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 recombination 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

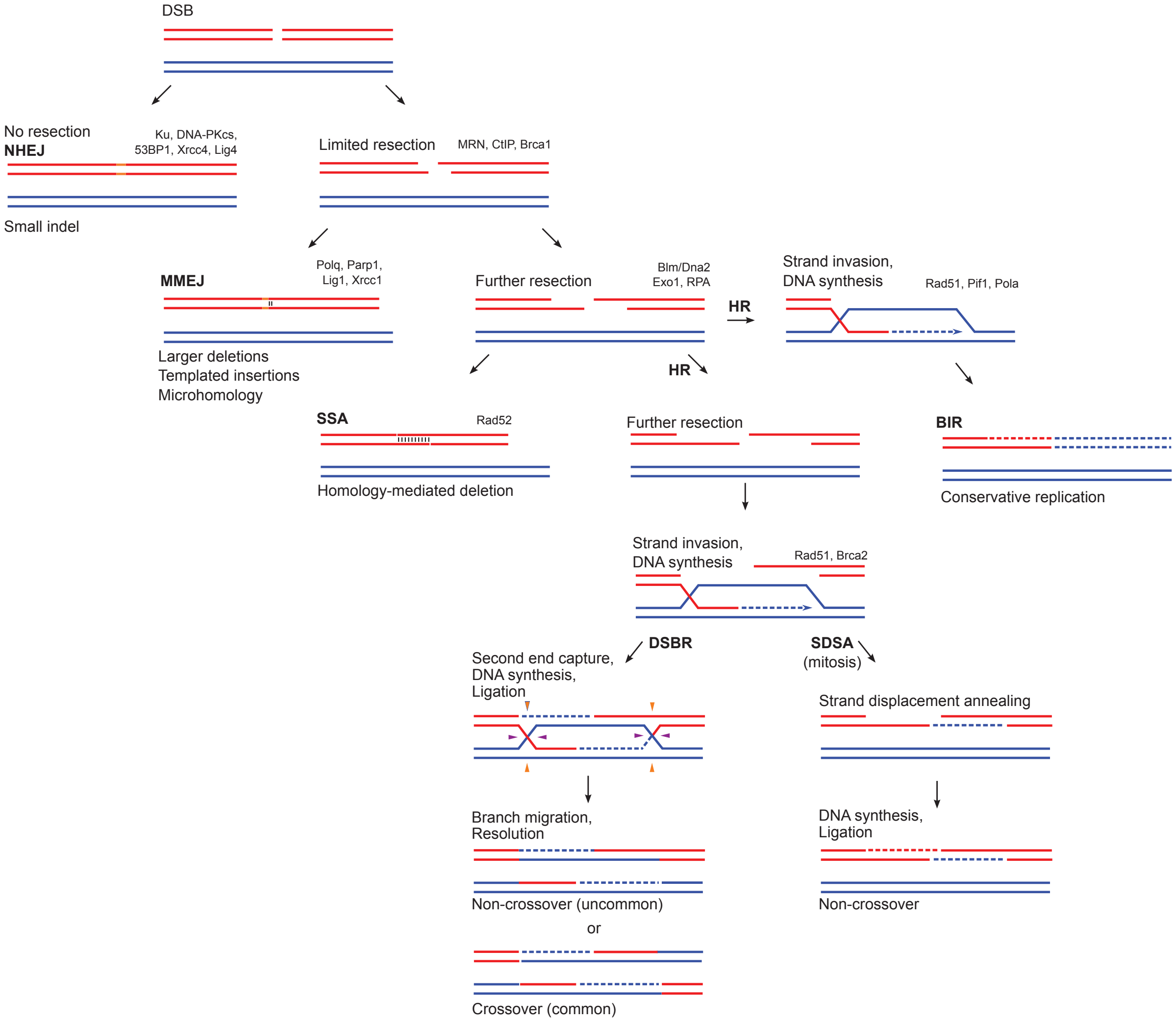

Figure 1.1: Pathways of DSB repair. Modified from Sung and Klein, 2006.

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).

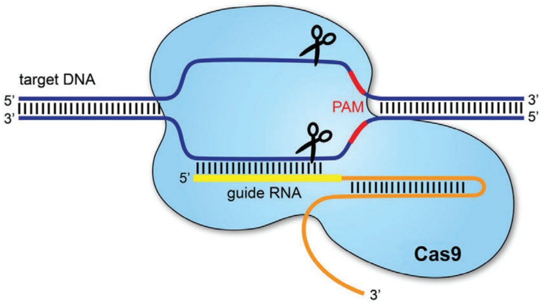

Figure 1.2: Schematic of Cas9 DNA cleavage mechanism. From Redman et al., 2016.

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 targeting 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).