also be similar. I mutagenized RPE1 cells with intronic guides at the *PIGA* gene, isolated PIGA-deficient single cell clones and resolved their alleles using long-range PCR (up to 12 kb) and Sanger sequencing. Only deletions on the active chromosome X would be selected for in these cells, so I expected the rate of deletions on this chromosome to approximate 85% (as in PigA-deficient male ES cells mutagenized with intronic guides). Conversely, selection should not affect the inactive chromosome, resulting in the unselected rate of 20% (as in ES cells mutagenized with the exonic guide). The observed frequency of deletions in RPE1 cells is 47% (40/85), very close to the expected 51% (85%*0.5 + 20%*0.5; Fig. 5.4a, 5.4c and 5.4b). The largest deletion spanned 6 kb. All but four deletions overlapped both the cut site and the nearest exon.

Frequent deletions were also observed in mouse ES cells edited with intronic gRNA #1 at the bi-allelic *Cd9* locus. The rate of deletions was highest in the $Cd9^{low}$ (42/43) and lowest in "true wild-type" clones (17/55). Intermediate levels of deletions were observed in bimodal and "low turned wild-type" clones expressing intermediate levels of Cd9 (as defined in chapter 4, Fig. 5.5). See subsection 5.2.5 for a more in-depth discussion taking into account the allelic composition of individual clones.

To show that large deletions at the *Cd9* locus occur regardless of the choice of gRNA, I mutagenized the biallelic *Cd9* locus using two single intronic guides (#1 and #86; two replicates) and two single exonic guides (#35 and #80; one replicate), sorted for cells expressing different Cd9 levels, reassessed the expression of isolated clones and genotyped the clones using long-range PCR (Table 5.1, "cbbcs1" and "cbbcs3" experiments, Fig. 5.1c). In all examined groups a substantial fraction of clones had at least one deletion (18-88%). In particular, cells mutagenized with intronic guides and sorted for loss of gene expression collectively exhibit higher rates of deletion clones (50-88%) than cells sorted for retention of gene expression (33-46%). "Low turned wild-type" and bimodal clones exhibited an intermediate frequency of deletions (43-71%). In clones mutagenized with exonic gRNAs a large deletion is not necessary to ablate gene expression, as a small indel would have the same effect. Consequently, there was not clear correlation between clone expression level and fraction of deletion clones (range: 17-42%).

### 5.2.2 Deletions in primary bone marrow cells

Mouse ES cells can maintain pluripotency in culture for many passages. However, culture conditions could temporarily influence the DDR in these cells. Therefore, I replicated my results in primary cells. I chose to work with progenitor cells from the bone marrow of mice expressing Cas9-GFP from a transgene at the homozygous *Rosa26* locus. Lineage-negative cells enriched by removal of differentiated cells on magnetic columns were electroporated with a crRNA:trRNA complex against the *GFP* locus. GFP-negative single cell clones were isolated and genotyped around the cut site with three different primer pairs spanning in total 5 kb (Fig. 5.1d). At least one large deletion product between 100 bp and ~3 kb in size was detected in 36% of clones (35/96; Table 5.1, "progenitor" experiment). I veri-

fied eight deletion products by Sanger sequencing across the deletion junction (Fig. 5.1e). Only wild-type-size products were detected in the remaining clones and none of the 96 control clones exhibited any deletion bands (data not shown).

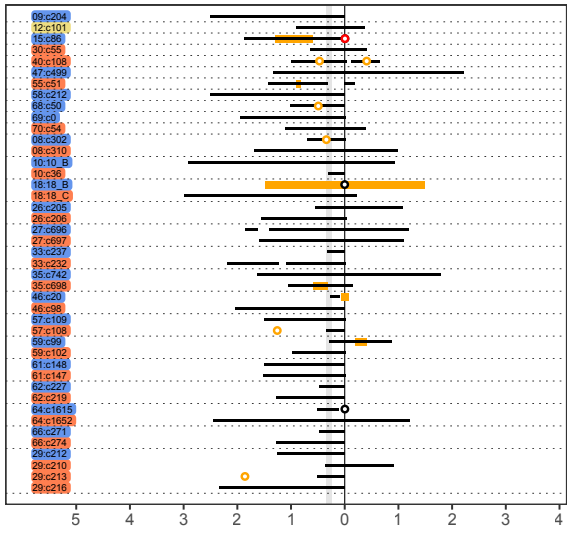
Observed frequency of deletions per clone was identical to the expected rate (36%), given 20% probability of each allele sustaining a deletion (as at the hemizygous *PigA* locus in mouse ES cells mutagenized with the exonic guide). However, the range covered by genotyping PCR was much smaller than in ES cells (5 kb vs 16 kb). While this result confirms large deletions are common, it also suggests that the real frequency is higher than expected, which may be a locus or cell-specific difference.

I have attempted to test this hypothesis by replicating the results at the *Cd45* locus (*Ptprc*) in an F1 cross combining two isoforms (*Cd45.1* and *Cd45.2*), whose expression can be distinguished through specific antibody staining. Initial experiments indicated that double knock-out is lethal in progenitor cells, as this population appeared only transiently in culture and did not form colonies when isolated by FACS (data not shown). Targeting *PigA* in progenitor cells failed to result in ablation of FLAER staining by day 11 post-delivery, at which point the experiment was terminated. It is possible that the effect would have been observed later. I decided not to target the *Cd9* locus, as without near 100% electroporation efficiency loss of Cd9 expression induced by intronic gRNAs would be very low. Furthermore, re-assessment of Cd9 expression status in outgrown colonies would not be possible using flow cytometry due to very low cell numbers. An immunofluorescence procedure could have been developed for that purpose.

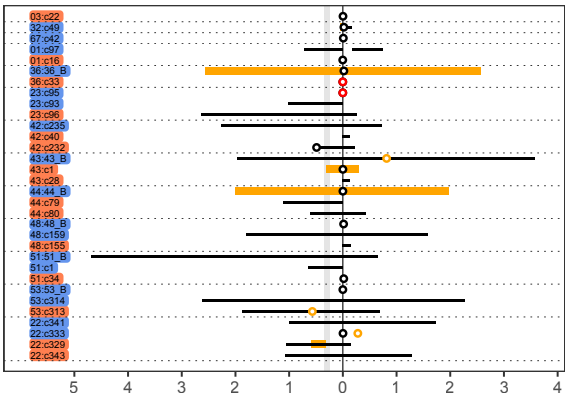
### 5.2.3 Insertions

Insertions (incl. duplications and inversions), defined as $\geq$10 bp fragments, which did not map in a linear fashion to the mutagenized locus, were present in 7-29% of resolved alleles from *PigA*, *PIGA* and *Cd9* loci. In almost all samples the most common origin of insertions was the edited locus (~62% of all insertions). This category ranged in size from small duplications <20 bp templated right next to the deletion breakpoint to perfect inversions of 3.9 kb (Fig. 5.4c). Fragments of *E.coli* genomic DNA and transfected plasmids up to 5 kb were found at all three examined loci, regardless of whether transfection was transient or involved mobilization of the PiggyBac transposon (Fig. 5.8). Distal insertions from introns and repetitive elements (predominantly LINE) were also present in a few samples. Notably, identity of four insertions of 13-29 bp could not be established, suggesting one non-templated or a few stitched, short, templated insertions.

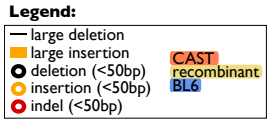
One of the alleles derived from PacBio sequencing of the edited *PigA* locus contained an insertion with a perfect match to four consecutive exons derived from the *Hmgn1* gene (Fig. 5.8a). It could represent a de novo insertion from the spliced and reverse-transcribed RNA, rather than from one of the pseudogenized forms of *Hmgn1*, as the pseudogenes diverge in sequence from the functional gene and thus from the observed insertion.

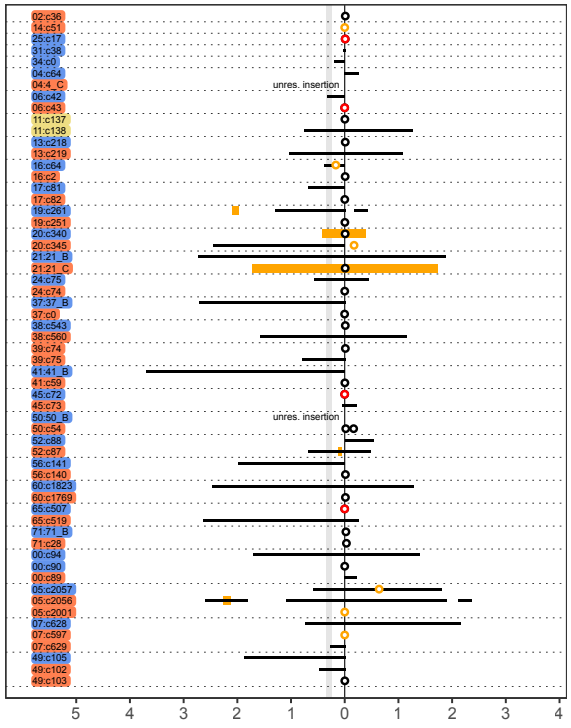


**(a)** Cd9$^{low}$ clones.

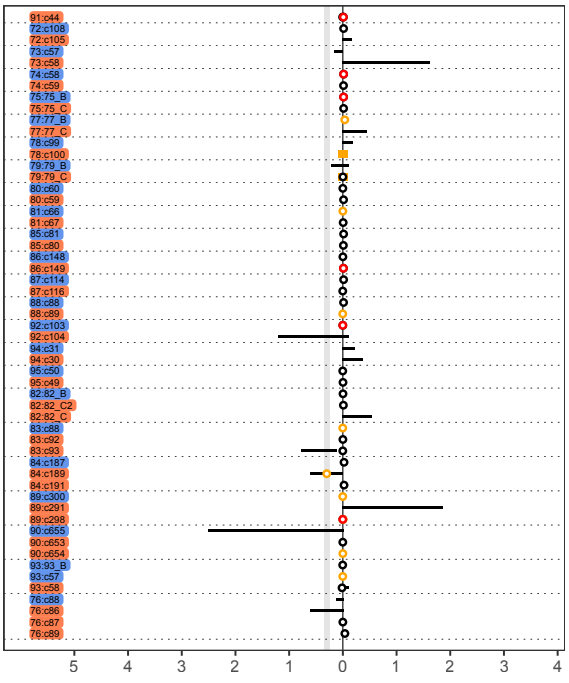


**(c)** Bimodal (mixed) clones.




**(b)** Wild-type clones, that were sorted for low expression ("low turned wild-type").



**(d)** Wild-type clones, that were sorted for wild-type expression ("true wild-type").

**Figure 5.5:** Alleles recovered by Sanger and PacBio sequencing from CAST/BL6 mouse ES cell clones mutagenized at the *Cd9* locus with the 3' intronic gRNA #1. PiggyBac constructs were stably delivered by transposition into Cas9-expressing cells. Gray shade represents exon 2. Dotted horizontal line separates clones. Clones are sorted by the number of alleles. Color behind allele names indicates strain of origin. Other display conventions as in Fig. 5.3.

**Table 5.1:** Results of PCR genotyping.

| Experiment | gRNA | Gene | Primer pairs | Amplicon size [bp] | Target region | Sorted population | Clone expression | Total clones | ≥1 del. | ≥1 ins. | No amp. | % del. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cbbcs1 | 1 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | intron | wt | wt | 24 | 9 | 2 | 0 | 38% |
| cbbcs1 | 1 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | intron | low | wt | 14 | 6 | 3 | 0 | 43% |
| cbbcs1 | 1 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | intron | low | low | 10 | 5 | 0 | 1 | 50% |
| cbbcs1 | 35 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | exon | wt | wt | 24 | 7 | 1 | 0 | 29% |
| cbbcs1 | 35 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | exon | medium | wt | 20 | 7 | 0 | 0 | 35% |
| cbbcs1 | 35 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | exon | medium | loss | 4 | 2 | 0 | 0 | 50% |
| cbbcs1 | 35 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | exon | loss | loss | 24 | 4 | 3 | 1 | 17% |
| cbbcs1 | 80 | Cd9 | 6F/6R, 3F/3R | 1266, 5865 | exon | medium | medium | 24 | 7 | 0 | 0 | 29% |
| cbbcs1 | 80 | Cd9 | 6F/6R, 3F/3R | 1266, 5865 | exon | loss | loss | 24 | 10 | 1 | 0 | 42% |
| cbbcs1 | 86 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | intron | wt | wt | 24 | 8 | 1 | 0 | 33% |
| cbbcs1 | 86 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | intron | low | wt | 2 | 1 | 0 | 0 | 50% |
| cbbcs1 | 86 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | intron | low | low | 22 | 17 | 1 | 0 | 77% |
| cbbcs3 | 1 | Cd9 | 4F/4R, 1F/1R, 2F/2R | 1263, 5554, 11968 | intron | wt | wt | 24 | 11 | 2 | 0 | 46% |
| cbbcs3 | 1 | Cd9 | 4F/4R, 1F/1R, 2F/2R | 1263, 5554, 11968 | intron | low | bimod. | 12 | 8 | 3 | 0 | 67% |
| cbbcs3 | 1 | Cd9 | 4F/4R, 1F/1R, 2F/2R | 1263, 5554, 11968 | intron | low | wt | 31 | 22 | 5 | 0 | 71% |
| cbbcs3 | 1 | Cd9 | 4F/4R, 1F/1R, 2F/2R | 1263, 5554, 11968 | intron | low | low | 29 | 25 | 1 | 2 | 86% |
| cbbcs3 | 86 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | intron | wt | wt | 24 | 10 | 1 | 0 | 42% |
| cbbcs3 | 86 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | intron | low | bimod. | 11 | 9 | 3 | 0 | 82% |
| cbbcs3 | 86 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | intron | low | wt | 29 | 16 | 4 | 0 | 55% |
| cbbcs3 | 86 | Cd9 | 5F/5R, 1F/1R | 1063, 5554 | intron | low | low | 32 | 28 | 3 | 0 | 88% |
| progenitor | 311 | GFP | 1F/2R, 1F/1R, 2F/2R | 1314, 2994, 3507 | exon | neg | N/A | 96 | 35 | 0 | 2 | 36% |
| cherry/gfp | 33 | GFP | 1F/1R, 1F/3R | 972, 2291 | exon | neg | neg | 89 | 11 | 4 | 13 | 12% |
| cherry/gfp | 34 | mCherry | 2F/2R, 2F/3R | 1258, 2968 | exon | neg | neg | 46 | 3 | 3 | 12 | 7% |
| cherry/gfp | 34 | mCherry | 2F/2R, 2F/3R | 1258, 2968 | exon | neg | pos | 2 | 0 | 0 | 0 | 0% |

Cells were edited with indicated guides, sorted for different gene expression levels ("Sorted population"), single cell cloned and reassessed for gene expression levels ("Clone expression"). **bimod.** - bimodal, **≥1 del.** - one or more deletion amplicons observed, **≥1 del.** - one or more insertion amplicons observed, **No amp.** - no amplicons, **% del.** fraction of clones with deletions amplicons.

### 5.2.4 Non-contiguous lesions

Notably, 13% of all alleles detected in single cell clones (56/428) contained additional lesions (SNPs, indels, large deletions and insertions) that were non-contiguous with the lesion at the cut site (Fig. 5.8b, c and d). This number is likely an underestimate due to stringent filtering of such variants at the *Cd9* locus (see Methods) and due to limited range of Sanger sequencing at the *PigA* and *PIGA* loci. For about 30% of non-contiguous lesions (17/56), the only exonic lesion detected was non-contiguous with the cut site. Furthermore, I observed alleles in which the intronic gRNA caused an inversion of a region containing the exon (Fig. 5.8c). Had the assessment been limited to the immediate vicinity of the cleavage site, such alleles would have been misclassified as wild type, and their phenotypic consequences would have been wrongly called.

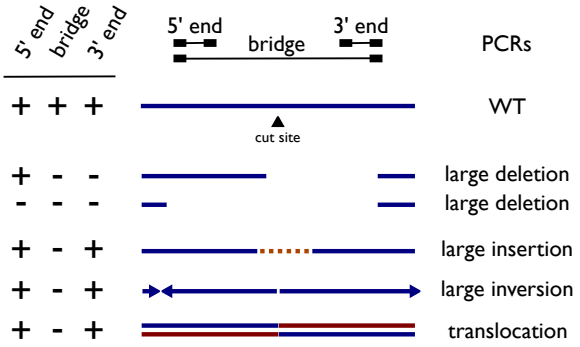



**Figure 5.6:** Results of "diagnostic" PCRs and their interpretations. Blue arrowheads indicate relative orientation of genomic fragments.

### 5.2.5 Unexpected genotypes of inconsistent clones

In mouse ES cells edited at the monoallelic *PigA* locus and sorted for loss of gene expression, I expected every clone to yield exactly one allele with a lesion overlapping the exon. One clone yielded two alleles, likely a result of a picking two closely growing colonies. Only seven out of remaining 164 clones did not contain a lesion overlapping the nearest exon (three were wild-type around the cut sites and four contained cut site, local lesions). They likely contained lesions in other exons or rearrangements outside of the amplified area that could ablate gene expression (e.g. large inversions containing the exon, insertions interfering with splicing, translocations within the gene). Since expression status of these clones was not ascertained after colony outgrowth, some of them could also be PigA proficient.

In ten cases, it was not possible to recover any product spanning the exon, even with a long-range PCR (16 kb). To understand this class of events, I performed additional, "diagnostic" PCRs targeting each end of the *PigA* locus (Fig. 5.1a, gray primers). In five cases, just one end or neither end of the locus could be amplified, suggesting a larger deletion. In the remaining five cases, both ends were amplified. Since no product connecting the two ends could be obtained, these are likely to be translocations, large inversions or large insertions (Fig. 5.3 and 5.6).

In female RPE1 cells edited at the *PIGA* locus and sorted for loss of gene expression, I expected every clone to yield exactly two alleles. At least one of them, presumably on the active chromosome X, should contain a lesion overlapping the nearest exon. No clone had more than two alleles and all clones with exactly two alleles had at least one exon-overlapping lesion, as expected. However, in about 32% of clones (14/44) only one allele was detected with PCR up to 12 kb. This could be due to a larger rearrangement (translocation, large deletion, insertion or inversion), which would explain loss of PIGA expression. Alternatively, five of the fourteen clones, in which an exon overlapping lesion was detected, could be monosomic or homologous for these lesions (there was no variants distinguishing the homologs). Therefore, the frequency of undetected alleles can range from 10% to 16% (9 or 14 alleles out of 88). The lower bound of this range is consistent with the rate of 8% (9/117) in mouse ES cell clones mutagenized with intronic gRNAs at the *PigA* locus, considering a slightly longer-range PCR was used (16 kb vs 12 kb). Higher rate could indicate a locus or cell-specific difference.

Clones derived from cells edited at the *Cd9* locus could be broadly classified into $Cd9^{low}$

and Cd9-positive (ie. bimodal, "low turned wild-type" and "true wild-type" clones; see chapter 4, Fig. 4.4c). The haplosufficient nature of the *Cd9* gene is demonstrated by the fact that I could detect at least one intact exon 2 in each one of the 67 Cd9-positive clones. Conversely, almost all $Cd9^{low}$ clones (25/26) had exon overlapping lesions in all detected alleles. The single exception contained an intronic insertion with a polyA signal. Furthermore, gene dosage could largely explain the difference between "true wild-type" and "low turned wild-type" clones. The first group usually contained at least two functional exon 2s (22/24), while the second group usually had exactly one (27/30), consistent with their 50% lower Cd9 expression (Fig. 5.5).

For experiments at the *Cd9* locus, I used mouse ES cells derived from an F1 cross between BL6 and CAST mouse strains, which allowed me to distinguish the homologous chromosomes. In no case was the repair outcome identical between homologs within a clone, despite 15 alleles reoccurring between clones. Just over half of the mutagenized clones (52/93) contained precisely one BL6 and one CAST allele, as expected. Notably, in 18 clones only one allele was detected with PCR spanning up to 12 kb, potentially due to a larger rearrangement (translocation, large deletion, insertions or inversions), monosomy or LOH. Some of the wild-type BL6 alleles removed as feeder-derived could be the missing alleles. An abnormal number of alleles (two from the same strain or more than two in total) was found in 21 clones, which could have resulted from picking two closely growing colonies, large duplication, repair events happening during clone outgrowth or aneuploidy (spontaneous or induced by Cas9 cutting).

Two clones contained recombinant BL6-CAST alleles (Fig. 5.9). In one case, a LOH event distal to the breakpoints converted part of the CAST allele to BL6. In another case, the BL6-CAST crossover boundary did not coincide with the breakpoint. I concluded that the creation of these alleles likely involved interhomolog strand invasion as they cannot be explained by a simple rejoining of the resected ends of two broken chromosomes.

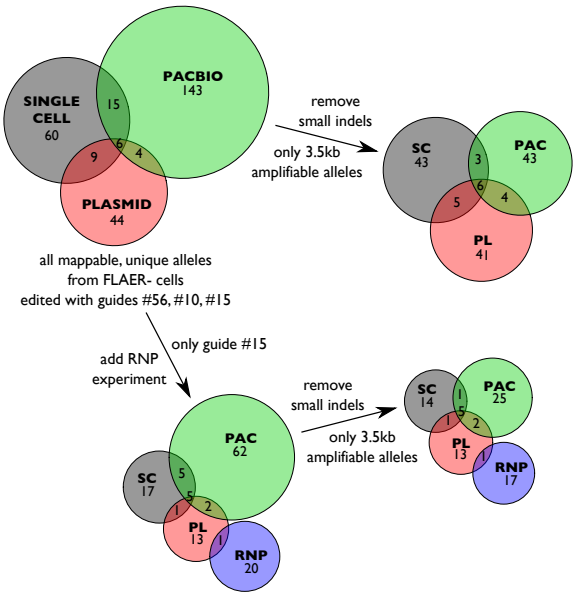



**Figure 5.7:** Overlap between unique *PigA* alleles derived using different methods. “PacBio” and “Single cell” refer to alleles shown in Fig. 5.2, 5.3a, 5.3c and 5.3d. “Plasmid” alleles were derived in the same experiment from subcloned PCR amplicons. “RNP” alleles were derived in an independent experiment only using guide #15 (Fig. 5.3b).

### 5.2.6 Diversity of resolved alleles at the *PigA* locus

To gauge the diversity of mutagenesis outcomes, I have compared unique, sequence resolved alleles that were derived in one experiment from PigA-deficient mouse ES cells edited with intronic gRNAs #10, #15 and an exonic gRNA #56 using following three methods:

- PacBio sequencing of 5.5 kb PCR products amplified from bulk DNA (2F/2R primer pair in Fig. 5.1a; results in Fig. 5.2)
- Sanger sequencing of single cell clones, with PCR product up to 16 kb in size (all primer pairs in Fig. 5.1a; results in Fig. 5.3a, 5.3c, 5.3d)
- Sanger sequencing of individual 3.5 kb PCR products amplified from bulk DNA

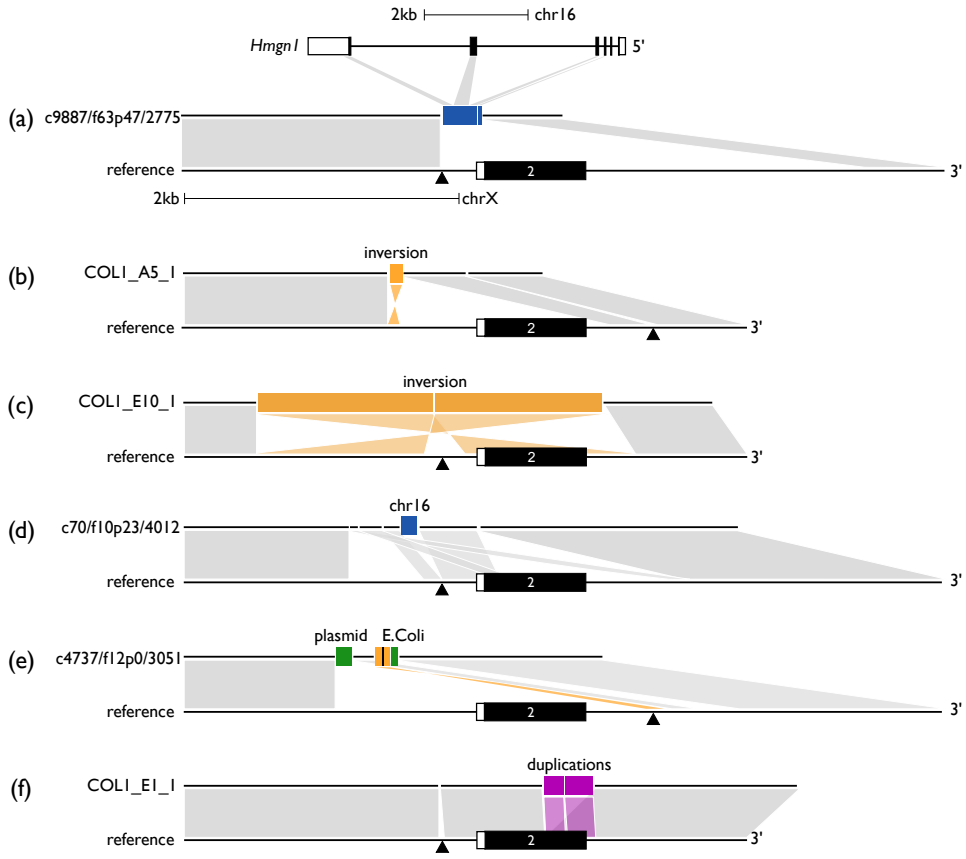



**Figure 5.8:** Examples of alleles. The bottom diagram of each panel represents the *PigA* reference allele around exon 2, the diagram immediately above shows the structure of the sequenced allele. Black horizontal line: direct reference match; orange bar: inversion; blue bar: insertion from another part of the genome; violet bar – duplication; black arrowhead: gRNA target site. Gray, orange and violet shadows represent, respectively, direct, inverted and duplicated match between the reference and the sequenced allele. Lack of shadow at the reference locus represents a deletion in the sequenced allele. (a) Putative insertion from a reverse transcribed RNA. The top diagram line shows the genomic structure of *Hmgn1*; note the scale differs from that of *PigA* gene. (b) Exonic lesion non-contiguous with the cut site. (c) Inversion of a region containing the exon. (d) "Scrambled" allele with insertion from chromosome 16. (e) Combined deletion, local inversion and insertion from *E.Coli* genome. (f) Duplication of a region containing the exon.

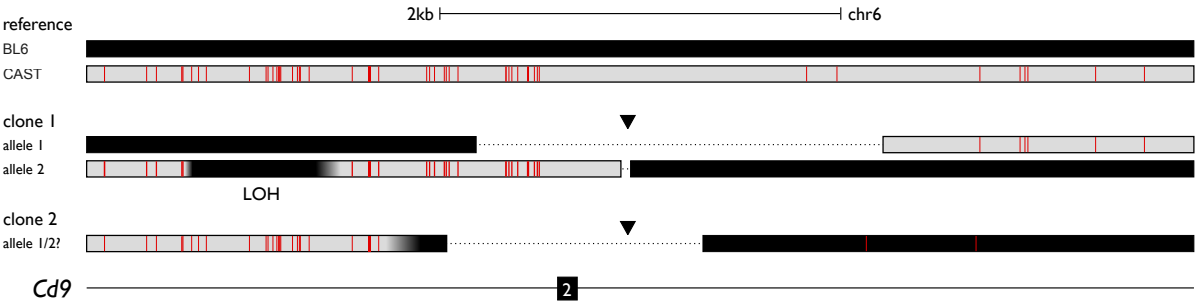



**Figure 5.9:** Recombinant *Cd9* alleles. Two of the sequenced single cell clones contained alleles indicative of a cross-over event between the homologous chromosomes. Red vertical bars in CAST allele (gray bar) indicate positions of sequence divergence from the BL6 reference genome (black bar), dotted black line indicates missing sequence (deletion), thin black line indicates an intron. LOH: loss of heterozygosity.

and cloned into plasmid vectors (1F/1R primer pair in Fig. 5.1a)

Clustering of PacBio reads from PigA-deficient samples yielded 168 unique alleles. The majority of the alleles recovered from single cell clones and plasmid cloned products were unique (90/130 and 63/75, respectively; wild-type alleles excluded). In total 281 unique alleles were recovered by the three methods, only 31 of which (11%) were shared between two or three methods (Fig. 5.7).

To make the comparison more reliable, I removed alleles which could not be recovered with the 3.5 kb primer pair (1F/1R) and small indels (<10 bp), which were depleted from PacBio clusters due to method-specific biases described in subsection 5.2.1. Out of the remaining 145 unique alleles only 18 (12%) were detected with more than one method (Fig. 5.7).

I also compared the alleles in this experiment with ones derived by single cell cloning of RNP mutagenized cells, keeping only the alleles mutagenized by the intronic gRNA #15. The only overlap observed was between one out of the 21 unique RNP alleles and one allele in the plasmid cloned group. I concluded that the large allelic diversity of mutagenic outcomes may be difficult to describe exhaustively using sequencing based methods.

### 5.2.7 Diversity of deletion fingerprints at the *PigA* locus

Diversity of deletion outcomes can be visualized by resolving PCR products from pools of edited cells on an agarose gel. If enough cells were used in each experiment to avoid stochastic undersampling of different deletion outcomes, the ladder-like pattern corresponding to different deletion sizes ("deletion fingerprint") should be similar between biologically independent replicates.

I repeated the original experiment four times using intronic gRNA #15 in two mouse ES cell lines, the original JM8 (also transfected with a PiggyBac Cas9 vector) and its subclone expressing Cas9 from a single-copy lentiviral transgene. I sorted PigA-deficient cells and performed 3.5 kb PCR (1F/1R primer pair) on bulk extracted DNA. I assumed 40% transfection and stable transposition efficiency of the gRNA plasmid, 15% frequency of PigA loss due to intronic mutagenesis (Fig. 4.2b), 20% plating efficiency of mouse ES cells and ability to amplify 80% of PigA-deficient alleles using the 3.5 kb PCR (35/43 among single cell clones). Starting with 1.5 million cells this translates into a transfection bottleneck of 15,000 individual cells with detectable deletions in the Cas9 expressing line (Fig. 5.10a). After antibiotic selection and population outgrowth, I sorted one million cells from each sample, ensuring more than 60x coverage of the bottleneck.

I asked whether sampling of 15,000 unique cells bearing >200 bp deletions is enough to cover the diversity of the possible deletion outcomes. I assumed that the sorting step (with 60x coverage) did not reduce this initial diversity. However, the PCR reaction itself could introduce stochastic noise into the procedure, if too few products were sampled from the pool of extracted genomic DNA in each individual reaction. In order to ensure that PCR was not the limiting factor I have performed a series of technical duplicate PCR reaction starting with 250, 2,500 and 12,500 DNA copies (80% of which should be possible to amplify with the “short” PCR) and compared their "deletion fingerprints" (Fig. 5.10). Sampling 250 copies led to loss of technical reproducibility, as PCR duplicates differed substantially. With 2,500 copies, the diversity of the biological replicate #4 was preserved, revealing it to be the least complex in the set. Sampling 12,500 copies preserved diversity of all replicates. Although some similarities could be observed across biological replicates and between the two cell lines in the same biological replicate, my general conclusion is that sampling ~15,000 cells did not sufficiently cover the diversity of deletion alleles.

## 5.3 Discussion

Using long-range PCR, I have genotyped in excess of 850 single cell clones mutagenized with Cas9,

**Table 5.2:** Summary classification of alleles.

| Gene | gRNA | Target | Expr. | Method | Indel | Deletion >50 bp | Insertion >10 bp | Multi -Lesion | Intact Exon | WT | Total alleles | Total clones |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PigA | 15 | intron | neg | RNP | 0 | 22 | 4 | 3 | 1 | 0 | 24 | 24 |
| PigA | 15 | intron | neg | PiggyBac | 0 | 40 | 9 | 10 | 1 | 1 | 48 | 48 |
| PigA | 10 | intron | neg | PiggyBac | 0 | 39 | 7 | 12 | 3 | 0 | 45 | 45 |
| PigA | 56 | exon | neg | PiggyBac | 32 | 10 | 6 | 1 | 2 | 2 | 48 | 47 |
| PIGA | 274 | intron | neg | transient | 7 | 19 | 4 | 2 | 12 | 3 | 32 | 16 |
| PIGA | 275 | intron | neg | transient | 12 | 16 | 6 | 5 | 15 | 0 | 34 | 19 |
| PIGA | 276 | intron | neg | transient | 6 | 5 | 4 | 0 | 8 | 2 | 22 | 9 |
| Cd9 | 1 | intron | low | PiggyBac | 0 | 42 | 9 | 10 | 1 | 0 | 43 | 26 |
| Cd9 | 1 | intron | bimod. | PiggyBac | 8 | 20 | 6 | 4 | 14 | 0 | 32 | 13 |
| Cd9 | 1 | intron | l-wt | PiggyBac | 24 | 30 | 9 | 5 | 33 | 0 | 59 | 30 |
| Cd9 | 1 | intron | wt | PiggyBac | 33 | 18 | 3 | 4 | 50 | 0 | 55 | 24 |

**Expr.:** expression class; **bimod.:** bimodal; **l-wt:** low turned wild-type; **indel:** small deletion and/or insertion only (<50 bp); **Intact exon:** intact exon in the correct orientation. Categories are not mutually exclusive.

about 300 of which were also sequenced at the mutagenized locus using Sanger and PacBio technologies. The results revealed a pervasive presence of large deletions (50 bp - 9.5 kb), which explains frequent loss of gene expression upon intronic cutting. Many complex rearrangements of the locus, including large insertions, inversion, translocation between homologs and non-contiguous lesions were also discovered.

### 5.3.1 Consequences of large deletions

Large deletions could be pathogenic in gene therapeutic context. Given that a target locus would presumably be transcriptionally active, such mutations could juxtapose it to the nearest oncogene, initiating neoplasia. A deletion inactivating a nearby tumor suppressor gene could predispose the cell to become cancerous, even if only one copy is affected (Santarosa and Ashworth, 2004). The effect might not be immediately obvious, as the lesions may constitute a carcinogenic first "hit". This is especially true for stem cells and progenitors, which have a long replicative lifespan and may become neoplastic with time. This would be similar to the activation of *LMO2* by pro-viral insertion in some of the early gene-therapy trials, which caused cancer in these patient (Hacein-Bey-Abina et al., 2003).

The closer the target site is to a cancer-driver gene, the higher the risk posed by deletions and other local rearrangements. I have not gathered enough unbiased data at any locus to accurately describe the frequency of "complex" lesions as a function of distance from the cut site. However, the gene expression data at the *PigA* locus comes close. A simple exponential model can be fitted using exon 2 proximal (100-500 bp) and distal (~2 kb) gRNAs. I assumed that loss of PigA expression caused by gRNAs close to exon 2 is exclusively due to damage to exon 2 (and not exon 1 or exon 3) and that the two gRNAs in the middle of intron 2 confer double the risk by affecting both exon 2 and 3 (data not shown). As a crude reality check, I asked if the model correctly predicts the tail of the distribution - the largest deletion in the Sanger sequencing dataset (which is 9.5 kb in total, 6.6 kb in one direction). Given 117 intronically edited alleles from PigA-deficient cells were tested, the model indicates on average 1.43 such lesions (or larger) should be found, which is consistent with reality.

The lesion frequency under this model halves with every kilobase of distance from the cut site. This implies that for every 100 million mutagenized cells (the scale of current gene therapeutic efforts), one lesion spanning 22.5 kb or more in one direction from the cut site would be expected,

on the average. While such calculations are subject to a very high statistical uncertainty and may not generalize to other loci, they could inform the design of future experiments with respect to investigated distances and numbers of cells.

### 5.3.2 Consequences of other complex lesions

Sequencing of single cell clones yielded large insertions, inversions, non-contiguous lesions, cross-overs and LOH events. Some of these were directly implicated in causing gene expression loss, notably inversions containing the exon, non-contiguous lesions within the exons and an intronic insertion containing polyA signal. Furthermore, the consequences of some of these lesions would have been underestimated, if only genotyping around the cut site was performed. This suggests that genotyping should not be limited to the immediate vicinity of the cut site and stresses the importance of careful phenotypic assessment, whenever possible.

The full extent of non-contiguous lesions is not known. Sanger sequencing was primarily performed to detect deletion breakpoints, resolve insertions and ensure integrity of the exon closest to the cut site, so more distal lesions could have been missed. I have observed some small, distal indels in alleles derived by PacBio sequencing of the *Cd9* locus. However, I decided to filter them out, as some of them consistently clustered at the ends of the read (indicating quality issues) and in low complexity regions (where accurate mapping turned out to be an issue). Such lesions could be investigated in the future using more reliable Sanger sequencing. As with large deletions, a quantitative description of the frequency of non-contiguous lesions as a function of distance from the cut site would be useful in gene therapeutic context.

Cas9-induced cross-overs would have minimal impact if products co-segregate on cell division (i.e. undergo a "z-segregation"). If instead they segregate away from each other ("x-segregation"), a cross-over would result in chromosome-scale LOH, which could uncover recessive alleles. If tumor suppressor genes are affected, this could initiate cancer.

Analysis of single cell clones has indicated presence of aneuploidies and alleles that could not be fully resolved, such as translocations. These events are often observed in cancers, due to their ability to juxtapose active promoters and oncogenes, amplify oncogenes and reduce the copy number of tumor suppressors. I have not performed a more detailed analysis of these clones, but copy-number screening by qPCR or digital droplet PCR, karyotyping to detect translocations and SNP array genotyping for large scale deletions and LOH events are warranted. Furthermore, it would be crucial to establish the causal relationship between Cas9 mutagenesis and aneuploidies, as ES cells in culture are known to acquire aneuploidies spontaneously.

Failure to detect one of the two lesions at an autosomal locus in a single cell clone (or a founder animal) can be easily mistaken for a homozygous lesion. While in the context of animal mutagenesis such mistake should be detected when animals fail to breed true (as discussed in Shin et al., 2017), it can significantly influence the interpretation of experiments using single cell clones, whose alleles cannot be easily isolated.

These considerations formed the basis of debate, in which I was involved, on the interpretation of a particular human embryo editing study. In that study, researchers used Cas9 to induce a DSB on the paternal allele in human zygotes and observed only the maternal allele in some blastomers isolated from the multicellular embryo three days later. In absence of further evidence, it was concluded that the maternal allele served as a template for the repair of the paternal allele, a process termed "interhomolog repair" (Ma et al., 2017). This conclusion has been challenged by two groups as equally consistent with a failure to detect the paternal allele due to destruction of primer binding sites (Adikusuma et al., 2018; Egli et al., 2018). One of these groups included data which showed edited mouse embryos exhibit high levels of large deletions. Reply to this criticism reported no deletions with PCR spanning up to

10 kb and established that at least some blastomers carry the expected heterozygous patterns of SNPs flanking the target site, which supports the original conclusion (Ma et al., 2018). Another group independently studying interhomolog repair in mouse embryos also carried out the prescribed checks (long-range PCRs, copy-number qPCR) and failed to observe large losses of genetic material (Wilde et al., 2018). With some other groups reporting detection of large deletions in human embryos and in differentiated animal tissues edited in vivo (personal communication), it remains to be established which conditions enable creation of complex lesions (also see the Discussion chapter).

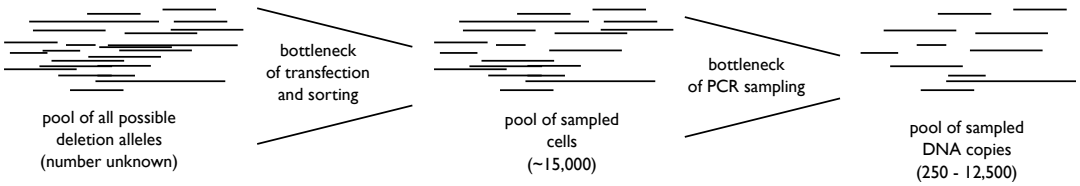
### 5.3.3 Stochasticity of large deletions

Indels induced by any gRNA are highly non-random, with a few indels of particular sizes forming a stable "indel profile". Such profiles are hypothesized to be related to local microhomologies guiding the repair process. I speculated an analogous "deletion profile" exists, potentially also guided by homologies or larger scale secondary structure of the DNA. Ladder pattern observed by resolving amplicons from long-range PCRs on pools of mutagenized cells initially seemed to confirm this hypothesis. However, the observed "profiles" differed between biological replicates.

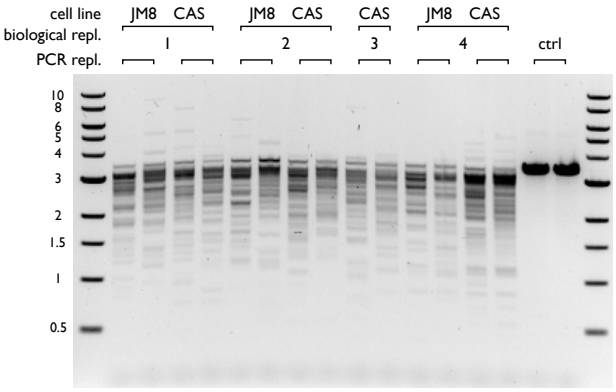
Two possible explanations exists for the lack of reproducibility of "deletion profiles". One is that the potential diversity of induced deletions far outstrips the number of transfected cells with deletion outcomes. This would lead to stochastic undersampling of deletion events in each transfection replicate, resulting in a "noisy" profile. If this model is correct, sampling more cells should eventually reduce the noise between biological replicates. This could be achieved at a scale by employing a non-leaky, inducible gRNA and Cas9 system. Another explanation could be clonal expansion due to stochastic genomic instability. In this model, cells which acquired a mutation that makes them grow faster (e.g. chromosome 8 triploidy) selectively amplify the Cas9-induced deletion they harbor. If this model is correct, a more karyotypically stable cell line should behave more predictably. If such genomic instability is independent of Cas9 mutagenesis, then even wild-type cells will exhibit strong clonal effects, which could be tested by random barcoding. Regardless, my results revealed a source of noise that needs to be taken into account when investigating "deletion profiles".

### 5.3.4 Other considerations

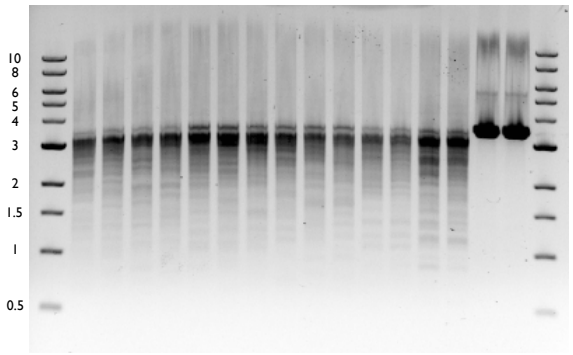
Most of the "low turned wild-type" clones edited at the *Cd9* locus had exactly one exon-overlapping and one non-overlapping lesion, as opposed to "true wild-type" clones, most of which did not have any exon-overlapping lesions. Although the difference between these populations was not immediately apparent in bulk cultures (Fig. 4.4a), improved culturing protocols and use of single cell clone controls could potentially allow systematic quantification and isolation of (or at least enrichment for) cells with monoallelic "complex" lesions at the *Cd9* locus.



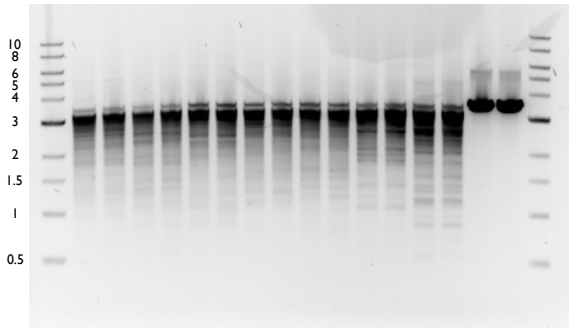

**(a)** Experimental considerations.




**(b)** Deletion fingerprint - 250 copies per reaction.




**(c)** Deletion fingerprint - 2,500 copies per reaction.




**(d)** Deletion fingerprint - 12,500 copies per reaction.

**Figure 5.10:** *PigA* locus was mutagenized using 5' intronic guide (#15) in biological quadruplicate. Duplicate PCR reactions spanning the cut site performed on DNA extracted from the bulk of PigA-deficient cells and resolved on an agarose gel (product size 3,500 bp, 1F/1R primer pair, Fig. 5.1a). JM8 – original mouse ES cell line (transfected with gRNA and Cas9 PiggyBac vectors); CAS – JM8 subclone stably expressing Cas9 (transfected only with a gRNA PiggyBac vector). Ladder scale is in kilobases.

**Table 5.3:** Genomic primer pairs.

| Gene | Name | Sequence | Chr | Start | End | Strand |
|---|---|---|---|---|---|---|
| PigA | 1F | CTTATGGGATGTACTGGGTCACTAG | X | 164421324 | 164421349 | + |
| PigA | 1R | CACCCCAGAAAATGTAACTGAGTTC | X | 164424799 | 164424824 | - |
| PigA | 2F | CTTTCATTTGGTTCATTATTTCTGTTCTTATC | X | 164420461 | 164420493 | + |
| PigA | 2R | CCTTAACTCAAGAGCTGAACTT | X | 164425873 | 164425895 | - |
| PigA | 3F | TTCGACCAGTTTGCTCTAACTCTTA | X | 164417878 | 164417903 | + |
| PigA | 3R | ATCAAAGTGTCCTCGAGTTAAT | X | 164430740 | 164430762 | - |
| PigA | 4F | AAGCTCTTAAGAGGAAAGGCTACAA | X | 164417360 | 164417385 | + |
| PigA | 4R | ATCACACCACAGCATTAGGA | X | 164418508 | 164418528 | - |
| PigA | 5F | TAACAGGTCACATATAGGATTTGGG | X | 164414904 | 164414929 | + |
| PigA | 6F | ATGTGGAAATCCTGTACCAGAAAGA | X | 164429755 | 164429780 | + |
| PigA | 6R | AACTGATTATCTGACCTTCCCT | X | 164423503 | 164423525 | - |
| PigA | 7F | AGGAGACTGAGGCCAGGAATAT | X | 164421983 | 164422005 | + |
| PIGA | 1F | CGGTTCACATGTTCCTGATTAAAGAA | X | 15328961 | 15328987 | + |
| PIGA | 1R | GTGGTCGAGAATTTTACGGTAATGT | X | 15334958 | 15334983 | - |
| PIGA | 2F | CTTTCCCGAACTTCTTCCAAAATGA | X | 15325931 | 15325956 | + |
| PIGA | 2R | AGGCAGGACACCATAATTAGAATCA | X | 15337669 | 15337694 | - |
| Cd9 | 1F | CTTTAGTGTCCTTTTGCACACTTCT | 6 | 125474857 | 125474882 | - |
| Cd9 | 1R | GGTATAACCAGTCCTTCTAGCACAT | 6 | 125469328 | 125469353 | + |
| Cd9 | 2F | CTGTCGTGAAATATTAGGAAAGGGC | 6 | 125477789 | 125477814 | - |
| Cd9 | 2R | AGTACCTCCCGTCTTGCTACC | 6 | 125465846 | 125465867 | + |
| Cd9 | 3F | ATCTGAAGAAGTCTCTCTGACCCTA | 6 | 125467206 | 125467231 | - |
| Cd9 | 3R | TCTTCTTTGGTGATTTGCTGATTCC | 6 | 125461366 | 125461391 | + |
| Cd9 | 4F | AGTTTTCTGGTGATTTTACCGCAAT | 6 | 125472672 | 125472697 | - |
| Cd9 | 4R | CCTTGTCAGAATGCTTTCTTGTCTT | 6 | 125471434 | 125471459 | + |
| Cd9 | 5F | ATCATTTGGCATCCTATTCAACACC | 6 | 125473010 | 125473035 | - |
| Cd9 | 5R | CTCCATCTCCATCCCCATTAATCTC | 6 | 125471972 | 125471997 | + |
| Cd9 | 6F | AGGTCTCAGTAAGTTAGCTCAAGTG | 6 | 125464803 | 125464828 | - |
| Cd9 | 6R | ATAAGGAGGTGTGATCAGTGGAAAA | 6 | 125463562 | 125463587 | + |
| Cas9-GFP | 1F | AGAAACTGAAGAGTGTGAAAGAGC | - | - | - | + |
| Cas9-GFP | 1R | CGTGCAATCCATCTTGTTCAATG | - | - | - | - |
| Cas9-GFP | 2F | GGCGGCAGGAAGATTTTTACCC | - | - | - | + |
| Cas9-GFP | 2R | GGGTGTTCTGCTGGTAGTGGT | - | - | - | - |
| cherry/gfp | 1F | GTAAACGGCCACAAGTTCAGC | - | - | - | + |
| cherry/gfp | 1R | GCTCAAGATGCCCCTGTTCT | - | - | - | - |
| cherry/gfp | 2F | GGAGGATAACATGGCCATCATCAAG | - | - | - | + |
| cherry/gfp | 2R | CTGATGCTCTTCGTCCAGATCA | - | - | - | - |
| cherry/gfp | 3R | TTGACCTATTCTGGCATTGTAGACA | - | - | - | - |

Genomic position is given with respect to the GRCm38 or GRCh38.

# Chapter 6

# Discussion

## 6.1 Causes of complex lesions

Induction of DSB by Cas9 and subsequent repair are clearly the necessary factors involved in creation of Cas9-induced complex lesions. However, mechanistic details are unknown. Here, I speculate on three mutually non-exclusive factors contributing to complex lesions and briefly review tentative support for them in my data. I also discuss how these factors can be studied further.

First, generation of some complex lesions may primarily involve double-strand resection. This would explain deletions as a simple resection-and-ligation reaction (presumably mediated in part by components responsible for NHEJ). Furthermore, a double-stand resection could expose repeat sequences distal to the cut site. This in turn could lead to SSA repair between exposed direct repeats (resulting in more deletions) and to non-allelic homologous recombination (NAHR) between repeats on different chromatids. NAHR is the proposed mechanism for many large-scale rearrangements in cancer (deletions, duplications, translocations and inversions).

Second, complex lesions could result from HR repair being subverted by simultaneous breakage of both sister chromatids at the same locus. Failing to find an intact repair template, HR may either mediate NAHR or the whole process may revert to MMEJ. Normally, MMEJ involves limited single-strand resection, but during an abortive attempt at HR the DNA may be resected much more extensively. This could expose distal homologies that are not normally available to MMEJ. Synapsis of these homologies could cause large deletions and inversions.

Third, complex lesions could be stimulated by Cas9-intrinsic properties, such as its exonuclease activity and interference with DDR machinery by staying bound to the DNA after making a cut. This may influence the repair process, making relatively non-mutagenic NHEJ and HR less likely, both in favor of more mutagenic processes, such as MMEJ or NAHR.

Data presented in this thesis does not clearly exclude any of the three above mentioned mechanisms. Preliminary analysis has revealed that the amount of exact homology at large deletion breakpoints varies from zero to 12 nt (data not shown), which indicates at least some deletions may be a product of double-strand resection. A more detailed comparison with the expected distribution of homology lengths will be necessary to tell, whether MMEJ is likely to be involved. Moreover, if recombination between repeats was a major driver of large deletions, I would expect them to emerge in "deletion fingerprinting" experiments as reproducible, highly enriched deletion bands. While these were not observed, it may be either due to low sensitivity of the assay or due to low repeat content of the studied *PigA* locus. Further analysis to explore inexact homologies (i.e. stretches of homology with infrequent mismatches, also called "homeologies") at deletion breakpoints could also help decide, if repeat recombination is involved. Finally, many small, locally templated insertions and small, non-contiguous lesions implicate single-strand resection and MMEJ.

Dissection of requirements for complex lesions by means of genetic screens based on the assay described in chapter 4 is warranted. Further-

more, translocations could potentially be enriched for and studied separately by targeting very long introns and selecting for loss of gene expression (although such selection will also enrich for very large deletions and inversions). Systematic study of different loci in the genome in this way would allow better understanding of targets of translocations and thus proper risk estimation in the context of gene therapy. A recent report found that distal insertions induced by Cas9 (i.e. mapping to other loci in the genome than the edited locus) are enriched for sequences close to the cut site in 3D space (Leenay et al., 2018). I predict this would also be the case for translocations. Finally, careful comparison between different precision nucleases may shed some light on whether Cas9-intrinsic properties are contributing to creation of complex lesions. However, such comparisons are often difficult due to many confounding factors, such as differential levels of protein expression and nuclease acitivity.

## 6.2 Ways to avoid complex lesions

Complex lesions are generally an undesirable outcome of gene editing, both in gene therapy (where they are potentially pathogenic, as discussed in chapter 5) and in basic research (where they may be confounding and difficult to genotype). Based on the discussed causes of complex lesions, I propose four broad ways of tackling this issue - complete avoidance (or more controlled induction) of the DSB, avoidance of simultaneous chromatid breakage, manipulation of DDR and re-engineering of the wild-type Cas9.

First, editing without DSB creation is possible in principle by using base editors. In practice, indels consistent with DSB creation are observed upon base editing at a rate of around 1% (Gaudelli et al., 2017; Komor et al., 2016). Further engineering of base editing tools (e.g. using deactivated instead of nickase Cas9, as in the first generation of base editing Cas9) may help abolish DSBs completely or at least reduce them to marginal levels. Another potential solution to the DSB problem is creation of the break in a controlled fashion, as for example during V(D)J recombination. These reactions do still lead to carcinogenic translocations, but at a very low rate (Alt et al., 2013). Discovery or engineering of a programmable recombinase could achieve this goal. Initial foray in this area has been made by fusing Cas9 and Gin recombinase, but this enzyme only operates on specific recombinase recognition sites (Chaikind et al., 2016). Finally, precise modification of the DNA has been demonstrated in yeast transfected with ssDNA complementary to the lagging strand (Barbieri et al., 2017). This approach could potentially be adapted to human cells. However, no systematic safety assessment of this method has been conducted yet.

Second, ensuring that only one sister chromatid is broken at any given time could be achieved e.g. by reducing the effective concentration of the nuclease, choosing low efficiency guides or reducing the activity of the Cas9-gRNA complex by engineering of its components. This does not necessarily require having one Cas9 molecule per cell, but only that on the average the activity should be low enough to induce one cut every few hours, to allow HR to finish the repair process. While initially the editing efficiency would drop, it may be possible to maintain a level of nuclease activity that is low enough to avoid breaking both chromatids, but high enough to keep the reaction going despite creation of new wild-type alleles due to DNA replication. Alternatively, the cutting could be confined to non-replicative stages of the cell cycles, e.g. by coupling Cas9 to cell stage specific degrons (Huang et al., 2017) or timing the delivery in synchronized cells (for analogous approaches trying to increase HR repair see Gutschner et al., 2016; Lin et al., 2014; Yang et al., 2016a).

Third, blocking excessively mutagenic DSB repair pathways or promoting 'safe' ones could be a promising way of avoiding complex lesions, providing these mechanisms are well understood. In basic research, such modulation could be achieved by overexpression and silencing of specific DDR proteins. However, this could have unexpected consequences at other loci than the edited one.

Coupling Cas9 to proteins involved in DNA repair or to the repair template (if templated repair is the goal, e.g. Savic et al., 2018; Shou et al., 2018) is a more "topical" solution, which may be compatible with gene therapy.

Fourth, re-engineering of Cas9 to abolish exonuclease activity or make it release DNA after cut could potentially improve its risk profile. A molecular evolution approach could perhaps be employed, if selective conditions against exonuclease activity / extended binding and for high endonucleotic activity can be obtained. However, there is a risk that these properties are inextricably linked. In particular, while a simple DSB would predominantly be repaired with no change to DNA sequence, a more complex, Cas9-blocked DSB may elicit more vigorous, mutagenic repair mechanisms.

## 6.3 Ways to exploit complex lesions

Cas9-induced cross-overs have already been used in yeast to enable genetic mapping at higher resolutions than allowed by natural recombination rate (Sadhu et al., 2016). "Distal" translocations, which cause a more dramatic reshufling of the genome, could be used to investigate the effect of putting different loci in a linkage disequilibrium and the significance of the particular chromosomal setup of a given organism. Other forms of genome engineering are already being used in this field, for example in a recent work in which all of yeast chromosomes were combined into one (or two) units (Luo et al., 2018a; Shao et al., 2018).

Single gRNA-induced deletions could be a useful tool for studying non-coding elements (enhancers, long non-coding RNAs etc.). In contrast to coding genes, which can be inactivated by introduction of a small, frameshifting indel, non-coding elements usually require more extensive mutagenesis. Such approach could complement currently available tools, which use CRISPRi, saturation mutagenesis or paired gRNA-induced deletions (Aparicio-Prat et al., 2015; Canver et al., 2015; Gasperini et al., 2017, 2018; Korkmaz et al., 2016; Zhu et al., 2016). Due to its simplicity, it could potentially match the throughput of the CRISPRi approach, if multiple gRNAs are multiplex per cell (despite the confounding risk of translocations between multiple cut sites). This would make it possible to study the effect of tens of thousands deletions in one experiment. Such deletions could also be isolated and studied in detail, as opposed to less well defined chromatin silencing induced by CRISPRi.

Single gRNA-induced deletions of varying size could be exploited to create genomic "deletion series". Normally, exonic deletions are created through Cre recombination between lox sites flanking the exon. This approach offers a high degree of precision. However, that also means potential confounding factor may not be discovered, if e.g. a regulatory element is consistently co-deleted with the exon. An exonic deletion series could easily be created by single gRNA in exons of moderate size (100-5000 bp). By introducing slight variability in resulting genotypes, this approach could make the experiment more robust. Another application of this technique could be to investigate the extent and the function of different protein domains. Normally, this is done by cloning of the cDNA of interest ("DNA complementary to the mRNA", containing only the coding part of the gene without introns) into a plasmid and deleting various elements in vitro. The modified product is then usually transiently expressed from the plasmid in the cells knocked-out for the endogenous gene. Using single gRNAs, a deletion series could be created either directly in the wild-type gene (if studying large exons) or at a locus in which the endogenous gene was replaced with its cDNA copy. The advantage of this approach over the current solution is that it could be done at a larger scale and in the endogenous regulatory context.

## 6.4 Probing protein isoform diversity using CRISPR/Cas9-based assays

I have detected a variety of discrete protein expression levels following Cas9 mutagenesis at the *Cd9* locus. Although I have not investigated them

in detail, it is likely they represent stable isoforms brought about by specific types of genomic damage (loss of exon, out-of-frame mutation, epitope modifying mutation). Systematic mutagenesis of the coding gene coupled to a continuous protein-related readout (e.g. abundance of protein measured by a flow cytometry, functional assay or depletion of specifically edited cells in case of an essential genes) could therefore be used as a tool to study the plasticity of protein isoforms. Abundance, and ideally the specific sequence, of mRNA could be measured in the same assay to provide additional layer of information and decouple transcriptional effects. In particular, such an assay would enable investigation of general splicing and folding rules, stability requirements, relationship between conservation and essentiality of specific protein domains as well as epitope malleability. A similar procedure investigating functionality of so-called variants of unknown significance in the essential *Brca1* gene has recently been described (Findlay et al., 2018; Starita et al., 2018).

## 6.5 New methods for genotyping of complex lesions

Reliable, unbiased, high throughput methods for genotyping of complex lesions are necessary to understand them in more detail, to monitor gene therapy applications and to guide the development of preventative measures.

Methods I used in this study suffer from a number of shortcomings. Single cell cloning and Sanger sequencing are low throughput. Flow cytometric assay can only be applied to a subset of transcriptionally active loci. PacBio sequencing of bulk DNA results in biased readouts. New methods need to be developed to enable more in-depth study and systematic monitoring of complex lesions at other loci, in other cell types or using other nucleases. Barcoding of PCR products in early cycles could improve PacBio readout by removing amplification and sequencing biases as well as allowing accurate genotyping of small indels. Oligos preventing amplification of wild-type and small indel products could be used to enrich for complex lesions without the need for selection based on a loss of gene expression. Depletion of wild-type products could also be performed post-amplification, e.g. by hybridization with complementary RNA and digestion using duplex-specific nuclease (Zhulidov et al., 2004) or using wild-type sequence-specific oligos conjugated to magnetic beads. Copy number profiling using qPCR would allow quick assessment of the extent of large deletions. RNA quantification by qPCR and full-length RNA isoform sequencing could also be used to approximately quantify the frequency of complex lesions (incl. translocations and small non-contiguous lesions) in genes whose activity cannot be assessed by flow cytometry.

## 6.6 Complex lesions and risk management

Direct consequences of complex lesions were discussed in chapter 5. Here, I discuss feasibility of Cas9 usage in basic research and gene therapy given these consequences.

As mentioned earlier, complex lesions can be considered a nuisance in most basic research applications. The presence of a large deletion, insertion, inversion or translocation can often be readily detected as failure to amplify any mutant genotype in the offspring of edited animals crossed to the wild-type. However, it does incurr a cost in time and money spent breeding and genotyping unsuitable lesions. Complex lesions can compromise interpretation of multiplexed editing, if animals are not sufficiently backcrossed to the wild-type, as lack of mutant detection in F1 animals cannot be interpreted as lack of editing. Non-contiguous lesion in F1 animals may also go unnoticed, unless a proper rescue experiment is performed. These problems compound when using single cell cloned cell lines, whose alleles cannot be separated by breeding. In practical terms, complex lesions highlight the necessity for well established experimental controls - careful assessment of expression and functional phenotypes, use of independently derived cell clones or edited animals and rescue experiments.

In terms of gene therapy, two factors seem crucial - whether editing is somatic or germline and whether the purpose of the gene editing experiment is curative or prophylactic. Whether the therapy is conducted ex vivo or in vivo is another important factor, since the former allows a higher degree of quality control and may reduce the risk of affecting long-lived proliferative stem cells. Currently, the discovery of complex lesions, our general understanding of off-target effects and dearth of studies on long-term consequences of gene editing in relevant animal models (such as monkeys) or in humans suggests extreme caution when considering gene therapeutic application beyond life-saving interventions. Therapies for terminal diseases (incl. most forms of cancer), where risk of cancer induced by Cas9 is outweighed by the imminent risk of death, fall squarely into the life-saving category. Consistently, therapies for cancer using ex vivo edited T cells are currently a major focus of gene therapeutic clinical trials. There is also a "grey zone" of applications that may not be life-saving, but are life-changing. One example may be the potential cure for patients with Hunter's or Hurler's syndrome, in which case a physician may elect together with the patient to accept the risk of carcinogenesis in exchange for a potential cure. Finally, prophylactic attempts aimed at reducing cholesterol levels (*Pcsk9* editing, not yet in clinical trials) or prevent (but not cure) HIV infection (*Ccr5* editing) and most forms of therapeutic embryo editing, which is bound to influence the germline (especially, when preimplantation diagnostics offers a viable alternative), may need to be deferred until our understanding of Cas9 mutagenesis improves. Finally, I want to acknowledge that the ultimate decision as to if and when to allow usage of Cas9 and other precision nucleases on patients should be primarily in the hands of people with extensive experience in healthcare related risk management - physicians running clinical trials and regulators specializing in drug safety.

# References


Adikusuma, F., Piltz, S., Corbett, M. A., Turvey, M., McColl, S. R., Helbig, K. J., Beard, M. R., Hughes, J., Pomerantz, R. T., and Thomas, P. Q. (2018). Large deletions induced by Cas9 cleavage. *Nature*, 560(7717):E8–E9.

Ahuja, A. K., Jodkowska, K., Teloni, F., Bizard, A. H., Zellweger, R., Herrador, R., Ortega, S., Hickson, I. D., Altmeyer, M., Mendez, J., and Lopes, M. (2016). A short G1 phase imposes constitutive replication stress and fork remodelling in mouse embryonic stem cells. *Nature Communications*, 7(May 2015):1–11.

Aiuti, A., Biasco, L., Scaramuzza, S., Ferrua, F., Cicalese, M. P., Baricordi, C., Dionisio, F., Calabria, A., Giannelli, S., Castiello, M. C., Bosticardo, M., Evangelio, C., Assanelli, A., Casiraghi, M., Di Nunzio, S., Callegaro, L., Benati, C., Rizzardi, P., Pellin, D., Di Serio, C., Schmidt, M., Von Kalle, C., Gardner, J., Mehta, N., Neduva, V., Dow, D. J., Galy, A., Miniero, R., Finocchi, A., Metin, A., Banerjee, P. P., Orange, J. S., Galimberti, S., Valsecchi, M. G., Biffi, A., Montini, E., Villa, A., Ciceri, F., Roncarolo, M. G., and Naldini, L. (2013). Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science*, 341(6148):1233151–1233151.

Akcakaya, P., Bobbin, M. L., Guo, J. A., Lopez, J. M., Clement, M. K., Garcia, S. P., Fellows, M. D., Porritt, M. J., Firth, M. A., Carreras, A., Baccega, T., Seeliger, F., Bjursell, M., Tsai, S. Q., Nguyen, N. T., Nitsch, R., Mayr, L., Pinello, L., Bohlooly-Y, M., Aryee, M. J., Maresca, M., and Joung, J. K. (2018). In vivo CRISPR-Cas gene editing with no detectable genome-wide off-target mutations. *bioRxiv*, page 272724.

Aladjem, M. I., Spike, B. T., Rodewald, L. W., Hope, T. J., Klemm, M., Jaenisch, R., and Wahl, G. M. (1998). ES cells do not activate p53-dependent stress responses and undergo p53- independent apoptosis in response to DNA damage. *Curr Biol*, 8(3):145–155.

Alt, F. W., Zhang, Y., Meng, F.-L., Guo, C., and Schwer, B. (2013). Mechanisms of programmed DNA lesions and genomic instability in the immune system. *Cell*, 152(3):417–29.

Ames, B. N. (1979). Identifying environmental chemicals causing mutations and cancer. *Science (New York, N.Y.)*, 204(4393):587–93.

Andersen, S. L. and Sekelsky, J. (2010). Meiotic versus mitotic recombination: two different routes for double-strand break repair: the different functions of meiotic versus mitotic DSB repair are reflected in different pathway usage and different outcomes. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 32(12):1058–66.

Andersson-Rolf, A., Mustata, R. C., Merenda, A., Kim, J., Perera, S., Grego, T., Andrews, K., Tremble, K., Silva, J. C. R., Fink, J., Skarnes, W. C., and Koo, B.-K. K. (2017). One-step generation of conditional and reversible gene knockouts. *Nature Methods*, 14(3):287–289.

Aparicio-Prat, E., Arnan, C., Sala, I., Bosch, N., Guigó, R., and Johnson, R. (2015). DECKO: Single-oligo, dual-CRISPR deletion of genomic elements including long non-coding RNAs. *BMC genomics*, 16(1):846.

Austin, C. P., Battey, J. F., Bradley, A., Bucan, M., Capecchi, M., Collins, F. S., Dove, W. F., Duyk, G., Dymecki, S., Eppig, J. T., Grieder, F. B., Heintz, N., Hicks, G., Insel, T. R., Joyner, A., Koller, B. H., Lloyd, K. C. K., Magnuson, T., Moore, M. W., Nagy, A., Pollock, J. D., Roses, A. D., Sands, A. T., Seed, B., Skarnes, W. C., Snoddy, J., Soriano, P., Stewart, D. J., Stewart, F., Stillman, B., Varmus, H., Varticovski, L., Verma, I. M., Vogt, T. F., von Melchner, H., Witkowski, J., Woychik, R. P., Wurst, W., Yancopoulos, G. D., Young, S. G., and Zambrowicz, B. (2004). The knockout mouse project. *Nature genetics*, 36(9):921–4.

Aylon, Y., Liefshitz, B., and Kupiec, M. (2004). The CDK regulates repair of double-strand breaks by homologous recombination during the cell cycle. *The EMBO journal*, 23(24):4868–75.

Bahassi, E. M., O'Dea, M. H., Allali, N., Messens, J., Gellert, M., and Couturier, M. (1999). Interactions of CcdB with DNA gyrase. Inactivation of Gyra, poisoning of the gyrase-DNA complex, and the antidote action of CcdA. *The Journal of biological chemistry*, 274(16):10936–44.

Barbieri, E. M., Muir, P., Akhuetie-Oni, B. O., Yellman, C. M., and Isaacs, F. J. (2017). Precise Editing at DNA Replication Forks Enables Multiplex Genome Engineering in Eukaryotes. *Cell*, 171(6):1453–1467.e13.

Barkal, A. A., Srinivasan, S., Hashimoto, T., Gifford, D. K., and Sherwood, R. I. (2016). Cas9 functionally opens chromatin. *PLoS ONE*, 11(3):1–8.

Bennardo, N., Cheng, A., Huang, N., and Stark, J. M. (2008). Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. *PLoS genetics*, 4(6):e1000110.

Bennett, C. B., Lewis, A. L., Baldwin, K. K., and Resnick, M. A. (1993). Lethality induced by a single site-specific double-strand break in a dispensable yeast plasmid. *Proceedings of the National Academy of Sciences of the United States of America*, 90(12):5613–7.

Bibikova, M., Beumer, K., Trautman, J. K., and Carroll, D. (2003). Enhancing gene targeting with designed zinc finger nucleases. *Science (New York, N.Y.)*, 300(5620):764.

Blackford, A. N. and Jackson, S. P. (2017). ATM, ATR, and DNA-PK: The Trinity at the Heart of the DNA Damage Response. *Molecular cell*, 66(6):801–817.

Blomen, V. A., Jae, L. T., Bigenzahn, J. W., Nieuwenhuis, J., Staring, J., Sacco, R., Diemen, F. R. V., Olk, N., Stukalov, A., Marceau, C., Janssen, H., Carette, J. E., Bennett, K. L., and Colinge, J. (2015). Gene essentiality and synthetic lethality in haploid human cells. *Science*, 350(6264).

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. *Science (New York, N.Y.)*, 326(5959):1509–12.

Bolton, H., Graham, S. J. L., Van der Aa, N., Kumar, P., Theunis, K., Fernandez Gallardo, E., Voet, T., and Zernicka-Goetz, M. (2016). Mouse model of chromosome mosaicism reveals lineage-specific depletion of aneuploid cells and normal developmental potential. *Nature Communications*, 7:11165.

Boroviak, K., Doe, B., Banerjee, R., Yang, F., and Bradley, A. (2016). Chromosome engineering in zygotes with CRISPR/Cas9. *Genesis*, 54(2):78–85.

Boroviak, K., Fu, B., Yang, F., Doe, B., and Bradley, A. (2017). Revealing hidden complexities of genomic rearrangements generated with Cas9. *Scientific Reports*, 7(1):1–8.

Boulton, S. J. and Jackson, S. P. (1996). Saccharomyces cerevisiae Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways. *The EMBO journal*, 15(18):5093–103.

Bradley, A., Evans, M., Kaufman, M. H., and Robertson, E. (1984). Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature*, 309(5965):255–256.

Brenneman, M. A., Wagener, B. M., Miller, C. A., Allen, C., and Nickoloff, J. A. (2002). XRCC3 controls the fidelity of homologous recombination: roles for XRCC3 in late stages of recombination. *Molecular cell*, 10(2):387–95.

Brenneman, M. A., Weiss, A. E., Nickoloff, J. A., and Chen, D. J. (2000). XRCC3 is required for efficient repair of chromosome breaks by homologous recombination. *Mutation research*, 459(2):89–97.

Brinkman, E. K., Chen, T., Amendola, M., and Van Steensel, B. (2014). Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Research*, 42(22):1–8.

Brinkman, E. K., Chen, T., de Haas, M., Holland, H. A., Akhtar, W., and van Steensel, B. (2018). Kinetics and Fidelity of the Repair of Cas9-Induced Double-Strand DNA Breaks. *Molecular Cell*.

Broach, J. R. (1982). The yeast plasmid $2\mu$ circle. *Cell*, 28(2):203–204.

Broll, S., Oumard, A., Hahn, K., Schambach, A., and Bode, J. (2010). Minicircle Performance Depending on S/MAR–Nuclear Matrix Interactions. *Journal of Molecular Biology*, 395(5):950–965.

Cai, Y., Bak, R. O., Krogh, L. B., Staunstrup, N. H., Moldt, B., Corydon, T. J., Schrøder, L. D., and Mikkelsen, J. G. (2014). DNA transposition by protein transduction of the piggyBac transposase from lentiviral Gag precursors. *Nucleic Acids Research*, 42(4).

Cain-Hom, C., Splinter, E., van Min, M., Simonis, M., van de Heijning, M., Martinez, M., Asghari, V., Cox, J., and Warming, S. (2017). Efficient mapping of transgene integration sites and local structural changes in Cre transgenic mice using targeted locus amplification. *Nucleic Acids Research*, 45(8):gkw1329.

Canela, A., Sridharan, S., Sciascia, N., Tubbs, A., Meltzer, P., Sleckman, B. P., and Nussenzweig, A. (2016). DNA Breaks and End Resection Measured Genome-wide by End Sequencing. *Molecular Cell*, 63(5):898–911.

Canver, M. C., Bauer, D. E., Dass, A., Yien, Y. Y., Chung, J., Masuda, T., Maeda, T., Paw, B. H., and Orkin, S. H. (2014). Characterization of Genomic Deletion Efficiency Mediated by CRISPR/Cas9 in Mammalian Cells. *The Journal of biological chemistry*.

Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G.-C. C., Zhang, F., Orkin, S. H., and Bauer, D. E. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*, 527(7577):192–197.

Casini, A., Olivieri, M., Petris, G., Montagna, C., Reginato, G., Maule, G., Lorenzin, F., Prandi, D., Romanel, A., Demichelis, F., Inga, A., and Cereseto, A. (2018). A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nature biotechnology*, 36(3):265–271.

Certo, M. T., Ryu, B. Y., Annis, J. E., Garibov, M., Jarjour, J., Rawlings, D. J., and Scharenberg, A. M. (2011). Tracking genome engineering outcome at individual DNA breakpoints. *Nature Methods*, 8(8):671–676.

Chaikind, B., Bessen, J. L., Thompson, D. B., Hu, J. H., and Liu, D. R. (2016). A programmable Cas9-serine recombinase fusion protein that operates on DNA sequences in mammalian cells. *Nucleic Acids Research*, 44(20):gkw707.

Chakrabarti, A. M., Henser-Brownhill, T., Monserrat, J., Poetsch, A. R., Luscombe, N. M., and Scaffidi, P. (2018). Target-Specific Precision of CRISPR-Mediated Genome Editing. *bioRxiv*, 0(0):387027.

Chang, H. H. Y., Pannunzio, N. R., Adachi, N., and Lieber, M. R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature Reviews Molecular Cell Biology*, 18(8):495–506.

Chang, H. H. Y., Watanabe, G., Gerodimos, C. A., Ochi, T., Blundell, T. L., Jackson, S. P., and Lieber, M. R. (2016). Different DNA End Configurations Dictate Which NHEJ Components Are Most Important for Joining Efficiency. *The Journal of biological chemistry*, 291(47):24377–24389.

Chavez, A., Pruitt, B. W., Tuttle, M., Shapiro, R. S., Cecchi, R. J., Winston, J., Turczyk, B. M., Tung, M., Collins, J. J., and Church, G. M. (2018). Precise Cas9 targeting enables genomic mutation prevention. *Proceedings of the National Academy of Sciences of the United States of America*, 115(14):3669–3673.

Chavez, A., Tuttle, M., Pruitt, B. W., Ewen-Campen, B., Chari, R., Ter-Ovanesyan, D., Haque, S. J., Cecchi, R. J., Kowal, E. J., Buchthal, J., Housden, B. E., Perrimon, N., Collins, J. J., and Church, G. (2016). Comparison of Cas9 activators in multiple species. *Nature Methods*, 13(7):563–567.

Chen, C.-h., Li, W., Xiao, T., Xu, H., Jiang, P., Meyer, C. A., Brown, M., and Liu, X. S. (2017). Integrative analysis and refined design of CRISPR knockout screens. *bioRxiv*, page 106534.

Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D., Monnat, R. J., and Stoddard, B. L. (2002). Design, Activity, and Structure of a Highly Specific Artificial Endonuclease. *Molecular Cell*, 10(4):895–905.

Chira, S., Jackson, C. S., Oprea, I., Ozturk, F., Pepper, M. S., Diaconu, I., Braicu, C., Raduly, L.-Z., Calin, G. A., and Berindan-Neagoe, I. (2015). Progresses towards safe and efficient gene therapy vectors. *Oncotarget*, 6(31):30675–30703.

Cho, S. W., Kim, S., Kim, J. M., and Kim, J.-S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature biotechnology*, 31(3):230–232.

Chu, V. T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., and Kühn, R. (2015). Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nature biotechnology*, 33(5):543–8.

Ciccia, A. and Elledge, S. J. (2010). The DNA damage response: making it safe to play with knives. *Molecular cell*, 40(2):179–204.

Clarke, R., Heler, R., MacDougall, M. S., Yeo, N. C., Chavez, A., Regan, M., Hanakahi, L., Church, G. M., Marraffini, L. A., and Merrill, B. J. (2018). Enhanced Bacterial Immunity and Mammalian Genome Editing via RNA-Polymerase-Mediated Dislodging of Cas9 from Double-Strand DNA Breaks. *Molecular cell*, 71(1):42–55.e8.

Cohen, S. N. (2013). DNA cloning: a personal view after 40 years. *Proceedings of the National Academy of Sciences of the United States of America*, 110(39):15521–9.

Cohen, S. N., Chang, A. C. Y., and Hsu, L. (1972). Nonchromosomal Antibiotic Resistance in Bacteria: Genetic Transformation of Escherichia coli by R-Factor DNA. *Proceedings of the National Academy of Sciences*, 69(8):2110–2114.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. a., and Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*, 339(6121):819–23.

Conway, A., Laganière, J., Paschon, D. E., Hacke, K., Kasahara, N., Gregory, P. D., Holmes, M. C., and Cost, G. J. (2014). HPRT As a Selectable Safe Harbor for Transgenesis. *Blood*, 124(21).

Daughtry, B. L., Rosenkrantz, J. L., Lazar, N. H., Fei, S. S., Redmayne, N., Torkenczy, K. A., Adey, A., Gao, L., Park, B., Nevonen, K. A., Carbone, L., and Chavez, S. L. (2018). Single-Cell Sequencing of Primate Preimplantation Embryos Reveals Chromosome Elimination Via Cellular Fragmentation and Blastomere Exclusion. *bioRxiv*, page 241851.

de Vree, P. J. P., de Wit, E., Yilmaz, M., van de Heijning, M., Klous, P., Verstegen, M. J. A. M., Wan, Y., Teunissen, H., Krijger, P. H. L., Geeven, G., Eijk, P. P., Sie, D., Ylstra, B., Hulsman, L. O. M., van Dooren, M. F., van Zutven, L. J. C. M., van den Ouweland, A., Verbeek, S., van Dijk, K. W., Cornelissen, M., Das, A. T., Berkhout, B., Sikkema-Raddatz, B., van den Berg, E., van der Vlies, P., Weening, D., den Dunnen, J. T., Matusiak, M., Lamkanfi, M., Ligtenberg, M. J. L., ter Brugge, P., Jonkers, J., Foekens, J. A., Martens, J. W., van der Luijt, R., van Amstel, H. K. P., van Min, M., Splinter, E., and de Laat, W. (2014). Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nature Biotechnology*, 32(10):1019–1025.

Deckbar, D., Birraux, J., Krempler, A., Tchouandong, L., Beucher, A., Walker, S., Stiff, T., Jeggo, P., and Löbrich, M. (2007). Chromosome breakage after G2 checkpoint release. *The Journal of cell biology*, 176(6):749–55.

Deem, A., Keszthelyi, A., Blackgrove, T., Vayl, A., Coffey, B., Mathur, R., Chabes, A., and Malkova, A. (2011). Break-induced replication is highly inaccurate. *PLoS biology*, 9(2):e1000594.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., and Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, 34(2):184–191.

Egli, D., Zuccaro, M. V., Kosicki, M., Church, G. M., Bradley, A., and Jasin, M. (2018). Inter-homologue repair in fertilized human eggs? *Nature*, 560(7717):E5–E7.

Elling, U., Wimmer, R. A., Leibbrandt, A., Burkard, T., Michlits, G., Leopoldi, A., Micheler, T., Abdeen, D., Zhuk, S., Aspalter, I. M., Handl, C., Liebergesell, J., Hubmann, M., Husa, A.-M., Kinzer, M., Schuller, N., Wetzel, E., van de Loo, N., Martinez, J. A. Z., Estoppey, D., Riedl, R., Yang, F., Fu, B., Dechat, T., Ivics, Z., Agu, C. A., Bell, O., Blaas, D., Gerhardt, H., Hoepfner, D., Stark, A., and Penninger, J. M. (2017). A reversible haploid mouse embryonic stem cell biobank resource for functional genomics. *Nature*, 550(7674):114.

Escribano-Díaz, C., Orthwein, A., Fradet-Turcotte, A., Xing, M., Young, J. T. F., Tkáč, J., Cook, M. A., Rosebrock, A. P., Munro, M., Canny, M. D., Xu, D., and Durocher, D. (2013). A cell cycle-dependent regulatory circuit composed of 53BP1-RIF1 and BRCA1-CtIP controls DNA repair pathway choice. *Molecular cell*, 49(5):872–83.

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., Janizek, J. D., Huang, X., Starita, L. M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, page 1.

Fonfara, I., Richter, H., BratoviÄ, M., Le Rhun, A., and Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature*, 532(7600):517–521.

Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M., and Joung, J. K. (2014). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature biotechnology*, 32(3):279–84.

Gallagher, D. N. and Haber, J. E. (2018). Repair of a Site-Specific DNA Cleavage: Old-School Lessons for Cas9-Mediated Gene Editing. *ACS Chemical Biology*, 13(2):397–405.

Gao, X., Tao, Y., Lamas, V., Huang, M., Yeh, W.-H., Pan, B., Hu, Y.-J., Hu, J. H., Thompson, D. B., Shu, Y., Li, Y., Wang, H., Yang, S., Xu, Q., Polley, D. B., Liberman, M. C., Kong, W.-J., Holt, J. R., Chen, Z.-Y., and Liu, D. R. (2017). Treatment of autosomal dominant hearing loss by in vivo delivery of genome editing agents. *Nature*, 553(7687):217–221.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 109(39):E2579–86.

Gasperini, M., Findlay, G. M., McKenna, A., Milbank, J. H., Lee, C., Zhang, M. D., Cusanovich, D. A., and Shendure, J. (2017). CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *American Journal of Human Genetics*, 101(2):192–205.

Gasperini, M., Hill, A., McFaline-Figueroa, J. L., Martin, B., Trapnell, C., Ahituv, N., and Shendure, J. (2018). crisprQTL mapping as a genome-wide association framework for cellular genetic screens. *bioRxiv*, page 314344.

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I., and Liu, D. R.

(2017). Programmable base editing of A • T to G • C in genomic DNA without DNA cleavage. *Nature Publishing Group*, 551(7681):464–471.

Georges-Labouesse, E., Messaddeq, N., Yehia, G., Cadalbert, L., Dierich, A., and Le Meur, M. (1996). Absence of integrin $\alpha$6 leads to epidermolysis bullosa and neonatal death in mice. *Nature Genetics*, 13(3):370–373.

Giannoukos, G., Ciulla, D. M., Marco, E., Abdulkerim, H. S., Barrera, L. A., Bothmer, A., Dhanapal, V., Gloskowski, S. W., Jayaram, H., Maeder, M. L., Skor, M. N., Wang, T., Myer, V. E., and Wilson, C. J. (2018). UDiTaS™, a genome editing detection method for indels and genome rearrangements. *BMC Genomics*, 19(1):212.

Gilbert, L., Horlbeck, M., Adamson, B., Villalta, J., Chen, Y., Whitehead, E., Guimaraes, C., Panning, B., Ploegh, H., Bassik, M., Qi, L., Kampmann, M., and Weissman, J. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*, 159(3):647–661.

Golic, K. G. and Lindquist, S. (1989). The FLP recombinase of yeast catalyzes site-specific recombination in the drosophila genome. *Cell*, 59(3):499–509.

Goto, T. and Wang, J. C. (1982). Yeast DNA topoisomerase II. An ATP-dependent type II topoisomerase that catalyzes the catenation, decatenation, unknotting, and relaxation of double-stranded DNA rings. *The Journal of biological chemistry*, 257(10):5866–72.

Guilinger, J. P., Thompson, D. B., and Liu, D. R. (2014). Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nature biotechnology*, 32(6):577–582.

Gutschner, T., Haemmerle, M., Genovese, G., Draetta, G. F., and Chin, L. (2016). Post-translational Regulation of Cas9 during G1 Enhances Homology-Directed Repair. *Cell Reports*, 14(6):1555–1566.

Haapaniemi, E., Botla, S., Persson, J., Schmierer, B., and Taipale, J. (2018). CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. *Nature Medicine*, page 1.

Hacein-Bey-Abina, S., von Kalle, C., Schmidt, M., Le Deist, F., Wulffraat, N., McIntyre, E., Radford, I., Villeval, J.-L., Fraser, C. C., Cavazzana-Calvo, M., and Fischer, A. (2003). A Serious Adverse Event after Successful Gene Therapy for X-Linked Severe Combined Immunodeficiency. *New England Journal of Medicine*, 348(3):255–256.

Hansen, A. S. and O'Shea, E. K. (2015). Cis Determinants of Promoter Threshold and Activation Timescale. *Cell Reports*, 12(8):1226–1233.

Hendel, A., Bak, R. O., Clark, J. T., Kennedy, A. B., Ryan, D. E., Roy, S., Steinfeld, I., Lunstad, B. D., Kaiser, R. J., Wilkens, A. B., Bacchetta, R., Tsalenko, A., Dellinger, D., Bruhn, L., and Porteus, M. H. (2015). Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nature Biotechnology*, 33(9):985–989.

Hicks, W. M., Kim, M., and Haber, J. E. (2010). Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. *Science (New York, N.Y.)*, 329(5987):82–5.

Hill, J. T., Demarest, B. L., Bisgrove, B. W., Su, Y. C., Smith, M., and Yost, H. J. (2014). Poly peak parser: Method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. *Developmental Dynamics*, 243(12):1632–1636.

Hirano, H., Gootenberg, J. S., Horii, T., Abudayyeh, O. O., Kimura, M., Hsu, P. D., Nakane, T., Ishitani, R., Hatada, I., Zhang, F., Nishimasu, H., and Nureki, O. (2016). Structure and Engineering of Francisella novicida Cas9. *Cell*, 164(5):950–961.

Hong, Y. and Stambrook, P. J. (2004). Restoration of an absent G1 arrest and protection from apoptosis in embryonic stem cells after ionizing radiation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14443–14448.

Horlbeck, M. A., Witkowsky, L. B., Guglielmi, B., Replogle, J. M., Gilbert, L. A., Villalta, J. E., Torigoe, S. E., Tjian, R., and Weissman, J. S. (2016). Nucleosomes impede cas9 access to DNA in vivo and in vitro. *eLife*, 5(MARCH2016).

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G., and Zhang, F. (2013a). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, 31(9):827–832.

Hsu, P. D., Scott, D. a., Weinstein, J. a., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. a., Bao, G., and Zhang, F. (2013b). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Nature biotechnology*, 31(9):827–32.

Hu, J. H., Miller, S. M., Geurts, M. H., Tang, W., Chen, L., Sun, N., Zeina, C. M., Gao, X., Rees, H. A., Lin, Z., and Liu, D. R. (2018). Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature*, 556(7699):57–63.

Huang, L. C., Clarkin, K. C., and Wahl, G. M. (1996). Sensitivity and selectivity of the DNA damage sensor responsible for activating p53-dependent G1 arrest. *Proceedings of the National Academy of Sciences of the United States of America*, 93(10):4827–32.

Huang, Y., McCann, C., Samsonov, A., Malkov, D., Davis, G. D., and Ji, Q. (2017). Modulation of Genome Editing Outcomes by Cell Cycle Control of Cas9 Expression. *bioRxiv*, page 127068.

Iliakis, G. (2009). Backup pathways of NHEJ in cells of higher eukaryotes: Cell cycle dependence. *Radiotherapy and Oncology*, 92(3):310–315.

Ira, G., Pellicioli, A., Balijja, A., Wang, X., Fiorani, S., Carotenuto, W., Liberi, G., Bressan, D., Wan, L., Hollingsworth, N. M., Haber, J. E., and Foiani, M. (2004). DNA end resection, homologous recombination and DNA damage checkpoint activation require CDK1. *Nature*, 431(7011):1011–7.

Iyer, V., Boroviak, K., Thomas, M., Doe, B., Ryder, E., and Adams, D. (2018). No unexpected CRISPR-Cas9 off-target activity revealed by trio sequencing of gene-edited mice. *bioRxiv*, page 263129.

Jasin, M. and Rothstein, R. (2013). Repair of Strand Breaks by Homologous Recombination. *Cold Spring Harbor Perspectives in Biology*, 5 VN - re(11):1–19.

Jensen, R. B., Carreira, A., and Kowalczykowski, S. C. (2010). Purified human BRCA2 stimulates RAD51-mediated recombination. *Nature*, 467(7316):678–83.

Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., Nogales, E., and Doudna, J. A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science (New York, N.Y.)*, 351(6275):867–871.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A Programmable Dual-RNA – Guided DNA Endonuclease in Adaptice Bacterial Immunity. *Science (New York, N.Y.)*, 337(August):816–822.

Johnson, R. D. and Jasin, M. (2000). Sister chromatid gene conversion is a prominent double-strand break repair pathway in mammalian cells. *The EMBO Journal*, 19(13):3398–407.

Julius, M. H., Masuda, T., and Herzenberg, L. A. (1972). Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter. *Proceedings of the National Academy of Sciences of the United States of America*, 69(7):1934–8.

Kearns, N. A., Pham, H., Tabak, B., Genga, R. M., Silverstein, N. J., Garber, M., and Maehr, R. (2015). Functional annotation of native enhancers with a Cas9–histone demethylase fusion. *Nature Methods*, 12(5):401–403.

Keeney, S. and Kleckner, N. (1995). Covalent protein-DNA complexes at the 5' strand termini of meiosis-specific double-strand breaks in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 92(24):11274–8.

Kent, T., Mateos-Gomez, P. A., Sfeir, A., and Pomerantz, R. T. (2016). Polymerase $\theta$ is a robust terminal transferase that oscillates between three different mechanisms during end-joining. *eLife*, 5(JUN2016):1–25.

Kim, H. K., Song, M., Lee, J., Menon, A. V., Jung, S., Kang, Y.-M., Choi, J. W., Woo, E., Koh, H. C., Nam, J.-W., and Kim, H. (2017a). In vivo high-throughput profiling of CRISPR–Cpf1 activity. *Nature Methods*, 14(2):153–159.

Kim, S., Bae, T., Hwang, J., and Kim, J.-S. (2017b). Rescue of high-specificity Cas9 variants using sgRNAs with matched 5' nucleotides. *Genome Biology*, 18(1):218.

Kim, S., Kim, D., Cho, S. W., Jungeun, K., Kim, J. S. J. J.-S. S., and Kim, J. S. J. J.-S. S. (2014). Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Research*, 24(6):1012–1019.

Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z., and Joung, J. K. (2016). High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, 529(7587):490–495.

Kleinstiver, B. P., Prew, M. S., Tsai, S. Q., Topkar, V. V., Nguyen, N. T., Zheng, Z., Gonzales, A. P. W., Li, Z., Peterson, R. T., Yeh, J.-R. J. R. J., Aryee, M. J., and Joung, J. K. (2015). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, 523(7561):481–485.

Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C., and Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature biotechnology*, 32(3):267–73.

Koller, B. H., Hagemann, L. J., Doetschman, T., Hagaman, J. R., Huang, S., Williams, P. J., First, N. L., Maeda, N., and Smithies, O. (1989). Germ-line transmission of a planned alteration made in a hypoxanthine phosphoribosyltransferase gene by homologous recombination in embryonic stem cells.

*Proceedings of the National Academy of Sciences*, 86(22):8927–8931.

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., and Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, 61(16):5985–91.

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., Nureki, O., and Zhang, F. (2014). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, 517(7536):583–8.

Korkmaz, G., Lopes, R., Ugalde, A. P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R., and Agami, R. (2016). Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nature biotechnology*, (August 2015):1–10.

Kosicki, M., Rajan, S. S., Lorenzetti, F. C., Wandall, H. H., Narimatsu, Y., Metzakopian, E., and Bennett, E. P. (2017). Dynamics of Indel Profiles Induced by Various CRISPR/Cas9 Delivery Methods. *Progress in Molecular Biology and Translational Science*, 152:49–67.

Kosicki, M., Tomberg, K., and Bradley, A. (2018). Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nature Biotechnology*, 36(8):765.

Kraft, K., Geuer, S., Will, A. J., Chan, W., Paliou, C., Borschiwer, M., Harabula, I., Wittler, L., Franke, M., Ibrahim, D. M., Kragesteen, B. K., Spielmann, M., Mundlos, S., Lupiáñez, D. G., and Andrey, G. (2015). Deletions, inversions, duplications: Engineering of structural variants using CRISPR/Cas in mice. *Cell Reports*, 10(5):833–839.

Le Naour, F. and Boucheix, C. (2000). Severely Reduced Female Fertility in CD9-Deficient Mice. *Science*, 287(5451):319–321.

Lee, E.-C., Liang, Q., Ali, H., Bayliss, L., Beasley, A., Bloomfield-Gerdes, T., Bonoli, L., Brown, R., Campbell, J., Carpenter, A., Chalk, S., Davis, A., England, N., Fane-Dremucheva, A., Franz, B., Germaschewski, V., Holmes, H., Holmes, S., Kirby, I., Kosmac, M., Legent, A., Lui, H., Manin, A., O'Leary, S., Paterson, J., Sciarrillo, R., Speak, A., Spensberger, D., Tuffery, L., Waddell, N., Wang, W., Wells, S., Wong, V., Wood, A., Owen, M. J., Friedrich, G. a., and Bradley, A. (2014). Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nature biotechnology*, 32(4):356–63.

Leenay, R. T., Aghazadeh, A., Hiatt, J., Tse, D., Hulquist, J., Krogan, N., Wu, Z., Marson, A., May, A. P., and Zou, J. (2018). Systematic characterization of genome editing in primary T cells reveals proximal genomic insertions and enables machine learning prediction of CRISPR-Cas9 DNA repair outcomes. *bioRxiv*, page 404947.

Lemos, B. R., Kaplan, A. C., Bae, J. E., Ferrazzoli, A. E., Kuo, J., Anand, R. P., Waterman, D. P., and Haber, J. E. (2018). CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. *Proceedings of the National Academy of Sciences*, page 201716855.

Lensing, S. V., Marsico, G., Hänsel-Hertsch, R., Lam, E. Y., Tannahill, D., and Balasubramanian, S. (2016). DSBCapture: in situ capture and sequencing of DNA breaks. *Nature Methods*, 13(10):855–7.

Lessard, S., Francioli, L., Alfoldi, J., Tardif, J.-C., Ellinor, P. T., MacArthur, D. G., Lettre, G., Orkin, S. H., and Canver, M. C. (2017). Human genetic variation alters CRISPR-Cas9 on- and off-targeting specificity at therapeutically implicated loci. *Proceedings of the National Academy of Sciences*, 114(52):E11257–E11266.

Levy, J. E., Jin, O., Fujiwara, Y., Kuo, F., and Andrews, N. C. (1999). Transferrin receptor is necessary for development of erythrocytes and the nervous system. *Nature Genetics*, 21(4):396–399.

Lin, F. L. and Sternberg, N. (1984). Homologous recombination between overlapping thymidine kinase gene fragments stably inserted into a mouse cell genome. *Molecular and Cellular Biology*, 4(5):852–861.

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

Listgarten, J., Weinstein, M., Kleinstiver, B. P., Sousa, A. A., Joung, J. K., Crawford, J., Gao, K., Hoang, L., Elibol, M., Doench, J. G., and Fusi, N. (2018). Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nature Biomedical Engineering*, 2(1):38–47.

Liu, X., Homma, A., Sayadi, J., Yang, S., Ohashi, J., and Takumi, T. (2016). Sequence features associated with the cleavage efficiency of CRISPR/Cas9 system. *Scientific Reports*, 6:1–9.

Liu, Z., Gerner, M. Y., Van Panhuys, N., Levine, A. G., Rudensky, A. Y., and Germain, R. N. (2015). Immune homeostasis enforced by co-localized effector and regulatory T cells. *Nature*, 528(7581):225–230.

Lonowski, L. A., Narimatsu, Y., Riaz, A., Delay, C. E., Yang, Z., Niola, F., Duda, K., Ober, E. A., Clausen, H., Wandall, H. H., Hansen, S. H., Bennett, E. P., and Frödin, M. (2017). Genome editing using FACS enrichment of nuclease-expressing cells and indel detection by amplicon analysis. *Nature protocols*, 12(3):581–603.

Luo, J., Sun, X., Cormack, B. P., and Boeke, J. D. (2018a). Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. *Nature*, 560(7718):392–396.

Luo, X., He, Y., Zhang, C., He, X., Yan, L., Li, M., Hu, T., Hu, Y., Jiang, J., Meng, X., Ji, W., Zhao, X., Zheng, P., Xu, S., and Su, B. (2018b). Trio deep-sequencing does not reveal unexpected mutations in Cas9-edited monkeys. *bioRxiv*, page 339143.

Ma, H., Marti-Gutierrez, N., Park, S.-W., Wu, J., Hayama, T., Darby, H., Van Dyken, C., Li, Y., Koski, A., Liang, D., Suzuki, K., Gu, Y., Gong, J., Xu, X., Ahmed, R., Lee, Y., Kang, E., Ji, D., Park, A.-R., Kim, D., Kim, S.-T., Heitner, S. B., Battaglia, D., Krieg, S. A., Lee, D. M., Wu, D. H., Wolf, D. P., Amato, P., Kaul, S., Belmonte, J. C. I., Kim, J.-S., and Mitalipov, S. (2018). Ma et al. reply. *Nature*, 560(7717):E10–E23.

Ma, M., Zhuang, F., Hu, X., Wang, B., Wen, X. Z., Ji, J. F., and Xi, J. J. (2017). Efficient generation of mice carrying homozygous double-floxp alleles using the Cas9-Avidin/Biotin-donor DNA system. *Cell Research*, 27(4):578–581.

Mahaney, B. L., Meek, K., and Lees-Miller, S. P. (2009). Repair of ionizing radiation-induced DNA double-strand breaks by non-homologous end-joining. *The Biochemical journal*, 417(3):639–50.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., and Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science*, 339(6121):823–826.

Mansour, S. L., Thomas, K. R., and Capecchi, M. R. (1988). Disruption of the proto-oncogene int-2 in mouse embryo-derived stem cells: a general strategy for targeting mutations to non-selectable genes. *Nature*, 336(6197):348–52.

Marrero, V. A. and Symington, L. S. (2010). Extensive DNA end processing by exo1 and sgs1 inhibits break-induced replication. *PLoS genetics*, 6(7):e1001007.

Maruyama, T., Dougan, S. K., Truttmann, M. C., Bilate, A. M., Ingram, J. R., and Ploegh, H. L. (2015). Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nature Biotechnology*, 33(5):538–542.

Mateos-Gomez, P. A., Kent, T., Deng, S. K., Mcdevitt, S., Kashkina, E., Hoang, T. M., Pomerantz, R. T., and Sfeir, A. (2017). The helicase domain of Pol$\theta$ counteracts RPA to promote alt-NHEJ. *Nature Structural and Molecular Biology*, 24(12):1116–1123.

Mehta, A. and Haber, J. E. (2014). Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harbor perspectives in biology*, 6(9):a016428.

Metzakopian, E., Strong, A., Iyer, V., Hodgkins, A., Tzelepis, K., Antunes, L., Friedrich, M. J., Kang, Q., Davidson, T., Lamberth, J., Hoffmann, C., Davis, G. D., Vassiliou, G. S., Skarnes, W. C., and Bradley, A. (2017). Enhancing the genome editing toolbox: Genome wide CRISPR arrayed libraries. *Scientific Reports*, 7(1):2244.

Mitra, R., Fain-Thornton, J., and Craig, N. L. (2008). piggyBac can bypass DNA synthesis during cut and paste transposition. *The EMBO journal*, 27(7):1097–109.

Miyata, T., Takeda, J., Iida, Y., Yamada, N., Inoue, N., Takahashi, M., Maeda, K., Kitani, T., and Kinoshita, T. (1993). The cloning of PIG-A, a component in the early step of GPI-anchor biosynthesis. *Science (New York, N.Y.)*, 259(5099):1318–20.

Mladenov, E. and Iliakis, G. (2011). Induction and repair of DNA double strand breaks: The increasing spectrum of non-homologous end joining pathways. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 711(1-2):61–72.

Moreno-Mateos, M. a., Vejnar, C. E., Beaudoin, J.-d., Fernandez, J. P., Mis, E. K., Khokha, M. K., and Giraldez, A. J. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nature methods*, 12(10):982–8.

Moscou, M. J. and Bogdanove, A. J. (2009). A simple cipher governs DNA recognition by TAL effectors. *Science (New York, N.Y.)*, 326(5959):1501.

Nassif, N., Penney, J., Pal, S., Engels, W. R., and Gloor, G. B. (1994). Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. *Molecular and cellular biology*, 14(3):1613–25.

Nygaard, S., Barzel, A., Haft, A., Major, A., Finegold, M., Kay, M. A., and Grompe, M. (2016). A universal system to select gene-modified hepatocytes in vivo. *Science translational medicine*, 8(342):342ra79.

Paquet, D., Kwart, D., Chen, A., Sproul, A., Jacob, S., Teo, S., Olsen, K. M., Gregg, A., Noggle, S., and Tessier-Lavigne, M. (2016). Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature*, 533(7601):125–129.

Parikh, B. A., Beckman, D. L., Patel, S. J., and White, J. M. (2015). Detailed Phenotypic and Molecular Analyses of Genetically Modified Mice Generated by CRISPR-Cas9-Mediated Editing. *PloS one*, pages 1–28.

Perrault, R., Wang, H., Wang, M., Rosidi, B., and Iliakis, G. (2004). Backup pathways of NHEJ are suppressed by DNA-PK. *Journal of Cellular Biochemistry*, 92(4):781–794.

Pettitt, S. J., Liang, Q., Rairdan, X. Y., Moran, J. L., Prosser, H. M., Beier, D. R., Lloyd, K. C., Bradley, A., and Skarnes, W. C. (2009). Agouti C57BL/6N embryonic stem cells for mouse genetic resources. *Nature methods*, 6(7):493–495.

Pierce, A. J., Johnson, R. D., Thompson, L. H., and Jasin, M. (1999). XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. *Genes & development*, 13(20):2633–8.

Platt, R., Chen, S., Zhou, Y., Yim, M., Swiech, L., Kempton, H., Dahlman, J., Parnas, O., Eisenhaure, T., Jovanovic, M., Graham, D., Jhunjhunwala, S., Heidenreich, M., Xavier, R., Langer, R., Anderson, D., Hacohen, N., Regev, A., Feng, G., Sharp, P., and Zhang, F. (2014). CRISPR-Cas9 Knockin Mice for Genome Editing and Cancer Modeling. *Cell*, 159(2):440–455.

Plessis, A., Perrin, A., Haber, J. E., and Dujon, B. (1992). Site-specific recombination determined by I-SceI, a mitochondrial group I intron-encoded endonuclease expressed in the yeast nucleus. *Genetics*, 130(3):451–60.

Polstein, L. R., Perez-Pinera, P., Kocak, D. D., Vockley, C. M., Bledsoe, P., Song, L., Safi, A., Crawford, G. E., Reddy, T. E., and Gersbach, C. A. (2015). Genome-wide specificity of DNA binding, gene regulation, and chromatin remodeling by TALE- and CRISPR/Cas9-based transcriptional activators. *Genome Research*, 25(8):1158–1169.

Prasher, D. C., Eckenrode, V. K., Ward, W. W., Prendergast, F. G., and Cormier, M. J. (1992). Primary structure of the Aequorea victoria green-fluorescent protein. *Gene*, 111(2):229–233.

Ramakrishna, S., Kwaku Dad, A.-B., Beloor, J., Gopalappa, R., Lee, S.-K., and Kim, H. (2014). Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. *Genome research*, 24(6):1020–7.

Ramírez-Solis, R., Davis, A. C., and Bradley, A. (1993). Gene targeting in embryonic stem cells. *Methods in enzymology*, 225:855–78.

Ran, F. A., Hsu, P. D., Lin, C.-Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., Scott, D. a., Inoue, A., Matoba, S., Zhang, Y., and Zhang, F. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, 154(6):1380–1389.

Redman, M., King, A., Watson, C., and King, D. (2016). What is CRISPR/Cas9? *Archives of disease in childhood. Education and practice edition*, 101(4):213–5.

Richardson, C. and Jasin, M. (2000). Frequent chromosomal translocations induced by DNA double-strand breaks. *Nature*, 405(51):697–700.

Richardson, C. D., Kazane, K. R., Feng, S. J., Zelin, E., Bray, N. L., Schäfer, A. J., Floor, S. N., and Corn, J. E. (2018). CRISPR–Cas9 genome editing in human cells occurs via the Fanconi anemia pathway. *Nature Genetics*, 50(8):1132–1139.

Richardson, C. D., Ray, G. J., Bray, N. L., and Corn, J. E. (2016). Non-homologous DNA increases gene disruption efficiency by altering DNA repair outcomes. *Nature Communications*, 7:1–7.

Riesenberg, S. and Maricic, T. (2018). Targeting repair pathways with small molecules increases precise genome editing in pluripotent stem cells. *Nature communications*, 9(1):2164.

Roberts, S. A., Sterling, J., Thompson, C., Harris, S., Mav, D., Shah, R., Klimczak, L. J., Kryukov, G. V., Malc, E., Mieczkowski, P. A., Resnick, M. A., and Gordenin, D. A. (2012). Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. *Molecular Cell*, 46(4):424–435.

Roos, W. P. and Kaina, B. (2006). DNA damage-induced cell death by apoptosis. *Trends in Molecular Medicine*, 12(9):440–450.

Rosen, L. E., Morrison, H. A., Masri, S., Brown, M. J., Springstubb, B., Sussman, D., Stoddard, B. L., and Seligman, L. M. (2006). Homing endonuclease I-CreI derivatives with novel DNA target specificities. *Nucleic acids research*, 34(17):4791–800.

Rouet, P., Smih, F., and Jasin, M. (1994). Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, 91(13):6064–8.

Ruff, P., Donnianni, R. A., Glancy, E., Oh, J., and Symington, L. S. (2016). RPA Stabilization of Single-Stranded DNA Is Critical for Break-Induced Replication. *Cell Reports*, 17(12):3359–3368.

Sadhu, M. J., Bloom, J. S., Day, L., and Kruglyak, L. (2016). CRISPR-directed mitotic recombination enables genetic mapping without crosses. *Science*, 352(6289):1113–1116.

Santaguida, S., Richardson, A., Iyer, D. R., M'Saad, O., Zasadil, L., Knouse, K. A., Wong, Y. L., Rhind, N., Desai, A., and Amon, A. (2017). Chromosome Mis-segregation Generates Cell-Cycle-Arrested Cells with Complex Karyotypes that Are Eliminated by the Immune System. *Developmental cell*, 41(6):638–651.e5.

Santarosa, M. and Ashworth, A. (2004). Haploinsufficiency for tumour suppressor genes: When you don't need to go all the way.

Savic, N., Ringnalda, F. C., Lindsay, H., Berk, C., Bargsten, K., Li, Y., Neri, D., Robinson, M. D., Ciaudo, C., Hall, J., Jinek, M., and Schwank, G. (2018). Covalent linkage of the DNA repair template to the CRISPR-Cas9 nuclease enhances homology-directed repair. *eLife*, 7:e33761.

Schaft, J., Ashery-Padan, R., van der Hoeven, F., Gruss, P., and Stewart, A. F. (2001). Efficient FLP recombination in mouse ES cells and oocytes. *Genesis (New York, N.Y. : 2000)*, 31(1):6–10.

Schröder, A. R., Shinn, P., Chen, H., Berry, C., Ecker, J. R., and Bushman, F. (2002). HIV-1 Integration in the Human Genome Favors Active Genes and Local Hotspots. *Cell*, 110(4):521–529.

Seligman, L. M., Chisholm, K. M., Chevalier, B. S., Chadsey, M. S., Edwards, S. T., Savage, J. H., and Veillet, A. L. (2002). Mutations altering the cleavage specificity of a homing endonuclease. *Nucleic acids research*, 30(17):3870–9.

Seol, J.-H., Shim, E. Y., and Lee, S. E. (2018). Microhomology-mediated end joining: Good, bad and ugly. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 809:81–87.

Sfeir, A. and Symington, L. S. (2015). Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends in biochemical sciences*, 40(11):701–714.

Shao, Y., Lu, N., Wu, Z., Cai, C., Wang, S., Zhang, L.-L., Zhou, F., Xiao, S., Liu, L., Zeng, X., Zheng, H., Yang, C., Zhao, Z., Zhao, G., Zhou, J.-Q., Xue, X., and Qin, Z. (2018). Creating a functional single-chromosome yeast. *Nature*, 560(7718):331–335.

Sharma, S., Javadekar, S. M., Pandey, M., Srivastava, M., Kumari, R., and Raghavan, S. C. (2015). Homology and enzymatic requirements of microhomology-dependent alternative end joining. *Cell death & disease*, 6(3):e1697.

Shimomura, O., Johnson, F. H., and Saiga, Y. (1962). Extraction, Purification and Properties of Aequorin, a Bioluminescent Protein from the Luminous Hydromedusan,Aequorea. *Journal of Cellular and Comparative Physiology*, 59(3):223–239.

Shin, H. Y., Wang, C., Lee, H. K., Yoo, K. H., Zeng, X., Kuhns, T., Yang, C. M., Mohr, T., Liu, C., and Hennighausen, L. (2017). CRISPR/Cas9 targeting events cause complex deletions and insertions at 17 sites in the mouse genome. *Nature Communications*, 8(May):1–10.

Shou, J., Li, J., Liu, Y., and Wu, Q. (2018). Precise and Predictable CRISPR Chromosomal Rearrangements Reveal Principles of Cas9-Mediated Nucleotide Insertion. *Molecular cell*, 71(4):498–509.e4.

Simonelli, V., Narciso, L., Dogliotti, E., and Fortini, P. (2005). Base excision repair intermediates are mutagenic in mammalian cells. *Nucleic acids research*, 33(14):4404–11.

Sinha, S., Li, F., Villarreal, D., Shim, J. H., Yoon, S., Myung, K., Shim, E. Y., and Lee, S. E. (2017). Microhomology-mediated end joining induces hypermutagenesis at breakpoint junctions. *PLoS Genetics*, 13(4):1–25.

Skarnes, W. C., Rosen, B., West, A. P., Koutsourakis, M., Bushell, W., Iyer, V., Mujica, A. O., Thomas, M., Harrow, J., Cox, T., Jackson, D., Severin, J., Biggs, P., Fu, J., Nefedov, M., de Jong, P. J., Stewart, A. F., and Bradley, A. (2011). A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474(7351):337–344.

Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. a., Yan, W. X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. *Science (New York, N.Y.)*, 351(6268):84–88.

Smithies, O., Gregg, R. G., Boggs, S. S., Koralewski, M. A., and Kucherlapati, R. S. (1985). Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. *Nature*, 317(6034):230–4.

Sotiriou, S. K., Kamileri, I., Lugli, N., Evangelou, K., Da-Ré, C., Huber, F., Padayachy, L., Tardy, S., Nicati, N. L., Barriot, S., Ochs, F., Lukas, C.,

Lukas, J., Gorgoulis, V. G., Scapozza, L., and Halazonetis, T. D. (2016). Mammalian RAD52 Functions in Break-Induced Replication Repair of Collapsed DNA Replication Forks. *Molecular Cell*, 64(6):1127–1134.

Starita, L. M., Islam, M. M., Banerjee, T., Adamovich, A. I., Gullingsrud, J., Fields, S., Shendure, J., and Parvin, J. D. (2018). A Multiplex Homology-Directed DNA Repair Assay Reveals the Impact of More Than 1,000 BRCA1 Missense Substitution Variants on Protein Function. *American journal of human genetics*, 0(0).

Stark, J. M., Pierce, A. J., Oh, J., Pastink, A., and Jasin, M. (2004). Genetic steps of mammalian homologous repair with distinct mutagenic consequences. *Molecular and cellular biology*, 24(21):9305–16.

Stemmer, M., Thumberger, T., del Sol Keyer, M., Wittbrodt, J., and Mateo, J. L. (2015). CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PLOS ONE*, 10(4):e0124633.

Stephenson, A. A., Raper, A. T., and Suo, Z. (2018). Bidirectional Degradation of DNA Cleavage Products Catalyzed by CRISPR/Cas9. *Journal of the American Chemical Society*, 140(10):3743–3750.

Sternberg, N. and Hamilton, D. (1981). Bacteriophage P1 site-specific recombination: I. Recombination between loxP sites. *Journal of Molecular Biology*, 150(4):467–486.

Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C., and Doudna, J. A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, 507(7490):62–67.

Stojic, L., Lun, A. T., Mangei, J., Mascalchi, P., Quarantotti, V., Barr, A. R., Bakal, C., Marioni, J. C., Gergely, F., and Odom, D. T. (2018). Specificity of RNAi, LNA and CRISPRi as loss-of-function methods in transcriptional analysis. *Nucleic Acids Research*.

Strogantsev, R., Krueger, F., Yamazawa, K., Shi, H., Gould, P., Goldman-roberts, M., Mcewen, K., Sun, B., Pedersen, R., and Ferguson-smith, A. C. (2015). Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression. *Genome Biology*, pages 1–18.

Sugawara, N. and Haber, J. E. (2012). Monitoring DNA recombination initiated by HO endonuclease. *Methods in Molecular Biology*, 920:349–370.

Sung, P. and Klein, H. (2006). Mechanism of homologous recombination: Mediators and helicases take on regulatory functions.

Sussman, D., Chadsey, M., Fauce, S., Engel, A., Bruett, A., Monnat, R., Stoddard, B. L., and Seligman, L. M. (2004). Isolation and Characterization of New Homing Endonuclease Specificities at Individual Target Site Positions. *Journal of Molecular Biology*, 342(1):31–41.

Sutherland, D. R., Kuek, N., Davidson, J., Barth, D., Chang, H., Yeo, E., Bamford, S., Chin-Yee, I., and Keeney, M. (2007). Diagnosing PNH with FLAER and multiparameter flow cytometry. *Cytometry. Part B, Clinical cytometry*, 72(3):167–177.

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., and Stahl, F. W. (1983). The double-strand-break repair model for recombination. *Cell*, 33(1):25–35.

Tadi, S. K., Sebastian, R., Dahal, S., Babu, R. K., Choudhary, B., and Raghavan, S. C. (2016). Microhomology-mediated end joining is the principal mediator of double-strand break repair during mitochondrial DNA lesions. *Molecular Biology of the Cell*, 27(2):223–235.

Taheri-Ghahfarokhi, A., Taylor, B. J., Nitsch, R., Lundin, A., Cavallo, A.-L., Madeyski-Bengtson, K., Karlsson, F., Clausen, M., Hicks, R., Mayr, L. M., Bohlooly-Y, M., and Maresca, M. (2018). Decoding non-random mutational signatures at Cas9 targeted sites. *Nucleic Acids Research*.

Takeda, J., Miyata, T., Kawagoe, K., Iida, Y., Endo, Y., Fujita, T., Takahashi, M., Kitani, T., and Kinoshita, T. (1993). Deficiency of the GPI anchor caused by a somatic mutation of the PIG-A gene in paroxysmal nocturnal hemoglobinuria. *Cell*, 73(4):703–711.

Tan, E. P., Li, Y., Del Castillo Velasco-Herrera, M., Yusa, K., and Bradley, A. (2015). Off-target assessment of CRISPR-Cas9 guiding RNAs in human iPS and mouse ES cells. *Genesis*, 53(2):225–236.

Taylor, J. and Woodcock, S. (2015). A Perspective on the Future of High-Throughput RNAi Screening: Will CRISPR Cut Out the Competition or Can RNAi Help Guide the Way? *Journal of Biomolecular Screening*, 20(8):1040–1051.

Thakore, P. I., D'Ippolito, A. M., Song, L., Safi, A., Shivakumar, N. K., Kabadi, A. M., Reddy, T. E., Crawford, G. E., and Gersbach, C. a. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature methods*, 12(12):1143–9.

Thomas, K. R. and Capecchi, M. R. (1987). Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell*, 51(3):503–512.

Thompson, S., Clarke, A. R., Pow, A. M., Hooper, M. L., and Melton, D. W. (1989). Germ line transmission and expression of a corrected HPRT gene produced by gene targeting in embryonic stem cells. *Cell*, 56(2):313–321.

Tichy, E. D., Pillai, R., Deng, L., Liang, L., Tischfield, J., Schwemberger, S. J., Babcock, G. F., and Stambrook, P. J. (2010). Mouse Embryonic Stem Cells, but Not Somatic Cells, Predominantly Use Homologous Recombination to Repair Double-Strand DNA Breaks. *Stem Cells and Development*, 19(11):1699–1711.

Tsai, S. Q. and Joung, J. K. (2016). Defining and improving the genome-wide specificities of CRISPR–Cas9 nucleases. *Nature Reviews Genetics*, 17(5):300–312.

Tsai, S. Q., Nguyen, N. T., Malagon-Lopez, J., Topkar, V. V., Aryee, M. J., and Joung, J. K. (2017). CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nature Methods*, 14(6):607–614.

Tsai, S. Q., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A. J., Le, L. P., Aryee, M. J., and Joung, J. K. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature biotechnology*, 33(2):187–197.

Urnov, F. D., Miller, J. C., Lee, Y.-L., Beausejour, C. M., Rock, J. M., Augustus, S., Jamieson, A. C., Porteus, M. H., Gregory, P. D., and Holmes, M. C. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*, 435(7042):646–51.

Uusi-Mäkelä, M. I., Barker, H. R., Bäuerlein, C. A., Häkkinen, T., Nykter, M., and Rämet, M. (2018). Chromatin accessibility is associated with CRISPR-Cas9 efficiency in the zebrafish (Danio rerio). *PLoS ONE*, 13(4):e0196238.

van der Lugt, N., Maandag, E. R., te Riele, H., Laird, P. W., and Berns, A. (1991). A pgk::hprt fusion as a selectable marker for targeting of genes in mouse embryonic stem cells: disruption of the T-cell receptor $\delta$-chain-encoding gene. *Gene*, 105(2):263–267.

van Gent, D. C., Mizuuchi, K., and Gellert, M. (1996). Similarities between initiation of V(D)J recombination and retroviral integration. *Science (New York, N.Y.)*, 271(5255):1592–4.

van Overbeek, M., Capurso, D., Carter, M. M., Thompson, M. S., Frias, E., Russ, C., Reece-Hoyes, J. S., Nye, C., Gradia, S., Vidal, B., Zheng, J., Hoffman, G. R., Fuller, C. K., and May, A. P. (2016). DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell*, 63(4):633–646.

Voet, T., Vanneste, E., and Vermeesch, J. R. (2011). The human cleavage stage embryo is a cradle of chromosomal rearrangements. *Cytogenetic and genome research*, 133(2-4):160–8.

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., Sabatini, D. M., Krupczak, M., Post, Y., Wei, J. J., Eric, S., and Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science*, 00013(6264):1–10.

Weitzman, M. D., Kyöstiö, S. R., Kotin, R. M., and Owens, R. A. (1994). Adeno-associated virus (AAV) Rep proteins mediate complex formation between AAV DNA and its integration site in human DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 91(13):5808–12.

Wilde, J. J., Aida, T., Wienisch, M., Zhang, Q., Qi, P., and Feng, G. (2018). RAD51 Enhances Zygotic Interhomolog Repair. *BioRxiv*.

Williams, R. L., Hilton, D. J., Pease, S., Willson, T. A., Stewart, C. L., Gearing, D. P., Wagner, E. F., Metcalf, D., Nicola, N. A., and Gough, N. M. (1988). Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. *Nature*, 336(6200):684–687.

Wright, A. V., Nuñez, J. K., and Doudna, J. A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell*, 164(1-2):29–44.

Xu, L., Park, K. H., Zhao, L., Xu, J., El Refaey, M., Gao, Y., Zhu, H., Ma, J., and Han, R. (2016). CRISPR-mediated Genome Editing Restores Dystrophin Expression and Function in mdx Mice. *Molecular Therapy*, 24(3):564–569.

Yagi, T., Nada, S., Watanabe, N., Tamemoto, H., Kohmura, N., Ikawa, Y., and Aizawa, S. (1993). A Novel Negative Selection for Homologous Recombinants Using Diphtheria Toxin A Fragment Gene. *Analytical Biochemistry*, 214(1):77–86.

Yang, D., Scavuzzo, M. A., Chmielowiec, J., Sharp, R., Bajic, A., and Borowiak, M. (2016a). Enrichment of G2/M cell cycle phase in human pluripotent stem cells enhances HDR-mediated gene repair with customizable endonucleases. *Scientific Reports*, 6(January):1–15.

Yang, S., Fujikado, N., Kolodin, D., Benoist, C., and Mathis, D. (2015). Regulatory T cells generated early in life play a distinct role in maintaining self-tolerance. *Science*, 348(6234):589–594.

Yang, Y., Sterling, J., Storici, F., Resnick, M. A., and Gordenin, D. A. (2008). Hypermutability of damaged single-strand DNA formed at double-strand breaks and uncapped telomeres in yeast Saccharomyces cerevisiae. *PLoS Genetics*, 4(11).

Yang, Y., Wang, L., Bell, P., McMenamin, D., He, Z., White, J., Yu, H., Xu, C., Morizono, H., Musunuru, K., Batshaw, M. L., and Wilson, J. M. (2016b). A dual AAV system enables the Cas9-mediated correction of a metabolic liver disease in newborn mice. *Nature biotechnology*, 34(3):334–338.

Yeung, A. T., Hattangadi, D., Blakesley, L., and Nicolas, E. (2005). Enzymatic mutation detection technologies.

Yusa, K., Rashid, S. T., Strick-Marchand, H., Varela, I., Liu, P.-Q., Paschon, D. E., Miranda, E., Ordóñez, A., Hannan, N. R. F., Rouhani, F. J., Darche, S., Alexander, G., Marciniak, S. J., Fusaki, N., Hasegawa, M., Holmes, M. C., Di Santo, J. P., Lomas, D. a., Bradley, A., and Vallier, L. (2011a). Targeted gene correction of $\alpha$1-antitrypsin deficiency in induced pluripotent stem cells. *Nature*, 478(7369):391–4.

Yusa, K., Zhou, L., Li, M. A., Bradley, A., and Craig, N. L. (2011b). A hyperactive piggyBac transposase for mammalian applications. *Proceedings of the National Academy of Sciences of the United States of America*, 108(4):1531–6.

Zetsche, B., Gootenberg, J. S. S., Abudayyeh, O. O. O., Slaymaker, I. M. M., Makarova, K. S. S., Essletzbichler, P., Volz, S. E. E., Joung, J., Van Der Oost, J., Regev, A., Koonin, E. V. V., Zhang, F., van der Oost, J., Regev, A., Koonin, E. V. V., Zhang, F., Van Der Oost, J., Regev, A., Koonin, E. V. V., and Zhang, F. (2015). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell*, 163(3):759–771.

Zhang, Y., Rajan, R., Seifert, H. S., Mondragón, A., and Sontheimer, E. J. (2015). DNase H Activity of Neisseria meningitidis Cas9. *Molecular Cell*, 60(2):242–255.

Zheng, Z., Liebers, M., Zhelyazkova, B., Cao, Y., Panditi, D., Lynch, K. D., Chen, J., Robinson, H. E., Shim, H. S., Chmielecki, J., Pao, W., Engelman, J. A., Iafrate, A. J., and Le, L. P. (2014). Anchored multiplex PCR for targeted next-generation sequencing. *Nature Medicine*, 20(12):1479–1484.

Zhu, S., Li, W., Liu, J., Chen, C.-h., Liao, Q., Xu, P., Xu, H., Xiao, T., Cao, Z., Peng, J., Yuan, P., Brown, M., Liu, X. S., and Wei, W. (2016). Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR–Cas9 library. *Nature Biotechnology*, 34(12):1279–1286.

Zhulidov, P. A., Bogdanova, E. A., Shcheglov, A. S., Vagner, L. L., Khaspekov, G. L., Kozhemyako, V. B., Matz, M. V., Meleshkevitch, E., Moroz, L. L., Lukyanov, S. A., and Shagin, D. A. (2004). Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Research*, 32(3):37e–37.

Zuckermann, M., Hovestadt, V., Knobbe-Thomsen, C. B., Zapatka, M., Northcott, P. A., Schramm, K., Belic, J., Jones, D. T., Tschida, B., Moriarity, B., Largaespada, D., Roussel, M. F., Korshunov, A., Reifenberger, G., Pfister, S. M., Lichter, P., Kawauchi, D., and Gronych, J. (2015). Somatic CRISPR/Cas9-mediated tumour suppressor disruption enables versatile brain tumour modelling. *Nature Communications*, 6(1):7391.

Zuris, J. A., Thompson, D. B., Shu, Y., Guilinger, J. P., Bessen, J. L., Hu, J. H., Maeder, M. L., Joung, J. K., Chen, Z.-Y., and Liu, D. R. (2015). Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nature biotechnology*, 33(1):73–80.