The genome editing revolution: A CRISPR-Cas TALE off-target story

Stefano Stella and Guillermo Montoya

In the last 10 years, we have witnessed a blooming of targeted genome editing systems and applications. The area was revolutionized by the discovery and characterization of the transcription activator-like effector proteins, which are easier to engineer to target new DNA sequences than the previously available DNA binding templates, zinc fingers and meganucleases. Recently, the area experimented a quantum leap because of the introduction of the clustered regularly interspaced short palindromic repeats (CRISPR)-associated protein (Cas) system (clustered regularly interspaced short palindromic sequence). This ribonucleoprotein complex protects bacteria from invading DNAs, and it was adapted to be used in genome editing. The CRISPR ribonucleic acid (RNA) molecule guides to the specific DNA site the Cas9 nuclease to cleave the DNA target. Two years and more than 1000 publications later, the CRISPR-Cas system has become the main tool for genome editing in many laboratories. Currently the targeted genome editing technology has been used in many fields and may be a possible approach for human gene therapy. Furthermore, it can also be used to modifying the genomes of model organisms for studying human pathways or to improve key organisms for biotechnological applications, such as plants, livestock genome as well as yeasts and bacterial strains.

Keywords:
CRISPR-Cas9; engineered nucleases; genome editing; specificity; TALE.

Engineered nucleases and DNA repair

The use of protein templates (engineered nucleases, hereafter ENs), which specifically recognize a DNA sequence, to efficiently edit the genome by introducing double-strand breaks (DSBs) have been available for the last 10–15 years [1]. This process takes advantage of the natural mechanisms that cells have developed to repair this deleterious damage. One of them is homologous recombination (HR), which uses an intact copy of the broken sequence as a template to repair the damage [2]. User-defined sequences next to two DNA regions homologous to the target site can be used as DNA template for HR to integrate the desired DNA segment at the specific site in the host genome. Because the target sequence is introduced at the specific sites in an error-free prone fashion, this procedure can be employed to replace a defective gene in gene therapy applications [2] (Fig. 1).

The other mechanism used to repair DSBs is nonhomologous end joining (NHEJ), which can result in the loss of some bases at the repair junction [3], thus generating genomic instability [4]. This non-conservative mechanism can solve the DSBs joining the two DNA ends. However, in the case of single stranded overhangs in DSBs, short homologous DNA sequences called microhomologies are used to guide repair [5, 6]. When the overhangs are perfectly compatible, NHEJ usually repairs the break accurately; however, imprecise repair leading to loss of nucleotides can also occur when the overhangs are not compatible.

Recent studies have indicated that NHEJ can repair DSBs in an error-free manner with a frequency of 75% and that probably this frequency is underestimated [7]. The generation of specific DNA damage to induce the repair by HR or NHEJ allows many different modifications at the targeted site (Fig. 1).

A short history of the DNA-binding template repertoire

About 10 years ago, the scientific community received the first reports showing how the use of customized zinc fingers nucleases [8, 9] and meganucleases [10–12] could enable researchers to pursue genomic modifications in different organisms, such as Drosophila or humans [8]. In the case of meganucleases, the template contains the recognition and cleavage functions in the same scaffold, while in the case of zinc fingers, the customizable
sequence-specific DNA-binding domain needs to be fused with a nuclease protein (FokI) that cleaves DNA. In zinc fingers, the modification of some residues near the recognition α-helix can change the DNA-binding specificity [13, 14]. They can be joined together into arrays that are capable of recognizing longer (a multiple of three base pairs) DNA sequences. In the case of meganucleases, extensive redesign of the I-CreI enzyme has been carried out without altering its nuclease activity [15]. This activity has focused the attention of few companies, such as Sangamo, Cellectis or Precision, which have developed platforms to provide a la carte ENs. However, the range of sequences that can be engineered with these templates and the challenge of setting these technologies in non-specialist laboratories have precluded them to be vastly implemented.

The situation changed when first transcription activator-like effector proteins (TALE) and then clustered regularly interspaced short palindromic repeats (CRISPR)-associate protein 9 (Cas9) were discovered. These new protein templates are flexible, faster and easier to customize and have basically little limitations regarding the DNA sequence that they could target. Thus, these templates have become the main editing tools in many laboratories.

The vast explosion of papers using these systems to modify a genome has recently reached the point where the germline of a human embryo may be edited. This has raised not only ethical questions but also concerns about our knowledge of these systems [16–19]. How much we know about their working mechanism?, and in particular; how safe they are in sensitive applications?, such as human gene therapy [20]. It seems clear that all ENs do not have a perfect DNA recognition, and this recognition event can depend in many different issues, such as the location of the targeted sequence in the genome, chromatin state, the type of cell, the expression of the protein in the target and several other factors.

Therefore, it is necessary to increase our basic mechanistic knowledge on them to guarantee that ENs introduce changes on the target sites without potentially dangerous off-target effects [1]. The use of unspecific ENs is rather toxic for the cell, and it can lead to chromosomal rearrangement, inversion and deletions but also to mutations at unwanted sites which could induce cytotoxic effects on the host cell. However, in sensitive applications, even very low unspecific ENs are dangerous because they can lead to subtle changes that may not be noticed at the genomic level, but they can induce undesirable phenotypes. Preventing, avoiding or at least minimizing these effects is crucial for the success of any genome editing application. Thus, the determination of the off-target activity of ENs is a
key for a successful use of these powerful tools. The wide use of TALEs and CRISPR-Cas9 system makes important to address in a systematic matter the off-target effect of these two scaffolds.

The transcription activator-like effector proteins template

Transcription activator-like effector proteins are proteins that Xanthomonas bacteria inject in plant cells to activate the expression of plant genes facilitating bacterial infection [21]. The DNA-binding activity of TALE is mediated by arrays of conserved 33–35 amino acid repeats, which recognize a specific base pair in the DNA by the amino acids located at position 13 of each repeat. The protein scaffold shows a web of non-specific interactions between each unit and the coding strand of the DNA target [22–24]. By modifying the amino acid at position 13 according to a known and simple base-pair amino acid code, the specificity of every repetitive unit can be exchanged (Fig. 2A) [25, 26]. Thus the repeats can be assembled to recognize any DNA sequence. A limitation for this system arises due to the fact that the repeat amino acid sequence is almost identical, making the assembly of these arrays challenging at the cloning level. Nevertheless, this issue has been overcome using different protocols [27–30]. On the other hand, the highly repetitive sequence of the TALE open reading frames is a limitation, in fact when TALENs (TALE fused to nuclease effector protein FokI) open reading frames are transduced in lentiviral systems, extensive deletions within TALE repeats have been observed because of reverse transcriptase template switching [31, 32].

These limitations can be avoided using an alternative TALE-like scaffold such as BuDs [33]. This scaffold does not show any preference for the nucleotide at the 5‘ of the binding site, in contrast to TALE that seem to prefer a Thymine [25, 26,34,35], the repeats are polymorphic and, in addition, in vivo and in vitro data have shown the BuD arrays to have higher DNA-binding affinity for the target DNA compared with the classic TALE arrays [36].

These arrays can be fused to many effector proteins in order to deliver enzymatic activities such as nucleases both dimeric and monomeric [24, 37–41], transcription factors [42] and site-specific recombinases [43]. TALE arrays fused to nuclease effector protein FokI (TALEN) have been used to induce DSBs in the target sites for genome editing purposes. Because the FokI nucleases dimerize to become catalytically active, two TALE arrays are needed to target a DNA site (Fig. 2A). Therefore, a TALENs recognition sites contain two binding sites of 15–21 base pair and a spacer sequence of 12–20 bp. This setting has been successfully used to modify many genomes from yeast to human (see review [1]).

The clustered regularly interspaced short palindromic repeat-associated protein 9 system

When phages or plasmids enter in prokaryotic cells, small pieces of these exogenous DNAs, called protosparers, are integrated in the genome of the bacterial host between copies of repeat sequences. This array of repetitive sequences is known as CRISPR [44]. In type II CRISPR system, this array is processed into two ribonucleic acid (RNAs) CRISPR RNA (crRNA) and trans activating CRISPR RNA (tracrRNA). The crRNA carries the sequence transcribed from the protoscaler and a part of the CRISPR repeat that hybridizes with the scaffolding tracrRNA. These two RNA molecules together with the endonuclease CRISPR-Cas9 are able to recognize and cut the exogenous DNA, which has been previously encountered and processed into protosparers. To prevent the CRISPR system to cleave the host DNA, a 5‘ protoscaler associated motif (PAM) sequence that is absent in the CRISPR array of the host, needs to be present in the invading DNA. Although the CRISPR system is present in many bacterial genomes [45], the most extensively characterized type II CRISPR system is from Streptococcus pyogenes. In this system, it is possible to fuse together the crRNA and the tracrRNA into a single RNA molecule called guide RNA (gRNA). This gRNA, in complex with Cas9, recognizes a 20 bp sequence complementing the protoscaler adjacent to the PAM DNA sequence NGG [46], and induces DSBs three bases away from the PAM sequence (Fig. 2B).

Associated protein 9 is a 160 kDa protein that contains an HNH-nuclease and the RuvC-like nuclease domains. These nucleases cleave the DNA coding and non-coding strands of the target site [47]. Unlike the other ENs, CRISPR-Cas9 is addressed to the target site by an RNA molecule based on Watson–Crick rules. Thus just by modifying the gRNA sequence without any protein engineering, the specificity of the CRISPR-Cas9 can be exchanged. Moreover, Cas9 is a native endonuclease and no extra effector proteins need to be added to allow phosphodiester bond hydrolysis.

Although the initial studies on CRISPR-Cas9 indicated that this system recognizes a 20 bp target site adjacent to the PAM sequence, the specific recognition is limited to 7–12 nucleotides in the target sequence next to the PAM site [46, 48–51]. This model would suggest that CRISPR-Cas9 could not be used for specific genome-modifications, because a 12-base pair plus the three bases of the PAM does not provide a DNA sequence long enough to be unique in complex genomes. In fact the complexity of a 15 bp sequence (4^15 = 1.07 x 10^9) is less than the size of the human genome (3.2 x 10^9). To extend the recognition sites of CRISPR-Cas9, a double nickase approach has been developed including the fusion of Cas9 to the FokI endonuclease domain. Cas9 variants that cut only one strand of the DNA target have also been developed. These variants have performed well in genome editing [52]. Introduction of the D10A mutation in the RuvC-like or H840A in the HNH-nuclease domains results in the generation of nickases that cut either the non-coding or the coding strand of the DNA target [46, 47]. The inactivation of both nuclease domains results in a catalytically inactive Cas9 (dCas9), which conserves the capacity to bind a DNA target associated to the gRNA sequence. Many studies have used this inactive dCas9 to regulate the expression of genes by fusing it with transcription factors [53–56] or for localization of a given genomic locus [57]. Like TALENs, dCas9 can also be fused to FokI nucleases to recognize a bipartite target sequence [58]. In addition, CRISPR-Cas9 can also be used in multiplex
Figure 2. Transcription activator-like effector arrays fused to nuclease effector protein FokI (TALENs) and clustered regularly interspaced short palindromic repeats-associated protein 9 (CRISPR-Cas9) nucleases. A: The identity of the amino acid at position 13th of the TALE repeats determine the specificity of the array. B: A model of the TALENs including the FokI dimer during double-strand break (DSB) cleavage. In this setting, the target site contains two binding sites separated by 14–24 bps. Each TALEN binds their specific half of the target site inducing FokI dimerization and DSB formation. C: Structure of CRISPR-Cas9. The Cas9 protein, colored in grey, of Streptococcus pyogenes is depicted in complex with the guide ribonucleic acid (gRNA) and the DNA target [92]. The target site contains an entire coding strand (colored in cyan) recognized by the protospacer and a portion of the non-coding strand (colored in cyan) that carry the PAM sequence (colored in red), the rest of the non-coding strand is indicated in black and the putative position of it is indicated by the arrow. The HNH and the RuvC-like domain are colored in green and in magenta, respectively. The cleavage sites in the sequence are indicated by the red triangles.
approach in which different sequences are simultaneously targeted by different gRNAs [50]. The CRISPR-Cas9 system has also been used to modify the genome of many organisms (see review [59]).

**Specificity and efficiency**

One of the most important features of any targeted genome editing technique is specificity. The optimal tool must introduce modifications just on the target site, thus leaving the rest of the genome unmodified. TALEN and CRISPR-Cas9 are vastly used in genome editing; however, none of them has perfect DNA recognition specificity [5], so possible breaks can occur on other DNA sites in the genome. This off-target effect can introduce undesired changes in sequences of the genome with unpredictable consequences for cells, organs, organisms and even environments. For this reason the issue of specificity is a key in every ENs genome modification applications.

Despite the large use of TALE and CRISPR-Cas9 technologies, their specificity and reliability is still an open debate, including a reliable method to identify and quantify the off-target sites. Researchers have made an extensive use of these tools during the last 3–5 years, but concerns about this issue raised recently after failed attempts to modify the genome of human embryos [16, 17,19]. Cellular assays of off-target nuclease activity such as staining repair foci or cell cytotoxicity [1, 60] provide only a broad evaluation of this problem. More in depth analysis and accurate procedures need to be developed to guarantee the safe use of these nucleases [61].

A bioinformatics approach to predict off-target sequences similar to the on-target has been employed to test ENs specificity [62, 63], polymerase chain reaction amplification, T7 assay and deep sequencing have also been used to follow changes that occur on a subset of these predicted secondary targets. The frequency of mutations at these target sites varies from very low to as high as in the on-target site [64, 65]. The limitation of this method arises from the selection of sequences based on the degree of homology with the on-target site; indeed, it has been shown that off-target events can also occur in sequences that are poorly related to the target site [66]. Nevertheless, this type of approach is extremely useful in the initial selection of target sequences to favor the selection of sites as unique as possible in the target genome.

A more general approach to determine specificity of ENs is to test their activity against every possible off-target sequence. Because engineered endonucleases recognize long sequences, a large library that represents all possible combinations is needed. For example, in the case of an 18 base pair target, a library of \(4^{18}(7 \times 10^{10})\) DNA fragments would need to be generated. These types of large libraries are very difficult to produce and to handle. Such libraries can be downsized by partially randomizing the base position [67] in a way that at each position the probability to have the on-target base is higher compared with the alternative options [66–68]. Once these libraries are ready, in vivo or in vitro, systematic evolution of ligands by exponential enrichment approaches can be used to ‘fish-out’ the specific DNA (Fig. 3A). This method has been used to test the specificity of ENs both TALEN [66] and CRISPR-Cas9 [68] system but also zinc-finger nucleases [67] on the on-target site using a library containing \(10^{12}\) DNA fragments related to the specific sites.

Following this approach, two different TALENs targeting CC-chemokine receptor 5 and ataxia telangiectasia mutated (ATM) genes were designed. The TALEN targeting CC-chemokine receptor 5A induced a frequency of modification in the targeted site equal to 47%; while on the off-target sites, the frequencies ranged 2.3% to 0.006%. On the other hand, the TALEN targeting ATM induced a modification in its site of 18% and on the off-target sites at a frequency ranging from 0.94% to 0.006% [66]. In the case of CRISPR-Cas9 system, the test was performed with the sequences CLTA1 and CLTA4 of the clathrin light chain A (CLTA) gene. When the CRISPR-Cas9 specific for CLTA4 sequence was used, a frequency of mutation equal to 76% was observed on the target site, while the frequencies in the off-target sites ranged from 24% to 0.003%. For the CLTA1 sequence, a frequency of modification of 0.34% in the targeted site and frequencies ranging from 0.15% to 0.001% in the off-target sites were observed [68].

Although this downsized library method covers a large number of potential off-target sites, it has been shown that it cannot account for all the off-target sites [69]. A library of \(10^{12}\) DNA fragments related to the target site were used to test the specificity of the zinc-finger nucleases specific for CCR5 gene, one of the first and most studied engineered nucleases [1]. The analysis identified a large number of off-target sites for this EN [67]. The off-target activity of the same EN was also tested in genome-wide approach using integrase-defective lentiviral vectors (IDLV), revealing that the number of off-target sites was lower compared with the large library method [69]. Remarkably, each method identified some off-target sites that were missed by the other [70]. This indicates that the large libraries approach does not represent a fully unbiased methodology to understand and quantify off-target effects. Furthermore, this type of approach cannot take into account the imperfect binding that can occur when one or more bases of the DNA targets are not recognized. This type of frame-shifted interaction may happen by the formation of bulge between the gRNA and the target site for CRISPR-Cas9, but it also naturally occurs in some TALE [71–73]. Moreover, many of the sequences present in this library do not exist in the genome [68]. Thus a more unbiased approach, in which the entire genome is taken into account regardless of the on-target sequence, is necessary to test off-target effect of ENs.

**Genome break analysis**

When a DSB is repaired by NHEJ pathway, it is possible to observe at the site insertions or deletion of some base pairs (InDels). The frequency of these error-prone NHEJ events is still unclear; a recent paper indicates that cells are able to correct DSBs by NHEJ with a frequency of error-free events equal to 75% and that this frequency is most likely underestimated [7]. High-coverage whole-genome sequencing has been used to explore whether it is possible to employ ENs to obtain targeted pluripotent stem cells clones and to determine the magnitude of accidental modifications due to off-target cleavage [74–76]. Both CRISPR-Cas9 and TALENs were used to target the genes of interest, and the whole-genome sequencing
Figure 3. Methods used to determine off-target frequencies. A: SELEX (systematic evolution of ligands by exponential enrichment). A size-down library of sequences related to the on-target sites is incubated with the engineered nucleases (ENs), the bound fragments are selected and enriched after several cycles. DNA adaptors for next generation sequencing (NGS) analysis are ligated to the final fragments. B: ChIP-seq. A specific and inactive ENs is transfected in the cell. All the proteins bound to the DNA are fixed in vivo, by crosslinking reagent. The genomic DNA is purified and then sheared in small pieces. A specific antibody for the ENs is used to immunoprecipitate DNA fragments that are bound to the ENs. These fragments, after reversing the crosslinking, are analyzed by NGS. C: Genome-wide, unbiased identification of DSBs enabled by (sequencing) GUIDE-seq. In this approach a unique double-stranded oligodeoxynucleotides (dsODN) is transfected in the cells together with a specific ENs. In these cells the repair of the DSBs by the nonhomologous end joining (NHEJ) pathway could result in the integration of the dsODN in the genome. A pool of transfected cells is collected and the genomic DNA purified and then sheared. Sequencing adaptors are added on this DNA. To map the integration sites of the dsODN, a polymerase chain reaction (PCR) reaction using specific primers for the dsODN and the adaptors is performed. The PCR products are analyzed by NGS. D: High-throughput, genome-wide translocation sequencing (HTGTS). The cells are transfected with a specific EN that cuts a known sequence (bait site). The repair of the DSBs could induce chromosomal translocations by fusion the ends of two DNA DSBs one of which could be the bait site. To map these translocation events, linear amplification-mediated PCR (LAM-PCR) using a specific primer for the bait sequence is performed. Adaptors barcodes are added to the PCR product, and the translocation events are identified by NGS. E: Integrase-defective lentiviral vectors (IDLV). A specific ENs is transfected in the cells following by a transduction of IDLV carrying a puromycin resistant gene. In this cell, the repaired of the DSBs by the NHEJ pathway could result into the integration of the IDLV particle and the puromycin resisted gene in the genome. A pool of this puromicine resistant cell is collected and the genomic DNA purified. To map the IDLV integration site, linear amplification-mediated PCR (LAM-PCR) is performed using a specific primer for LTR (long terminal repeat) sequence present in the IDLV. The PCR product is analyzed by next generation sequencing (NGS) after adding the adaptors barcodes. F: Digenome-seq. Purified genomic DNA is cut in vitro by purified ENs protein. The digested and non-digested DNA are sheared and then sequenced by whole-genome sequence. The alignment of the non-digested DNA to a reference genome produced staggered alignment because of the formation of random DNA fragments during the shearing step. The alignment of the digested DNA generates a vertical alignment of the cleaved sites because the DNA fragments are specifically produced by an EN.
between 30× and 60× coverage was performed. These studies concluded that the mutations induced by ENs at off-target sites in individual human pluripotent stem cells clones are quite rare and that it is possible to isolate single human pluripotent stem cells containing very few ENs associated mutations. However, it is possible to see that the modified clones are not isogenic compared with the parental line and that each clone has acquired different new mutations during clonal expansion. Thus additional studies using more clones modified by ENs targeting different sites need to be analyzed to understand these observations [77]. Furthermore, whole-genome sequencing is unlikely to reveal the complete spectrum of off-target mutations induced by ENs, because it is technically difficult and economically expensive to use this technique to identify medium low/frequency mutations. In fact, in order to have a 95% chance to find an off-target mutation that occurs at a frequency of 10%, at least 15 diploid genomes need to be sequenced. If the mutation occurs at a frequency of 0.1%, 1500 diploid genomes need to be sequenced to have 95% of probability of finding this mutation at least once [77].

In addition, bioinformatics cleanup filters and other technical limitations, such as short-read sequencing, make difficult to distinguish the modifications induced by off-target activity of ENs from artifacts because of sequencing procedures [75–77]. Another limitation of this approach is that the cells correct DSBs mostly by NHEJ, which is supposed to be a DNA repair pathway that may introduce errors at the DNA site with high frequency [78]; however, it does not seem to be so error prone [7]. A study using an engineered meganuclease alone or transfected together with proteins that impact the NHEJ pathway reveals that when the NHEJ pathway is disturbed, a larger number of meganucleases cleavage events are detected and the efficiency of mutagenesis goes from 1.6% to 31%. [79].

The study also indicates that NHEJ repairs DSBs mostly in an error-free fashion, in fact the use of proteins that altered this pathway such as polymerases (Tdt) or exonucleases (Trex2) can induce up to 30-fold more mutations [79]. Thus, a genome-wide method that is sensitive enough to detect low frequency mutations and at the same time insensitive to the technical issues of the whole-genome sequencing may be a good option to study off-target effects of ENs.

### Labeling the binding sites

Chromatin immunoprecipitation sequencing (ChIP-seq) (Fig. 3B) has been used to determine the interaction of dCas9 in complex with different gRNAs in the context of the entire genomes [80–82]. The results reveal that for each gRNA, a clear ChIP-seq peak was observed at the intended target site, but other peaks at different sites were also identified. For some gRNA the number of non-specific sites has been reported to be more than 1200. Furthermore, while in general, the binding intensity of the CRISPR-dCas9 to the on-target site was greater than on the off-target sites; in some cases, some off-target sites showed a higher binding compared with the on-target [81].

**Determination of the frequency of InDels, by targeted deep-sequencing, on the on-targets sites and at some of off-target sites, identified by the ChIP-seq experiment, reveals that while the specific sites are modified at a frequency between 12% and 37% the off-target sites show an extremely low percentage of InDels formation [82].** These results indicate that DNA binding by CRISPR-Cas9 is more promiscuous than the cleavage activity [59]. However, because the formation of InDels upon repair of the DSBs by NHEJ is a quite rare event, the search for these variants is not a sensitive method to analyze the off-target activity of the ENs.

The ChIP-seq approach using monomeric transcription activator-like (TAL) fused to 3× FLAG epitope revealed the top-ranked binding site to correspond to the on-target site and no other ChIP-seq peaks that were reproducibly detected in the two biological replicates were identified. In addition, manual scanning of the genome for sequence motifs with one or two mismatches from the TAL monomer recognition site did not show any notable ChIP-seq peaks enrichment [83].

### Labeling the breaks

The nuclease activity can be also used to address off-target issues in the entire genome [71, 84–86]. These methods are unbiased because they capture DSBs without making assumptions about the off-target sequences (e.g., presuming that the off-target site is closely related in sequence to the on-target site). The advantage of this approach is that when DSBs events occur, they get ‘marked’, so they can be identified and in theory low frequency events can be easily revealed [71, 84–86].

The GUIDE-seq approach (genome-wide, unbiased identification of DSBs enabled by sequencing) relies on NHEJ-mediated capture of double-stranded oligodeoxynucleotides (dsODN) into induced DSBs by CRISPR-Cas9 (Fig. 3C). These methods consist of two steps: first, the DSBs induced by CRISPR-Cas9 in vivo are tagged by integration of a blunt dsODN; then, the dsODN integration sites in the genomic DNA are mapped by amplification and next-generation sequencing. GUIDE-seq reveals that the number of off-target sites varies greatly among different gRNAs from 0 to more than 150 sites [84]. Off-target sites found by GUIDE-seq were compared with the sites previously identified by ChIP-seq methods using the same gRNAs [81]; very little overlapping was found. Only three sites among the 149 off-target sites identified by GUIDE-seq were also observed by ChIP-seq experiments with dCas9. The authors concluded that binding, which is measured by using dCas9, and cleavage, measured by using active Cas9, are two different biological processes. However, GUIDE-seq failed to identify the seven well-defined off-target sites for these four gRNAs; these sites showed both binding by ChIP-seq and cleavage by InDels generation [81]. GUIDE-seq was not used to assess TALENs specificity because the breaks induced by TALENs generate 5' overhangs, and an additional modification of the dsODNs is needed to optimize their efficient capture into such breaks [84].

Another method for labeling the breaks is high-throughput genome-wide translocation sequencing (HTGTS). This procedure detects DSBs generated in the genome by engineered nucleases based on their translocation to other broken DNA sequence both from endogenous and exogenous events. These translocation events are identified by amplification of the known ‘bait’ sequences (i.e. the genomic sequence adjacent to the known nuclease target site), which is joined to an
unknown ‘prey’ sequences (i.e., sequences adjacent to off-target sites) (Fig. 3D). In that way, endogenous DSBs in lymphocyte lineage cells have been assessed by HTGTS and this method has also been used to reveal I-SceI meganuclease off-target sites in the mouse genome [87]. Based on this ability, the HTGTS technique was implemented together with linear amplification-mediated PCR [88] to study off-target cleavage of both CRISPR-Cas9, and TALEN [85]. The authors used four different gRNAs specific for recombination activating gene 1 to find between 0 and 40 translocation sites in the genome. When this technique was used to evaluate off-target activity of the previously described gRNAs targeting empty spiracles homeobox 1 (EMX1) and Vascular Endothelial Growth Factor A (VEGFA) genes [64], some of the sites previously identified by InDels analysis and ChIP-seq were found together with 12 and 34 novel off-target sites, respectively. In the case of TALENs targeting ATM or c-Myc, by HTGTS, it was possible to identify 522 and 384 off-target sites, respectively. Many of these sites are pseudo-palindromic sequences that corresponded to variants of the recognition site of a single TALEN monomer. Nevertheless, the use of forced heterodimerization for this template should dramatically decrease the off-target effect [66]. In addition, all the TALENs off-target sites, identified by HTGTS, were cut with lower frequency than Cas9 off-target sites [89]. This is the only method that suggested that TALEN displays higher off-target activity than CRISPR-Cas9. Similarly to GUIDE-seq, HTGTS required a bioinformatics filtering-out of the ‘catch’ sites. This filter is based on the degree of homology with the target site to discard the many false positive sites, which arise from randomly occurring DSBs. During this filtering process, some of the real off-target sites that are poorly related to the target site will be excluded for the further analysis [71].

Another approach to indicate the breaks caused by ENs is the IDLV capture and linear amplification-mediated PCR (Fig. 3E), which has been used to examine the off-target activity of TALEN and CRISPR-Cas9 after DNA cleavage. In this procedure, a linear double-stranded IDLV genome is incorporated preferentially into DNA DSBs by NHEJ [71, 90]. Six CRISPR-Cas9 and four TALENs targeting the same genomic regions in the human Wiskott-Aldrich syndrome (WAS) or tyrosine aminotransferase (TAT) genes were investigated. No off-target sites were detected for TALENs; in the case of the CRISPR-Cas9 system, 1–7 off-target sites were identified [71, 90]. The major limitation of this method is its sensitivity. In theory, it can identify mutations that occur with a frequency of 5% or higher; thus, mutations with lower frequencies are not detected. In addition, both GUIDE-seq and IDLV capture depend on NHEJ pathway, which has shown to seal the breaks with lower mutation frequencies than expected [7].

Finally, digenome-seq (in vitro nuclease-digested genome sequencing) uses cell free genomic DNA to analyze off-target sites in CRISPR-Cas9. In this approach, the genomic DNA from a haploid cell line is purified and digested in vitro by a purified Cas9 in complex with the gRNA (300 nM concentration of the CRISPR-Cas9 is used to ensure full cleavage) [86]. Because the digested DNAs should produce many fragments with the same 5’ ends at the cleavage site, after whole-genome sequencing, the cleaved fragments are computationally vertically aligned. The sequences that are not cleaved by ENs or the undigested DNA are aligned in a staggered manner.

**Figure 4.** Comparison of different off-target detection methods. Venn diagram comparing the off-target sites of a clustered regularly interspaced short palindromic repeats-associated protein 9 (CRISPR-Cas9) variant specific for the vascular endothelial growth factor A (VEGFA) gene identified by genome-wide, unbiased identification of DSBs enabled by sequencing (GUIDE-seq), chromatin immunoprecipitation sequencing (ChIP-seq), high-throughput genome-wide translocation sequencing (HTGTS) and (nuclease-digested genome sequencing) digenome-seq.

(Fig. 3F), thus revealing the differences in the DNA sites [86]. The Digenome-seq using an isolated CRISPR-Cas9 targeting the human β-globin gene on purified genomic DNA of cells showed 78 cleavages sites. However, when the same CRISPR-Cas9 was transfected in the cell, this procedure revealed the presence of 125 sites. This difference could arise from the expression of the ENs in the cell or the chromatin state of the cells, but other causes can be the origin of this variability. The 74 overlapped sites were further analyzed to identify off-target InDel formation. Seven sites showed InDels at a frequency between 87% and 0.11% [86]. This data suggested that although the CRISPR-Cas9 nuclease targets the human β-globin gene, it was also able to cut in many other sites. However, according to the in vitro assay, only in seven of these sites, it was possible to identify InDels formation in vivo. This result may be due to the action of the NHEJ repair pathway [79].

**Conclusions and future directions**

The genome-editing field has faced a large revolution in the last 5 years. In this work, we comment one of the most important aspect of targeted genome editing, the precision of the ENs used to generate the DSB. In particular, we have systematically analyzed the procedures currently available to determine off-target effects of ENs such as TALE and CRISPR-Cas9, which are the main tools used for this purpose in many laboratories. A systematic comparison of the sensitivity of these methods could be performed with the CRISPR-Cas9 system targeting the VEGFA gene. In this example, the same gRNA was used in four different methods InDels targeted by ChIP-seq [64], GUIDE-seq [84], HTGTS [85] and Digenome-seq [86] (Fig. 4). InDels sites targeted by ChIP-seq are 55, while the number of sites identifies by GUIDE-seq or HTGTS or Digenome-seq are 21, 38 and 87, respectively. The examination of the results shows that each technique revealed a new set of off-target
sites that were not identified with the other methods. Although some overlapping sites were found, the bottom line of this comparison suggests that none of the methods is able to identify all possible off-targets sites (Fig. 4). Furthermore, to fully compare all the procedures, the use of the same target site is most likely not sufficient. The cell lines, the delivery method of the ENs and other technical aspects need to be taken into account because they have been proved to be quite significant in determining the specificity of ENs [19]. Thus, despite the intense research in this area, there is no approach to completely address the off-target effects of ENs. This problem imposes a high hurdle in the translation of these techniques for biotechnological and therapeutic applications.

The power of these methods, the new possibilities open by their applications and the fact that the new templates to target specific DNA sites are easy to redesign and use, have raised concerns in the scientific community regarding biosafety. The fact that it may be difficult to identify whether an organism has been mutated or it has been genetically engineered makes the situation even more complicated. The economic interest of the possible biotechnological applications of these techniques is obvious and it does not help the debate. Although it is clear that the huge potential of the gene editing technology is going to progress and fuel the area, perhaps is also time to go back to basics in order to understand better the effect of these amazing tools in the organism. In fact some members of the scientific community involved in different aspects of genome editing have proposed a moratorium for the use of these techniques [19]. Sometimes, it is better to learn to walk before you run in order to avoid accidents.

Acknowledgement

The Novo Nordisk Foundation Center for Protein Research is supported financially by the Novo Nordisk Foundation (Grant agreement NNF14CC0001).

The authors declare no conflict of interests with this manuscript.

References

1 Carroll D. 2014. Genome engineering with targetable nucleases. Annu Rev Biochem 83: 409–39.
2 Wyman C, Kanaar R. 2006. DNA double-strand break repair: all’s well that ends well. Annu Rev Genet 40: 363–83.
3 Lieber MR. 2010. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. Annu Rev Biochem 79: 181–211.
4 Mimitou EP, Symington LS. 2009. Nucleases and helicases take center stage in homologous recombination. Trends Biochem Sci 34: 264–72.
5 Moore JK, Haber JE. 1996. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. Mol Cell Biol 16: 2164–73.
6 Boulton SJ, Jackson SP. 1996. Saccharomyces cerevisiae Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways. EMBO J 15: 5093–103.
7 Bétermin M, Bertrand P, Lopez BS. 2014. Is non-homologous end-joining really an inherently error-prone process? PLoS Genet 10: e1004086.
8 Urnov FD, Miller JC, Lee YL, Beausejour CM, et al. 2005. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature 435: 646–51.
9 Bibikova M, Beumer K, Trautman JK, Carroll D. 2003. Enhancing gene targeting with designed zinc finger nucleases. Science 300: 764–8.
10 Redondo P, Prieto J, Muñoz-IGG, Alibés A, et al. 2008. Molecular basis of xeroderma pigmentosum group C DNA recognition by engineered meganucleases. Nature 456: 107–11.
11 Muñoz-IGG, Prieto J, Subramanian S, Coloma J, et al. 2011. Molecular basis of engineered meganuclease targeting of the endogenous human RAG1 locus. Nucleic Acids Res 39: 729–43.
12 Stoddard BL. 2011. Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. Structure 19: 7–15.
13 Pabo CO, Peisach E, Grant RA. 2001. Design and selection of novel Cys2His2 zinc finger proteins. Annu Rev Biochem 70: 313–40.
14 Wolfe SA, Nekludova L, Pabo CO. 2000. DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct 29: 183–212.
15 Arnold U, Perez C, Cabaniols J-P, Smith J, et al. 2007. Engineered I-CreI derivatives cleaving sequences from the human XPC gene can induce highly efficient gene correction in mammalian cells. J Mol Biol 371: 49–65.
16 Liang P, Xu Y, Zhang X, Ding C, et al. 2015. CRISPR/Cas9-mediated gene editing in human trionuclear zygoites. Protein Cell 6: 363–72.
17 Cyranoski D, Reardon S. 2015. Embryo editing sparks epic debate. Nature 520: 593–4.
18 Lephrier E, Urnov F, Haecker SE, Werner M, et al. 2015. Don’t edit the human germ line. Nature 519: 410–1.
19 Baltimore D, Berg P, Botchan M, Carroll D, et al. 2015. Biotechnology. A prudent path forward for genomic engineering and germine gene modification. Science 348: 36–8.
20 Ledford H. 2015. CRISPR, the disruptor. Nature 522: 20–4.
21 Boch J, Bonas U. 2010. Xanthomonas AvrBs3 family-type III effectors: discovery and function. Annu Rev Phytopathol 48: 419–36.
22 Mak AN, Bradley P, Cermadas RA, Bogdanove AJ, et al. 2012. The crystal structure of TAL effector PXD101 bound to its DNA target. Science 335: 716–9.
23 Deng D, Yan C, Pan X, Mahfouz M, et al. 2012. Structural basis for sequence-specific recognition of DNA by TAL effectors. Science 335: 720–3.
24 Stella S, Molina R, Yefimenko I, Prieto J, et al. 2013. Structure of the AvrBs3-DNA complex provides new insights into the initial thymine-methylation recognition mechanism. Acta Crystallogr D Biol Crystallogr 69: 1707–16.
25 Boch J, Scholze H, Schornack S, Landgraf A, et al. 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. Science 326: 1509–12.
26 Moscou MJ, Bogdanove AJ. 2009. A simple cipher governs DNA recognition by TAL effectors. Science 326: 1501.
27 Briggs AW, Rios X, Charri R, Yang L, et al. 2012. Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. Nucleic Acids Res 40: e117.
28 Cermak T, Doyle EL, Christian M, Wang L, et al. 2011. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res 39: e82.
29 Schmid-Burgk JL, Schmidt T, Kaiser V, Höning K, et al. 2013. A ligation-independent cloning technique for high-throughput assembly of transcription activator–like effector genes. Nat Biotechnol 31: 76–81.
30 Reyno D, Tsai SQ, Khayter C, Foden JA, et al. 2012. FLASH assembly of TALENs for high-throughput genome editing. Nat Biotechnol 30: 460–5.
31 Yang L, Guell M, Byrne S, Yang JL. 2013. Optimization of scarless human stem cell genome editing. Nucleic Acids Res 41: 9049–61.
32 Holkers M, Maggio I, Liu J, Janssen JM, et al. 2013. Differential integrity of TALE nuclelease genes following adenovalend and lentiviral vector gene transfer into human cells. Nucleic Acids Res 41: e63.
33 Juillerat A, Bertonati C, Dubois G, Guyot V, et al. 2014. BURH: a new modular DNA binding protein for genome engineering. Sci Rep 4: e003831.
34 Doyle EL, Hummel AW, Demorest ZL, Stacker CG, et al. 2013. TAL effector specificity for base 0 of the DNA target is determined in a complex, effector- and assay-dependent manner by substitutions for the tryptophan in cryptic repeat –1. PLoS One 8: e82120.
35 Christian ML, Demorest ZL, Stacker CG, Osborn MJ, et al. 2012. Targeting G with TAL effectors: a comparison of activities of TALENs constructed with NN and NK repeat variable di-residues. PLoS One 7: e53833.
36 Stella S, Molina R, López-Méndez B, Juillerat A, et al. 2014. BuDH, a helix-loop-helix DNA-binding domain for genome modification. Acta Crystallogr D Biol Crystallogr 70: 2042–9.
37 Li T, Huang S, Jiang WZ, Wright D, et al. 2011. TAL nucleases (TALENs); hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. Nucleic Acids Res 39: 359–72.
38 Christian M, Cermak T, Doyle EL, Schmidt C, et al. 2010. Targeting DNA double-strand breaks with TAL effector nucleases. Genetics 186: 757–61.
of genomic loci in living human cells by an optimized CRISPR/Cas9

system. Mol Cell 2013; 52: 991–997.

Chen H, Nucleic Acids Res 2014; 42: 8864–68.

Sternberg SH, Tsai SQ, 58

Jinek M

Mojica FJ

Ran FA

Jiang W

51

Perez-Pinera P

Doyle EL

Mercer AC

Cong L

Miller JC

Qi LS

Bioessays

June 15, 2015. Unbiased detection of offtarget cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. Nat Biotechnol 2013; 175–8.

Lin Y, Cradick TJ, Brown MT, Deshmukh H, et al. 2014. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic Acids Res 42: 7473–85.

Richter A, Streubel J, Blücher C, Szurek B, et al. 2014. A TAL effector repeat architecture for frameshift binding. Nat Commun 8: 3447.

Smith C, Gore A, Yan W, Abalde-Atristain L, et al. 2014. Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. Cell Stem Cell 15: 12–3.

Xie Y, Yuan T, Goebel A, Tang S, et al. 2014. Targeted gene correction minimally impacts whole-genome homologous with melting in vitro selection. Cell Rep 3: 1319–24.

Kuscu C, Arslan S, Singh R, Thorpe J, et al. 2014. Genome-wide analysis reveals characteristics of off-target sites bound by CRISPR endonucleases. Cell Stem Cell 2014; 77: 977–9.

Wu X, Scott D, Krix A, Chu A, et al. 2014. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat Biotechnol 2014; 32: 670–6.

Mendell J, et al. 2013. Locus-specific editing of histone modifications at endogenous enhancers. Nat Biotechnol 2013; 1133–6.

Tsai SQ, Zhang Z, Nguyen NT, Liebers M, et al. 2015. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol 2015; 33: 187–97.

Frock RL, Hu J, Meyers RM, Ho Y-JJ, et al. 2015. Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. Nat Biotechnol 2015; 1179–90.

Kim D, Bae S, Park J, Kim E, et al. 2015. Digene-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nat Methods 12: 237–43.

Chiarle R, Zhang Y, Frock RL, Lewis SM, et al. 2011. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell 147: 107–19.

Schmidt M, Schwarzwälder K, Bartholomae C, Zasoi K, et al. 2007. High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat Methods 4: 1051–7.

Gabriel R, von Kalle C, Schmidt M. 2015. Mapping the precision of genome editing. Nat Biotechnol 2015; 33: 150–2.

Osborn MJ, Starker CG, McCloy AN, Webber BR, et al. 2013. TALEN-based gene correction for epimutation bullosa. Mol Ther 21: 1151–9.

Maggio I, Gonçalves MA. 2015. Genome editing at the crossroads of delivery, specificity, and fidelity. Trends Biotechnol 2013; 30: 80–91.

Anders C, Niewoehner O, Duerst A, Jinek M. 2014. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature 513: 569–73.