Genome editing

Introduction

Until now, we have discussed ways to alter a genome of an organism that rely on either random integration or on targeted integration by homologous recombination. With the development of targeted mutagenesis (as discussed in the mouse knockout approach) it was possible to specifically alter a gene of interest. However, all gene-targeting methods rely on homologous recombination, and recombination events are rather rare: while still relatively frequent in yeast (one in 10^4 cells undergoes homologous recombination), only one homologous recombination event happens in 10^7 mouse cells. The effective number of cells that underwent homologous recombination leading to a successful targeted gene replacement is therefore very low. For yeast and mouse embryonic stem cells (ES cells), selection in cell cultures allows the recovery of the rare cells with the desired event, but for many model organisms, ES cells are not available, screening or selection procedures are not adequate, and therefore, the development of useful gene-targeting approaches is hindered by the low frequency of recombination. Thus, one of the big challenges was to increase the frequency of recombination. Here, we will present three approaches that researchers developed to tackle this problem.

The goal: Increasing the frequency of homologous recombination

The main goal in improving the success of targeted gene replacement is to increase the absolute frequency of homologous recombination. Here, the key process is the induction of DNA double-strand breaks (DSB) at the target site. DSB are detected in cells as potentially lethal damage, and one natural pathway of DSB repairs is copying from a homologous template. Normally, in a diploid cell, the homologous template would be the homologous chromosome. Thus, the approach of gene targeting by introducing DSB simply provides an exogenous template for a naturally occurring repair process (see figure 5-1). Alternatively, DSB can be repaired by joining the broken ends directly without a need for a homologous template, a process called non-homologous end joining. Sometimes, this ligation is carried out imprecisely and can create deletions, insertions, or substitutions at the break site and is therefore a tool for targeted mutagenesis (see figure 5-1, right).

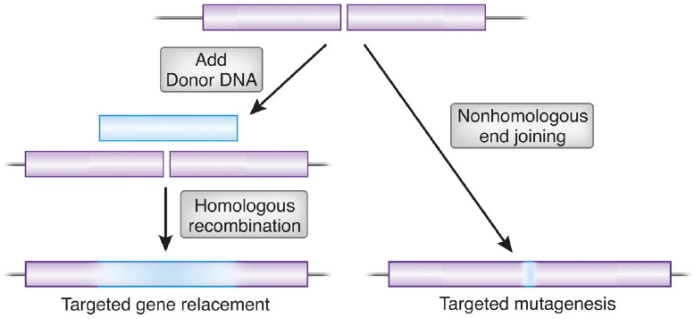

Figure 5-1 Repair outcomes of a genomic DSB. *Left*: If a homologous donor DNA is provided, homology-directed repair can proceed by homologous recombination using the donor DNA as template. The amount of donor sequence ultimately incorporated is highest at the site of the DSB. *Right*: Alternatively, the break can be repaired by non-homologous end joining, leading to mutations, e.g., deletions, insertions, and base substitutions at the cleavage site. (Adapted from D. Carroll, *Gen. Rev.*, 2011)

Thus, if DSB are stimulated in the presence of an exogenous template, the natural repair process

is used for gene targeting. While non-homologous end joining stimulates targeted mutagenesis, the repair of DSB by homologous recombination can be utilized for targeted gene replacement by supplying a donor DNA sequence.

However, double-strand breaks occur rather rarely (10-40 breaks per cell cycle in a human cell) and random; thus, the probability for a certain donor DNA to be inserted into the genome of an organism at the correct place by homologous recombination is very low. To achieve efficient gene targeting, it is desired to increase the number of DSB, since this increases the likelihood of homologous recombination. Therefore, geneticists thought of strategies to increase the number of DSB, which we will discuss here.

Method 1: Zinc-finger nucleases (ZFN)

How can one induce DSB into a DNA strand? DNA is cut by DNA endonucleases; thus, by targeting such an enzyme to a specific genomic sequence, DSB can be induced at the desired location, followed by the generation of desired modifications during subsequent DNA repair. For this purpose, a nuclease was created by fusing the DNA-cleavage domain of a bacterial endonuclease with a DNA-binding domain, a zinc-finger domain. Zinc fingers are small structural motifs (domains) present in proteins. Zinc-finger domains contain multiple finger-like protrusions that make tandem contacts with their target molecule. Their binding properties depend on the amino-acid sequence of the finger domains and on the linker between fingers, as well as on the number of fingers. Zinc-finger proteins typically function as interaction modules that bind DNA, RNA, or proteins.

The zinc-finger nucleases (ZFN) created for gene targeting are based on the structure of the natural restriction endonuclease *FokI* (from *Flavobacterium okeanokoites*). The DNA-binding domain can be engineered to target a specific DNA sequence such that the zinc-finger nucleases can be directed to the desired sequence where it induces DSB. The most useful design for a synthetic DNA-binding domain proved to be a set of three Cys_2His_2 zinc fingers, with each finger primarily binding to only 3 base pairs of DNA. It is important to note that two *FokI* DNA-cleavage domains must dimerize in order to cut the DNA strand, thus, two ZFN molecules need to bind simultaneously with their DNA-binding domains to their recognition sequence on the genome (see figure 5-2).

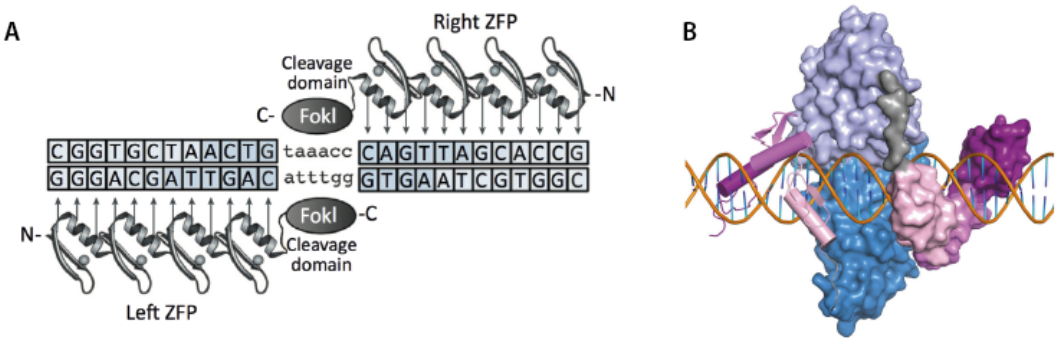

Figure 5-2 Illustration of a pair of ZFNs bound to DNA. (A) Two ZFNs have to dimerize across the DNA for the *Fok*I cleavage domains to be able to cut the DNA at their cleavage sites. Each *Fok*I cleavage domain is connected by a linker to the DNA-binding domain which consists of multiple zinc fingers. The arrows indicate the binding of the zinc ingers to the nucleotides. (Adapted from T. Gaj *et al.*, *Trends in Biotechn.*, 2013) **(B)** Model of a pair of ZFNs bound to DNA, with each of three zinc fingers shown in a shade of pink (ribbon representation on the left and space-filling representation on the right), the linkers shown in grey, and the *Fok*I cleavage domains shown in shades of blue. (Adapted from D. Carroll, *Gen. Rev.*, 2011)

How are the engineered zinc-finger nucleases introduced into animals to generate gene deletions? ZFN-based genetic engineering does not require the use of ES cells, because ZFNs can be injected

directly into early-stage embryos. Therefore, fertilized single-cell embryos are injected with plasmids containing the ZFN (see figure 5-3). In the nucleus, the ZFN locates the target sequence and creates a DSB. The DSB stimulates the cellular process of non-homologous end joining and results in the mis-repair of the DNA sequence. The resulting mutation usually gives rise to a knockout. The embryos are then further cultivated (e.g., for *Drosophila*) or implanted into a foster mother (e.g., mice) and allowed to divide and grow into whole organisms. At birth, the animals are screened for mutations and the animals containing the targeted knockout are identified.

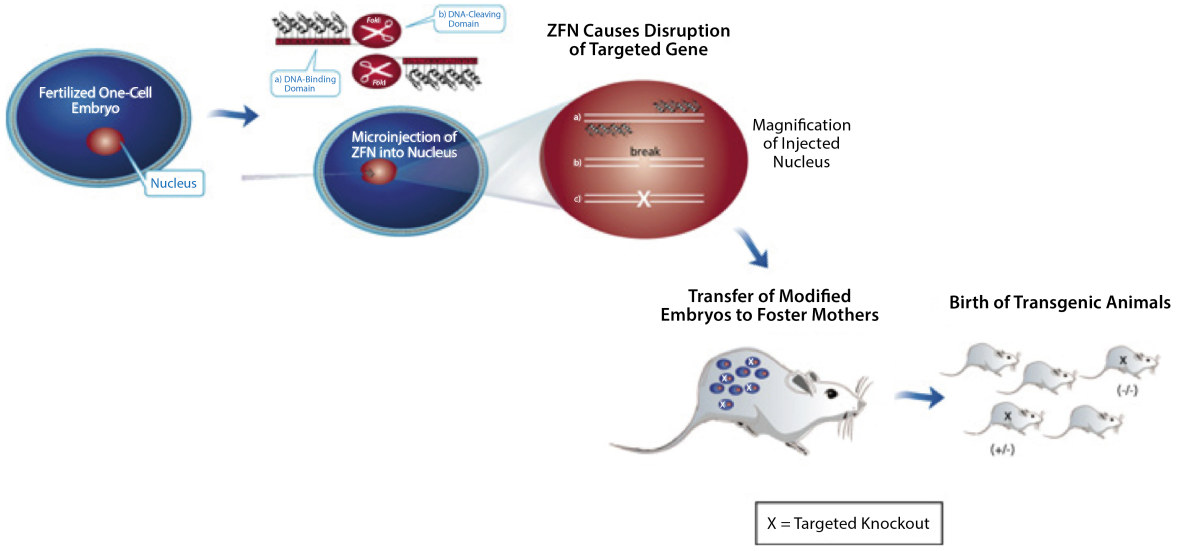

Figure 5-3 Knockout mouse generation by microinjection of plasmids containing an engineered ZFN, which recognizes a specific DNA sequence to induce DSB. This method can be used for many organisms, including *Drosophila* or *C. elegans* by adapting the injection procedure accordingly (for injection methods for these animals, see lesson 2). (Adapted from sigmaaldrich.org)

ZFNs have successfully been used for targeted gene replacement in a variety of organisms, including *Drosophila*, *C. elegans*, and mice, various plants, as well as several cell types such as human, mouse, and pig. ZFNs also provide an alternative strategy to embryonic-stem-cell-based homologous recombination, and can, thanks to their low immunogenicity, be used for gene therapy. Since ZFN-based genetic engineering does not require the use of ES cells, this allows targeted gene disruption in a wider spectrum of organisms and in a shorter time. Knockout rats and mice can be created in as little as 2-3 months compared to the ES-cell method in mice that can take up to 12-18 months. Further, the ZFN method results in efficient germline transmission of targeted genetic mutations without incorporation of foreign DNA sequences (unlike in the ES-based knockout approach). Unfortunately, the design and production of ZFNs are expensive (several 1000 US$) and time-consuming ($>$ 1 month), since context-dependent effects, e.g., neighboring fingers can alter the specificity, and off-target effects have to be taken into account. The three zinc-fingers themselves pose rather inflexible molecules, and can therefore only be targeted to sequences of about 500 base pairs. Additionally, a substantial portion of ZFNs fails *in vivo* due to the limited capacity of target cells towards the ZFNs. Furthermore, not every nucleotide triplet has a corresponding zinc finger, and interactions between zinc fingers within an array can reduce their specificity.

Method 2: Transcription activator-like effector nucleases (TALEN)

The ZFN technique had already been used for many years when researchers developed a method that is also based on engineering the *FokI* nuclease to direct it to specific sequences by fusing it to different

DNA-binding domains. Instead of zinc fingers, they used transcription activator-like (TAL) effectors, proteins that are secreted by the bacteria of the genus *Xanthomonas* when they infect various plant species. These proteins can bind promoter sequences in the host plant and activate the expression of plant genes that aid bacterial infection. They recognize DNA sequences through a central repeat domain consisting of a variable number of 34 amino-acid repeats. The sequence of each repeat is similar but not identical: it differs at amino-acid positions 12 and 13. This amino-acid pair is called repeat-variable di-residue (RVD), and each RVD recognizes a specific base. Upon folding of the TALE protein, the different RVDs are exposed at the protein surface within a sequence. This RVD sequence allows the TAL effector to specifically bind the sequence of corresponding bases, and thus binding to the target DNA sequence (see figure 5-4A). Therefore, it is this RVD that determines which single nucleotide the TAL effector will recognize: HD targets cytosine, NI targets adenine, NG targets thymine, and NN targets guanine (though NN can also bind adenine with lower specificity).

Knowing how certain RVDs "code" for certain bases, TALEs can specifically be designed to bind a particular DNA sequence of interest. Like ZFNs, TALEs are fused to a DNA-cleaving domain, such as *FokI*, and thus provide a tool for genome editing through targeted gene replacement (see figure 5-4B). The advantage of TAL effectors is that they have a potential cleavage sites approximately every 35 base pairs in the genome, while potential target sites for zinc fingers are roughly every 500 base pairs, because zinc-finger targets are confined to sequences composed of triplets with corresponding zinc fingers. They are context-independent and appear to show little off-target effects and low immunogenicity.

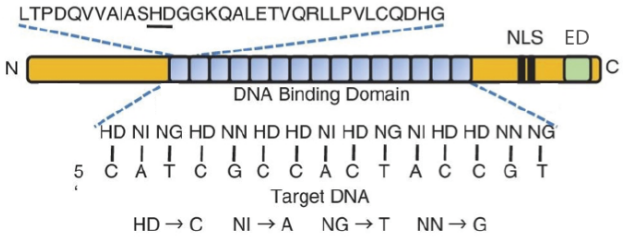

Figure 5-4 Structure of TALENs. The DNA-binding domain of natural TALEs (blue) consists of multiple repeats, each containing an RVD at positions 12 and 13 (underlined in the top sequence) that binds one base on the target DNA. The TALEs enter the nucleus by an incorporated nuclear localization sequence (NLS), and contain an effector domain (ED) to, e.g., the nuclease *Fok*I. (Adapted from T. Cermak *et al.*, *Nucl. Ac. Res.*, 2011)

Furthermore, the designed TALEs can be fused to another protein in order to achieve a targeted effect. For example, by fusing a designed TALE to transcriptional activation domains (AD) or to repression domains (RD), the designed fusion protein serves as an artificial switch for gene regulation *in vivo* (figure 5-5).

Method 3: The CRISPR/Cas9 system

The two methods presented above, the zinc-finger nucleases (ZFNs) and TAL-effector nucleases (TALENs) use the principles of DNA-protein recognition to induce site-specific double-strand breaks to increase homologous recombination. A very recent development has revolutionized genome engineering with the discovery of a system that uses an RNA-programmable system to introduce double-strand breaks: the CRISPR-Cas9 technology.

This technology originates from the type-II CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)/Cas system, which provides bacteria with adaptive immunity to viruses and plasmids.

The CRISPR/Cas9 immune defense has two important features. First, the host cell can specifically incorporate short sequences from invading genetic elements (virus or plasmid) into a specific region of its genome, the so-called CRISPR-array locus. This locus contains repeat sequences, which are called clustered regularly interspaced short palindromic repeats (CRISPR), and pieces of the foreign DNA integrated between these repeats (called protospacer sequences). Second, when these sequences are transcribed and precisely processed into small RNAs, they guide a multifunctional protein complex (Cas proteins) to recognize and cleave incoming foreign genetic material.

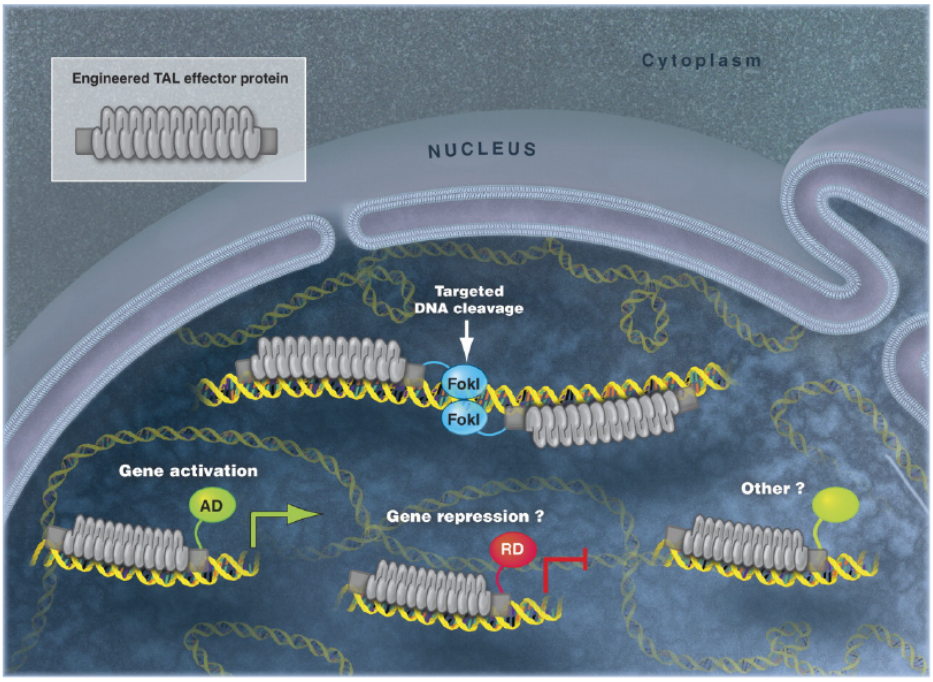

Figure 5-5 Various applications of TALEs. Fusion of TAL effector proteins to *Fok*I creates sequence-specific nucleases that enable targeted DNA cleavage for gene knockouts and genome editing. TAL effector proteins fused to transcriptional activation domains (AD) and putatively to repression domains (RD) provide artificial switches for gene regulation in vivo. (Adapted from A.J. Bogdanove *et al., Science,* 2011)

In bacteria and archaea, three different CRISPR/Cas9 systems exist that act similarly: here, we will discuss the type-II system of the bacterium Streptococcus pyogenes, which served as the basis to develop the CRISPR/Cas9 genome-engineering system. In brief, transcription of the CRISPR array leads to the formation of a long RNA that is subsequently cleaved into short crisper RNAs (crRNAs). These crRNAs contain the sequence information about the invader (since they are transcripts of the foreign DNA). These RNAs bind to another RNA called the transactivating RNA (tracrRNA), and this RNA hybrid is able to build a complex with a cellular endonuclease, the Cas9 protein. If this RNA-Cas9 complex encounters invasive DNA present in the cell, the crRNA directs the Cas9 nuclease to the invasive DNA by complementary base pairing. The Cas9 nuclease then cuts and destroys the incoming foreign DNA (see figure 5-6 and 5-7A).

The mechanism of generating this immunity is, even in the simple type-II CRISPR/Cas9 system of *S. pyogenes*, quite complex: Transcription of the CRISPR array leads to a long RNA that is subsequently processed into crRNAs, each harboring a variable sequence transcribed from the invading DNA (called the protospacer sequences, marked in different colors in figure 5-7) and part of the CRISPR repeat

(black in figure 5-7). Each crRNA hybridizes with a second RNA, the tracrRNA, through a short region that is complementary to the CRISPR repeat present on the crRNA. These two RNAs together complex with the Cas9 nuclease. The tracrRNA and Cas9 form a complex with each unique crRNA. Accordingly, each crRNA:tracrRNA:Cas9 complex seeks out the DNA sequence complementary to the crRNA. The protospacer sequence present on the crRNA directs the complex to the complementary sequence on the DNA, where the Cas9 nuclease cleaves the target DNA. After the crRNA-tracrRNA:Cas9 complex binds, the Cas9 separates the double-stranded DNA target and cleaves both strands (see figure 5-7A). The crRNA:tracrRNA:Cas9 complex unbinds after inducing the DSB.

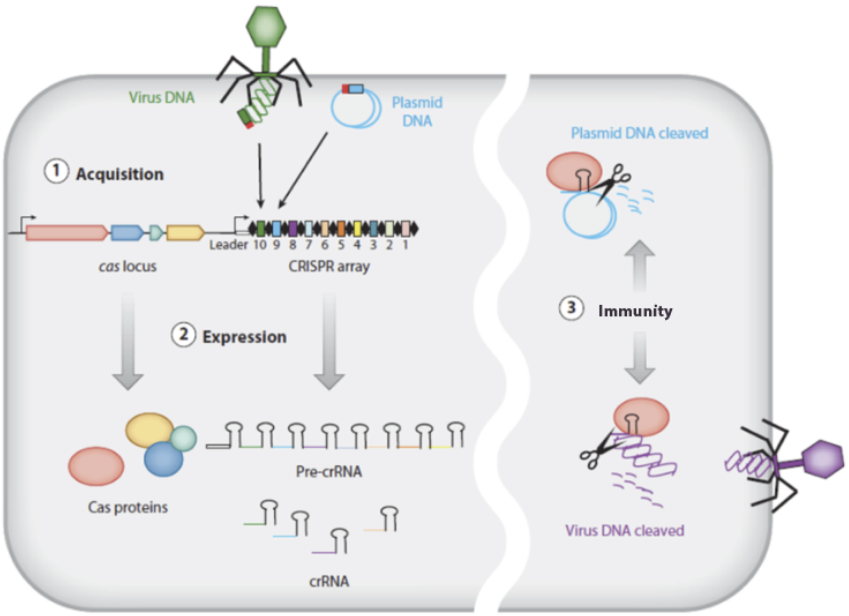

Figure 5-6 The CRISPR/Cas9 system as a microbial defense mechanism. **(1)** Sequences of pathogenic foreign DNA are integrated into the CRISPR locus of the host genome in the form of protospacer sequences between organism-specific (often palindromic) repeat sequences. A diverse group of *cas* genes are located in the vicinity of a CRISPR locus, which encode proteins (generically called Cas proteins) required for the defense against invasive genetic elements. A CRISPR array consists of unique spacers (colored boxes; spacers are numbered sequentially with the most recently acquired spacer having the highest number) interspaced between repeats (black diamonds). **(2)** The pre-crRNA, transcribed from the entire CRISPR locus, will be cleaved into individual crRNAs, each containing a partial repeat and a single protospacer sequence. **(3)** By forming an RNA-protein complex with Cas9, which provides nuclease activity, this system allows targeting and cleavage of the invasive DNA in the future. (Adapted from A. Bhaya *et al.*, *Annu. Rev. Genet.*, 2011)

In the CRISPR-Cas9 genome-engineering technology, the dual tracrRNA:crRNA was engineered as a single-guide RNA (sgRNA) that retains two critical features: a sequence at the 5'-side that determines the DNA target site (the protospacer, green in figure 5-7B) and a duplex-RNA structure at the 3'-side that binds to Cas9. By specifically designing the protospacer sequence to target a genomic sequence of interest, only the synthetic sgRNA and Cas9 have to be provided to a cell in order for the system to target and efficiently induce genomic DSB (see figure 5-7B).

This development created a simple two-component system in which changes in the guide sequence of the sgRNA program Cas9 to target any DNA sequence of interest. After the Cas9-induced double-strand break occurred, the DNA can either be repaired by non-homologous end joining (NHEJ), which

can induce insertions or deletions, or by homology-driven repair (HDR). The latter can be used to insert desired sequences through recombination of the target locus with supplied DNA donor templates (see figure 5-8B). If pairs of guideRNA (gRNA) are used, the Cas9 nuclease can induce large deletions or chromosomal rearrangements, such as inversions (see figure 5-8C). Furthermore, by fusing the Cas9 nuclease to protein domains that can activate the expression of proteins, the expression of proteins can be specifically induced (see figure 5-8D). The fusion of Cas9 with other effector domains may be used to alter DNA modifications (see figure 5-8E). Finally, if Cas9 is fused to fluorescent proteins, specific genomic loci can be marked and analyzed by imaging (see figure 5-8F).

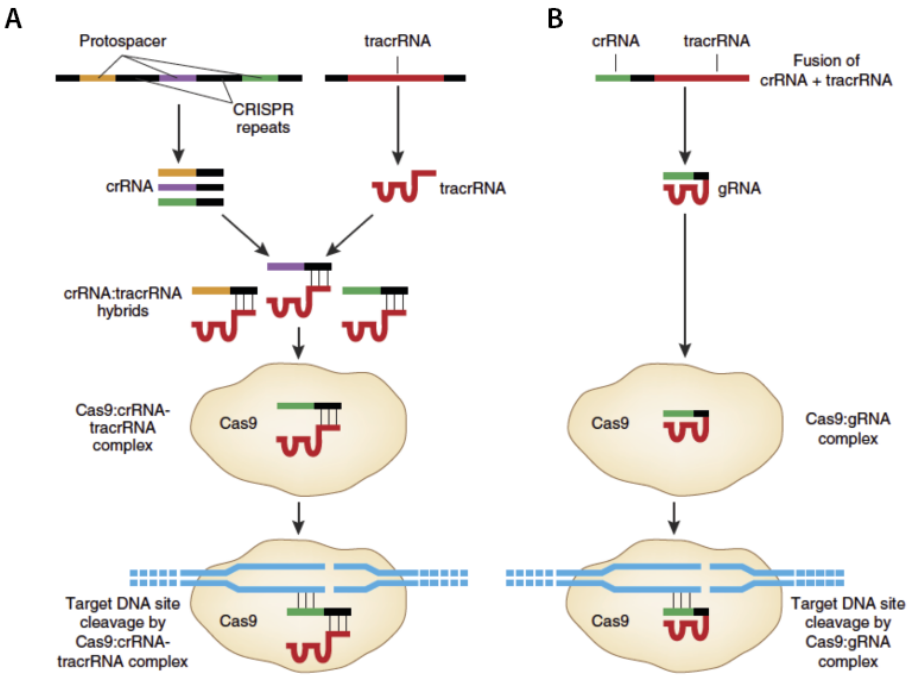

Figure 5-7 Naturally occurring and engineered CRISPR-Cas systems. (A) Naturally occurring CRISPR systems incorporate foreign DNA sequences into CRISPR arrays, which then produce crRNAs bearing protospacer sequences that are complementary to the foreign DNA site. crRNAs hybridize with tracrRNAs (also encoded by the CRISPR system) and this pair of RNAs can associate with the Cas9 nuclease. crRNA-tracrRNA:Cas9 complexes recognize and cleave foreign DNAs bearing the protospacer sequences. **(B)** The most widely used engineered CRISPR-Cas system utilizes sgRNA, a fusion between a crRNA and part of the tracrRNA sequence. This sgRNA complexes with Cas9 to mediate cleavage of target DNA sites that are complementary to the sgRNA. (Adapted from J.D. Sander and J.K. Joung, *Nature Biotechn. Rev.*, 2014)

Comparing the three methods for targeting DSB to specific genomic loci

All three methods presented here aim at inducing double-strand breaks by employing molecular scissors using endonucleases from bacterial systems (see figure 5-9). In the ZFN and TALEN approach, the endonuclease is directed to a cleave at a specific site with the help of protein domains that bind DNA (zinc-finger domains or TALE). In the CRISPR/Cas9 system, an RNA is used to guide the endonuclease to the desired DNA sequence.

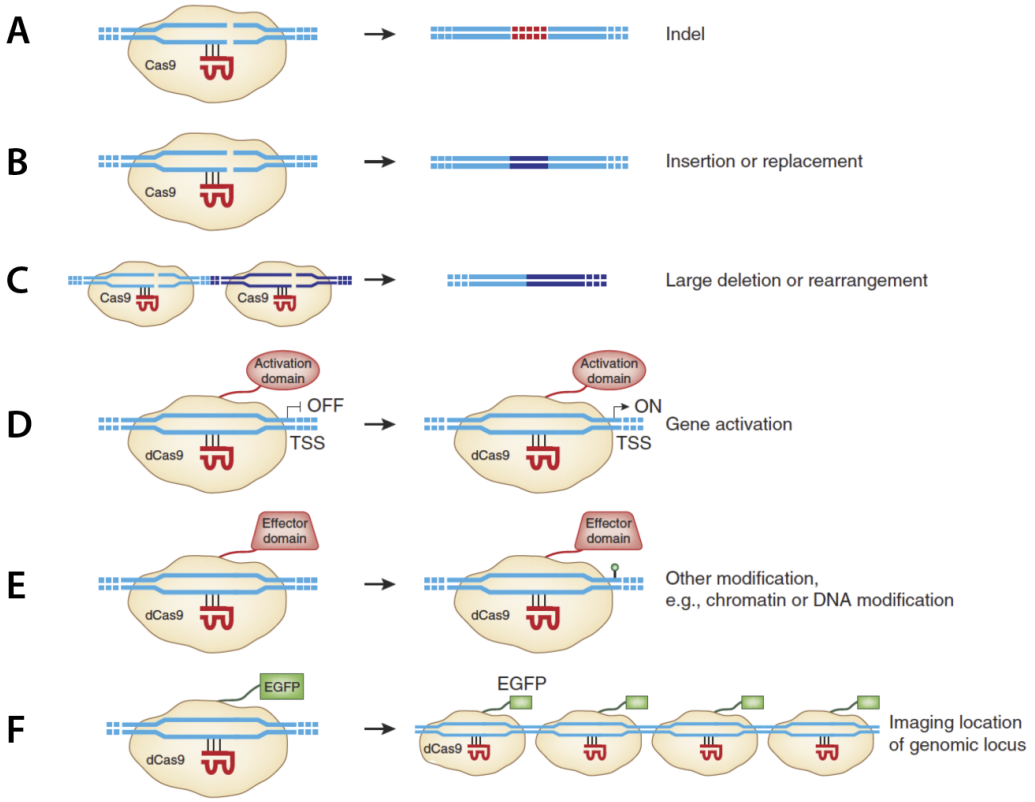

Figure 5-8 Overview of various Cas9-based applications. (A,B) gRNA-directed Cas9 nuclease can induce insertions or deletions (indel) mutations **(A)** or specific sequence replacement or insertion **(B)**. **(C)** Pairs of gRNA-directed Cas9 nucleases can stimulate large deletions or genomic rearrangements. **(D-F)** gRNA-directed dCas9 can be fused to activation domains **(D)** to mediate upregulation of specific endogenous genes, heterologous effector domains **(E)** to alter histone modifications or DNA methylation, or **(F)** fluorescent proteins to enable imaging of specific genomic loci. (Adapted from J.D. Sander and J.K. Joung, *Nature Biotechn. Rev.*, 2014)

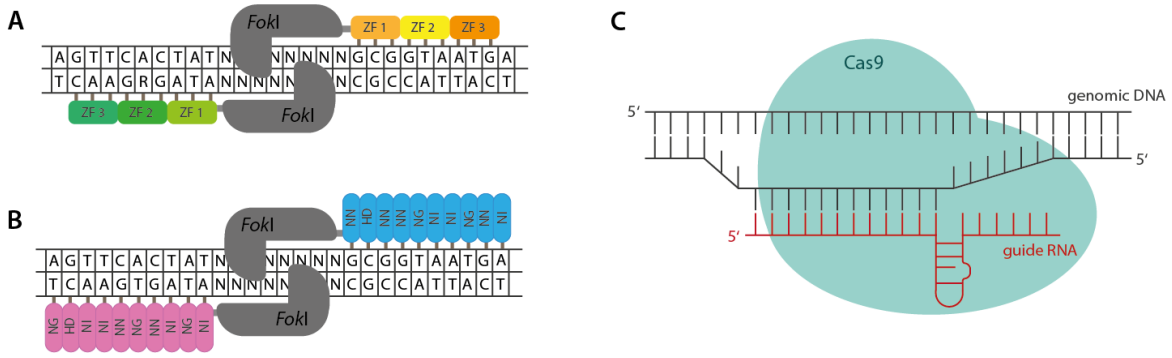

Figure 5-9 Comparison of the three methods to induce DNA double-strand breaks using molecular scissors. (A) Zinc-finger nucleases (ZFNs) consist of three tandem zinc-finger domains (ZF 1-3, yellow and green boxes) fused to a *FokI* nuclease (gray). *FokI* acts as a dimer, and specifically cleaves genomic DNA when two ZFNs are targeted in tandem. **(B)** TALENs are similar to ZFNs, except that each domain of the TALE protein recognizes a single nucleotide, e.g., NG to T, HD to C, NI to A, and NN to G. Equally, the nuclease is *FokI* (gray), functioning as a dimer. **(C)** The CRISPR system uses a single-guide RNA (red) to target a region of interest for cleavage. The Cas9 nuclease (green) recognizes aspects of the secondary structure of the single-guide RNA and the RNA:DNA hybrid.