CHAPTER THREE

Determining the Specificities of TALENs, Cas9, and Other Genome-Editing Enzymes

Vikram Pattanayak*,1, John P. Guilinger†,1, David R. Liu†,2

*Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA
†Department of Chemistry & Chemical Biology, Harvard University, Cambridge, Massachusetts, USA
Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts, USA
1These authors contributed equally to this work.
2Corresponding author: e-mail address: drliu@fas.harvard.edu

Contents
1. Introduction 48
   1.1 Introduction to programmable nucleases for genome editing 48
   1.2 Overview of methods to study specificity of genome-editing agents 49
   1.3 Insights and improvements from ZFN specificity studies 57
   1.4 Insights and improvements from TALEN specificity studies 59
   1.5 Insights and improvements from Cas9 specificity studies 61
2. Methods 65
   2.1 Overview of in vitro selection-based nuclease specificity profiling 65
   2.2 Pre-selection library design 65
   2.3 In vitro selection protocol 66
   2.4 Confirmation of in vitro-identified genomic off-target sites 71
3. Conclusion 73
Acknowledgments 73
References 74

Abstract
The rapid development of programmable site-specific endonucleases has led to a dramatic increase in genome engineering activities for research and therapeutic purposes. Specific loci of interest in the genomes of a wide range of organisms including mammals can now be modified using zinc-finger nucleases, transcription activator-like effector nucleases, and CRISPR-associated Cas9 endonucleases in a site-specific manner, in some cases requiring relatively modest effort for endonuclease design, construction, and application. While these technologies have made genome engineering widely accessible, the ability of programmable nucleases to cleave off-target sequences can limit their applicability and raise concerns about therapeutic safety. In this chapter, we review methods to evaluate and improve the DNA cleavage activity of programmable site-specific endonucleases and describe a procedure for a comprehensive off-target profiling method based on the in vitro selection of very large (~10^{12}-membered) libraries of potential nuclease substrates.
1. INTRODUCTION

1.1. Introduction to programmable nucleases for genome editing

Programmable site-specific nucleases such as zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and CRISPR-associated Cas9 nucleases can be designed to target any gene of interest and therefore are powerful research tools with significant therapeutic implications. In cells, a targeted double-strand break can lead to gene modification or insertion through homology-directed repair (HDR) with exogenous DNA or to gene knockout via nonhomologous end-joining (NHEJ). In the HDR pathway, the creation of a double-strand break at a chromosomal DNA locus by a sequence-specific endonuclease can increase the efficiency of insertion of an exogenous donor DNA template by several orders of magnitude (Choulika, Perrin, Dujon, & Nicolas, 1995). If no donor template is provided, endogenous NHEJ pathways that repair the break will often introduce missense mutations that abrogate production of functional protein product (Lukacsovich, Yang, & Waldman, 1994; Rouet, Smih, & Jasin, 1994). Programmable nucleases have been used to modify the genomes of a variety of organisms and human cell lines, as has been reviewed extensively (Carroll, 2011; Joung & Sander, 2013; Sander & Joung, 2014). In addition to engineering the genomes of cells or organisms for direct biological interrogation, genetic screens have recently been performed with these enzymes in human tissue culture to uncover genetic factors underlying specific cellular processes in an unbiased manner (Koike-Yusa, Li, Tan, Velasco-Herrera Mdel, & Yusa, 2014; Shalem et al., 2014; Wang, Wei, Sabatini, & Lander, 2014; Zhou et al., 2014).

These nucleases also serve as the promising basis of a new generation of human therapeutics. Clinical trials of two site-specific nucleases are currently underway as potential treatments for HIV and glioblastoma. Researchers have completed and are conducting clinical trials using a ZFN, developed by Sangamo BioSciences, that targets a sequence in the CCR5 gene (Tebas et al., 2014). CCR5 is a co-receptor used by HIV in early-stage infection (Scarlatini et al., 1997), and mutation of CCR5 (CCR5Δ32) is known to confer resistance to HIV infection (Huang et al., 1996; Liu et al., 1996; Samson et al., 1996).

The second ZFN in clinical trials, also developed by Sangamo BioSciences, disrupts the gene for the glucocorticoid receptor (Reik et al.,
2008) as part of a potential treatment for glioblastoma. The target cells for the ZFN are T cells modified by other methods to express a cell-surface receptor that specifically recognizes malignant glioblastoma cells (Kahlon et al., 2004). The therapeutic cells, however, are rendered inactive by glucocorticoids, which are often also a component of therapy. ZFN-mediated modification of the glucocorticoid receptor in the therapeutic cells confers resistance to glucocorticoid treatment, while maintaining anti-glioblastoma activity, allowing the cells to recognize their malignant targets. These and other examples demonstrate that, in addition to serving as powerful research tools, programmable nucleases are promising platforms for clinically relevant genetic manipulation.

1.2. Overview of methods to study specificity of genome-editing agents

Specificity is a crucial feature of programmable endonucleases, and a high (though currently undefined) level of specificity is desired for the vast majority of therapeutic applications. Until recently, however, few methods existed to study the DNA cleavage specificity of active, site-specific nucleases. An ideal study of off-target activities of site-specific endonucleases would measure nuclease activity against each of the >10^9 potential off-target sites for every target site in the human genome. While whole-exome sequencing has been used in studies of site-specific endonuclease specificity (Cho et al., 2014; Ding et al., 2013; Li et al., 2011b), sequencing offers limited sensitivity in detecting rare off-target events, and exomes represent only a small fraction of genomic DNA containing potential off-target sites. Therefore, the general study of off-target activities of site-specific endonucleases has relied on the experimental identification of likely off-target sites. Off-target studies have taken one of three general forms: discrete off-target site testing, genome-wide selections, and minimally biased in vitro selections (Fig. 3.1).

1.2.1 Discrete off-target site testing

Perhaps, the most obvious approach to evaluating the sequence specificity of nucleases is by assaying discrete potential off-target substrates, either in a low- or high-throughput format. While the methods summarized below are not a comprehensive list of such efforts, they are representative examples of this strategy.

Homing endonucleases such as I-SceI were the subjects of some of the earliest studies of the specificity of nucleases that recognize sites sufficiently long to be unique in the human genome, even though the presence of
integrated binding and cleavage domains complicates engineering homing endonucleases with tailor-made specificities (Chen, Wen, Sun, & Zhao, 2009; Chen & Zhao, 2005; Doyon, Pattanayak, Meyer, & Liu, 2006; Gimble, Moure, & Posey, 2003). In early studies of I-SceI homing endonuclease specificity, Dujon and coworkers interrogated a subset of the 54 potential single-mutant individual off-target sequences of the 18 base pair target site (Colleaux, D’Auriol, Galibert, & Dujon, 1988).

The throughput of this approach was increased in the multi-target ELISA method, initially applied to zinc-fingers, developed by Barbas and coworkers (Segal, Dreier, Beerli, & Barbas, 1999), in which 96 biotinylated oligonucleotides or oligonucleotide pools are plated individually in the streptavidin-coated wells of a 96-well plate. Fusions to maltose-binding protein of a DNA-binding domain of interest are incubated with the oligonucleotides in the wells. After a wash step to remove unbound protein, the wells are incubated with a primary antibody that recognizes maltose-binding protein, followed by a secondary antibody that allows visualization of wells containing bound protein.

Church and coworkers (Bulyk, Huang, Choo, & Church, 2001) have used a microarray approach to study zinc-finger DNA-binding specificity. They prepared DNA microarrays containing all 64 possible three-base pair
subsequences within a longer target site. The microarrays were incubated with M13 phage displaying the DNA-binding domain of interest, washed, and visualized with primary and secondary antibody staining to reveal DNA-binding specificities. A variant of this method, developed by Bulyk and coworkers (Philippakis, Qureshi, Berger, & Bulyk, 2008), has been extended to profiling 10-base pair subsequences of transcription factor binding sites. Another microarray-based method (Carlson et al., 2010) has also been used to profile the DNA-binding specificity of engineered zinc fingers. More recently, discrete testing of potential single- and double-mutant off-target sequences has been used in human cells to study the sequence preferences of Cas9. In these methods, a single target site in human cells is assayed for its ability to be modified through NHEJ by a set of endonucleases that are targeted to cleave either the target site or discrete single- or multiple-mutant variants of the target site. At least two separate studies have used this strategy. In one study by Joung and coworkers, an eGFP reporter is the target of a collection of Cas9:guide RNA complexes containing mutant (mismatched) guide RNAs (Fu et al., 2013). In this approach, off-target endonuclease activity leads to the loss of cellular GFP expression. A second study, developed by Zhang and coworkers (Hsu et al., 2013), assayed the ability of a set of Cas9:guide RNA complexes to cleave the \textit{EMX1} gene. Cleavage activity was detected as NHEJ events at the \textit{EMX1} locus using high-throughput sequencing. Although if Cas9 cleaved with perfect specificity the site would not be modified by Cas9:guide RNA complexes containing mutated guide RNAs, many of the mutated guide RNAs resulted in NHEJ, thereby demonstrating off-target activity. In both methods, other potential genomic off-target sites are extrapolated from the small set of off-target sites directly screened. The results of this approach applied to Cas9 are summarized in Section 1.5 and further demonstrate the utility of simple, discrete-off-target site testing to identify genomic off-target sites.

### 1.2.2 Genome-wide selections

In contrast to discrete screening assays of potential off-target sequences to be cleaved by a nuclease of interest, genome-wide selections have also been used to identify those sequences in a population of human cells that can bind to or are cleaved by a nuclease of interest. In assessments of genome-wide binding of Cas9, Adli, Sharp, Zhang, and their respective colleagues (Kuscu, Arslan, Singh, Thorpe, & Adli, 2014; Wu et al., 2014) used chromatin immunoprecipitation followed by sequencing (ChIP-seq) to study the
ability of inactive Cas9 to bind off-target sequences in the genome. In this method, hemagglutinin-tagged, catalytically inactive Cas9 is expressed in human cells. A crosslinking step covalently attaches the tagged Cas9 to any DNA target sites it is bound to in the cell. The bound DNA is then fractionated, the crosslinks are reversed, and high-throughput sequencing of the resulting DNA reveals the genomic sequences bound by Cas9. While these studies show that Cas9 is capable of extensively binding off-target sites, they also suggest that most of the off-target sites bound are not modified.

Genome-wide selections for DNA cleavage, rather than binding alone, have been achieved by exploiting the tendency of certain viruses to preferentially integrate at sites of double-strand breaks. The endonuclease of interest is expressed in cultured human cells, creating double-strand breaks at cleaved genomic sites. Cells are then exposed to a virus that preferentially integrates at double-strand breaks. Genomic DNA sequences containing integrated virus are then identified through selection or direct DNA sequencing. In a selection method developed by Miller and coworkers, adeno-associated virus packaged with antibiotic resistance markers and an E. coli plasmid origin is used as an integration marker (Petek, Russell, & Miller, 2010). Any on-target and off-target substrates in the genome containing the integration marker would therefore contain a plasmid origin and antibiotic resistance markers. Genomic DNA is then isolated from infected cells, fragmented with a cocktail of restriction enzymes, circularized, and transformed into E. coli. Only fragments containing integrated adeno-associated virus have an E. coli origin of replication and the appropriate antibiotic resistant markers, and therefore only fragments containing integrated virus survive. Sequencing of the plasmid reveals the viral-chromosomal junctions, which contain the off-target sites of the endonuclease. Subsequent studies by von Kalle, Tolar, and their respective coworkers to study the specificities of ZFNs and TALENs have extended this approach, using instead integrase-deficient lentiviral vectors (IDLVs) and read-out of integration sites using high-throughput sequencing (Gabriel et al., 2011; Osborn et al., 2013).

Advantages of viral integration methods include their abilities to study specificity directly in the context of the target genome and the unbiased nature of the selection, allowing for identification of off-target sites that are not highly similar in sequence to the on-target site. Results from these methods should be interpreted carefully, however, as integration can occur at double-strand breaks that arise naturally, independent of nuclease activity. As with whole-genome sequencing, viral integration methods may not be
sufficiently sensitive to detect low-level off-target modification. In addition, abstraction of general properties of endonuclease specificity could be complicated by cellular factors such as DNA accessibility, which varies from site to site and between cell types (Maeder et al., 2008; Wu et al., 2014).

1.2.3 Minimally biased selections in vitro and in cells

The most general method to determine site-specific endonuclease specificity would test the activity of a given endonuclease against each potential off-target sequence. Since therapeutic endonucleases target long sequences (≥ ~20 base pairs) to ensure uniqueness in the genome, a truly comprehensive specificity study would require an assay with at least $4^{20}(\sim 1 \times 10^{12})$ different substrates. Since libraries of this size are challenging to generate and process even using in vitro methods, selections to determine site-specific endonuclease specificity either rely on the use of “minimally biased” libraries or focus on smaller subsets of the DNA substrate to be studied. Minimally biased libraries are randomized across the nucleotide positions being studied, but the composition of nucleotides at each position is biased toward the target sequence rather than fully randomized. For example, if a particular target site of a three-base pair specific endonuclease is ATG, a fully randomized library would contain equal proportions of all sequences (NNN). A minimally biased library contains higher proportions of sequences that are similar to the target site. In this example, the most common sequence in the library would be ATG, followed by the single-mutant sequences (cTG, gTG, tTG, AaG, AcG, AgG, ATa, ATc, ATt), the double-mutant sequences, and then triple-mutant sequences, which are the rarest in the library. Biasing is accomplished through the incorporation of mixtures of nucleoside phosphoramidites at each position during DNA synthesis, such that the on-target base is incorporated at a higher frequency than the other off-target bases (Fig. 3.2A). In other variants of this approach, portions of the sequence are fixed, while subsets are fully randomized (for example, nTG, AnG, or ATn).

Using minimally biased libraries, several methods have studied the binding specificities of monomeric zinc-finger domains, in the absence of cleavage domains and dimeric binding partners. In the bacterial one-hybrid system developed by Wolfe and coworkers (Meng, Brodsky, & Wolfe, 2005), a DNA target site library is placed upstream of a selectable marker on a plasmid. The DNA-binding domain of interest is expressed in E. coli as a fusion to the α-subunit of RNA polymerase. In each individual bacterium, which each contains only one member of the target site library, RNA
Figure 3.2 (See figure legend on next page.)
polymerase is recruited to the promoter of the selectable marker if the DNA-binding domain is able to bind to the library DNA sequence present. Only cells that have target sites that can be bound by the DNA-binding domain will express the selectable marker and survive. Wolfe and coworkers have used this approach to assay libraries of up to $10^8$ molecules for DNA binding (Meng, Thibodeau-Beganny, Jiang, Joung, & Wolfe, 2007). A computational structure-based approach developed by Bradley and colleagues (Yanover & Bradley, 2011) using the Rosetta algorithm has also been used to study monomeric zinc-finger domain specificity and has accurately predicted DNA-binding profiles that were obtained by the bacterial one-hybrid system.

Church and coworkers recently used a variant of the bacterial one-hybrid approach to study the specificity of Cas9 in human cells (Mali et al., 2013). In this method, a library of target sites was placed upstream of a reporter gene in a plasmid. Instead of using active Cas9 as an endonuclease in the selection, an inactive variant was expressed as a DNA-binding domain alone, fused to the VP64 activation domain. Therefore, any inactive Cas9 that could bind to a library member caused expression of the reporter gene. The results of this study are summarized in Section 1.5.

Larger libraries, covering more potential off-target sites, have been used to evaluate DNA-binding domain specificity in vitro. Applying an in vitro SELEX approach (Oliphant, Brandl, & Struhl, 1989) to large (~$10^{14}$-membered) libraries of randomized target site DNA (Miller et al., 2011), Struhl and coworkers, and later several other groups, enriched DNA sequences that can bind a given DNA-binding domain of interest.

**Figure 3.2** In vitro selection scheme for profiling the specificity of site-specific nucleases. (A) Example sequences biased toward a target sequence for both the left- and right-half sites of TALEN targeting the human CCR5 gene. The on-target sequences are in bold and below are examples of variant sequences from minimally biased libraries. (B) A single-stranded library of DNA oligonucleotides containing partially randomized target sites (gray box) and constant region (thick black line) is circularized, then transformed into concatenemeric repeats by rolling-circle amplification. The concatenemeric repeats of double-stranded DNA (double arrows) target site variants are incubated in vitro with a site-specific nuclease of interest. The resulting cleaved ends are blunted and then are ligated to adapter #1. The ligation products are amplified by PCR using one primer consisting of adapter #1 and the other primer consisting of adapter #2—constant sequence, which anneals to the constant regions of the library. From the resulting ladder of amplicons containing 0.5, 1.5, 2.5, ... repeats of a target site, amplicons corresponding to 1.5 target-sites in length are isolated by gel purification and subjected to high-throughput DNA sequencing and computational analysis.
(Thiesen & Bach, 1990; Zykovich, Korf, & Segal, 2009). In this approach, the DNA-binding domain is immobilized and incubated with a randomized target site library. After washing steps to remove unbound DNA, the bound DNA is eluted, amplified, and cycled through the procedure several times before being sequenced.

The bacterial one-hybrid and SELEX methods described above study DNA-binding domains alone, outside of the context of catalysis. Since site-specific endonucleases involve DNA cleavage in addition to DNA binding, and since DNA-binding specificities may not exactly predict DNA cleavage specificities, methods to study the specificity of DNA cleavage reactions are desirable. Monnat and coworkers developed a gel electrophoresis-based method (Argast, Stephens, Emond, & Monnat, 1998) in which an active endonuclease is incubated \textit{in vitro} with a target site library that was cloned into a circular plasmid. Cleavage of library members results in linearization of the plasmid, and the pool of cleavable, linearized DNA sequences is separated from uncleaved, circular DNA through agarose gel electrophoresis and gel purification. The linear DNAs containing bona fide substrate sequences are ligated back into circles and amplified in \textit{E. coli}. After several rounds of enrichment of a pre-selection library with a theoretical complexity of $10^8$ to $10^9$ members (constrained by the need to introduce library members into \textit{E. coli}), the post-selection library is sequenced and analyzed.

To combine the benefits of both large library sizes and the context of cleavage selection, Liu and coworkers developed a fully \textit{in vitro} selection strategy to profile the DNA cleavage specificity of ZFNs, TALENs, and Cas9 using libraries of $10^{11}$ to $10^{12}$ potential off-target sites (Guilinger et al., 2014; Pattanayak et al., 2013; Pattanayak, Ramirez, Joung, & Liu, 2011). In this strategy, library construction is performed entirely \textit{in vitro} and therefore is not bottlenecked by cell transformation efficiency. This method, which is described in detail in Section 2, uses the generation of $5'$ phosphates upon DNA cleavage to selectively tag and amplify library members that are cleaved by nucleases. These cleaved library members are then revealed by high-throughput DNA sequencing (Fig. 3.2B).

When applied to ZFNs, this \textit{in vitro} DNA cleavage specificity profiling strategy demonstrated that a SELEX study on the specificity of individual DNA-binding domains, in the absence of dimerization and cleavage, did not detect some genomic off-target sites. Analysis of hundreds of thousands of off-target sites cleaved \textit{in vitro} suggested that interactions between ZFN monomers affect DNA cleavage specificity and explain differences with the SELEX study (Pattanayak et al., 2011). For the CCR5-targeting
ZFNs described above, the in vitro cleavage selection also identified more genomic off-target sites than a genome-wide selection method on the same ZFN using IDLVs reported by von Kalle and coworkers (Gabriel et al., 2011). However, each method identified off-target sites that were missed by the other method. As was also demonstrated for SELEX results (Perez et al., 2008), additional computational analysis of in vitro selection results improves the sensitivity of the in vitro cleavage selection method to determine off-target sites. Joung and coworkers (Sander et al., 2013) applied a machine-learning classifier algorithm to in vitro cleavage selection results for the CCR5-targeting ZFN, and identified 26 more off-target sites than had previously been identified, including all of the previously determined off-target sites. These studies collectively demonstrate how in vitro selection methods and genome-wide selection methods can serve as complementary tools in the determination of gene-editing nuclease specificities.

1.3. Insights and improvements from ZFN specificity studies

ZFNs (Kim, Cha, & Chandrasegaran, 1996) are dimeric fusions of the non-specific FokI restriction endonuclease cleavage domain (Hirsch, Wah, Dorner, Schildkraut, & Aggarwal, 1997) with zinc-finger DNA-binding domains (Fig. 3.3A). The FokI cleavage domain must dimerize to be active; therefore, ZFNs can cleave DNA only after dimerizing and bridging two half-sites (Vanamee, Santagata, & Aggarwal, 2001) that are separated by an unspecified DNA spacer sequence. Target site specificity is therefore determined by two zinc-finger DNA-binding domains, each of which consists of three or more tandem repeats of individual zinc fingers. Each individual zinc finger recognizes three base pairs (Beerli, Segal, Dreier, & Barbas, 1998), and a zinc-finger DNA-binding domain in total recognizes at least nine base pairs. Therefore, in total, ZFNs recognize sites that are at least 18 bp long (not including the spacer).

The DNA-binding specificity of ZFNs is programmed by the composite individual zinc fingers. Each individual zinc finger consists of a compact ββα fold with a hydrophobic core stabilized by a zinc ion coordinated by two cysteines and two histidines. While a great deal of progress has been reported in the design of zinc fingers that can target any DNA triplet, primarily by Barbas, Joung, Klug, Pabo, and their respective coworkers (Beerli et al., 1998; Choo, Sanchez-Garcia, & Klug, 1994; Dreier, Beerli, Segal, Flippin, & Barbas, 2001; Dreier, Segal, & Barbas, 2000; Dreier et al., 2005; Maeder et al., 2008; Rebar & Pabo, 1994; Sander et al., 2011;
Figure 3.3 Architecture of ZFN, TALEN, and Cas9 programmable nucleases. (A) A ZFN monomer is a fusion of a FokI nuclease cleavage domain (purple; dark gray in print version) to a set of adjoined zinc fingers (four, in this example) each targeting three base pairs (for a total of 12 base pairs recognized, in this example). Two different ZFN monomers bind their corresponding half-sites, allowing FokI dimerization and DNA cleavage between the half-sites. (B) A TALEN monomer contains an N-terminal domain followed by an array of TALE repeats (filled boxes), a C-terminal domain, and a FokI nuclease cleavage domain (purple; dark gray in print version). The 12th and 13th amino acids (the RVD, red; dark gray in print version) of each TALE repeat recognize a specific DNA base pair. Two different TALEN monomers bind their corresponding half-sites, allowing FokI dimerization and DNA cleavage between the half-sites. (C) Cas9 protein (yellow; light gray in print version) binds to target DNA in complex with a single guide RNA (sgRNA, green; light gray in print version). The S. pyogenes Cas9 protein and sgRNA complex recognizes the PAM sequence NGG (blue; light gray in print version). Black triangles indicate the cleavage points in the target DNA three bases from the PAM on both DNA strands.
Wu, Yang, & Barbas, 1995), designing the multi-finger domains of a ZFN often requires selection (Greisman & Pabo, 1997; Isalan, Klug, & Choo, 2001; Maeder et al., 2008) or computational approaches (Sander et al., 2011) such as those described by Joung and coworkers.

Initial studies of ZFN specificity using SELEX on zinc-finger binding domains alone (Perez et al., 2008) suggested that ZFNs are highly specific, especially when heterodimeric versions, first developed by Rebar and Cathomen, are used (Miller et al., 2007; Szczepk et al., 2007). The heterodimeric ZFNs have mutations in the FokI cleavage domain that only allow dimerization between different ZFN monomers (Doyon et al., 2011; Miller et al., 2007; Szczepk et al., 2007). While many CCR5 ZFN off-target sites have been identified (Gabriel et al., 2011; Pattanayak et al., 2011, Sander et al., 2013), to date, no toxicity has been reported in clinical trials (Tebas et al., 2014).

In addition to identifying genomic off-target sites, in vitro selections on two different ZFNs by Liu and coworkers also illuminated several general properties of ZFN specificity (Pattanayak et al., 2011). Like other enzymes, ZFNs exhibit concentration-dependent specificity, such that a larger set of off-target sites can be cut when the ZFN is at higher concentration. In general, ZFN off-target sites with a small number of mutations (for example, for the CCR5 ZFN, three or fewer mutations out of 24 target base pairs) can be recognized and cleaved. Although no sequence preference in the spacer region between half-sites was observed, sites with disfavored four- and seven-base pair spacers were generally recognized with greater specificity than sites with the more favored five- and six-base pair spacers. Finally, off-target sites with several mutations in one half-site likely contain few to no mutations in the other half-site. All of these observations are consistent with a model in which ZFN:DNA-binding energy must meet a minimum threshold for cleavage to take place, and that off-target cleavage activity arises from excess binding energy between a ZFN and DNA that can tolerate the energetic penalty incurred by protein–DNA mismatches.

1.4. Insights and improvements from TALEN specificity studies

Like ZFNs, TALENs are engineered fusions of DNA-binding domains with FokI nuclease domains (Fig. 3.1B). In the case of TALENs, the DNA-binding domains consist of TALE repeat arrays (Christian et al., 2010; Li et al., 2011a, 2011b; Miller et al., 2011). TALE repeats are naturally found in the plant pathogen Xanthomonas and are part of transcriptional activator
proteins that lead to gene expression upon binding to specific promoter elements in the plant host cell (Gu et al., 2005; Kay, Hahn, Marois, Hause, & Bonas, 2007; Yang, Sugio, & White, 2006). Canonical TALE repeats are 34-amino acid sequences that each recognize one base pair of DNA. The DNA-binding specificity of each repeat is determined by two amino acids referred to as the repeat-variable di-residue (RVD) (Boch et al., 2009; Moscou & Bogdanove, 2009). Examples of RVDs that recognize each of the four DNA base pairs are known. The only known sequence constraint on TALE repeat domains is a requirement for the 5′ end of the target site to contain deoxythymidine (T). Beyond this requirement, TALEs can be designed to target virtually any DNA sequence, and have been successfully used to manipulate genomes in a variety of organisms (Cermak et al., 2011; Moore et al., 2012; Tesson et al., 2011; Wood et al., 2011) and cell lines (Hockemeyer et al., 2011; Mussolino et al., 2011; Reyon et al., 2012).

Multiple studies, using genome-wide studies and minimally biased selections, have demonstrated that TALEN-mediated genome modification can be accompanied by very rare off-target effects. Whole-genome sequencing of TALEN-treated yeast strains (Li et al., 2011b) and whole-exome sequencing of human cell lines derived from TALEN-treated cells (Ding et al., 2013) revealed no evidence of TALEN-induced genomic off-target mutations. However, whole-genome sequencing may not be sensitive to detecting rare mutations in the absence of sequencing the genomic DNA from an impractically large number of treated cells.

Discrete DNA cleavage studies using homology to on-target sequences to predict potential off-target sites found no TALEN-induced modification of potential off-target sites in Xenopus (Lei et al., 2012) and human cell lines (Kim et al., 2013). Several groups have studied the specificity of the TALE repeat DNA-binding domains in isolation, in the absence of cleavage domains. Initial minimally biased selection experiments using SELEX and TALE activator binding (Hockemeyer et al., 2011; Mali et al., 2013; Miller et al., 2011; Tesson et al., 2011) on monomeric TALE repeat array domains demonstrated strong preferences for the intended target base pair at each position in the binding site, and a study by Duchateau and coworkers using a cellular GFP reporter assay found that relatively few mismatches can be accommodated (Juillerat et al., 2014).

Several studies, in human cell lines (Mussolino et al., 2011), zebrafish (Dahlem et al., 2012), and rats (Tesson et al., 2011) have demonstrated TALEN-mediated off-target modification of multiple genomic sites that differ from the on-target site at two to six base pairs. The detection of these sites
is not thought to be a general problem of TALEN specificity, since for many applications a TALEN on-target site (up to 36 bp long) can be chosen to be at least seven mutations from any other site in the human genome. However, at least three studies have uncovered off-target sites modified in cells with more than seven mutations from the target site. In one study, Jaenisch and coworkers used DNA-binding SELEX results on TALE repeat domains in isolation to computationally predict potential genomic off-target sites of a fully active heterodimeric TALEN. Of the 19 predicted sites assayed, two off-target sites containing 9 or 10 mutations relative to the on-target site, were modified in cultured human cells (Hockemeyer et al., 2011). Tolar and coworkers used genome-wide selection with IDLVs (Gabriel et al., 2011; Osborn et al., 2013) to capture off-target double-strand break sites in cells, resulting in the identification of three off-target sites in the genome with up to 12 mutations from the target sequence.

Finally, Liu and colleagues applied the in vitro cleavage selection method described above to reveal 16 sites confirmed to be off-target sites in human cells with modification efficiencies ranging from 0.03% to 2.3% (Guilinger et al., 2014). The 16 off-target sites contained 8 to 12 mutations compared to the on-target site, demonstrating that TALENs can have appreciable off-target activities in human cells even at loci that are quite distant from the on-target sequence. Similar to the model developed to describe ZFN specificity, the in vitro cleavage results of Liu and coworkers suggested that reducing the cationic charge of the canonical 63-aa TALE C-terminal domain or the canonical N-terminal TALE domain could improve specificity by reducing nonspecific DNA-binding energy. Consistent with this hypothesis, the ability of off-target sites to survive the in vitro selection decreased as these cationic residues were mutated to neutral amino acids. Many of these charge-engineered TALENs demonstrated improved specificity across all positions in the target site. Specificity profiles generated using the in vitro selection method applied to charge-engineered TALENs indeed showed ~10 to ~100-fold improved specificity from assays of on-target and off-target activity both in vitro and in cells.

1.5. Insights and improvements from Cas9 specificity studies
In contrast to ZFNs and TALENs, RNA-guided Cas9 nucleases (referred to below as Cas9) do not require the design of separate DNA-binding domains for each new target site (Fig. 3.3C). Cas9 is a member of the CRISPR/Cas family of proteins that naturally defend bacterial genomes through
endonuclease activity against foreign DNA sequences. In contrast to ZFNs and TALENs, the target DNA specificity of Cas9 is programmed by hybridization of the target DNA to a Cas9-bound guide RNA sequence (sgRNA) (Jinek et al., 2012). Similar to TALENs, Cas9 target sequences are constrained at one end. All Cas9-targeted sequences require a sequence motif called a protospacer adjacent motif (PAM), the identity of which depends on the species of the Cas9 protein. For example, the most commonly used Cas9, from *S. pyogenes*, cleaves most efficiently target sequences containing an NGG PAM. Unlike ZFNs and TALENs, Cas9 target sites described to date consist of at most 20 base pairs, not including the PAM sequence.

Early studies of Cas9 specificity in nature by Siksyns, Severinov, Maraffini, and their respective coworkers (Cong et al., 2013; Jiang, Bikard, Cox, Zhang, & Marraffini, 2013; Jinek et al., 2012; Sapranauskas et al., 2011; Semenova et al., 2011) suggested that specific recognition of target DNA by Cas9 was limited to a 7–12 base pair subsequence adjacent to the PAM end of the target site. Further *in vitro* study using discrete off-target site testing by Doudna, Charpentier, and colleagues (Jinek et al., 2012) also supported the model. In this model of Cas9 specificity, mismatches were thought to be tolerated at the non-PAM end of the molecule. This model would suggest that Cas9 could not be used for specific genome modification, since a 12-base pair sequence plus two base pair PAM is not long enough to specify a unique sequence in the human genome. Several studies had shown, however, that Cas9 could be used for genome modification in several organisms without adverse effects; for example, Joung and coworkers reported that Cas9-mediated gene modification in zebrafish embryos exhibited a similar rate of off-target toxicity as ZFNs and TALENs (Hwang et al., 2013).

Four subsequent studies, two using discrete off-target testing in human cell culture by Joung (Fu et al., 2013), Zhang (Hsu et al., 2013), and their respective coworkers, one using a minimally biased selection in cells by Church and coworkers (Mali et al., 2013), and one using a minimally biased selection *in vitro* by Liu and coworkers (Pattanayak et al., 2013), investigated Cas9 specificity and showed that while Cas9 specificity is sufficient for at least some genome-editing applications, several off-target cleavage sites could be detected for most Cas9 target sites tested. While the magnitude of off-target activity varied in the four studies, Joung and coworkers observed that some off-target sites could be modified at similar frequencies to the on-target site.

All four studies showed that Cas9 specificity extended past the 7–12 base pair subsequence near the PAM, and that specificity is decreased at the end of the target site farthest from the PAM. The subsequence near the PAM, while
highly specified, tolerates certain single-base pair mismatches in an unpredictable fashion depending on the target site. These functional observations of cleavage specificity have been supported both by a molecular dynamics study of Cas9 by Doudna and colleagues (Sternberg, Redding, Jinek, Greene, & Doudna, 2014), as well as crystallographic models of Cas9 elucidated by Doudna, Nureki, and their respective colleagues (Jinek et al., 2014; Nishimasu et al., 2014). Genomic binding site profiling of inactive Cas9 by Adli, Sharp, Zhang, and their respective colleagues, in addition to confirming that the entire Cas9 target site is necessary for cleavage, suggests that Cas9 can bind many more sites in the genome than it actually cleaves (Kuscu et al., 2014; Wu et al., 2014).

Within the PAM itself, certain mismatches can also be tolerated. While the \textit{S. pyogenes} Cas9 specifies an NGG PAM, studies in bacteria by Marraffini and coworkers (Jiang et al., 2013) using a fully randomized PAM library demonstrated a tolerance of NAG and NNGGN PAM sequences. Further observations by Church, Liu, Zhang, and their respective coworkers in the specificity studies described above showed that an NAG PAM can also be recognized, and \textit{in vitro}, an NNG or an NGN PAM can be recognized with weak activity when the rest of the target sequence is fully complementary to the sgRNA (Pattanayak et al., 2013). The observation that an NAG PAM can be tolerated has also been supported by crystallographic studies by Jinek and coworkers (Anders, Niewoehner, Duerst, & Jinek, 2014). A more recent study on Cas9 specificity by Bao and colleagues (Lin et al., 2014) also suggests that Cas9 can tolerate single-base pair insertions or deletions in the target sequence relative to the sgRNA, though with reduced activity. In addition, several studies have established that specificity is dependent on Cas9 concentration (Fu et al., 2013; Hsu et al., 2013; Pattanayak et al., 2013) and guide RNA architecture (Fu, Sander, Reyon, Cascio, & Joung, 2014; Hsu et al., 2013; Pattanayak et al., 2013).

Given the significant off-target activity of Cas9 endonucleases, numerous groups have engineered Cas9 or guide RNA variants with enhanced specificity. Joung and coworkers improved the specificity of the Cas9:sgRNA complex by truncating the sgRNA to target less than the canonical 20-bp target sites (Fig. 3.4A) (Fu et al., 2014). By analogy to a study by Kim and colleagues on dimeric zinc-finger nickases (Kim et al., 2012), Church, Zhang, and their respective coworkers demonstrated that mutant Cas9 proteins that cleave only a single strand of dsDNA can be used to nick opposite strands of two nearby target sites, generating what is effectively a double-strand break with reduced off-target activity (Fig. 3.4B) (Ran et al., 2013; Cho et al., 2014; Mali et al., 2013).
Nickases even when bound to off-target loci as monomers retain their ability to nick DNA, which can result in a low level of undesired genome modification (Cho et al., 2014; Fu et al., 2014; Ran et al., 2013), as has previously been described for single zinc-finger nickases (Ramírez et al., 2012; Wang et al., 2012). Therefore, Liu, Joung, and their respective coworkers developed engineered Cas9 variants that are only able to cleave DNA when two monomers are adjacently bound to a target locus by fusing a FokI nuclease (red) fused to catalytically inactive Cas9 bind to separate sites within a target locus. Only adjacently bound FokI–dCas9 monomers can assemble a catalytically active FokI nuclease dimer, triggering dsDNA cleavage.

Figure 3.4 Engineered Cas9 components with improved DNA cleavage specificity. (A) A truncated guide RNA (tru-gRNA, right), contains 17–18 base pairs of complementarity to its DNA target site, rather than 20 base pairs in a canonical sgRNA (left). The base pairs in the sgRNA that are not present in the tru-gRNA are colored black. (B) Mutant Cas9 proteins that cleave only a single strand of dsDNA (nCas9) can be targeted to opposite strands of adjacent sites as pairs to cause double-strand breaks. (C) Monomers of FokI nuclease (red) fused to catalytically inactive Cas9 bind to separate sites within a target locus. Only adjacently bound FokI–dCas9 monomers can assemble a catalytically active FokI nuclease dimer, triggering dsDNA cleavage.
restriction endonuclease cleavage domain to a catalytically inactive Cas9 (dCas9) (Guilinger, Thompson, & Liu, 2014; Tsai et al., 2014), analogous to dimeric ZFNs and TALENs. In discrete off-target studies, the FokI–dCas9 fusions maintain substantial on-target DNA modification with a large reduction in off-target modification at known Cas9 off-target sites.

Collectively, studies that reveal in detail the DNA cleavage specificity of Cas9, together with the engineering of improved Cas9 variants, demonstrate the potential of Cas9 as an accessible and specific genome engineering tool.

2. METHODS

2.1. Overview of in vitro selection-based nuclease specificity profiling

The in vitro selection method developed by our group to profile the DNA cleavage specificity of a nuclease comprises three major steps: pre-selection library construction, in vitro selection, and high-throughput sequencing and analysis. Briefly, synthetic 5′-phosphorylated oligonucleotides are converted into concatemeric repeats of a library of potential off-target sites through intramolecular circularization followed by rolling-circle amplification. The resulting pre-selection libraries are then incubated in vitro with the appropriate nuclease, either in purified form or used directly from in vitro translation systems. Cleaved library members, which contain free 5′ phosphates, are captured by adapter ligation enabling their separation from uncleaved pre-selection library members, which do not contain 5′ phosphates. Cleaved post-selection library members are then amplified by PCR prior to high-throughput DNA sequencing.

2.2. Pre-selection library design

While it would be ideal to use a pre-selection library that consists of all possible off-target sequences of a given length (for example, an N_{22} library for Cas9, including PAM), the in vitro selection method has an upper limit of approximately 10^{12} sequences in the pre-selection library. Since an N_{22} library would contain 4^{22} (∼10^{13} sequences), a library biased in favor of sequences resembling the on-target recognition site is used instead. Library biasing is accomplished through the use of randomized nucleotide mixtures at all target-site base pairs during library construction. We and others (Argast et al., 1998; Doyon et al., 2006; Guilinger et al., 2014; Mali et al., 2013; Pattanayak et al., 2013, 2011) have had success using mixtures that contain 79% on-target base pair at each targeted position, with the remaining 21% of the mixture comprising of the three off-target base pairs. For Cas9, this
approach results in a pre-selection library that in theory contains at least ten copies of each potential off-target sequence containing eight or fewer mutations relative to the on-target site. For a 36-base pair TALEN on-target site, the pre-selection library provides at least 10-fold coverage of all sequences with six or fewer mutations. The partially randomized on-target site is also flanked by fully randomized base pairs on each side to test for patterns of specificity beyond the canonical target site.

The concentration of nuclease used to digest the pre-selection library should be enough to produce sufficient cleaved sequences for robust detection but not enough to completely digest highly cleaved sequences. The use of high nuclease concentrations will augment the detection of rare off-target cleavage events and could result in lower apparent nuclease specificity. Therefore, careful consideration of the nuclease concentrations used, or at least reporting the percentage of on-target sequences cleaved under the assay conditions, is required when studying and describing specificity.

2.3. **In vitro selection protocol**

2.3.1 **Before Day 1: Design and synthesize pre-selection library oligonucleotides**

For a selection using a guide RNA (CLTA4) targeted to the human clathrin gene (*CLTA*) (Pattanayak et al., 2013), the library oligonucleotide was ordered from Integrated DNA Technologies and was of the form:

```
/5Phos/TTG TGT NNN NC*C* NT*G* T*G*G* A*A*A* C*A*C*
T*A*C* A*T*C* T*G*C* NNN NAC CTG CCG AGT TGT GT
```

“/5Phos/” refers to a 5’ phosphate modification. The underlined sequence refers to the target site library, where each asterisk denotes a position that was ordered as a mixture of bases with 79% of the mixture corresponding to the base preceding the asterisk and 7% each corresponding to the other three bases. For Cas9, we found that this target site orientation (with the reverse complement of the PAM at the 5’ end of the oligonucleotide) yielded higher quality data than the reverse complement of this orientation. The italicized sequences denote a repeated six-base pair barcode that can be used to identify the target site used in the selection, if multiple selections are assayed at once. Other barcodes that we have used include AAC ACA, TCT TCT, and AGA GAA. Any barcodes can be used, although we recommend a minimum of two base pairs differing between each barcode. The sequence in bold denotes a constant region that remains the same for all selections. The constant sequence includes a **BspMI** restriction site that is used for pre-selection library preparation for high-throughput sequencing.
2.3.2 Day 1: Circularize library oligonucleotides

Dilute library oligonucleotides to 10 \( \mu M \) in 1 mM Tris, pH 8.0. In PCR tubes, add 1 \( \mu L \) of library oligo (10 pmol), 2 \( \mu L \) of 10 \( \times \) CircLigase II 10 \( \times \) Reaction Buffer, 1 \( \mu L \) of 50 mM MnCl\(_2\), 15 \( \mu L \) water, and 1 \( \mu L \) CircLigase II ssDNA Ligase (100 U) (Epicentre #CL9021K). Incubate 16 h at 60 \(^\circ\)C, followed by a 10 min inactivation step at 85 \(^\circ\)C.

2.3.3 Day 2: Confirm circularization of library oligonucleotides and perform rolling-circle amplification

On a 15% TBE-Urea polyacrylamide gel, load 2.5 pmol of uncircularized library oligo and 2.5 \( \mu L \) (1.25 pmol) of the CircLigase mixture without purification. Run for 75 min at 200 V. Stain the gel in 100 mL of 0.5 \( \times \) TBE containing 10 \( \mu L \) SYBR Gold (Invitrogen #11494) for 2 h. Rinse with water before imaging the gel. Under these conditions, the circularized oligonucleotides should migrate more slowly than the linear oligonucleotide control.

We use the Illustra TempliPhi Amplification Kit (GE Healthcare #25-6400-10) for the rolling-circle amplification reaction. Combine 5 \( \mu L \) (2.5 pmol) of each unpurified CircLigase reaction and 50 \( \mu L \) of TempliPhi sample buffer. Incubate 3 min at 95 \(^\circ\)C. Cool at 0.5 \(^\circ\)C/s to 4 \(^\circ\)C. Add 50 \( \mu L \) TempliPhi reaction buffer and 1 \( \mu L \) TempliPhi enzyme mix. Incubate 16 h at 30 \(^\circ\)C. Heat-inactivate for 10 min at 65 \(^\circ\)C. The rolling-circle amplification step can be halved or doubled in scale, with the final pre-selection library size determined by the amount of CircLigase reaction that is used.

2.3.4 Day 3: Quantify and digest pre-selection library

Since the pre-selection library is used in the selection without purification, a double-stranded DNA quantification reagent (Quant-it PicoGreen dsDNA Assay Kit, Invitrogen) is used to quantify the amplified double-stranded DNA. To quantify, add 1 \( \mu L \) rolling-circle amplified DNA or 10–200 ng of lambda DNA standard to 200 \( \mu L \) of 1 mM Tris, pH 8.0. Incubate 10 min at room temperature in the dark, before reading fluorescence in a plate reader (excitation wavelength \( \sim 480 \) nm, emission wavelength \( \sim 520 \) nm). Create a standard curve relating DNA concentration to fluorescence using, for example, pre-quantitated phage lambda DNA and calculate the concentration of the rolling-circle amplified pre-selection library.

To perform the in vitro selection, digest the pre-selection library with purified or in vitro translated site-specific endonuclease. For Cas9, 200 nM amplified pre-selection library was incubated with 100 nM Cas9 and
100 nM sgRNA or 1000 nM Cas9 and 1000 nM sgRNA in Cas9 cleavage buffer (20 mM HEPES, pH 7.5, 150 mM potassium chloride, 10 mM magnesium chloride, 0.1 mM EDTA, 0.5 mM dithiothreitol) for 10 min at 37 °C. Separately, the pre-selection library is also incubated with 2 U of BspMI restriction endonuclease (NEB) in NEBuffer 3 (100 mM NaCl, 50 mM Tris–HCl, 10 mM MgCl₂, 1 mM dithiothreitol, pH 7.9) for 1 h at 37 °C. Both nuclease-digested and restriction-digested libraries are purified with the QIAQuick PCR Purification Kit (Qiagen).

For site-specific nucleases that leave overhangs, such as ZFNs and TALENs, an additional step to convert the cut overhangs into blunt ends is performed before adapter ligation. In this step, 50 μL of purified, digested DNA is incubated with 3 μL of 10 mM dNTP mix (10 mM dATP, 10 mM dCTP, 10 mM dGTP, 10 mM dTTP) (NEB), 6 μL of 10 x NEBuffer 2, and 1 μL of 5 U/μL Klenow Fragment DNA Polymerase (NEB) for 30 min at room temperature. The blunt-ended mixture is purified with the QIAQuick PCR Purification Kit (Qiagen).

Once the cut ends have been made blunt, either by Cas9, or for TALENs by Klenow polymerase, sequencing adapters are ligated. For post-selection blunt libraries, adapter 1 (5’ AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TAC ACT CTG GAT) and adapter 2 (5’ TGT TAG ATC GGA AGA GCG TCG TGT AGG GAA AGA GTG TAG ATC TCG GTG GTG G) are used to incorporate sequences for Illumina sequencing. The reverse complementary sequences in italics can be varied to barcode multiple reaction conditions. The rest of the adapter sequences can also be appropriately varied to be used with other high-throughput sequencing platforms. In the ligation step, 10 pmol each of adapter 1 and adapter 2 are incubated with the blunt-ended post-selection library and 1000 U T4 DNA Ligase (NEB) in NEB T4 DNA Ligase Reaction Buffer (50 mM Tris–HCl, pH 7.5, 10 mM magnesium chloride, 1 mM ATP, 10 mM dithiothreitol) overnight at room temperature. For the restriction-digested pre-selection library, the ligation protocol is the same, with the exception of the use of lib adapter 1 (5’ GAC GGC ATA CGA GAT) and lib adapter 2 (5’ TTG TAC TCT GTA TGC CGT CTT CTG CTG G). Of note, the first four bases of lib adapter 2 (in italics) must match the first four bases of the library oligonucleotide barcode (see Section 2.3.1), since BspMI digestion will leave an overhang that is specific to the barcode used. Therefore, if multiple target sites are tested in the same selection run, multiple lib adapter 2’s must be used.
2.3.5 Day 4: PCR of pre- and post-selection libraries

The PCR amplification step prior to high-throughput sequencing must be well controlled to minimize the potential effects of PCR bias on the final sequencing results. Prior to PCR, the adapter ligation mixtures from Day 3 are purified with the QIAQuick PCR Purification Kit (Qiagen) and eluted with 50 μL 1 mM Tris, pH 8.0. We use Phusion Hot Start Flex DNA Polymerase (NEB) in Buffer HF with an annealing temperature of 60 °C and an extension temperature of 72 °C for 1 min per cycle. For the nuclease-digested post-selection PCR, primers sel PCR (5′ CAA GCA GAA GAC GGC ATA CGA GAT ACA CAA CTC GGC AGG T) and PE2 short (5′ AAT GAT ACG GCG ACC ACC GA) are used. For the restriction-digested pre-selection library PCR, use the same PCR cycling conditions with primers lib fwd PCR (5′ CAA GCA GAA GAC GGC ATA CGA GAT) and lib seq PCR (5′ AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TNN NNA CCT ACC TGC CGA G) T). Four Ns are included in the lib seq PCR to provide a randomized initiation sequence to maintain compatibility with Illumina sequencing requirements. Of note, if multiple target sites are used in the same selection run, multiple sel PCR primers must be used, with the four base sequence in sel PCR and the six base sequence in lib seq PCR listed in italics should be modified to maintain complementarity to the original library oligonucleotide backbone (see Section 2.3.1).

Before PCR of the full volume of post-selection library, we suggest performing a test PCR or test qPCR with 1 μL of purified post-selection library under the PCR conditions listed above to determine the number of cycles required to reach saturation. If doing a test PCR, remove aliquots every 5 cycles for 35 cycles and visualize the reaction products. Determine the point at which the PCR amplification saturates and subtract an appropriate number of cycles when the PCR is scaled up. For example, if scaling from 1 to 32 μL of purified post-selection library, subtract five cycles (2^5 = 32).

After PCR, there may be a ladder of products, corresponding to amplified post-selection library members that contain 0.5, 1.5, 2.5, etc., repeats of a given library member (Fig. 3.5). The variation in PCR product size results from the concatemeric nature of the pre-selection library. During PCR, the sel PCR primer can also anneal to one of multiple repeats, also leading a distribution of PCR product sizes. To standardize the analysis, only those PCR products that contain 1.5 repeats are analyzed. Therefore, before high-throughput sequencing, a final gel purification step is used to enrich the
amplified post-selection library members that contain exactly 1.5 repeats and to remove any remaining free adapters and primers.

### 2.3.6 Day 5: High-throughput sequencing and analysis

The gel-purified post-selection and pre-selection libraries can be quantified using the KAPA Library Quantification Kit–Illumina (KAPA Biosystems) before high-throughput sequencing. We used single-read sequencing on an Illumina HiSeq or Illumina MiSeq, though the selection should be compatible with any high-throughput sequencing platform as long as the adapter
sequences and PCR primers are modified appropriately. For Cas9, a minimum of 66 bases must be sequenced to capture the entire library member. If using a selection condition barcode (for example, AACA below), we recommend spiking in a PhiX library control at 25% with the sequencing run to provide appropriate initial base-calling diversity if using Illumina sequencing.

The sequencing output can be binned using a simple scripting language, such as C++ or Python. The components of the sequencing read are illustrated below:

\[
\text{AACAcatgggtcgACACAAACACAACTCGGCAGGTACTTGCGAGATGTAGTCTTTCCACATGGGTCGACACAAACACAACTCGG}
\]

\[
\text{CAGGTATCTCGTATGCCAACA}
\]

\[
\text{AACA} \text{is the four-base pair barcode for selection conditions. catgggtcg is the cut “half” of the library target site. ACACAAACACAA is the Cas9 target barcode. CTCGGCAGGT is the constant sequence (the reverse complement of the bold sequence in the Library Design section). ACTTGAGTATGTCTTTCCACATGGGTCGACACAAACACAACTCGGCAGGTATCTCGTATGCCAACA is the full sequence of the post-selection library member. This sequence can be recognized and cut in the selection. The nonunderlined portion of the sequence consists of the eight random base pairs (four on each side) that flanked the target site library. Once sequences are binned, standard analyses can be performed on the set of target sites (bold and underlined). For example, specificity profiles can be represented as heat maps of specificity scores calculated as the enrichment level of each possible base pair at every position in the post-selection sequences relative to the pre-selection sequences, normalized to the maximum possible enrichment of that base pair (Fig. 3.6).}

2.4. Confirmation of in vitro-identified genomic off-target sites

To identify genomic off-target sites, the set of target sites identified in vitro by selection and high-throughput sequencing can be searched for sequences that appear in the human genome. In addition to this simple comparison, a machine-learning algorithm can be trained on the in vitro dataset to assist the identification of potential genomic off-target sequences (Sander et al., 2013). The tested site-specific nuclease is then expressed in cultured human cells, along with a parallel experiment with a control, inactive form of the same site-specific nuclease. Genomic DNA is isolated, followed by PCR with primers specific to each potential off-target site. A primer design tool,
such as NCBI Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/), can be useful in the design of primers that lead to specific amplification of the target site of interest. Portions of the high-throughput adapter sequences can be incorporated into the primers (for example, for Illumina, 5’ ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT at the 5’ end of one primer and 5’ GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT on the other) for the initial PCR. When assaying multiple off-target sites at once, PCRs can be pooled in equimolar ratios, purified, and then reamplified using primers PE1-barcode (5’ CAA GCA GAA GAC GGC ATA CGA GAT ATA TCA GTG TGA CTG GAG TTC AGA CGT GTG CT) and PE2-barcode (5’ AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA TTA CTC GAC ACT CTT TCC CTA CAC GAC). The number of cycles of the re-amplification PCR should be minimized to avoid introducing significant PCR bias. The italicized bases in PE1-barcode and PE2-barcode correspond to barcodes that can be used in Illumina sequencing. Different barcodes should be used for PCR products derived from active-nuclease-treated DNA compared to inactive-nuclease-treated DNA.

Following high-throughput sequencing, nuclease-modified off-target sequences can be identified through sequence alignment or through computational methods. One algorithm for identifying modified-sequences involves searching for the 20 base pairs flanking each off-target site for each high-throughput sequencing read. For example:

Figure 3.6 An in vitro selection-derived specificity profile. The heat map shows the specificity profile resulting from a selection performed on Cas9:sgRNA targeting the human CLTA gene. Specificity scores of 1.0 (dark blue) and −1.0 (dark red) corresponds to 100% enrichment for and against, respectively, a particular base pair at a particular position. Black boxes denote the intended target nucleotides.
5’ CAATCTCCGCATGCGCTCACTCTCCCTCAAG CAGGCCCGCTGGTGCACTGAAGAGGCCACCCCTGTGAAACAC TACATCTGGAATATCTTAATCCTAATCCTAATCAGTGAAGCTCTTCAC AGTCATTGGAATATTTATGTGAGTTCTTTTTGGACC

The flanking sequences (underlined) can be used to identify the off-target site being assayed (in bold). In the reference genome sequence, the sequence between the underlined flanking regions is 5’ CCCTGTGAAACACTA CATCTGC. In this example, the sequence between the underlined flanking regions is 5’ CCCTG-TGAAACACTACATCTGC, where the dash indicates a one base pair deletion. For each potential off-target site tested, the fraction of sequences with insertions and deletions can be calculated and compared between active-nuclease and inactive-nuclease experiments. For target sites with high modification efficiencies, it may be necessary to use flanking sequences (the underlined sequences above) that are more distal to the target site, in case NHEJ leads to deletion of a region that is larger than the off-target site (the bold sequence).

3. CONCLUSION

Genome engineering in the last few years has become more facile through the use of programmable site-specific nucleases such as TALENs and Cas9, which can be designed to target nearly any DNA sequence. As the use of ZFNs, TALENs, and Cas9 in research and clinical settings continues to grow, efforts to reveal in depth the DNA cleavage specificity of programmable nucleases will become increasingly important. Efforts to characterize programmable nuclease specificity have ranged from discrete target-site assays to in vitro selections to genome-wide selections, all of which have been applied recently to study TALEN and Cas9 specificity. The findings from these methods will continue to deepen our understanding of the basis of the DNA cleavage specificity of these important proteins, inform the development of programmable nucleases with improved specificity, and perhaps eventually enable the broad application of these or other programmable nucleases to treat human genetic diseases.

ACKNOWLEDGMENTS

V. P., J. P. G., and D. R. L. were supported by DARPA HR.0011-11-2-0003, DARPA N66001-12-C-4207, NIH/NIGMS R01 GM095501 (D. R. L.), and the Howard Hughes Medical Institute. D. R. L. was supported as a HHMI Investigator. V. P. was supported by NIGMS T32GM007753. We are grateful to J. Keith Joung for helpful comments.
REFERENCES


