Profiling and Improving the Specificity of Site-Specific Nucleases

Profiling and Improving the Specificity of Site-Specific Nucleases

Abstract

Programmable site-specific endonucleases are useful tools for genome editing and may lead to novel therapeutics to treat genetic diseases. TALENs can be designed to cleave chosen DNA sequences. To better understand TALEN specificity and engineer TALENs with improved specificity, we profiled 30 unique TALENs with varying target sites, array length, and domain sequences for their ability to cleave any of 10^{12} potential off-target DNA sequences using *in vitro* selection and high-throughput sequencing. Computational analysis of the selection results predicted 76 off-target substrates in the human genome, 16 of which were accessible and modified by TALENs in human cells. The results collectively suggest that (*i*) TALE repeats bind DNA relatively independently; (*ii*) longer TALENs are more tolerant of mismatches, yet are more specific in a genomic context; and (*iii*) excessive DNA-binding energy can lead to reduced TALEN specificity in cells. We engineered a TALEN variant, Q3, that exhibits equal on-target cleavage activity but 10-fold lower average off-target activity in human cells. Our results demonstrate that identifying and mutating residues that contribute to non-specific DNA-binding can yield genome engineering agents with improved DNA specificities.

Cas9 cleaves double-stranded DNA in cells at a sequence targeted by a short guide RNA (gRNA). We used *in vitro* selection to determine the abilities of eight Cas9:guide RNA complexes to cleave off-target DNA sequences. The selection results predicted five off-target substrates in the human genome that were confirmed to undergo genome cleavage in cells treated

with Cas9:guide RNA complexes. To improve the specificity of Cas9:guided RNA complexes we describe the development of a *Fok*I nuclease fusion to a catalytically dead Cas9 that requires simultaneous DNA binding and association of two *Fok*I-dCas9 monomers to cleave DNA. Off-target DNA cleavage of the engineered *Fok*I-dCas9 is further reduced by the requirement that only sites flanked by two gRNAs ~15 or 25 base pairs apart are cleaved. In human cells, fCas9 modified target and off-target DNA sites with comparable efficiency to wild-type Cas9 nucleases but with > 140-fold higher specificity.

Profiling and Improving the Specificity of Site-Specific Nucleases

Table of Contents

Introduction

I.1 Introduction to site-specific nucleases for genome modification

Site-specific nucleases that can be designed to target any sequence of DNA in a cell are powerful research tools with significant therapeutic implications. In cells, site-specific nucleases bind a target DNA sequence. A site-specific nucleases will then cleave both strands of DNA either within or nearby the targeted DNA resulting in double strand breaks.[1,2] Double-strand breaks can be repaired through the cellular process of non-homologous end joining (NHEJ). Erroneous repair by NHEJ can cause stochastic DNA insertions or deletions that will likely shift the coding frame of a gene effectively resulting in targeted gene knockout.[3–6] Alternatively, the double-strand break can efficiently recombine with exogenous DNA template homologous to the target site through homology-directed repair (HDR) for precise alterations to a target genomic sequence.[7–10] Indeed, site-specific nucleases have already been used to modify targeted sequences in the genomes of a variety of organisms[11–17] and human cell lines.[18–20,8,4–6]

In addition to engineering the genomes of cells or organisms for direct biological interrogation, genetic screens have recently been performed with site-specific nucleases to uncover genetic factors underlying specific cellular processes in an unbiased manner[21,22]. Site-specific nucleases could also serve as the basis of a new generation of human therapeutics and are currently in clinical trials as a potential cure for HIV.[18] In this therapy,[23] the *CCR5* gene is knocked out with site-specific nuclease in a patients own T-cells, resulting in T-cells that lack the co-receptor encoded by *CCR5* that is required for HIV infection. Thus, site-specific nucleases represent an effective tool for clinically relevant genetic manipulation.

In order to effectively modify a given DNA sequence, an optimal site-specific nuclease will be easily programmed to target many different DNA sequences, highly efficient at cleaving the target DNA and specifically cleave only the target DNA. First, an ideal site-specific nuclease can be easily designed to target any DNA sequence with as little sequence constraints as possible. Maximizing the targetable sequence space increases the number of potential genomic loci that can be modified for use in both basic science and therapeutics. Second, a site-specific nuclease should be able to efficiently and effectively cleave the target DNA sequence to maximize the amount of genetic modification. Third, a site-specific nuclease should be as specific as possible cleaving only the target DNA sequence and not cleaving other sequences (**Figure I.1**).

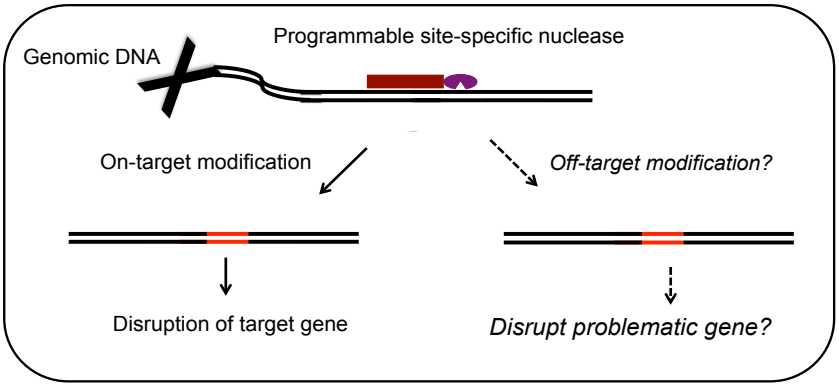


Figure I.1. Site-specific genome modification. A programmable site-specific nuclease is shown that cleaves genomic DNA resulting in modification. Modification of the target site and a potential off-target site is shown.

The specificity of site-specific nucleases is imperative since the DNA-binding specificity of endonucleases has been inversely correlated with their cellular toxicity.[24] The specificity of a site-specific nuclease is also critical because cleavage at unintended, off-target sites could confound any biological inquiry[21] or worse introduce genetic malefactors in human therapies. Previous reports of gene therapies relying on viruses to deliver genes that integrate randomly into the genome highlight the need for more specific and precise genome modification. Retroviral delivery and resulting random integration of a gene to express interleukin receptor gamma chain in children with severe combined immunodeficiency-X1 (SCID-X1) effectively cured nine out of ten children treated.[25] However, two of the children developed leukemia likely from retroviral integration near the promoter of the *LMO2* oncogene.[26] Given the potential for integration into oncogenic loci, it is imperative that genes integrate at specific sequences known to be safe for integration. Apart from directing integration of exogenous genes into defined genomic loci, it is also important that site-specific nucleases do not modify tumor suppressor genes potentiating malignant cells.

Identifying genuine off-target sites cleaved by site-specific nucleases is challenging given the shear size of the human genome. There will be at least ~$3x10^9$ potential "off-target" sites for every target site in the human genome, although the vast majority of these "off-target" sites will be very different from the target site, a fraction of these off-target sites will be similar to the on-target site to some degree. To identify off-target modification caused by a site-specific nuclease, ideally one would high-throughput sequence genomic DNA from cells or organisms treated with the site-specific nuclease searching for any off-target modification across the entire genome in an

unbiased study. However, to sequence enough genomes to detect rare off-target modification events on the order of 1/10,000, it would cost a prohibitive amount of well over $300,000.[27] Unbiased high-throughput sequencing of the exome of human cells expanded from a single cell treated with a site-specific nuclease revealed that at least some site-specific nucleases do not cause widespread or highly abundant off-target cleavage.[28,29] Another method to interrogate the specificity of site-specific nucleases in cells uses integrase-deficient lentiviral vectors (IDLVs),[30,31] which can integrate at double-strand breaks caused by site-specific nucleases in cells. While IDLV can be used to identify off-target cleavage in cells, it could be complicated by cellular factors such as DNA accessibility, which varies from site to site and between cell types,[32] or DNA repair and integration pathways after cleavage that could obscure the determination of nuclease specificity. IDLV may also suffer from high-background integration at double-strand breaks that arise naturally, independent of nuclease activity. Purely cellular studies are also inherently limited to the stochastic handful of off-target sites in a given genome that are similar to the target sequence.

Studies measuring the nuclease activity of a handful of closely related off-target substrates can yield useful information but cannot cover the number of potential off-target sites in genomes, which for some site-specific nucleases number more than 10,000.[33,34] In order to evaluate the ability of site-specific nucleases to cleave a very large number of off-target sites required to estimate, predict and reliably identify off-target cleavage in cells, a broad and in-depth study of specificity is necessary. Sequential enrichment of ligands by exponential (SELEX) is a method used to identify numerous off-target sequences resulting in a more comprehensive specificity profile.[4,8] Off-target DNAs are uncovered from interrogating an immense library potential off-target DNAs (~10^{14}) for those members that bind to an immobilized DNA-binding domain of the site-specific nuclease. Alternatively, a high-throughput study of DNA-binding domains fused to a transcriptional activator domain (in place of the nuclease domain) can also describe specificity of the monomeric DNA-binding domains of site-specific nucleases.[35] While both the above activator and SELEX experiments can describe the DNA-binding specificities of the monomeric DNA-binding domains of some site-specific nucleases, the DNA cleavage specificities of active, dimeric nucleases can differ from the specificities of their component monomeric DNA-binding domains.[36] Thus, the Liu lab has previously developed an *in vitro* selection to interrogate site-specific nucleases for their abilities

to cleave 10^{12} potential off-target DNA substrates.[36] While the selection is preformed *in vitro* and not in cells, it does report on the specificity of active nucleases on a large number (~10^{12}) of potential off-target sites and can be used to identify cellular off-target sites.[36] Chapters 1 and 2 apply this *in vitro* selection to profile the specificity of two different types of designer site-specific nucleases, transcription activator-like effector nucleases (TALENs) and RNA-guided Cas9 nucleases.

There are three major types of programmable site-specific nucleases: zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and RNA-guided Cas9 nucleases (Cas9:gRNA). ZFNs are fusions of the non-specific *Fok*I restriction endonuclease cleavage domain with a zinc finger DNA-binding domain. Zinc finger DNA-binding domains consist of three to four individual zinc fingers in tandem repeats. Since each individual zinc finger specifically recognizing three base pairs,[37] a zinc finger DNA-binding domain in total recognizes 9 bp or 12 bp. ZFNs can be engineered to be active only as heterodimers through the use of obligate heterodimeric *Fok*I variants.[38,39] In this configuration, two distinct ZFN monomers are each designed to bind one target half-site resulting in cleavage within the DNA spacer sequence between the two half-sites. While sophisticated methods[32,40] have been developed to generate zinc finger DNA-binding domains, the ease of designer zinc finger construction and the targetable sequence space of ZFNs remains limited. More troubling for the use of ZFNs as genome editing agents, reports demonstrate potential widespread off-target activity of ZFNs in cells.[30,33,36]

TALENs are another type of designer site-specific nuclease which are fusions of the *Fok*I restriction endonuclease cleavage domain with a DNA-binding TALE repeat array (**Figure I.2**).[41]

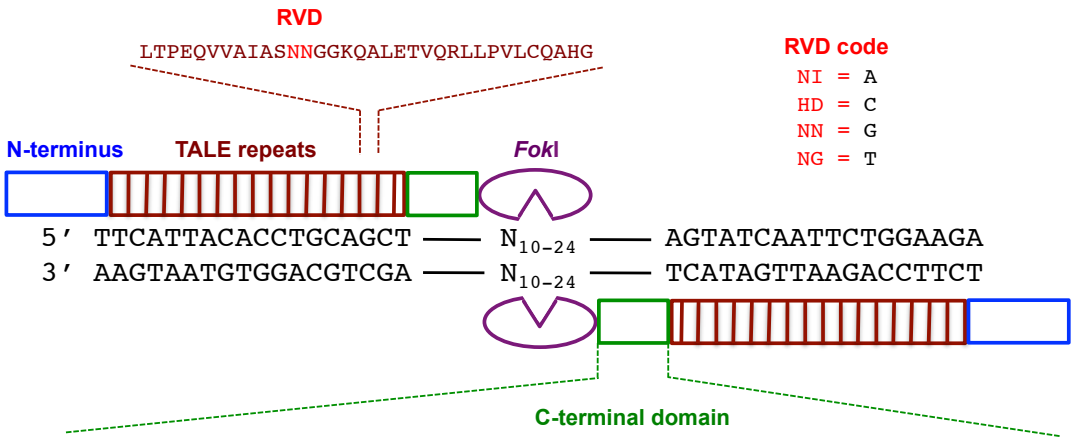


Figure I.2. TALEN architecture. A TALEN monomer contains an N-terminal domain (blue) followed by an array of TALE repeats (brown), a C-terminal domain (green), and a *Fok*I nuclease cleavage domain (purple). The 12th and 13th amino acids (the RVD, red) of each TALE repeat recognize a specific DNA base pair. Two different TALENs bind their corresponding half-sites, allowing *Fok*I dimerization and DNA cleavage.

These arrays consist of multiple 34-amino acid TALE repeat sequences, each of which uses a repeat-variable di-residue (RVD), the amino acids at positions 12 and 13, to recognize a single DNA nucleotide.[42,43] Examples of RVDs that recognize each of the four DNA base pairs are known, enabling arrays of TALE repeats to be constructed that can bind virtually any DNA sequence. Like ZFNs, TALENs can be engineered to be active only as heterodimers through the use of obligate heterodimeric *Fok*I variants. In this configuration, two distinct TALEN monomers are each designed to bind one target half-site resulting in cleavage within the DNA spacer sequence between the two half-sites.

Numerous studies have reported on the specificities of individual TALE repeats.[44–48] The specificity of individual repeats determines the specificity of the array of repeats, but it is not clear exactly how repeats may interact with each other or if there is interedepence between the specific binding of TALE repeats. Chapter 1 demonstrates that individual TALE repeats bind their respective DNA base pairs relatively independently. Other studies[8,31,49,50] and Chapter 1 demonstrate that TALENs can induce off-target modification in cells. Chapter 1 also broadly profiles the specificity of TALENs using the specificity profiles to predict and estimate off-target cleavage in cells. Chapter 1 also discusses a model of excess DNA-binding energy derived from the specificity profiling of TALENs, which is then used to engineer TALEN variants with improved specificity in cells.

RNA-guided Cas9 nucleases are another type of site-specific nucleases with powerful genome-modifying capability. Cas9 protein in complex with a guide RNA uses simple RNA:DNA hybridization to direct nuclease activity to a target DNA sequence (**Figure I.3**).[2]

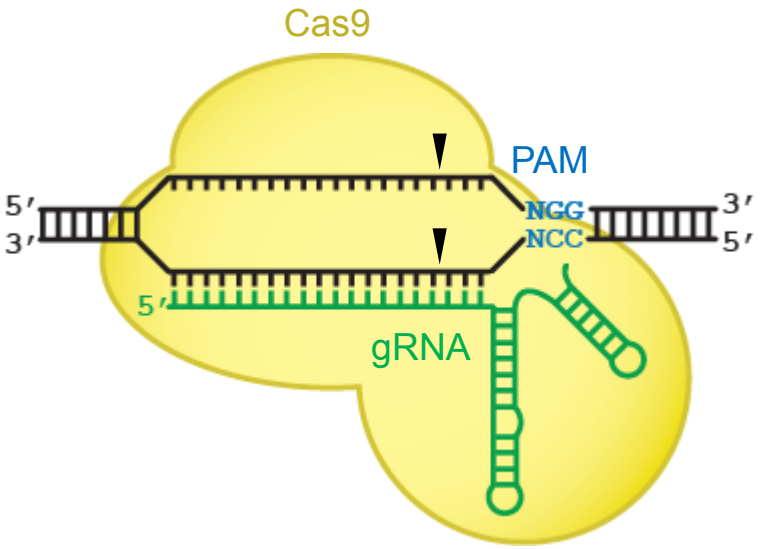


Figure I.3. RNA-guided Cas9 architecture. Cas9 protein (yellow) binds to target DNA in complex with a guide RNA (gRNA, green). The *S. pyogenes* Cas9 protein recognizes the PAM sequence NGG (blue), initiating unwinding of dsDNA and gRNA:DNA base pairing. Black triangles indicate the cleavage points three bases from the PAM on both top and bottom strands.

Because the guide RNA (gRNA) is short, ~100 bases, distinct gRNAs each targeting a different DNA site can easily be designed and constructed. Thus, Cas9:gRNA genome editing agents are especially easy to generate compared to generating much larger sequences coding for ZFNs or TALENs proteins. Chapter 2 and other studies[29,51–53] demonstrate that Cas9:gRNA is capable of significant off-target activity. While a dimeric, Cas9 nickase strategy can drastically improve specificity,[15,29,35] Chapter 3 reports an alternative strategy improving the specificity of genome modification by fusing a dimeric *Fok*I domain to catalytically inactive Cas9. Future avenues of research into the specificity of site-specific nucleases should focus on a comparison between TALENs and the more specific dimeric RNA-guided Cas9 nuclease strategies.

I.2 References cited in introduction

1. Vanamee, É. S., Santagata, S. & Aggarwal, A. K. FokI requires two specific DNA sites for cleavage. *J. Mol. Biol.* **309,** 69–78 (2001).
2. Jinek, M. *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337,** 816–821 (2012).
3. Rouet, P., Smih, F. & Jasin, M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol. Cell. Biol.* **14,** 8096–8106 (1994).
4. Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol* **29,** 143–8 (2011).
5. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2,** e00471–e00471 (2013).
6. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science* **339,** 823–826 (2013).
7. Choulika, A., Perrin, A., Dujon, B. & Nicolas, J. F. Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. *Mol. Cell. Biol.* **15,** 1968–1973 (1995).
8. Hockemeyer, D. *et al.* Genetic engineering of human pluripotent cells using TALE nucleases. *Nat Biotechnol* **29,** 731–4 (2011).
9. Bedell, V. M. *et al.* In vivo genome editing using a high-efficiency TALEN system. *Nature* **491,** 114–8 (2012).
10. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339,** 819–823 (2013).
11. Doyon, Y. *et al.* Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nat. Biotechnol.* **26,** 702–708 (2008).
12. Meng, X., Noyes, M. B., Zhu, L. J., Lawson, N. D. & Wolfe, S. A. Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nat. Biotechnol.* **26,** 695–701 (2008).
13. Tesson, L. *et al.* Knockout rats generated by embryo microinjection of TALENs. *Nat Biotechnol* **29,** 695–6 (2011).
14. Friedland, A. E. *et al.* Heritable genome editing in C. elegans via a CRISPR-Cas9 system. *Nat. Methods* **10,** 741–743 (2013).
15. Ran, F. A. *et al.* Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* **154,** 1380–1389 (2013).
16. Yang, H. *et al.* One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154,** 1370–1379 (2013).
17. Niu, Y. *et al.* Generation of Gene-Modified Cynomolgus Monkey via Cas9/RNA-Mediated Gene Targeting in One-Cell Embryos. *Cell* **156,** 836–843 (2014).
18. Perez, E. E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat. Biotechnol.* **26,** 808–816 (2008).
19. Zou, J. *et al.* Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells. *Cell Stem Cell* **5,** 97–110 (2009).
20. Hockemeyer, D. *et al.* Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nat. Biotechnol.* **27,** 851–857 (2009).

21. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343,** 80–84 (2014).
22. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343,** 84–87 (2014).
23. NCT00842634, NCT01044654, NCT01252641.
24. Cornu, T. I. *et al.* DNA-binding specificity is a major determinant of the activity and toxicity of zinc-finger nucleases. *Mol. Ther. J. Am. Soc. Gene Ther.* **16,** 352–358 (2008).
25. Cavazzana-Calvo, M. *et al.* Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science* **288,** 669–672 (2000).
26. Hacein-Bey-Abina, S. *et al.* LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302,** 415–419 (2003).
27. Sheridan, C. Illumina claims $1,000 genome win. *Nat. Biotechnol.* **32,** 115 (2014).
28. Ding, Q. *et al.* A TALEN genome-editing system for generating human stem cell-based disease models. *Cell Stem Cell* **12,** 238–51 (2013).
29. Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24,** 132–141 (2013).
30. Gabriel, R. *et al.* An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nat Biotechnol* **29,** 816–23 (2011).
31. Osborn, M. J. *et al.* TALEN-based gene correction for epidermolysis bullosa. *Mol Ther* **21,** 1151–9 (2013).
32. Maeder, M. L. *et al.* Rapid 'open-source' engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol Cell* **31,** 294–301 (2008).
33. Sander, J. D. *et al.* In silico abstraction of zinc finger nuclease cleavage profiles reveals an expanded landscape of off-target sites. *Nucleic Acids Res.* **41,** e181–e181 (2013).
34. Guilinger, J. P. *et al.* Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. *Nat. Methods* (2014). doi:10.1038/nmeth.2845
35. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31,** 833–838 (2013).
36. Pattanayak, V., Ramirez, C. L., Joung, J. K. & Liu, D. R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nat. Methods* **8,** 765–770 (2011).
37. Miller, J., McLachlan, A. D. & Klug, A. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J.* **4,** 1609–1614 (1985).
38. Doyon, Y. *et al.* Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *Nat Methods* **8,** 74–9 (2011).
39. Cade, L. *et al.* Highly efficient generation of heritable zebrafish gene mutations using homo- and heterodimeric TALENs. *Nucleic Acids Res* **40,** 8001–10 (2012).
40. Sander, J. D. *et al.* Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nat. Methods* **8,** 67–69 (2011).
41. Christian, M. *et al.* Targeting DNA Double-Strand Breaks with TAL Effector Nucleases. *Genetics* **186,** 757–761 (2010).
42. Moscou, M. J. & Bogdanove, A. J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326,** 1501 (2009).
43. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326,** 1509–12 (2009).

44. Cong, L., Zhou, R., Kuo, Y. C., Cunniff, M. & Zhang, F. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nat Commun* **3,** 968 (2012).
45. Streubel, J., Blucher, C., Landgraf, A. & Boch, J. TAL effector RVD specificities and efficiencies. *Nat Biotechnol* **30,** 593–5 (2012).
46. Christian, M. L. *et al.* Targeting G with TAL effectors: a comparison of activities of TALENs constructed with NN and NK repeat variable di-residues. *PLoS One* **7,** e45383 (2012).
47. Valton, J. *et al.* Overcoming Transcription Activator-like Effector (TALE) DNA Binding Domain Sensitivity to Cytosine Methylation. *J. Biol. Chem.* **287,** 38427–38432 (2012).
48. Lamb, B. M., Mercer, A. C. & Barbas, C. F. Directed evolution of the TALE N-terminal domain for recognition of all 5' bases. *Nucleic Acids Res.* **41,** 9779–9785 (2013).
49. Dahlem, T. J. *et al.* Simple methods for generating and detecting locus-specific mutations induced with TALENs in the zebrafish genome. *PLoS Genet* **8,** e1002861 (2012).
50. Mussolino, C. *et al.* A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res* **39,** 9283–93 (2011).
51. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31,** 827–832 (2013).
52. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31,** 822–826 (2013).
53. Cradick, T. J., Fine, E. J., Antico, C. J. & Bao, G. CRISPR/Cas9 systems targeting -globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.* **41,** 9584–9592 (2013).

Chapter 1:
Broad Specificity Profiling of TALENs Results in Engineered Nucleases with Improved DNA Cleavage Specificity

John P. Guilinger, Vikram Pattanayak, Deepak Reyon, Shengdar Q. Tsai, Jeffry D. Sander, J. Keith Joung, and David R. Liu

Deepak Reyon, in Keith Joung's laboratory, cloned the initial TALEN expression constructs and designed the primers for amplifying genomic off-target sites. Shengdar Q. Tsai, in Keith Joung's laboratory, expressed TALENs in U2OS-GFP cells and isolated genomic DNA. Jeffry D. Sander, in Keith Joung's laboratory, predicted potential genomic off-target sites using a computational classifier trained on selection data. Vikram Pattanayak, in David R. Liu's laboratory, designed the selection. I preformed all other experiments and analyzed all of the data presented in this chapter.

Text in this chapter appeared in *Nature Methods*, 2014, Published Online (doi:10.1038/nmeth.2845)

1.1 Introduction to TALEN specificity

The ability to engineer site-specific changes in genomes is a powerful research capability with significant therapeutic implications. Transcription activator-like effector nucleases (TALENs) are fusions of the *Fok*I restriction endonuclease cleavage domain with a DNA-binding TALE repeat array (**Figure 1.1**).

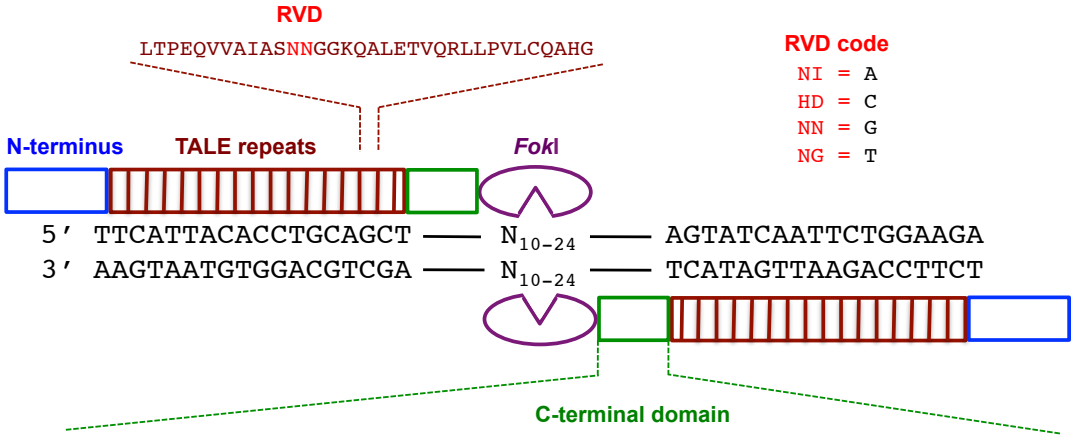


Figure 1.1. TALEN architecture. A TALEN monomer contains an N-terminal domain (blue) followed by an array of TALE repeats (brown), a C-terminal domain (green), and a *Fok*I nuclease cleavage domain (purple). The 12th and 13th amino acids (the RVD, red) of each TALE repeat recognize a specific DNA base pair. Two different TALENs bind their corresponding half-sites, allowing *Fok*I dimerization and DNA cleavage.

These arrays consist of multiple 34-amino acid TALE repeat sequences, each of which uses a repeat-variable di-residue (RVD), the amino acids at positions 12 and 13, to recognize a single DNA nucleotide.[1, 2] Examples of RVDs that recognize each of the four DNA base pairs are known, enabling arrays of TALE repeats to be constructed that can bind virtually any DNA sequence. TALENs can be engineered to be active only as heterodimers through the use of obligate heterodimeric *Fok*I variants.[3, 4] In this configuration, two distinct TALEN monomers are each designed to bind one target half-site resulting in cleavage within the DNA spacer sequence between the two half-sites. In cells, TALEN-induced double-strand breaks can result in targeted gene knockout through non-homologous end joining (NHEJ)[5] or precise targeted genomic sequence alteration through homology-directed repair (HDR) using an exogenous DNA template.[6, 7] TALENs have been successfully used to manipulate genomes in a variety of organisms[6, 8-11] and cell lines.[5, 7, 12, 13]

Although TALENs are sufficiently specific to show activity against their intended target sites without causing widespread and highly abundant genomic off-target modification,[4, 14-16]

TALEN-mediated DNA cleavage at off-target sites can result in unintended mutations at genomic loci. While recent studies identify closely related off-target substrates containing two or fewer mismatches in zebrafish[17] and in human cell lines,[13] more distant off-target substrates are of particular interest since one would expect a typical 36-bp target site to be approximately eight or more mutations away from any sequence in the human genome. Two distant genomic off-target sites were identified from 19 potential off-target sites predicted using SELEX,[7] an *in vitro* method to identify binding sites of DNA-binding domains in isolation. Only a single heterodimeric off-target site was identified through the use of integrase-deficient lentiviral vectors (IDLVs)[18, 19] to capture off-target double-strand break sites in cells. The limited number of off-target TALEN sites previously characterized in studies performed to date suggest that further research is needed both to better understand the extent of TALEN-induced genomic off-target mutations and to improve TALEN specificity to minimize these unwanted effects.

The underlying principles that determine the specificities of TALEN proteins remain poorly characterized. While SELEX experiments and a high-throughput study of TALE activator specificity have described the DNA-binding specificities of monomeric TALE proteins[5, 7, 9] and a single TALE activator,[20] respectively, the DNA cleavage specificities of active, dimeric nucleases can differ from the specificities of their component monomeric DNA-binding domains.[21] For example, zinc finger nucleases (ZFNs), another type of engineered dimeric nuclease, demonstrate compensation effects between monomers.[21] Cellular methods to study off-target genomic modification such as whole-genome sequencing or IDLV capture could be complicated by cellular factors such as DNA accessibility, which varies from site to site and between cell types,[22] or DNA repair and integration pathways after cleavage that could obscure the determination of intrinsic TALEN protein specificity. Purely cellular studies are also inherently limited to the stochastic handful of off-target sites in a given genome that are similar to the target sequence and thus are unable to evaluate the ability of TALENs to cleave a very large number of off-target sites necessary for a broad and in-depth study of TALEN specificity.

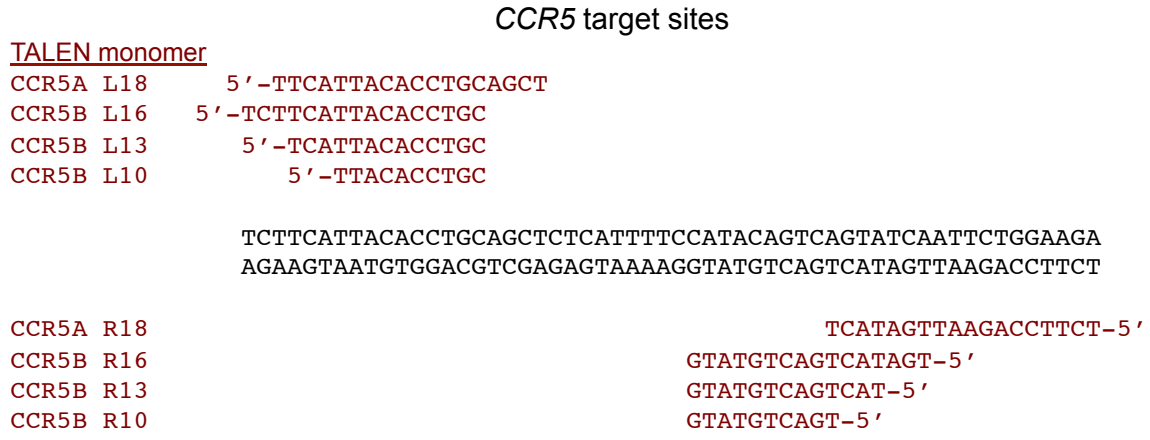
Using a previously described *in vitro* selection method,[21, 23] we interrogated TALENs for their abilities to each cleave 10^{12} potential off-target DNA substrates related to their intended target sequences. The resulting data provide the first comprehensive profiles of TALEN cleavage specificities in a manner that is not limited to the small number of typical target-related sites in a genome. The selection results suggest a model in which excess non-specific DNA-

binding energy gives rise to greater off-target cleavage relative to on-target cleavage. Based on this model, we engineered TALENs with substantially improved DNA cleavage specificity *in vitro*. In human cells, these engineered TALENs exhibit 24- to > 120-fold greater specificity for the most readily cleaved off-target site than currently used TALEN constructs.

1.2 Specificity profiling of TALENs targeting human genes

We profiled the specificities of 30 unique heterodimeric TALEN pairs (hereafter referred to as TALENs) harboring different C-terminal, N-terminal and *Fok*I domain variants and targeted to sites with half-sites of various lengths. Throughout this report, the number of base pairs recognized by each half site is listed to include the 5' T nucleotide recognized by the N-terminal domain. Most of the TALENs tested contained the obligate heterodimeric EL/KK *Fok*I domain, although the more active heterodimeric ELD/KKR and homodimeric *Fok*I nuclease domain were also used, as specified below.[3, 24] TALENs were constructed as previously reported[12] and designed to target one of three distinct sequences, which we refer to as CCR5A, CCR5B, or ATM, in two different human genes, *CCR5* and *ATM* (**Figure 1.2**).

A



B

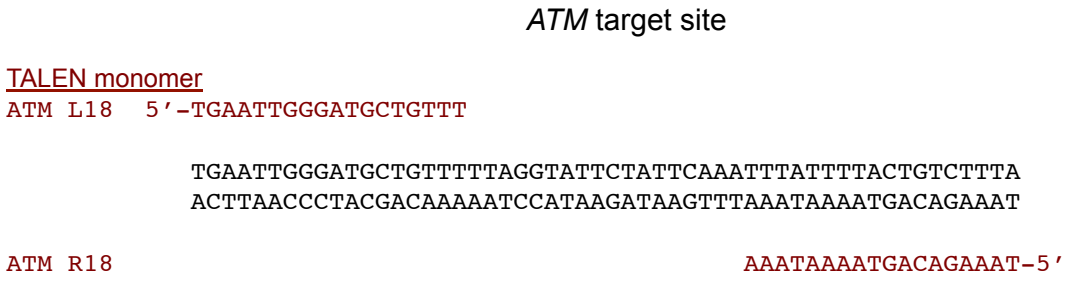


Figure 1.2. Target DNA sequences in human *CCR5* and *ATM* genes. The target DNA sequences for the TALENs used in this study are shown in black. The N-terminal TALEN end recognizing the 5′ T for each half-site target is noted (5′) and TALENs are named according to number of base pairs targeted. TALENs targeting the CCR5 L18 and R18 shown are referred to as CCR5A TALENs while TALENs targeting the L10, L13, L16, R10, R13 or R16 half-sites shown are referred to as CCR5B TALENs.

The specificity profiles were generated using a previously described in vitro selection method.21, 23 Briefly, pre-selection libraries of > 10^{12} DNA sequences each were digested with 3 nM to 40 nM of an *in vitro*-translated TALEN. TALEN concentrations in the selection were empirically determined to avoid overdigestion, since the selection isolates DNA products 1.5 target sites in length. For example, CCR5A TALEN digestion of concatemerized repeats of the CCR5 DNA target sites under the conditions used in the selection yielded the expected ladder of integral units of target sites (**Figure 1.3**).

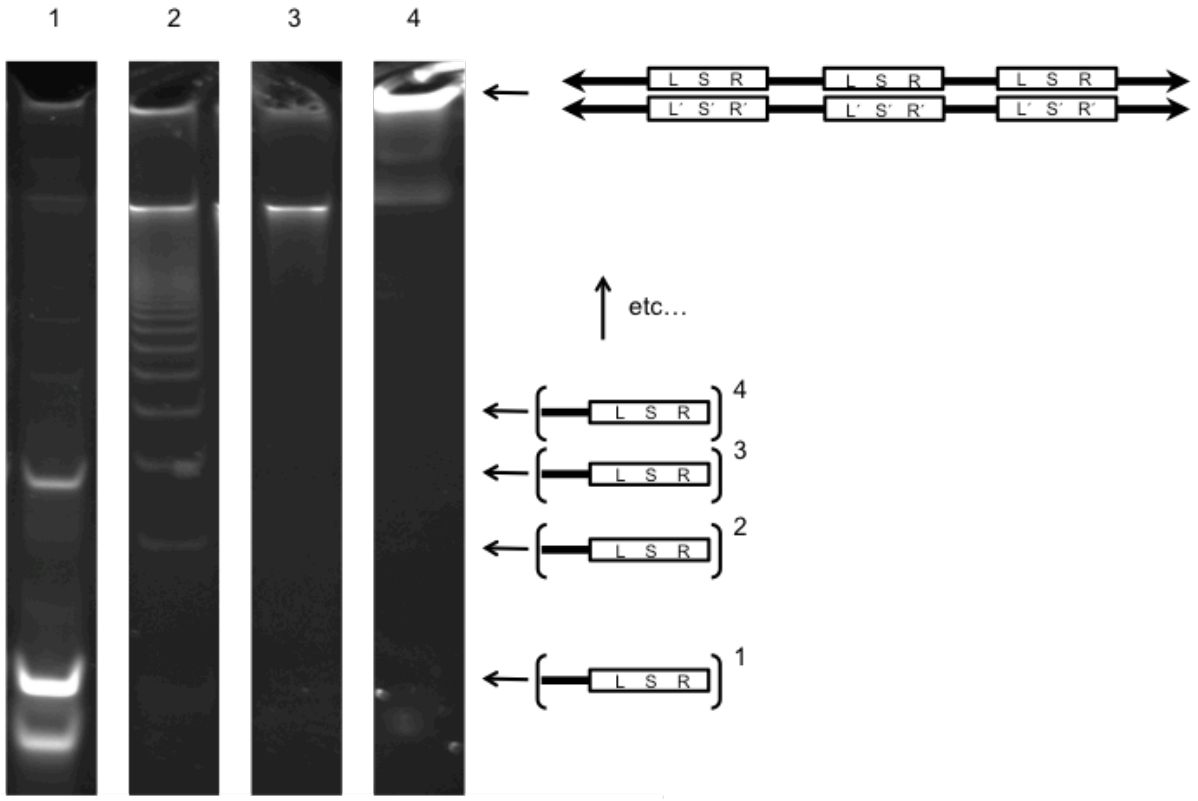


Figure 1.3. TALEN and restriction enzyme digestion of concatemerized on-target or mutant DNA sites. A single-stranded discrete DNA oligonucleotide containing a left half-site (L), spacer (S), right half-site (R), and constant region (represented by a thick black line) was circularized, then concatemerized by rolling circle amplification. The protocol for *In Vitro* Selection for DNA Cleavage (Online Methods) was preformed using CCR5Aon-Circ oligos (CCR5 on-target DNA sites) or CCR5Amut-Circ oligos (CCR5 mutant DNA sites) except after TALEN digestion samples were analyzed by PAGE gel. For lanes 1, 2, and 4 the DNA target site was the CCR5 on-target sequence but for lane 3 the DNA sequence contained a completely randomized right half-site. For lanes 1, 2, and 3 the resulting concatemerized DNAs were incubated with in vitro-translated CCR5A TALENs and the samples were purified by M-column (Qiagen) and analyzed by PAGE. Lane 1 shows digestion of concatemerized on-target DNA sites with a restriction enzyme (*Tsp*45I); lane 2 shows digestion of concatemerized on-target DNA sites with a CCR5A TALEN; lane 3 shows digestion of concatemerized mutant DNA sites with a CCR5A TALEN; and lane 4 shows the concatemerized DNA substrate before digestion or column purification.

Under these conditions, there is little cleavage to monomeric target sites compared to digestion of the same concatemerized DNA with a restriction enzyme that yields mostly monomeric sites. TALEN concentrations used in the selection (3 nM to 40 nM) correspond to ~20 to ~200 dimeric TALEN molecules per human cell nucleus,[25] which is in the lower range of cellular protein expression[26, 27] and therefore approaches the lower limit of TALEN concentration possible in a cell.

A completely random library of 25% of each base pair at each position would not yield sufficient library coverage of potential target sites with relatively few mutations relative to the on-target site. Therefore, a partially randomized library of 79% on-target base pair at each position was used. At this randomization frequency, 0.021% of the library contains an on-target site (no mutations), based on the expected binomial distribution of 79% on-target base pairs across the 36-bp target site. 14% of the library will have six mutations, with ~100 copies on average of all possible six-mutation sequences possible in a 36-bp site. (In theory, the library will have 1.0 x10^{12} molecules with exactly six mutations which is ~100-fold more than the 1.4 x10^{9} different six-mutation sequences possible in a 36-bp site.) Thus, pre-selection DNA libraries were sufficiently large that they each contain, in theory, at least ten copies of all possible DNA sequences with six or fewer mutations relative to the on-target sequence.

Incubation of the pre-selection DNA libraries with TALENs results in cleavage of preferred sequences. Cleaved library members harbored a free 5' monophosphate that enabled them to be captured by adapter ligation (**Figure 1.4**).

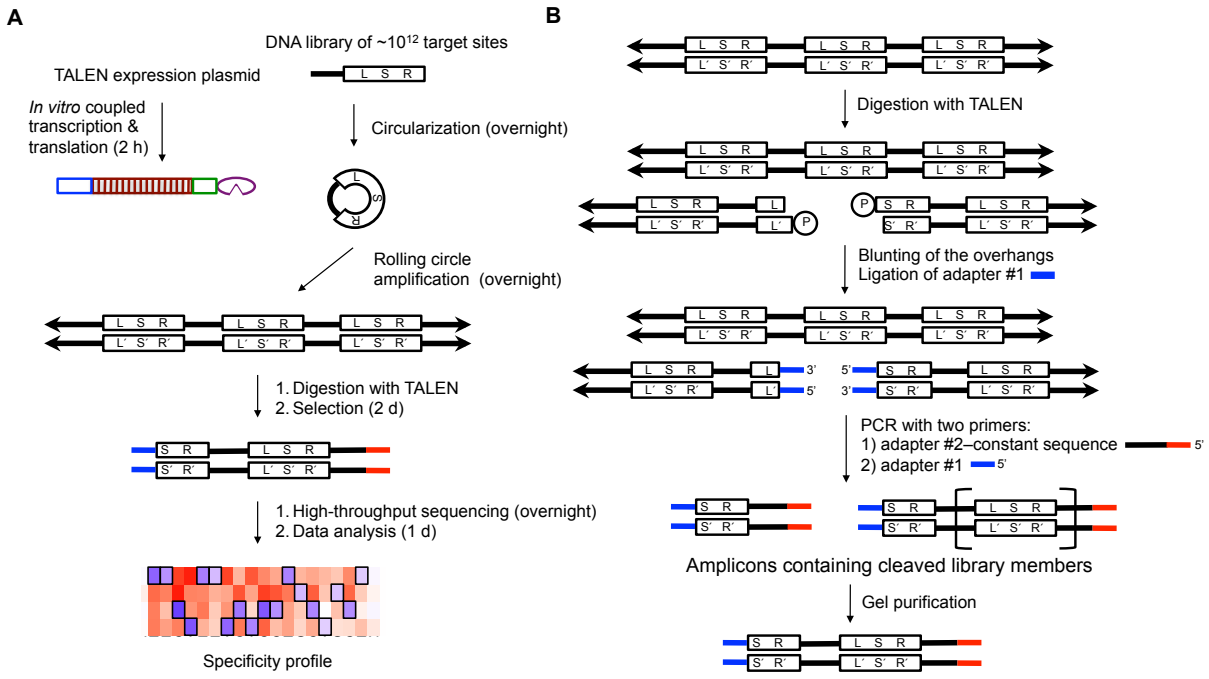


Figure 1.4. Selection scheme. (A) A single-stranded library of DNA oligonucleotides containing partially randomized left half-site (L), spacer (S), right half-site (R) and constant region (thick black line) was circularized, then concatemerized by rolling circle amplification. The concatemerized double stranded DNA (double arrows) contained repeated target sites with L' S' R' representing the reverse sequence complement of L S R. (B) The concatemerized DNA libraries of mutant target sites were incubated with an in vitro-translated TALEN of interest. Cleaved library members were blunted and ligated to adapter #1. The ligation products were 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. From the resulting ladder of amplicons containing a half-site with an integral number (*n*) of repeats of a target site (represented by brackets), amplicons corresponding to 1.5 target-sites in length were isolated by gel purification and subjected to high-throughput DNA sequencing and computational analysis.

DNA fragments of length corresponding to 1.5 target sites (an intact target site and a repeated half-site up to the point of TALEN-induced DNA cleavage) were isolated by gel purification. High-throughput sequencing and computational analysis of TALEN-treated or control samples surviving this selection process revealed the abundance of all TALEN-cleaved sequences as well as the abundance of the corresponding sequences before selection. In the control sample, all members of the pre-selection library were cleaved by a restriction endonuclease at a constant sequence to enable them to be captured by adapter ligation and isolated by gel purification. The enrichment value for each library member surviving selection was calculated by dividing its post-selection sequence abundance by its pre-selection abundance.

For ATM and CCR5A TALEN variants, the DNA that survived the selection contained significantly fewer mean mutations in the targeted half-sites than were present in the pre-selection libraries (**Figure 1.5**).

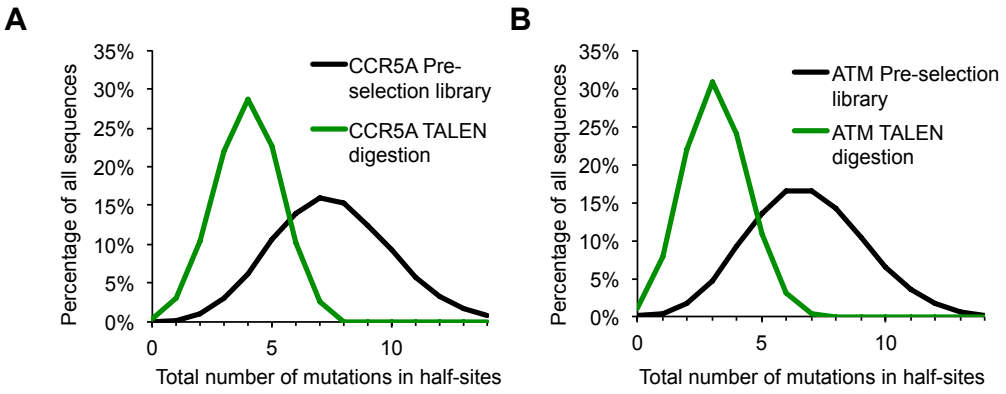


Figure 1.5. ***In vitro*** **selection results.** The fraction of sequences surviving selection (green) and before selection (black) are shown for CCR5A TALENs (A) and ATM TALENs (B) with EL/KK *Fok*I domains as a function of the number of mutations in both half-sites (left and right half-sites combined excluding the spacer).

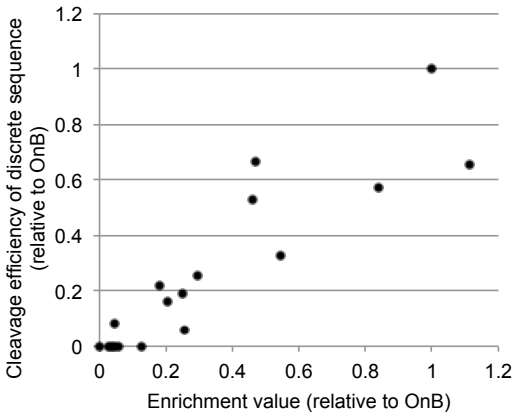
For example, the mean number of combined mutations in the 18-bp left half-site and 18-bp right half-site among DNA sequences surviving selection after treatment with TALENs was 4.06 for CCR5A and 3.18 for ATM sequences, respectively, compared to 7.54 and 6.82 mutations in the corresponding pre-selection libraries (**Figure 1.5**). For all selections, the on-target sequences were enriched by 8- to 640-fold with an average enrichment value of 110-fold (Supplementary

Table S4). To validate our selection results *in vitro*, we assayed the ability of the CCR5B TALEN targeting 13-bp left and right half-sites (L13+R13) to cleave each of 16 diverse off-target substrates (**Figure 1.6**).

A

DNA	Enrichment	Left half-site	Right half-site
OnB	1.0	TCATTACACCTGC	CATACAGTCAGTA
B1	0.47	TCATTACACCTGt	CATACAGTCAGTA
B2	0.46	TCATaACACCTGC	CATACAGTCAGTA
B3	0.13	TCATTACACCcGC	CATACAGTCAGTA
B4	1.1	TCATTACACCTGC	CATACAGTaAGTA
B5	0.84	TCATTACACCTGC	gATACAGTCAGTA
B6	0.21	TCATTACACCTGC	CATtCAGTCAGTA
B7	0.03	TCATTACACCTGC	CATACAGTCgGTA
B8	0.25	TCATaACACCTGt	CATACAGTCAGTA
B9	0.25	TCATTAtACCTaC	CATACAGTCAGTA
B10	0.00	TCATTgCACCcGC	CATACAGTCAGTA
B11	0.55	TCATTACACCTGC	gATACAGTaAGTA
B12	0.04	TCATTACACCTGC	CAcACAtTCAGTA
B13	0.03	TCATTACACCTGC	CATACtaTCAGTA
B14	0.30	TCATTACACCTGt	CATACAGTaAGTA
B15	0.18	TCATaACACCTGC	gATACAGTCAGTA
B16	0.04	TCATTACACCTGt	CATtCAGTCAGTA

C



B

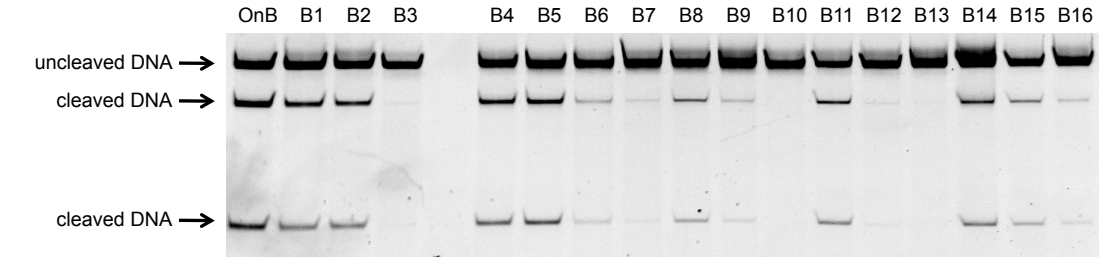


Figure 1.6. ***In vitro*** **selection results.** (A) Enrichment values from the selection of L13+R13 CCR5B TALEN for 16 mutant DNA sequences (mutations in red) relative to on-target DNA (OnB). (B) Discrete assays of on-target and off-target sequences from (A) as analyzed by PAGE. (C) Correspondence between discrete *in vitro* TALEN cleavage efficiency (cleaved DNA as a fraction of total DNA) for the sequences listed in (B) normalized to on-target cleavage (= 1) versus their enrichment values in the selection normalized to the on-target enrichment value (= 1). The Pearson's r coefficient of correlation between normalized cleavage efficiency and normalized enrichment value is 0.90.

The efficiencies with which each of these 16 putative off-target substrates were cleaved by the TALEN in these discrete *in vitro* assays correlated well ($r = 0.90$) with the observed enrichment values from the selection (**Figure 1.6**).

To quantify the DNA cleavage specificity at each position in the TALEN target site for all four possible base pairs, a specificity score was calculated as the difference between pre-selection and post-selection base pair frequencies, normalized to the maximum possible change of the pre-selection frequency from complete specificity (defined as 1.0) to complete anti-

specificity (defined as –1.0). For ATM and CCR5A TALENs tested, the targeted base pair at every position in both half-sites is preferred, with the sole exception of the base pair closest to the spacer for some ATM TALENs at the right-half site (**Figure 1.7**).

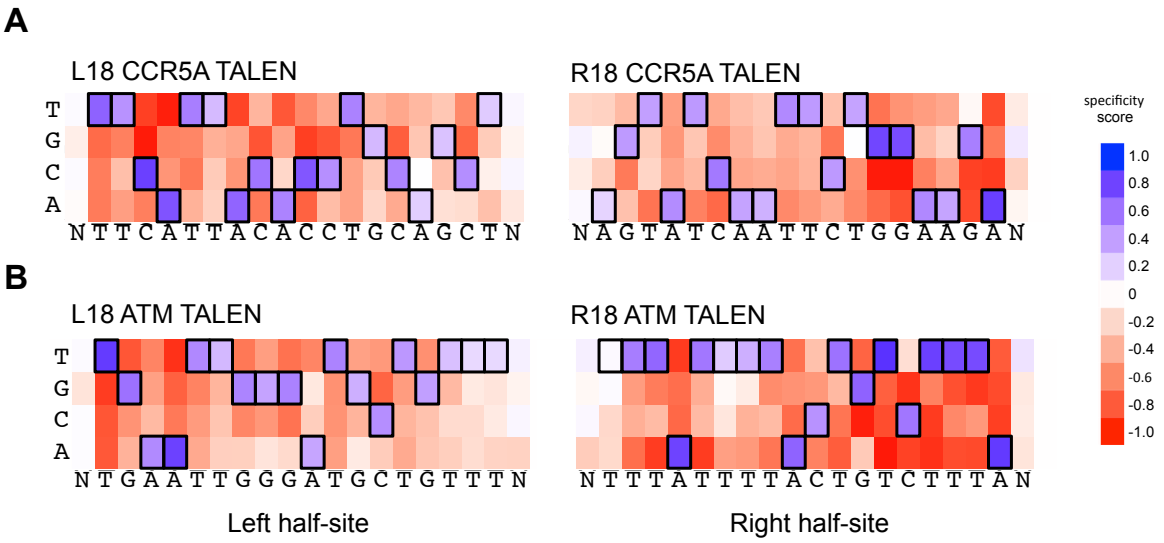


Figure 1.7. ***In vitro*** **selection results.** (A) Specificity scores for the CCR5A TALENs at all positions in the target half-sites plus a single flanking position. The colors range from dark blue (maximum specificity score of 1.0) to white (no specificity, score of 0) to dark red (maximum negative score of –1.0); see the main text for details. Boxed bases represent the intended target base. Note for the right half-site, the R18 TALENs, the sense strand is shown. (B) Same as (A) for the ATM TALENs.

The 5' T recognized by the N-terminal domain is highly specified, and the 3' DNA end (targeted by the C-terminal TALEN end) generally tolerates more mutations than the 5' DNA end; both of these observations are consistent with previous reports.[33, 34] All 12 of the positions targeted by the NN RVDs in the ATM and CCR5A TALENs were enriched for G, confirming previous reports[5, 7, 33, 35] that the NN RVD specifies G. Taken together, these results show that the selection data accurately predicts the efficiency of off-target TALEN cleavage *in vitro*, and that TALENs are overall quite specific across the entire target sequence.

Selection results for all TALEN variants and under all tested conditions demonstrated similar trends for the initial for selection results for ATM and CCR5A TALENs. The on-target sequences were enriched by 8- to 640-fold with an average enrichment value of 110-fold (**Figures 1.8 - 1.10 and Tables 1.1-1.10**).

A

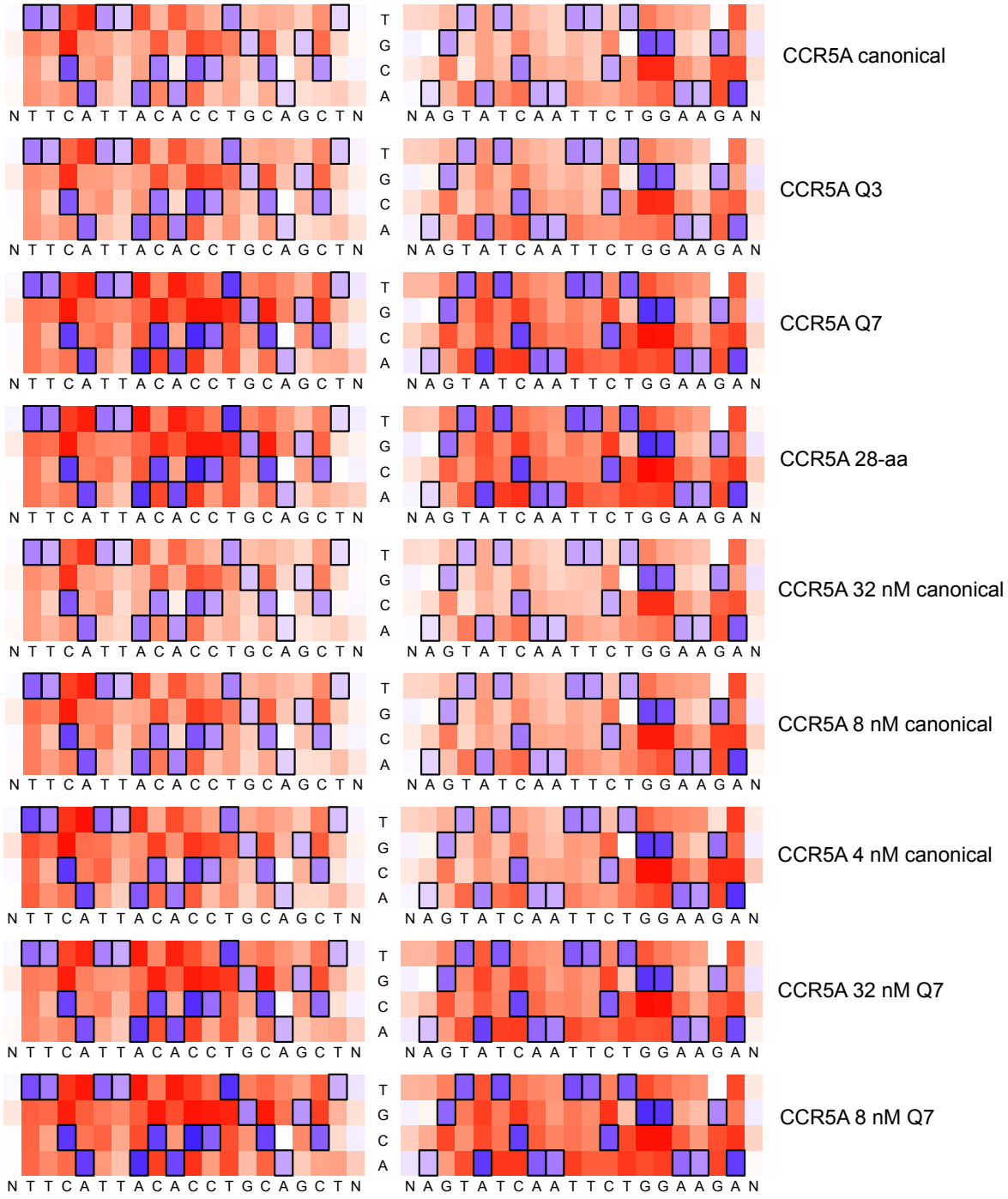


Figure 1.8 (Continued). Specificity profiles from all CCR5A TALEN selections as heat maps.

B

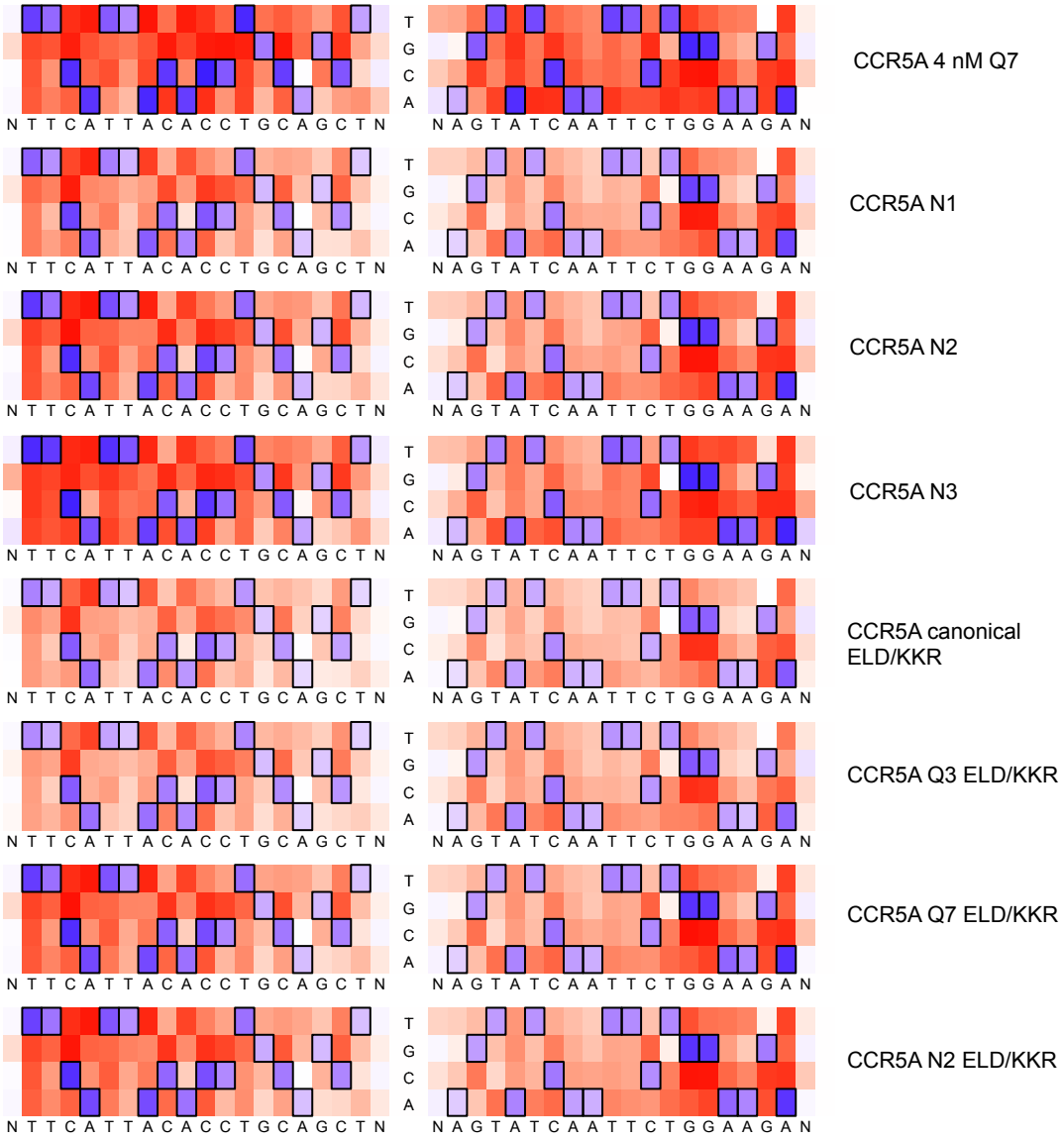


Figure 1.8 (Continued). Specificity profiles from all CCR5A TALEN selections as heat maps. Specificity scores for every targeted base pair in selections of CCR5A TALENs are shown. Specificity scores for the L18+R18 CCR5A TALEN at all positions in the target half-sites plus a single flanking position. The colors range from dark blue (maximum specificity score of 1.0) to white (score of 0, no specificity) to dark red (maximum negative score of -1.0); see the main text for details. Boxed bases represent the intended target base. The titles to the right indicate if the TALEN used in the selection differs from the canonical TALEN architecture, which contains a canonical C-terminal domain, wild-type N-terminal domain, and EL/KK *Fok*I variant. (A) Specificity profiles of canonical, Q3, Q7, 28-aa, 32 nM canonical, 8 nM canonical, 4 nM canonical, 32 nM Q7 and 8 nM Q7 CCR5A TALEN selections. (B) Specificity profiles of 4 nM Q7, N1, N2, N3, canonical ELD/KKR, Q3 ELD/KKR, Q7 ELD/KKR and N2 ELD/KKR CCR5A TALEN selections. When not specified, TALEN concentration was 16 nM.

A

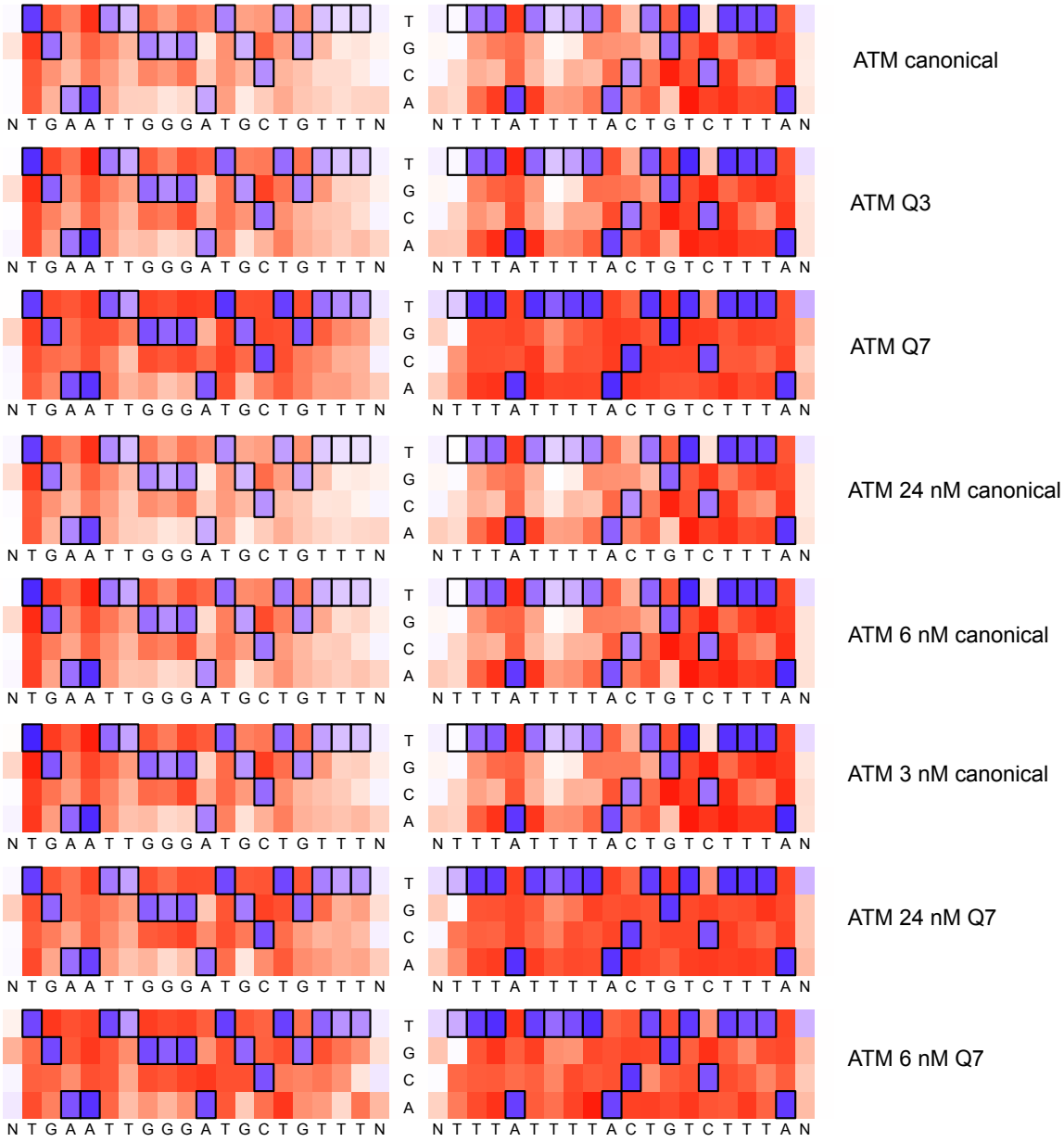


Figure 1.9 (Continued). Specificity profiles from all ATM TALEN selections as heat maps.

B

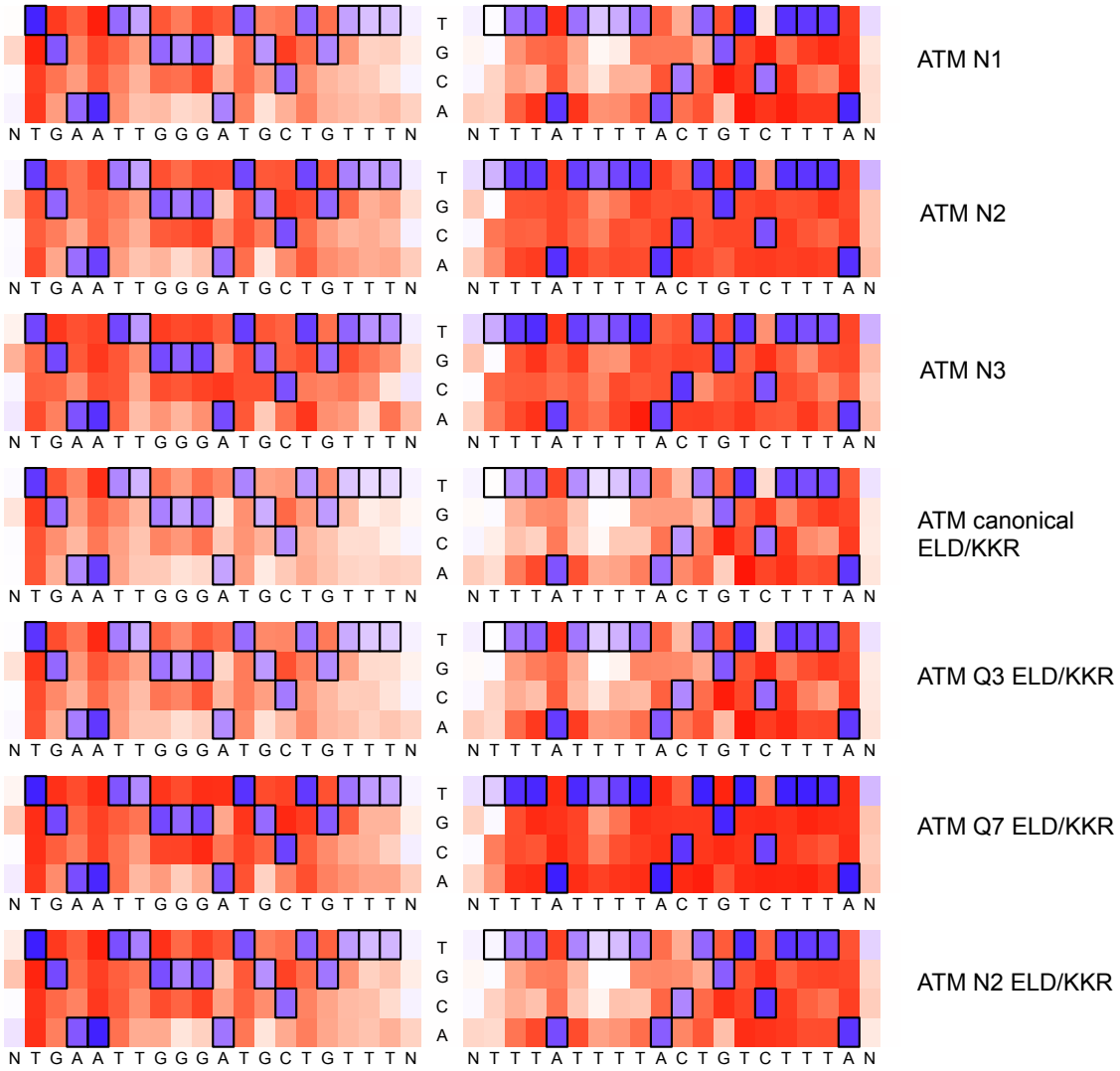


Figure 1.9 (Continued). Specificity profiles from all ATM TALEN selections as heat maps. Specificity scores for every targeted base pair in selections of ATM TALENs are shown. Specificity scores for the L18+R18 ATM TALEN at all positions in the target half-sites plus a single flanking position. The colors range from dark blue (maximum specificity score of 1.0) to white (score of 0, no specificity) to dark red (maximum negative score of -1.0); see the main text for details. Boxed bases represent the intended target base. The titles to the right indicate if the TALEN used in the selection differs from the canonical TALEN architecture, which contains a canonical C-terminal domain, wild-type N-terminal domain, and EL/KK *Fok*I variant. Selections correspond to conditions listed in Supplementary Table S1. (A) Specificity profiles of (12 nM) canonical, Q3, (12 nM) Q7, 24 nM canonical, 6 nM canonical, 3 nM canonical, 24 nM Q7, and 6 nM Q7 ATM TALEN selections. (B) Specificity profiles of N1, N2, N3, canonical ELD/KKR, Q3 ELD/KKR, Q7 ELD/KKR, and N2 ELD/KKR ATM TALEN selections. When not specified, TALEN concentration was 12 nM.

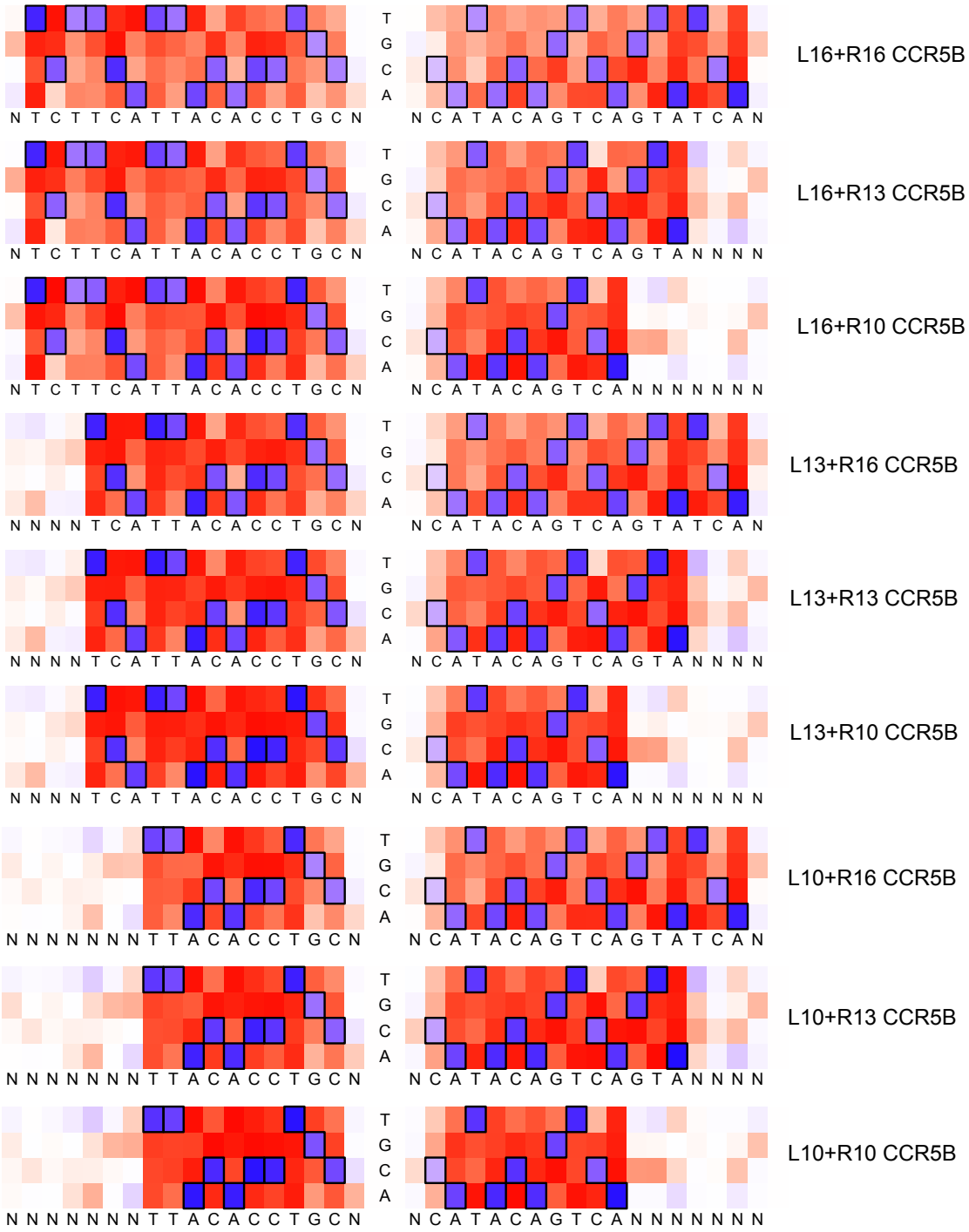


Figure 1.10. Specificity profiles from all CCR5B TALEN selections as heat maps. Specificity scores for every targeted base pair in selections of CCR5B TALENs are shown. Specificity scores for CCR5B TALENs targeting all possible combinations of the left (L10, L13, L16) and right (R10, R13, R16) half-sites at all positions in the target half-sites plus a single flanking position. The colors range from dark blue (maximum specificity score of 1.0) to white (score of 0, no specificity) to dark red (maximum negative score of -1.0); see the main text for details. Boxed bases represent the intended target base

Selection name	Target site	Left+Right half-site	Site length	N-terminal domain	C-terminal domain	*Fok*I domain	TALEN conc. (nM)
CCR5A 32 nM canonical	CCR5A	L18+R18	36	canonical	Canonical	EL/KK	32
CCR5A 16 nM canonical (or CCR5A canonical)	CCR5A	L18+R18	36	canonical	Canonical	EL/KK	16
CCR5A 8 nM canonical	CCR5A	L18+R18	36	canonical	Canonical	EL/KK	8
CCR5A 4 nM canonical	CCR5A	L18+R18	36	canonical	Canonical	EL/KK	4
CCR5A Q3	CCR5A	L18+R18	36	canonical	Q3	EL/KK	16
CCR5A 32 nM Q7	CCR5A	L18+R18	36	canonical	Q7	EL/KK	32
CCR5A 16 nM Q7 (or CCR5A Q7)	CCR5A	L18+R18	36	canonical	Q7	EL/KK	16
CCR5A 8 nM Q7	CCR5A	L18+R18	36	canonical	Q7	EL/KK	8
CCR5A 4 nM Q7	CCR5A	L18+R18	36	canonical	Q7	EL/KK	4
CCR5A 28-aa	CCR5A	L18+R18	36	canonical	28-aa	EL/KK	16
CCR5A N1	CCR5A	L18+R18	36	N1	Canonical	EL/KK	16
CCR5A N2	CCR5A	L18+R18	36	N2	Canonical	EL/KK	16
CCR5A N3	CCR5A	L18+R18	36	N3	Canonical	EL/KK	16
CCR5A canonical ELD/KKR	CCR5A	L18+R18	36	canonical	Canonical	ELD/KKR	16
CCR5A Q3 ELD/KKR	CCR5A	L18+R18	36	canonical	Q3	ELD/KKR	16
CCR5A Q7 ELD/KKR	CCR5A	L18+R18	36	canonical	Q7	ELD/KKR	16
CCR5A N2 ELD/KKR	CCR5A	L18+R18	36	N2	Canonical	ELD/KKR	16

Table 1.1. CCR5A TALEN constructs and concentrations used in the selections. For each selection using TALENs targeting the CCR5A target sequence, the selection name, the target DNA site, the TALEN N-terminal domain, the TALEN C-terminal domain, the TALEN *Fok*I domain, and the TALEN concentration (conc.) are shown.

Selection name	Target site	Left + Right half-site	Site length	N-terminal domain	C-terminal domain	*Fok*I domain	TALEN conc. (nM)
ATM 24 nM canonical	ATM	L18+R18	36	canonical	Canonical	EL/KK	24
ATM 12 nM canonical (or ATM canonical)	ATM	L18+R18	36	canonical	Canonical	EL/KK	12
ATM 6 nM canonical	ATM	L18+R18	36	canonical	Canonical	EL/KK	6
ATM 3 nM canonical	ATM	L18+R18	36	canonical	Canonical	EL/KK	3
ATM Q3	ATM	L18+R18	36	canonical	Q3	EL/KK	12
ATM 24 nM Q7	ATM	L18+R18	36	canonical	Q7	EL/KK	24
ATM 12 nM Q7 (or ATM Q7)	ATM	L18+R18	36	canonical	Q7	EL/KK	12
ATM 6 nM Q7	ATM	L18+R18	36	canonical	Q7	EL/KK	6
ATM N1	ATM	L18+R18	36	N1	Canonical	EL/KK	12
ATM N2	ATM	L18+R18	36	N2	Canonical	EL/KK	12
ATM N3	ATM	L18+R18	36	N3	Canonical	EL/KK	12
ATM canonical ELD/KKR	ATM	L18+R18	36	canonical	Canonical	ELD/KKR	12
ATM Q3 ELD/KKR	ATM	L18+R18	36	canonical	Q3	ELD/KKR	12
ATM Q7 ELD/KKR	ATM	L18+R18	36	canonical	Q7	ELD/KKR	12
ATM N2 ELD/KKR	ATM	L18+R18	36	N2	Canonical	ELD/KKR	12

Table 1.2. ATM TALEN constructs and concentrations used in the selections. For each selection using TALENs targeting the ATM target sequence, the selection name, the target DNA site, the TALEN N-terminal domain, the TALEN C-terminal domain, the TALEN *Fok*I domain, and the TALEN concentration (conc.) are shown.

Selection name	Target site	Left + Right half-site	Site length	N-terminal domain	C-terminal domain	*Fok*I domain	TALEN conc. (nM)
L16+R16 CCR5B	CCR5B	L16+R16	32	canonical	Canonical	EL/KK	10
L16+R13 CCR5B	CCR5B	L16+R13	29	canonical	Canonical	EL/KK	10
L16+R10 CCR5B	CCR5B	L16+R10	26	canonical	Canonical	EL/KK	10
L13+R16 CCR5B	CCR5B	L13+R16	29	canonical	Canonical	EL/KK	10
L13+R13 CCR5B	CCR5B	L13+R13	26	canonical	Canonical	EL/KK	10
L13+R10 CCR5B	CCR5B	L13+R10	23	canonical	Canonical	EL/KK	10
L10+R16 CCR5B	CCR5B	L10+R16	26	canonical	Canonical	EL/KK	10
L10+R13 CCR5B	CCR5B	L10+R13	23	canonical	Canonical	EL/KK	10
L10+R10 CCR5B	CCR5B	L10+R10	20	canonical	Canonical	EL/KK	10

Table 1.3. CCR5B TALEN constructs and concentrations used in the selections. For each selection using TALENs targeting the CCR5B target sequence, the selection name, the target DNA site, the TALEN N-terminal domain, the TALEN C-terminal domain, the TALEN *Fok*I domain, and the TALEN concentration (conc.) are shown.

Selection name	Seq. count	Mean mut.	Stdev mut.	Mut./bp	P-value vs. library	P-value vs. other TALENs
CCR5A 32 nM canonical	53883	4.327	1.483	0.120	3.3E-10	vs. CCR5A canonical ELD/KKR = 0.26
CCR5A 16 nM canonical	28940	4.061	1.438	0.113	5.4E-10	vs. CCR5A Q3 ELD/KKR = 0.028
CCR5A 8 nM canonical	29568	3.751	1.394	0.104	3.3E-10	
CCR5A 4 nM canonical	34355	3.347	1.355	0.093	1.5E-10	
CCR5A Q3	51694	3.841	1.380	0.107	1.7E-10	
CCR5A 32 nM Q7	48473	2.718	1.197	0.076	4.4E-11	
CCR5A 16 nM Q7	56593	2.559	1.154	0.071	3.1E-11	
CCR5A 8 nM Q7	43895	2.303	1.157	0.064	3.0E-11	
CCR5A 4 nM Q7	43737	2.018	1.234	0.056	2.1E-11	
CCR5A 28-aa	47395	2.614	1.203	0.073	4.0E-11	
CCR5A N1	64257	3.721	1.379	0.103	1.1E-10	vs. CCR5A 8 nM canonical =0.039
CCR5A N2	45467	3.148	1.306	0.087	8.2E-11	
CCR5A N3	24064	2.474	1.493	0.069	8.1E-11	
CCR5A canonical ELD/KKR	46998	4.336	1.491	0.120	4.0E-10	
CCR5A Q3 ELD/KKR	56978	4.098	1.415	0.114	2.2E-10	
CCR5A Q7 ELD/KKR	54903	3.234	1.330	0.090	7.3E-11	
CCR5A N2 ELD/KKR	79632	3.286	1.341	0.091	5.2E-11	

Table 1.4. Statistics of sequences selected by CCR5A TALEN digestion. Statistics are shown for each TALEN selection on the CCR5A target sequence . Seq. counts: total counts of high-throughput sequenced and computationally filtered selection sequences. Mean mut.: mean mutations in selected sequences. Stdev. mut.: standard deviation of mutations in selected sequences. Mut./bp: mean mutation normalized to target site length (bp). P-value vs. library: P-values between the TALEN selection sequence distributions to the corresponding pre-selection library sequence distributions were determined as previously reported[14] using a (left) one-sided *t*-test. P-value vs. other TALENs: all pairwise comparisons between each distribution of TALEN-digested sequences were calculated (e.g., comparing CCR5A 16 nM canonical and CCR5A 32 nM canonical). All P-values *above* the lower P-value limits are shown. Lower P-values limits were based on multiple comparison correction and thus any P-value below the lower value is considered significant. For CCR5A lower P-value limits were 0.0125 based on multiple comparisons correction using the Bonferroni method. All post-selection sequences were assumed to be binomially distributed.

Selection name	Seq. count	Mean mut.	Stdev mut.	Mut./bp	P-value vs. library	P-value vs. other TALENs
ATM 24 nM canonical	89571	3.262	1.360	0.091	6.54E-11	vs. ATM canonical ELD/KKR =0.046
ATM 12 nM canonical (or ATM canonical)	96703	3.181	1.307	0.088	5.36E-11	
ATM 6 nM canonical	78852	2.736	1.259	0.076	3.63E-11	
ATM 3 nM canonical	82527	2.552	1.258	0.071	2.71E-11	
ATM Q3	96582	2.551	1.248	0.071	2.31E-11	
ATM 24 nM Q7	10166	1.885	2.125	0.052	2.06E-10	
ATM 12 nM Q7 (or ATM Q7)	4662	1.626	2.083	0.045	5.31E-10	vs. ATM 6 nM Q7 =0.069
ATM 6 nM Q7	1290	1.700	2.376	0.047	7.16E-09	
ATM N1	84402	2.627	1.318	0.073	2.92E-11	vs. ATM N3 =0.039
ATM N2	62470	2.317	1.516	0.064	2.69E-11	
ATM N3	1605	2.720	2.363	0.076	2.69E-08	vs. ATM 6 nM canonical =0.36 vs. Q3 ELD/KKR =0.017
ATM canonical ELD/KKR	107970	3.279	1.329	0.091	5.48E-11	
ATM Q3 ELD/KKR	104099	2.846	1.244	0.079	3.15E-11	
ATM Q7 ELD/KKR	21108	1.444	1.56	0.040	3.02E-11	
ATM N2 ELD/KKR	70185	2.45	1.444	0.06805	2.82E-11	

Table 1.5. Statistics of sequences selected by ATM TALEN digestion. Statistics are shown for each TALEN selection on the ATM target sequence. Seq. counts: total counts of high-throughput sequenced and computationally filtered selection sequences. Mean mut.: mean mutations in selected sequences. Stdev. mut.: standard deviation of mutations in selected sequences. Mut./bp: mean mutation normalized to target site length (bp). P-value vs. library: P-values between the TALEN selection sequence distributions to the corresponding pre-selection library sequence distributions) were determined as previously reported[14] using a (left) one-sided *t*-test. P-value vs. other TALENs: all pairwise comparisons between each distribution of TALEN-digested sequences were calculated. All P-values *above* the lower P-value limits are shown. Lower P-values limits were based on multiple comparison correction and thus any P-value below the lower value is considered significant. For ATM lower P-value limits were 0.0125 based on multiple comparisons correction using the Bonferroni method. All post-selection sequences were assumed to be binomially distributed. Note that for the 3 nM Q7 ATM and the 28-aa ATM selection not enough sequences were obtained to interpret, although these selections were performed.

Selection name	Seq. count	Mean mut.	Stdev mut.	Mut./bp	P-value vs. library	P-value vs. other TALENs
L16+R16 CCR5B	34904	2.134	1.168	0.067	4.7E-11	
L16+R13 CCR5B	38229	1.581	1.142	0.055	2.7E-11	
L16+R10 CCR5B	37801	1.187	0.949	0.046	2.2E-11	
L13+R16 CCR5B	46608	1.505	1.090	0.052	1.7E-11	
L13+R13 CCR5B	53973	0.996	1.025	0.038	8.8E-12	
L13+R10 CCR5B	60550	0.737	0.884	0.032	7.4E-12	
L10+R16 CCR5B	36927	1.387	0.971	0.053	3.0E-11	
L10+R13 CCR5B	58170	0.839	0.882	0.036	9.1E-12	
L10+R10 CCR5B	57331	0.646	0.779	0.032	1.0E-11	

Table 1.6. Statistics of sequences selected by CCR5B TALEN digestion. Statistics are shown for each TALEN selection on the CCR5B target sequences. Seq. counts: total counts of high-throughput sequenced and computationally filtered selection sequences. Mean mut.: mean mutations in selected sequences. Stdev. mut.: standard deviation of mutations in selected sequences. Mut./bp: mean mutation normalized to target site length (bp). P-value vs. library: P-values between the TALEN selection sequence distributions to the corresponding pre-selection library sequence distributions) were determined as previously reported[14] using a (left) one-sided *t*-test. P-value vs. other TALENs: all pairwise comparisons between each distribution of TALEN-digested sequences were calculated. All P-values *above* the lower P-value limits are shown. Lower P-values limits were based on multiple comparison correction and thus any P-value below the lower value is considered significant. For CCR5B lower P-value limits were 0.0055 based on multiple comparisons correction using the Bonferroni method. All post-selection sequences were assumed to be binomially distributed.

Library name	Target site	Left + Right half-site	Site length	Seq. count	Mean mut.	Stdev mut.	Mut./bp
CCR5A Library	CCR5A	L18+R18	36	158643	7.539	2.475	0.209
ATM Library	ATM	L18+R18	36	212661	6.820	2.327	0.189
CCR5B Library	CCR5B	L16+R16	32	280223	6.500	2.441	0.203
CCR5B Library	CCR5B	L16+R13	29	280223	5.914	2.336	0.204
CCR5B Library	CCR5B	L16+R10	26	280223	5.273	2.218	0.203
CCR5B Library	CCR5B	L13+R16	29	280223	5.969	2.340	0.206
CCR5B Library	CCR5B	L13+R13	26	280223	5.383	2.230	0.207
CCR5B Library	CCR5B	L13+R10	23	280223	4.742	2.106	0.206
CCR5B Library	CCR5B	L10+R16	26	280223	5.396	2.217	0.208
CCR5B Library	CCR5B	L10+R13	23	280223	4.810	2.100	0.209
CCR5B Library	CCR5B	L10+R10	20	280223	4.169	1.971	0.208

Table 1.7. Statistics of sequences from pre-selection libraries. For each pre-selection library containing a distribution of mutant sequences of the CCR5A target sequence, ATM target sequence and CCR5B target sequences. Seq. counts: total counts of high-throughput sequenced and the computationally filtered selection sequences. Mean mut.: mean mutations of sequences. Stdev. mut.: standard deviation of sequences. All pre-selection sequences were assumed to be binomially distributed. Mut./bp: mean mutation normalized to target site length (bp).

Selection	Enrichment value							
	0 Mut.	1 Mut.	2 Mut.	3 Mut.	4 Mut.	5 Mut.	6 Mut.	7 Mut.
CCR5A 32 nM canonical	9.879	9.191	8.335	6.149	4.205	2.269	1.005	0.325
CCR5A 16 nM canonical	12.182	13.200	10.322	7.195	4.442	2.127	0.748	0.216
CCR5A 8 nM canonical	19.673	17.935	13.731	8.505	4.512	1.756	0.531	0.116
CCR5A 4 nM canonical	36.737	29.407	19.224	9.958	4.047	1.242	0.302	0.058
CCR5A Q3	18.550	16.466	12.024	8.070	4.632	1.938	0.572	0.126
CCR5A 32 nM Q7	60.583	54.117	31.082	11.031	2.640	0.469	0.073	0.013
CCR5A 16 nM Q7	62.294	64.689	35.036	10.538	2.183	0.322	0.046	0.010
CCR5A 8 nM Q7	97.020	91.633	38.634	8.974	1.485	0.189	0.029	0.010
CCR5A 4 nM Q7	197.239	130.497	38.361	6.535	0.896	0.120	0.025	0.019
CCR5A 28-aa	70.441	62.213	33.481	10.488	2.317	0.402	0.064	0.012
CCR5A N1	19.038	18.052	13.858	8.788	4.546	1.697	0.499	0.115
CCR5A N2	41.715	35.752	22.638	10.424	3.777	0.989	0.194	0.038
CCR5A N3	173.897	88.392	31.503	8.770	1.853	0.350	0.089	0.036
CCR5A canonical ELD/KKR	8.101	10.012	8.220	6.147	4.119	2.291	1.019	0.330
CCR5A Q3 ELD/KKR	14.664	12.975	9.409	6.819	4.544	2.235	0.797	0.198
CCR5A Q7 ELD/KKR	37.435	32.922	21.033	10.397	3.867	1.087	0.238	0.046
CCR5A N2 ELD/KKR	35.860	31.469	20.135	10.189	3.983	1.155	0.260	0.050

Table 1.8. CCR5A enrichment values of sequences as a function of number of mutations. For each TALEN selection on the *CCR5A* target sequence, enrichment values calculated by dividing the fractional abundance of post-selection sequences from a TALEN digestion by the fractional abundance of pre-selection sequences as a function of total mutations (Mut.) in the half-sites.

Selection	Enrichment value							
	0 Mut.	1 Mut.	2 Mut.	3 Mut.	4 Mut.	5 Mut.	6 Mut.	7 Mut.
ATM 24 nM canonical	19.900	16.881	12.162	6.318	2.629	0.884	0.228	0.057
ATM 12 nM canonical	20.472	17.645	12.724	6.549	2.606	0.803	0.189	0.039
ATM 6 nM canonical	41.141	29.522	17.153	6.551	1.872	0.431	0.082	0.017
ATM 3 nM canonical	56.152	37.152	18.530	6.196	1.562	0.308	0.058	0.015
ATM Q3	50.403	36.687	19.031	6.245	1.513	0.294	0.057	0.016
ATM 24 nM Q7	353.148	90.350	13.475	1.531	0.186	0.128	0.116	0.118
ATM 12 nM Q7	513.385	89.962	11.310	0.860	0.190	0.093	0.115	0.092
ATM 6 nM Q7	644.427	82.074	7.650	0.677	0.170	0.205	0.163	0.164
ATM N1	57.218	35.388	17.808	6.124	1.644	0.383	0.076	0.023
ATM N2	119.240	53.618	18.977	4.742	0.992	0.233	0.076	0.044
ATM N3	201.158	55.468	15.244	3.187	0.764	0.307	0.154	0.173
ATM canonical ELD/KKR	19.356	15.692	11.855	6.403	2.706	0.899	0.224	0.054
ATM Q3 ELD/KKR	32.816	25.151	16.172	6.727	2.095	0.506	0.095	0.018
ATM Q7 ELD/KKR	447.509	93.166	13.505	1.543	0.170	0.053	0.049	0.045
ATM N2 ELD/KKR	90.625	45.525	18.683	5.369	1.267	0.274	0.076	0.035

Table 1.9. ATM Enrichment values of sequences as a function of number of mutations. For each TALEN selection on the *ATM* target sequence enrichment values calculated by dividing the fractional abundance of post-selection sequences from a TALEN digestion by the fractional abundance of pre-selection sequences as a function of total mutations (Mut.) in the half-sites.

Selection	Enrichment value							
	0 Mut.	1 Mut.	2 Mut.	3 Mut.	4 Mut.	5 Mut.	6 Mut.	7 Mut.
L16+R16 CCR5B	59.422	35.499	13.719	3.770	0.737	0.132	0.024	0.011
L16+R13 CCR5B	80.852	31.434	7.754	1.380	0.218	0.040	0.022	0.016
L16+R10 CCR5B	64.944	20.056	3.867	0.515	0.056	0.010	0.006	0.006
L13+R16 CCR5B	101.929	34.255	8.131	1.299	0.167	0.033	0.016	0.011
L13+R13 CCR5B	113.102	22.582	3.037	0.315	0.044	0.022	0.017	0.017
L13+R10 CCR5B	74.085	11.483	1.270	0.121	0.022	0.013	0.011	0.013
L10+R16 CCR5B	60.186	22.393	5.286	0.777	0.084	0.012	0.006	0.006
L10+R13 CCR5B	74.204	13.696	1.673	0.152	0.021	0.011	0.010	0.009
L10+R10 CCR5B	43.983	7.018	0.740	0.061	0.013	0.007	0.007	0.008

Table 1.10. CCR5B enrichment values of sequences as a function of number of mutations. For each TALEN selection on the CCR5B target sequence, enrichment values calculated by dividing the fractional abundance of post-selection sequences from a TALEN digestion by the fractional abundance of pre-selection sequences as a function of total mutations (Mut.) in the half-sites.

1.3 TALEN off-target cleavage in cells

For TALENs targeting 36 total base pairs, potential off-target sites in the human genome are expected on the average to contain approximately eight or more mutations relative to the on-target site (**Table 1.11**), more mutations than theoretically are covered in the *in vitro* selection.

Mutations in site	Off-target sites to CCR5A	Statistically expected
0	1	1
1	0	0.0
2	0	0.0
3	0	0.0
4	0	0.0
5	0	0.0
6	0	0.0
7	0	0.3
8	8	3.6
9	70	34.1
10	634	275.9
11	4338	1956.3
12	27114	12226.7
13	149005	67716.9
14	648230	333747.3
15	2657598	1468488.3
16	9783617	5782172.6

Table 1.11. Genomic off-target site abundance in the human genome. Column 2 shows the number of sites in the human genome related to the CCR5A on-target sequence, allowing for a spacer length from 12 to 25 bps between the two half-sites. Column 3 shows the statistically expected average numbers of off-target sites that would be found in a human-sized genome for a 36-bp target sequence assuming a random distribution of an equal ratio of A:C:G:T across the genome utilizing the binomial distribution of mutant sites multiplied by the number of spacer lengths allowed (14). Number of sites with m mutations = genome size ($2 \times 3.1 \times 10^9$) x spacer length allowance (14) x binomial distribution of mutations in a 36-bp target site with a 0.75 probability of a mutation occurring [$(36! / (m! \times (36 - m)!)) \times (0.75)^m \times (0.25)^{36-m}$].

Therefore, we used a machine-learning algorithm[31] trained on the tens of thousands of off-target sites revealed by the *in vitro* selection to identify rare TALEN candidate off-target sites in the human genome. The "classifier" algorithm calculates the posterior probability of each nucleotide in each position of a target to occur in a sequence that was cleaved by the TALENs in opposition to sequences from the target library that were not observed to be cleaved.[28] These posterior probabilities were then used to score the likelihood that the TALEN used to train the

algorithm would cleave every possible target sequence in the human genome with monomer spacing of 10 to 30 bps. Since sites containing the longest tested DNA spacers (24 bp) were productively cleaved during the *in vitro* selection, target sites with even longer spacers were considered for potential off-target modification in the genome.

Using this "classifier" algorithm, we identified the 36 best-scoring heterodimeric candidate off-target sites for the ATM TALENs and 48 of the best-scoring candidate off-target sites for the CCR5A TALENs (**Table 1.12**).

A

CCR5A Site	Score	Mut.	Left half-site	Spacer length	Right half-site
OnCCR5A	0.008	0	TTCATTACACCTGCAGCT	18	AGTATCAATTCTGGAAGA
OffC-1	0.747	9	TaCATcACAtaTGCAaaT	29	tGTATCAtTTCTGGgAGA
OffC-2	0.747	9	TaCATcACAtaTGCAaaT	29	tGTATCAtTTCTGGgAGA
OffC-3	0.747	9	TaCATcACAtaTGCAaaT	29	tGTATCAtTTCTGGgAGA
OffC-4	0.747	11	TcCATaACACaTctttCT	10	tGcATCAtTcCTGGAAGA
OffC-5	0.804	11	TcCAaTACctCTGCcaCa	14	AGgAgCAAcTCTGGgAGA
OffC-6	0.818	10	TTCAgTcCAtCTGaAaac	16	gGTATCAtTTCTGGAgGA
OffC-7	0.834	14	TaCAaaACcCtTGCcaaa	27	taTATCAATTtgGGgAGA
OffC-8	0.837	12	TcCAagACACCTGCttac	26	tcTATCAATTtgGGgAGA
OffC-9	0.874	10	TTCATaACAtCTtaAaaT	27	AaTAcCAAcTCTGGAtGA
OffC-10	0.89	12	TcCAaaACAtCTGaAaaT	25	tGgATCAAaTtgGGAAGA
OffC-11	0.896	12	TTCAgaACACaTGactac	21	tGTATCAgTTaTGGAtGA
OffC-12	0.904	13	TcCATaAtAtCTtCctCT	28	gGgATtAATTtgGGAgGA
OffC-13	0.905	11	TgCAaTAtACCTGttGaT	16	ctcATCAATTCTGGgtGA
OffC-14	0.906	12	TTCATaACACtccacctT	16	gGTATCAAaTCTGGggGA
OffC-15	0.906	12	TcCATgACACaaaagaCT	26	gGTATCtATcCTGGAAtA
OffC-16	0.906	9	TTCcTTcCACCaGtgtCc	28	AGcATCAATcCTGGAAGA
OffC-17	0.907	10	TTaATaACAtCTcCAaCT	24	gGcAcCAAaTCTGGAtGA
OffC-18	0.909	13	TcCATcACcCCTcCctCc	10	gGTgcCAgcTCTGGAgGA
OffC-19	0.909	8	TTCATTACtCCTcCttCT	30	ctTATCAcTTtTGGAAGA
OffC-20	0.912	10	TgCATTACACaTtatGtg	17	AGcAgCAcTTCTGGAAGA
OffC-21	0.913	11	TTCAaaACACaTaCAtCT	28	AacAaCAtTcCTGtAAGA
OffC-22	0.913	10	TcCATTACcaCTGCAGaT	25	gacATCAgTTaTGGAtGA
OffC-23	0.925	13	TTCcagACcCCTtCctCa	13	gacATCAAaTCTGGgAGA
OffC-24	0.927	12	TTCcaaACACCcGCttCc	26	taTATCctTTCTGGAAtA
OffC-25	0.93	12	TgaAaTACACCTGCctaT	13	gGccTCAAggCTGGAtGA
OffC-26	0.93	12	TgCcaaACctCTGtcaCc	22	AGgATCAcTTCTGGAAGA
OffC-27	0.931	12	TgCcaaACctCTGtcaCc	22	AGgATCAcTTCTGGAAGA

Table 1.12 (Conitinued). Predicted ATM and CCR5A off-target sites in the human genome.

OffC-28	0.931	8	TTtATTACACtTcCAGaT	19	gaTATCctTTCTGGAAGA
OffC-29	0.932	13	TaCAaaAaACtTtCtGag	27	tGTATCAATTtgGGgAGA
OffC-30	0.932	11	TcCAaaACACCcaCAGac	19	gGTATagATTgTGGAAGA
OffC-31	0.934	13	TTCATTcCACaTcCccac	25	gtTATCAAcatgGGAAGA
OffC-32	0.934	11	TTCAaTAtgCCaaCAGCT	11	AGctTCAATctgGGAgGA
OffC-33	0.934	12	TTCAaTACACtTGtctaT	12	tGTgTCAtTTCTGGgttA
OffC-34	0.935	11	TTCAacACACCTtCAaaa	12	tGTgTCAtTaaTGGAAGA
OffC-35	0.935	10	TTCAaaACAtCTGacatT	10	AaTAgaAATTCTGGAAGA
OffC-36	0.935	11	cTCcTaAtACCTGCAaaT	21	gaTATtAtTTCTGGAgGA
OffC-38	0.939	10	TTCATaACAaCTtaAaCa	16	tGgATCAATTgTGGAAtA
OffC-39	0.941	9	TTaATTAtAttTttAaCa	25	AGTAcCAATTCTGGAAGA
OffC-40	0.941	9	TTaATTAtAttTttAaCa	25	AGTAcCAATTCTGGAAGA
OffC-42	0.941	9	TcaATTAtACCTaCAtaT	24	AGTtTCAATTtaGGAAGA
OffC-45	0.943	10	aTCATTgCcCCTGCAGag	23	gGTATCttTcCTGGAtGA
OffC-49	0.946	9	TTCAaTACACCTaCAtCa	15	AGTcaCAtTTCTGGAAtt
OffC-56	0.951	9	TgCATTACAaCTGaAGac	19	AGcATtAcTTCaGGAAGA
OffC-65	0.956	9	TaCATTcCACCTaCttCc	22	AGTtTgAcTTCTGGAAGA
OffC-69	0.958	9	TTCATTACACCcaCtGCT	23	caTtTCtAgTCTGGAttA
OffC-76	0.96	9	TcCATaAtcCCatCAGCT	19	AGTgTCAtTTCTGGAAGg
OffC-137	0.974	9	TTCATTAtgCCTGCAGta	14	AGcATCAATTCatGAtGA
OffC-150	0.975	9	TTCATTgCACCTGataCT	16	gGaAcCtATTCTGGAAGt
B					
OnATM	0.000	0	TGAATTGGGATGCTGTTT	18	TTTATTTTACTGTCTTTA
OffA-1	0.595	7	TGAATaGGaAataTaTTT	20	TTTATTTTACTGTtTTTA
OffA-2	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA
OffA-3	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA
OffA-4	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA
OffA-5	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA
OffA-6	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA
OffA-7	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA
OffA-8	0.7	8	TGcATaGGaATGCTaaTT	10	TTTATTTTACTaTtTaTA
OffA-9	0.708	10	TGAATTaaaATcCTGcTT	19	gTTATaTgACTaTtTTTA
OffA-10	0.711	10	TccATTaaaATaCTaTTT	18	TTTATTTTAtTaTtTTTA
OffA-11	0.715	10	TGAATTGaGAgaagcaTT	16	TTTATTTTAtTaTtTTTA
OffA-12	0.725	10	TGAAgTGGGATaCTGTTa	29	ggTATaTTAtaaTtTTTA
OffA-13	0.729	9	TGAATTatGAaGCTacTT	17	TTTATTgTAaTaTtTTTA
OffA-14	0.731	9	TGAATaaGGATGCTaTTa	25	TTTATTTattTaTtTTTA
OffA-15	0.744	10	TGAATgGGGAcaCaGcca	29	TTTATTTTAtTaTtTTTA
OffA-16	0.752	9	TaAATgGaaATGCTGTTc	24	aTTATTTTAtTGTtTTTt
OffA-17	0.761	9	gGAAaTGGGATaCTGagT	15	TTTATgTTACTaTtTcTA
OffA-18	0.781	11	TGgATcGaagTGaTtaTT	23	TTTATTTTAtTaTtTTTA
OffA-19	0.792	11	TGAATTGaGATtCacagc	23	TTTATTTTttTaTtTTTA

Table 1.12 (Continued). Predicted ATM and CCR5A off-target sites in the human genome.

OffA-20	0.803	8	TGAATTaGGAatCTGaTT	10	TTTATTTTAtTaTtaTTA
OffA-21	0.807	12	TaAATTaaaATaCTccag	23	aTTATTTTAaTGTtTTTA
OffA-22	0.811	10	TGAATaGGaATatTcTTT	12	TTTATTTattTaTtTTTA
OffA-23	0.811	9	TagATTGaaATGCTGTTT	15	TTTtTaTTAtTaTtTTTA
OffA-24	0.816	10	TGAcTaGaaATGaTGaTT	25	TTTATTTTctTaTtTTTA
OffA-25	0.817	12	TGAATTtaaAaaaTGTcc	13	aTTATTTTAtTaTtTTTA
OffA-26	0.817	12	TGAATTtaaAaaaTGTcc	13	aTTATTTTAtTaTtTTTA
OffA-27	0.817	10	TGgATccaGATaCTcTTT	10	TTTATTTTttTaTtTTTA
OffA-28	0.819	7	TGgAgTGaGATcCTGTTT	21	TTTATTTTAtTGTtaTTA
OffA-29	0.824	8	TGAAcTtGGATGaTaTaT	24	TTTATTTgAtTaTCTTTA
OffA-30	0.832	9	TGtATTGGGATaCcaTTT	26	TcTATTTTAtTaTtTTTt
OffA-31	0.833	9	TcAATTGGGATGaTcaTa	23	TTTATTcTAtTtTtTTTA
OffA-32	0.835	9	TGAAagGGaAaGtTGgaT	23	TTTATTTTACTaTtTTTA
OffA-33	0.841	9	TGgtTTGGGATcCTGTgT	27	TTTATgTTttTaTtTTTA
OffA-34	0.841	9	TGAAaTGGGATGagcTTg	28	TTTATTTTAtTaTtTTaA
OffA-35	0.844	10	TGAATTGGGATaCTGTag	29	cTTAaaTaAaTaTtTTTA
OffA-36	0.844	10	TGAATTGtGgTatTGccT	18	TTTATggTttTGTCTTTA

Table 1.12 (Continued). Predicted ATM and CCR5A off-target sites in the human genome. (A) Using a machine-learning "classifier" algorithm trained on the output of the *in vitro* CCR5A TALEN selection,[15] mutant sequences of the target site allowing for spacer lengths of 10 to 30 base pairs were scored. The resulting 36 predicted off-targets sites with the best scores for the CCR5A TALENs and the next 12 best-scoring off-target sites with nine or ten mutations are shown with their respective classifier scores, mutation numbers, left and right half-site sequences (mutations from on-target in lower case), and the length of the spacer between half-sites in base pairs. (B) The 36 predicted off-targets sites with the best scores for the ATM TALENs for ATM TALENs (without the next 12 best-scoring off-target sites with nine or ten mutations) are listed as OffA.

These sites differ from the on-target sequence at seven to fourteen positions. These 84 predicted off-target sites for CCR5A and ATM TALENs were amplified from genomic DNA purified from human U2OS-EGFP cells expressing either CCR5A or ATM TALENs.[12] Sequences containing insertions or deletions of three or more base pairs in the DNA spacer of the potential genomic off-target sites and present in significantly greater numbers in the TALEN-treated samples versus the untreated control sample were considered TALEN-induced modifications. Consistent with a previous report[3], CCR5A or ATM TALENs containing ELD/KKR and homodimeric *Fok*I domains demonstrated increased on-target activity compared to EL/KK *Fok*I domains. Of the 45 CCR5A off-target sites that we successfully amplified, we identified nine off-target sites with TALEN-induced modifications; likewise, of the 31 *ATM* off-

target sites that we successfully amplified, we observed seven off-target sites with TALEN-induced modifications (**Table 1.13**).

A

Site	No TALEN	CCR5A EL/KK *Fok*I	CCR5A ELD/KKR *Fok*I	CCR5A Homo *Fok*I
OnCCR5A	<0.006%	9.8%	28%	47%
OffC-5	<0.006%	0.53%	2.3%	2.3%
OffC-15	<0.020%	<0.014%	0.23%	0.043%
OffC-16	<0.006%	<0.006%	0.031%	<0.006%
OffC-28	<0.009%	0.014%	0.16%	0.056%
OffC-36	<0.006%	<0.006%	0.15%	0.028%
OffC-38	<0.006%	ND	ND	0.067%
OffC-49	<0.006%	ND	ND	0.110%
OffC-69	<0.010%	ND	ND	0.089%
OffC-76	<0.006%	ND	ND	0.149%

B

Site	No TALEN	ATM EL/KK *Fok*I	ATM ELD/KKR *Fok*I	ATM Homo *Fok*I
OnATM	0.007%	6.8%	16%	18%
OffA-1	<0.006%	<0.006%	0.026%	0.077%
OffA-11	<0.006%	<0.006%	0.036%	0.39%
OffA-13	<0.006%	0.008%	0.025%	<0.006%
OffA-16	<0.006%	<0.006%	<0.006%	0.057%
OffA-17	<0.051%	<0.14%	<0.17%	0.94%
OffA-23	0.018%	<0.006%	0.29%	0.23%
OffA-35	<0.006%	<0.006%	<0.006%	0.070%

Table 1.13. Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites. (A) For cells treated with either no TALEN or CCR5A TALENs containing heterodimeric EL/KK, heterodimeric ELD/KKR, or the homodimeric (Homo) *Fok*I variants, cellular modification rates are shown as the percentage of observed insertions or deletions (indels) consistent with TALEN cleavage relative to the total number of sequences for on-target (On) and predicted off-target sites (Off). See the main text for details. ND refers to no data collected since the cellular modification of off-target sites OffC-38, OffC-49, OffC-69 and OffC-76 was not assayed for CCR5A TALENs containing EL/KK and ELD/KKR FokI domains. (B) Same as (A) for ATM TALENs.

The inspection of modified on-target and off-target sites yielded a prevalence of deletions ranging from three to dozens of base pairs (**Figure 1.11**), consistent with previously described

characteristics of TALEN-induced genomic modification.[36]

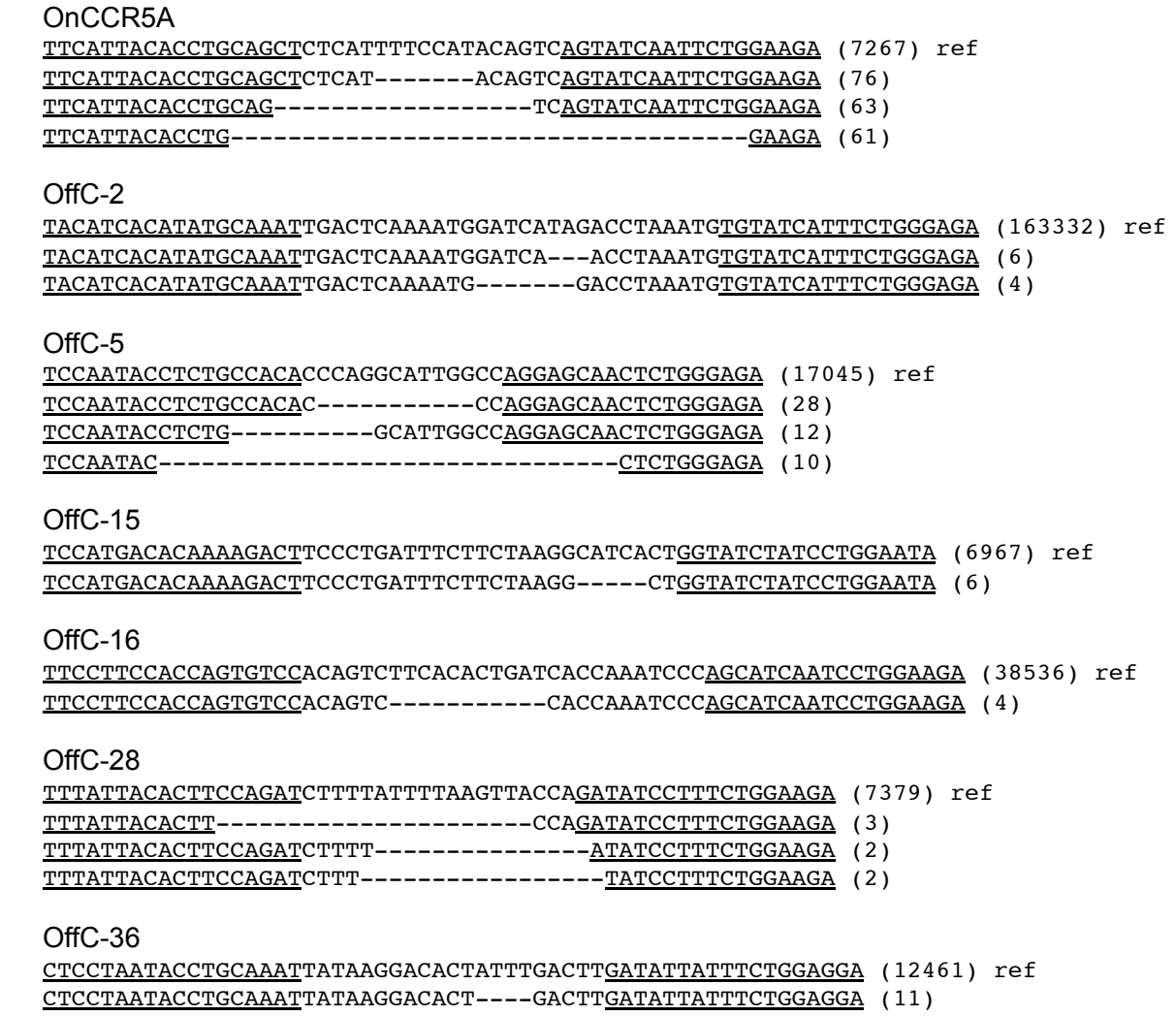


Figure 1.11. Modifications induced by TALENs at on-target and predicted off-target genomic sites. Examples of modified sequences at the on-target site and off-target sites for cells treated with CCR5A TALENs containing the ELD/KKR *Fok*I domains. For each example shown, the unmodified reference genomic site (ref) is the first sequence, followed by the top three sequences containing deletions. The numbers in parentheses indicate sequencing counts and the half-sites are underlined and bolded.

These results collectively indicate *in vitro* selection data processed through a machine-learning algorithm, can predict bona fide off-target substrates that undergo TALEN-induced modification in human cells.

To better understand the likelihood of detecting genomic off-target sites using other methods, three previously reported TALENs targeting separate sites in the human HDAC1,

PMS2, and SDHD genes[12] were constructed with ELD/KKR *Fok*I domains. These TALENs were chosen because their target sites are only three to five mutations away from potential off-target sites in the human genome (**Table 1.14**).

Site	Mut.	Left half-site	Spacer length	Right half-site
OnHDAC	0	TGGCGCAGACGCAGGGC	17	TTACTACTACGACGGTGA
OffHDAC-1	3	TGaCGCAGACaCAGGGC	17	TTACTACTACGACGGgGA
OnPMS	0	TCGGGTGTTGCATCCATG	18	AGGTGAGCGGGGCTCGCA
OffPMS-1	4	TCcGGTGTTGCATCCtTG	18	AGGTGAGCtGGGCTCGCg
OffPMS-2	4	TCcGGTGTTGCATCCtTG	18	AGGTGAGCtGGGCTCGCg
OnSDHD	0	TCAGGAACGAGATGGCGG	17	GCCGTTTGCGGTGCCCTA
OffSDHD-1	1	TCAGGAACGAGATGGCGG	17	GCCGTTTGCGGTGCCCaA
OffSDHD-3	3	TCAGGAACGAGATGGCGG	17	GCCcTTTGCaGTGCCCaA
OffSDHD-4	3	TCAGGAACGAGATGGCGG	17	GCCcTTTaCGGTGCCCaA
OffSDHD-5	5	TCAGGAAtGAGATGGCGG	17	aCCcTTcGCGGTGCCCaA

Table 1.14. Predicted HDAC1, PMS2, and SDHD off-target sites in the human genome. List of genomic on-target and off-targets sites with one to five mutations from the on-target site of the HDAC1, PMS2, and SDHD TALENs. Sites were identified by a simple computational search for off-target sites closely related to the target site in the human genome. Sites are shown with mutation numbers, left and right half-site sequences (with mutations from on-target in lower case), and the length of the spacer between half-sites in base pairs.

The HDAC1, PMS2, and SDHD putative off-target and on-target sites were amplified from genomic DNA isolated from cells expressing the appropriate TALENs. Since only three out of six of these more closely related off-target sites containing three to five mutations were significantly modified (**Table 1.15**), it is likely that more distant off-target sites with 8 to 12 mutations would be very difficult to predict and detect, consistent with reports[14-16, 38] that fail to identify TALEN-induced off-target site modification in cells.

Site	No TALEN	HDAC1 TALEN
OnHDAC	< 0.001%	5.22%
OffHDAC-1	< 0.001%	0.052%

Site	No TALEN	SHDD TALEN
OnSDHD	< 0.001%	33.10%
OffSDHD-1	< 0.001%	0.246%
OffSDHD-2	< 0.001%	< 0.001%
OffSDHD-3	< 0.001%	0.002%
OffSDHD-4	< 0.001%	< 0.001%

Site	No TALEN	PMS2 TALEN
OnPMS	0.006%	20.36%
OffPMS-1	< 0.001%	1.439%
OffPMS-2	0.002%	3.930%

Table 1.15. Cellular modification induced by HDAC1, PMS2, and SDHD TALENs at on-target and predicted off-target genomic sites. (A) For cells treated with either no TALEN or HDAC1, PMS2, or SDHD TALENs containing heterodimeric ELD/KKR *Fok*I variants, cellular modification rates are shown as the percentage of observed insertions or deletions (indels) consistent with TALEN cleavage relative to the total number of sequences for on-target (On) and predicted off-target sites (Off). See the main text for details.

Thus, the challenge of identifying bona fide genomic off-target sites is compounded by the presence of more than 10,000 potential genomic off-target sites containing 8 to 12 mismatches out of 36 recognized bases revealed by both computational identification and statistical modeling (**Table 1.11**).

To compare genomic off-target site prediction by our classifier to a purely computational approach, we used a recently developed *ab initio* genomic off-target site prediction algorithm, TALENoffer,[37] to identify the 32 best-scoring genomic off-target sites for the CCR5A TALEN. These TALENoffer-predicted off-target sites were amplified from genomic DNA purified from cells expressing the CCR5A TALEN with homodimeric FokI domains, and assayed for cellular modification (**Table 1.16**).

Site	Score	Mut.	Left half-site	Spacer length	Right half-site	Classifier
CP_CCRoff-1	-1.503	9	aTaATTACACCTGCAaCT	24	AGTgTCAgTTtgGtAAGt	
CP_CCRoff-2	-1.524	9	TTCATTACAaaTaCAGaa	23	AGaATttATTCTGtAAGA	
CP_CCRoff-3	-1.53	10	TTCtTTAaACCTaaAaCT	21	AtTATtAtTTtTGGAAGt	
CP_CCRoff-4	-1.534	8	TTCATTACtCCTcCttCT	30	ctTATCAcTTtTGGAAGA	OffC-19
CP_CCRoff-5	-1.537	8	TTttTTACtCCTGtAaCa	21	AGTcTCAATTtTGGAAGA	
CP_CCRoff-6	-1.571	9	TTCATTcaACCTtCAaCT	21	AcagTgAATTtTGGAAGA	
CP_CCRoff-7	-1.583	9	TcaATTAtACCTaCAtaT	24	AGTtTCAATTtaGGAAGA	OffC-42
CP_CCRoff-8	-1.583	9	gcCAcTgCAaCTcCAGCc	30	AGTATCAtTTtTGGAAGA	
CP_CCRoff-9	-1.594	9	TaCATcACAtaTGCAaaT	29	tGTATCAtTTCTGGgAGA	OffC-1
CP_CCRoff-10	-1.594	9	TaCATcACAtaTGCAaaT	29	tGTATCAtTTCTGGgAGA	OffC-2
CP_CCRoff-11	-1.594	9	TaCATcACAtaTGCAaaT	29	tGTATCAtTTCTGGgAGA	OffC-3
CP_CCRoff-12	-1.594	9	TaCATTcCACCTaCttCc	22	AGTtTgAcTTCTGGAAGA	
CP_CCRoff-13	-1.606	11	TcCcTTcCACCcaCAaCT	26	AGTATtAgTTCTGGggtA	
CP_CCRoff-14	-1.606	9	TTCATTtCAtCTcaAaCT	13	AaTgTCAtTTCTGtAAGA	
CP_CCRoff-15	-1.611	9	TcCccTACctCTGCAGCT	28	AGTgTCAcTTCTGGgAGg	
CP_CCRoff-16	-1.611	8	TTCATcttcCCTGCAGCg	27	AGaATCAAaTCTGtAAGA	
CP_CCRoff-17	-1.617	9	TTCATTACAaaTGCAGta	29	tGTATgAATTtgaGAAGA	
CP_CCRoff-18	-1.62	8	TTtATTACACtTcCAGaT	19	gaTATCctTTCTGGAAGA	OffC-28
CP_CCRoff-19	-1.621	10	TTCtaTAaAaCTcCAaCT	11	AGTATgAATTCTttAAtA	
CP_CCRoff-20	-1.622	9	TTCtTTACACCTcCAGCT	22	AaaATCAAagtTtGAtGA	
CP_CCRoff-21	-1.635	10	TTCATTAtcaCTcCAaCT	10	AtTAggAgTcCTGGAAGA	
CP_CCRoff-22	-1.637	10	TTCtcTACACCTGaAaCc	12	AGcATtgtgTCTGGAAGA	
CP_CCRoff-23	-1.638	8	TTaATTAaACCTGCAGtT	24	tacATtAATTCTGGAAaA	
CP_CCRoff-24	-1.65	11	TTCATTACAaCcaaAatT	29	AGTAatAATgtTGGgAGA	
CP_CCRoff-25	-1.651	10	TTCcTTACcCCTGCAtCT	21	tGTtTCAtTTtTttgAGA	
CP_CCRoff-26	-1.653	8	TggAcTgCACCTGCAGCT	23	AGctTtgATTCTGGAAGA	
CP_CCRoff-27	-1.653	11	TTCATTtCACaTaCAcCc	20	AtTATgAAgTtTtGAtGA	
CP_CCRoff-28	-1.654	10	TTCATTAaACaTGaAaCT	16	AGTATgAgcTCatGgAGA	
CP_CCRoff-29	-1.655	9	TTCAcTACACCTGCAGCc	26	AtaATgAcTTCTGGctGg	
CP_CCRoff-30	-1.655	10	TTaATTcCACCTGCAGCT	29	gGTcaCAggTtTtGgAGA	
CP_CCRoff-31	-1.659	8	TTCtaaACACCTGtAGCT	27	AGgAaCAATaCTGGAtGA	
CP_CCRoff-32	-1.662	10	TTCATTtgACCTcCcaCT	17	tGTgggAATTCTGGgAGA	

Table 1.16. Purely computationally predicted off-target sites in the human genome using TALENoffer program. Computationally predicted genomic off-target sites from TALENoffer program[9] allowing for spacer lengths of 10 to 30 base pairs and only searching for heterodimeric off-target sites. The resulting 32 predicted off-targets sites with the best scores for the CCR5A TALENs are shown with their respective classifier scores, mutation numbers, left and right half-site sequences (with mutations from on-target in lower case), and the length of the spacer between half-sites in base pairs.

Since the TALENoffer sites were only eight to 11 mutations from the on-target site, they were on average less distant from the on-target sequence than the off-target sites predicted by our

classifier (9.2 versus 11.1 mean mutations in the off-target sites from the TALENoffer versus from our classifier, respectively). Therefore, from our classifier prediction of genomic off-target sites for the CCR5A TALEN, only CCR5A genomic off-target sites containing eight to 11 mutations were considered for comparison to TALENoffer sites. Both sets of sites were assayed for cellular modification by the CCR5A TALEN with homodimeric FokI domains (**Table 1.17 and Table 1.18**).

C-terminal domain:		No TALEN		Canonical			
FokI domain:		No TALEN		Homo			
TALENoffer site	Classifier site	Indels	Total	Indels	Total	% Modified	P-value
OnCCR5A		0	9997	3006	9773	30.758%	< 1.0E-250
CP_CCRoff-1		0	9135	2	17468	0.011%	
CP_CCRoff-2		0	63390	0	36666	< 0.006%	
CP_CCRoff-3		14	11594	7	10460	0.067%	
CP_CCRoff-4	OffC-19	0	28732	0	26314	< 0.006%	
CP_CCRoff-5		0	21413	0	21538	< 0.006%	
CP_CCRoff-6		0	22764	0	28279	< 0.006%	
CP_CCRoff-8		2	24054	0	25702	< 0.006%	
CP_CCRoff-9	OffC-1	1	68283	6	59642	0.010%	
CP_CCRoff-10	OffC-2	0	68421	4	59558	0.007%	
CP_CCRoff-11	OffC-3	0	68421	4	59558	0.007%	
CP_CCRoff-12		0	14364	0	13806	< 0.007%	
CP_CCRoff-13		0	15016	0	24585	< 0.006%	
CP_CCRoff-14		0	28025	1	27546	< 0.006%	
CP_CCRoff-15		0	16105	8	13019	0.061%	1.59E-03
CP_CCRoff-16		0	26453	1	29619	< 0.006%	
CP_CCRoff-17		0	21155	1	28839	< 0.006%	
CP_CCRoff-18	OffC-28	0	28111	52	12591	0.413%	2.34E-19
CP_CCRoff-19		0	35891	0	33962	< 0.006%	
CP_CCRoff-20		0	64345	1	118954	< 0.006%	
CP_CCRoff-21		0	14857	0	11150	< 0.009%	
CP_CCRoff-22		0	14368	0	37008	< 0.006%	
CP_CCRoff-23		3	22876	7	20671	0.034%	
CP_CCRoff-24		4	129051	1	50695	< 0.006%	
CP_CCRoff-25		0	0	1	39845	< 0.006%	
CP_CCRoff-26		0	21677	0	24695	< 0.006%	
CP_CCRoff-27		0	24710	0	43452	< 0.006%	

Table 1.16. Cellular modification induced by CCR5A TALENs at on-target and predicted off-target genomic sites generated by TALENoffer.

CP_CCRoff-28	0	10269	0	11496	< 0.009%
CP_CCRoff-29	3	121960	2	100659	< 0.006%
CP_CCRoff-30	0	18320	0	12945	< 0.008%
CP_CCRoff-31	0	74541	1	89994	< 0.006%
CP_CCRoff-32	0	29550	1	102087	< 0.006%

Table 1.16 (Continued). Cellular modification induced by CCR5A TALENs at on-target and predicted off-target genomic sites generated by TALENoffer. Results from sequencing CCR5A on-target and each genomic off-target site predicted by TALENoffer that amplified from 50 ng genomic DNA isolated from human cells treated with either no TALEN or TALENs containing canonical C-terminal domains and homodimeric (Homo) *Fok*I domains. Indels: the number of observed sequences containing insertions or deletions consistent with TALEN-induced cleavage. Total: total number of sequence counts. Modified: number of indels divided by total number of sequences, expressed as percentages. Upper limits of potential modification were calculated for sites with no observed indels by assuming there is less than one indel then dividing by the total sequence count to arrive at an upper limit modification percentage, or taking the theoretical limit of detection (1/16,400), whichever value was larger. P-values: calculated as previously reported[15] using a (right) one-sided Fisher's exact test between each canonical C-terminal domain TALEN-treated sample and the untreated control sample. P-values of < 0.00161 were considered significant and are shown. The significance cut off of 0.00161 was based on the multiple comparison correction from the Benjamini-Hochberg method.[16, 17] Specificity is the ratio of on-target to off-target genomic modification frequency for each site.

C-terminal domain:	No TALEN			Canonical		
FokI domain:	No TALEN			Homo		
Classifier site	Indels	Total	Indels	Total	% Modified	P-value
OffC-1	0	51248	1	87496	< 0.006%	
OffC-2	6	124356	0	163322	< 0.006%	
OffC-3	6	124356	0	163322	< 0.006%	
OffC-4	0	45377	0	40452	< 0.006%	
OffC-5	0	27009	395	17077	2.313%	1.20E-164
OffC-6	0	10766	0	6560	< 0.015%	
OffC-9	0	40603	0	30771	< 0.006%	
OffC-13	0	65518	0	78546	< 0.006%	
OffC-16	0	36228	0	32636	< 0.006%	
OffC-17	0	32112	0	31299	< 0.006%	
OffC-19	1	22868	0	28478	< 0.006%	
OffC-20	0	23335	0	18972	< 0.006%	
OffC-21	0	34302	0	21161	< 0.006%	
OffC-22	1	81037	0	104857	< 0.006%	
OffC-28	0	28111	52	12591	0.056%	2.34E-19
OffC-30	0	11840	0	6285	< 0.015%	
OffC-32	0	1944	0	19115	< 0.006%	
OffC-34	0	9052	0	9072	< 0.011%	
OffC-35	0	23839	0	11897	< 0.008%	
OffC-36	1	23412	5	18052	0.028%	
OffC-38	0	16396	9	13351	0.067%	7.38E-04
OffC-39	0	51962	0	13562	< 0.007%	
OffC-40	0	3910	0	24711	< 0.006%	
OffC-49	0	26333	27	24497	0.110%	2.74E-09
OffC-56	0	25357	0	26999	< 0.006%	
OffC-65	0	14364	0	13806	< 0.007%	
OffC-69	0	10407	25	27950	0.089%	4.90E-04
OffC-76	0	61760	59	39617	0.149%	8.19E-25
OffC-137	0	26470	0	23318	< 0.006%	
OffC-150	0	22058	5	20952	0.024%	

Table 1.17 (Continued). Cellular modification induced by CCR5A TALENs at on-target and predicted off-target genomic sites generated by the classifier trained on selection results.

Table 1.17 (Continued). Cellular modification induced by CCR5A TALENs at on-target and predicted off-target genomic sites generated by the classifier trained on selection results. Results from sequencing CCR5A on-target and each genomic off-target site with eight to 11 mutations predicted by the classifier and the next 10 best-scoring amplified off-target sites with nine or ten mutations (OffC-38 to OffC-150) after the first 36 top-scoring genomic off-target site predicted by the classifer. Sites that amplified from 50 ng genomic DNA isolated from human cells treated with either no TALEN or TALENs containing canonical C-terminal domains and homodimeric (Homo) *Fok*I domains. Indels: the number of observed sequences containing insertions or deletions consistent with TALEN-induced cleavage. Total: total number of sequence counts. Modified: number of indels divided by total number of sequences, expressed as percentages. Upper limits of potential modification were calculated for sites with no observed indels by assuming there is less than one indel then dividing by the total sequence count to arrive at an upper limit modification percentage, or taking the theoretical limit of detection (1/16,400), whichever value was larger. P-values: calculated as previously reported[15] using a (right) one-sided Fisher's exact test between each canonical C-terminal domain TALEN-treated sample and the untreated control sample. P-values < 0.0040 were considered significant and are shown. The significance cut off 0.0040 was based on the multiple comparison correction from the Benjamini-Hochberg method.[16, 17] Specificity is the ratio of on-target to off-target genomic modification frequency for each site.

One genomic off-target site predicted by both TALENoffer and our classifier was modified in cells. Of the remaining 25 amplified genomic off-target sites predicted exclusively by TALENoffer, only one was significantly modified in cells (**Tables 1.16 and 1.17**). In contrast, 5 of the 25 amplified genomic off-target sites exclusively predicted by our classifier were significantly modified in cells (**Table 1.17**). These results indicate that our classifier outperformed purely computational prediction for off-target CCR5A TALEN substrates. We directly compared our combined *in vitro* selection and machine learning method with TALENoffer, a recently described purely computational prediction method[37] and found that our approach outperforms the purely computational approach for the identification of TALEN-induced off-target substrates in cells.

1.4 TALEN specificity as function of array length, interdependence of mismatches and estimation of total genomic TALEN cleavage

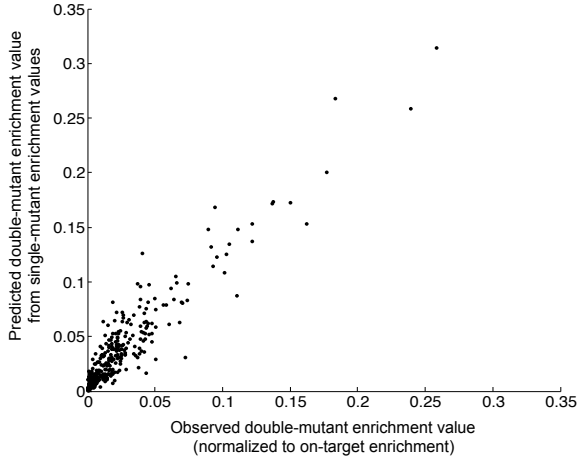
The extensive number of quantitatively characterized off-target substrates in the selection data enabled us to address several key questions about TALEN specificity. First, in order to assess whether mutations at one position in the target sequence affect the ability of TALEN repeats to productively bind other positions, we generated an expected enrichment value for

every possible double-mutant sequence for the L13+R13 CCR5B TALENs assuming independent contributions from the two corresponding single-mutation enrichments. In general, the predicted enrichment values closely resembled the actual observed enrichment values for each double-mutant sequence (**Figure 1.12**), suggesting that component single mutations independently contributed to the overall cleavability of double-mutant sequences.

A

Sequence	Observed enrichment value from selection	Predicted enrichment value assuming independence
TCAT**a**ACACCTGC	0.46	
TCATTACACCTG**t**	0.47	
TCAT**a**ACACCTG**t**	0.25	0.46 x 0.47 = 0.22

B



C

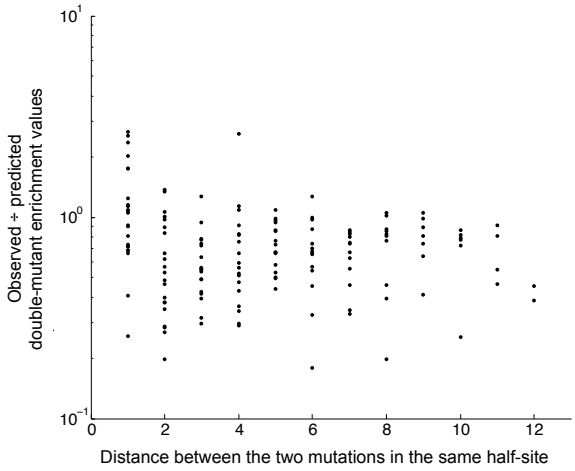


Figure 1.12. Observed versus predicted double-mutant sequence enrichment values. (A) An example calculation of a predicted double-mutant enrichment value from the L13+R13 CCR5A TALEN selection. All enrichment values were normalized to the on-target enrichment value (= 1.0 by definition). Observed enrichment values of each single-mutant are shown and were multiplied together to calculate the predicted enrichment value of the corresponding double-mutant sequence. This predicted enrichment value can then be compared to the observed enrichment value for the same double-mutant sequence. (B) For the L13+R13 CCR5A TALEN selection, the observed double-mutant enrichment values of individual sequences (post-selection sequence abundance ÷ pre-selection sequence abundance) were normalized to the on-target enrichment value (= 1.0 by definition) and plotted against the corresponding predicted double-mutant enrichment values calculated by multiplying the enrichment value of the component single-mutants normalized to the on-target enrichment. The predicted double-mutant enrichment values assume independent contributions from each single mutation to the double-mutant's enrichment value. (C) The observed double-mutant sequence enrichment divided by the predicted double-mutant sequence enrichment plotted as a function of the distance (in base pairs) between the two mutations. Only sequences with two mutations in the same half-site were considered.

The difference between the observed and predicted double-mutant enrichment values was relatively independent of the distance between the two mutations, except that two adjacent mismatches were slightly better tolerated than would be expected (**Figure 1.12**).

To determine the potential interdependence of more than two mutations, we evaluated the relationship between selection enrichment values and the number of mutations in the post-selection target for the L13+R13 CCR5B TALEN (**Figure 1.13**).

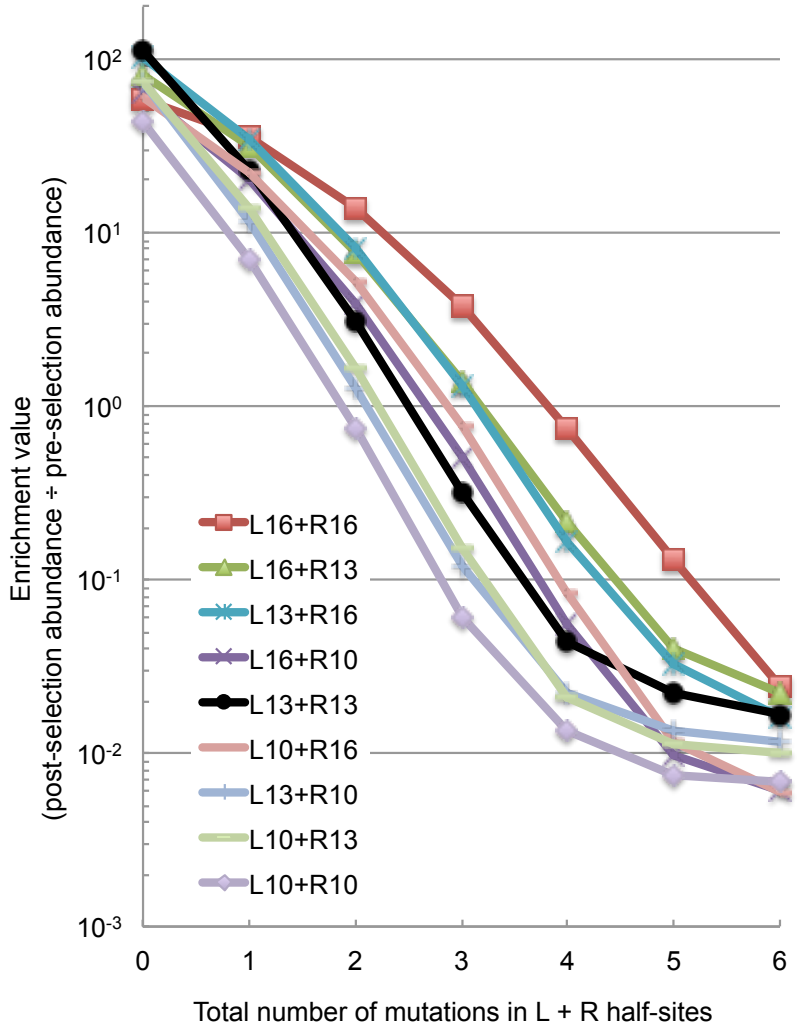


Figure 1.13. ***In vitro*** **specificity as a function of TALEN length**. The enrichment value of on-target (zero mutation) and off-target sequences containing one to six mutations are shown for CCR5B TALENs of varying TALE repeat array lengths with EL/KK *Fok*I domains. The TALENs targeted DNA sites of 32 bp (L16+R16), 29 bp (L16+R13 or L13+R16), 26 bp

(L16+R10 or L13+R13 or L10+R16), 23 bp (L13+R10 or L10+R13) or 20 bp (L10+R10) in length.
For 0 to 5 mutations, enrichment values closely followed a simple exponential function of the mean number of mutations (m) (**Table 1.18**).

TALEN selection	a	b	R^2
L13+R10 CCR5B	1.00	-1.88	0.999937
L10+R10 CCR5B	1.00	-1.85	0.999901
L10+R13 CCR5B	1.00	-1.71	0.999822
L13+R13 CCR5B	1.00	-1.64	0.999771
L13+R16 CCR5B	1.00	-1.15	0.998286
L16+R10 CCR5B	1.00	-1.24	0.998252
L10+R16 CCR5B	1.01	-1.08	0.996343
L16+R13 CCR5B	1.01	-1.04	0.995844
L16+R16 CCR5B	1.03	-0.70	0.977880
L18+R18 ATM	1.08	-0.36	0.913087
L18+R18 CCR5A	1.13	-0.21	0.798923

Table S1.18. Exponential fitting of enrichment values as function of mutation number. Enrichment values of post-selection sequences as function of mutation were normalized relative to on-target enrichment (= 1.0 by definition). Normalized enrichment values of sequences with zero to four mutations were fit to an exponential function, $a \times e^{b}$, with R^2 reported using the non-linear least squares method.

This relationship is consistent with a model in which each successive mutation reduces the binding energy by a constant amount ($\Delta\Delta G$), resulting in an exponential decrease in TALEN binding ($K_{eq}(m)$) such that $K_{eq}(m) \sim e^{\Delta\Delta G * m}$. The observed exponential relationship therefore suggests that the mean reduction in binding energy from a typical mismatch is independent of the number of mismatches already present in the TALEN:DNA interaction. Collectively, these results indicate that TALE repeats bind their respective DNA base pairs independently beyond a slightly increased tolerance for adjacent mismatches.

To characterize the interdependence between the two TALE arrays that comprise a TALEN pair, we calculated the enrichment value for sequences with the same number of mutations in the left half-site and the same number of mutations in the right-half site (**Figure 1.14**).

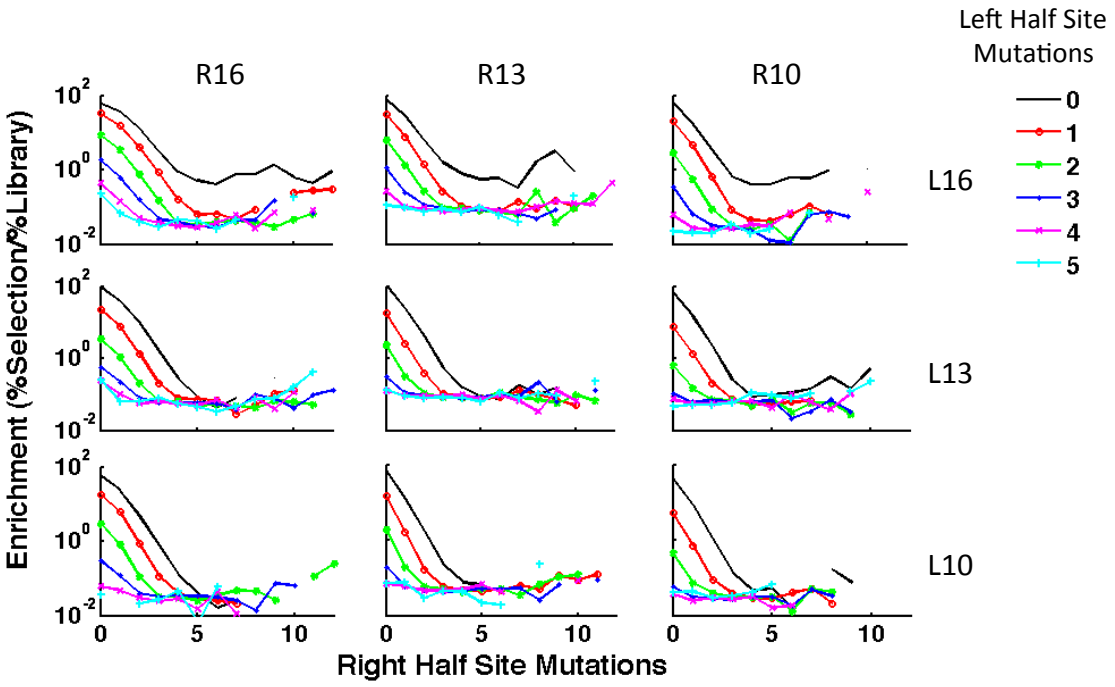


Figure 1.14. ***In vitro*** **specificity as a function of TALEN length by half-sites**. The enrichment value of on-target (zero mutation) and off-target sequences containing one to ten mutations are shown for CCR5B TALENs of varying TALE repeat array lengths with EL/KK *Fok*I domains. Enrichment values are for sequences with the identical number of mutations in the right half-site (x-axis) and identical number of mutations in the left-half site (colored lines). The TALENs targeted DNA sites of 32 bp (L16+R16), 29 bp (L16+R13 or L13+R16), 26 bp (L16+R10 or L13+R13 or L10+R16), 23 bp (L13+R10 or L10+R13) or 20 bp (L10+R10) in length.

In general, the enrichment value drops per mutation of the left half-sites closely resembled the enrichment values decreases per mutation of the right half-site. Thus a regular drop in enrichment is observed regardless of whether a mutation occurs in the left-half site or right-half site and regardless of whether there is already a mutation in the same or other half-site. This suggests that component mutations in either half-site independently contribute to the overall cleavability of a sequence.

Of note, a similar analysis of considering the number of mutations in each half-site (as opposed to the total in both half-sites combined) was used to demonstrate that the selection methodology cannot be used to effectively profile the specificity of TALENs containing homodimeric *Fok*I domains (**Figure 1.15**). Because sequences with highly on-target half-sites become enriched with little effect from two or mutations in the other-half site beyond, it is likely that two identical TALENs bind to repeats of a highly on-target site on the same molecule of

concatmeric half-sites. They are then able to bridge the entire spacer, constant sequence and right half-site to dimerize and cleave.

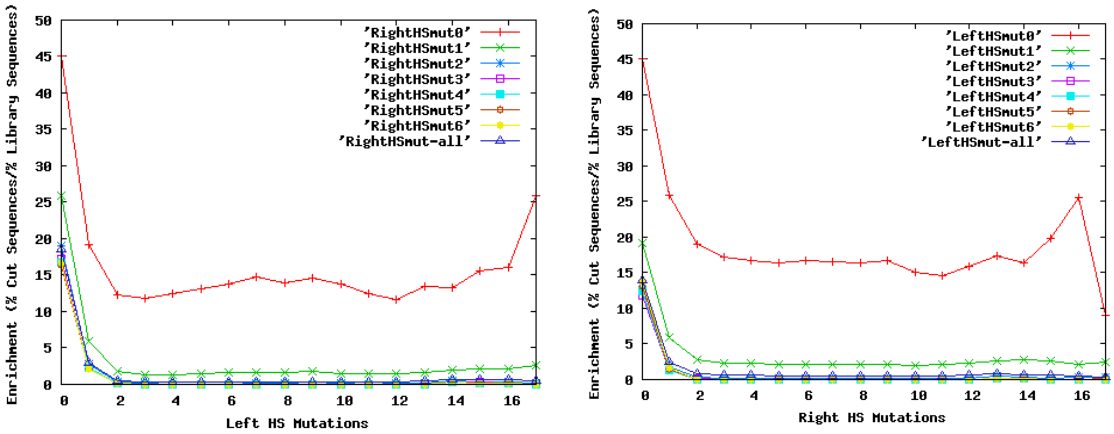


Figure 1.15. *In vitro* specificity as a function of TALEN with homodimeric *Fok*I domains by half-sites. The enrichment value of on-target (zero mutation) and off-target sequences containing one to ten mutations are shown for CCR5B TALENs of varying TALE repeat array lengths with homodimeric *Fok*I domains. Enrichment values are for sequences with the identical number of mutations in the right half-site (x-axis) and identical number of mutations in the left-half site (colored lines). The TALENs targeted DNA sites of 32 bp (L16+R16), 29 bp (L16+R13 or L13+R16), 26 bp (L16+R10 or L13+R13 or L10+R16), 23 bp (L13+R10 or L10+R13) or 20 bp (L10+R10) in length.

The independent binding of TALE repeats simplistically predicts that TALEN specificity per base pair is independent of target-site length. To experimentally characterize the relationship between TALE array length and off-target cleavage, we constructed TALENs targeting 10, 13, or 16 bps (including the 5' T) for both the left (L10, L13, L16) and right (R10, R13, R16) half-sites. TALENs representing all nine possible combinations of left and right CCR5B TALENs were subjected to *in vitro* selection. The results revealed that shorter TALENs have greater specificity per targeted base pair than longer TALENs (**Table 1.6**). For example, sequences cleaved by the L10+R10 TALEN contained a mean of 0.032 mutations per recognized base pair, while those cleaved by the L16+R16 TALEN contained a mean of 0.067 mutations per recognized base pair.

For selections with the longest CCR5B TALENs targeting 16+16 base pairs or CCR5A and ATM TALENs targeting 18+18 bp, the mean selection enrichment values do not follow a simple exponential decrease as function of mutation number (**Figure 1.13 and Table 1.18**). It is possible these TALENs have greater affinity than is required to substantially bind and cleave the target site (referred to below as "excess DNA-binding energy"). Thus, we hypothesize that excess DNA-binding energy from the larger number of TALE repeats in longer TALENs reduces

specificity by enabling the cleavage of sequences with more mutations, without a corresponding increase in the cleavage of sequences with fewer mutations, because the latter are already nearly completely cleaved. Indeed, the *in vitro* cleavage efficiencies of discrete DNA sequences for these longer TALENs are independent of the presence of a small number of mutations in the target site (**Figure 1.22**), suggesting there is nearly complete binding and cleavage of sequences containing few mutations. Likewise, higher TALEN concentrations also result in decreased enrichment values of sequences with few mutations while increasing the enrichment values of sequences with many mutations (**Tables 1.7 and 1.8**). These results together support a model in which excessive TALEN binding arising from either long TALE arrays or high TALEN concentrations decreases the observed TALEN DNA cleavage specificity for each recognized base pair.

Although longer TALENs are more tolerant of mismatched sequences (**Figure 1.13 and Table 1.6**) than shorter TALENs, in the human genome there are far fewer closely related off-target sites for a longer target site than for a shorter target site (**Figure 1.16**).

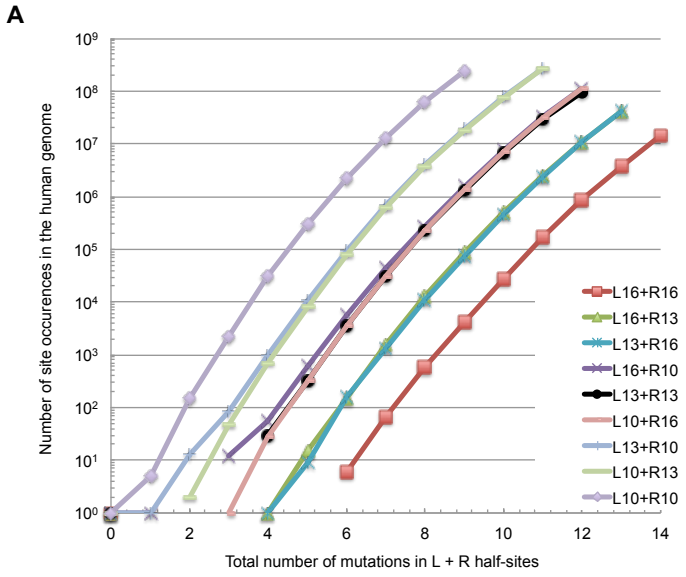


Figure 1.16. Predicted off-target genomic cleavage as a function of TALEN length considering both TALEN specificity and off-target site abundance in the human genome

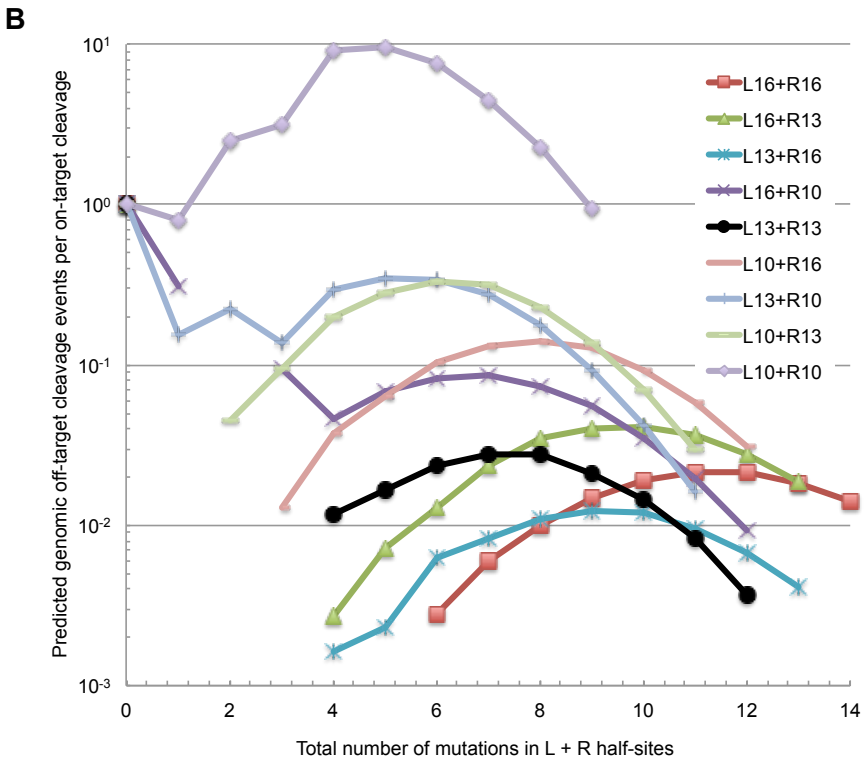


Figure 1.16 (Continued). Predicted off-target genomic cleavage as a function of TALEN length considering both TALEN specificity and off-target site abundance in the human genome. (A) Number of sites in the human genome related to each of the nine CCR5B on-target sequences (L10, L13, or L16 combined with R10, R13, or R16), allowing for a spacer length from 12 to 25 bps between the two half-sites (Supplementary Algorithm). The TALENs targeted DNA sites of 32 bp (L16+R16), 29 bp (L16+R13 or L13+R16), 26 bp (L16+R10 or L13+R13 or L10+R16), 23 bp (L13+R10 or L10+R13) or 20 bp (L10+R10) in length (B) For all nine CCR5B TALENs, overall genomic off-target cleavage frequency was predicted by multiplying the number of sites in the human genome containing a certain number of mutations by the enrichment value of off-target sequences containing that same number of mutations shown in (**Figure 1.13**). Because enrichment values level off at high mutation numbers likely due to the limit of sensitivity of the selection, it was necessary to extrapolate high-mutation enrichment values by fitting low-mutation enrichment values as function of mutation number. The overall predicted genomic cleavage was calculated only for mutation numbers with sites observed to occur more than once in the human genome. For L16+R10 there are no genomic sequences with two mutations, causing the break in the corresponding line.

Since off-target site abundance and cleavage efficiency both contribute to the number of off-target cleavage events in a genomic context, we calculated overall genome cleavage specificity as a function of TALEN length by multiplying the extrapolated mean enrichment value of mutant sequences of a given length with the number of corresponding mutant sequences in the human genome (we note that this estimation assumes that extrapolated mean enrichments of highly

mutant off-target substrates correlate with cleavage rates in cells). The decrease in potential off-target site abundance resulting from the longer target site length is large enough to outweigh the decrease in specificity per recognized base pair observed for longer TALENs (**Figure 1.16**). As a result, longer TALENs are predicted to be more specific against the set of potential cleavage sites in the human genome than shorter TALENs for the tested TALEN pairs targeting a total of 20 to 32 base pairs. Thus, despite being less specific per base pair, TALENs designed to cleave longer target sites are estimated to haves higher overall specificity than those that target shorter sites when considering the number of potential off-target sites in the human genome.

1.5 Engineering and profiling TALENs with improved specificity

The findings above suggest that TALEN specificity could be improved by reducing non-specific DNA binding energy to only what is needed to support efficient on-target cleavage. The most widely used 63-aa C-terminal domain between the TALE repeat array and the *Fok*I nuclease domain contains ten cationic residues.[4, 5, 7, 9, 10, 12] A related C-terminal domain variant (89% homology), containing 11 cationic residues, has also been used in other studies.[6, 19, 34] We hypothesized that reducing the cationic charge of the canonical 63-aa TALE C-terminal domain would decrease non-specific DNA binding[39] and improve the specificity of TALENs.

We constructed two C-terminal domain variants in which three ("Q3", consisting of K788Q, R792Q, and R801Q) or seven ("Q7", consisting of K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, and R801Q) cationic Arg or Lys residues in the canonical 63-aa C-terminal domain were mutated to Gln (**Figure 1.17**).

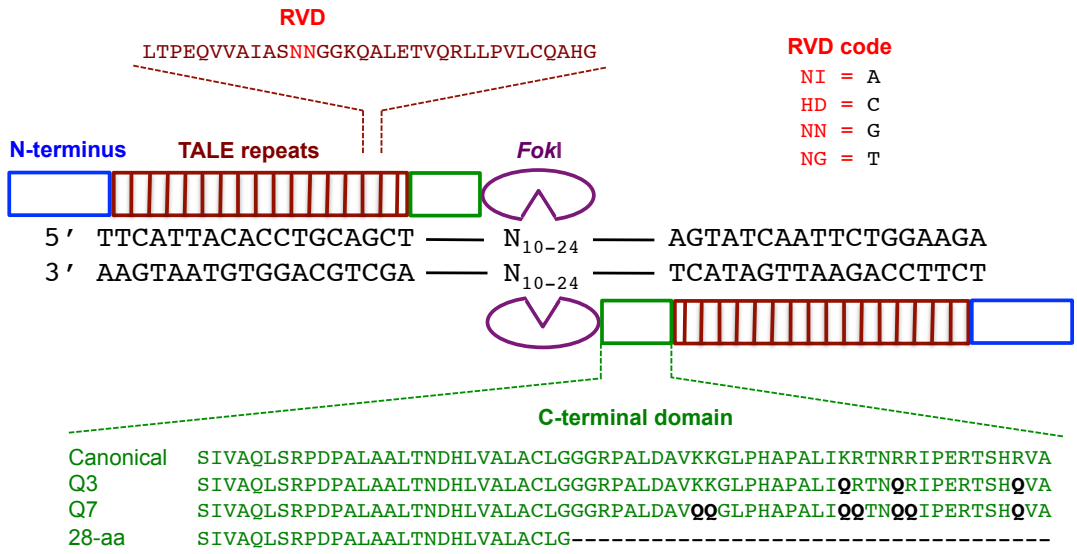


Figure 1.17. TALEN architecture. The C-terminal domain variants used in this study are shown in green with mutations in black.

We performed *in vitro* selections on CCR5A and ATM TALENs containing the canonical 63-aa, the engineered Q3, and the engineered Q7 C-terminal domains, as well as a previously reported 28-aa truncated C-terminal domain[5] with a theoretical net charge (-1) identical to that of the Q7 C-terminal domain. The on-target sequence enrichment values for the CCR5A and ATM selections increased substantially as the net charge of the C-terminal domain decreased (**Figure 1.18**).

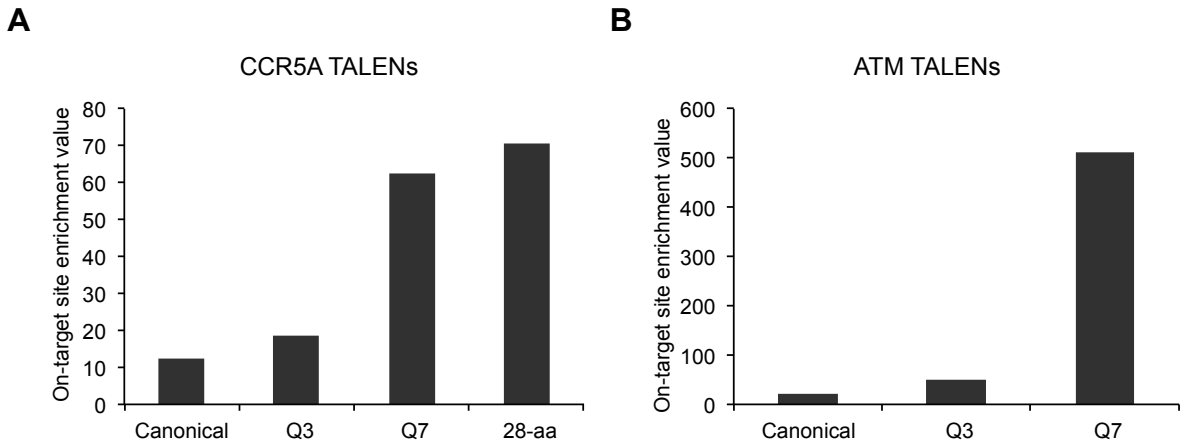


Figure 1.18. *In vitro* specificity and discrete cleavage efficiencies of TALENs containing canonical or engineered C-terminal domains. On-target enrichment values for selections of (A) CCR5A TALENs containing canonical, Q3, Q7, or 28-aa C-terminal domains with EL/KK *Fok*I domains or (B) ATM TALENs containing canonical, Q3 or Q7 C-terminal domains with EL/KK *Fok*I domains.

For example, the ATM selections resulted in on-target enrichment values of 510, 50, and 20 for the Q7, Q3, and canonical 63-aa C-terminal variants, respectively. These results suggest that the TALEN variants in which cationic residues in the C-terminal domain have been partially replaced by neutral residues or completely removed are substantially more specific *in vitro* than the TALENs that contain the canonical, commonly used 63-aa C-terminal domain.

The model of TALEN binding and specificity described in the main text predicts that reducing excess TALEN binding energy will increase TALEN DNA cleavage specificity. To further test this prediction and potentially further augment TALEN specificity, we mutated one ("N1", K150Q), two ("N2", K150Q and K153Q), or three ("N3", K150Q, K153Q, and R154Q) Lys or Arg residues to Gln in the N-terminal domain of TALENs targeting CCR5A and ATM. These N-terminal residues have been shown in previous studies to bind non-specifically to DNA, and mutations at these specific residues to neutralize the cationic charge decrease non-specific DNA binding energy.[40] We hypothesized the reduction in non-specific binding energy from these N-terminal mutations would decrease excess TALEN binding energy resulting in increased specificity. *In vitro* selections on these three TALEN variants revealed that the less cationic N-terminal TALENs indeed exhibit greater enrichment values of on-target cleavage (**Figures 1.8 and 1.9 and Tables 1.8 and 1.9**).

All TALEN constructs tested specifically recognize the intended base pair across both half-sites, except that some of the ATM TALENs do not specifically interact with the base pair adjacent to the spacer (targeted by the most C-terminal TALE repeat) (**Figures 1.8 -1.10**). To compare the broad specificity profiles of canonical TALENs with those containing engineered C-terminal or N-terminal domains, the specificity scores of each target base pair from selections using CCR5A and ATM TALENs with the canonical, Q3, or Q7 C-terminal domains and N1, N2, or N3 N-terminal domains were subtracted by the corresponding specificity scores from selections on the canonical TALEN (canonical 63-aa C-terminal domain, wild-type N-terminal domain) (**Figure 1.19**).

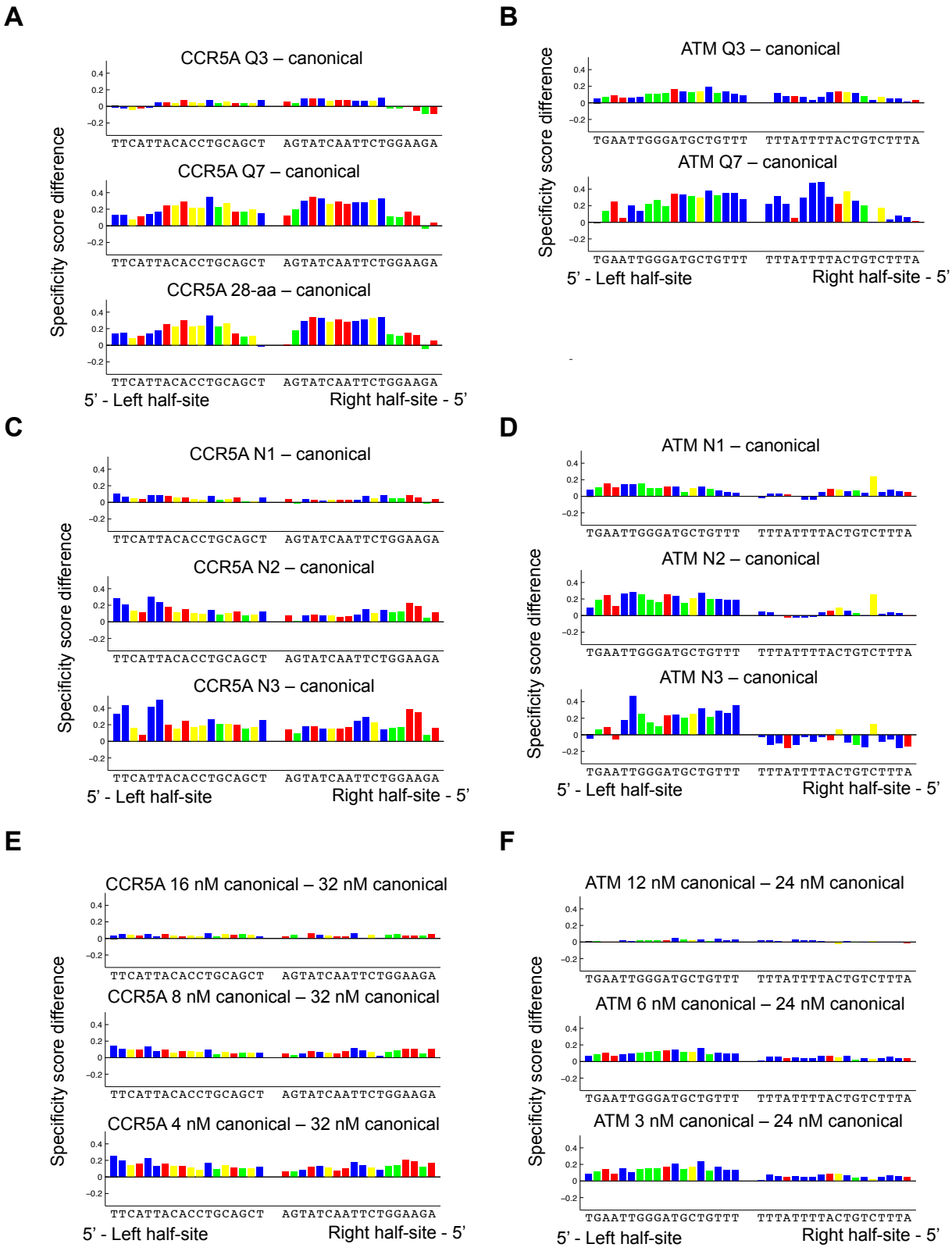


Figure 1.19. Effects of engineered TALEN domains and TALEN concentration on specificity.

Figure 1.19 (Continued). Effects of engineered TALEN domains and TALEN concentration on specificity. (A) The specificity score of the targeted base pair at each position of the CCR5A site was calculated for CCR5A TALENs containing the canonical, Q3, Q7, or 28-aa C-terminal domains. The specificity scores of the Q3, Q7, or 28-aa C-terminal domain TALENs subtracted by the specificity scores of the TALEN with the canonical C-terminal domain are shown. (B) Same as (A) but for CCR5A TALENs containing engineered N-terminal domains N1, N2, or N3. (C) Same as (A) but comparing specificity scores differences of the canonical CCR5A TALEN assayed at 16 nM, 8 nM or 4 nM subtracted by the specificity scores of canonical CCR5A TALENs assayed at 32 nM. (D-F) Same as (A-C) but for ATM TALENs.

Mutations in the C-terminal domain that increase specificity did so most strongly in the middle and at the C-terminal end of each half-site. Likewise, the specificity-increasing mutations in the N-terminus tended to increase specificity most strongly at positions near the TALEN N-terminus (5' DNA end) although mutations in the N-terminus of ATM TALEN targeting the right half-site did not significantly alter specificity. These results are consistent with a local binding compensation model in which weaker binding at either terminus demands increased specificity in the TALE repeats near this terminus. To characterize the effects of TALEN concentration on specificity, the specificity scores from selections of ATM and CCR5A TALENs performed at three different concentrations ranging from 3 nM to 16 nM were each subtracted by the specificity scores of corresponding selections performed at the highest TALEN concentration assayed, 24 nM for ATM, or 32 nM for CCR5A. The results (**Figure 1.19**) indicate that specificity scores increase fairly uniformly across the half-sites as the concentration of TALEN is decreased.

To assess the spacer-length preference of various TALEN architectures (C-terminal mutations, N-terminal mutations, and *Fok*I variants) and various TALEN concentrations, the enrichment values of library members with 10- to 24- base pair spacer lengths in each of the selections with CCR5A and ATM TALEN with various combinations of the canonical, Q3, Q7, or 28-aa C-terminal domains; N1, N2, or N3 N-terminal mutations; and the EL/KK or ELD/KKR *Fok*I variants at 4 nM to 32 nM CCR5A and ATM TALEN were calculated (**Figure 1.20**).

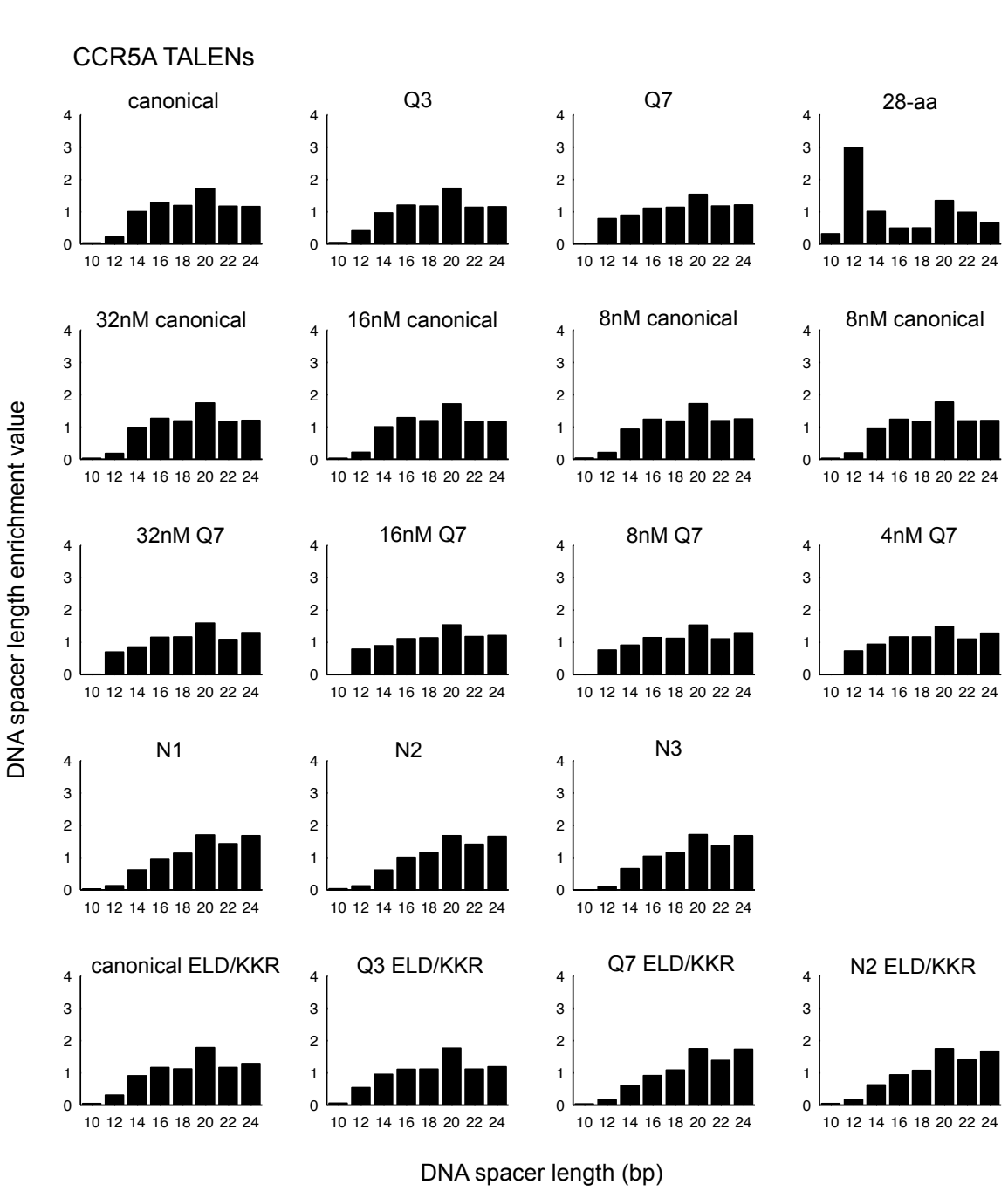


Figure 1.20. Spacer-length preferences of TALENs.

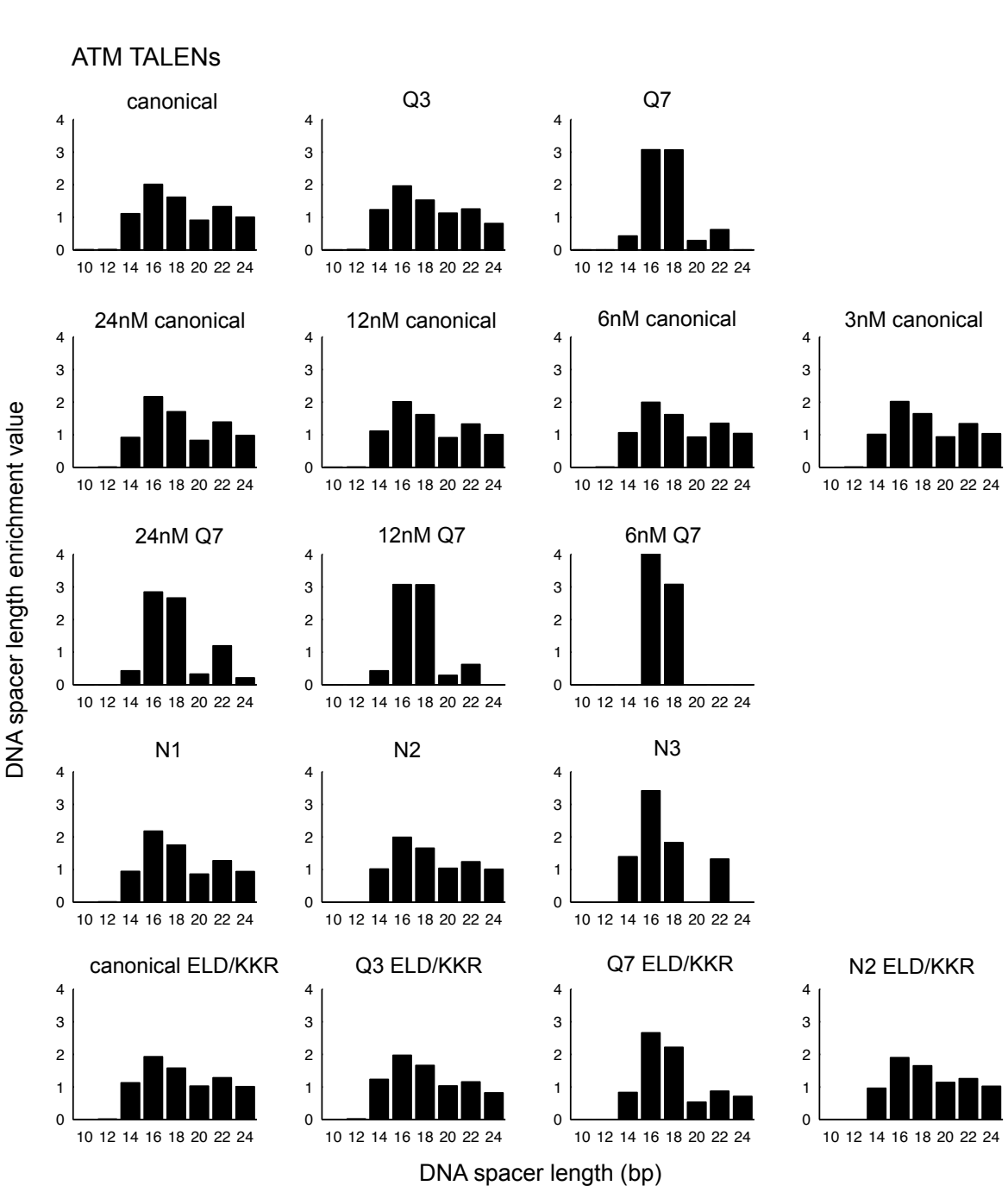


Figure 1.20 (Continued). Spacer-length preferences of TALENs. (A) For each selection with CCR5A TALENs containing various combinations of the canonical, Q3, Q7, or 28-aa C-terminal domains; N1, N2, or N3 N-terminal mutations; and the EL/KK or ELD/KKR *Fok*I variants and at 4, 8, 16, or 32 nM, the DNA spacer-length enrichment values were calculated by dividing the abundance of DNA spacer lengths in post-selection sequences by the abundance of DNA spacer lengths in the pre-selection library sequences. (B) Same as (A) but for ATM TALENs.

All of the tested concentrations, N-terminal variants, C-terminal variants, and *Fok*I variants demonstrated a broad DNA spacer-length preference ranging from 14- to 24- base pairs with three notable exceptions. First, the CCR5A 28-aa C-terminal domain exhibited a much narrower DNA spacer-length preference than the broader DNA spacer-length preference of the canonical C-terminal domain, consistent with previous reports.[5, 14, 41] Second, the CCR5A TALENs containing Q7 C-terminal domains showed an increased tolerance for 12-base spacers compared to the canonical C-terminal domain variant (**Figure 1.20**). This slightly broadened spacer-length preference may reflect greater conformational flexibility in the Q7 C-terminal domain, perhaps resulting from a smaller number of non-specific protein:DNA interactions along the TALEN:DNA interface. Third, the ATM TALENs with Q7 C-terminal domains and the ATM TALENs with N3 mutant N-terminal domains showed a narrowed spacer preference. We speculate that these more specific TALENs with lower DNA-binding affinity may have faster off-rates that are competitive with the rate of cleavage of non-optimal DNA spacer lengths, altering the observed spacer-length preference. While previous reports have focused on the length of the TALEN C-terminal domain as a primary determinant of DNA spacer-length preference, these results suggest the net charge of the C-terminal domain as well as overall DNA-binding affinity can also affect TALEN spacer-length preference. For many of the TALENs assayed, the spacer preferences revealed in the *in vitro* selection results are broader than a previous report by Miller *et al*[5] of a comparably narrower spacer length preference in cells. However, the broader spacer preferences revealed in the *in vitro* selection are consistent with a previous report by Mussolino *et al*.[13] This discrepancy could be explained by higher affinity TALENs, or by a higher concentration of TALENs used in our study and used in the study of Mussolino *et al*., leading to saturation of binding sites that allows cleavage of otherwise non-optimal DNA spacer lengths.

We also characterized the location of TALEN DNA cleavage within the spacer. We created histograms reporting the number of spacer DNA bases observed preceding the right half-site in each of the sequences from the selections with CCR5A and ATM TALEN with various combinations of the canonical, Q3, Q7, or 28-aa C-terminal domains; N1, N2, or N3 N-terminal mutations; and the EL/KK or ELD/KKR *Fok*I variants (**Figure 1.21**).

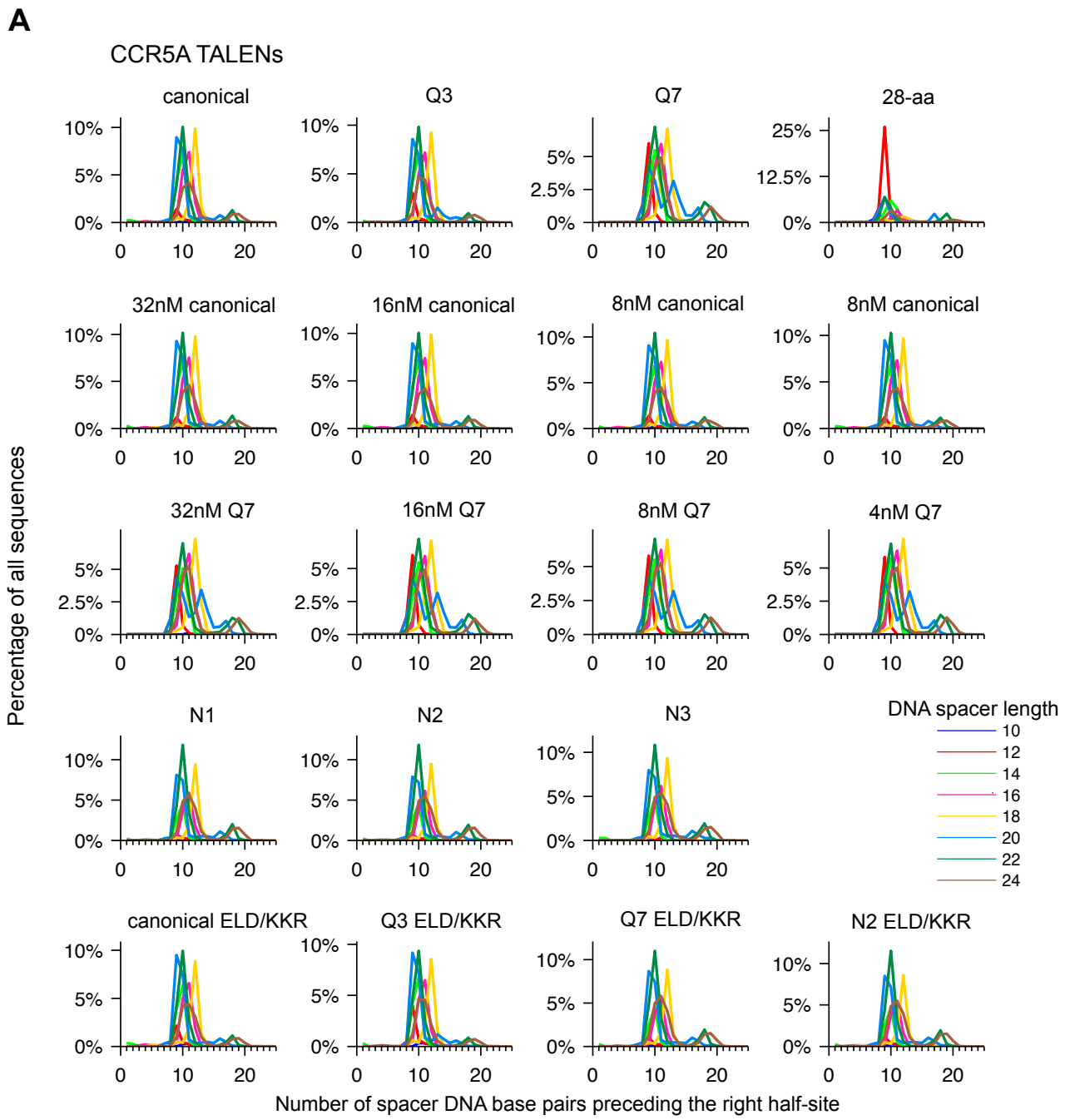


Figure 1.21. DNA cleavage-site preferences of TALENs.

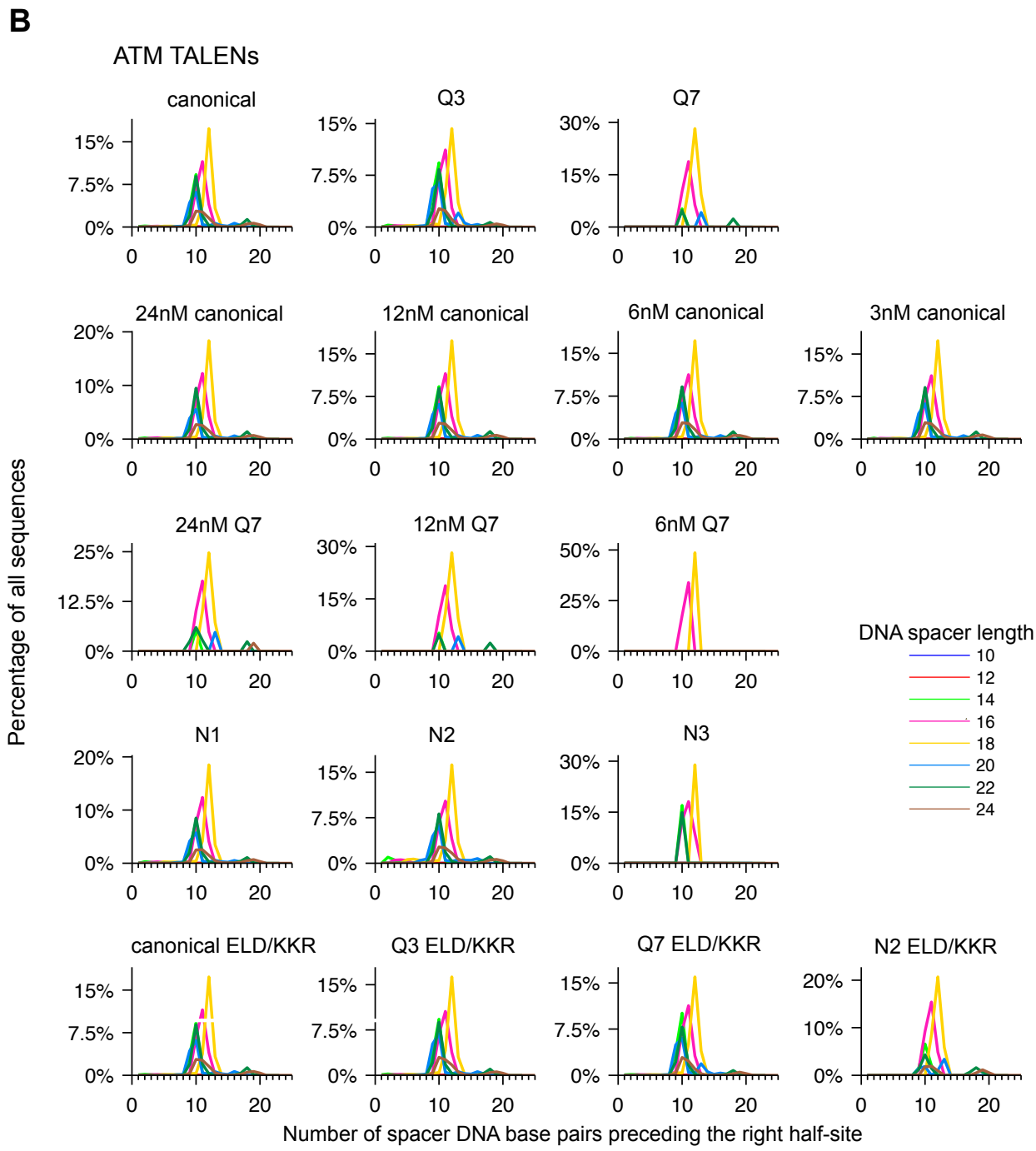


Figure 1.21 (Continued). DNA cleavage-site preferences of TALENs. (A) For each selection with CCR5A TALENs with various combinations of canonical, Q3, Q7, or 28-aa C-terminal domains; N1, N2, or N3 N-terminal mutations; and the EL/KK or ELD/KKR *Fok*I variants and at 4, 8, 16, or 32 nM, histograms of the number of spacer DNA base pairs preceding the right half-site for each possible DNA spacer length, normalized to the total sequence counts of the entire selection, are shown. (B) Same as (A) for ATM TALENs.

The peaks in the histogram were interpreted to represent the most likely locations of DNA cleavage within the spacer. The cleavage positions are dependent on the length of the DNA spacer between the TALEN binding half-sites, as might be expected from conformational constraints imposed by the TALEN C-terminal domain and DNA spacer lengths.

In order to confirm the greater DNA cleavage specificity of Q7 over canonical 63-aa C-terminal domains *in vitro*, a representative set of 16 off-target DNA substrates was digested *in vitro* with TALENs containing either canonical 63-aa or engineered Q7 C-terminal domains. ATM TALENs with the canonical 63-aa C-terminal domain demonstrated comparable *in vitro* cleavage activity on target sites with zero, one, or two mutations (**Figure 1.22**).

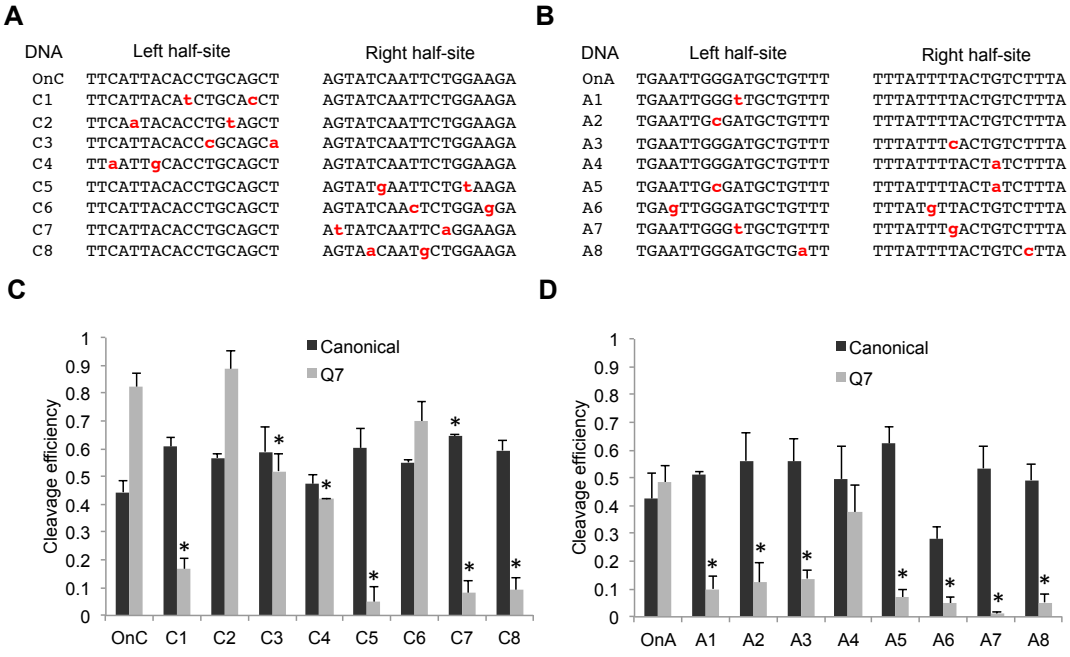


Figure 1.22. *In vitro* specificity and discrete cleavage efficiencies of TALENs containing canonical or engineered C-terminal domains. (A) CCR5A on-target sequence (OnC) and double-mutant sequences with mutations in red. For CCR5A, sequences containing two mutations were assayed because one-mutation and zero-mutation sequences were similarly enriched (**Tables 1.8 and 1.9**). (B) ATM on-target sequence (OnA), single-mutant sequences, and double-mutant sequences with mutations in red. (C) Discrete *in vitro* cleavage efficiency of DNA sequences listed in (D) with CCR5A TALENs containing either canonical or engineered Q7 C-terminal domains with EL/KK FokI domains. Error bars reflect s.d. from three biological replicates, except two replicates for C4. * Indicates that the efficiency of cleaving a mutant sequence was significantly different (P value < 0.01) from cleaving the on-target sequence (with the same TALEN type).

CCR5A TALENs with the canonical 63-aa C-terminal domain TALEN demonstrate comparable *in vitro* cleavage activity on target sites with zero or two mutations. In contrast, for 11 of the 16 off-target substrates tested, the engineered Q7 TALEN variants showed substantially higher (~4-fold or greater) discrimination against off-target DNA substrates with one or two mutations than the canonical 63-aa C-terminal domain TALENs, even though the Q7 TALENs cleaved their respective on-target sequences *in vitro* with comparable or greater efficiency than TALENs with the canonical 63-aa C-terminal domains (**Figure 1.22**). For both the ATM and CCR5A Q7 C-terminal TALENs, some sequences are cleaved with greater specificity than others. Sequence-dependent specificity is expected based on the variable specificity at each position (**Figures 1.8 and 1.9**). Overall, the discrete cleavage assays are consistent with the selection results and indicate that TALENs with engineered Q3 or Q7 C-terminal domains can be substantially more specific than TALENs with canonical 63-aa C-terminal domains *in vitro*.

To explore the greater on-target DNA cleavage activity of CCR5A TALENs with Q7 C-terminal domains compared to the lower activity of canonical C-terminal domains (**Figure 1.22**), a time course of *in vitro* cleavage of discrete on-target DNA with CCR5A TALENs in excess over DNA was preformed (**Figure 1.23**).

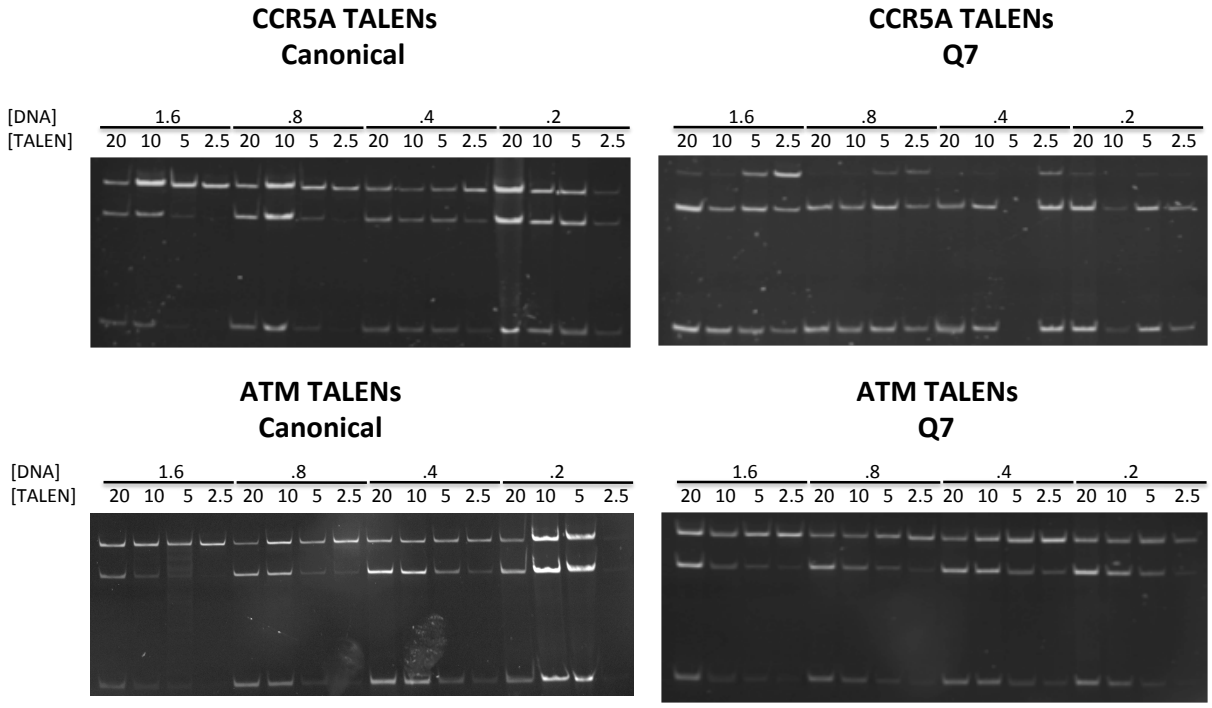


Figure 1.23. *In vitro* specificity and discrete cleavage efficiencies of TALENs containing canonical or engineered C-terminal domains. Discrete *in vitro* cleavage efficiency of CCR5A or ATM on-target DNA sequences with CCR5A or ATM TALENs containing either canonical

or Q7 C-terminal domains with EL/KK *Fok*I domains. The TALEN protein concentration ([TALEN]) was varied from 20 nM to 2.5 nM. On-target DNA concentration ([DNA]) was varied from 1.6 nM to .2 nM.

CCR5A TALENs with canonical C-terminal domains showed cleavage independent of time on the scale of the experiment (minutes), consistent with a model where TALENs bound almost all DNA but only a fraction of DNAs, ~.5, were quickly cleaved by active TALENs while an inactive form of TALEN bound and inhibited cleavage on the other ~.5 of DNAs over the entire course of the experiment. This model would require TALENs to have very tight binding affinities, well below the ~10 nM TALEN utilized in the digestion, consistent with previously reported K_ds of 3 nM for TALE DNA binding domains targeting 19 bases pairs [5]. Since TALENs have a longer, more positively charged linker and a *Fok* I domain, it is entirely likely that TALENs have substantially lower K_ds compared to these TALE constructs. In this model, TALENs would also be expected to cleave quickly, much faster than 1 min^{-1}, consistent with the previously reported rate of endogenous *Fok* I DNA cleavage of .5 s^{-1} [6]. CCR5A TALENs with the Q7 C-terminal domains cleave to the same initial point in 1 minute as the canonical C-terminal domain TALENs, ~.5, but then there was an increase in cleavage (**Figure 1.23**). Again, it is likely TALENs bound almost all DNA but only a fraction of DNAs, ~.5, were quickly cleaved by active TALENs while an inactive form of TALEN initially bound and inhibited cleavage on the other ~.5 of DNAs but for TALENs with Q7 C-terminal domains the inactive TALEN falls off over the course of the experiment allowing active TALEN to bind and cleave. Thus, the slow increase in cleavage by the Q7 C-terminal domains could be a result of a decrease in affinity from the Q7 C-terminal domain increasing the off-rate of the CCR5A TALENs with Q7 C-terminal domains. To preclude the possibility that TALENs are begin inactivated over the course of the experiment, TALENs were pre-incubated in reaction buffer without DNA for 9 min and then used to digest DNA normally resulting in no change in activity from TALENs not pre-incubated (data not shown). This model of inactive TALEN falling off allowing a catalytically active TALEN to bind and cleave is consistent with the observation of some single mutation sites having enrichments above wild type (**Figure 1.16 and Tables 1.8 and 1.9**) with TALENs binding these single mutation with excess but lower overall binding affinities and the resulting faster off-rates leading to some cleavage of otherwise inaccessible DNA target sites. Taken

together, it seems likely, at least *in vitro,* there is an inactive form of TALEN binding to DNA with implications in cleavage efficiency for both on-target and off-target sites.

1.6 Improved specificity of engineered TALENs in human cells

To determine if the increased specificity of the engineered TALENs observed *in vitro* also occurs in human cells, TALEN-induced modification rates of the on-target and top 36 predicted off-target sites were measured for CCR5A and ATM TALENs containing all six possible combinations of the canonical 63-aa, Q3, or Q7 C-terminal domains and the EL/KK or ELD/KKR *Fok*I domains (12 TALENs total). We did not analyze TALENs containing a 28-aa C-terminal domain in these experiments because both the ATM and CCR5A on-target sites have DNA spacer lengths of 18 bp, which lies outside the 28-aa C-terminal domain's preferred DNA spacer length range (**Figure 1.20**). For both *Fok*I variants, the TALENs with Q3 C-terminal domains demonstrate significant on-target activities ranging from 8% to 24% modification, comparable to the activity of TALENs with the canonical 63-aa C-terminal domains. TALENs with canonical 63-aa or Q3 C-terminal domains and the ELD/KKR *Fok*I domain are both more active in modifying the CCR5A and ATM on-target site in cells than the corresponding TALENs with the Q7 C-terminal domain by 5- to 9-fold (**Figure 1.24**).

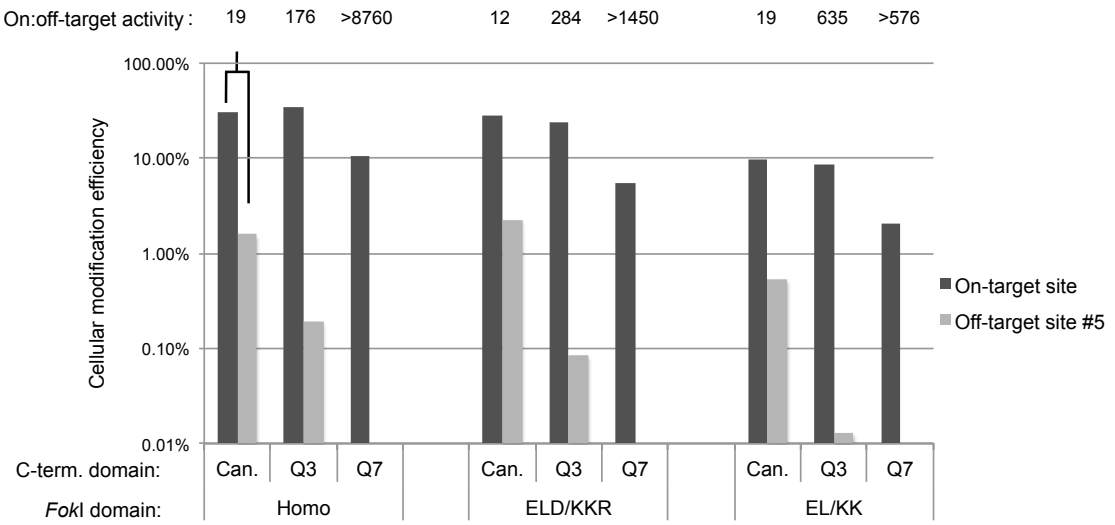


Figure 1.24. Specificity of engineered ATM and CCR5A TALENs in human cells. (A) The cellular modification efficiency of canonical and engineered TALENs expressed as a percentage of indels consistent with TALEN-induced modification out of total sequences is shown for the on-target CCR5A site (OnCCR5A) and for CCR5A off-target site #5 (OffC5), the most highly cleaved off-target substrate tested. All pairwise P-values comparing the number of observed sequences containing insertions or deletions consistent with TALEN-induced cleavage vs. the total number of sequences were calculated with a Fischer exact test between samples (see Supplementary Table S7). P-values are < 0.005 for samples of canonical vs. Q3 vs. Q7 TALENs in the same *Fok*I background for both on-target and off-target sites with the exception of off-target site #5 modified with Q3 vs. Q7 TALENs in the EL/KK *Fok*I background (P-value < 0.087). On:off target activity, defined as the ratio of on-target to off-target modification, is shown above each pair of bars.

Compared to the canonical 63-aa C-terminal domains, TALENs with Q3 C-terminal domains demonstrate a mean increase in on-target:off-target activity ratio of more than 12-fold and more than 9-fold for CCR5A and ATM sites, respectively, with the ELD/KKR *Fok*I domain (**Table 1.20**).

A

	CCR5A TALEN C-terminal domain		
Site	Can.	Q3	Q7
OnCCR5A	(27.6%)*	(24.3%)*	(5.57%)*
OffC-5	12	284	>1450
OffC-15	120	>2221	>335
OffC-16	886	769	>835
OffC-28	170	526	>835
OffC-36	181	2392	>835

B

	ATM TALEN C-terminal domain		
Site	Can.	Q3	Q7
OnATM	(16.2%)*	(17.1%)*	(1.83%)*
OffA-2	627	>2564	>274
OffA-11	448	>2564	>274
OffA-13	649	>2564	>225
OffA-23	56	>1210	>40

Table 1.20. Specificity of engineered ATM and CCR5A TALENs in human cells. (A) The on:off target activity of the canonical, Q3, and Q7 TALENs for each detected genomic off-target substrate of the CCR5A TALEN with the ELD/KKR *Fok*I domain are shown. The absolute genomic modification frequency for the on-target site is in parentheses with *. (B) Same as (B) for the ATM TALENs and off-target sites. (D) The on:off target activities of the canonical, Q3, and Q7 TALENs for each detected genomic off-target substrate of the PMS2, SDHD, and HDAC1_TALENs with the ELD/KKR *Fok*I domain are shown. The absolute genome modification frequency for the on-target site is in parentheses.

These mean improvements can only be expressed as lower limits due to the absence or near-absence of observed cleavage events by the engineered TALENs for many off-target sequences. For the ATM TALENs containing Q7 C-terminal domains, the cleavage efficiency of both the on-target and off-target sites is so low that their specificity cannot be determined (**Table 1.20**). For the most abundantly cleaved off-target site (CCR5A off-target site #5), the Q3 C-terminal domain is 24-fold more specific, and the Q7 C-terminal domain is > 120-fold more specific (**Figure 1.21 and Table 1.20**), than the canonical 63-aa C-terminal domain. Consistent with the improved on-target:off-target activity ratio observed *in vitro*, the engineered Q7 TALENs are more specific than the Q3 variants, which in turn are more specific than the canonical 63-aa C-terminal domain TALENs.

To determine if the increased specificity of the engineered TALENs observed for CCR5A and ATM TALENs applies more generally, three new TALENs targeting sequences in the

PMS2, SDHD, and HDAC1 genes[12] were constructed using the canonical 63-aa, Q3, or Q7 C-terminal domains and ELD/KKR *Fok*I domains. Of the 64 TALENs reported previously in Reyon et al, these three TALENs had target sequences with closely homologous genomic off-target sites containing one to five mutations. For each of these TALENs, modification rates were measured for genomic on-target and off-target sites. PMS2, SDHD, and HDAC-1 TALENs with Q3 C-terminal domains demonstrate on-target activities ranging from 6% to 28% modification, comparable to the activity of TALENs with the canonical 63-aa C-terminal domains (**Table 1.21**).

	HDAC1 TALEN C-terminal domain		
Site	Can.	Q3	Q7
OnHDAC	(5.22%)*	(6.09%)*	(1.42%)*
OffHDAC-1	100	487	> 1160

	SDHD TALEN C-terminal domain		
Site	Can.	Q3	Q7
OnSDHD	(33.10%)*	(28.08%)*	(0.28%)*
OffSDHD-1	135	722	> 231

	PMS2 TALEN C-terminal domain		
Site	Can.	Q3	Q7
OnPMS	(20.36%)*	(14.80%)*	(2.87%)*
OffPMS-1	14	83	748
OffPMS-2	5	36	320

Table 1.21. Specificity of engineered PMS2, SDHD, and HDAC1 TALENs in human cells. The on:off target activities of the canonical, Q3, and Q7 TALENs for each detected genomic off-target substrate of the PMS2, SDHD, and HDAC1 TALENs with the ELD/KKR *Fok*I domain are shown. The absolute genome modification frequency for the on-target site is in parentheses.

While demonstrating similar on-target activities to TALENs with canonical domains, PMS2, SDHD, and HDAC1 TALENs with Q3 C-terminal domains demonstrated a 5- to 7-fold increase in on-target:off-target activity ratio. For the PMS2 TALENs, the Q7 C-terminal domains demonstrated a 53- and 64-fold increase in on-target:off-target activity ratio in cells, although as observed above the Q7 TALENs were less active on the target site than TALENs containing the canonical and Q3 C-terminal domains (**Table 1.21**).

Together, these results reveal that for five families of TALENs targeting the *CCR5, ATM, PMS2, SDHD, and HDAC1* genes, replacing the canonical 63-aa C-terminal domain with the engineered Q3 C-terminal domain results in comparable activity for the on-target site in cells, and an average 10-fold increase in specificity for all assayed off-target sites. From the total results of sequencing genomic sites from cells treated with CCR5 and ATM TALENs (**Table 1.22**) and with PMS2, SDHD, and HDAC1 TALENs (**Table 1.23**), the engineered Q7 C-terminal domain can offer additional gains in specificity beyond that of the Q3 TALENs, but with reduced on-target activity. Collectively, these results validate a method to evaluate the specificity of TALEN variants that also revealed underlying principles resulting in engineered TALENs with improved DNA cleavage specificity in cells.

A

C-terminal domain	No TALEN	Q7	Q7	Q3	Q3	Canonical	Canonical	Canonical
*Fok*I domain	No TALEN	EL/KK	ELD/KKR	EL/KK	ELD/KKR	EL/KK	ELD/KKR	Homo
CCR5A Sites								
OnC								
Indels	1	147	705	1430	3731	841	2004	3943
Total	42042	7192	12667	16843	15381	8546	7267	8422
% Modified	<0.006%	2.044%	5.566%	8.490%	24.257%	9.841%	27.577%	46.818%
P-value		(5.5E-122)	(<1.0E-250)	(<1.0E-250)	(<1.0E-250)	<1.0E-250	<1.0E-250	<1.0E-250
Specificity								
OffC-1								
Indels	0	0	1	1	0	0	1	1
Total	51248	38975	79857	35490	77804	34227	87496	42497
% Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

Table 1.22 (Continued). Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites.

OffC-2								
Indels	6	1	6	2	1	0	11	3
Total	124356	96280	157387	93337	159817	85603	163322	114663
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	0.006%	<0.006%
P-value								
Specificity								
OffC-3								
Indels	6	1	6	2	1	0	11	3
Total	124356	96280	157387	93337	159817	85603	163322	114663
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-4								
Indels	0	1	0	0	0	0	0	0
Total	45377	44674	52876	35133	53909	26034	42284	40452
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-5								
Indels	0	0	0	3	22	134	385	395
Total	27009	28172	26036	22432	25800	25273	17045	17077
Modified	<0.004%*	<0.004%*	<0.004%*	0.013%	0.085%	0.530%	2.259%	2.313%
P-value					(1.4E-07)	4.1E-43	1.2E-160	1.2E-164
Specificity		>576	>1450	635	284	19	12	20
OffC-6								
Indels	0	0	0	0	0	0	0	0
Total	10766	12309	10886	9240	10558	10500	5943	6560
Modified	<0.009%	<0.008%	<0.009%	<0.011%	<0.009%	<0.010%	<0.017%	<0.015%
P-value								
Specificity								
OffC-7								
Total	0	0	0	0	0	0	0	0
Modified	15626	28825	22138	31742	19577	11902	33200	15400
P-value	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.008%	<0.006%	<0.006%
Specificity								
OffC-9								
Indels	0	0	0	1	0	0	0	0
Total	40603	39765	47974	51595	44002	34520	25211	30771
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-10								
Indels	0	0	0	0	0	0	0	0
Total	4142	9591	5187	1413	7975	4378	2215	3779
Modified	<0.024%	<0.010%	<0.019%	<0.071%	<0.013%	<0.023%	<0.045%	<0.026%
P-value								
Specificity								

Table 1.22 (Continued). Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites.

OffC-11								
Indels	0	0	0	0	0	0	0	0
Total	71180	55455	65015	44847	70907	50967	65257	60191
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-12								
Indels	0	0	0	0	0	0	0	0
Total	3242	1784	30274	14006	4897	19830	9747	12910
Modified	<0.031%	<0.056%	<0.006%	<0.007%	<0.020%	<0.006%	<0.010%	<0.008%
P-value								
Specificity								
OffC-13								
Indels	0	0	0	0	0	0	0	0
Total	65518	52459	53413	38156	61600	47922	57211	78546
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-14								
Indels	0	0	0	0	0	0	2	0
Total	34607	7217	26301	8339	29845	1081	9471	19026
Modified	<0.006%	<0.014%	<0.006%	<0.012%	<0.006%	<0.093%	0.021%	<0.006%
P-value								
Specificity								
OffC-15								
Indels	0	0	0	0	0	0	16	2
Total	4989	4880	6026	9370	9156	7371	6967	4662
Modified	<0.020%	<0.020%	<0.017%	<0.011%	<0.011%	<0.014%	0.230%	0.043%
P-value							1.9E-04	
Specificity		>100	>335	>796	>2221	>725	120	1091
OffC-16								
Indels	0	1	1	1	14	1	12	0
Total	36228	34728	34403	34866	44362	38384	38536	32636
Modified	<0.006%	<0.006%	<0.006%	<0.006%	0.032%	<0.006%	0.031%	<0.006%
P-value					(2.5E-04)		5.2E-04	
Specificity		>307	>835	>1274	769	>1476	886	>7023
OffC-17								
Indels	0	0	0	0	0	0	0	0
Total	32112	23901	31273	33968	27437	29670	27133	31299
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

Table 1.22 (Continued). Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites.

OffC-18								
Indels	0	0	0	0	0	0	0	0
Total	9437	9661	13505	14900	13848	12720	6624	12804
Modified	<0.011%	<0.010%	<0.007%	<0.007%	<0.007%	<0.008%	<0.015%	<0.008%
P-value								
Specificity								
OffC-19								
Indels	1	1	1	2	2	2	1	0
Total	22868	11479	22702	15258	20733	17449	14638	28478
Modified	<0.006%	0.009%	<0.006%	0.013%	0.010%	0.011%	0.007%	<0.006%
P-value								
Specificity								
OffC-20								
Indels	0	0	0	0	0	1	0	0
Total	23335	26164	30782	15261	20231	21184	14144	18972
Modified	<0.006%	<0.006%	<0.006%	<0.007%	<0.006%	<0.006%	<0.007%	<0.006%
P-value								
Specificity								
OffC-21								
Indels	0	0	0	0	0	0	0	0
Total	34302	27573	31694	24451	25826	27192	18110	21161
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-22								
Indels	1	0	0	0	0	0	0	0
Total	81037	86687	74274	79004	93477	92089	75359	104857
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-23								
Indels	0	0	0	0	0	0	0	0
Total	18812	19337	23034	25603	25023	28615	17172	21033
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-24								
Indels	0	0	1	0	0	0	0	1
Total	23538	21673	24594	27687	18343	29113	21709	26610
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-25								
Indels	0	0	0	0	0	0	0	0
Total	28941	25326	25871	10641	21422	20171	18946	18711
Modified	<0.006%	<0.006%	<0.006%	<0.009%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

Table 1.22 (Continued). Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites.

OffC-26								
Indels	0	0	1	0	0	0	0	0
Total	71831	48494	62650	45801	60175	65137	28795	64632
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-27								
Indels	0	0	0	0	0	0	0	0
Total	12181	2423	11258	7188	5126	4003	2116	4603
% Modified	<0.008%	<0.041%	<0.009%	<0.014%	<0.020%	<0.025%	<0.047%	<0.022%
P-value								
Specificity								
OffC-28								
Indels	0	0	0	0	6	1	12	5
Total	10651	6410	16179	13980	13022	7232	7379	8998
% Modified	<0.009%	<0.016%	<0.006%	<0.007%	0.046%	0.014%	0.163%	0.056%
P-value							2.2E-05	
Specificity		>131	>835	>1187	526	712	170	843
OffC-29								
Indels	0	0	0	0	0	0	0	0
Total	4262	3766	4228	6960	3234	1516	2466	1810
% Modified	<0.023%	<0.027%	<0.024%	<0.014%	<0.031%	<0.066%	<0.041%	<0.055%
P-value								
Specificity								
OffC-30								
Indels	0	0	0	0	0	0	0	0
Total	11840	12257	9617	34097	20507	5029	22248	6285
% Modified	<0.008%	<0.008%	<0.010%	<0.006%	<0.006%	<0.020%	<0.006%	<0.016%
P-value								
Specificity								
OffC-31								
Indels	0	0	0	0	0	0	0	0
Total	64522	67791	50085	50056	56241	48287	72230	100410
% Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-32								
Indels	0	0	0	0	0	0	0	0
Total	1944	6888	9330	3207	4591	6699	13607	19115
% Modified	<0.051%	<0.015%	<0.011%	<0.031%	<0.022%	<0.015%	<0.007%	<0.006%
P-value								
Specificity								

Table 1.22 (Continued). Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites.

OffC-33								
Indels	0	0	0	0	0	0	0	0
Total	34475	27039	18547	33467	15745	17075	4	18844
% Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<25.000%	<0.006%
P-value								
Specificity								
OffC-34								
Indels	0	0	0	0	0	0	0	0
Total	9052	18858	13647	11796	6945	6114	4979	9072
% Modified	<0.011%	<0.006%	<0.007%	<0.008%	<0.014%	<0.016%	<0.020%	<0.011%
P-value								
Specificity								
OffC-35								
Indels	0	0	0	0	0	0	0	0
Total	23839	22290	25133	24190	10	10459	22554	11897
% Modified	<0.006%	<0.006%	<0.006%	<0.006%	<10.000%	<0.010%	<0.006%	<0.008%
P-value								
Specificity								
OffC-36								
Indels	1	0	0	1	2	1	19	5
Total	23412	24394	23427	24132	19723	28369	12461	18052
Modified	<0.006%	<0.006%	<0.006%	<0.006%	0.010%	<0.006%	0.152%	0.028%
P-value							2.5E-08	
Specificity		>307	>835	>1274	2392	>1476	181	1690
B								
C-term. Domain:	No TALEN	Q7	Q7	Q3	Q3	Canonical	Canonical	Canonical
*Fok*I Domain:	No TALEN	EL/KK	ELD/KKR	EL/KK	ELD/KKR	EL/KK	ELD/KKR	Homo
ATM Sites								
On-A								
Indels	1	0	46	104	309	1289	410	909
Total	15116	1869	2520	1198	1808	19025	2533	5003
Modified	0.007%	<0.006%	1.825%	8.681%	17.091%	6.775%	16.186%	18.169%
P-value	0		(4.4E-27)	(1.4E-93)	(5.6E-243)	2.6E-214	<1.0E-250	<1.0E-250
Specificity								
OffA-1								
Indels	0	0	1	0	1	0	13	34
Total	52490	45383	34195	32325	47589	39704	50349	44056
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	0.026%	0.077%
P-value							9.3E-05	2.6E-12
Specificity		>0	>274	>1302	>2564	>1016	627	235
OffA-2								
Indels	0	0	0	0	0	0	0	0
Total	8777	11846	11362	12273	20704	3776	5650	5025
Modified	<0.011%	<0.008%	<0.009%	<0.008%	<0.006%	<0.026%	<0.018%	<0.020%
P-value								
Specificity								

Table 1.22 (Continued). Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites.

OffA-3								
Indels	0	0	0	0	1	0	0	0
Total	47338	14352	21253	17777	26512	19483	43728	29469
Modified	<0.006%	<0.007%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffA-4								
Indels	0	0	0	0	0	0	0	0
Total	12292	532	1383	2597	861	2598	1356	3573
Modified	<0.008%	<0.188%	<0.072%	<0.039%	<0.116%	<0.038%	<0.074%	<0.028%
P-value								
Specificity								
OffA-5								
Indels	0	0	0	0	0	0	0	0
Total	60859	22846	25573	19054	25315	31754	66622	60925
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffA-6								
Indels	0	0	0	0	0	0	0	0
Total	60859	22846	25573	19054	25315	31754	66622	60925
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffA-7								
Indels	0	0	0	0	0	0	0	0
Total	60859	22846	25573	19054	25315	31754	66622	60925
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffA-8								
Indels	0	0	0	0	0	0	0	0
Total	9170	1614	5934	3215	2450	12750	10120	13003
Modified	<0.011%	<0.062%	<0.017%	<0.031%	<0.041%	<0.008%	<0.010%	<0.008%
P-value								
Specificity								
OffA-9								
Indels	0	0	0	0	0	0	0	3
Total	8753	12766	9504	10114	11086	10676	9013	11110
Modified	<0.011%	<0.008%	<0.011%	<0.010%	<0.009%	<0.009%	<0.011%	0.027%
P-value								
Specificity								
OffA-10								
Indels	1	0	0	2	2	3	5	7
Total	8151	16888	8804	7061	8891	32138	14889	40120
Modified	0.012%	<0.006%	<0.011%	0.028%	0.022%	0.009%	0.034%	0.017%
P-value								
Specificity								

Table 1.22 (Continued). Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites.

OffA-11								
Indels	0	0	1	0	0	0	9	76
Total	41343	32352	28834	28709	26188	32519	24894	19586
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	0.036%	0.388%
P-value							1.5E-04	3.1E-38
Specificity		>0	>274	>1302	>2564	>1016	448	47
OffA-12								
Indels	0	0	0	0	0	0	0	0
Total	13186	2326	13981	12911	21134	9220	7792	8068
Modified	<0.008%	<0.043%	<0.007%	<0.008%	<0.006%	<0.011%	<0.013%	<0.012%
P-value								
Specificity								
OffA-13								
Indels	0	0	0	0	0	2	9	0
Total	32704	32015	12312	23645	26315	24078	36111	22364
Modified	<0.006%	<0.006%	<0.008%	<0.006%	<0.006%	0.008%	0.025%	<0.006%
P-value							4.3E-03	
Specificity		>0	>225	>1302	>2564	816	649	>2725
OffA-15								
Indels	0	0	0	0	1	0	0	0
Total	14654	15934	12313	6581	13053	18996	10916	21519
Modified	<0.007%	<0.006%	<0.008%	<0.015%	0.008%	<0.006%	<0.009%	<0.006%
P-value								
Specificity								
OffA-16								
Indels	1	0	0	0	0	0	0	12
Total	65190	35639	37252	30378	31489	22590	13594	20922
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.007%	0.057%
P-value								4.3E-07
Specificity		>0	>274	>1302	>2564	>1016	>2200	317
OffA-17								
Indels	0	0	0	0	0	0	0	6
Total	1972	606	1439	2113	2862	728	597	636
Modified	<0.051%	<0.165%	<0.069%	<0.047%	<0.035%	<0.137%	<0.168%	0.943%
P-value								2.0E-04
Specificity		>0	>26	>183	>489	>49	>97	19
OffA-18								
Indels	0	0	0	0	0	0	0	0
Total	5425	995	1453	1831	3132	1934	1534	5816
Modified	<0.018%	<0.101%	<0.069%	<0.055%	<0.032%	<0.052%	<0.065%	<0.017%
P-value								
Specificity								

Table 1.22 (Continued). Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites.

OffA-19								
Indels	1	2	0	1	1	1	1	3
Total	31094	41252	33213	29518	32337	25904	27575	38711
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	0.008%
P-value								
Specificity								
OffA-21								
Indels	0	0	0	0	0	0	0	0
Total	15297	9710	16719	12119	15483	21692	16558	16418
Modified	<0.007%	<0.010%	<0.006%	<0.008%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffA-22								
Indels	27	41	38	46	32	50	55	57
Total	9406	11150	11516	10269	13814	14057	11685	14291
Modified	0.287%	0.368%	0.330%	0.448%	0.232%	0.356%	0.471%	0.399%
P-value								
Specificity								
OffA-23								
Indels	1	0	0	0	0	0	10	20
Total	5671	9363	2203	7011	7078	12068	3484	8619
Modified	0.018%	<0.011%	<0.045%	<0.014%	<0.014%	<0.008%	0.287%	0.232%
P-value							4.5E-04	4.9E-04
Specificity		>0	>40	>609	>1210	>818	56	78
OffA-24								
Indels	4	0	0	1	0	1	0	2
Total	17288	7909	14261	29936	6943	6333	14973	19953
Modified	0.023%	<0.013%	<0.007%	<0.006%	<0.014%	0.016%	<0.007%	0.010%
P-value								
Specificity								
OffA-25								
Indels	0	0	0	0	0	0	0	0
Total	20089	45320	50758	108581	11574	20948	123827	74151
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.009%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffA-27								
Indels	0	0	0	0	1	0	0	0
Total	47338	14352	21253	17777	26512	19483	43728	29469
Modified	<0.006%	<0.007%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

Table 1.22 (Continued). Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites.

OffA-29								
Indels	0	0	0	0	0	0	0	0
Total	5174	12618	36909	18063	16486	17934	9999	35072
Modified	<0.019%	<0.008%	<0.006%	<0.006%	<0.006%	<0.006%	<0.010%	<0.006%
P-value								
Specificity								
OffA-30								
Indels	4	4	0	7	4	4	0	3
Total	45082	56631	36333	88651	69652	20362	29180	21350
Modified	0.009%	0.007%	<0.006%	0.008%	<0.006%	0.020%	<0.006%	0.014%
P-value								
Specificity								
OffA-32								
Indels	0	0	0	0	0	0	0	0
Total	13405	6721	14013	7513	14136	22376	6407	13720
Modified	<0.007%	<0.015%	<0.007%	<0.013%	<0.007%	<0.006%	<0.016%	<0.007%
P-value								
Specificity								
OffA-33								
Indels	0	0	0	0	1	1	0	4
Total	108222	46866	157329	48611	92559	152094	201408	225805
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffA-34								
Indels	0	0	0	0	0	0	0	2
Total	3889	3158	2903	2235	2112	3022	2322	2481
Modified	<0.026%	<0.032%	<0.034%	<0.045%	<0.047%	<0.033%	<0.043%	0.081%
P-value								
Specificity								
OffA-35								
Indels	0	0	0	1	0	0	0	33
Total	48482	37431	38043	31033	44803	37257	41073	47273
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	0.070%
P-value								7.6E-11
Specificity		>0	>274	>1302	>2564	>1016	>2428	260
.								
OffA-36								
Indels	0	0	2	0	0	0	0	0
Total	27115	17075	45425	35059	22298	19610	12620	27170
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.008%	<0.006%
P-value								
Specificity								

Table 1.22 (Continued). Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites.

Table 1.22 (Continued). Cellular modification induced by ATM and CCR5A TALENs at on-target and predicted off-target genomic sites.
(A) Results from sequencing CCR5A on-target and each predicted genomic off-target site that amplified from 50ng genomic DNA isolated from human cells treated with either no TALEN or TALENs containing canonical, Q3 or Q7 C-terminal domains, and either EL/KK heterodimeric, ELD/KKR heterodimeric, or homodimeric (Homo) *Fok*I domains. Indels: the number of observed sequences containing insertions or deletions consistent with TALEN-induced cleavage. Total: total number of sequence counts. Modified: number of indels divided by total number of sequences as percentages. Upper limits of potential modification were calculated for sites with no observed indels by assuming there is less than one indel then dividing by the total sequence count to arrive at an upper limit modification percentage, or taking the theoretical limit of detection (1/16,400), whichever value was larger. For the OffC-5 site, the limit of detection was 1/32,800 (see Methods). P-values: calculated as previously reported[15] using a (right) one-sided Fisher's exact test between each canonical C-terminal domain TALEN-treated sample and the untreated control sample. P-values of < 0.00025 were considered significant and shown. Cut off at 0.00025 was based on multiple comparison correction from the Benjamini-Hochberg method.[16, 17] For sites with a significant P-value in a canonical C-terminal domain TALEN-treated sample, significant P-values for other TALEN-treated samples are shown in parenthesis. Specificity is the ratio of on-target to off-target genomic modification frequency for each site. (B) Same as (A) for the *ATM* target sites; P-values < 0.005 were considered significant and are shown.

C-term. Domain:	No TALEN	Q7	Q3	Canonical
FokI Domain:	No TALEN	ELD/KKR	ELD/KKR	ELD/KKR
HDAC1 Sites				
OnHDAC				
Indels	0	697	7774	4595
Total	55902	49232	127610	88068
Modified	< 0.001%	1.416%	6.092%	5.218%
P-value		2.98E-229	< 1.00E-250	< 1.00E-250
Specificity				
OffHDAC-1				
Indels	0	2	30	169
Total	261086	292020	239664	322463
Modified	< 0.001%	< 0.001%	0.013%	0.052%
P-value			2.51E-10	4.50E-44
Specificity		> 1160	487	100
PMS2 Sites				
OnPMS				
Indels	1	591	4364	6941
Total	17023	20623	29496	34092
Modified	0.006%	2.866%	14.795%	20.360%
P-value		4.31E-151	< 1.00E-250	< 1.00E-250
Specificity				
OffPMS-1				
Indels	2	10	485	3740
Total	259211	260985	270713	259935
Modified	< 0.001%	0.004%	0.179%	1.439%
P-value		3.86E-02	1.24E-137	< 1.00E-250
Specificity		748	83	14
OffPMS-2				
Indels	1	6	334	3000
Total	63593	67027	81473	76343
Modified	0.002%	0.009%	0.410%	3.930%
P-value			8.22E-82	< 1.00E-250
Specificity		320	36	5
SDHD sites				
OnSDHD				
Indels	0	263	16217	24188
Total	94384	93203	57748	73085
Modified	< 0.001%	0.282%	28.082%	33.096%
P-value		1.55E-80	< 1.00E-250	< 1.00E-250
Specificity				
OffSDHD-1				
Indels	0	0	31	191
Total	65095	70344	79721	77706
Modified	< 0.001%	< 0.001%	0.039%	0.246%
P-value			9.88E-09	4.48E-51
Specificity		> 231	722	135

Table 1.23 (Continued). Cellular modification induced by HDAC1, PMS2, and SDHD TALENs at on-target and off-target genomic sites.

OffSDHD-2				
Indels	0	0	0	0
Total	146373	142097	153028	153595
Modified	< 0.001%	< 0.001%	< 0.001%	< 0.001%
P-value				
Specificity				
OffSDHD-3				
Indels	0	0	1	2
Total	86283	85294	114095	87973
Modified	< 0.001%	< 0.001%	< 0.001%	0.002%
P-value				
Specificity				
OffSDHD-4				
Indels	0	0	0	0
Total	53497	57541	65379	62535
Modified	< 0.001%	< 0.001%	< 0.001%	< 0.001%
P-value				
Specificity				

Table 1.23 (Continued). Cellular modification induced by HDAC1, PMS2, and SDHD TALENs at on-target and off-target genomic sites. (A) Results from sequencing HDAC1, PMS2, and SDHD on-target and genomic off-target sites amplified from 250 ng of genomic DNA isolated from human cells treated with either no TALEN, or with corresponding TALENs containing canonical, Q3, or Q7 C-terminal domains with ELD/KKR heterodimeric *Fok*I domains. Indels: the number of observed sequences containing insertions or deletions consistent with TALEN-induced cleavage. Total: total number of sequence counts. Modified: number of indels divided by total number of sequences as percentages. Upper limits of potential modification were calculated for sites with no observed indels by assuming there is less than one indel then dividing by the total sequence count to arrive at an upper limit modification percentage, or taking the theoretical limit of detection (1/82,000), whichever value was larger. P-values: calculated as previously reported[15] using a (right) one-sided Fisher's exact test between each canonical C-terminal domain TALEN-treated sample and the untreated control sample. P-values of < 0.025 were considered significant and shown. Specificity is the ratio of on-target to off-target genomic modification frequency for each site.

Discussion of TALEN specificity

The *in vitro* selection results for 30 unique TALENs each challenged with 10^{12} closed related off-target sequences and subsequent analysis inform our understanding of TALEN specificity through four key findings: (i) TALENs are highly specific for their intended target base pair at 103 of the 104 positions profiled with specificity increasing near the N-terminal TALEN end of each TALE repeat array (corresponding to the 5' end of the bound DNA); (ii) longer TALENs are more specific in a genomic context while shorter TALENs have higher specificity per nucleotide; (iii) TALE repeats each bind their respective base pairs relatively

independently; and (iv) excess DNA-binding affinity leads to increased TALEN activity against off-target sites and therefore decreased specificity.

The 16 confirmed TALEN off-target sites containing eight to 12 mutations identified from the 76 predicted sites assayed in this study represent more bona fide genomic off-target sites in the human genome than have been revealed collectively to date by other methods. These 16 sites were modified at efficiencies ranging from 2.3% to 0.03% in human cells, demonstrating that TALENs can have appreciable off-target activities in human cells even at sites that are eight or more mutations away from the on-target sequence. Site accessibility in cells, mediated by histone proteins, transcription factors, and DNA modification,[22] likely account for at least some of the difference between our *in vitro*, computational, and cell-based results.

While previous reports[7,19] have investigated TALEN specificity by identifying genomic off-target modifications for only a single TALEN pair, our study profiles the genomic specificity of two independent TALEN pairs (CCR5A and ATM) and is the first to compare different TALEN variants (canonical 63-aa vs. Q3 vs. Q7 C-terminal domains). It is difficult to directly compare our study with those that characterize different TALENs, although we note that our study has identified 16 bona fide heterodimeric off-target sites with eight to 12 mutations modified by TALENs in cells while the previous studies, using SELEX[7] or IDLV[19], collectively identified only three such sites that are distant from their corresponding on-target sequences. Furthermore, the same strategy used in this study has been previously shown to identify more off-target sites of a ZFN[21, 31] than the purely cellular IDLV study[18] or using SELEX data to predict off-target sites.[32] Of the 76 total IDLV insertion sites identified in cells treated with TALENs,[19] seven were at the target site while only three off-target sites were consistent with nuclease-mediated IDLV insertion. Only one of these three genomic off-target sites was heterodimeric (the use of TALENs containing homodimeric *Fok1* domains likely allowed for homodimeric off-target sites in which both right and left half-sites were targeted by the same TALEN). The on-target:off-target activity ratios of 170 and 1,140 for the two bona fide TALEN genomic off-targets sites out of 19 identified by SELEX[7] are similar to the 12 to 890 on:off-target activity ratios observed in this study.

The observed decrease in specificity for TALENs with more TALE repeats or more cationic residues in the C-terminal domain or N-terminus are consistent with a model in which excess TALEN binding affinity leads to increased promiscuity. This excess binding energy model may explain reports that NN RVDs bind either A or G.[2, 34, 30] These studies used TALE

arrays of more than 14 RVDs, which could have created a scenario in which excess DNA-binding energy masked a suboptimal NN RVD interaction with A compared to G. We observed NN RVDs to discriminate between A and G, consistent with reports using shorter TALE arrays of 13 RVDs[35] and by direct biochemical interrogation.[33] Excess DNA-binding energy could also explain the previously reported promiscuity at the 5' terminal T of TALENs with longer C-terminal domains[29] and is consistent with a report of higher TALEN protein concentrations resulting in more off-target site cleavage *in vivo*.[9] While decreasing TALEN protein expression in cells in theory could reduce off-target cleavage, TALE arrays are reported with on-target DNA binding affinities as high as K_d = 2.8 nM,[33] which is sufficient to theoretically saturate target sites even when expressed at modest, mid-nM concentrations in the cell. The difficulty of improving the specificity of such TALENs by lowering their expression levels, coupled with the need to maintain sufficient TALEN concentrations to effect desired levels of on-target cleavage, highlight the value of engineering TALENs with higher intrinsic specificity such as those described in this work.

Our findings suggest that mutant C-terminal domains with reduced non-specific DNA binding may be used to alter the DNA-binding affinity of TALENs such that on-target sequences are cleaved efficiently but with minimal excess DNA-binding energy, resulting in better discrimination between on-target and off-target sites. Since TALENs targeting up to 46 total base pairs have been shown to be active in cells,[14] it may be possible to further improve specificity by engineering TALENs with a combination of mutant N-terminal and C-terminal domains that impart reduced non-specific DNA-binding, a greater number of TALE repeats to contribute additional on-target DNA binding, and lower-affinity RVDs such as the NK RVD to recognize G.[34, 35] It is tempting to speculate that the strategy of mutating residues that contribute to non-specific DNA binding to improve DNA specificity may also apply to other genome engineering proteins including Cas9 and ZFNs. While a comparison of the specificity of various programmable genome-editing technologies including ZFNs, TALENs and CRISPRs would be of interest, in order for such a comparison to be rigorous, many new lines of experiments are needed so that all three technologies are evaluated on their ability to cleave the same target sequences *in vitro* and in cells under the same conditions.

Our model and the resulting improved TALENs would have been difficult to derive or validate using purely cellular off-target cleavage methods. The ability of our profiling method to

reveal the broad, unobscured DNA cleavage specificity of TALENs in the absence of cellular complications enabled the elucidation of the inherent DNA-cleavage specificity of TALENs. Studies of cellular off-target cleavage are also intrinsically limited by the small number of sequences closely related to a target sequence of interest that are present in a genome. In contrast, each active, dimeric TALEN in this study was evaluated for its ability to cleave any of 10^{12} close variants of its on-target sequence, a library size several orders of magnitude greater than the number of different sequences in a mammalian genome. This dense coverage of off-target sequence space enabled the elucidation of detailed relationships between DNA-cleavage specificity and target base pair position, TALE repeat length, TALEN concentration, mismatch location, and engineered TALEN composition. These results collectively reveal principles for characterizing and improving TALENs with greater specificity that may enable a wider range of genome engineering applications.

1.7 Methods used to study TALEN specificity

Oligonucleotides, PCR and DNA Purification.

PCR was performed with 0.4 µL of 2 U/µL Phusion Hot Start II DNA polymerase (Thermo-Fisher) in 50 µL with 1x HF Buffer, 0.2 mM dNTP mix (0.2 mM dATP, 0.2 mM dCTP, 0.2 mM dGTP, 0.2 mM dTTP) (NEB), 0.5 µM to 1 µM of each primer and a program of: 98 ˚C, 1 min; 35 cycles of [98 ˚C, 15 s; 62 ˚C, 15 s; 72 ˚C, 1 min] unless otherwise noted. Many DNA reactions were purified with a QIAquick PCR Purification Kit (Qiagen) referred to below as Q-column purification or MinElute PCR Purification Kit (Qiagen) referred to below as M-column purification. All oligonucleotides were purchased from Integrated DNA Technologies (IDT). Oligonucleotide sequences are listed below in **Table 1.24.**

A

oligonucleotide name	oligonucleotide sequence (5'->3')
CCR5Aon-Circ	CACCACTNTTCATTACACCTGCAGCTNNNNNNNNNNNNAGTATCAATTCTGGAAGANCGTCACGCT
CCR5Amut-Circ	CACCACTNTTCATTACACCTGCAGCTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCGTCACGCT
TAL-Nrev	5Phos/CAGCAGCTGCCCGGT
TAL-N1fwd	5Phos/cAGATCGCGAAGAGAGGGGGAGTAACAGCGGTAG
TAL-N2fwd	5Phos/cAGATCGCGcAGAGAGGGGGAGTAACAGCGGTAG
TAL-N3fwd	5Phos/cAGATCGCGcAGcagGGGGGAGTAACAGCGGTAG
TAL-Cifwd	ATC GTA GCC CAA TTG TCC A
TAL-Cirev	GTTGGTTCTTTGGATCAATGCG
TAL-Q3	AAGTTCTCTCGGGAATCCGTTGGTTGGTTCTTTGGATCA
TAL-Q7	GAAGTTCTCTCGGGAATTTGTTGGTTGGTTTGTTGGATCAATGCGGGAGCATGAGGCAGACCTTGTTGGA CTGCATC
TAL-Ciirev	CTTTTGACTAGTTGGGATCCCGCGACTTGATGGGAAGTTCTCTCGGGAAT
CCR5A Library10	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNNA%G%T% A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library12	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNNNNA%G% T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library14	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNNNNNNNA% G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library16	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNNNNNNNNA %G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library18	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNNNNNNNNN NNA%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library20	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNNNNNNNNN NNNNA%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library22	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNNNNNNNNN NNNNNNA%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library24	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNNNNNNNNN NNNNNNNNA%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5B Library10	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNC%A%T%A%C% A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library12	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNC%A%T%A% C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT

Table 1.24 (Continued). **Oligonucleotides used in this study**.

CCR5B Library14	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNNC%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library16	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNNNNNC%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library18	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNNNNNNNC%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library20	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNNNNNNNNC%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library22	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNNNNNNNNNNC%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library24	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNNNNNNNNNNNC%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
ATM Library10	Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library12	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library14	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library16	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNNNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library18	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNNNNNNNNNT%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library20	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNNNNNNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library22	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNNNNNNNNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library24	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNNNNNNNNNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
#1 adapter-fwd**1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGT
#1 adapter-rev**1	ACAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGAA
#1 adapter-rev**2	TTCAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCAA
#1 adapter-rev**3	TTGCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGACT
#1 adapter-rev**4	AGTCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCATT
#1 adapter-rev**5	AATGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**6	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCATGA
#1 adapter-rev**6	TCATGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**7	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGCT
#1 adapter-rev**7	AGCATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**8	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTAGT
#1 adapter-rev**8	ACTAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**9	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTAA
#1 adapter-rev**10	TTAGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**10	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGTA
#1 adapter-rev**11	TACTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**11	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTACT
#1 adapter-rev**12	AGTACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**12	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGT
#1 adapter-rev**13	ACAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**13	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTAA
#1 adapter-rev**14	TTAGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**14	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGTA
#1 adapter-rev**14	TACTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**15	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTACT
#1 adapter-rev**15	AGTACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**16	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGT
#1 adapter-rev**16	ACAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG

Table 1.24 (Continued). **Oligonucleotides used in this study**.

#2A primer-fwd	AATGATACGGCGACCAC
#2A primer-rev*CCR5A	GTTCAGACGTGTGCTCTTCCGATCTNNNNAGTGGTGAGCGTGACG
#2A primer-rev*ATM	GTTCAGACGTGTGCTCTTCCGATCTNNNNACGCGGAGTGGGGTACC
#2A primer-rev*CCR5B	CAGACGTGTGCTCTTCCGATCNNNNAGCGTGGAGTCCCGAGG
#2B primer-fwd	AATGATACGGCGACCAC
#2B primer-rev**1	CAAGCAGAAGACGGCATACGAGATTGTTGACTGTGACTGGAGTTCAGACGTGTGCTCTTC
#2B primer-rev**2	CAAGCAGAAGACGGCATACGAGATACGGAACTGTGACTGGAGTTCAGACGTGTGCTCTTC
#2B primer-rev**3	CAAGCAGAAGACGGCATACGAGATTCTAACATGTGACTGGAGTTCAGACGTGTGCTCTTC
#2B primer-rev**4	CAAGCAGAAGACGGCATACGAGATCGGGACGGGTGACTGGAGTTCAGACGTGTGCTCTTC
	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCG
#1 Lib. adapter-fwd*CCR5A	GTACCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG
#1 Lib. adapter-rev*CCR5A	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTG
#1 Lib. adapter-fwd*ATM	GTACGATGCGATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG
#1 Lib. adapter-rev*ATM	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGCATC
#1 Lib. adapter-fwd*CCR5B	TCGGGAACGTGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGTCTAATCTCGTATGCCGTCTTCTGCTTG
#1 Lib. adapter-rev*CCR5B	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCACGTT
#2A Lib. primer-rev	CAAGCAGAAGACGGCATACGA
#2A Lib. primer-fwd*CCR5A	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCGTCACGCTCACCACT
#2A Lib. primer-fwd*ATM	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGGTACCCCACTCCGCGT
#2A Lib. primer-fwd*CCR5B	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCCTCGGGACTCCACGCT
#2B Lib. primer-rev	CAAGCAGAAGACGGCATACGA
#2B Lib. primer-fwd	AATGATACGGCGACCAC
G adapter-fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
G adapter-rev	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCA
G-B primer-fwd	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC
G-B primer-rev**1	CAAGCAGAAGACGGCATACGAGATGTGCGGACGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**2	CAAGCAGAAGACGGCATACGAGATCGTTTCACGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**3	CAAGCAGAAGACGGCATACGAGATAAGGCCACGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**4	CAAGCAGAAGACGGCATACGAGATTCCGAAACGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**5	CAAGCAGAAGACGGCATACGAGATTACGTACGGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**6	CAAGCAGAAGACGGCATACGAGATATCCACTCGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**7	CAAGCAGAAGACGGCATACGAGATAAAGGAATGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**8	CAAGCAGAAGACGGCATACGAGATATATCAGTGTGACTGGAGTTCAGACGTGTGCT
CCR5AonCfwd	CGACGGTCTAGAGTCTTCATTACACCTGCAGCTCTCATTTTCCATACAGT
CCR5Amut1fwd	CGACGGTCTAGAGTCTTCATTACAtCTGCAcCTCTCATTTTCCATACAGT
CCR5Amut2fwd	CGACGGTCTAGAGTCTTCAaTACACCTGtAGCTCTCATTTTCCATACAGT
CCR5Amut3fwd	CGACGGTCTAGAGTCTTCgTTACACCTGCAtCTCTCATTTTCCATACAGT
CCR5Amut4fwd	CGACGGTCTAGAGTCTTaATTgCACCTGCAGCTCTCATTTTCCATACAGT
CCR5AonCrev	CCGACGAAGCTTTTCTTCCAGAATTGATACTGACTGTATGGAAAATGA
CCR5Amut1rev	CCGACGAAGCTTTTCTTaCAGAATTcATACTGACTGTATGGAAAATGA
CCR5Amut2rev	CCGACGAAGCTTTTCcTCCAGAgTTGATACTGACTGTATGGAAAATGA
CCR5Amut3rev	CCGACGAAGCTTTTCTTCCtGAATTGATAaTGACTGTATGGAAAATGA
CCR5Amut4rev	CCGACGAAGCTTTTCTTCCAGcATTGtTACTGACTGTATGGAAAATGA
ATMonAfwd	CGACGGTCTAGATTTGAATTGGGATGCTGTTTTTAGGTATTCTATTCAAATT
ATMmut1fwd	CGACGGTCTAGATTTGAATTGGGtTGCTGTTTTTAGGTATTCTATTCAAATT
ATMmut2fwd	CGACGGTCTAGATTTGAATTGcGATGCTGTTTTTAGGTATTCTATTCAAATT
ATMmut3wd	CGACGGTCTAGATTTGAgTTGGGATGCTGTTTTTAGGTATTCTATTCAAATT
ATMmut4fwd	CGACGGTCTAGATTTGAATTGGGATGCTGaTTTTAGGTATTCTATTCAAATT

Table 1.24 (Continued). **Oligonucleotides used in this study**.

ATMonArev	CCGACGAAGCTTAATAAAGACAGTAAAATAAATTTGAATAGAATACCTAAAA
ATMmut1rev	CCGACGAAGCTTAATAAAGACAGTgAAATAAATTTGAATAGAATACCTAAAA
ATMmut2rev	CCGACGAAGCTTAATAAAGAtAGTAAAATAAATTTGAATAGAATACCTAAAA
ATMmut3rev	CCGACGAAGCTTAATAAAGACAGTAAgATAAATTTGAATAGAATACCTAAAA
ATMmut4rev	CCGACGAAGCTTAATAAcGACAGTAAAATAAATTTGAATAGAATACCTAAAA
CCR5BonBfwd	CGACGGTCTAGAAAGGTCTTCATTACACCTGCAGCTCTCATTTTCCATACAGTCA
CCR5Bmut1fwd	CGACGGTCTAGAGTCTTCATTACACCTGtAGCTCTCATTTTC
CCR5Bmut2fwd	CGACGGTCTAGAGTCTTCATaACACCTGCAGCTCTCATTTTC
CCR5Bmut3fwd	CGACGGTCTAGAGTCTTCATTACACCcGCAGCTCTCATTTTC
CCR5Bmut4fwd	CGACGGTCTAGAGTCTTCATaACACCTGtAGCTCTCATTTTC
CCR5Bmut5fwd	CGACGGTCTAGAGTCTTCATTAtACCTaCAGCTCTCATTTTC
CCR5Bmut6fwd	CGACGGTCTAGAGTCTTCATTgCACCcGCAGCTCTCATTTTC
CCR5BonBrev	CCGACGAAGCTTTCTTCCAGAATTGATACTGACTGTATGGAAAATGAGAGCT
CCR5Bmut1rev	CCGACGAAGCTTTCTTCCAGAATTGATACTaACTGTATGGAAAATGAGAGCT
CCR5Bmut2rev	CCGACGAAGCTTTCTTCCAGAATTGATACTGACTGTATcGAAAATGAGAGCT
CCR5Bmut3rev	CCGACGAAGCTTTCTTCCAGAATTGATACTGACTGaATGGAAAATGAGAGCT
CCR5Bmut4rev	CCGACGAAGCTTTCTTCCAGAATTGATACcGACTGTATGGAAAATGAGAGCT
CCR5Bmut5rev	CCGACGAAGCTTTCTTCCAGAATTGATACTaACTGTATcGAAAATGAGAGCT
CCR5Bmut6rev	CCGACGAAGCTTTCTTCCAGAATTGATACTGAaTGTgTGGAAAATGAGAGCT
CCR5Bmut7rev	CCGACGAAGCTTTCTTCCAGAATTGATACTGAtaGTATGGAAAATGAGAGCT
pUC19Ofwd	GCGACACGGAAATGTTGAATACTCAT
pUC19Orev	CAGCGAGTCAGTGAGCGA

B

Site	Fwd primer	Rev primer	PCR
OnCCR5A	TCACTTGGGTGGTGGCTGTG	GACCATGACAAGCAGCGGCA	
OffC-1	AGTCCAAGACCAGCCTGGGG	AAGAACCTGTTGTCTAATCCAGCA	
OffC-2	GAACCTGTTGTCTAATCCAGCGTC	CTGCAAAGAAGGCCAGGCA	
OffC-3	GAACCTGTTGTCTAATCCAGCGTC	CTGCAAAGAAGGCCAGGCA	
OffC-4	TGACCTGTTTGTTCAGGTCTTCC	CCATATGGTCCCTGTCGCAA	
OffC-5	TCCAGTTGCTGTCCCTTCAGA	ACAGGGAGAGCCACCAATGC	
OffC-6	GCCCGGCCTGTCCTGTATTT	CACCCACACATGCACTTCCC	
OffC-7	TGGCTATTCTAGTTCTTTTGCAT	CCATGCCCTAGGGATTTGTGGA	
OffC-8	CGCTGAAGGCTGTCACCCTAA	TGGACCTAAGAGTCCTGCCCAT	ND
OffC-9	CCACCACCACACAACTTCACA	CAGCTGGCGAGAACTGCAAA	
OffC-10	TTCCAGGTCCTTTGCACAAATA	GCAAGGTCGTTGGATAGAAGTTGA	
OffC-11	CACCGAAAGCAACCCATTCC	TGATCTGCCCACCCCAGACT	
OffC-12	TTCATTCTCACCATCTGGAATTGG	TCTGGCTGGACTGCTCTGGTT	
OffC-13	TGGCATGTGGATCAGTACCCA	TAGAACATGCCCGCGAACAG	
OffC-14	CTGACGTCCATGTCAACGGG	TTTGAATTCCCCCTCCCCAT	
OffC-15	GCTCCTTTCTGAGAAGCACCCAT	GGCAGATGGTGGCAGGTCTT	
OffC-16	ATGAGGGCTTGGATTGGCTG	CCACCTCCCCCACTGCAATA	
OffC-17	GGAGGCCTTCATTGTGTCACG	AACTCCACCTGGGTGCCCTA	
OffC-18	CGTGGTCCCCCAGAAATCAC	GGAGCAGGAGTTGGTGGCAT	
OffC-19	GATTGCATAGGTTAGCATTGCC	GCCCCTGTTGGTTGACTCCC	
OffC-20	TTCCAGCGAATGGAAAGTGCT	AAGCCCAGGAATAAGGGCCA	
OffC-21	AAGCATGCTCACACTGTGGTGTA	TTGCTTGAGGCGGAAGTTGC	
OffC-22	TGACCCTCCAGCAAAGGTGA	CCCCAGGGACTGAGCATGAG	
OffC-23	GCTTTGCTTGCACTGTGCCTT	GGGGACAGACTGTGAGGGCT	
OffC-24	TCAAAAGGATGTGATCTGCCACA	GGCCTCTTTGAGGGCCAGTT	

Table 1.24 (Continued). **Oligonucleotides used in this study**

OffC-25	CCAGGGCTCAATTCTTAGACCG	AAAAGAGCAGGGCTGCCATC	
OffC-26	TGTTCATGCCTGCACAGTGG	TGGATGTGCCCTCTACCACA	
OffC-27	TTTGGCAAGGAATTCACAGTTC	TCATGCCTGCACAGTGGTTG	
OffC-28	GGAGGATGTCTTTGTGGTAGGGG	CGCTGCCAAGCAAACTCAAA	
OffC-29	TCCCCCAACTTCACTGTTTTT	GCAATGAGCATGTGGACACCA	
OffC-30	TTCTCTGTTTCCAGTGATTTCAGA	GTCGCAAAACAGCCAGTTGC	+DMSO
OffC-31	TGGCTTGGTTAATGGACAATGG	CCTGCAAGGAGCAAGGCTTC	+DMSO
OffC-32	TGGGCTTCGTTGACTTAAAGAG	GGACAAGAGGGCCAGGGTTT	
OffC-33	TCTTAAACATGTGGAACCCAGTCAT	TGAAAACCCACAGAGTGGGAGA	
OffC-34	GCAGATTCATTAGCGTTTGTGGC	TGCATGGGTGTAAATGTAGCAGAAA	
OffC-35	CCAAGGATCAATACCTTTGGAGGA	GCCCTCCCTTGAATCAGGCT	
OffC-36	TTCCCCTAACCAGGGGCAGT	GTGGTGAGTGGGTGTGGCAG	+DMSO
OffC-38	CGCCCATGAGAAAGAGTCCA	CTACACCCCCTCCCCAAAGG	
OffC-39	GCTGTCATTTCCAAAGCCGC	GGCCTCTAGAGTGAGGGGGTTG	
OffC-40	GCTGTCATTTCCAAAGCGGC	GGCCAACAAGATGCCAAGGT	
OffC-42	AGTTCTCAGCAACTAAAGTATTGA	CTGGGATTACAGGCGTGAGC	ND
OffC-45	CCATCATTGGCATCATGGGA	TGGAATGGAGTGGCCAACCT	ND
OffC-49	GTCAGGACCACACAAGAAAATAAAA	GGGAATGCCAGTCTTTGCCA	
OffC-56	CCCAGCCGATGACCAGAAAT	GCCAGCTGAGCTCTTTGCTGTA	
OffC-65	ACCCATTGATAAGGCACATTCT	AGTCAAACTCATCTAACACTCCAG	
OffC-69	TCTCCCACACCCCACATTCA	TGCGATAAGGTTGCAATGACA	
OffC-76	TAGGGCGCCTCAGATCCACT	GAGCTGCAGGCCTTGAATGG	
OffC-137	AGCCACAAGGGCCTGCTGTA	ACTGCGCCCAGCCTATCACT	
OffC-150	GCTTTTCTTTTCAGCCAAATGAAAGTT	ACGTGCGTCCCTGTCACTCA	
OnATM	AGCGCCTGATTCGAGATCCT	ATGCCAAATTCATATGCAAGGC	
OffA-1	CCTGCCATTGAATTCCAGCCT	TGTCTGCCTTTCCTGTCCCC	
OffA-2	GACTGCCACTGCACTCCCAC	GGATACCCTTGCCTCCCCAC	
OffA-3	CCTCCCATTTTCCTTCCTCCA	CTGGGAGACACAGGTGGCAG	
OffA-4	TCCTCCAATTTTCCTTCCTCCA	CTGGGAGACACAGGTGGCAG	
OffA-5	CTGGGAGACACAGGTGGCAG	AGGACCAATGGGGCCAATCT	
OffA-6	CTGGGAGACACAGGTGGCAG	AGGACCAATGGGGCCAATCT	
OffA-7	CTGGGAGACACAGGTGGCAG	AGGACCAATGGGGCCAATCT	
OffA-8	GCATGCCAAAGAAATTGTAGGC	TTCCCCCTGTCATGGTCTTCA	
OffA-9	GCATCTCTGCATTCCTCAGAAGTGG	AGAAACTGAGCAAGCCTCAGTCAA	
OffA-10	GGGATACCAAAGAGCTTTTGTTTTGTT	CAGAGGCTGCATGATGCCTAATA	
OffA-11	TGCAGCTACGGATGAAAACCAT	TCAGAATACCTCCCCGCCAG	
OffA-12	GCATAAAGCACAGGATGGGAGA	TCCCTCTTTAACGGTTATGTTGGC	
OffA-13	TGGGTTAAGTAATTTCGAAAGGAGAA	ATGTGCCCCACACATTGCC	+DMSO
OffA-14	GAGTGAGCCACTGCACCCAG	CGTGTGGTGGTGGCACAAG	ND
OffA-15	CCTCCCTCTGGCTCCCTCCC	ACCAGGGCCTGTTGGGGGTT	
OffA-16	TGCTCCCTGACCTTCCTGAGA	CCATTGGAATGAGAACCTTCTGG	
OffA-17	GGTGGAACAATCCACCTGTATTAGC	GAATGTGACACCACCACCGC	
OffA-18	GGCTTTGCAAACATAAACACTCA	CCTTCTGAGCAGCTGGGACAA	
OffA-19	CACTGGAACCCAGGAGGTGG	CCTCCCATTGGAGCCTTGGT	ND
OffA-20	CAGCCTGCCTGGGTGACAG	CATCTGAGCTCAAAACTGCTGC	+DMSO
OffA-21	GCCACTGCATTGCATTTTCC	TGAGGGCAGGTCTGTTTCCTG	

Table 1.24 (Continued). **Oligonucleotides used in this study**

OffA-22	GGGAGGATCTCTCGAGTCCAGG	CCTTGCCTGACTTGCCCTGT	
OffA-23	TGTTTAGTAATTAAGACCCTGGCTTTC	GCGACAGGTACAAAGCAGTCCAT	
OffA-24	GCCCTTTGATTTCATCTGTTTCCC	CATTGCTGCCATTGCACTCC	
OffA-25	AAACTGGCACATGTACTCCT	ACATGATTTGATTTTTCATGTGTTT	
OffA-26	GGGTGGAAGGTGAGAGGAGATT	CGCAGATGGGCATGTTATTG	ND
OffA-27	CCTCCCATTTTCCTTCCTCCA	GACTGCCACTGCACTCCCAC	
OffA-28	AGCCAAGATTGCACCATTGC	GTCCCTGACGGAGGCTGAGA	ND
OffA-29	TGGTTGGATTTTGGCTCTGTCAC	TGTCAATATCAATACCCTGCTTTCCTC	
OffA-30	TGGTTACTTTTAAAGGGTCATGATGGA	AAAAATGGATGCAAAGCCAAA	+DMSO
OffA-31	GGGACACAGAGCCAAACCGT	TGTGCACATGTACCCTAAAACT	ND
OffA-32	CAGTCATTGTTTCTAGGTAGGGGA	TTGGCAATTTGGGTGCAACA	
OffA-33	TGGATAACCTGCAGATTTGTTTCTG	TGAGCCCAGGAGTTTCAGGC	
OffA-34	TCGTGTGTGTGTGTTTGCTTCA	CAGTGGTTCGGGAAACAGCA	
OffA-35	TGGGAATGTAAATCTGACTGGCTG	CTGGAACTCTGGGCATGGCT	
OffA-36	GCTGCAATTGCTTTTTGGCA	TGGACCCCTCCCTTACACC	
CP_CCRoff-1	TTGTTCCAACCAGCTTCATGATAA	CCTCCTTAAAGCTTCTTGCCA	
CP_CCRoff-2	TGTGACACAAACCATTGCATTC	CTATGAAGAGCTATTGATGCAGAA	
CP_CCRoff-3	TCCAGAATAATCACTGTGGCTGC	ACCAGGGGAAACTAGTGGGAGG	
CP_CCRoff-4	AACTTACTATCTGCTGGACACATTG	CTAGAGCCCCTGTTGGTTGACT	
CP_CCRoff-5	GCTGGGCTAGACCACAATTTTT	CAAGGTTCAGTTTCCCTGCTCT	
CP_CCRoff-6	TGTAATAGCAAGGCTTCAGGAC	AGAAAAGTTTCAGTGAAGAAAAACG	
CP_CCRoff-7	AGTTCTCAGCAACTAAAGTATTGA	CTGGGATTACAGGCGTGAGC	ND
CP_CCRoff-8	GGAGGCTGAGGTGAGAGGTT	ATCACCCGTCTTCTGCATCG	
CP_CCRoff-9	ACCTGTTGTCTAATCCAGCGTC	GAGACCAGGAGTCCGAGACC	
CP_CCRoff-10	ACCTGTTGTCTAATCCAGCGTC	GAGACCAGGAGTCCGAGACC	
CP_CCRoff-11	CAGGAGTCCAAGACCAGCCT	AACCTGTTGTCTAATCCAGCA	
CP_CCRoff-12	ACCCATTGATAAGGCACATTCT	AGTCAAACTCATCTAACACTCCAG	
CP_CCRoff-13	TAGGGTCTCACCCACTTGCTC	GCCTTGGGTCAGTCCTGGAG	
CP_CCRoff-14	TTGGAGGGAATGAGTTGGCTG	AGTATTTGGCACAGTGATGGG	
CP_CCRoff-15	AGAAGCAGGTACCTTCCCACC	GCATGTTAAGCCATAGAAAGGGC	
CP_CCRoff-16	AAAAAGCTCCAGCCCAGTCTC	TTTGGAATTGATCCATTGAATTTGA	
CP_CCRoff-17	TCTGTCCCTCCTACTGAGGC	TGGAAGGCTCATTTTGTTGTTTCC	
CP_CCRoff-18	GGATGTCTTTGTGGTAGGGGT	TGCTGTCACTTGGGAAAAGAA	
CP_CCRoff-19	TGTCTCTAGAAGCTAATCACTTTTT	AGAATAGTCTGCAGCTCTTTCAA	
CP_CCRoff-20	CCATGTGCATCTGTGCCGC	AACCCTTGAGAGATGAGAAGAGA	
CP_CCRoff-21	GCATTTCAGGTGGTGCTGGA	AAGAGAAGAGATTCCTTGGGGG	
CP_CCRoff-22	AAAGCACCGACAAGCTCCG	AAAGACACCCCCAGTCTGCT	
CP_CCRoff-23	TGATTTTGAGAGCATTGATTTTCAT	AGCACAGCGTCAGTGATCT	
CP_CCRoff-24	AGGAAAAACAAAGTTGAGTGAGA	TAGCTAGAATCAGTTAAGTTCCTGT	
CP_CCRoff-25	GATTGTGCCACTGCACTCCA	CCCATCCCTCCTCCACCCTA	
CP_CCRoff-26	CCTGTGAAGTGTGCCGAAGA	TTTCTTCAGGAGGGGTGCCT	
CP_CCRoff-27	TGTGATCCCCACAGAAGCCA	ACCAAGGGTCTCCTCTGTAACC	
CP_CCRoff-28	AGGGTCATCTGGGAATATAGCTC	TCTGGCTGGAAATTCAAAATGA	
CP_CCRoff-29	CACTAATCATCAGGGAAATGCAA	TGAAACTGTTTTCCATAGACGTAGT	
CP_CCRoff-30	ACCATCTGTGGTAGGCAGAAT	GCTCACCTGATAACCCAGGC	
CP_CCRoff-31	CTGTCCATGGTTGATGGATGCT	CCCCCTTCTCTTTTCACTCCCT	

Table 1.24 (Continued). **Oligonucleotides used in this study**

Table 1.24 (Continued). **Oligonucleotides used in this study**. (A) All oligonucleotides were purchased from Integrated DNA Technologies. '/5Phos/' indicates 5' phosphorylated oligonucleotides. A % symbol indicates that the preceding nucleotide was incorporated as a mixture of phosphoramidites consisting of 79 mol% of the phosphoramidite corresponding to the preceding nucleotide and 7 mol% of each of the other three canonical phosphoramidites. An (*) indicates that the oligonucleotide primer was specific to a selection sequence (either CCR5A, ATM or CCR5B). An (**) indicates that the oligonucleotide adapter or primer had a unique sequence identifier to distinguish between different samples (selection conditions or cellular TALEN treatment). (B) Combinations of oligonucleotides used to construct discrete DNA substrates used in TALEN digestion assays. (C) Primer pairs for PCR amplifying on-target and off-target genomic sites. +DMSO: DMSO was used in the PCR; ND: no correct DNA product was detected from the PCR reaction.

TALEN Construction

The canonical TALEN plasmids were constructed by the FLASH method[12] with each TALEN targeting 10-18 base pairs. N-terminal mutations were cloned by PCR with Q5 Hot Start Master Mix (NEB) [98 ˚C, 22 s; 62 ˚C, 15 s; 72 °C, 7 min]) using phosphorylated TAL-N1fwd (for N1), phosphorylated TAL-N2fwd (for N2), or phosphorylated TAL-N3fwd (for N3) and phosphorylated TAL-Nrev as primers. 1 µL *Dpn*I (NEB) was added and the reaction was incubated at 37 °C for 30 min then M-column purified. ~25 ng of eluted DNA was blunt-end ligated intramolecularly in 10 µL 2x Quick Ligase Buffer, 1 µL of Quick Ligase (NEB) in a total volume of 20 µL at room temperature (~21 °C) for 15 min. 1 µL of this ligation reaction was transformed into Top10 chemically competent cells (Invitrogen). C-terminal domain mutations were cloned by PCR using TAL-Cifwd and TAL-Cirev primers, then Q-column purified. ~1 ng of this eluted DNA was used as the template for PCR with TAL-Cifwd and either TAL-Q3 (for Q3) or TAL-Q7 (for Q7) for primers, then Q-column purified. ~1 ng of this eluted DNA was used as the template for PCR with TAL-Cifwd and TAL-Ciirev for primers, then Q-column purified. ~1 µg of this DNA fragment was digested with *Hpa*I and *Bam*HI in 1x NEBuffer 4 and cloned[21] into ~2 µg of desired TALEN plasmid pre-digested with *Hpa*I and *Bam*HI. TALENs containing the N-terminal mutant domains, the Q3 C-terminal domains and the Q7 C-terminal will be available from Addgene.

***In Vitro* TALEN Expression**

TALEN proteins, all containing a 3xFLAG tag, were expressed by *in vitro* transcription/translation. 800 ng of TALEN-encoding plasmid or no plasmid ("empty lysate" control) was added to an *in vitro* transcription/translation reaction using the TNT® Quick Coupled Transcription/Translation System, T7 Variant (Promega) in a final volume of 20 µL at 30 °C for 1.5 h. Western blots were used to visualize protein using the anti-FLAG M2 monoclonal antibody (Sigma-Aldrich). TALEN concentrations were calculated by comparison to standard curve of 1 ng to 16 ng N-terminally FLAG-tagged bacterial alkaline phosphatase (Sigma-Aldrich).

In Vitro Selection for DNA Cleavage

Pre-selection libraries were prepared with 10 pmol of oligo libraries containing partially randomized target half-site sequences (CCR5A, ATM, or CCR5B) and fully randomized 10- to 24-bp spacer sequences (**Table 1.24**). Oligonucleotide libraries were separately circularized by incubation with 100 units of CircLigase II ssDNA Ligase (Epicentre) in 1x CircLigase II Reaction Buffer (33 mM Tris-acetate, 66 mM potassium acetate, 0.5 mM dithiothreitol, pH 7.5) supplemented with 2.5 mM $MnCl_2$ in 20 µL total for 16 h at 60 °C then incubated at 80 °C for 10 min. 2.5 µL of each circularization reaction was used as a substrate for rolling-circle amplification at 30 °C for 16 h in a 50-µL reaction using the Illustra TempliPhi 100 Amplification Kit (GE Healthcare). The resulting concatemerized libraries were quantified with Quant-iT™ PicoGreen ® dsDNA Kit (Invitrogen) and libraries with different spacer lengths were combined in an equimolar ratio.

For selections on the CCR5B sequence libraries, 500 ng of pre-selection library was digested for 2 h at 37 °C in 1x NEBuffer 3 with *in vitro* transcribed/translated TALEN plus empty lysate (30 µL total). For all CCR5B TALENs, *in vitro* transcribed/translated TALEN concentrations were quantified by Western blot (during the blot, TALENs were stored for 16 h at 4 °C) and then TALEN was added to 40 nM final concentration per monomer. For selections on CCR5A and ATM sequence libraries, the combined pre-selection library was further purified in a 300,000 MWCO spin column (Sartorius) with three 500-µL washes in 1x NEBuffer 3. 125 ng pre-selection library was digested for 30 min at 37 °C in 1x NEBuffer 3 with a total 24 µL of fresh *in vitro* transcribed/translated TALENs and empty lysate. For all CCR5A and ATM TALENs, 6 µL of *in vitro* transcription/translation left TALEN and 6 µL of right TALEN were

used, corresponding to a final concentration in a cleavage reaction of 16 nM ± 2 nM or 12 nM ± 1.5 nM for CC5A or ATM TALENs, respectively. These TALEN concentrations were quantified by Western blot performed in parallel with digestion.

For all selections, the TALEN-digested library was incubated with 1 µL of 100 µg/µL RNase A (Qiagen) for 2 min and then Q-column purified. 50 µL of purified DNA was 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 10x NEBuffer 2, and 1 µL of 5 U/µL Klenow Fragment DNA Polymerase (NEB) for 30 min at room temperature and Q-column purified. 50 µL of the eluted DNA was ligated with 2 pmol of heated and cooled #1 adapters containing barcodes corresponding to each sample (selections with different TALEN concentrations or constructs) (**Table 1.24**). Ligation was performed in 1x T4 DNA Ligase Buffer (50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM ATP, 10 mM DTT, pH 7.5) with 1 µL of 400 U/µL T4 DNA ligase (NEB) in 60 µL total volume for 16 h at room temperature, then Q-column purified.

6 µL of the eluted DNA was amplified by PCR in 150 µL total reaction volume (divided into 3x 50 µL reactions) for 14 to 22 cycles using the #2A adapter primers in **Table 1.24**. The PCR products were purified by Q-column. Each DNA sample was quantified with Quant-iT™ PicoGreen® dsDNA Kit (Invitrogen) and then pooled into an equimolar mixture. 500 ng of pooled DNA was run a 5% TBE 18-well Criterion PAGE gel (BioRad) for 30 min at 200 V and DNAs of length ~230 bp (corresponding to 1.5 target site repeats plus adapter sequences) were isolated and purified by Q-column. ~2 ng of eluted DNA was amplified by PCR for 5 to 8 cycles with #2B adapter primers (**Table 1.24**) and purified by M-column.

10 µL of eluted DNA was purified using 12 µL of AMPure XP beads (Agencourt) and quantified with an Illumina/Universal Library Quantification Kit (Kapa Biosystems). DNA was prepared for high-throughput DNA sequencing according to Illumina instructions and sequenced using a MiSeq DNA Sequencer (Illumina) using a 12 pM final solution and 156-bp paired-end reads. To prepare the pre-selection library for sequencing, the pre-selection library was digested with 1 µL to 4 µL of appropriate restriction enzyme (CCR5A = *Tsp*45I, ATM = *Acc*65I, CCR5B = *Ava*I (NEB)) for 1 h at 37 °C then ligated as described above with 2 pmol of heated and cooled #1 library adapters (**Table 1.24**). Pre-selection library DNA was prepared as described above using #2A library adapter primers and #2B library adapter primers in place of #2A adapter

primers and #2B adapter primers, respectively (**Table 1.24**). The resulting pre-selection library DNA was sequenced together with the TALEN-digested samples.

Discrete *In Vitro* TALEN Cleavage Assays

Discrete DNA substrates for TALEN digestion were constructed by combining pairs of oligonucleotides as specified in **Table 1.24** with restriction cloning[21] into pUC19 (NEB). Corresponding cloned plasmids were amplified by PCR (59 °C annealing for 15 s) for 24 cycles with pUC19Ofwd and pUC19Orev primers (**Table 1.24**) and Q-column purified. 50 ng of amplified DNAs were digested in 1x NEBuffer 3 with 3 μL each of *in vitro* transcribed/translated TALEN left and right monomers (corresponding to a ~16 nM to ~12 nM final TALEN concentration), and 6 μL of empty lysate in a total reaction volume of 120 μL. The digestion reaction was incubated for 30 min at 37 °C, then incubated with 1 μL of 100 μg/μL RNase A (Qiagen) for 2 min and purified by M-column. The entire 10 μL of eluted DNA with glycerol added to 15% was analyzed on a 5% TBE 18-well Criterion PAGE gel (Bio-Rad) for 45 min at 200 V, then stained with 1x SYBR Gold (Invitrogen) for 10 min. Bands were visualized and quantified on an AlphaImager HP (Alpha Innotech).

Cellular TALEN Cleavage Assays

TALENs were cloned into mammalian expression vectors[12] and the resulting TALEN vectors transfected into U2OS-EGFP cells as previously described.[12] Genomic DNA was isolated after 2 days as previously described.[12] For each assay, 50 ng of isolated genomic DNA was amplified by PCR [98 °C, 15 s 67.5 °C, 15 s; 72 °C, 22s] for 35 cycles with pairs of primers with or without 4% DMSO as specified in **Table 1.24**. Two PCR reactions were performed for OffC-5 to improve the limit of detection. The relative dsDNA content of the PCR reaction for each genomic site was quantified with Quant-iT™ PicoGreen ® dsDNA Kit (Invitrogen) and then pooled into an equimolar mixture, keeping no-TALEN and all TALEN-treated samples separate. DNA corresponding to 150 to 350 bp was purified by PAGE as described above.

44 μL of eluted DNA was incubated with 5 μL of 1x T4 DNA Ligase Buffer and 1 μL of 10 U/μL Polynucleotide kinase (NEB) for 30 min at 37 °C and Q-column purified. 43 μL of eluted DNA was incubated with 1 μL of 10 mM dATP (NEB), 5 μL of 10x NEBuffer 2, and 1 μL of 5 U/μL DNA Klenow Fragment (3'→ 5' exo$^-$) (NEB) for 30 min at 37 °C and purified by

M-column. 10μL of eluted DNA was ligated as above with 10 pmol of heated and cooled G (genomic) adapters (**Table 1.24**) and purified by Q-column. 8 μL of eluted DNA was amplified by PCR for 6 to 8 cycles with G-B primers containing barcodes corresponding to each sample. Each sample DNA was quantified with Quant-iT™ PicoGreen ® dsDNA Kit (Invitrogen) and then pooled into an equimolar mixture. The combined DNA was subjected to high-throughput sequencing using a MiSeq as described above.

Data Analysis

.DNA sequences can be found at the NCBI's Sequence Read Archive with Accession Cdoe SRP035232. Specificity scores were calculated as previously described.[21] Sample sizes for sequencing experiments were maximized (within practical experimental considerations) to ensure greatest power to detect effects. Statistical analysis on the distribution of number of mutations in various TALEN selections was performed as previously described[21]. Statistical analysis of TALEN modified genomic sites was performed as previously described[31] with multiple comparison correction using the Benjamini-Hochberg method.[42, 43]

To determine extrapolated mean enrichment curves mutation enrichment value as function of mutation number were fit to an exponential function, $a*e^b$, with R^2 reported utilizing the non-linear least squares method. These exponential decrease, b, were used to extrapolate all mean enrichment values beyond five mutations to determine the extrapolated mean enrichment.

Computational Filtering of Pre-selection Sequences and Selected Sequences

For Pre-selection Sequences

1) Search for 16 bp constant sequence (CCR5A = CGTCACGCTCACCACT, CCR5B = CCTCGGGACTCCACGCT, ATM = GGTACCCCACTCCGCGT) immediately after first 4 bases read (random bases), accepting only sequences with the 16bp constant sequence allowing for one mutation.

2) Search for 9 bp final sequence at a position at least the minimum possible full site length away and up to the max full site length away from constant sequence to confirm the presence of a full site, accept only sequences with this 9 bp final sequence. (Final sequence: CCR5A = CGTCACGCT, CCR5B = CCTCGGGAC, ATM = GGTACGTGC)

3) Search for instances of each half-site with the least amount of mutations from the target half-site in the full site, accept any sequences with proper left and right half-site order of left then right.

4) Determine DNA spacer sequence between the two half sites, the single flanking nucleotide to left of the left half-site and single flanking nucleotide to right of the right half-site (sequence between half sites and constant sequences).

5) Filter by sequencing read quality scores, accepting sequences with quality scores of 'A' or better across three fourths of the half site positions.

For Selected Sequences

1) Output to separate files all sequence reads and position quality scores of all sequences starting with correct 5 bp barcodes corresponding to different selection conditions.

2) Search for the initial 16 bp sequence immediately after the 5 bp barcode repeated at a psotion at least the minimum possible full site length away and up to the max full site length away from initial sequence to confirm the presence of a full site with repeated sequence, accept only sequences with a 16bp repeat allowing for 1 mutation.

3) Search for 16 bp constant sequence within the full site, accept only sequences with a constant sequence allowing for one mutation. Parse sequence to start with constant sequence plus 5' sequence to second instance of repeated sequence then initial sequence after barcode to constant sequence resulting in constant sequences sandwiching the equivalent of one full site:
CONSTANT – LFLANK – LHS – SPACER – RHS – RFLANK – CONSTANT
LFLANK = Left Flank Sequence (designed as a single random base)
LHS = Left Half Site Sequence
RHS = Right Half Site Sequence
RFLANK = Right Flank Sequence (designed as a single random base)
CONSTANT = Constant Sequence (CCR5A = CGTCACGCTCACCACT, CCR5B = CCTCGGGACTCCACGCT, ATM = GGTACCCCACTCCGCGT)

4) Search for instances of each half-site with the fewest number of mutations from the target half-site in the full site, accept any sequences with proper left and right half-site order of left then right.

5) With half site positions determine corresponding spacer (sequence between the two half sites), left flank and right flank sequences (sequence between half sites and constant sequences).

6) Determine sequence end by taking sequence from the start of read after the 5 bp barcode sequence to the beginning of the constant sequence.
SEQUENCESTART – RHS – RFLANK – CONSTANT

7) Filter by sequencing read quality scores, accepting sequences with quality scores of A or better across three fourths of the half site positions.

8) Selected sequences were filtered by sequence end, by accepting only sequences with sequence ends in the spacer that were 2.5-fold more abundant than the amount of sequence end background calculated as the mean of the number of sequences with ends zero to five base pairs into each half-site from the spacer side (sequence end background number was calculated for both half sites with the closest half site to the sequence end utilized as sequence end background for comparison).

Computational Search for Genomic Off-Target Sites Related to the CCR5B Target Site

1) The Patmatch program[18] was used to search the human genome (GRCh37/hg19 build) for pattern sequences as follows: CCR5B left half-site sequence (L16, L13 or L10) NNNNNNNNN… CCR5B right half-site sequence (R16, R13 or R10)[M,0,0] where number of Ns varied from 12 to 25 and M (indicating mutations allowed) varied from 0 to 14.

2) The number of output off-target sites were deconvoluted since the program outputs all sequences with X or fewer mutations, resulting in the number of off-target sites in the human genome that are a specific number of mutations away from the target site.

Identification of Indels in Sequences of Genomic Sites

1) For each sequence the primer sequence was used to identify the genomic site.

2) Sequences containing the reference genomic sequence corresponding to 8 bp to the left of the target site and reference genomic sequence 6 to 10 bp (or 6 bp for genomic sites at the very end of sequencing reads and 10bp for genomic sites with low complexity or highly repetitive regions) to the right of the full target site were considered target site sequences.

3) Any target site sequences corresponding to the same size as the reference genomic site were considered unmodified and any sequences not the reference size were aligned with ClustalW[19] to the reference genomic site.

4) Aligned sequences with more than two insertions or two deletions in the DNA spacer sequence between the two half-site sequences were considered indels. Since high-throughput sequencing can result in insertions or deletions of one or two base pairs (mis-phasing) at a low but relevant rate - we only considering indels of three bp that are more likely to arise from TALEN induced modifications.

1.8 References cited in TALEN specificity study

1. Moscou, M.J. & Bogdanove, A.J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326**, 1501 (2009).
2. Boch, J. et al. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509-1512 (2009).
3. Doyon, Y. et al. Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *Nature methods* **8**, 74-79 (2011).
4. Cade, L. et al. Highly efficient generation of heritable zebrafish gene mutations using homo- and heterodimeric TALENs. *Nucleic acids research* **40**, 8001-8010 (2012).
5. Miller, J.C. et al. A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **29**, 143-148 (2011).
6. Bedell, V.M. et al. In vivo genome editing using a high-efficiency TALEN system. *Nature* **491**, 114-118 (2012).
7. Hockemeyer, D. et al. Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology* **29**, 731-734 (2011).
8. Cermak, T. et al. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic acids research* **39**, e82 (2011).
9. Tesson, L. et al. Knockout rats generated by embryo microinjection of TALENs. *Nature biotechnology* **29**, 695-696 (2011).
10. Moore, F.E. et al. Improved somatic mutagenesis in zebrafish using transcription activator-like effector nucleases (TALENs). *PloS one* **7**, e37877 (2012).
11. Wood, A.J. et al. Targeted genome editing across species using ZFNs and TALENs. *Science* **333**, 307 (2011).
12. Reyon, D. et al. FLASH assembly of TALENs for high-throughput genome editing. *Nat Biotechnol* **30**, 460-465 (2012).
13. Mussolino, C. et al. A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic acids research* **39**, 9283-9293 (2011).
14. Li, T. et al. Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic acids research* **39**, 6315-6325 (2011).
15. Lei, Y. et al. Efficient targeted gene disruption in Xenopus embryos using engineered transcription activator-like effector nucleases (TALENs). *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17484-17489 (2012).
16. Kim, Y. et al. A library of TAL effector nucleases spanning the human genome. *Nature biotechnology* **31**, 251-258 (2013).
17. Dahlem, T.J. et al. Simple methods for generating and detecting locus-specific mutations induced with TALENs in the zebrafish genome. *PLoS genetics* **8**, e1002861 (2012).
18. Gabriel, R. et al. An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nature biotechnology* **29**, 816-823 (2011).
19. Osborn, M.J. et al. TALEN-based gene correction for epidermolysis bullosa. *Mol Ther* **21**, 1151-1159 (2013).
20. Mali, P. et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology* **31**, 833-838 (2013).

21. Pattanayak, V., Ramirez, C.L., Joung, J.K. & Liu, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature methods* **8**, 765-770 (2011).
22. Maeder, M.L. et al. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Molecular cell* **31**, 294-301 (2008).
23. Pattanayak, V. et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature biotechnology* **31**, 839-843 (2013).
24. Miller, J.C. et al. An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature biotechnology* **25**, 778-785 (2007).
25. Maul, G.G. & Deaven, L. Quantitative determination of nuclear pore complexes in cycling cells with differing DNA content. *The Journal of cell biology* **73**, 748-760 (1977).
26. Huang, B. et al. Counting low-copy number proteins in a single cell. *Science* **315**, 81-84 (2007).
27. Beck, M. et al. The quantitative proteome of a human cell line. *Mol Syst Biol* **7**, 549 (2011).
28. Witten, I.H. & Frank, E. Data mining: practical machine learning tools and techniques, Edn. 2nd. (Morgan Kaufman, San Francisco; 2005).
29. Sun, N., Liang, J., Abil, Z. & Zhao, H. Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. *Molecular bioSystems* **8**, 1255-1263 (2012).
30. Streubel, J., Blucher, C., Landgraf, A. & Boch, J. TAL effector RVD specificities and efficiencies. *Nature biotechnology* **30**, 593-595 (2012).
31. Sander, J.D. et al. In silico abstraction of zinc finger nuclease cleavage profiles reveals an expanded landscape of off-target sites. *Nucleic acids research* **41**, e181 (2013).
32. Perez, E.E. et al. Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature biotechnology* **26**, 808-816 (2008).
33. Meckler, J.F. et al. Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic acids research* **41**, 4118-4128 (2013).
34. Christian, M.L. et al. Targeting G with TAL effectors: a comparison of activities of TALENs constructed with NN and NK repeat variable di-residues. *PloS one* **7**, e45383 (2012).
35. Cong, L., Zhou, R., Kuo, Y.C., Cunniff, M. & Zhang, F. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature communications* **3**, 968 (2012).
36. Kim, Y., Kweon, J. & Kim, J.S. TALENs and ZFNs are associated with different mutation signatures. *Nature methods* **10**, 185 (2013).
37. Grau, J., Boch, J. & Posch, S. TALENoffer: genome-wide TALEN off-target prediction. *Bioinformatics* **29**, 2931-2932 (2013).
38. Ding, Q. et al. A TALEN genome-editing system for generating human stem cell-based disease models. *Cell stem cell* **12**, 238-251 (2013).
39. McNaughton, B.R., Cronican, J.J., Thompson, D.B. & Liu, D.R. Mammalian cell penetration, siRNA transfection, and DNA transfection by supercharged proteins. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 6111-6116 (2009).
40. Gao, H., Wu, X., Chai, J. & Han, Z. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res* **22**, 1716-1720 (2012).

41. Mahfouz, M.M. et al. De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. *Proc Natl Acad Sci U S A* **108**, 2623-2628 (2011).
42. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289-300 (1995).
43. Noble, W.S. How does multiple testing correction work? *Nature biotechnology* **27**, 1135-1137 (2009).

Chapter 2:
Broad Off-target DNA Cleavage Profiling Reveals RNA-Guided Cas9 Nuclease Specificity

Vikram Pattanayak, Steven Lin, John P. Guilinger, Enbo Ma, Jennifer A. Doudna, and David R. Liu

Steven Lin and Enbo Ma, in the laboratory of Jenneifer Dounda, purified Cas9 protein, preformed and analyze initial *in vitro* off-target cleavage assays. Steven Lin and Enbo Ma isolated genomic DNA from cells treated with Cas9 and gRNA expression vectors. Vikram Pattanayak, in the laboratory of David R. Liu, designed, preformed and analyzed the selection results. Together Vikram Pattanayak and I preformed and analyzed the *in vitro* off-target cleavage assays and cellular off-target modification assays.

Text in this chapter appeared in *Nature Biotechnology*, 2013, **31**, 839-843.

2.1 Introduction to RNA-guided Cas9 nuclease specificity

Sequence-specific endonucleases including zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) have become important tools to modify genes in induced pluripotent stem cells (iPSCs),[1-3] in multi-cellular organisms,[4-8] and in *ex vivo* gene therapy clinical trials.[9, 10] Although ZFNs and TALENs have proved effective for such genetic manipulation, a new ZFN or TALEN protein must be generated for each DNA target site. In contrast, the RNA-guided Cas9 endonuclease uses RNA:DNA hybridization to determine target DNA cleavage sites, enabling a single monomeric protein to cleave, in principle, any sequence specified by the guide RNA (**Figure 2.1**).[11]

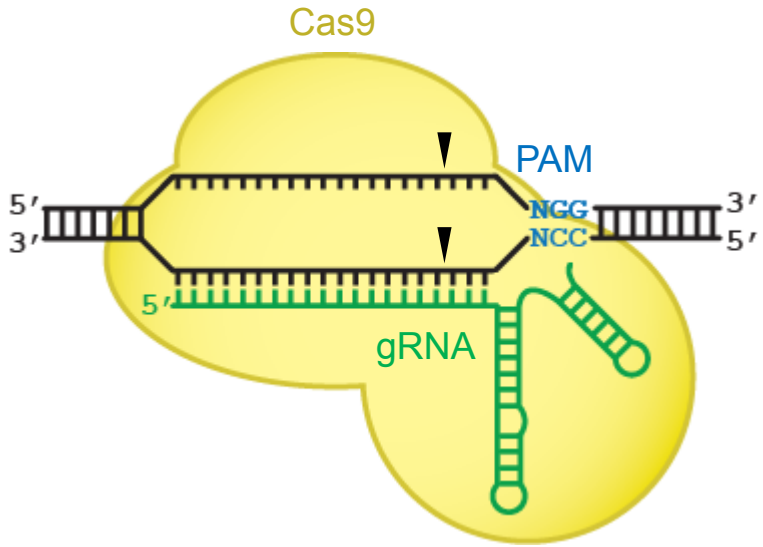


Figure 2.1. Architecture of Cas9. (a) Cas9 protein (yellow) binds to target DNA in complex with a guide RNA (gRNA, green). The *S. pyogenes* Cas9 protein recognizes the PAM sequence NGG (blue), initiating unwinding of dsDNA and gRNA:DNA base pairing. Black triangles indicate the cleavage points three bases from the PAM on both top and bottom strands.

Previous studies[12-17] demonstrated that Cas9 mediates genome editing at sites complementary to a 20-nucleotide sequence in a bound guide RNA. In addition, target sites must include a protospacer adjacent motif (PAM) at the 3' end adjacent to the 20-nucleotide target site; for *Streptococcus pyogenes* Cas9, the PAM sequence is NGG. Cas9-mediated DNA cleavage specificity both *in vitro* and in cells has been inferred previously based on assays against small collections of potential single-mutation off-target sites. These studies suggested

that perfect complementarity between guide RNA and target DNA is required in the 7-12 base pairs adjacent to the PAM end of the target site (3' end of the guide RNA) and mismatches are tolerated at the non-PAM end (5' end of the guide RNA).[11, 12, 17-19]

Although such a limited number of nucleotides specifying Cas9:guide RNA target recognition would predict multiple sites of DNA cleavage in genomes of moderate to large size ($> \sim 10^7$ bp), Cas9:guide RNA complexes have been successfully used to modify both cells[12, 13, 15] and organisms.[14] A study using Cas9:guide RNA complexes to modify zebrafish embryos observed toxicity at a rate similar to that of ZFNs and TALENs.[14] A recent, broad study of the specificity of DNA binding (transcriptional repression) in *E. coli* of a catalytically inactive Cas9 mutant using high-throughput sequencing found no detectable off-target transcriptional repression in the relatively small *E. coli* transcriptome.[20] While these studies have substantially advanced our basic understanding of Cas9, a systematic and comprehensive profile of Cas9:guide RNA-mediated DNA cleavage specificity generated from measurements of Cas9 cleavage on a large number of related mutant target sites has not been described. Such a specificity profile is needed to understand and improve the potential of Cas9:guide RNA complexes as research tools and future therapeutic agents.

2.2 Profiling the specificity of RNA-guided Cas9 nucleases

We modified our previously published *in vitro* selection,[21] adapted to process the blunt-ended cleavage products produced by Cas9 instead of the overhang-containing products of ZFN cleavage, to determine the off-target DNA cleavage profiles of Cas9:single guide RNA (sgRNA)[11] complexes. Each selection experiment used DNA substrate libraries containing $\sim 10^{12}$ sequences, a size sufficiently large to include ten-fold coverage of all sequences with eight or fewer mutations relative to each 22-base pair target sequence (including the two-base pair PAM) (**Figure 2.1**). We used partially randomized nucleotide mixtures at all 22 target-site base pairs to create a binomially distributed library of mutant target sites with an expected mean of 4.62 mutations per target site. In addition, target site library members were flanked by four fully randomized base pairs on each side to test for specificity patterns beyond those imposed by the canonical 20-base pair target site and PAM.

Pre-selection libraries of 10^{12} individual potential off-target sites were generated for each of four different target sequences in the human clathrin light chain A (*CLTA*) gene (**Figure 2.2**).

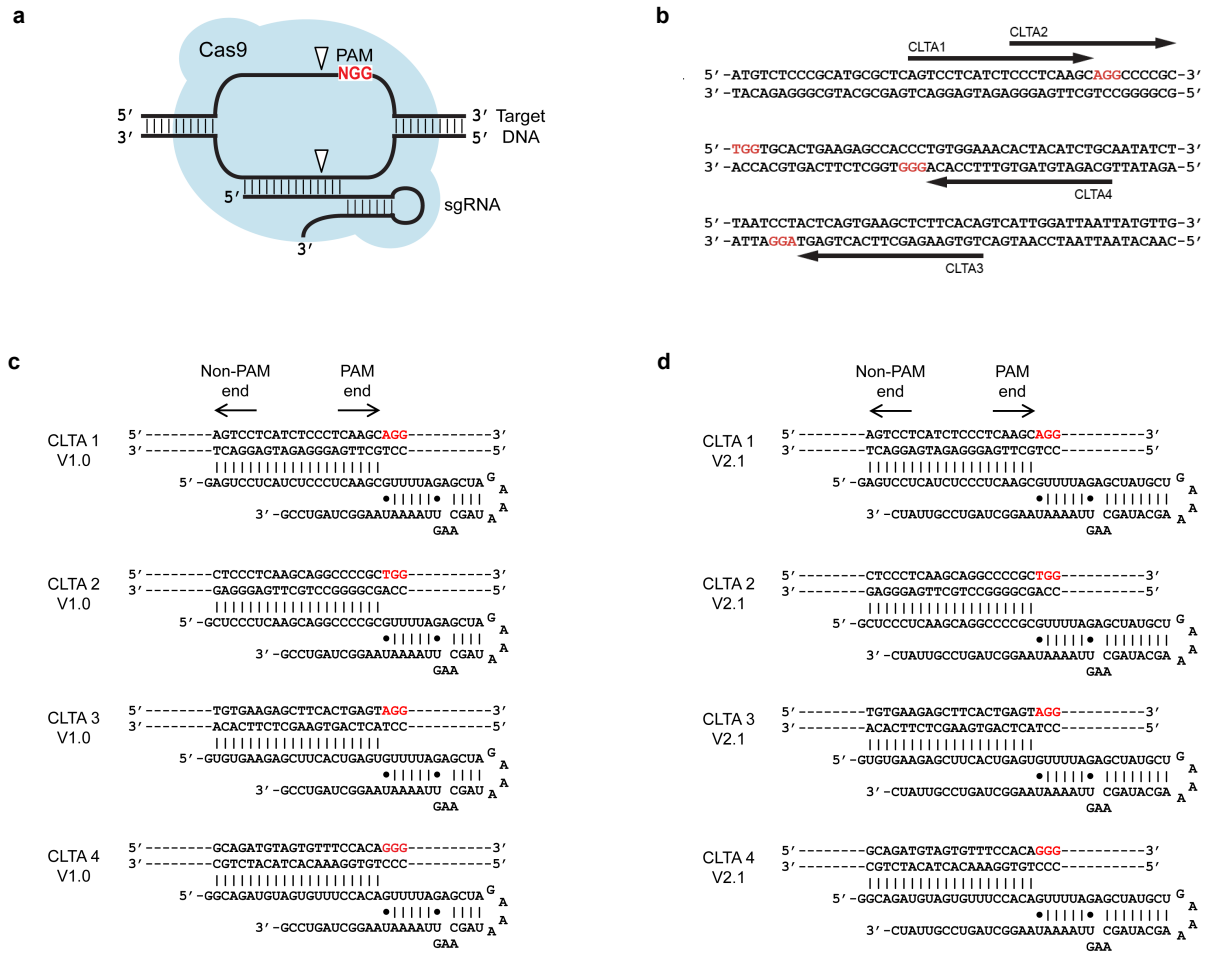


Figure 2.2. Target sites profiled in human *CLTA* gene. (**a**) The 5' end of the sgRNA has 20 nucleotides that are complementary to the target site. The target site contains an NGG motif (PAM) adjacent to the region of RNA:DNA complementarity. (**b**) Four human clathrin gene (CLTA) target sites are shown. (**c, d**) Four human clathrin gene (CLTA) target sites are shown with sgRNAs. sgRNA v1.0 is shorter than sgRNA v2.1. The PAM is shown in red for each site. The non-PAM end of the target site corresponds to the 5' end of the sgRNA.

Synthetic 5'-phosphorylated 53-base oligonucleotides were self-ligated into circular single-stranded DNA *in vitro*, then converted into concatemeric 53-base pair repeats through rolling-circle amplification. The resulting pre-selection libraries were incubated with their corresponding Cas9:sgRNA complexes. Cleaved library members containing free 5' phosphates were separated from intact library members through the 5' phosphate-dependent ligation of non-phosphorylated double-stranded sequencing adapters. The ligation-tagged post-selection libraries were amplified by PCR. The PCR step generated a mixture of post-selection DNA fragments containing 0.5, 1.5, or 2.5, etc. repeats of library members cleaved by Cas9, resulting

from amplification of an adapter-ligated cut half-site with or without one or more adjacent corresponding full sites (**Figure 2.3**).

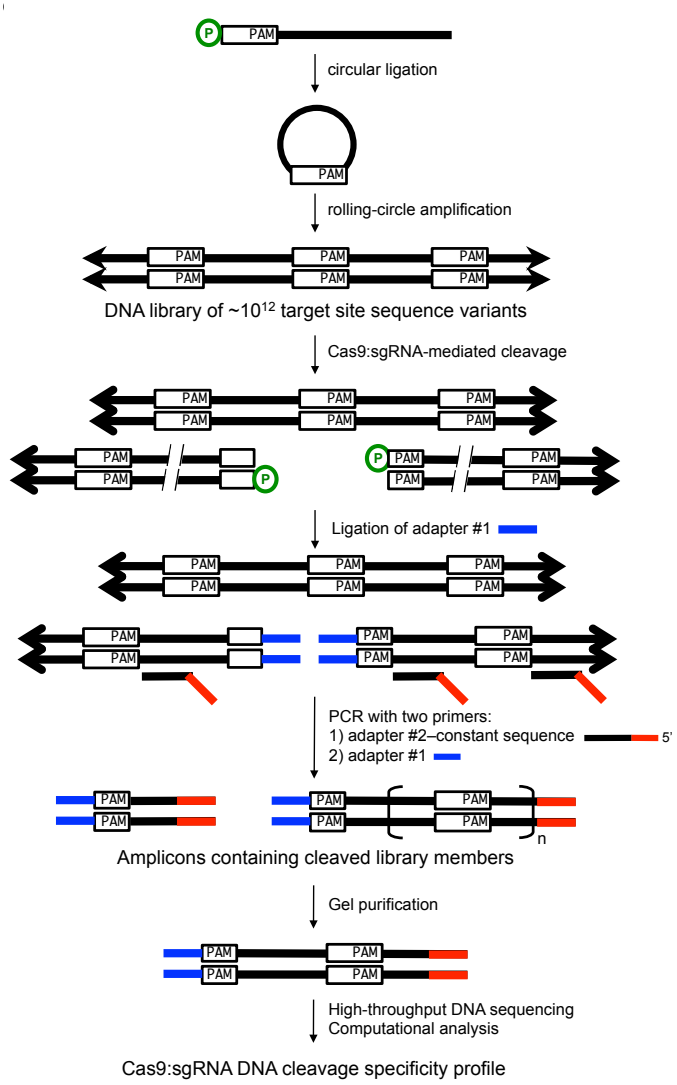


Figure 2.3. ***In vitro*** **selection overview.** A modified version of our previously described *in vitro* selection was used to comprehensively profile Cas9 specificity. A concatemeric pre-selection DNA library in which each molecule contains one of 10^{12} distinct variants of a target DNA sequence (white rectangles) was generated from synthetic DNA oligonucleotides by ligation and rolling-circle amplification. This library was incubated with a Cas9:sgRNA complex of interest. Cleaved library members contain 5' phosphate groups (green circles) and therefore are substrates for adapter ligation and PCR. The resulting amplicons were subjected to high-throughput DNA sequencing and computational analysis.

Post-selection library members with 1.5 target-sequence repeats were isolated by gel purification and analyzed by high-throughput sequencing. In a final computational selection step to

minimize the impact of errors during DNA amplification or sequencing, only sequences with two identical copies of the repeated cut half-site were analyzed.

Pre-selection libraries were incubated under enzyme-limiting conditions (200 nM target site library, 100 nM Cas9:sgRNA v2.1) or enzyme-excess conditions (200 nM target site library, 1000 nM Cas9:sgRNA v2.1) for each of the four guide RNAs targets tested (CLTA1, CLTA2, CLTA3, and CLTA4) (**Figure 2.2**). A second guide RNA construct, sgRNA v1.0, which is less active than sgRNA v2.1, was assayed under enzyme-excess conditions alone for each of the four guide RNA targets tested (200 nM target site library, 1000 nM Cas9:sgRNA v1.0). The two guide RNA constructs differ in their length (**Figure 2.2**) and in their DNA cleavage activity level under the selection conditions, consistent with previous reports[15] (**Figure 2.4).**

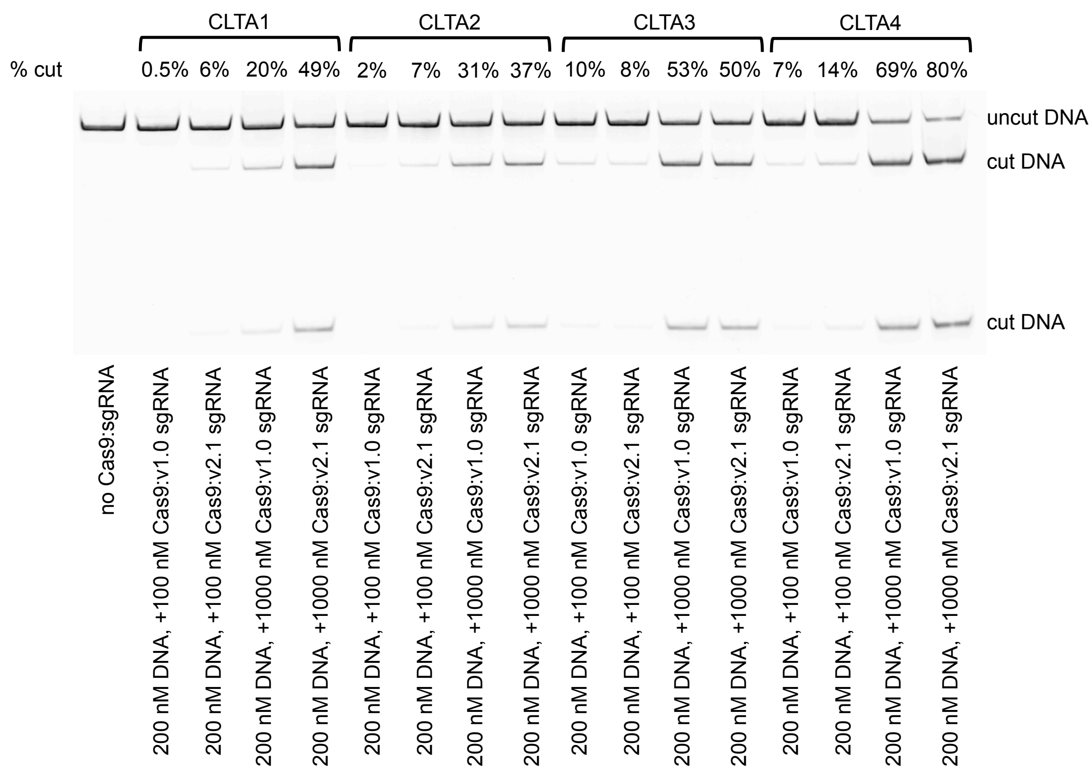


Figure 2.4. **Cas9:guide RNA cleavage of on-target DNA sequences *in vitro***. Discrete DNA cleavage assays on an approximately 1-kb linear substrate were performed with 200 nM on-target site and 100 nM Cas9:v1.0 sgRNA, 100 nM Cas9:v2.1 sgRNA, 1000 nM Cas9:v1.0 sgRNA, and 1000 nM Cas9:v2.1 sgRNA for each of four CLTA target sites. For CLTA1, CLTA2, and CLTA4, Cas9:v2.1 sgRNA shows higher activity than Cas9:v1.0 sgRNA. For CLTA3, the activities of the Cas9:v1.0 sgRNA and Cas9:v2.1 sgRNA were comparable.

Both pre-selection and post-selection libraries were characterized by high-throughput DNA sequencing and computational analysis. As expected, library members with fewer mutations were significantly enriched in post-selection libraries relative to pre-selection libraries.

We calculated specificity scores to quantify the enrichment level of each base pair at each position in the post-selection library relative to the pre-selection library, normalized to the maximum possible enrichment of that base pair. Positive specificity scores indicate base pairs that were enriched in the post-selection library and negative specificity scores indicate base pairs that were de-enriched in the post-selection library. For example, a score of +0.5 indicates that a base pair is enriched to 50% of the maximum enrichment value, while a score of –0.5 indicates that a base pair is de-enriched to 50% of the maximum de-enrichment value.

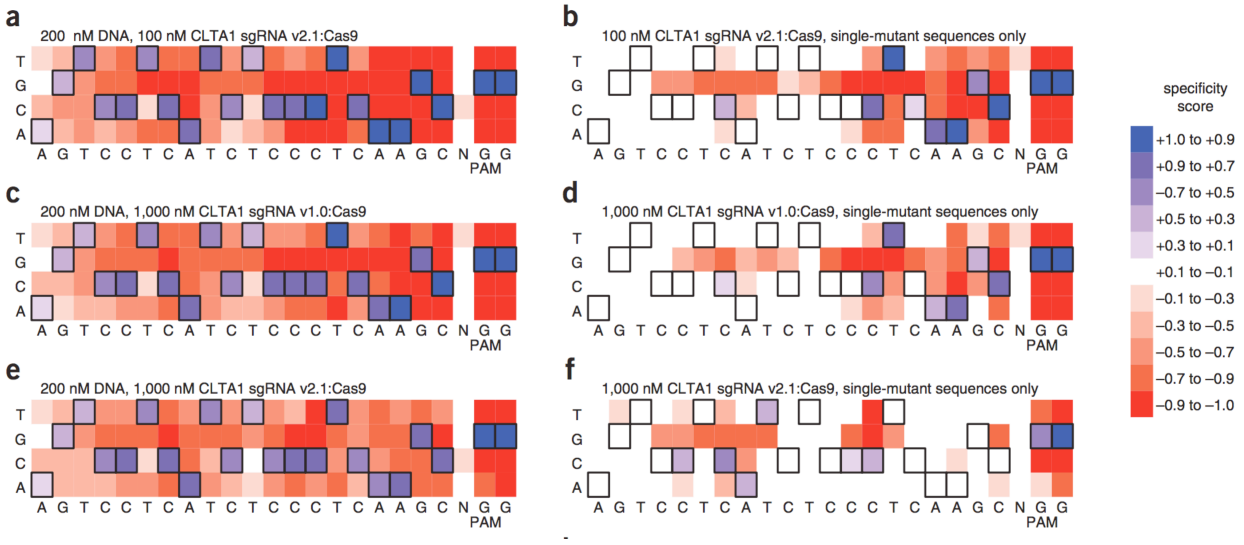


Figure 2.5. ***In vitro*** **selection results for Cas9:CLTA1 sgRNA**. Heat maps[21] show the specificity profiles of Cas9:CLTA1 sgRNA v2.1 under enzyme-limiting conditions (**a, b**), Cas9:CLTA1 sgRNA v1.0 under enzyme-excess conditions (**c, d**), and Cas9:CLTA1 sgRNA v2.1 under enzyme-excess conditions (**e, f**). Heat maps show all post-selection sequences (**a, c, e**) or only those sequences containing a single mutation in the 20-base pair sgRNA-specified target site and two-base pair PAM (**b, d, f**). 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.

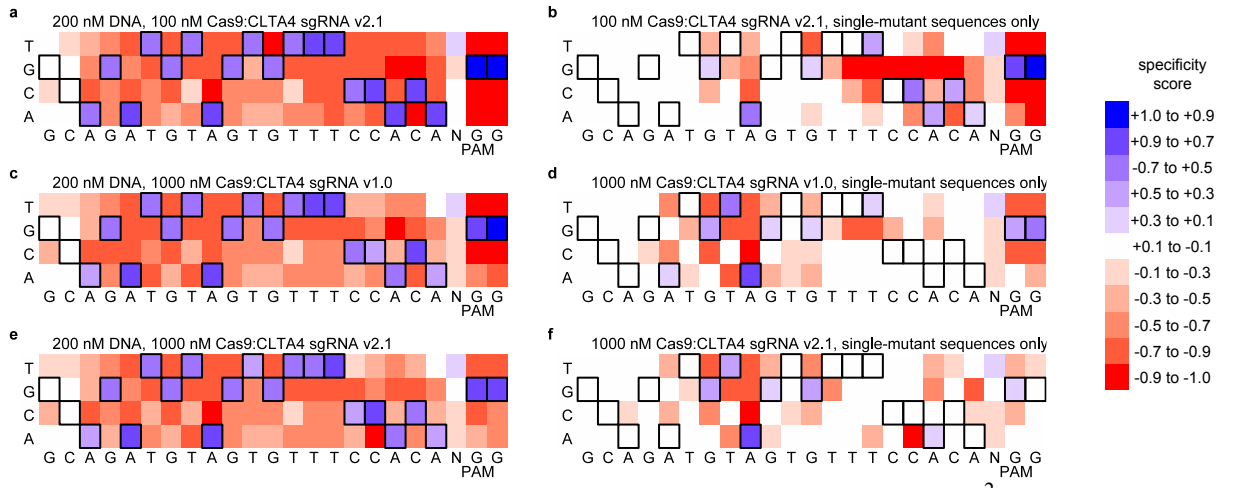


Figure 2.6. ***In vitro*** **selection results for Cas9:CLTA4 sgRNA**. Heat maps[2] show the specificity profiles of Cas9:CLTA4 sgRNA v2.1 under enzyme-limiting conditions (**a, b**), Cas9:CLTA4 sgRNA v1.0 under enzyme-excess conditions (**c, d**), and Cas9:CLTA4 sgRNA v2.1 under enzyme- saturating conditions (**e, f**). Heat maps show all post-selection sequences (**a, c, e**) or only those sequences containing a single mutation in the 20-base pair sgRNA-specified target site and two-base pair PAM (**b, d, f**). 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.

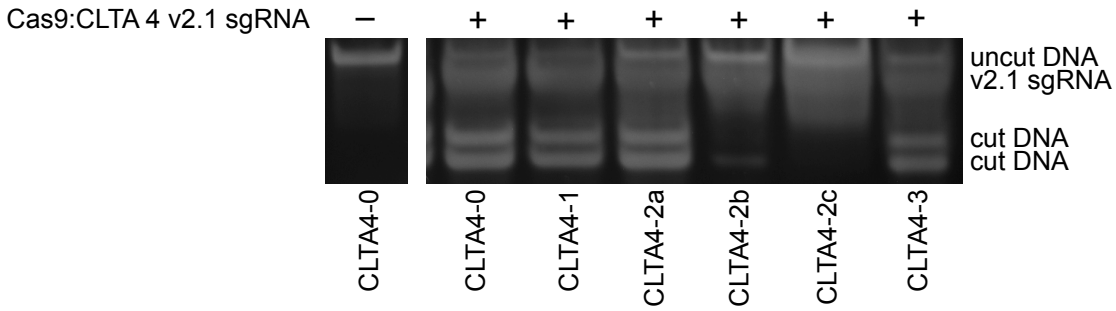
In addition to the two base pairs specified by the PAM, all 20 base pairs targeted by the guide RNA were enriched in the sequences from the CLTA1 selections (**Figure 2.5**). For the CLTA4 selections (**Figure 2.6**), guide RNA-specified base pairs were enriched at all positions except for the two most distal base pairs from the PAM (5' end of the guide RNA), respectively. At these non-specified positions farthest from the PAM, at least two of the three alternate base pairs were nearly as enriched as the specified base pair. Our finding that the entire 20 base-pair target site and two base pair PAM can contribute to Cas9:sgRNA DNA cleavage specificity contrasts with the results from previous single-substrate assays suggesting that only 7-12 base pairs and two base pair PAM are specified.[11, 12, 15]

All single-mutant pre-selection ($n \geq 14{,}569$) and post-selection library members ($n \geq 103{,}660$) were computationally analyzed to provide a selection enrichment value for every possible single-mutant sequence. The results of this analysis (**Figure 2.5 and Figure 2.6**) show that when only single-mutant sequences are considered, the six to eight base pairs closest to the PAM are generally highly specified and the non-PAM end is poorly specified under enzyme-limiting conditions, consistent with previous findings.[11, 12, 17-19] Under enzyme-excess conditions, however, single mutations even in the six to eight base pairs most proximal to the

PAM are tolerated, suggesting that the high specificity at the PAM end of the DNA target site can be compromised when enzyme concentrations are high relative to substrate (**Figure 2.5 and Figure 2.6**). The observation of high specificity against single mutations close to the PAM only applies to sequences with a single mutation and the selection results do not support a model in which any combination of mutations is tolerated in the region of the target site farthest from the PAM.

Importantly, the selection results also reveal that the choice of guide RNA hairpin affects cleavage of off-target sites. The shorter, less-active sgRNA v1.0 constructs are less tolerant of mutations than the longer, more-active sgRNA v2.1 constructs when assayed under identical, enzyme-excess conditions that reflect an excess of enzyme relative to substrate in a cellular context (**Figure 2.5 and Figure 2.6**). Thus, these results indicate that different guide RNA architectures result in different off-target DNA cleavage activities, and that guide RNA-dependent changes in specificity do not affect all positions in the target site equally. Given the inverse relationship between Cas9:sgRNA concentration and specificity described above, we speculate that the differences in off-target activities between guide RNA architectures arises from differences in their overall level of DNA-cleavage activities.

To confirm that the *in vitro* selection results accurately reflect the cleavage behavior of Cas9 *in vitro*, we performed discrete cleavage assays of six CLTA4 off-target substrates containing one to three mutations in the target site. We calculated enrichment values for all sequences in the post-selection libraries for the Cas9:CLTA4 v2.1 sgRNA under enzyme-excess conditions by dividing the abundance of each sequence in the post-selection library by the calculated abundance in the pre-selection library. Under enzyme-excess conditions, the single one, two, and three mutation sequences with the highest enrichment values (27.5, 43.9, and 95.9) were cleaved to $\geq$ 71% completion (**Figure 2.7**).



	sequence	*In vitro* selection enrichment value	% cut
CLTA4-0	GCAGATGTAGTGTTTCCACAGGG	7.9	85%
CLTA4-1	GaAGATGTAGTGTTTCCACAGGG	27.5	84%
CLTA4-2a	GaAGATGTAGTGTTTCCACtGGG	43.9	79%
CLTA4-2b	GCAGATGgAGgGTTTCCACAGGG	1.0	35%
CLTA4-2c	GCAGATGTAGTGTTaCCAgAGGG	0.064	none detected
CLTA4-3	GggGATGTAGTGTTTCCACtGGG	95.9	72%

Figure 2.7. **Cas9:guide RNA cleavage of off-target DNA sequences *in vitro***. Discrete DNA cleavage assays on a 96-bp linear substrate were performed with 200 nM DNA and 1000 nM Cas9:CLTA4 v2.1 sgRNA for the on-target CLTA4 site (CLTA4-0) and five CLTA4 off-target sites identified by *in vitro* selection. Enrichment values shown are from the *in vitro* selection with 1000 nM Cas9:CLTA4 v2.1 sgRNA. CLTA4-1 and CLTA4-3 were the most highly enriched sequences under these conditions. CLTA4-2a, CLTA4-2b, and CLTA4-2c are two-mutation sequences that represent a range of enrichment values from high enrichment to no enrichment to high de-enrichment. Red lowercase letters indicate mutations relative to the on-target CLTA4 site. The enrichment values are qualitatively consistent with the observed amount of cleavage *in vitro*.

A two-mutation sequence with an enrichment value of 1.0 was cleaved to 35%, and a two-mutation sequence with an enrichment value near zero (0.064) was not cleaved. The three-mutation sequence, which was cleaved to 77% by CLTA4 v2.1 sgRNA, was cleaved to a lower efficiency of 53% by CLTA4 v1.0 sgRNA (**Figure 2.8**).

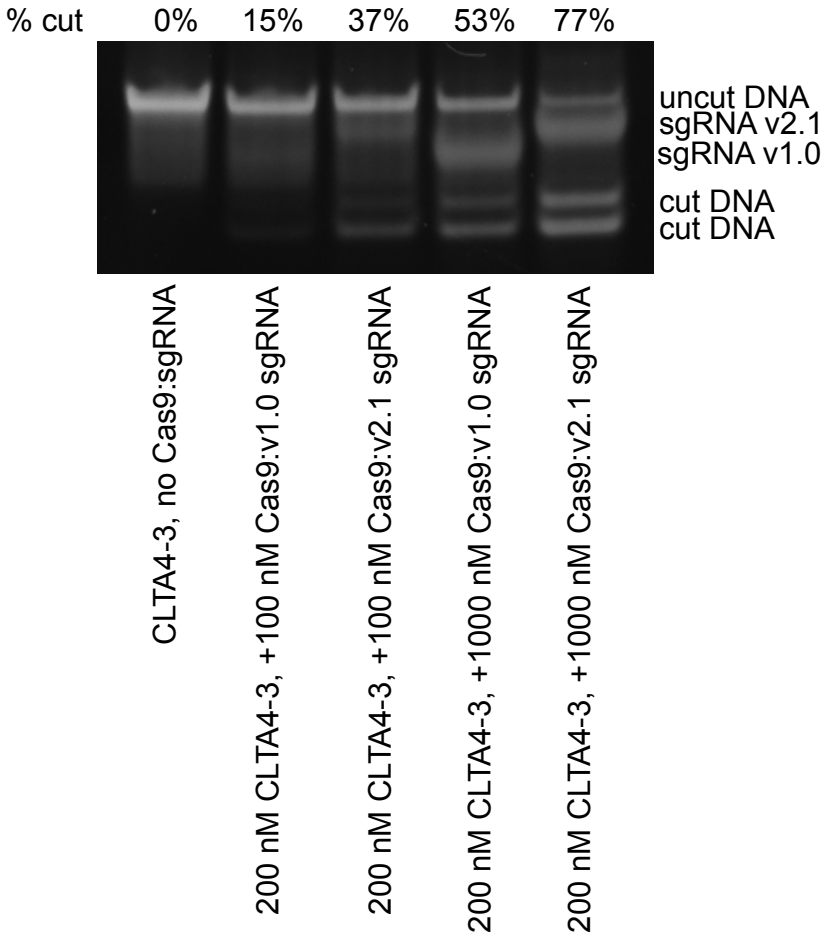


Figure 2.8. **Effect of guide RNA architecture and Cas9:sgRNA concentration on *in vitro* cleavage of an off-target site**. Discrete DNA cleavage assays on a 96-bp linear substrate were performed with 200 nM DNA and 100 nM Cas9:v1.0 sgRNA, 100 nM Cas9:v2.1 sgRNA, 1000 nM Cas9:v1.0 sgRNA, or 1000 nM Cas9:v2.1 sgRNA for the CLTA4-3 off-target site (5' GggGATGTAGTGTTTCCACtGGG - mutations are shown in lowercase letters). DNA cleavage is observed under all four conditions tested, and cleavage rates are higher under enzyme-excess conditions, or with v2.1 sgRNA compared with v1.0 sgRNA.

These results indicate that the selection enrichment values of individual sequences are predictive of *in vitro* cleavage efficiencies.

To determine if results of the *in vitro* selection and *in vitro* cleavage assays pertain to Cas9:guide RNA activity in human cells, we identified 51 off-target sites (19 for CLTA1 and 32 for CLTA4) containing up to eight mutations that were both enriched in the *in vitro* selection and present in the human genome. We expressed Cas9:CLTA1 sgRNA v1.0, Cas9:CLTA1 sgRNA v2.1, Cas9:CLTA4 sgRNA v1.0, Cas9:CLTA4 sgRNA v2.1, or Cas9 without sgRNA in HEK293T cells by transient transfection and used genomic PCR and high-throughput DNA sequencing to look for evidence of Cas9:sgRNA modification at 46 of the 51 off-target sites as well as at the on-target loci; no specific amplified DNA was obtained for five of the 51 predicted off-target sites (three for CLTA1 and two for CLTA4).

Deep sequencing of genomic DNA isolated from HEK293T cells treated with Cas9:CLTA1 sgRNA or Cas9:CLTA4 sgRNA identified sequences evident of non-homologous end-joining (NHEJ) at the on-target sites and at five of the 49 tested off-target sites (CLTA1-1-1, CLTA1-2-2, CLTA4-3-1, CLTA4-3-3, and CLTA4-4-8). The CLTA4 target site was modified by Cas9:CLTA4 v2.1 sgRNA at a frequency of 76%, while off-target sites, CLTA4-3-1 CLTA4-3-3, and CLTA4-4-8, were modified at frequencies of 24%, 0.47% and 0.73%, respectively. The CLTA1 target site was modified by Cas9:CLTA1 v2.1 sgRNA at a frequency of 0.34%, while off-target sites, CLTA1-1-1 and CLTA1-2-2, were modified at frequencies of 0.09% and 0.16%, respectively (**Table 2.1**).

a

				in vitro enrichment		modification frequency in HEK293T cells			P-value	
	number of mutations	sequence	gene	v1.0	v2.1	no sgRNA	v1.0	v2.1	v1.0	v2.1
CLTA1-0-1	0	AGTCCTCATCTCCCTCAAGCAGG	CLTA	41.4	23.3	0.003%	0.042%	0.337%	1.1E-05	6.9E-55
CLTA1-1-1	1	AGTCCTCAaCTCCCTCAAGCAGG	TUSC3	25.9	14	0.003%	0.031%	0.091%	2.6E-03	2.0E-10
CLTA1-2-1	2	AGcCCTCATtTCCCTCAAGCAGG	CACNA2D3	15.4	26.2	0%	0%	0%		
CLTA1-2-2	2	AcTCCTCATCcCCCTCAAGCCGG	ACAN	29.2	18.8	0.014%	0.005%	0.146%		4.6E-08
CLTA1-2-3	2	AGTCaTCATCTCCCTCAAGCAGa		0.06	1.27	n.t.	n.t.	n.t.		
CLTA1-3-1	3	cGTCCTCcTCTCCCcCAAGCAGG		0	2.07	0.004%	0%	0%		
CLTA1-3-2	3	tGTCCTCtTCTCCCTCAAGCAGa	BC029598	0	1.47	0%	0%	0%		
CLTA1-4-1	4	AagCtTCATCTCtCTCAAGCTGG				0%	0.006%	0%		
CLTA1-4-2	4	AGTaCTCtTtTCCCTCAgGCTGG	ENTPD1			0.007%	0.004%	0.015%		
CLTA1-4-3	4	AGTCtTaAatTCCCTCAAGCAGG				0.003%	0%	0.001%		
CLTA1-4-4	4	AGTgCTCATCTaCCagAAGCTGG				0.008%	0%	0%		
CLTA1-4-5	4	ccTCCTCATCTCCCTgcAGCAGG				0.013%	0.004%	0%		
CLTA1-4-6	4	ctaCaTCATCTCCCTCAAGCTGG				0%	0.008%	0.016%		
CLTA1-4-7	4	gGTCCTCATCTCCCTaAAaCAGa	POLQ (coding)			0.011%	0.037%	0%		
CLTA1-4-8	4	tGTCCTCATCggCCTCAgGCAGG				0%	0%	0.016%		
CLTA1-5-1	5	AGaCacCATCTCCCTtgAGCTGG	PSAT1			0%	0.003%	0.006%		
CLTA1-5-2	5	AGgCaTCATCTaCaTCAAGtTGG				0%	0%	0%		
CLTA1-5-3	5	AGTaaTCActTCCaTCAAGCCGG	ZDHHC3, EXOSC7			n.t.	n.t.	n.t.		
CLTA1-5-4	5	tccCCTCAcCTCCCTaAAGCAGG				0.008%	0.029%	0.004%		
CLTA1-5-5	5	tGTCtTtATtTCCCTCtAGCTGG				0%	0%	0.009%		
CLTA1-6-1	6	AGTCCTCATCTCCCTCAAGCAGG				0%	0%	0%		

Table 2.1 (Continued). Cellular modification induced by Cas9:CLTA1 sgRNA and Cas9:CLTA4 sgRNA.

b

				in vitro enrichment		modification frequency in HEK293T cells			P-value	
	# of mutations	sequence	gene	v1.0	v2.1	no sgRNA	v1.0	v2.1	v1.0	v2.1
CLTA4-0-1	0	GCAGATGTAGTGTTTCCACAGGG	CLTA	20	7.95	0.021%	11%	76%	<1E-55	<1E-55
CLTA4-3-1	3	aCAtATGTAGTaTTTCCACAGGG		16.5	12.5	0.006%	0.055%	24%	6.0E-04	<1E-55
CLTA4-3-2	3	GCAtATGTAGTGTTTCCAaATGt		2.99	6.97	0.017%	0%	0.014%		
CLTA4-3-3	3	cCAGATGTAGTaTTcCCACAGGG	CELF1	1.00	4.95	0%	0%	0.469%		2.5E-21
CLTA4-3-4	3	GCAGtTtTAGTGTTTtCACAGGG	BC073807	0.79	3.12	0%	0%	0%		
CLTA4-3-5	3	GCAGAgtTAGTGTTTCCACACaG	MPPED2	0	1.22	0.005%	0.015%	0.018%		
CLTA4-3-6	3	GCAGATGgAGgGTTTtCACAGGG	DCHS2	1.57	1.17	0.015%	0.023%	0.021%		
CLTA4-3-7	3	GgAaATtTAGTGTTTCCACAGGG		0.43	0.42	0.005%	0.012%	0.003%		
CLTA4-4-1	4	aaAGAaGTAGTaTTTCCACATGG				n.t.	n.t.	n.t.		
CLTA4-4-2	4	aaAGATGTAGTcaTTCCACAAGG				0.004%	0%	0.005%		
CLTA4-4-3	4	aaAtATGTAGTcTTTCCACAGGG				0.004%	0.009%	0%		
CLTA4-4-4	4	atAGATGTAGTGTTTCCAaAGGa	NR1H4			0.032%	0.006%	0.052%		
CLTA4-4-5	4	cCAGAgGTAGTGcTcCCACAGGG				0.005%	0.006%	0.007%		
CLTA4-4-6	4	cCAGATGTgagGTTTCCACAAGG	XKR6			0.018%	0%	0.007%		
CLTA4-4-7	4	ctAcATGTAGTGTTTCCAtATGG	HKR1			0.006%	0%	0.008%		
CLTA4-4-8	4	ctAGATGaAGTGcTTCCACATGG	CDK8			0.009%	0.013%	0.730%		9.70E-21
CLTA4-4-9	4	GaAaATGgAGTGTTTaCACATGG				0%	0%	0.004%		
CLTA4-4-10	4	GCAaATGaAGTGTcaCCACAAGG				0.004%	0%	0%		
CLTA4-4-11	4	GCAaATGTAtTaTTTCCACtAGG	NOV			0%	0.00%	0%		
CLTA4-4-12	4	GCAGATGTAGctTTTgtACATGG				0%	0.00%	0%		
CLTA4-4-13	4	GCAGcTtaAGTGTTTtCACATGG	GRHL2			0.020%	0.02%	0.030%		
CLTA4-4-14	4	ttAcATGTAGTGTTTaCACACGG	LINC00535			n.t.	n.t.	n.t.		
CLTA4-5-1	5	GaAGAgGaAGTGTTTgCcCAGGG	RNH1			0.004%	0.01%	0.006%		
CLTA4-5-2	5	GaAGATGTgGaGTTgaCACATGG	FZD3			0.004%	0.00%	0%		
CLTA4-5-3	5	GCAGAaGTAcTGTTgttACAAGG				0.002%	0.00%	0.003%		
CLTA4-5-4	5	GCAGATGTgGaaTTaCaACAGGG	SLC9A2			0%	0.00%	0%		
CLTA4-5-5	5	GCAGtcaTAGTGTaTaCACATGG				0.004%	0.00%	0.005%		
CLTA4-5-6	5	taAGATGTAGTaTTTCCAaAAGt				0.007%	0.01%	0%		
CLTA4-6-1	6	GCAGcTGgcaTtTcTCCACACGG				n.t.	n.t.	n.t.		
CLTA4-6-2	6	GgAGATcTgaTGgTTCtACAAGG				0.007%	0.00%	0.009%		
CLTA4-6-3	6	taAaATGcAGTGTaTCCAtATGG	SMA4			0.015%	0.00%	0%		
CLTA4-7-1	7	GCcagaaTAGTtTTTCaACAAGG	SEPHS2			0%	0.00%	0.007%		
CLTA4-7-2	8	ttgtATtTAGaGaTTgCACAAGG	RORB			0%	0.00%	0%		

Table 2.1 (Continued). Cellular modification induced by Cas9:CLTA1 sgRNA and Cas9:CLTA4 sgRNA. (**a**) 20 human genomic DNA sequences were identified that were enriched in the Cas9:CLTA1 v2.1 sgRNA *in vitro* selections under enzyme-limiting or enzyme-excess conditions. Sites shown in red contain insertions or deletions (indels) that are consistent with significant Cas9:sgRNA-mediated modification in HEK293T cells. *In vitro* enrichment values for selections with Cas9:CLTA1 v1.0 sgRNA or Cas9:CLTA1 v2.1 sgRNA are shown for sequences with three or fewer mutations. Enrichment values were not calculated for sequences with four or more mutations due to low numbers of *in vitr*o selection sequence counts. Modification frequencies (number of sequences with indels divided by total number of sequences) in HEK293T cells treated with Cas9 without sgRNA ("no sgRNA"), Cas9 with CLTA1 v1.0 sgRNA, or Cas9 with CLTA1 v2.1 sgRNA. P-values are listed for those sites that show significant modification in v1.0 sgRNA- or v2.1 sgRNA-treated cells compared to cells treated with Cas9 without sgRNA. P-values were calculated using a one-sided Fisher exact text. "Not tested (n.t.)" indicates that PCR of the genomic sequence failed to provide specific amplification products. (**b**) Same as (**a**) for CLTA4 sgRNA with 33 human genomic DNA sequences were identified that were enriched in the Cas9:CLTA4 v2.1 sgRNA *in vitro* selections under enzyme-limiting or enzyme-excess conditions

Under enzyme-excess conditions with the v2.1 sgRNA, the two verified CLTA1 off-target sites, CLTA1-1-1 and CLTA1-2-2, were two of the three most highly enriched sequences identified in the *in vitro* selection. CLTA4-3-1 and CLTA4-3-3 were the highest and third-highest enriched sequences of the seven CLTA4 three-mutation sequences enriched in the *in*

vitro selection that are also present in the genome. The *in vitro* selection enrichment values of the four-mutation sequences were not calculated, since 12 out of the 14 CLTA4 sequences in the genome containing four mutations, including CLTA4-4-8, were observed at a level of only one sequence count in the post-selection library. Taken together, these results confirm that several of the off-target substrates identified in the *in vitro* selection that are present in the human genome are indeed cleaved by Cas9:sgRNA complexes in human cells, and also suggest that the most highly enriched genomic off-target sequences in the selection are modified in cells to the greatest extent.

The off-target sites we identified in cells were among the most-highly enriched in our *in vitro* selection and contain up to four mutations relative to the intended target sites. While it is possible that heterochromatin or covalent DNA modifications could diminish the ability of a Cas9:guide RNA complex to access genomic off-target sites in cells, the identification of five out of 49 tested cellular off-target sites in this study, rather than zero or many, strongly suggests that Cas9-mediated DNA cleavage is not limited to specific targeting of only a 7-12-base pair target sequence, as suggested in recent studies.[11, 12, 19]

The cellular genome modification data are also consistent with the tradeoff between activity and specificity of sgRNA v1.0 compared to sgRNA v2.1 sgRNAs observed in the *in vitro* selection data and discrete assays. The on-target CLTA4-0-1 site had a modification frequency that was seven-fold lower (11% vs. 76%) in cells expressing Cas9:sgRNA v1.0 compared to cells expressing Cas9:sgRNA v2.1. Although the CLTA4-3-3, and CLTA4-4-8 sites were modified by the Cas9-sgRNA v2.1 complexes, no evidence of modification at any of these three sites was detected in Cas9:sgRNA v1.0-treated cells. The CLTA4-3-1 site, which was modified at 32% of the frequency of on-target CLTA4 site modification in Cas9:v2.1 sgRNA-treated cells, was modified at only 0.5% of the on-target modification frequency in v1.0 sgRNA-treated cells, representing a 62-fold change in selectivity. Taken together, these results demonstrate that guide RNA architecture can have a significant influence on both Cas9 activity and specificity in cells. Our specificity profiling findings present an important caveat to recent and ongoing efforts to improve the overall DNA modification activity of Cas9:guide RNA complexes through guide RNA engineering.[11, 15]

Overall, the off-target DNA cleavage profiling of Cas9 and subsequent analyses show that (*i*) Cas9:guide RNA recognition extends to 18-20 specified target site base pairs and a two-

base pair PAM for the four target sites tested; (*ii*) increasing Cas9:guide RNA concentrations can decrease DNA-cleaving specificity *in vitro*; (*iii*) using more active sgRNA architectures can increase DNA-cleavage activity both *in vitro* and in cells but also increase cleavage of off-target sites both *in vitro* and in cells; and (*iv*) as predicted by our *in vitro* results, Cas9:guide RNA can modify off-target sites in cells with up to four mutations relative to the on-target site. Our findings provide key insights to our understanding of RNA-programmed Cas9 specificity, and reveal a previously unknown role for sgRNA architecture in DNA-cleavage specificity. The principles revealed in this study may also apply to Cas9-based effectors engineered to mediate functions beyond DNA cleavage.

2.3 Methods used to study RNA-guided Cas9 nuclease specificity

Oligonucleotides

All oligonucleotides used in this study were purchased from Integrated DNA Technologies. Oligonucleotide sequences are listed in **Table 2.2**.

oligonucleotide name	oligonucleotide sequence (5'->3')
CLTA1 v2.1 template fwd	TAA TAC GAC TCA CTA TAG GAG TCC TCA TCT CCC TCA AGC GTT TTA GAG CTA TGC TG
CLTA2 v2.1 template fwd	TAA TAC GAC TCA CTA TAG GCT CCC TCA AGC AGG CCC CGC GTT TTA GAG CTA TGC TG
CLTA3 v2.1 template fwd	TAA TAC GAC TCA CTA TAG GTG TGA AGA GCT TCA CTG AGT GTT TTA GAG CTA TGC TG
CLTA4 v2.1 template fwd	TAA TAC GAC TCA CTA TAG GGC AGA TGT AGT GTT TCC ACA GTT TTA GAG CTA TGC TG
v2.1 template rev	GAT AAC GGA CTA GCC TTA TTT TAA CTT GCT ATG CTT TTC AGC ATA GCT CTA AAA C
CLTA1 v1.0 template	CGG ACT AGC CTT ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC GCT TGA GGG AGA TGA GGA CTC CTA TAG TGA GTC GTA TTA
CLTA2 v1.0 template	CGG ACT AGC CTT ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC GCG GGG CCT GCT TGA GGG AGC CTA TAG TGA GTC GTA TTA
CLTA3 v1.0 template	CGG ACT AGC CTT ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC ACT CAG TGA AGC TCT TCA CAC CTA TAG TGA GTC GTA TTA
CLTA4 v1.0 template	CGG ACT AGC CTT ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TGT GGA AAC ACT ACA TCT GCC CTA TAG TGA GTC GTA TTA
T7 promoter oligo	TAA TAC GAC TCA CTA TAG G
CLTA1 lib	/5Phos/AAC ACA NNN NC*C* NG*C* T*T*G* A*G*G* G*A*G* A*T*G* A*G*G* A*C*T* NNN NAC CTG CCG AGA ACA CA
CLTA2 lib	/5Phos/TCT TCT NNN NC*C* NG*C* G*G*G* G*C*C* T*G*C* T*T*G* A*G*G* G*A*G* NNN NAC CTG CCG AGT CTT CT
CLTA3 lib	/5Phos/AGA GAA NNN NC*C* NA*C* T*C*A* G*T*G* A*A*G* C*T*C* T*T*C* A*C*A* NNN NAC CTG CCG AGA GAG AA
CLTA4 lib	/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
CLTA1 site fwd	CTA GCA GTC CTC ATC TCC CTC AAG CAG GC
CLTA1 site rev	AGC TGC CTG CTT GAG GGA GAT GAG GAC TG
CLTA2 site fwd	CTA GTC TCC CTC AAG CAG GCC CCG CTG GT
CLTA2 site rev	AGC TAC CAG CGG GGC CTG CTT GAG GGA GA
CLTA3 site fwd	CTA GCT GTG AAG AGC TTC ACT GAG TAG GA
CLTA3 site rev	AGC TTC CTA CTC AGT GAA GCT CTT CAC AG
CLTA4 site fwd	CTA GTG CAG ATG TAG TGT TTC CAC AGG GT
CLTA4 site rev	AGC TAC CCT GTG GAA ACA CTA CAT CTG CA
test fwd	GCG ACA CGG AAA TGT TGA ATA CTC AT
test rev	GGA GTC AGG CAA CTA TGG ATG AAC G
off-target CLTA4-0 fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GCA GAT GTA GTG TTT CCA CAG GGT
off-target CLTA4-1 fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GAA GAT GTA GTG TTT CCA CAG GGT
off-target CLTA4-2a fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GAA GAT GTA GTG TTT CCA CTG GGT
off-target CLTA4-2b fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GCA GAT GGA GGG TTT CCA CAG GGT
off-target CLTA4-2c fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GCA GAT GTA GTG TTA CCA GAG GGT
off-target CLTA4-3 fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GGG GAT GTA GTG TTT CCA CTG GGT
off-target CLTA4-0 rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CTG TGG AAA CAC TAC ATC TGC
off-target CLTA4-1 rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CTG TGG AAA CAC TAC ATC TTC
off-target CLTA4-2a rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CAG TGG AAA CAC TAC ATC TTC
off-target CLTA4-2b rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CTG TGG AAA CCC TCC ATC TGC
off-target CLTA4-2c rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CTC TGG TAA CAC TAC ATC TGC
off-target CLTA4-3 rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CAG TGG AAA CAC TAC ATC CCC
adapter1(AACA)	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TAA CA
adapter2(AACA)	TGT TAG ATC GGA AGA GCG TCG TGT AGG GAA AGA GTG TAG ATC TCG GTG G
adapter1(TTCA)	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TTT CA
adapter2(TTCA)	TGA AAG ATC GGA AGA GCG TCG TGT AGG GAA AGA GTG TAG ATC TCG GTG G
adapter1	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T
adapter2	AGA TCG GAA GAG CGT CGT GTA GGG AAA GAG TGT AGA TCT CGG TGG

Table 2.2 (Continued). **Oligonucleotides used in this study**.

lib adapter1	GAC GGC ATA CGA GAT
CLTA1 lib adapter2	AAC AAT CTC GTA TGC CGT CTT CTG CTT G
CLTA2 lib adapter2	TCT TAT CTC GTA TGC CGT CTT CTG CTT G
CLTA3 lib adapter2	AGA GAT CTC GTA TGC CGT CTT CTG CTT G
CLTA4 lib adapter2	TTG TAT CTC GTA TGC CGT CTT CTG CTT G
CLTA1 sel PCR	CAA GCA GAA GAC GGC ATA CGA GAT TGT GTT CTC GGC AGG T
CLTA2 sel PCR	CAA GCA GAA GAC GGC ATA CGA GAT AGA AGA CTC GGC AGG T
CLTA3 sel PCR	CAA GCA GAA GAC GGC ATA CGA GAT TTC TCT CTC GGC AGG T
CLTA4 sel PCR	CAA GCA GAA GAC GGC ATA CGA GAT ACA CAA CTC GGC AGG T
PE2 short	AAT GAT ACG GCG ACC ACC GA
CLTA1 lib seq PCR	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 GAA CAC A
CLTA2 lib seq PCR	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 GTC TTC T
CLTA3 lib seq PCR	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 GAG AGA A
CLTA4 lib seq PCR	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 GTT GTG T
lib fwd PCR	CAA GCA GAA GAC GGC ATA CGA GAT
CLTA1-0-1 (Chr. 9) fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CAA GTC TAG CAA GCA GGC CA
CLTA1-0-1 (Chr. 12) fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CAG GCA CTG AGT GGG AAA GT
CLTA1-1-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TAA CCC CAA GTC AGC AAG CA
CLTA1-2-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TTG CTG GTC AAT ACC CTG GC
CLTA1-2-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGA GTA CCC CTG AAA TGG GC
CLTA1-3-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCG CTA CCA ATC AGG GCT TT
CLTA1-3-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCA TTG CCA CTT GTT TGC AT
CLTA1-4-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCT ACC CCC ACA ACT TTG CT
CLTA1-4-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GTG TAC ATC CAG TGC ACC CA
CLTA1-4-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCG GAA AGG ACT TTG AAT ACT TGT
CLTA1-4-4 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CGG CCC AAG ACC TCA TTC AC
CLTA1-4-5 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GTC CTC TCT GGG GCA GAA GT
CLTA1-4-6 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT AGC TGA GTC ATG AGT TGT CTC C
CLTA1-4-7 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CTG CCA GCT TCT CAC ACC AT
CLTA1-4-8 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CTG AAG GAC AAA GGC GGG AA
CLTA1-5-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT AAG GTG CTA AAG GCT CCA CG
CLTA1-5-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GAC CAT TGG TGA GCC CAG AG
CLTA1-5-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TTT TTC GGG CAA CTG CTC AC
CLTA1-5-4 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GCA AGC CTT CTC TCC TCA GA
CLTA1-5-5 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT ACA CAA ACT TCC CTG AGA CCC
CLTA1-6-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGA GTT AGC CCT GCT GTT CA
CLTA4-0-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGA AGA GCT TCA CTG AGT AGG A
CLTA4-3-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCC CCT TAC AGC CAA TTT CGT
CLTA4-3-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGC TGA TGA AAT GCA ATT AAG AGG T
CLTA4-3-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GGT CCC TGC AAG CCA GTA TG
CLTA4-3-4 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT ATC AAA GCC TTG TAT CAC AGT T
CLTA4-3-5 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCC AAA TAA TGC AGG AGC CAA
CLTA4-3-6 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CTG CCT TTA GTG GGA CAG ACT T
CLTA4-3-7 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT AGT AAC CCT AGT AGC CCT CCA
CLTA4-4-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CAT TGC AGT GAG CCG AGA TTG
CLTA4-4-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGG CAA AGT TCA CTT CCA TGT
CLTA4-4-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGC TCT GTG ATG TCT GCC AC
CLTA4-4-4 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGT GTA GGA TTG TGA ACC AGC A
CLTA4-4-5 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCC CAG CCC AGC ATT TTT CT
CLTA4-4-6 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT AGG TTG CTT TGT GCA CAG TC
CLTA4-4-7 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCT GGC TTG GGA TGT TGG AA
CLTA4-4-8 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TTG CCC AAG GTC ATA CTG CT
CLTA4-4-9 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT ACC CAC TAG GTA GCC ATA ATC CA

Table 2.2 (Continued). Oligonucleotides used in this study.

CLTA4-4-10 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CGG TCA TGT CGC TTG GAA GA
CLTA4-4-11 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TTG GCC CAT ATT GCT TTA TGC TG
CLTA4-4-12 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT ATT AGG GGT TGG CTG CAT GA
CLTA4-4-13 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCA AGA CGT GTT GCA TGC TG
CLTA4-4-14 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGG GAG GTG ATA AAT TCC CTA AAT
CLTA4-5-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCA GAG ACA AAG GTG GGG AG
CLTA4-5-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCA TAC AGA AGA GCA AAG TAC CA
CLTA4-5-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CAA AGA GGG GTA TCG GGA GC
CLTA4-5-4 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT AAA TGG AAG AAC CAA GTA GAT GAA
CLTA4-5-5 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TTT TGG TTG ACA GAT GGC CAC A
CLTA4-5-6 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCT TAC TTG TGT GAT TTT AGA ACA A
CLTA4-6-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GAT GGT TCA TGC AGA GGG CT
CLTA4-6-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GCT GGT CTT TCC TGA GCT GT
CLTA4-6-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CTC CAT CAG ATA CCT GTA CCC A
CLTA4-7-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GGG AAA ACA CTC TCT CTC TGC T
CLTA4-7-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GGA GGC CAC GAC ACA CAA TA
CLTA1-0-1 (Chr. 9) rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAC AGG GTG GCT CTT CAG TG
CLTA1-0-1 (Chr. 12) rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGC ACA TGT TTC CAC AGG GT
CLTA1-1-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGT GTT TCC AGG AGC GGT TT
CLTA1-2-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AAG CCT CAG GCA CAA CTC TG
CLTA1-2-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TAG GGG AGG GGC AAA GAC A
CLTA1-3-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGG AAC AGT GGT ATG CTG GT
CLTA1-3-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGT GTG GAC ACT GAC AAG GAA
CLTA1-4-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TCA CTG CCT GGG TGC TTT AG
CLTA1-4-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TAC CCC AGC CTC CAG CTT TA
CLTA1-4-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGA CTA CTG GGG AGC GAT GA
CLTA1-4-4 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGG CTG TTA TGC AGG AAA GGA A
CLTA1-4-5 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GCG GTT GAG GTG GAT GGA AG
CLTA1-4-6 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGC AGC ATC CCT TAC ATC CT
CLTA1-4-7 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGA AAA AGC TTC CCC AGA AAG GA
CLTA1-4-8 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CTG CAC CAA CCT CTA CGT CC
CLTA1-5-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CTG GAG AGG GCA TAG TTG GC
CLTA1-5-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGG AAG GCT CTT TGT GGG TT
CLTA1-5-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TTC CTA GCG GGA ACT GGA AA
CLTA1-5-4 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGG CTA ATG GGG TAG GGG AT
CLTA1-5-5 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGT CCA TGT TGG CTG AGG TG
CLTA1-6-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAG GCC AAC CTT GAC AAC TT
CLTA4-0-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGC AGG CCA AAG ATG TCT CC
CLTA4-3-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TCT GCT CTT GAG GTT ATT TGT CC
CLTA4-3-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGG ACC AAT TTG CTA CTC ATG G
CLTA4-3-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGG AGG CTG TAA ACG TCC TG
CLTA4-3-4 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGC TAT GAT TTG CTG AAT TAC TCC T
CLTA4-3-5 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GCA ATT TTG CAG ACC ACC ATC
CLTA4-3-6 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGC AGC TTG CAA CCT TCT TG
CLTA4-3-7 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TCA TGA GAG TTT CCC CAA CA
CLTA4-4-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT ACT TGA GGG GGA AAA AGT TTC TTA
CLTA4-4-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGG TCC CTG TCT GTC ATT GG
CLTA4-4-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AAG CGA GTG ACT GTC TGG GA
CLTA4-4-4 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAT GGG TGG GAC ACG TAG TT
CLTA4-4-5 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGC TTT CCT GGA CAC CCT ATC
CLTA4-4-6 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGA GCG AGG GAG CGA TGT A
CLTA4-4-7 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TTG TGG ACC ACT GCT TAG TGC
CLTA4-4-8 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAA CTA CCC TGA GGC CAC C
CLTA4-4-9 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGT CAG CAC TCC TCA GCT TT
CLTA4-4-10 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGG AGG ATG CAT GCC ACA TT

Table 2.2 (Continued). **Oligonucleotides used in this study**.

CLTA4-4-11 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CCC AGC CTC TTT GAC CCT TC
CLTA4-4-12 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CCC ACA CCA GGC TGT AAG G
CLTA4-4-13 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TAG ATA TAT GGG TGT GTC TGT ACG
CLTA4-4-14 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TTC CAA AGT GGC TGA ACC AT
CLTA4-5-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CCC ACA GGG CTG ATG TTT CA
CLTA4-5-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TTG TAA TGC AAC CTC TGT CAT GC
CLTA4-5-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CCA GCT CCA GCA ATC CAT GA
CLTA4-5-4 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TTT GGG AAA GAT AGC CCT GGA
CLTA4-5-5 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAA TGA AAC AGC GGG GAG GT
CLTA4-5-6 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT ACA ATC ACG TGT CCT TCA CT
CLTA4-6-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAG ATC CCT CCT GGG CAA TG
CLTA4-6-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GTC AGG AGG CAA GGA GGA AC
CLTA4-6-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT ACT TCC TTC CTT TTG AGA CCA AGT
CLTA4-7-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GCG GCA GAT TCC TGG TGA TT
CLTA4-7-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGT CAC CAT CAG CAC AGT CA
PE1-barcode1	CAA GCA GAA GAC GGC ATA CGA GAT ATA TCA GTG TGA CTG GAG TTC AGA CGT GTG CT
PE1-barcode2	CAA GCA GAA GAC GGC ATA CGA GAT TTT CAC CGG TGA CTG GAG TTC AGA CGT GTG CT
PE1-barcode3	CAA GCA GAA GAC GGC ATA CGA GAT CCA CTC ATG TGA CTG GAG TTC AGA CGT GTG CT
PE1-barcode4	CAA GCA GAA GAC GGC ATA CGA GAT TAC GTA CGG TGA CTG GAG TTC AGA CGT GTG CT
PE1-barcode5	CAA GCA GAA GAC GGC ATA CGA GAT CGA AAC TCG TGA CTG GAG TTC AGA CGT GTG CT
PE1-barcode6	CAA GCA GAA GAC GGC ATA CGA GAT ATC AGT ATG TGA CTG GAG TTC AGA CGT GTG CT
PE2-barcode1	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA TTA CTC GAC ACT CTT TCC CTA CAC GAC
PE2-barcode2	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CCG GAG AAC ACT CTT TCC CTA CAC GAC
PE2-barcode3	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC GCT CAT TAC ACT CTT TCC CTA CAC GAC

Table 2.2 (Continued). **Oligonucleotides used in this study**. All oligonucleotides were purchased from Integrated DNA Technologies. An asterisk (*) indicates that the preceding nucleotide was incorporated as a hand mix of phosphoramidites consisting of 79 mol% of the phosphoramidite corresponding to the preceding nucleotide and 4 mol% of each of the other three canonical phosphoramidites. "/5Phos/" denotes a 5' phosphate group installed during synthesis.

Expression and Purification of *S. pyogenes* Cas9

E. coli Rosetta (DE3) cells were transformed with plasmid pMJ806[11], encoding the *S. pyogenes cas9* gene fused to an N-terminal 6xHis-tag/maltose binding protein. The resulting expression strain was inoculated in Luria-Bertani (LB) broth containing 100 μg/mL of ampicillin and 30 μg/mL of chloramphenicol at 37 °C overnight. The cells were diluted 1:100 into the same growth medium and grown at 37 °C to OD_{600} ~0.6. The culture was incubated at 18 °C for 30 min, and isopropyl β-D-1-thiogalactopyranoside (IPTG) was added at 0.2 mM to induce Cas9 expression. After ~17 h, the cells were collected by centrifugation at 8,000 *g* and resuspended in lysis buffer (20 mM tris(hydroxymethyl)-aminomethane (Tris)-HCl, pH 8.0, 1 M KCl, 20 % glycerol, 1 mM tris (2-carboxyethyl)phosphine (TCEP)). The cells were lysed by sonication (10 sec pulse-on and 30 sec pulse-off for 10 min total at 6 W output) and the soluble lysate was obtained by centrifugation at 20,000 *g* for 30 min. The cell lysate was incubated with nickel-

nitriloacetic acid (nickel-NTA) resin (Qiagen) at 4 °C for 20 min to capture His-tagged Cas9. The resin was transferred to a 20-mL column and washed with 20 column volumes of lysis buffer. Cas9 was eluted in 20 mM Tris-HCl (pH 8), 0.1 M KCl, 20 % glycerol, 1 mM TCEP, and 250 mM imidazole, and concentrated by Amicon ultra centrifugal filter (Millipore, 30-kDa molecular weight cut-off) to ~50 mg/mL. The 6xHis tag and maltose-binding protein were removed by TEV protease treatment at 4 °C for 20 h and captured by a second Ni-affinity purification step. The eluent, containing Cas9, was injected into a HiTrap SP FF column (GE Healthcare) in purification buffer containing 20 mM Tris-HCl (pH 8), 0.1 M KCl, 20 % glycerol, and 1 mM TCEP. Cas9 was eluted with purification buffer containing a linear KCl gradient from 0.1 M to 1 M over five column volumes. The eluted Cas9 was further purified by a HiLoad Superdex 200 column in purifiation buffer, snap-frozen in liquid nitrogen, and stored in aliquots at -80 °C.

In Vitro RNA Transcription

100 pmol CLTA(#) v2.1 fwd and v2.1 template rev were incubated at 95 °C and cooled at 0.1 °C/s to 37 °C in NEBuffer2 (50 mM sodium chloride, 10 mM Tris-HCl, 10 mM magnesium chloride, 1 mM dithiothreitol, pH 7.9) supplemented with 10 μM dNTP mix (Bio-Rad). 10 U of Klenow Fragment (3'→5' exo$^-$) (NEB) were added to the reaction mixture and a double-stranded CLTA(#) v2.1 template was obtained by overlap extension for 1 h at 37 °C. 200 nM CLTA(#) v2.1 template alone or 100 nM CLTA(#) template with 100 nM T7 promoter oligo was incubated overnight at 37 °C with 0.16 U/μL of T7 RNA Polymerase (NEB) in NEB RNAPol Buffer (40 mM Tris-HCl, pH 7.9, 6 mM magnesium chloride, 10 mM dithiothreitol, 2 mM spermidine) supplemented with 1 mM rNTP mix (1 mM rATP, 1 mM rCTP, 1 mM rGTP, 1 mM rUTP). *In vitro* transcribed RNA was precipitated with ethanol and purified by gel electrophoresis on a Criterion 10% polyacrylamide TBE-Urea gel (Bio-Rad). Gel-purified sgRNA was precipitated with ethanol and redissolved in water.

In Vitro Library Construction

10 pmol of CLTA(#) lib oligonucleotides were separately circularized by incubation with 100 units of CircLigase II ssDNA Ligase (Epicentre) in 1x CircLigase II Reaction Buffer (33 mM Tris-acetate, 66 mM potassium acetate, 0.5 mM dithiothreitol, pH 7.5) supplemented with

2.5 mM manganese chloride in a total reaction volume of 20 µL for 16 hours at 60 °C. The reaction mixture was incubated for 10 minutes at 85 °C to inactivate the enzyme. 5 µL (5 pmol) of the crude circular single-stranded DNA were converted into the concatemeric pre-selection libraries with the illustra TempliPhi Amplification Kit (GE Healthcare) according to the manufacturer's protocol. Concatemeric pre-selection libraries were quantified with the Quant-it PicoGreen dsDNA Assay Kit (Invitrogen).

In Vitro Cleavage of On-Target and Off-Target Substrates

Plasmid templates for PCR were constructed by ligation of annealed oligonucleotides CLTA(#) site fwd/rev into *Hin*dIII/*Xba*I double-digested pUC19 (NEB). On-target substrate DNAs were generated by PCR with the plasmid templates and test fwd and test rev primers, then purified with the QIAquick PCR Purification Kit (Qiagen). Off-target substrate DNAs were generated by primer extension. 100 pmol off-target (#) fwd and off-target (#) rev primers were incubated at 95 °C and cooled at 0.1 °C/s to 37 °C in NEBuffer2 (50 mM sodium chloride, 10 mM Tris-HCl, 10 mM magnesium chloride, 1 mM dithiothreitol, pH 7.9) supplemented with 10 µM dNTP mix (Bio-Rad). 10 U of Klenow Fragment (3'→5' exo-) (NEB) were added to the reaction mixture and double-stranded off-target templates were obtained by overlap extension for 1 h at 37 °C followed by enzyme inactivation for 20 min at 75 °C, then purified with the QIAquick PCR Purification Kit (Qiagen). 200 nM substrate DNAs were incubated with 100 nM Cas9 and 100 nM (v1.0 or v2.1) sgRNA or 1000 nM Cas9 and 1000 nM (v1.0 or v2.1) sgRNA in Cas9 cleavage buffer (200 mM HEPES, pH 7.5, 1.5 M potassium chloride, 100 mM magnesium chloride, 1 mM EDTA, 5 mM dithiothreitol) for 10 min at 37 °C. On-target cleavage reactions were purified with the QIAquick PCR Purification Kit (Qiagen), and off-target cleavage reactions were purified with the QIAquick Nucleotide Removal Kit (Qiagen) before electrophoresis in a Criterion 5% polyacrylamide TBE gel (Bio-Rad).

In Vitro Selection

200 nM concatemeric pre-selection libraries were incubated with 100 nM Cas9 and 100 nM sgRNA or 1000 nM Cas9 and 1000 nM sgRNA in Cas9 cleavage buffer (200 mM HEPES, pH 7.5, 1.5 M potassium chloride, 100 mM magnesium chloride, 1 mM EDTA, 5 mM dithiothreitol) for 10 min at 37 °C. Pre-selection libraries were also separately 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. Blunt-ended post-selection library members or sticky-ended pre-selection library members were purified with the QIAQuick PCR Purification Kit (Qiagen) and ligated to 10 pmol adapter1/2(AACA) (Cas9:v2.1 sgRNA, 100 nM), adapter1/2(TTCA) (Cas9:v2.1 sgRNA, 1000 nM), adapter1/2 (Cas9:v2.1 sgRNA, 1000 nM), or lib adapter1/CLTA(#) lib adapter 2 (pre-selection) with 1,000 U of 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 (> 10 h) at room temperature. Adapter-ligated DNA was purified with the QIAquick PCR Purification Kit and PCR-amplified for 10-13 cycles with Phusion Hot Start Flex DNA Polymerase (NEB) in Buffer HF (NEB) and primers CLTA(#) sel PCR/PE2 short (post-selection) or CLTA(#) lib seq PCR/lib fwd PCR (pre-selection). Amplified DNAs were gel purified, quantified with the KAPA Library Quantification Kit-Illumina (KAPA Biosystems), and subjected to single-read sequencing on an Illumina MiSeq or Rapid Run single-read sequencing on an Illumina HiSeq 2500 (Harvard University FAS Center for Systems Biology Core facility, Cambridge, MA).

Selection Analysis

Pre-selection and post-selection sequencing data were analyzed as previously described[21], with modification using scripts written in C++. Specificity scores were calculated with the formulae: positive specificity score = (frequency of base pair at position[post-selection] - frequency of base pair at position[pre-selection]) / (1 - frequency of base pair at position[pre-selection]) and negative specificity score = (frequency of base pair at position[post-selection] - frequency of base pair at position[pre-selection]) / (frequency of base pair at position[pre-selection]). Normalization for sequence logos was performed as previously described[22].

Cellular Cleavage Assays

HEK293T cells were split at a density of 0.8×10^5 per well (6-well plate) before transcription and maintained in Dulbecco's modified eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS) in a 37°C humidified incubator with 5% CO_2. After 1 day, cells were transiently transfected using Lipofectamine 2000 (Invitrogen) following the manufacturer's protocols. HEK293T cells were transfected at 70% confluency in each well of 6-well plate with

1.0 μg of the Cas9 expression plasmid (Cas9-HA-2xNLS-GFP-NLS) and 2.5 μg of the single-strand RNA expression plasmid pSiliencer-CLTA (version 1.0 or 2.1). The transfection efficiencies were estimated to be ~70%, based on the fraction of GFP-positive cells observed by fluorescence microscopy. 48 h after transfection, cells were washed with phosphate buffered saline (PBS), pelleted and frozen at -80 °C. Genomic DNA was isolated from 200 μL cell lysate using the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's protocol.

Off-Target Site Sequence Determination

100 ng genomic DNA isolated from cells treated with Cas9 expression plasmid and single-strand RNA expression plasmid (treated cells) or Cas9 expression plasmid alone (control cells) were amplified by PCR with 10 s 72°C extension for 35 cycles with primers CLTA(#)-(#)-(#) fwd and CLTA(#)-(#)-(#) rev and Phusion Hot Start Flex DNA Polymerase (NEB) in Buffer GC (NEB), supplemented with 3% DMSO. Relative amounts of crude PCR products were quantified by gel, and Cas9-treated (control) and Cas9:sgRNA-treated PCRs were separately pooled in equimolar concentrations before purification with the QIAquick PCR Purification Kit (Qiagen). Purified DNA was amplified by PCR with primers PE1-barcode# and PE2-barcode# for 7 cycles with Phusion Hot Start Flex DNA Polymerase (NEB) in Buffer HF (NEB). Amplified control and treated DNA pools were purified with the QIAquick PCR Purification Kit (Qiagen), followed by purification with Agencourt AMPure XP (Beckman Coulter). Purified control and treated DNAs were quantified with the KAPA Library Quantification Kit-Illumina (KAPA Biosystems), pooled in a 1:1 ratio, and subjected to paired-end sequencing on an Illumina MiSeq.

Statistical Analysis

Statistical analysis was performed as previously described[21]. *P*-values from cellular off-target modification assays were calculated for a one-sided Fisher exact test.

2.4 References cited in RNA-guided Cas9 nuclease specificity study

1. Hockemeyer, D. et al. Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology* **29**, 731-734 (2011).
2. Zou, J. et al. Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells. *Cell stem cell* **5**, 97-110 (2009).
3. Hockemeyer, D. et al. Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nature biotechnology* **27**, 851-857 (2009).
4. Doyon, Y. et al. Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nature biotechnology* **26**, 702-708 (2008).
5. Meng, X., Noyes, M.B., Zhu, L.J., Lawson, N.D. & Wolfe, S.A. Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nature biotechnology* **26**, 695-701 (2008).
6. Sander, J.D. et al. Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nature biotechnology* **29**, 697-698 (2011).
7. Tesson, L. et al. Knockout rats generated by embryo microinjection of TALENs. *Nature biotechnology* **29**, 695-696 (2011).
8. Cui, X. et al. Targeted integration in rat and mouse embryos with zinc-finger nucleases. *Nature biotechnology* **29**, 64-67 (2011).
9. Perez, E.E. et al. Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature biotechnology* **26**, 808-816 (2008).
10. NCT00842634, NCT01044654, NCT01252641, NCT01082926.
11. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).
12. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823 (2013).
13. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826 (2013).
14. Hwang, W.Y. et al. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature biotechnology* **31**, 227-229 (2013).
15. Jinek, M. et al. RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).
16. Dicarlo, J.E. et al. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic acids research* (2013).
17. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L.A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology* **31**, 233-239 (2013).
18. Sapranauskas, R. et al. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic acids research* **39**, 9275-9282 (2011).
19. Semenova, E. et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10098-10103 (2011).
20. Qi, L.S. et al. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152**, 1173-1183 (2013).

21. Pattanayak, V., Ramirez, C.L., Joung, J.K. & Liu, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature methods* **8**, 765-770 (2011).
22. Doyon, J.B., Pattanayak, V., Meyer, C.B. & Liu, D.R. Directed evolution and substrate specificity profile of homing endonuclease I-SceI. *Journal of the American Chemical Society* **128**, 2477-2484 (2006).

Chapter 3:
Fusion of Inactivated Cas9 to *Fok*I Nuclease Improves Genome Modification Specificity

John P. Guilinger, David B. Thompson, and David R. Liu

David B. Thompson and I, in the laboratory of David R. Liu, designed, preformed and analyzed all the experiments together.

3.1 Introduction to specificity of RNA-guided Cas9 nucleases and nickases

The recent development of robust, predictable, and user-friendly methods for the generation of sequence-specific DNA-binding proteins has led to a rapid expansion of the field of genome editing. Today, user-defined site-specific genome modification has become a powerful tool in biological research,[1,2] and holds significant potential to serve as the basis of a new generation of human therapeutics.[3] One such programmable endonuclease system uses the CRISPR-derived Cas9 nuclease complexed with a guide RNA (gRNA) to target dsDNA sequences for cleavage.[4] The gRNA programs Cas9 DNA cleavage specificity through ~17 to 20 bp of complementarity between the gRNA and the target DNA sequence when complexed with Cas9.[5] Provided that the target sequence is adjacent to a short 3' motif, the protospacer adjacent motif (PAM) required for initial binding and Cas9 activation[4] (**Figure 3.1**), any locus can in principle be targeted.

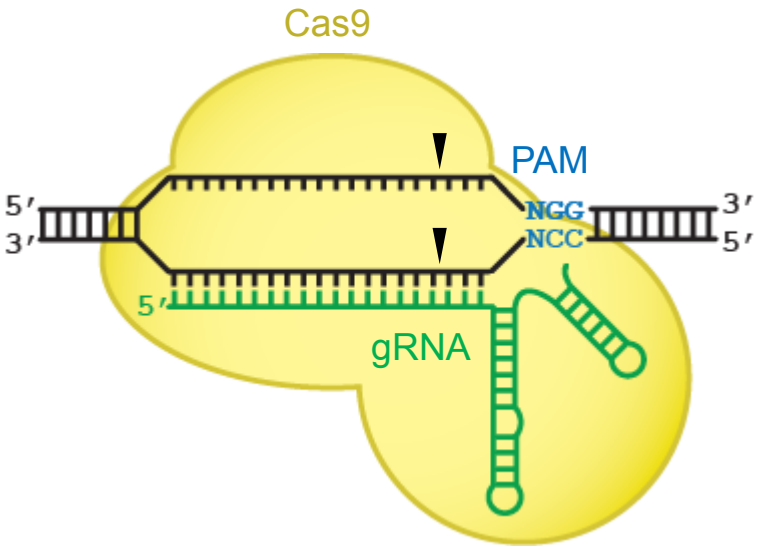


Figure 3.1. Architecture of Cas9. (a) Cas9 protein (yellow) binds to target DNA in complex with a guide RNA (gRNA, green). The *S. pyogenes* Cas9 protein recognizes the PAM sequence NGG (blue), initiating unwinding of dsDNA and gRNA:DNA base pairing. Black triangles indicate the cleavage points three bases from the PAM on both top and bottom strands.

In cells, Cas9:gRNA-induced double strand breaks can result in functional gene knockout through non-homologous end joining (NHEJ) or alteration of a target locus to virtually any sequence through homology-directed repair (HDR) with an exogenous DNA template.[6–8] Cas9 is an especially convenient genome editing platform,[9] as a genome editing agent for each new

target site of interest can be accessed by simply generating the corresponding gRNA. This approach has been widely used to create targeted knockouts and gene insertions in cells and model organisms, and has also been recognized for its potential therapeutic relevance.

While Cas9:gRNA systems provide an unprecedented level of programmability and ease of use, our group[10] and others[11–14] have reported the ability of Cas9 to cleave off-target genomic sites, resulting in modification of unintended loci that can limit the usefulness and safety of Cas9 as a research tool and as a potential therapeutic. We hypothesized that engineering Cas9 variants to cleave DNA only when two simultaneous, adjacent Cas9:DNA binding events take place could substantially improve genome editing specificity since the likelihood of two adjacent off-target binding events is much smaller than the likelihood of a single off-target binding event (approximately $1/n^2$ vs. $1/n$). Such an approach is distinct from the recent development of mutant Cas9 proteins that cleave only a single strand of dsDNA ("nickases"). Nickases can be used to nick opposite strands of two nearby target sites, generating what is effectively a double strand break, and paired Cas9 nickases can effect substantial on-target DNA modification with reduced off-target modification.[5,14,15] Because each of the component Cas9 nickases remains catalytically active[4,8,16] and single-stranded DNA cleavage events are weakly mutagenic,[17,18] nickases can induce genomic modification even when acting as monomers.[7,14,19] Indeed, Cas9 nickases have been previously reported to induce off-target modifications in cells.[5,15] Moreover, since paired Cas9 nickases can efficiently induce dsDNA cleavage-derived modification events when bound up to ~100 bp apart,15 the statistical number of potential off-target sites for paired nickases is larger than that of a more spatially constrained dimeric Cas9 cleavage system.

3.2 Screening and optimizing *Fok*I-dCas9 architectures for genome modification

To further improve the specificity of the Cas9:gRNA system, we sought to engineer an obligate dimeric Cas9 system analogous to previously developed dimeric zinc-finger nucleases (ZFNs) and TALENs. These nucleases have been widely used as research tools in cell culture and *in vivo*[20], and ZFNs are currently in clinical trials as potential human therapeutics.[3] Based on ZFN and TALEN examples, we speculated that fusing the *Fok*I restriction endonuclease cleavage domain to a catalytically dead Cas9 (dCas9) could create an obligate dimeric Cas9 that

would cleave DNA only when two distinct *Fok*I-dCas9:gRNA complexes bind to adjacent sites ("half-sites") with particular spacing constraints (**Figure 3.2**).

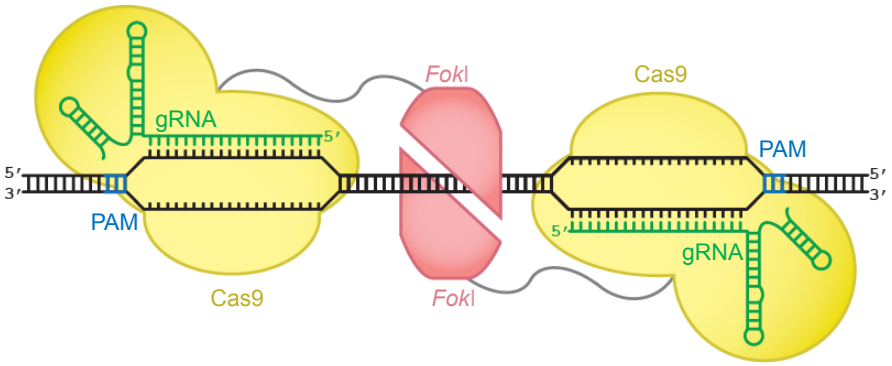


Figure 3.2. Architectures of *Fok*I-dCas9 fusion. Monomers of *Fok*I nuclease (red) fused to dCas9 bind to separate sites within the target locus. Only adjacently bound *Fok*I-dCas9 monomers can assemble a catalytically active *Fok*I nuclease dimer, triggering dsDNA cleavage

In contrast with Cas9 nickases, in which single-stranded DNA cleavage by monomers takes place independently, the DNA cleavage of *Fok*I-dCas9 requires simultaneous binding of two distinct *Fok*I-dCas9 monomers because monomeric *Fok*I nuclease domains are not catalytically competent.[21] In principle this approach should increase the specificity of DNA cleavage relative to wild-type Cas9 by doubling the number of specified target bases contributed by both monomers of the *Fok*I-dCas9 dimer, and should also offer improved specificity compared to nickases due to inactivity of monomeric *Fok*I-dCas9:gRNA complexes, and the more stringent spatial requirements for assembly of a *Fok*I-dCas9 dimer.

While fusions of dCas9 to short functional peptide tags have been described to enable gRNA-programmed transcriptional regulation,[22] no fusions of Cas9 with active enzyme domains have been previously reported to our knowledge. Therefore we began by constructing and characterizing a wide variety of *Fok*I-dCas9 fusion proteins with distinct configurations of a *Fok*I nuclease domain, dCas9 containing inactivating mutations D10A and H840A, and a nuclear localization sequence (NLS). We fused *Fok*I to either the N- or C-terminus of dCas9, and varied the location of the NLS to be at either terminus or between the two domains (**Figure 3.3).**

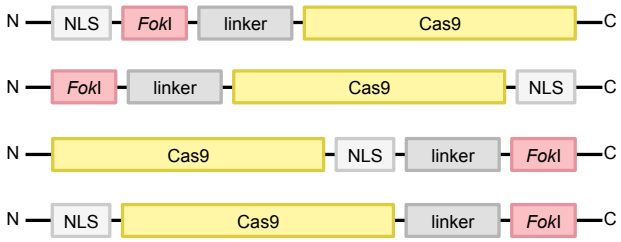


Figure 3.3. Architectures of *Fok*I-dCas9 fusion variants. *Fok*I-dCas9 fusion architectures tested. Four distinct configurations of NLS, *Fok*I nuclease, and dCas9 were assembled. 17 protein linker variants were also tested (see main text).

We further varied the length of the linker sequence as either one or three repeats of Gly-Gly-Ser (GGS) between the *Fok*I and dCas9 domains. Since previously developed dimeric nuclease systems are sensitive to the length of the spacer sequence between half-sites,[23,24] we also tested a wide range of spacer sequence lengths between two gRNA binding sites within a test target gene, Emerald GFP (referred to hereafter as GFP). Two sets of gRNA binding-site pairs with different orientations were chosen within GFP (**Figure 3.4)**.

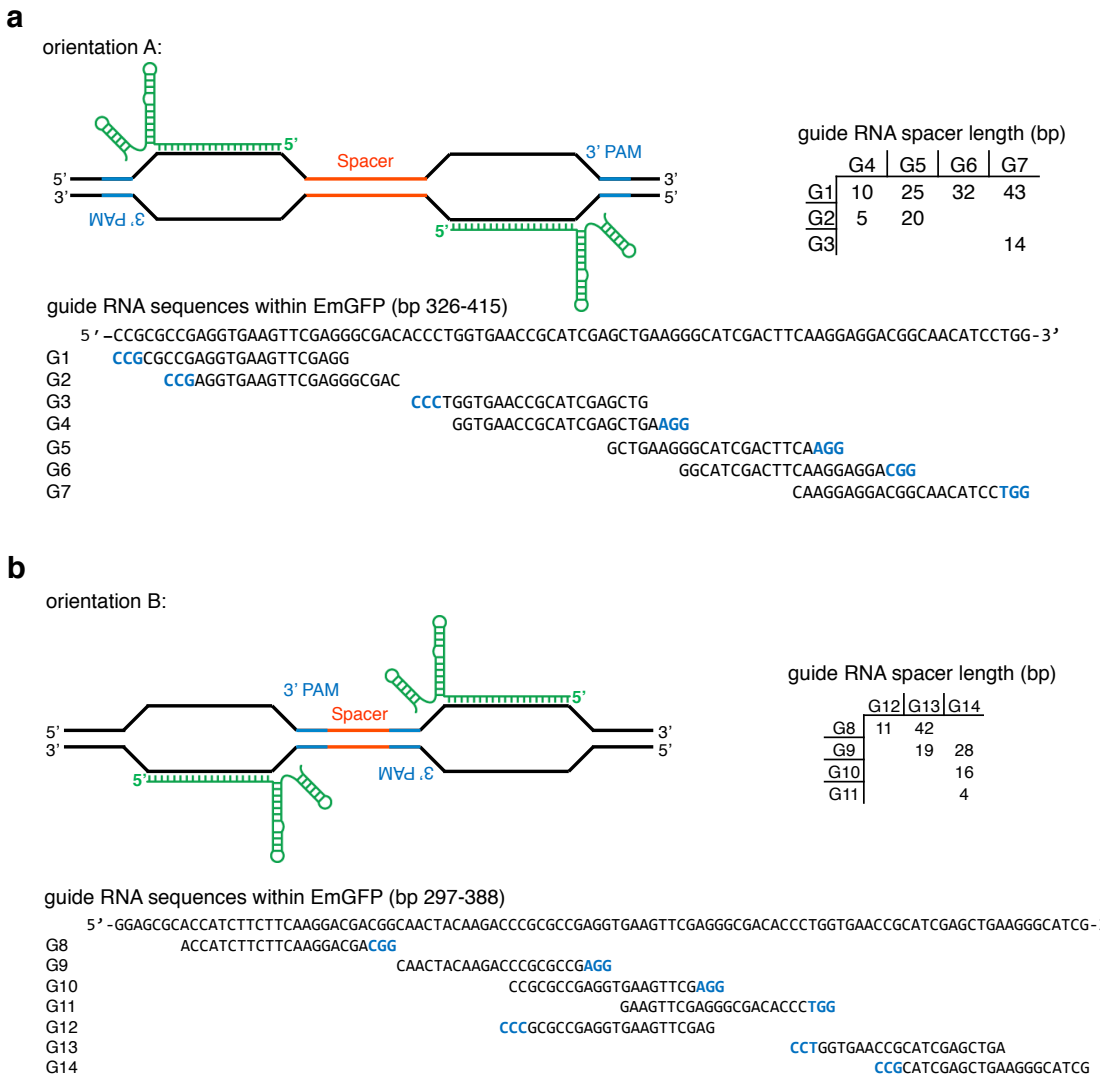


Figure 3.4. gRNA pairs targeting DNA sequences in a genomic GFP gene. gRNA target sites tested within GFP. Seven gRNA target sites were chosen to test *Fok*I-dCas9 activity in an orientation in which the PAM is distal from the cleaved spacer sequence (orientation A). Together, these seven gRNAs enabled testing of *Fok*I-dCas9 fusion variants across seven spacer lengths ranging from 5 to 43 bp. (b) Seven gRNA target sites were chosen to test *Fok*I-dCas9 candidate activity in an orientation in which the PAM is adjacent from the cleaved spacer sequence (orientation B). Together, these seven gRNAs enabled testing of *Fok*I-dCas9 fusion variants across six spacer lengths ranging from 4 to 42 bp.

One set placed the pair of NGG PAM sequences distal from the spacer sequence, with the 5' end of the gRNA adjacent to the spacer (orientation A) (**Figure 3.4**), while the other placed the PAM sequences immediately adjacent to the spacer (orientation B) (**Figure 3.4**). By pairwise combination of the gRNA targets, we tested seven and six spacer lengths in both dimer orientations, ranging from 5 to 43 bp in orientation A, and 4 to 42 bp in orientation B. In total,

DNA constructs corresponding to 104 pairs of *Fok*I-dCas9:gRNA complexes were generated and tested, exploring four fusion architectures, 17 protein linker variants (described below), both gRNA orientations and 13 spacer lengths between half-sites.

To assay the activities of these candidate *Fok*I-dCas9:gRNA pairs, we used a previously described flow cytometry-based fluorescence assay[5,12] in which DNA cleavage and NHEJ of a stably integrated constitutively expressed GFP gene in HEK293 cells leads to loss of cellular fluorescence (**Figure 3.5**).

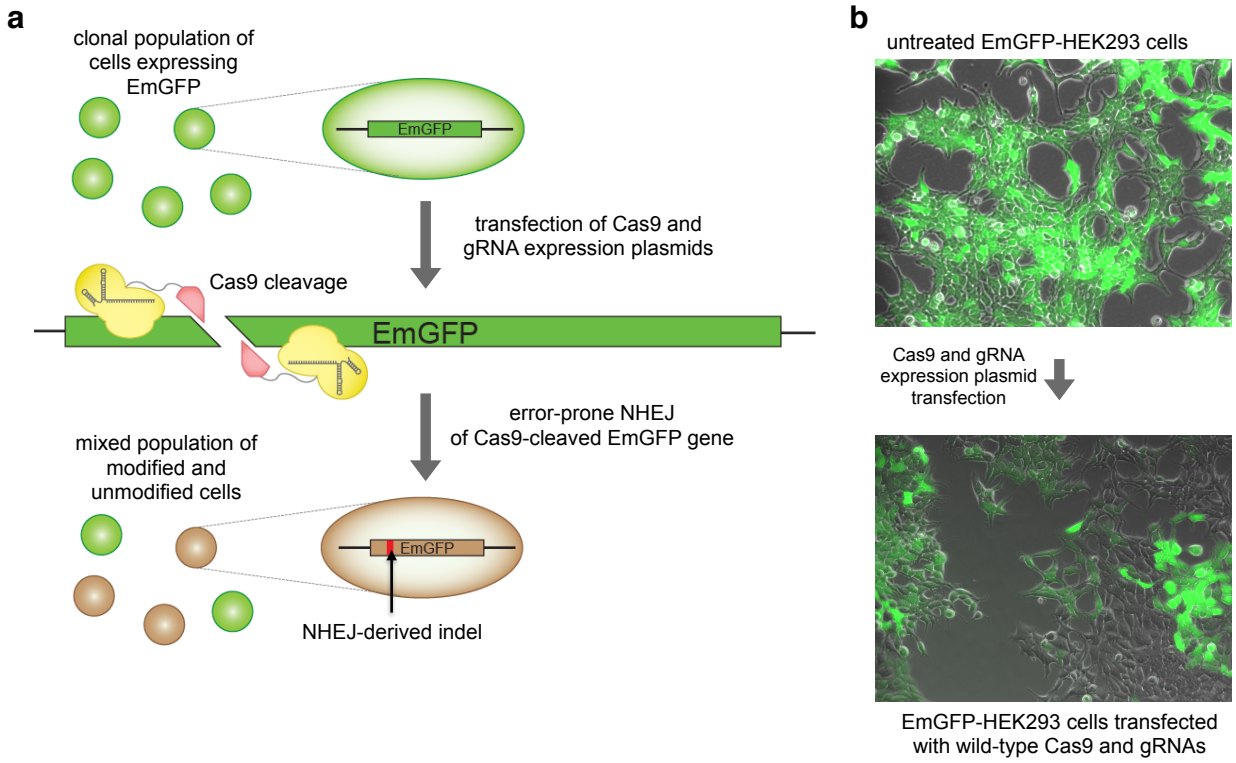


Figure 3.5. GFP disruption assay for measuring genomic DNA-modification activity. (**a**) A HEK293-derived cell line constitutively expressing a genomically integrated EmGFP gene was used to test the activity of candidate *Fok*I-dCas9 fusion constructs. Co-transfection of these cells with appropriate nuclease and gRNA expression plasmids leads to dsDNA cleavage within the EmGFP coding sequence, stimulating error-prone NHEJ and generating indels that can disrupt the expression of GFP, leading to loss of cellular fluorescence. The fraction of cells displaying a loss of GFP fluorescence is then quantitated by flow cytometry. (**b**) Typical epifluorescence microscopy images at 200x magnification of EmGFP-HEK293 cells before and after co-transfection with wild-type Cas9 and gRNA expression plasmids.

For comparison, we assayed the initial set of *Fok*I-dCas9 variants side-by-side with the corresponding Cas9 nickases and wild-type Cas9 in the same expression plasmid across both gRNA spacer orientation sets A and B. Cas9 protein variants and gRNA were generated in cells

by transient co-transfection of the corresponding Cas9 protein expression plasmids together with the appropriate pair of gRNA expression plasmids. The *Fok*I-dCas9 variants, nickases, and wild-type Cas9 all targeted identical DNA sites using identical gRNAs.

Most of the initial *Fok*I-dCas9 fusion variants were inactive or very weakly active. The NLS-*Fok*I-dCas9 architecture (listed from N to C terminus), however, resulted in a 10% increase of GFP-negative cells above corresponding the no-gRNA control when used in orientation A, with PAMs distal from the spacer (**Figure 3.6**).

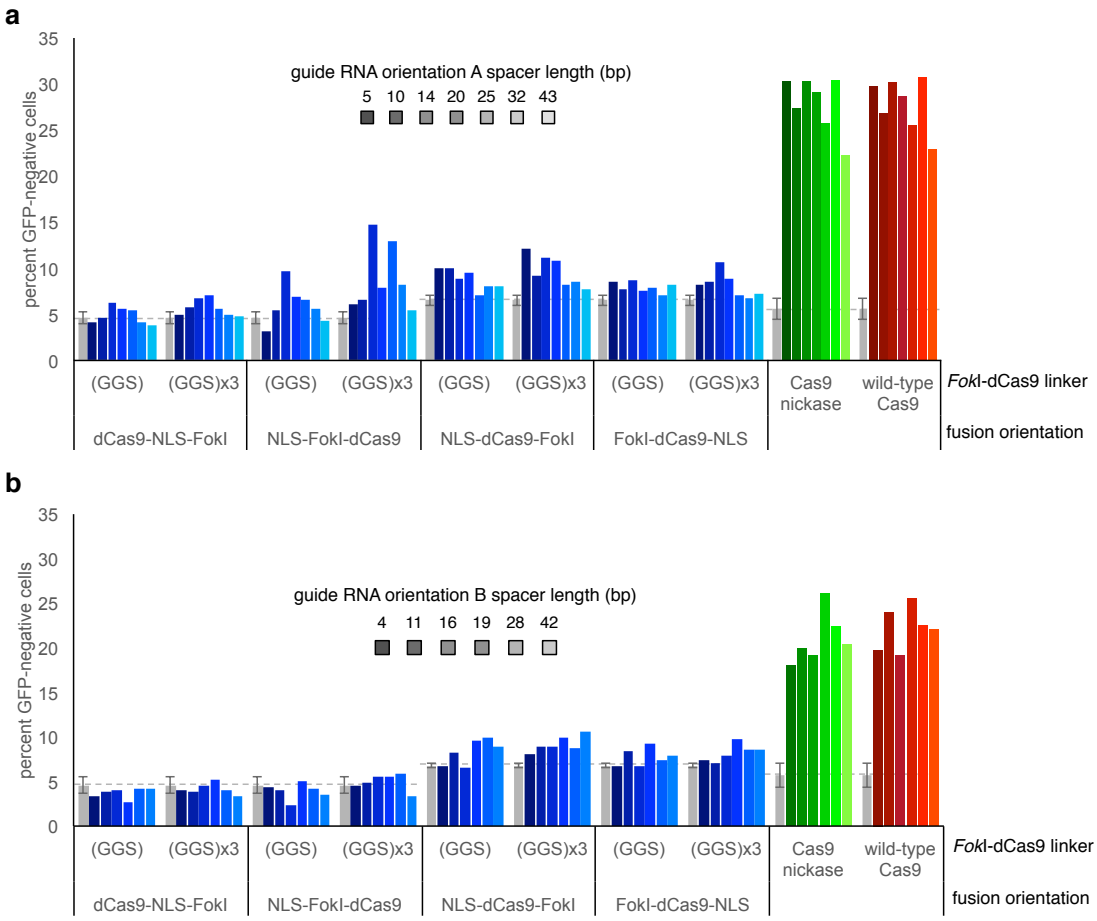


Figure 3.6. Activities of *Fok*I-dCas9 fusion candidates combined with gRNA pairs of different orientations and varying spacer lengths. Activity of *Fok*I-dCas9 fusion orientations in GFP disruption assay. The fusion architectures described in **Figure 3.3** were tested for functionality by flow cytometry using the GFP loss-of-function reporter across all (**a**) orientation A gRNA spacers and (**b**) orientation B gRNA spacers. All *Fok*I-dCas9 fusion data shown are the results of single trials. Wild-type Cas9 and Cas9 nickase data are the average of two replicates, while the 'no treatment' negative control data is the average of 6 replicates, with error bars representing one standard deviation. The gray dotted line across the Y-axis corresponds to the average of the 'no treatment' controls performed on the same day

.

In contrast, NLS-*Fok*I-dCas9 activity was undetectable when used on gRNA pairs with PAMs adjacent to the spacer (**Figure 3.6**). Examination of the recently reported Cas9 structures[25,26] reveals that the Cas9 N-terminus protrudes from the RuvCI domain, which contacts the 5' end of the gRNA:DNA duplex. We speculate that this arrangement places an N-terminally fused *Fok*I distal from the PAM, resulting in a preference for gRNA pairs with PAMs distal from the cleaved spacer. While other *Fok*I-dCas9 fusion pairings and the other gRNA orientation in some cases showed modest activity (**Figure 3.6**), we chose NLS-*Fok*I-dCas9 with gRNAs in orientation A for further development.

Next we optimized the protein linkers between the NLS and *Fok*I domain, and between the *Fok*I domain and dCas9 in the NLS-*Fok*I-dCas9 architecture. We tested 17 linkers with a wide range of amino acid compositions, predicted flexibilities, and lengths varying from 9 to 21 residues (**Figure 3.7**).

a

name	NLS-**linker**-Fok1	Fok1-**linker**-dCas9
*Fok*I-(GGS)x3	GGS	GGSGGSGGS
*Fok*I-(GGS)x6	GGS	GGSGGSGGSGGSGGSGGS
*Fok*I-L0	GGS	-
*Fok*I-L1	GGS	MKIIEQLPSA
*Fok*I-L2	GGS	VRHKLKRVGS
*Fok*I-L3	GGS	VPFLLEPDNINGKTC
*Fok*I-L4	GGS	GHGTGSTGSGSS
*Fok*I-L5	GGS	MSRPDPA
*Fok*I-L6	GGS	GSAGSAAGSGEF
*Fok*I-L7	GGS	SGSETPGTSESA
*Fok*I-L8	GGS	SGSETPGTSESATPES
*Fok*I-L9	GGS	SGSETPGTSESATPEGGSGGS
NLS-(GGS)	GGS	GGSM
NLS-(GGS)x3	GGSGGSGGS	GGSM
NLS-L1	VPFLLEPDNINGKTC	GGSM
NLS-L2	GSAGSAAGSGEF	GGSM
NLS-L3	SIVAQLSRPDPA	GGSM
wild-type Cas9	N/A	N/A
Cas9 nickase	N/A	N/A

b

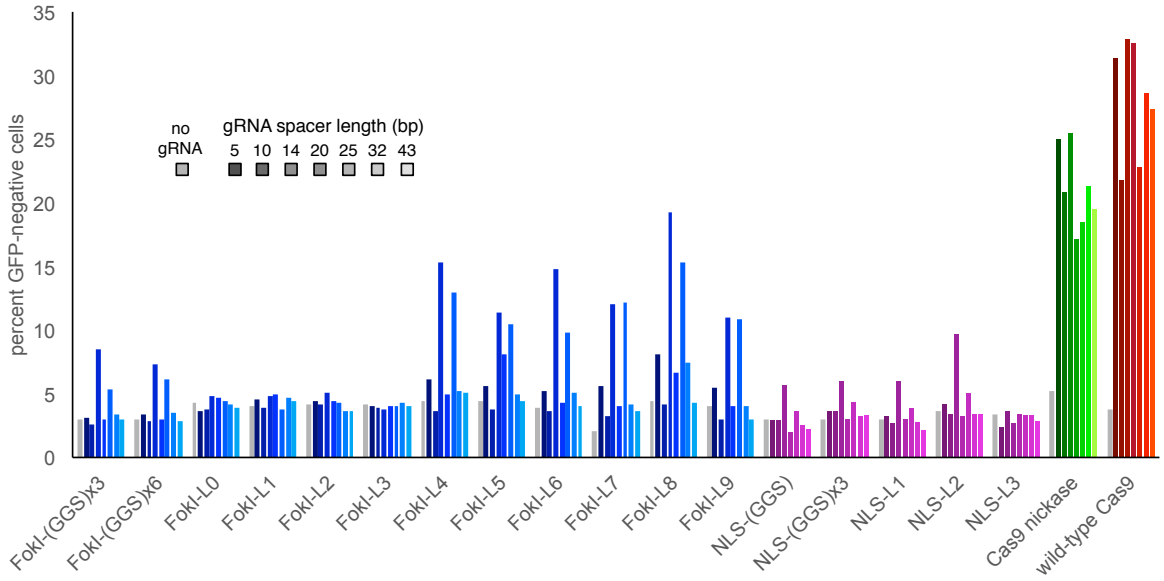


Figure 3.7. Optimization of protein linkers in NLS-*Fok*I-dCas9. (**a**) All linker variants tested. Wild-type Cas9 and Cas9 nickase were included for comparison. The initial active construct NLS-*Fok*I-dCas9 with a $(GGS)_3$ linker between *Fok*I and dCas9 was tested across a range of alternate linkers. The final choice of linkers for fCas9 is highlighted in blue. (**b**) The activity of *Fok*I-dCas9 fusions with linker variants. Each variant was tested across a range of spacer lengths from 5 to 43 bp using gRNA pair orientation A. A control lacking gRNA (grey) was included for each separate fusion construct. NLS-*Fok*I-dCas9 variant L8 showed the best activity, approaching the activity of Cas9 nickase. Variants L4 through L9 show peak activity with 14- and 25-bp spacer lengths, suggesting two optimal spacer lengths roughly one helical turn of dsDNA apart.

Between the *Fok*I domain and dCas9 we identified a flexible 18-residue linker, $(GGS)_6$, and a 16-residue "XTEN" linker (*Fok*I-L8 in **Figure 3.7**) based on a previously reported engineered protein with an open, extended conformation,[27] as supporting the highest levels of genomic GFP modification (**Figure 3.7**).

The XTEN protein was originally designed to extend the serum half-life of translationally fused biologic drugs by increasing their hydrodynamic radius, acting as protein-based functional analog to chemical PEGylation. Possessing a chemically stable, non-cationic, and non-hydrophobic primary sequence, and an extended conformation, we hypothesized that a portion of XTEN could functional as a stable, inert linker sequence for fusion proteins. The sequence of the XTEN protein tag from E-XTEN was analyzed, and repeating motifs within the amino acid sequence were aligned. The sequence used in the *Fok*I-dCas9 fusion construct *Fok*I-L8 (**Figure 3.7**) was derived from the consensus sequence of a common E-XTEN motif, and a 16 amino acid sequence was chosen from within this motif to test as a *Fok*I-dCas9 linker.

Many of the *Fok*I-dCas9 linkers tested including the optimal XTEN linker resulted in nucleases with a marked preference for spacer lengths of ~15 and ~25 bp between half-sites, with all other spacer lengths, including 20 bp, showing substantially lower activity (**Figure 3.7**). This pattern of linker preference is consistent with a model in which the *Fok*I-dCas9 fusions must bind to opposite faces of the DNA double helix to cleave DNA, with optimal binding taking place ~1.5 or 2.5 helical turns apart. The variation of NLS-*Fok*I linkers did not strongly affect nuclease performance, especially when combined with the XTEN *Fok*I-dCas9 linker (**Figure 3.7**). In addition to assaying linkers between the *Fok*I domain and dCas9 in the NLS-*Fok*I-dCas9 architecture, we also tested four linker variants between the N-terminal NLS and the *Fok*I domain (**Figure 3.7**). Although a NLS-GSAGSAAGSGEF-*Fok*I-dCas9 linker exhibited nearly 2-fold better GFP gene modification than the other NLS-*Fok*I linkers tested when a simple GGS linker was used between the *Fok*I and dCas9 domains (**Figure 3.7**), the GSAGSAAGSGEF linker did not perform substantially better when combined with the XTEN linker between the *Fok*I and dCas9 domains.

3.3 Characterizing the activity and specificity of *Fok*I-dCas9 (fCas9)

The NLS-GGS-FokI-XTEN-dCas9 construct consistently exhibited the highest activity among the tested candidates, inducing loss of GFP in ~15% of cells, compared to ~20% and ~30% for Cas9 nickases and wild-type Cas9 nuclease, respectively (**Figure 3.8**).

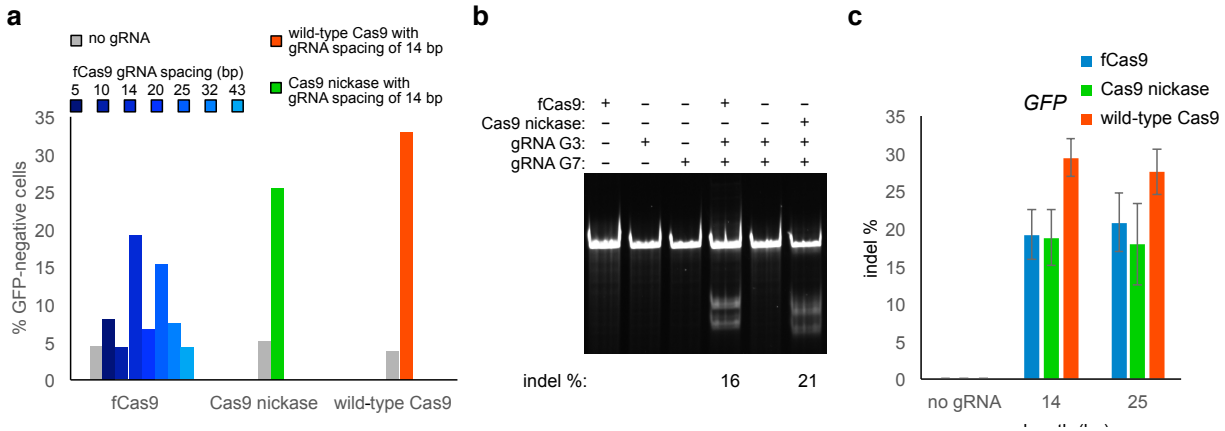
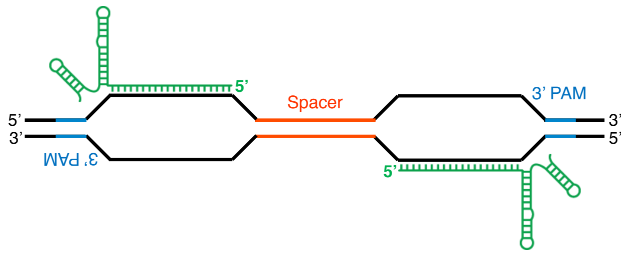


Figure 3.8. Genomic DNA modification by fCas9, Cas9 nickase, and wild-type Cas9 at GFP target site. (a) GFP disruption activity of fCas9, Cas9 nickase, or wild-type Cas9 with either no gRNA, or gRNA pairs of variable spacer length targeting the GFP gene in orientation A. (b) Indel modification efficiency from PAGE analysis of a Surveyor cleavage assay of renatured target-site DNA amplified from cells treated with fCas9, Cas9 nickase, or wild-type Cas9 and two gRNAs spaced 14 bp apart targeting the GFP site (gRNAs G3 and G7), each gRNA individually, or no gRNAs. The indel modification percentage is shown below each lane for samples with modification above the detection limit (~2%). (c) Indel modification efficiency for (c) two pairs of gRNAs spaced 14 or 25bp apart targeting the GFP site.

All subsequent experiments were performed using this construct, hereafter referred to as fCas9. To confirm the ability of fCas9 to efficiently modify genomic target sites, we used the T7 endonuclease I Surveyor assay[28] to measure the amount of mutation at each of seven target sites within the integrated GFP gene in HEK293 cells treated with fCas9, Cas9 nickase, or wild-type Cas9 and either two distinct gRNAs in orientation A or no gRNAs as a negative control. Consistent with our flow cytometry-based studies, fCas9 was able to modify the GFP target sites with optimal spacer lengths of ~15 or ~25 bp at a rate of ~20%, comparable to the efficiency of nickase-induced modification and approximately two-thirds that of wild-type Cas9 (**Figure 3.8)**.

Next we evaluated the ability of the optimized fCas9 to modify four distinct endogenous genomic loci by Surveyor assay. *CLTA* (two sites), *EMX* (two sites), *HBB* (six sites) *VEGF* (three sites), and were targeted with two gRNAs per site in orientation A spaced at various lengths (**Figure 3.9)**.

orientation A:




```
                                                                                                       Spacer (bp)
CLTA-1
CCCCAAGTCTAGCAAGCAGGCCAAAGATGTCTCCCGCATGCGCTCAGTCCTCATCTCCCTCAAGCAGG
CCCCAAGTCTAGCAAGCAGGCCA (C1)                 AGTCCTCATCTCCCTCAAGCAGG (C2)                        22

CLTA-2
CCCTGTGGAAACACTACATCTGCAATATCTTAATCCTACTCAGTGAAGCTCTTCACAGTCATTGG
CCCTGTGGAAACACTACATCTGC (C3)              GTGAAGCTCTTCACAGTCATTGG (C4)                           19

HBC
CCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGACAGGTTTAAGG
CCGTTACTGCCCTGTGGGGCAAG(H1) CGTGGATGAAGTTGGTGGTGGG (H2)                                         5
                                  TGAAGTTGGTGGTGAGGCCCTGG (H3)                                   11
                                       TTGGTGGTGAGGCCCTGGGCAGG (H4)                              16
                                           TGGTGAGGCCCTGGGCAGGTTGG (H5)                          20
                                                   CCCTGGGCAGGTTGGTATCAAGG (H6)                  28
                                                                      AAGGTTACAAGACAGGTTTAAGG (H7) 47

EMX
CCCTTCTTCTTCTGCTCGGACTCAGGCCCTTCCTCCTCCAGCTTCTGCCGTTTGTACTTTGTCCTCCGGTTCTGG
CCCTTCTTCTTCTGCTCGGACTC (E1)                  GCCGTTTGTACTTTGTCCTCCGG (E2)                       23
                                                    TGTACTTTGTCCTCCGGTTCTGG (E3)                 29

VEGF
CCAGGAGCAAACTCCCCCCACCCCCTTTCCAAAGCCCATTCCCTCTTTAGCCAGAGCCGGGGTGTGCAGACGG
CCAGGAGCAAACTCCCCCCACCC (V1)         ATTCCCTCTTTAGCCAGAGCCGG (V2)                                14
                                       TCCCTCTTTAGCCAGAGCCGGGG (V3)                              16
                                                  CCAGAGCCGGGGTGTGCAGACGG (V4)                   27
```

Figure 3.9. Target DNA sequences in endogenous human *EMX*, *VEGF*, *CLTA,* and *HBB* genes. gRNA target sites tested within endogenous human *EMX*, *VEGF*, *CLTA,* and *HBB* genes. Thirteen gRNA target sites were chosen to test the activity of the optimized fCas9 fusion in an orientation in which the PAM is distal from the cleaved spacer sequence (orientation A). Together, these 13 gRNAs enabled testing of fCas9 fusion variants across eight spacer lengths ranging from 5 to 47 bp.

Consistent with the results of the experiments targeting GFP, at appropriately spaced target half-sites fCas9 induced efficient modification of all four genes, ranging from 8% to 22% target chromosomal site modification (**Figure 3.10**).

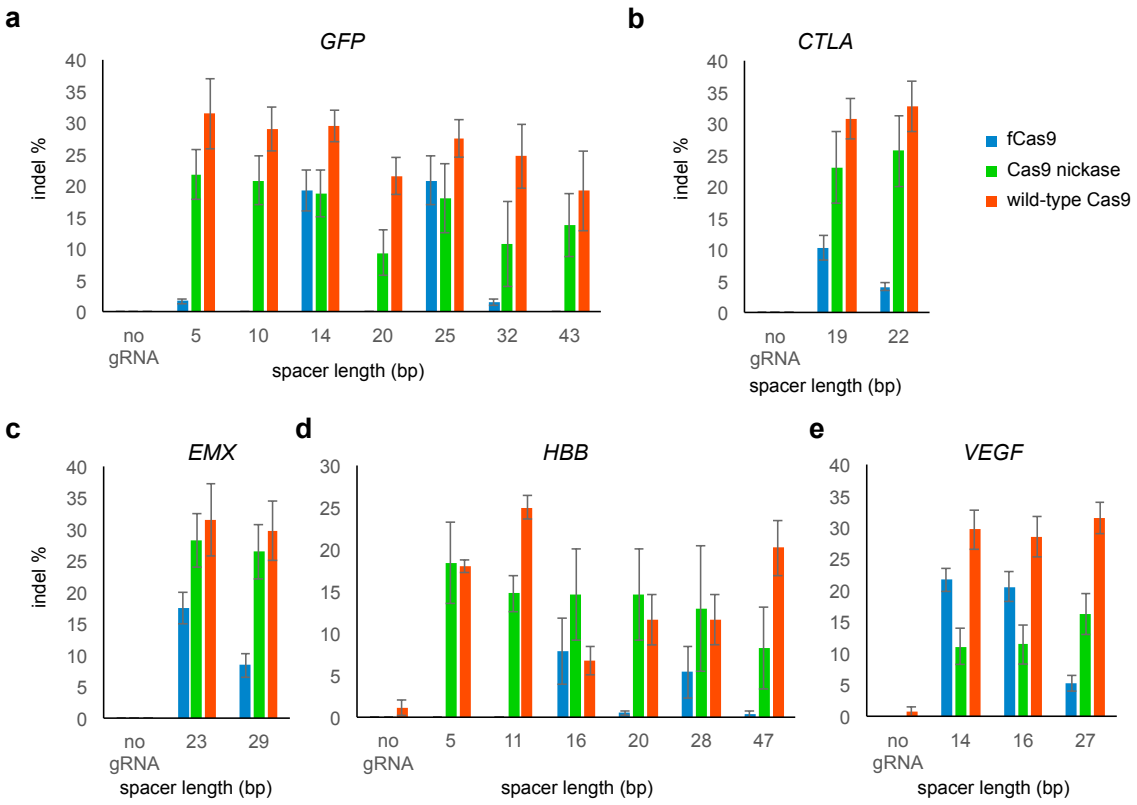


Figure 3.10. Genomic DNA modification by fCas9, Cas9 nickase, and wild-type Cas9 at endogenous human genes. Indel modification efficiency for (**a**) six pairs of gRNAs targeting the *GFP* site, (**b**) two pairs of gRNAs targeting the *CLTA* site, (**c**) two pairs of gRNAs targeting the *EMX* site (**d**) six pairs of gRNAs targeting the *HBB* site, and (**e**) three pairs of gRNAs targeting the *VEGF* site. Error bars reflect standard error of the mean from three biological replicates performed on different days.

Among the gRNA spacer lengths resulting in the highest modification at each of the five genes targeted (including GFP), fCas9 induced on average 15.6% (± 6.3% s.d.) modification, while Cas9 nickase and wild-type Cas9 induced on average 22.1% (± 4.9% s.d.) and 30.4% (± 3.1% s.d.) modification, respectively, from their optimal gRNA pairs for each gene. Because decreasing the amount of Cas9 expression plasmid and gRNA expression plasmid during transfection generally did not proportionally decrease genomic modification activity for Cas9 nickase and fCas9 (**Figure 3.11**), expression was likely not limiting under the conditions tested.

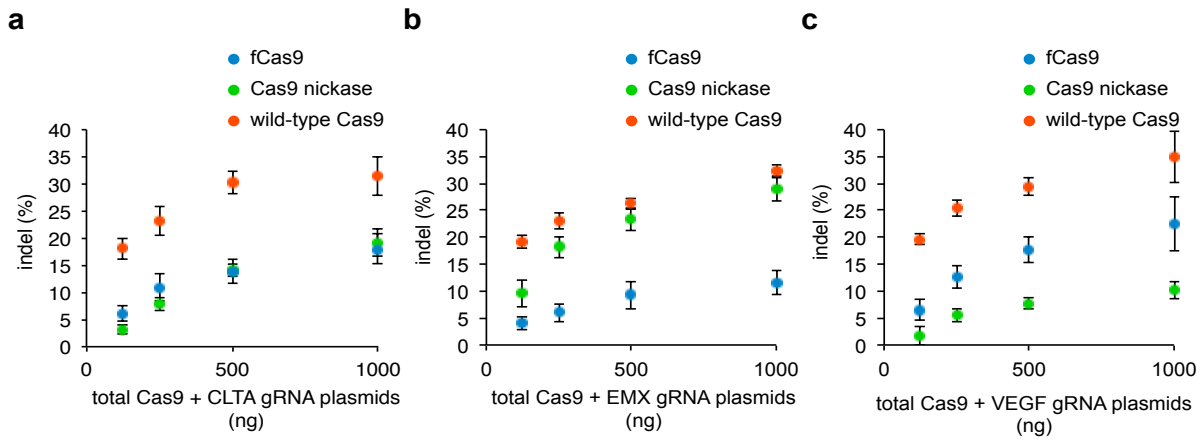


Figure 3.11. Efficiency of genomic DNA modification by fCas9, Cas9 nickase, and wild-type Cas9 with varying amounts of Cas9 and gRNA expression plasmids. Indel modification efficiency from a Surveyor assay of renatured target-site DNA amplified from a population of cells treated with fCas9, Cas9 nickase, or wild-type Cas9 and two target site gRNAs. Titrating the total amount of expression plasmids (Cas9 expression + gRNA expression plasmid), 700/250, 350/125, 175/62.5, 88/31 ng of Cas9 expression plasmid/ng of gRNA expression plasmid were combined with inert carrier plasmid to ensure uniform transfection of 950 ng of plasmid across all treatments. Indel modification efficiency for (**a**) gRNAs spaced 19-bp apart targeting the *CLTA* site, (**b**) gRNAs spaced 23 bp apart targeting the *EMX* site, and (**c**) gRNAs spaced 14 bp apart targeting the *VEGF* site. Error bars represent the standard error of the mean from three biological replicates performed on separate days.

As the gRNA requirements of fCas9 potentially restricts the number of potential off-target substrates of fCas9, we compared the effect of guide RNA orientation on the ability of fCas9, Cas9 nickase, and wild-type Cas9 to cleave target GFP sequences. Consistent with previous reports,[9,14,15] Cas9 nickase efficiently cleaved targets when guide RNAs were bound either in orientation A or orientation B, similar to wild-type Cas9 (**Figure 3.12**). In contrast, fCas9 only cleaved the GFP target when guide RNAs were aligned in orientation A (**Figure 3.12**). This orientation requirement further limits opportunities for undesired off-target DNA cleavage.

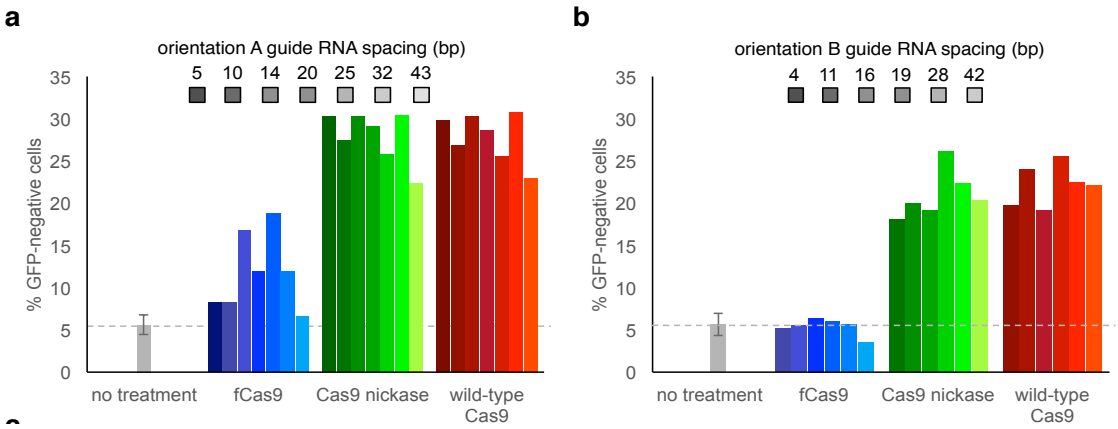


Figure 3.12. DNA modification of fCas9, Cas9 nickase, and wild-type Cas9 as function of gRNA spacer length and orientation. (**a**) GFP gene disruption by wild-type Cas9, Cas9 nickase, fCas9 using gRNA pairs in orientation A. High activity of fCas9 requires spacer lengths of ~15 and 25 bp, roughly one DNA helical turn apart. (**b**) GFP gene disruption using gRNA pairs in orientation B. Cas9 nickase, but not fCas9, accepts either orientation of gRNA pairs.

Importantly, no modification was observed by GFP disruption or Surveyor assay when any of four single gRNAs were expressed individually with fCas9, as expected since two simultaneous binding events are required for *Fok*I activity (**Figure 3.8 and Figure 3.13).** In contrast, GFP disruption resulted from expression of any single gRNA with wild-type Cas9 (as expected) and, for two single gRNAs, with Cas9 nickase (**Figure 3.13**).

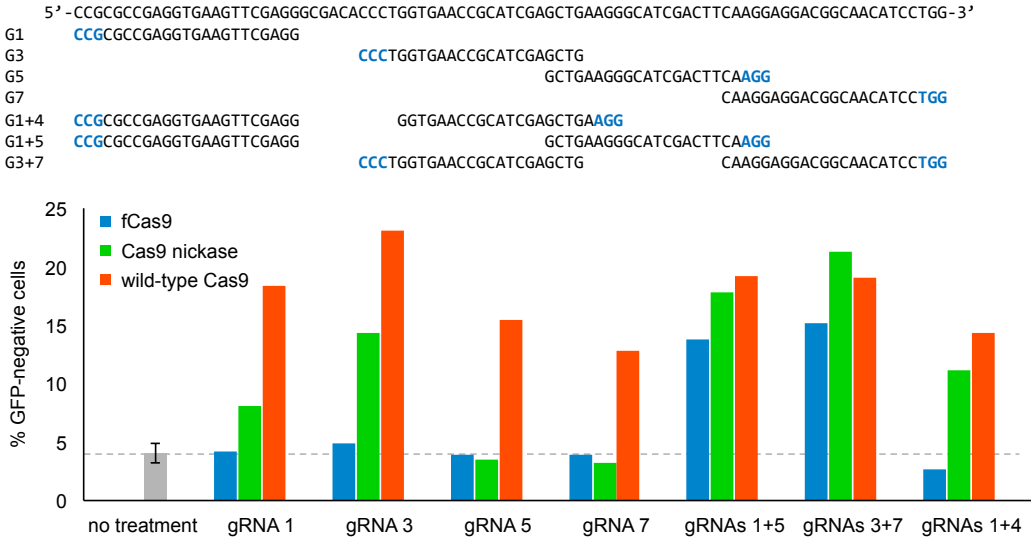


Figure 3.13. Ability of fCas9, Cas9 nickase, and wild-type Cas9 to disrupt GFP in the presence of a single gRNA. GFP gene disruption by fCas9, but not Cas9 nickase or wild-type Cas9, depends on the presence of two gRNAs. Four single gRNAs were tested along with three gRNA pairs of varying spacer length. In the presence of gRNA pairs in orientation A with spacer lengths of 14 or 25 bp (gRNAs 1+5, and gRNAs 3+7, respectively), fCas9 is active, but

not when a gRNA pair with a 10-bp spacer (gRNAs 1+4) is used. "no treatment" refers to cells receiving no plasmid DNA.

Surprisingly, Surveyor assay revealed that although GFP was heavily modified by wild-type Cas9 with single gRNAs, neither fCas9 nor Cas9 nickase showed detectable modification (< ~2%) in cells treated with single gRNAs (**Figure 3.14**).

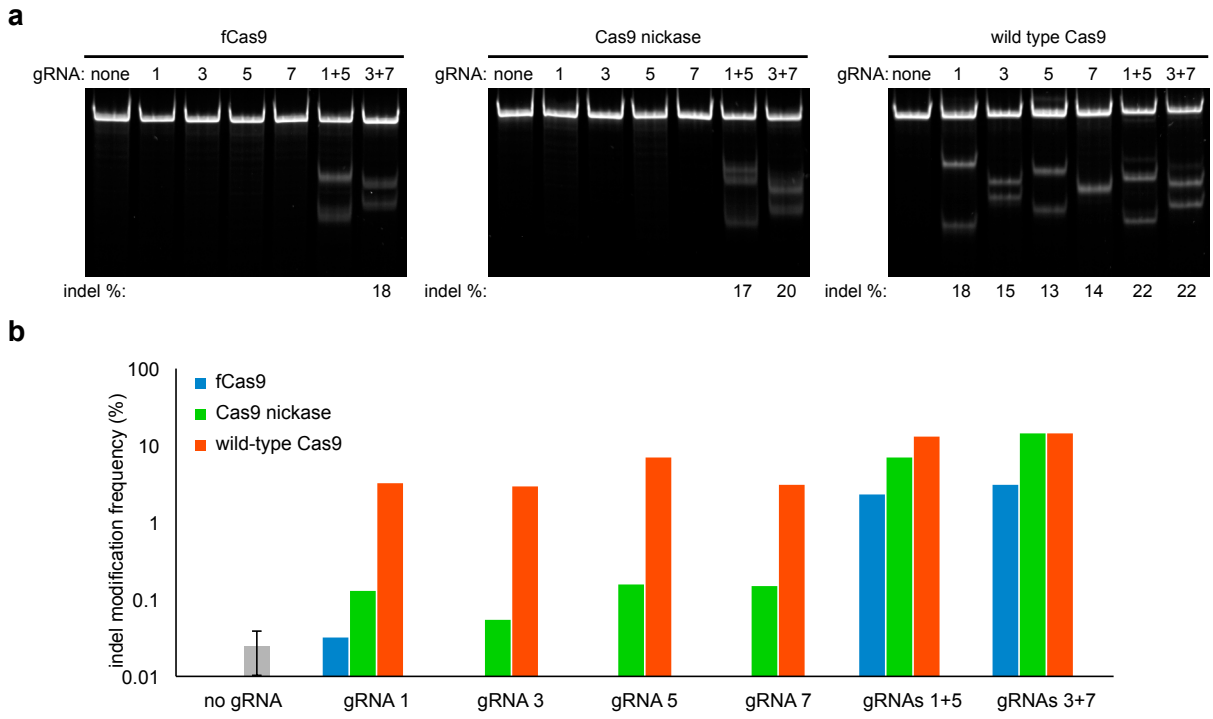


Figure 3.14. Ability of fCas9, Cas9 nickase, and wild-type Cas9 to modify genomic DNA in the presence of a single gRNA. (**a**) Surveyor assay of a genomic GFP target from DNA of cells treated with the indicated combination of Cas9 protein and gRNA(s). Single gRNAs do not induce genome modification at a detectable level (< 2% modification) for both fCas9 and Cas9 nickase. Wild-type Cas9 effectively modifies the GFP target for all tested single and paired gRNAs. For both fCas9 and Cas9 nickase, appropriately paired gRNAs induce genome modification at levels comparable to those of wild-type Cas9. (**b**) Results from sequencing *GFP* on-target sites amplified from 150 ng genomic DNA isolated from human cells treated with a plasmid expressing either wild-type Cas9, Cas9 nickase, or fCas9 and either a single plasmid expressing a single gRNAs (G1, G3, G5 or G7), or two plasmids each expressing a different gRNA (G1+G5, or G3+G7). As a negative control, transfection and sequencing were performed in triplicate as above without any gRNA expression plasmids. Error bars represent s.d. Sequences with more than one insertion or deletion at the GFP target site (the start of the G1 binding site to the end of the G7 binding site) were considered indels. Indel percentages were calculated by dividing the number of indels by total number of sequences. While wild-type Cas9 produced indels across all gRNA treatments, fCas9 and Cas9 nickase produced indels efficiently (> 1%) only when paired gRNAs were present. Indels induced by fCas9 and single gRNAs were

not detected above the no-gRNA control, while Cas9 nickase and single gRNAs modified the target GFP sequence at an average rate of 0.12%.

High-throughput sequencing to detect indels at the GFP target site in cells treated with a single gRNA and fCas9, Cas9 nickase, or wild-type Cas9 revealed the expected substantial level of modification by wild-type Cas9 (3-7% of sequence reads). Modification by fCas9 in the presence of any of the four single gRNAs was not detected above background (< ~0.03% modification), consistent with the requirement of fCas9 to engage two gRNAs in order to cleave DNA. In contrast, Cas9 nickases in the presence of single gRNAs resulted in modification levels ranging from 0.05% to 0.16% at the target site (**Figure 3.14**). The detection of bona fide indels at target sites following Cas9 nickase treatment with single gRNAs confirms the mutagenic potential of genomic DNA nicking, consistent with previous reports.[14,17–19]

The observed rate of nickase-induced DNA modification, however, did not account for the much higher GFP disruption signal in the flow cytometry assay (**Figure 3.13 and Figure 3.14**). Since the gRNAs that induced GFP signal loss with Cas9 nickase (gRNAs G1 and G3) both target the non-template strand of the GFP gene, and since targeting the non-template strand with dCas9 in the coding region of a gene is known to mediate efficient transcriptional repression,[29] we speculate that Cas9 nickase combined with the G1 or G3 single guide RNAs induced substantial transcriptional repression, in addition to a low level of genome modification. The same effect was not seen for fCas9, suggesting that fCas9 may be more easily displaced from DNA by transcriptional machinery. Taken together, these results indicate that fCas9 can modify genomic DNA efficiently and in a manner that requires simultaneous engagement of two guide RNAs targeting adjacent sites, unlike the ability of wild-type Cas9 and Cas9 nickase to cleave DNA when bound to a single guide RNA.

The above results collectively reveal much more stringent spacer, gRNA orientation, and guide RNA pairing requirements for fCas9 (**Figure 3.15)** compared with Cas9 nickase (**Figure 3.10 and Figure 3.12**). In contrast with fCas9, Cas9 nickase cleaved sites across all spacers assayed (5- to 47- bp in orientation A and 4 to 42 bp in orientation B in this work).

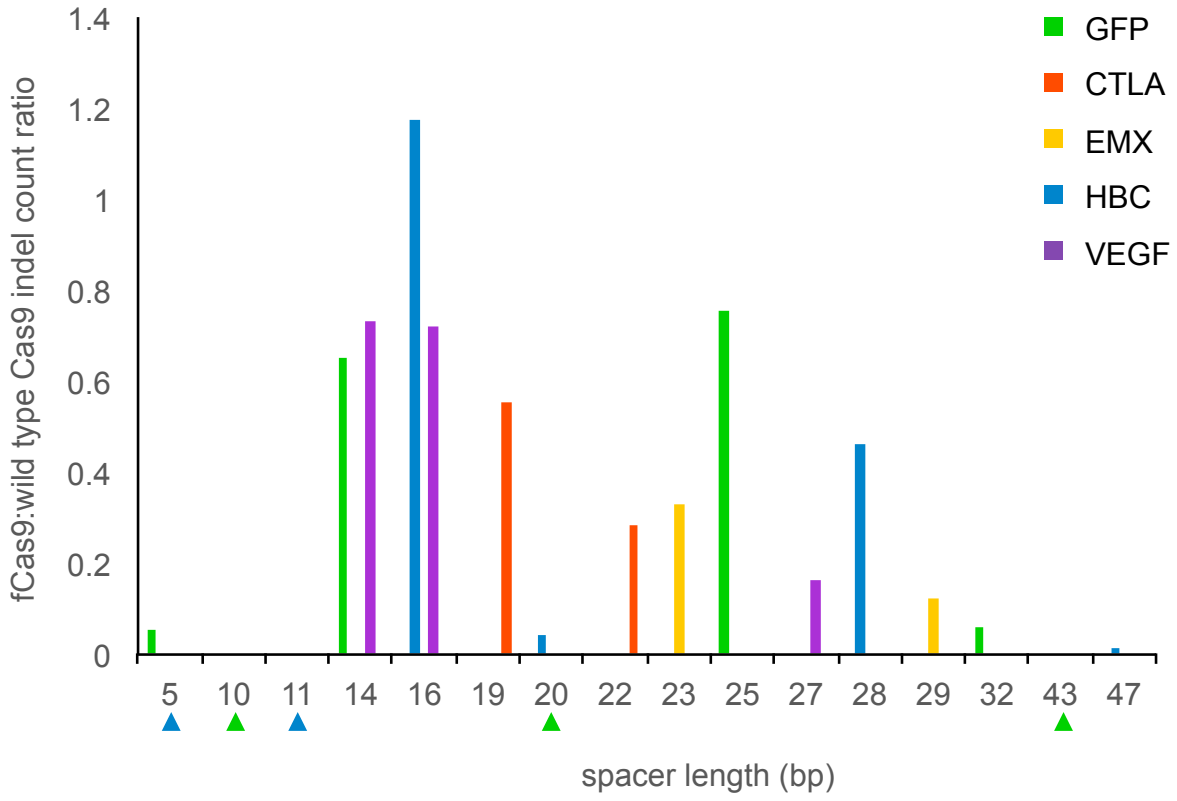


Figure 13.15. fCas9 indel frequency of genomic targets reflects gRNA pair spacer length preference. The graph shows the relationship between spacer length (number of bp between two gRNAs) and the indel modification efficiency of fCas9 normalized to the indel modification efficiency of the same gRNAs co-expressed with wild-type Cas9 nuclease. Colored triangles below the X-axis denote spacer lengths that were tested but which yielded no detectable indels for the indicated target gene. These results suggest that fCas9 requires ~15 bp or ~25 bp between half-sites to efficiency cleave DNA.

These observations are consistent with previous reports of Cas9 nickases modifying sites targeted by gRNAs with spacer lengths up to 100 bp apart.[15] The more stringent spacer and gRNA orientation requirements of fCas9 compared with Cas9 nickase reduces the number of potential genomic off-target sites of the former by approximately 10-fold (**Table 3.1**).

Spacer length (b)	Number of paired gRNA sites in orientation A	Number of paired gRNA sites in orientation B
-8	6874293	NC
-7	6785996	NC
-6	6984064	NC
-5	7023260	NC
-4	6487302	NC
-3	6401348	NC
-2	6981383	NC
-1	7230098	NC
0	7055143	NC
1	6598582	NC
2	6877046	NC
3	6971447	NC
4	6505614	5542549
5	6098107	5663458
6	6254974	6819289
7	6680118	6061225
8	7687598	5702252
9	6755736	7306646
10	6544849	6387485
11	6918186	6172852
12	6241723	5799496
13	6233385	7092283
14	6298717	7882433
15	6181422	7472725
16	6266909	6294684
17	6647352	6825904
18	6103603	6973590
19	5896092	6349456
20	6000683	5835825
21	5858015	6056352
22	6116108	6531913
23	5991254	6941816
24	6114969	6572849
25	6135119	5671641

b

Cas9 variant	Preferred spacer lengths (bp)	Total sites
fCas9	13 to 19, or 22 to 29, in orientation A	92354891
Cas9 nickase	-8 to 100 in orientation A 4 to 42 in orientation B	953048977

Table 3.1 (Continued). Paired gRNA target site abundances for fCas9 and Cas9 nickase in the human genome.

Table 3.1 (Continued). Paired gRNA target site abundances for fCas9 and Cas9 nickase in the human genome. (**a**) Column 2 shows the number of sites in the human genome with paired gRNA binding sites in orientation A allowing for a spacer length from -8 bp to 25 bp (column 1) between the two gRNA binding sites. gRNA binding sites in orientation A have the NGG PAM sequences distal from the spacer sequence ($CCNN_{20}$-spacer-$N_{20}NGG$). Column 3 shows the number of sites in the human genome with paired gRNA binding sites in orientation B allowing for a spacer length from 4 to 25 bp (column 1) between the two gRNA binding sites. gRNA binding sites in orientation B have the NGG PAM sequences adjacent to the spacer sequence ($N_{20}NGG$ spacer $CCNN_{20}$) . NC indicates the number of sites in the human genome was not calculated. Negative spacer lengths refer to target gRNA binding sites that overlap by the indicated number of base pairs. (**b**) Sum of the number of paired gRNA binding sites in orientation A with spacer lengths of 13 to 19 bp, or 22 to 29 bp, the spacer preference of fCas9 (**Figure 3.15**). Sum of the number of paired gRNA binding sites with spacer lengths of -8 bp to 100 bp in orientation A, or 4 to 42 bp in orientation B, the spacer preference of Cas9 nickases (4 to 42 bp in orientation B is based on **Figure 3.12**, and -8 bp to 100 bp in orientation A is based on previous reports[2,3]).

Although the more stringent spacer requirements of fCas9 also reduce the number of potential targetable sites, sequences that conform to the fCas9 spacer and dual PAM requirements exist in the human genome on average once every 34 bp (9.2×10^7 sites in 3.1×10^9 bp). We also anticipate that the growing number of Cas9 homologs with different PAM specificities[30] will further increase the number of targetable sites using the fCas9 approach.

To evaluate the DNA cleavage specificity of fCas9, we measured the modification of known Cas9 off-target sites of *CLTA*, *EMX*, and *VEGF* genomic target sites.[5,10,12,15] The target site and its corresponding known off-target sites (**Table 3.1**) were amplified from genomic DNA isolated from HEK293 cells treated with fCas9, Cas9 nickase, or wild-type Cas9 and two gRNAs spaced 19 bp apart targeting the *CLTA* site, two gRNAs spaced 23 bp apart targeting the *EMX* site, two gRNAs spaced 14 bp apart targeting the *VEGF* site, or two gRNAs targeting an unrelated site (GFP) as a negative control.

	Genomic target site
EMX_On	GAGTCCGAGCAGAAGAAGAAGGG
EMX_Off1	GAGgCCGAGCAGAAGAAagACGG
EMX_Off2	GAGTCCtAGCAGgAGAAGAAGaG
EMX_Off3	GAGTCtaAGCAGAAGAAGAAGaG
EMX_Off4	GAGTtaGAGCAGAAGAAGAAAGG
VEG_On	GGGTGGGGGGAGTTTGCTCCTGG
VEG_Off1	GGaTGGaGGGAGTTTGCTCCTGG
VEG_Off2	GGGaGGGtGGAGTTTGCTCCTGG
VEG_Off3	cGGgGGaGGGAGTTTGCTCCTGG
VEG_Off4	GGGgaGGGGaAGTTTGCTCCTGG
CLT2_On	GCAGATGTAGTGTTTCCACAGGG
CLT2_Off1	aCAaATGTAGTaTTTCCACAGGG
CLT2_Off2	cCAGATGTAGTaTTcCCACAGGG
CLT2_Off3	ctAGATGaAGTGcTTCCACATGG

Table 3.2. Known off-target substrates of Cas9 target sites in *EMX*, *VEGF,* and *CLTA*. List of genomic on-target and off-targets sites of the EMX, VEGF, and CLTA are shown with mutations from on-target in lower case and red. PAMs are shown in blue.

In total 11 off-target sites were analyzed by high-throughput sequencing. Sequences containing insertions or deletions of two or more base pairs in potential genomic off-target sites and present

in significantly greater numbers (P value < 0.005, Fisher's exact test) in the target gRNA-treated samples versus the control gRNA-treated samples were considered Cas9 nuclease-induced genome modifications. The sensitivity of the high-throughput sequencing method for detecting genomic off-target cleavage is limited by the amount genomic DNA (gDNA) input into the PCR amplification of each genomic target site. A 1 ng sample of human gDNA represents only ~330 unique genomes, and thus only ~330 unique copies of each genomic site are present. PCR amplification for each genomic target was performed on a total of 150 ng of input gDNA, which provides amplicons derived from at most 50,000 unique gDNA copies. Therefore, the high-throughput sequencing assay cannot detect rare genome modification events that occur at a frequency of less than 1 in 50,000, or 0.002%.

For 10 of the 11 off-target sites assayed, fCas9 did not result in any detectable genomic off-target modification within the sensitivity limit of our assay (< 0.002%), while demonstrating substantial on-target modification efficiencies of 5% to 10% (**Figure 13.16 and Table 3.3**).

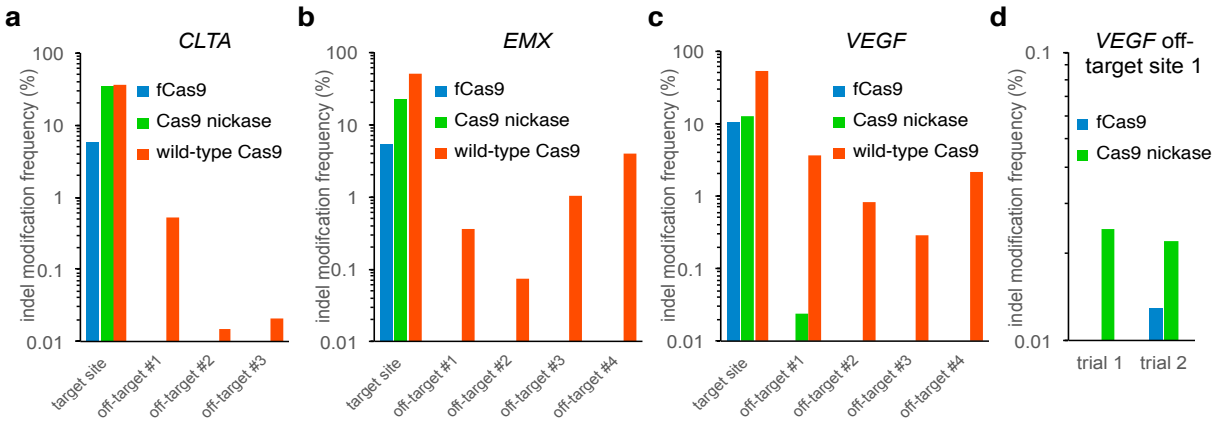


Figure 3.16. DNA modification specificity of fCas9, Cas9 nickase, and wild-type Cas9. The indel mutation frequency from high-throughput DNA sequencing of amplified genomic on-target sites and off-target sites from human cells treated with fCas9, Cas9 nickase, or wild-type Cas9 and (**a**) two gRNAs spaced 19 bp apart targeting the *CLTA* site (gRNAs C1 and C2), (**b**) two gRNAs spaced 23 bp apart targeting the *EMX* site (gRNAs E1 and E2), or (**c, d**) two gRNAs spaced 14 bp apart targeting the *VEGF* site (gRNAs V1 and V2). (**d**) Two in-depth trials to measure genome modification at *VEGF* off-target site 1. Trial 1 used 150 ng of genomic input DNA and > 8 x 10^5 sequence reads for each sample; trial 2 used 600 ng of genomic input DNA and > 23 x 10^5 sequence reads for each sample. All significant (P value < 0.005 Fisher's Exact Test) indel frequencies are shown. Each on- and off-target sample was sequenced once with > 10,000 sequences analyzed per on-target sample and an average of 76,260 sequences analyzed per off-target sample.

a

Nuclease type:	wt Cas9	wt Cas9	Cas9 nickase	fCas9	wt Cas9	Cas9 nickase	fCas9
gRNA pair target:	*CLTA*	*CLTA*	*CLTA*	*CLTA*	*GFP*	*GFP*	*GFP*
Total expression plasmids (ng):	1000	125	1000	1000	1000	1000	1000
CLTA Sites							
CLT2_On							
Indels	3528	1423	3400	575	3	13	5
Total	10000	10000	10000	10000	10000	10000	10000
Modified (%)	35.280	14.230	34.000	5.750	0.030	0.130	0.050
P-value	<1.0E-300	<1.0E-300	<1.0E-300	1.4E-163			
On:off specificity	1	1		1			
CLT2_Off1							
Indels	316	44	2	2	1	3	3
Total	60620	64755	71537	63079	93883	91306	82055
Modified (%)	0.521	0.068	0.003	0.003	<0.002	0.003	0.004
P-value	1.3E-126	2.1E-16					
On:off specificity	68	209		>2850			
CLT2_Off2							
Indels	11	5	3	1	1	1	2
Total	72596	51093	59632	35541	69114	64412	39978
Modified (%)	0.015	0.010	0.005	0.003	<0.002	<0.002	0.005
P-value	6.5E-03						
On:off specificity	2328	1454		>2850			
CLT2_Off3							
Indels	11	10	0	0	1	1	1
Total	52382	44212	54072	48668	55670	58707	54341
Modified (%)	0.021	0.023	<0.002	<0.002	<0.002	<0.002	<0.002
P-value	2.7E-03	3.5E-03					
On:off specificity	1680	629		>2850			

Table 3.3 (Continued). Cellular modification induced by wild-type Cas9, Cas9 nickase, and fCas9 at on-target and off-target genomic sites.

b

Nuclease type:	wt Cas9	wt Cas9	Cas9 nickase	fCas9	wt Cas9	Cas9 nickase	fCas9
gRNA pair:	*EMX*	*EMX*	*EMX*	*EMX*	*GFP*	*GFP*	*GFP*
Total expression plasmids (ng):	1000	125	1000	1000	1000	1000	1000
EMX Site							
EMX_On							
Indels	5111	2683	2267	522	0	0	2
Total	10000	10000	10000	10000	10000	10000	10000
Modified (%)	51.110	26.830	22.670	5.220	<0.002	<0.002	0.020
P-value	<1.0E-300	<1.0E-300	<1.0E-300	1.0E-154			
On:off specificity	1	1	1	1			
EMX_Off1							
Indels	386	122	7	1	4	9	7
Total	109787	83420	124564	88424	102817	90020	96526
Modified (%)	0.352	0.146	0.006	<0.002	0.004	0.010	0.007
P-value	1.3E-103	2.8E-37					
On:off specificity	145	183	>11222	>2584			
EMX_Off2							
Indels	74	58	3	6	3	0	4
Total	98568	94108	105747	78871	81717	79469	79193
Modified (%)	0.075	0.062	0.003	0.008	0.004	<0.002	0.005
P-value	3.2E-16	1.4E-12					
On:off specificity	681	435	>11222	>2584			
EMX_Off3							
Indels	736	178	20	14	12	11	17
Total	72888	65139	82348	59593	74341	73408	75080
Modified (%)	1.010	0.273	0.024	0.023	0.016	0.015	0.023
P-value	2.5E-202	3.1E-44					
On:off specificity	51	98	>11222	>2584			
EMX_Off4							
Indels	4149	620	3	3	6	7	5
Total	107537	91695	91368	91605	111736	119643	128088
Modified (%)	3.858	0.676	0.003	0.003	0.005	0.006	0.004
P-value	<1.0E-300	1.9E-202					
On:off specificity	13	40	>11222	>2584			

Table 3.3 (Continued). Cellular modification induced by wild-type Cas9, Cas9 nickase, and fCas9 at on-target and off-target genomic sites.

c

Nuclease type:	wt Cas9	wt Cas9	Cas9 nickase	fCas9	wt Cas9	Cas9 nickase	fCas9
gRNA pair:	*VEGF*	*VEGF*	*VEGF*	*VEGF*	*GFP*	*GFP*	*GFP*
Total expression plasmids (ng):	1000	125	1000	1000	1000	1000	1000
VEGF Sites							
VEG_On							
Indels	5253	2454	1230	1041	8	0	1
Total	10000	10000	10000	10000	10000	10000	10000
Modified (%)	52.530	24.540	12.300	10.410	0.080	<0.002	0.010
P-value	<1.0E-300	<1.0E-300	<1.0E-300	6.6E-286			
On:off specificity	1	1	1	1			
VEG_Off1							
Indels	2950	603	22	0	0	4	1
Total	82198	71163	90434	77557	74765	79738	74109
Modified (%)	3.589	0.847	0.024	<0.002	<0.002	0.005	<0.002
P-value	<1.0E-300	3.2E-188	2.5E-06				
On:off specificity	15	29	506	>5150			
VEG_Off2							
Indels	863	72	3	3	0	2	1
Total	102501	49836	119702	65107	54247	65753	61556
Modified (%)	0.842	0.144	0.003	0.005	<0.002	0.003	<0.002
P-value	3.5E-159	9.6E-24					
On:off specificity	62	170	>6090	>5150			
VEG_Off3							
Indels	260	33	3	2	3	1	0
Total	91277	83124	90063	84385	62126	68165	69811
Modified (%)	0.285	0.040	0.003	0.002	0.005	<0.002	<0.002
P-value	6.8E-54	1.0E-05					
On:off specificity	184	618	>6090	>5150			
VEG_Off4							
Indels	1305	149	3	2	3	2	4
Total	59827	41203	65964	57828	60906	61219	62162
Modified (%)	2.181	0.362	0.005	0.003	0.005	0.003	0.006
P-value	<1.0E-300	2.7E-54					
On:off specificity	24	68	>6090	>5150			

Table 3.3 (Continued). Cellular modification induced by wild-type Cas9, Cas9 nickase, and fCas9 at on-target and off-target genomic sites.

d

Nuclease type:	Cas9 nickase	fCas9	Cas9 nickase	fCas9
gRNA pair:	*VEGF*	*VEGF*	*GFP*	*GFP*
Total expression plasmids (ng):	1000	1000	1000	1000
VEGF Sites				
VEG_On				
Indels	2717	2122	10	13
Total	10000	10000	10000	10000
Modified (%)	27.170	21.220	0.100	0.130
P-value	<1.0E-300	<1.0E-300		
On:off specificity	1	1		
VEG_Off1				
Indels	67	30	3	2
Total	302573	233567	204454	190240
Modified (%)	0.022	0.013		
P-value	5.9E-12	2.5E-06		
On:off specificity	1227	1652		

Table 3.3 (Continued). Cellular modification induced by wild-type Cas9, Cas9 nickase, and fCas9 at on-target and off-target genomic sites. (**a**) Results from sequencing *CLTA* on-target and previously reported genomic off-target sites amplified from 150 ng genomic DNA isolated from human cells treated with a plasmid expressing either wild-type Cas9, Cas9 nickase, or fCas9 and a single plasmid expressing two gRNAs targeting the *CLTA* on-target site (gRNA C3 and gRNA C4). As a negative control, transfection and sequencing were performed as above, but using two gRNAs targeting the *GFP* gene on-target site (gRNA G1, G2 or G3 and gRNA G4, G5, G6 or G7. Indels: the number of observed sequences containing insertions or deletions consistent with any of the three Cas9 nuclease-induced cleavage. Total: total number of sequence counts while only the first 10,000 sequences were analyzed for the on-target site sequences. Modified: number of indels divided by total number of sequences as percentages. Upper limits of potential modification were calculated for sites with no observed indels by assuming there is less than one indel then dividing by the total sequence count to arrive at an upper limit modification percentage, or taking the theoretical limit of detection (1/49,500), whichever value was larger. P-values: For wild-type Cas9 nuclease, Cas9 nickase or fCas9 nuclease, P-values were calculated as previously reported[18] using a two-sided Fisher's exact test between each sample treated with two gRNAs targeting the *CLTA* on-target site and the control sample treated with two gRNAs targeting the *GFP* on-target site. P-values of < 0.0045 were considered significant and shown based on conservative multiple comparison correction using the Bonferroni method. On:off specificity is the ratio of on-target to off-target genomic modification frequency for each site. (**b**) Experimental and analytic methods as in (**a**) applied to *EMX* target sites using a single plasmid expressing two gRNAs targeting the *EMX* on-target site (gRNA E1 and gRNA E2). (**c**) Experimental and analytic methods as in (**a**) applied to *VEGF* target sites using a single plasmid expressing two gRNAs targeting the *VEGF* on-target site (gRNA V1 and gRNA v2).

The detailed inspection of fCas9-modified *VEGF* on-target sequences (**Figure 13.17**) revealed a prevalence of deletions ranging from two to dozens of base pairs consistent with cleavage occurring in the DNA spacer between the two target binding sites, similar to the effects of *Fok*I nuclease domains fused to zinc finger or TALE DNA-binding domains.[31]

a

```
Wild-type Cas9 nuclease modifications of VEGF on-target site:
4747 gctgtttgggaggtcagaaataggggtCCAGGAGCAAACTCCCCCCACCCcctttccaaagcccATTCCCTCTTTAGCCAGAGCCGGggtgtgcagacggcagtc (ref)
4577 gctgtttgggaggtcagaaataggggtccagga----------------------------------------------------agccggggtgtgcagacggcagtc
  58 gctgtttgggaggtcagaaataggggtccaggagcaaactccccccaccccctttccaaagcccattccctctttagc------cggggtgtgcagacggcagtc
  54 gctgtttgggaggtcagaaataggggtccaggag----------------------------------------------agccggggtgtgcagacggcagtc
  43 gctgtttgggaggtcagaaatag-----------------------------------------------------------ccggggtgtgcagacggcagtc
  33 gctgtttgggaggtcagaaataggggtccaggagc-----------------------------------------------cggggtgtgcagacggcagtc
  23 gctgtttgggaggtcagaaataggggtccag---------------------------------------------------ccggggtgtgcagacggcagtc
  22 gctgtttgggaggtcagaaataggggtccaggagcaaactccccccaccccctttccaaagcccattccctctttagccag------ggtgtgcagacggcagtc
  18 gctgtttgggaggtcagaaataggggtccagga-------------------t---------------------------agccggggtgtgcagacggcagtc

Cas9 nickase modifications of VEGF on-target site:
8770 gctgtttgggaggtcagaaataggggtCCAGGAGCAAACTCCCCCCACCCcctttccaaagcccATTCCCTCTTTAGCCAGAGCCGGggtgtgcagacggcagtc (ref)
  78 gctgtttgggaggtcagaaataggggtccag--------------------------------------------------------------------acggcagtc
  60 gctgtttgggaggtcagaaataggggtccaggagcaaactccccccaccccctttccaaagcc----------------------ggggtgtgcagacggcagtc
  58 gctgtttgggaggtcagaaataggggtcca---------------------------aagcccattccctctttagccagagccggggtgtgcagacggcagtc
  56 gctgtttgggaggtcagaaataggggt-------------------------------------------------------------gtgcagacggcagtc
  49 gctgtttgggaggtcagaaataggggtccag---------------------------------------------------ccggggtgtgcagacggcagtc
  37 gctgtttgggaggtcagaaataggggtccagg-----------------------------------------------------gtgtgcagacggcagtc
  36 gctgtttgggaggtcagaaataggggtccaggagc-----------------------------------------------cggggtgtgcagacggcagtc
  27 gctgtttgggaggtcagaaatag-----------------------------------------------------------ccggggtgtgcagacggcagtc

fCas9 nuclease modifications of VEGF on-target site:
8959 gctgtttgggaggtcagaaataggggtCCAGGAGCAAACTCCCCCCACCCcctttccaaagcccATTCCCTCTTTAGCCAGAGCCGGggtgtgcagacggcagtc (ref)
 125 gctgtttgggaggtcagaaataggggtccaggagcaaactccccca------------agcccattccctctttagccagagccggggtgtgcagacggcagtc
 121 gctgtttgggaggtcagaaataggggtccaggagcaaactccccccaccccctt-----------ttccctctttagccagagccggggtgtgcagacggcagtc
  77 gctgtttgggaggtcagaaataggggtccaggagcaaactccccccaccccct------------ttccctctttagccagagccggggtgtgcagacggcagtc
  73 gctgtttgggaggtcagaaataggggtccaggagcaaactcccccca-------------gcccattccctctttagccagagccggggtgtgcagacggcagtc
  48 gctgtttgggaggtcagaaataggggtccaggagcaaactccccccaccccc------------attccctctttagccagagccggggtgtgcagacggcagtc
  44 gctgtttgggaggtcagaaataggggtccaggagcaaactccccccaccccctttt--------------------agccagagccggggtgtgcagacggcagtc
  24 gctgtttgggaggtcagaaataggggtccaggagcaaactccccccaccccctttt--aaagcccattccctctttagccagagccggggtgtgcagacggcagtc
  22 gctgtttgggaggtcagaaataggggtccaggagcaaactccccc-------------aagcccattccctctttagccagagccggggtgtgcagacggcagtc
```

Figure 3.17 (Continued). Modifications induced by Cas9 nuclease, Cas9 nickases, or fCas9 nucleases at endogenous loci.

b

```
Wild-type Cas9 nuclease modifications of VEG_Off1:
 79248 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccCCAGGAGCAAACTCCCtCCAtCCCacaaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat (ref)
   800 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccagg--caaactccctccatcccacaaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
   239 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccca----------------tcccacaaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
   155 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccag---caaactccctccatcccacaaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
    90 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccca----------------------caaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
    71 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacac------------tccctccatcccacaaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
    54 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccc-------------tccatcccacaaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
    53 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccagga----------------------------------------------------------tgat
    47 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgac---------------tccctccatcccacaaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat

Cas9 nickase modifications of VEG_Off1:
302573 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccCCAGGAGCAAACTCCCtCCAtCCCacaaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat (ref)
    28 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccagga-----c----------------atccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
    13 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgac---------------tccctccatcccacaaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
    11 cattcaacagatacttactgaatgctaatgtctcagacaggacattctga-------------------------------tccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
     4 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccagg--ca---tc--tg-atg--acaaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
     2 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccaggagcaaactccctccatcccacaaatccgtccttagatgtgca-cccaacctcctaagaaatagaaggatgat
     1 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccaggagcaaactccctccatcccacaaatccgtcctt--atgtgcacacccaacctcctaagaaatagaaggatgat
     1 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccagga---------------------------------tgtgcacacccaacctcctaagaaatagaaggatgat
     1 cattcaacagatacttactgaatgctaatgtctcagacaggacattctga-------------------------------aatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat

fCas9 nuclease modifications of VEG_Off1:
233567  cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccCCAGGAGCAAACTCCCtCCAtCCCacaaatccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat (ref)
     6 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccaggagcaaactccctcc-------------gtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
     5 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccaggagcaaactccctc-----------tccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
     4 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccaggagcaaactccctccat--------tccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
     3 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccaggagcaaactccctccatcccacaaatccgtccttagatgtgca--cccaacctcctaagaaatagaaggatgat
     3 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccaggagcaaactccctcc----------tccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
     2 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccaggagcaaactccctcc---------atccgtccttagatgtgcacacccaacctcctaagaaatagaaggatgat
     1 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccaggagcaaactccctccatcccacaaatccgtccttagatgtgcacaccc----tcctaagaaatagaaggatgat
     1 cattcaacagatacttactgaatgctaatgtctcagacaggacattctgacaccccaggagcaaactccctccatcccacaaatccgtcc--agatgtgcacacccaacctcctaagaaatagaaggatgat
```

Figure 3.17 (Continued). Modifications induced by Cas9 nuclease, Cas9 nickases, or fCas9 nucleases at endogenous loci. (**a**) Examples of modified sequences at the *VEGF* on-target site with wild-type Cas9 nuclease, Cas9 nickases, or fCas9 nucleases and a single plasmid expressing two gRNAs targeting the *VEGF* on-target site (gRNA V1 and gRNA V2). For each example shown, the unmodified genomic site is the first sequence, followed by the top eight sequences containing deletions. The numbers before each sequence indicate sequencing counts. The gRNA target sites are bold and capitalized. (**b**) Identical analysis as in (**a**) for *VEGF* off-target site 1VEG_Off1. (**c**) Potential binding mode of two gRNAs to *VEGF* off-target site 1. The top strand is bound in a canonical mode, while the bottom strand binds the second gRNA, gRNA V2, through gRNA:DNA base pairing that includes G:U base pairs.

In contrast, genomic off-target DNA cleavage was observed for wild-type Cas9 at all 11 sites assayed. Using the detection limit of the assay as an upper bound for off-target fCas9 activity, we calculated that fCas9 has a much lower off-target modification rate than wild-type Cas9 nuclease. At the 11 off-target sites modified by wild-type Cas9 nuclease, fCas9 resulted in on-target:off-target modification ratios at least 140-fold higher than that of wild-type Cas9 (**Figure 13.16 and Table 3.3**).

Consistent with previous reports,[5,14,15] Cas9 nickase also induced substantially fewer off-target modification events (1/11 off-target sites modified at a detectable rate) compared to wild-type Cas9. An initial high-throughput sequencing assay revealed significant (P value < 10^{-3}, Fisher's Exact Test) modification induced by Cas9 nickases in 0.024% of sequences at *VEGF* off-target site 1. This genomic off-target site was not modified by fCas9 despite similar *VEGF* on-target modification efficiencies of 12.3% for Cas9 nickase and 10.4% for (**Figure 13.16 and Table 3.3**). Because Cas9 nickase-induced modification levels were within an order of

magnitude of the limit of detection and fCas9 modification levels were undetected, we repeated the experiment with a larger input DNA samples and a greater number of sequence reads (150 versus 600 ng genomic DNA and > 8 x 10^5 versus > 23 x 10^5 reads for the initial and repeated experiments, respectively) to detect off-target cleavage at this site by Cas9 nickase or fCas9. From this deeper interrogation, we observed Cas9 nickase and fCas9 to both significantly modify (P value < 10^{-5}, Fisher's Exact Test) *VEGF* off-target site 1 (**Figure 13.16 and Table 3.3**). For both experiments interrogating the modification rates at *VEGF* off-target site 1, fCas9 exhibited a greater on-target:off-target DNA modification ratio than that of Cas9 nickase (> 5,150 and 1,650 for fCas9, versus 510 and 1,230 for Cas9 nickase, **Figure 3g**).

On either side of *VEGF* off-target site 1 there exist no other sites with six or fewer mutations from either of the two half-sites of the *VEGF* on-target sequence. We speculate that the first 11 bases of one gRNA (V2) might hybridize to the single-stranded DNA freed by canonical Cas9:gRNA binding within VEGF off-target site 1 (**Figure 3.18**).

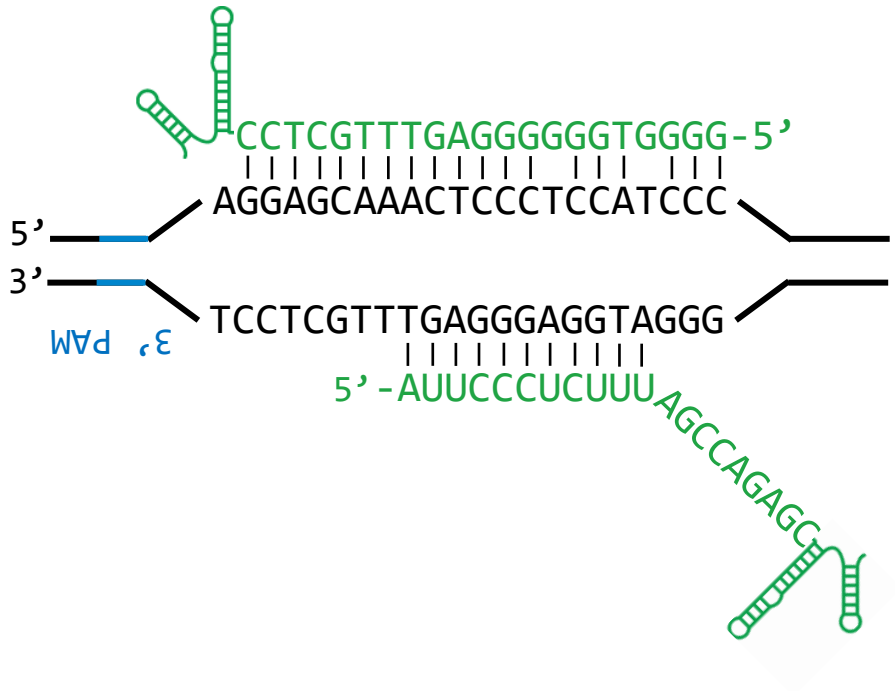


Figure 3.18. Model for dimeric cleavage at *VEGF* off-target site 1. Potential binding model of two gRNAs to *VEGF* off-target site 1. The top strand is bound in a canonical mode, while the bottom strand binds the second gRNA, gRNA V2, through gRNA:DNA base pairing that includes G:U base pairs.

Through this gRNA:DNA hybridization it is possible that a second Cas9 nickase or fCas9 could be recruited to modify this off-target site at a very low, but detectable level. Judicious gRNA pair design could eliminate this potential mode of off-target DNA cleavage, as *VEGF* off-target site 1 is highly unusual in its ability to form 11 consecutive potential base pairs with the second gRNA of a pair. In general, fCas9 was unable to modify the genomic off-target sites tested because of the absence of any adjacent second binding site required to dimerize and activate the *Fok*I nuclease domain.

The optimized *Fok*I-dCas9 fusion architecture developed in this work modified all five genomic loci targeted, demonstrating the generality of using fCas9 to induce genomic modification in human cells, although modification with fCas9 was somewhat less efficient than with wild-type Cas9. The use of fCas9 is straightforward, requiring only that PAM sequences be present with an appropriate spacing and orientation, and using the same gRNAs as wild-type Cas9 or Cas9 nickases. The observed low off-target:on-target modification ratios of fCas9, > 140-fold lower than that of wild-type Cas9, likely arises from the distinct mode of action of dimeric *Fok*I, in which DNA cleavage proceeds only if two DNA sites are occupied simultaneously by two *Fok*I domains at a specified distance (here, ~15 bp or ~25 bp apart) and in a specific half-site orientation. The resulting unusually low off-target activity of fCas9 may enable applications of Cas9:gRNA-based technologies that require a very high degree of target specificity, such as *ex vivo* or *in vivo* therapeutic modification of human cells. This work also provides a foundation for future studies to characterize in greater detail and further improve the DNA cleavage activity and specificity of fCas9 *in vitro* and *in vivo*.

3.4 Methods used to study *Fok*I-dCas9 fusions

Oligonucleotides and PCR

PCR was performed with 0.4 μL of 2 U/μL Phusion Hot Start Flex DNA polymerase (NEB) in 50 μL with 1x HF Buffer, 0.2 mM dNTP mix (0.2 mM dATP, 0.2 mM dCTP, 0.2 mM dGTP, 0.2 mM dTTP) (NEB), 0.5 μM of each primer and a program of: 98 ˚C, 1 min; 35 cycles of [98 ˚C, 15 s; 65 ˚C, 15 s; 72 ˚C, 30 s] unless otherwise noted. All oligonucleotides were purchased from Integrated DNA Technologies (IDT). Oligonucleotide sequences are listed in **Table 3.4**.

dCas9-NLS-FokI primers:	sequence
Cas9_Exp_CNF_Fok1+Plas-Fwd	CGGCGAGATAAACTTTTAA TGACCGGTCATCATCACCA
Cas9_Exp_CNF_Cas9coD10-Rev	CCAACGGAATTAGTGCCGATAGCTAAACCAATAGAATACTTTTTATC
Cas9_Exp_CNF_Cas9coD10-Fwd	GATAAAAAGTATTCTATTGGTTTAGCTATCGGCACTAATTCCGTTGG
Cas9_Exp_CNF_Cas9coH850-Rev	TTCAAAAAGGATTGGGGTACAATGGCATCGACGTCGTAATCAGATAAAC
Cas9_Exp_CNF_Cas9coH850-Fwd	GTTTATCTGATTACGACGTCGATGCCATTGTACCCCAATCCTTTTTGAA
Cas9_Exp_CNF_(Cas9)NLS+GGS-Fok-Rev	TTGGGATCCAGAACCTCCTCCTGCAGCCTTGTCATCG
Cas9_Exp_CNF_(Cas9)NLS+GGS3-Fok-Rev	TTGGGATCCAGAACCTCC GCTGCCGCCACTTCCACCTGA TCCTGCAGCCTTGTCATCG
Cas9_Exp_CNF_(Cas9)NLS+GGS-Fok-Fwd	CGATGACAAGGCTGCAGGAGGAGGTTCTGGATCCCAA
Cas9_Exp_CNF_(Cas9)NLS+GGS3-Fok-Fwd	CGATGACAAGGCTGCAGGA TCAGGTGGAAGTGGCGGCAGC GGAGGTTCTGGATCCCAA
Cas9_Exp_CNF_Fok1+Plas-Rev	TGGTGATGATGACCGGTCA TTAAAAGTTTATCTCGCCG
NLS-dCas9-FokI primers:	
Cas9_Exp_NCF_Fok1+Plas-Fwd	CGGCGAGATAAACTTTTAA TGACCGGTCATCATCACCA
Cas9_Exp_NCF_PlasS+FLAG(NLS-Fok1-Rev	TAGGGAGAGCCGCCACCATGGACTACAAAGACCATGACGG
Cas9_Exp_NCF_NLS+Cas9coD10-Rev	TAAACCAATAGAATACTTTTTATC CATAGGTACCCCGCGGTGAATG
Cas9_Exp_NCF_Cas9coD10-Fwd	GATAAAAAGTATTCTATTGGTTTAGCTATCGGCACTAATTCCGTTGG
Cas9_Exp_NCF_Cas9coH850-Rev	TTCAAAAAGGATTGGGGTACAATGGCATCGACGTCGTAATCAGATAAAC
Cas9_Exp_NCF_Cas9coH850-Fwd	GTTTATCTGATTACGACGTCGATGCCATTGTACCCCAATCCTTTTTGAA
Cas9_Exp_NCF_Cas9End+GGS-Fok-Rev	TTGGGATCCAGAACCTCCGTCACCCCCAAGCTGTG
Cas9_Exp_NCF_Cas9End+GGS3-Fok-Rev	TTGGGATCCAGAACCTCC GCTGCCGCCACTTCCACCTGA GTCACCCCCAAGCTGTG
Cas9_Exp_NCF_Cas9End+GGS-Fok-Fwd	CACAGCTTGGGGGTGACGGAGGTTCTGGATCCCAA
Cas9_Exp_NCF_Cas9End+GGS3-Fok-Fwd	CACAGCTTGGGGGTGAC TCAGGTGGAAGTGGCGGCAGC GGAGGTTCTGGATCCCAA
Cas9_Exp_NCF_Fok1+Plas-Rev	TGGTGATGATGACCGGTCA TTAAAAGTTTATCTCGCCG
FokI-dCas9-NLS primers:	
Cas9_Exp_FCN_PlasS+Fok-Fwd	TAGGGAGAGCCGCCACCATGGGATCCCAACTAGTCAAAAG

Table 3.4 (Continued). Oligonucleotides used in this study.

Cas9_Exp_FCN_Fok1GGS+Cas-Rev	ACCAATAGAATACTTTTTATCCATGCTGCCACCAAAGTTTATCTC
Cas9_Exp_FCN_Fok1GGS3+Cas-Rev	ACCAATAGAATACTTTTTATCCATGCTGCCGCCACTTCCACCTG
Cas9_Exp_FCN_Cas9coD10-Fwd	GATAAAAAGTATTCTATTGGTTTAGCTATCGGCACTAATTCCGTTGG
Cas9_Exp_FCN_Cas9coH850-Rev	CCAACGGAATTAGTGCCGATAGCTAAACCAATAGAATACTTTTTATC
Cas9_Exp_FCN_Cas9coH850-Fwd	GTTTATCTGATTACGACGTCGATGCCATTGTACCCCAATCCTTTTTGAA
Cas9_Exp_FCN_Cas9End+PlasmidEn-Rev	TGGTGATGATGACCGGTCA GTCACCCCCAAGCTGTG
Cas9_Exp_FCN_Cas9End+PlasmidEn-Fwd	CACAGCTTGGGGGTGAC TGACCGGTCATCATCACCA
Cas9_Exp_FCN_PlasS+Fok-Rev	CTTTTGACTAGTTGGGATCCCATGGTGGCGGCTCTCCCTA
gRNA_G1-top	ACACCCCTCGAACTTCACCTCGGCGG
gRNA_G2-top	ACACCGTCGCCCTCGAACTTCACCTG
gRNA_G3-top	ACACCCAGCTCGATGCGGTTCACCAG
gRNA_G4-top	ACACCGGTGAACCGCATCGAGCTGAG
gRNA_G5-top	ACACCGCTGAAGGGCATCGACTTCAG
gRNA_G6-top	ACACCGGCATCGACTTCAAGGAGGAG
gRNA_G7-top	ACACCCAAGGAGGACGGCAACATCCG
gRNA_G8-top	ACACCACCATCTTCTTCAAGGACGAG
gRNA_G9-top	ACACCCAACTACAAGACCCGCGCCGG
gRNA_G10-top	ACACCCCGCGCCGAGGTGAAGTTCGG
gRNA_G11-top	ACACCGAAGTTCGAGGGCGACACCCG
gRNA_G12-top	ACACCTTCGAACTTCACCTCGGCGCG
gRNA_G13-top	ACACCTCAGCTCGATGCGGTTCACCG
gRNA_G14-top	ACACCCGATGCCCTTCAGCTCGATGG
gRNA_G1-bottom	AAAACCGCCGAGGTGAAGTTCGAGGG
gRNA_G2-bottom	AAAACAGGTGAAGTTCGAGGGCGACG
gRNA_G3-bottom	AAAACTGGTGAACCGCATCGAGCTGG
gRNA_G4-bottom	AAAACTCAGCTCGATGCGGTTCACCG
gRNA_G5-bottom	AAAACTGAAGTCGATGCCCTTCAGCG
gRNA_G6-bottom	AAAACTCCTCCTTGAAGTCGATGCCG
gRNA_G7-bottom	AAAACGGATGTTGCCGTCCTCCTTGG
gRNA_G8-bottom	AAAACTCGTCCTTGAAGAAGATGGTG
gRNA_G9-bottom	AAAACCGGCGCGGGTCTTGTAGTTGG
gRNA_G10-bottom	AAAACCGAACTTCACCTCGGCGCGGG
gRNA_G11-bottom	AAAACGGGTGTCGCCCTCGAACTTCG
gRNA_G12-bottom	AAAACGCGCCGAGGTGAAGTTCGAAG
gRNA_G13-bottom	AAAACGGTGAACCGCATCGAGCTGAG
gRNA_G14-bottom	AAAACCATCGAGCTGAAGGGCATCGG
gRNA_C1-top	ACACCTGGCCTGCTTGCTAGACTTGG
gRNA_C3-top	ACACCGCAGATGTAGTGTTTCCACAG
gRNA_H1-top	ACACCCTTGCCCCACAGGGCAGTAAG
gRNA_E1-top	ACACCGAGTCCGAGCAGAAGAAGAAG
gRNA_V1-top	ACACCGGGTGGGGGGAGTTTGCTCCG

Table 3.4 (Continued). Oligonucleotides used in this study.

gRNA_C1-bottom	AAAACCAAGTCTAGCAAGCAGGCCAG
gRNA_C3-bottom	AAAACTGTGGAAACACTACATCTGCG
gRNA_H1-bottom	AAAACTTCTTCTTCTGCTCGGACTCG
gRNA_E1-bottom	AAAACTTACTGCCCTGTGGGGCAAGG
gRNA_V1-bottom	AAAACGGAGCAAACTCCCCCCACCCG
PCR_Pla-fwd	AGG AAA GAA CAT GTG AGC AAA AG
PCR_Pla-rev	CAGCGAGTCAGTGAGCGA
PCR_gRNA-fwd1	CTGTACAAAAAAGCAGGCTTTA
PCR_gRNA-rev1	AACGTAGGTCTCTACCGCTGTACAAAAAAGCAGGCTTTA
PCR_gRNA-rev2	AAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAG CTCTAAAAC
PCR_gRNA_G1	TTGCTATTTCTAGCTCTAAAACCGCCGAGGTGAAGTTCGAGGGGTGTTTCGTCCTTTCCA
PCR_gRNA_G2	TTGCTATTTCTAGCTCTAAAACAGGTGAAGTTCGAGGGCGACGGTGTTTCGTCCTTTCCA
PCR_gRNA_G3	TTGCTATTTCTAGCTCTAAAACTGGTGAACCGCATCGAGCTGGGTGTTTCGTCCTTTCCA
PCR_gRNA_G4	TTGCTATTTCTAGCTCTAAAACTCAGCTCGATGCGGTTCACCGGTGTTTCGTCCTTTCCA
PCR_gRNA_G5	TTGCTATTTCTAGCTCTAAAACTGAAGTCGATGCCCTTCAGCGGTGTTTCGTCCTTTCCA
PCR_gRNA_G6	TTGCTATTTCTAGCTCTAAAACTCCTCCTTGAAGTCGATGCCGGTGTTTCGTCCTTTCCA
PCR_gRNA_G7	TTGCTATTTCTAGCTCTAAAACGGATGTTGCCGTCCTCCTTGGGTGTTTCGTCCTTTCCA
PCR_gRNA_C2	TTGCTATTTCTAGCTCTAAAACGCTTGAGGGAGATGAGGACTGGTGTTTCGTCCTTTCCA
PCR_gRNA_C4	TTGCTATTTCTAGCTCTAAAACATGACTGTGAAGAGCTTCACGGTGTTTCGTCCTTTCCA
PCR_gRNA_E2	TTGCTATTTCTAGCTCTAAAACGAGGACAAAGTACAAACGGCGGTGTTTCGTCCTTTCCA
PCR_gRNA_E3	TTGCTATTTCTAGCTCTAAAACGAACCGGAGGACAAAGTACAGGTGTTTCGTCCTTTCCA
PCR_gRNA_H2	TTGCTATTTCTAGCTCTAAAACCACCACCAACTTCATCCACGGGTGTTTCGTCCTTTCCA
PCR_gRNA_H3	TTGCTATTTCTAGCTCTAAAACGGGCCTCACCACCAACTTCAGGTGTTTCGTCCTTTCCA
PCR_gRNA_H4	TTGCTATTTCTAGCTCTAAAACGCCCAGGGCCTCACCACCAAGGTGTTTCGTCCTTTCCA
PCR_gRNA_H5	TTGCTATTTCTAGCTCTAAAACACCTGCCCAGGGCCTCACCAGGTGTTTCGTCCTTTCCA
PCR_gRNA_H6	TTGCTATTTCTAGCTCTAAAACTGATACCAACCTGCCCAGGGGGTGTTTCGTCCTTTCCA
PCR_gRNA_H7	TTGCTATTTCTAGCTCTAAAACTAAACCTGTCTTGTAACCTTGGTGTTTCGTCCTTTCCA
PCR_gRNA_V2	TTGCTATTTCTAGCTCTAAAACGCTCTGGCTAAAGAGGGAATGGTGTTTCGTCCTTTCCA
PCR_gRNA_V3	TTGCTATTTCTAGCTCTAAAACCGGCTCTGGCTAAAGAGGGAGGTGTTTCGTCCTTTCCA
PCR_gRNA_V4	TTGCTATTTCTAGCTCTAAAACTCTGCACACCCCGGCTCTGGGGTGTTTCGTCCTTTCCA
Survey_GFP-fwd	TACGGCAAGCTGACCCTGAA
Survey_GFP-rev	GTCCATGCCGAGAGTGATCC
Survye_CLTA-fwd	GCCAGGGGCTGTTATCTTGG
Survye_CLTA-rev	ATGCACAGAAGCACAGGTTGA
Survey_EMX-fwd	CTGTGTCCTCTTCCTGCCCT
Survey_EMX-rev	CTCTCCGAGGAGAAGGCCAA
Survey_HBB-fwd	GGTAGACCACCAGCAGCCTA
Survey_HBB-rev	CAGTGCCAGAAGAGCCAAGG
Survey_VEGF-fwd	CCACACAGCTTCCCGTTCTC
Survey_VEGF-rev	GAGAGCCGTTCCCTCTTTGC

Table 3.4 (Continued). Oligonucleotides used in this study.

HTS_EXM_ON-fwd	CCTCCCCATTGGCCTGCTTC
HTS_EXM_Off1-fwd	TCGTCCTGCTCTCACTTAGAC
HTS_EXM_Off2-fwd	TTTTGTGGCTTGGCCCCAGT
HTS_EXM_Off3-fwd	TGCAGTCTCATGACTTGGCCT
HTS_EXM_Off4-fwd	TTCTGAGGGCTGCTACCTGT
HTS_VEFG_ON-fwd	ACATGAAGCAACTCCAGTCCCA
HTS_EXM_Off1-fwd	AGCAGACCCACTGAGTCAACTG
HTS_EXM_Off2-fwd	CCCGCCACAGTCGTGTCAT
HTS_EXM_Off3-fwd	CGCCCCGGTACAAGGTGA
HTS_EXM_Off4-fwd	GTACCGTACATTGTAGGATGTTT
HTS_CLTA2_ON-fwd	CCTCATCTCCCTCAAGCAGGC
HTS_CLTA2_Off1-fwd	ATTCTGCTCTTGAGGTTATTTGT
HTS_CLTA2_Off2-fwd	CACCTCTGCCTCAAGAGCAGAAAA
HTS_CLTA2_Off3-fwd	TGTGTGTGTGTGTGTGTAGGACT
HTS_EXM_ON-rev	TCATCTGTGCCCCTCCCTCC
HTS_EXM_Off-rev	CGAGAAGGAGGTGCAGGAG
HTS_EXM_Off-rev	CGGGAGCTGTTCAGAGGCTG
HTS_EXM_Off-rev	CTCACCTGGGCGAGAAAGGT
HTS_EXM_Off-rev	AAAACTCAAAGAAATGCCCAATCA
HTS_VEFG_ON-rev	AGACGCTGCTCGCTCCATTC
HTS_EXM_Off1-rev	ACAGGCATGAATCACTGCACCT
HTS_EXM_Off2-rev	GCGGCAACTTCAGACAACCGA
HTS_EXM_Off3-rev	GACCCAGGGGCACCAGTT
HTS_EXM_Off4-rev	CTGCCTTCATTGCTTAAAAGTGGAT
HTS_CLTA2_ON-rev	ACAGTTGAAGGAAGGAAACATGC
HTS_CLTA2_Off1-rev	GCTGCATTTGCCCATTTCCA
HTS_CLTA2_Off2-rev	GTTGGGGGAGGAGGAGCTTAT
HTS_CLTA2_Off3-rev	CTAAGAGCTATAAGGGCAAATGACT

Table 3.4 (Continued). Oligonucleotides used in this study. All oligonucleotides were purchased from Integrated DNA Technologies. '/5Phos/' indicates 5' phosphorylated oligonucleotides.

Construction of *Fok*I-dCas9, Cas9 Nickase and gRNA Expression Plasmids

The human codon-optimized *streptococcus pyogenes* Cas9 nuclease with NLS and 3xFLAG tag (Addgene plasmid 43861)[12] was used as the wild-type Cas9 expression plasmid. PCR (72 °C, 3 min) products of wild-type Cas9 expression plasmid as template with Cas9_Exp primers listed in **Table 3.4** were assembled with Gibson Assembly Cloning Kit (New England Biolabs) to construct Cas9 and *Fok*I-dCas9 variants. Expression plasmids encoding a single gRNA construct (gRNA G1 through G13) were cloned as previously described. Briefly, gRNA

oligonucleotides listed in **Table 3.4** containing the 20-bp protospacer target sequence were annealed and the resulting 4-bp overhangs were ligated into BsmBI-digested gRNA expression plasmid. gRNA expression plasmids encoding expression of two separate gRNA constructs from separate promoters on a single plasmid were cloned in a two-step process. First, one gRNA (gRNA E1, V1, C1, C3, H1, G1, G2 or G3) was cloned as above and used as template for PCR (72 °C, 3 min) with PCR_Pla-fwd and PCR_Pla-rev primers, 1 μl DpnI (NEB) was added, and the reaction was incubated at 37 °C for 30 min and then subjected to QIAquick PCR Purification Kit (Qiagen) for the "1st gRNA + vector DNA". PCR (72 °C, 3 min) of 100 pg of BsmBI-digested gRNA expression plasmid as template with PCR_gRNA-fwd1, PCR_gRNA-rev1, PCR_gRNA-rev2 and appropriate PCR_gRNA primer listed in **Table 3.4** was DpnI treated and purified as above for the "2nd gRNA instert DNA". ~200 ng of "1st gRNA + vector DNA" and ~200 ng of "2nd gRNA instert DNA" were blunt-end ligated in 1× T4 DNA Ligase Buffer, 1 μl of T4 DNA Ligase (400 U/μl, NEB) in a total volume of 20 μl at room temperature (~21 °C) for 15 min. For all cloning, 1 μl of ligation or assembly reaction was transformed into Mach1 chemically competent cells (Life Technologies). Protein and DNA sequences are listed in **Table 3.4**. *Fok*I-dCas9 expression plasmids will be available from Addgene.

Modification of Genomic GFP

HEK293-GFP stable cells (GenTarget) were used as a cell line constitutively expressing an Emerald GFP gene (GFP) integrated on the genome. Cells were maintained in Dulbecco's modified Eagle medium (DMEM, Life Technologies) supplemented with 10% (vol/vol) fetal bovine serum (FBS, Life Technologies) and penicillin/streptomycin (1x, Amresco). 5×10^4 HEK293-GFP cells were plated on 48-well collagen coated Biocoat plates (Becton Dickinson). One day following plating, cells at ~75% confluence were transfected with Lipofecatmine 2000 (Life Technologies) according to the manufacturer's protocol. Briefly, 1.5 μL of Lipofecatmine 2000 was used to transfect 950 ng of total plasmid (Cas9 expression plasmid plus gRNA expression plasmids). 700 ng of Cas9 expression plasmid, 125 ng of one gRNA expression plasmid and 125 ng of the paired gRNA expression plasmid with the pairs of targeted gRNAs. Separate wells were transfected with 1 μg of a near-infrared iRFP670 (Addgene plasmid 45457)[32] as a transfection control. 3.5 days following transfection, cells were trypsinized and resuspended in DMEM supplemented with 10% FBS and analyzed on a C6 flow

cytometer (Accuri) with a 488 nm laser excitation and 520 nm filter with a 20 nm band pass. For each sample, transfections and flow cytometry measurements were performed once.

T7 Endonuclease I Surveyor Assays of Genomic Modifications

HEK293-GFP stable cells were transfected with Cas9 expression and gRNA expression plasmids as described above. A single plasmid encoding two separate gRNAs was transfected. For experiments titrating the total amount of expression plasmids (Cas9 expression + gRNA expression plasmid), 700/250, 350/125, 175/62.5, 88/31 ng of Cas9 expression plasmid/ng of gRNA expression plasmid were combined with inert carrier plasmid, pUC19 (NEB), as necessary to reach a total of 950 ng transfected plasmid DNA.

Genomic DNA was isolated from cells 2 days after transfection using a genomic DNA isolation kit, DNAdvance Kit (Agencourt). Briefly, cells in a 48-well plate were incubated with 40 μL of tryspin for 5 min at 37 °C. 160 uL of DNAdvance lysis solution was added and the solution incubated for 2 hr at 55 °C and the subsequent steps in the Agencourt DNAdvance kit protocol were followed. 40 ng of isolated genomic DNA was used as template to PCR amplify the targeted genomic loci with flanking Survey primer pairs specified in the **Table 3.4**. PCR products were purified with a QIAquick PCR Purification Kit (Qiagen) and quantified with Quant-iT™ PicoGreen ® dsDNA Kit (Life Technologies). 250ng of purified PCR DNA was combined with 2 μL of NEBuffer 2 (NEB) in a total volume of 19 μL and denatured then re-annealed with thermocycling at 95 ˚C for 5 min, 95 to 85 ˚C at 2 ˚C/s; 85 to 20 ˚C at 0.2 ˚C/s. The re-annealed DNA was incubated with 1 μl of T7 Endonuclease I (10 U/μ1, NEB) at 37 °C for 15 min. 10 μL of 50% glycerol was added to the T7 Endonuclease reaction and 12 μL was analyzed on a 5% TBE 18-well Criterion PAGE gel (Bio-Rad) electrophoresed for 30 min at 150 V, then stained with 1x SYBR Gold (Life Technologies) for 30 min. Cas9-induced cleavage bands and the uncleaved band were visualized on an AlphaImager HP (Alpha Innotech) and quantified using ImageJ software.[33] The peak intensities of the cleaved bands were divided by the total intensity of all bands (uncleaved + cleaved bands) to determine the fraction cleaved which was used to estimate gene modification levels as previously described.[28] For each sample, transfections and subsequent modification measurements were performed in triplicate on different days.

High-throughput Sequencing of Genomic Modifications

HEK293-GFP stable cells were transfected with Cas9 expression and gRNA expression plasmids, 700 ng of Cas9 expression plasmid plus 250 ng of a single plasmid expression a pair of gRNAs were transfected (high levels) and for just Cas9 nuclease, 88 ng of Cas9 expression plasmid plus 31 ng of a single plasmid expression a pair of gRNAs were transfected (low levels). Genomic DNA was isolated as above and pooled from three biological replicates. 150 ng or 600 ng of pooled genomic DNA was used as template to amplify by PCR the on-target and off-target genomic sites with flanking HTS primer pairs specified in the **Table 3.4**. Relative amounts of crude PCR products were quantified by gel electrophoresis and samples treated with different gRNA pairs or Cas9 nuclease types were separately pooled in equimolar concentrations before purification with the QIAquick PCR Purification Kit (Qiagen). ~500 ng of pooled DNA was run a 5% TBE 18-well Criterion PAGE gel (BioRad) for 30 min at 200 V and DNAs of length ~125 bp to ~300 bp were isolated and purified by QIAquick PCR Purification Kit (Qiagen). Purified DNA was PCR amplified with primers containing sequencing adaptors, purified and sequenced on a MiSeq high-throughput DNA sequencer (Illumina) as described previously.[10]

Data Analysis

Illumina sequencing reads were filtered and parsed with scripts written in Unix Bash as outlined in **Table 3.4**. DNA sequences will be deposited in NCBI's Sequencing Reads Archive (SRA) and source code can be found in **Supplementary Software**. Sample sizes for sequencing experiments were maximized (within practical experimental considerations) to ensure greatest power to detect effects. Statistical analyses for Cas9-modified genomic sites were performed as previously described[34] with multiple comparison correction using the Bonferroni method.

3.5 References cited in *FokI*-dCas9 (fCas9) study

1. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343,** 80–84 (2014).
2. Shalem, O. *et al*. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **343,** 84–87 (2013).
3. Perez, E. E. *et al*. Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat. Biotechnol.* **26,** 808–816 (2008).
4. Jinek, M. *et al*. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337,** 816–821 (2012).
5. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* (2014). doi:10.1038/nbt.2808
6. Jinek, M. *et al*. RNA-programmed genome editing in human cells. *eLife* **2,** e00471–e00471 (2013).
7. Mali, P. *et al*. RNA-Guided Human Genome Engineering via Cas9. *Science* **339,** 823–826 (2013).
8. Cong, L. *et al*. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339,** 819–823 (2013).
9. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10,** 957–963 (2013).
10. Pattanayak, V. *et al*. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31,** 839–843 (2013).
11. Hsu, P. D. *et al*. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31,** 827–832 (2013).
12. Fu, Y. *et al*. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31,** 822–826 (2013).
13. Cradick, T. J., Fine, E. J., Antico, C. J. & Bao, G. CRISPR/Cas9 systems targeting -globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.* **41,** 9584–9592 (2013).
14. Cho, S. W. *et al*. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24,** 132–141 (2013).
15. Ran, F. A. *et al*. Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* **154,** 1380–1389 (2013).
16. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci.* **109,** E2579–E2586 (2012).
17. Ramirez, C. L. *et al*. Engineered zinc finger nickases induce homology-directed repair with reduced mutagenic effects. *Nucleic Acids Res.* **40,** 5560–5568 (2012).
18. Wang, J. *et al*. Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme. *Genome Res.* **22,** 1316–1326 (2012).
19. Mali, P. *et al*. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31,** 833–838 (2013).
20. Gaj, T., Gersbach, C. A. & Barbas, C. F. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31,** 397–405 (2013).

21. Vanamee, É. S., Santagata, S. & Aggarwal, A. K. FokI requires two specific DNA sites for cleavage. *J. Mol. Biol.* **309,** 69–78 (2001).
22. Maeder, M. L. *et al.* CRISPR RNA–guided activation of endogenous human genes. *Nat. Methods* **10,** 977–979 (2013).
23. Pattanayak, V., Ramirez, C. L., Joung, J. K. & Liu, D. R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nat. Methods* **8,** 765–770 (2011).
24. Guilinger, J. P. *et al.* Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. *Nat. Methods* (2014). doi:10.1038/nmeth.2845
25. Nishimasu, H. *et al.* Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell* (2014). doi:10.1016/j.cell.2014.02.001
26. Jinek, M. *et al.* Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. *Science* (2014). doi:10.1126/science.1247997
27. Schellenberger, V. *et al.* A recombinant polypeptide extends the in vivo half-life of peptides and proteins in a tunable manner. *Nat. Biotechnol.* **27,** 1186–1190 (2009).
28. Guschin, D. Y. *et al.* in *Eng. Zinc Finger Proteins* (Mackay, J. P. & Segal, D. J.) **649,** 247–256 (Humana Press, 2010).
29. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152,** 1173–1183 (2013).
30. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* **10,** 1116–1121 (2013).
31. Kim, Y., Kweon, J. & Kim, J.-S. TALENs and ZFNs are associated with different mutation signatures. *Nat. Methods* **10,** 185–185 (2013).
32. Shcherbakova, D. M. & Verkhusha, V. V. Near-infrared fluorescent proteins for multicolor in vivo imaging. *Nat. Methods* **10,** 751–754 (2013).
33. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9,** 671–675 (2012).
34. Sander, J. D. *et al.* In silico abstraction of zinc finger nuclease cleavage profiles reveals an expanded landscape of off-target sites. *Nucleic Acids Res.* **41,** e181–e181 (2013).