A TOOLKIT FOR ANALYSIS OF GENE EDITING AND OFF-TARGET EFFECTS OF ENGINEERED NUCLEASES

A Dissertation
Presented to
The Academic Faculty

by

Eli J. Fine

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
Wallace H. Coulter Department of Biomedical Engineering

Georgia Institute of Technology
May 2015

A TOOLKIT FOR ANALYSIS OF GENE EDITING AND OFF-TARGET EFFECTS OF ENGINEERED NUCLEASES

Approved by:

Professor Gang Bao, Advisor
Department of Biomedical Engineering
Georgia Institute of Technology

Professor Peng Qiu
Department of Biomedical Engineering
Georgia Institute of Technology

Professor Francesca Storici
School of Biology
Georgia Institute of Technology

Professor H. Trent Spencer
Department of Pediatrics
Emory University

Professor May Wang
Department of Biomedical Engineering
Georgia Institute of Technology

Date Approved: 03/18/2015

To my loving fiancée, Laura

whose patience and dedication has always been my foundation

ACKNOWLEDGEMENTS

First, I would like to thank the members of my thesis committee for the invaluable support and feedback over the past years—Dr. Gang Bao, Dr. Francesca Storici, Dr. May Wang, Dr. Peng Qiu, and Dr. Trent Spencer. Much of the success and potential impact of my thesis project would have been impossible without their individually unique, and invaluable perspectives and areas of expertise.

My fellow lab members provided tremendous support both in technical training and in making the lab an enjoyable and entertaining place to work. In particular, Dr. Yanni Lin and Dr. TJ Cradick were instrumental in teaching me how to work with nucleases and to conduct high-throughput experiments.

Two of the undergraduate students I mentored, Caleb Appleton and Sean Song, both worked in the lab for several years and their capable hands provided a huge boost to the pace of my research.

The nature of my thesis was highly collaborative and would not have been possible without the work of:

- Dr. Ayal Hendel, Dr. Eric Kildebeck, Joesph Clark, Gabe Washington, and Dr. Matthew Porteus at Stanford University
- Dr. Claudio Mussolino and Dr. Toni Cathomen at the University of Freiberg in Germany
- Dr. Cecilia Abarrategui-Pontes and Dr. Tuy Nguyen at the University of Nantes in France
- Dr. Zhong Chen and Dr. Steffen Meiler at Georgia Regents University

- Dr. Patrick Hsu, David Scott, and Dr. Feng Zhang at the Massachusetts Institute of Technology

Of course, none of this work would be possible with the generous funding from the National Institutes of Health and the National Science Foundation Graduate Research Fellowship program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Advances in genome engineering technology over the past four years have triggered a renaissance for gene therapy prospects. But while researchers have "cured" numerous single gene disorders in cellular models by repairing the underlying defect in the genome [64, 103, 84, 70, 49, 133, 41], clinical translation of these findings has progressed much more slowly. "Off-target" effects of gene therapy remain a major safety concern as the negative outcomes of the X-SCID clinical trials [46] continue to haunt the field. However, most researchers do not perform an analysis of the off-target effects of the nucleases they use or design their nucleases to rationally limit potential off-target effects. This is primarily attributed to the lack of user-friendly tools and methods that could be easily applied by researchers who do not have particular expertise in genome-wide bioinformatics analysis. Furthermore, efforts to optimize homology directed repair (a more versatile genome engineering approach for gene therapy applications) in primary cells have been muted due to the lack of sensitive assays which can detect incremental—but potentially clinically meaningful—improvements in efficiency at endogenous loci. Therefore, there is a clear need to develop tools that can be placed into the hands of eager experts in various genetic diseases to help them translate cellular proof-of-concept studies into more clinically meaningful results. To address this need, my research developed computational tools to accurately predict locations of off-target effects and novel DNA sequencing-based methods to sensitively measure homology directed repair.

The goal of accurately predicting nuclease off-target activity through computational approaches (as opposed to experimentally examining a nuclease in order to predict off-target activity) is not new. However, numerous previous attempts failed to uncover any off-target sites through their computational approaches [54, 129, 26, 103, 66, 134]. To date, no study

has uncovered unexpected off-target activity of ZFNs or TALENs using computational predictive modeling apart from the work described in this thesis. We created more advanced models of ZFN and TALEN binding guided by biological principles of the protein-DNA interactions which allowed us to make the first report of ZFN and TALEN off-target activity discovered through computational modeling (**Fine EJ** et al. [34]). Through this analysis, we discovered that alternative 'NK' forms of TALENs can have substantially reduced off-target activity. We then provided this model as an online webtool which allowed any researcher to easily search for potential off-target sites of their nuclease in a wide variety of different genomes.

Although our algorithms were able to accurately locate off-target activity, there was ample room for improvement in the predictive power of the algorithms (particularly with respect to false-positives). To achieve this, we investigated a large number of nucleases for off-target activity, achieving an overall ~230% increase in the number of ZFNs and TALENs with known off-target sites from 2011 levels ([79, 1], and other manuscripts in preparation). Through these analyses, we discovered that adding additional zinc finger modules does not always reduce off-target activity, that rational design of TALEN binding sites can allow for enhanced discrimination between two similar sequences, that the locations of off-target are similar between different cell types, and that the frequencies of off-target activity are very consistent within the same cell type. With a larger training set of data from which to construct more advanced algorithms, we applied machine learning techniques to develop a robust framework for ZFN and TALEN off-target prediction and achieved a ~10% improvement in the predictive power of the algorithms. The large dataset of nuclease off-target sites also allowed for the first comprehensive comparison of the genomic locations of off-target activity of engineered nucleases compared to lentiviral vectors (**Fine EJ** and Bao G, submitted). Through this analysis, we discovered that while lentiviral vectors tend to induce off-target activity within cancer-associated genes more frequently than engineered nucleases, nucleases are much more prone to have off-target activity within

exons.

Leveraging the long read lengths of SMRT sequencing, we developed a highly sensitive DNA-sequencing based approach to simultaneously measure rates of homology directed repair and non-homologous end-joining in any cell type, at any genomic locus, using any type of nuclease (Hendel A*, Kildebeck EJ*, **Fine EJ*** et al. [48]). We validated the approach against the 'gold standard' of sequencing single cell clones and found our new method to be both precise and accurate, with minimal variance introduced by the experimental processes and the bioinformatics analysis pipeline. We discovered substantial variance in the ratios of the two DNA repair pathways at different genomic loci, an effect that had previously been undiscovered because only artificial reporter constructs had been able to measure both pathways [15]. We used the system to quantify low (<1%) rates of homology directed repair occurring in primary human $CD34^+$ and embryonic stem cells. Furthermore, we demonstrated the use of our system in optimizing gene editing parameters, including plasmid masses and ratios and the length of donor homology arms. We then used the method specifically to investigate a wide variety of nucleases targeting the *HBB* gene and found that CRISPR systems achieved the most favorable homology directed repair rates. Finally, we observed rare cases of exogenous DNA (from sources such as bovine and *E. coli* genomes) integrating at the nuclease target site—a phenomenon only observable through the SMRT-based approach—that warrants caution when choosing reagents for human therapeutic purposes.

Genome engineering has enormous potential to impact human health by curing genetic disorders. However, limitations in available methods have slowed the translation of this technology. Together, this project has established a toolkit of robust and user-friendly methods that can be applied by researchers to tackle a wide variety of genetic diseases in different clinically relevant cell types to optimize on-target activity and minimize off-target effects.

CHAPTER I

INTRODUCTION

Single gene disorders have a global prevalence of 1/100 births and are estimated to account for up to 40% of pediatric hospital care in developed countries [99]. While transplants are curative for some diseases, the discovery that engineered nucleases could create targeted modifications to the genomes of human cells [90] opened the possibility of curing genetic disorders by correcting the underlying mutation in the DNA sequence of a patients cells. Recent advances have made the design of these nucleases much easier [78, 72, 77], allowing many new researchers to join the field and prompting a rapid increase in the number of genetic diseases that have been corrected in cellular experiments [64, 103, 84, 70, 49, 133, 41]. However, progress in understanding nuclease off-target effects and increasing modification efficiency in relevant cell types has been limited due to the complexity of current assays required to obtain those types of data. This thesis aims to fill this gap by developing a user friendly analysis toolkit that can be used by any lab along with basic molecular biology techniques to answer questions about off-target effects and gene modification efficiency. I hypothesized that **nuclease off-target effects can be reasonably predicted by bioinformatic modeling and that Single Molecule Real-Time (SMRT) sequencing can yield sensitive gene modification rates without the use of specialized computing platforms or complex experimental protocols**.

Recent experiments in unbiased assessments of nuclease off-target effects [88, 44] and analysis of DNA binding domains [78, 28] produced theories that formed the basis of my new bioinformatic predictive models. Improvements made to the SMRT sequencing platform now allow it to achieve sufficiently long read lengths to make measurement of gene modification possible [117].

The objective was accomplished and the central hypothesis was tested through completion of the following specific aims:

Specific Aim 1: Develop and validate an off-target prediction tool An online search interface to allow exhaustive searching of a genome for potential nuclease off-target sites was implemented. Previously discovered off-target sites were collated and ranking algorithms developed that preferentially score validated off-target sites higher than other predictions. HEK-293T cells transfected with newly developed TALENs and ZFNs targeting the beta-globin gene were analyzed at the off-target sites predicted by the tool.

Specific Aim 2: Expand off-target data set and refine off-target prediction algorithms Many samples of genomic DNA from cells treated with different ZFNs and TALENs were analyzed for off-target effects to generate a greatly expanded training set of bona fide off-target sites. Modifications to the off-target prediction algorithm parameters were evaluated for improvement through Precision-Recall analysis and several other metrics.

Specific Aim 3: Develop a method to provide molecular data on gene editing events An analysis pipeline was developed to process SMRT reads to simultaneously measure the rates of different DNA repair mechanisms by directly examining the DNA sequences. K562 cells were transfected with different types of nucleases and donor repair templates in order to optimize conditions for repairing the beta-globin gene.

This thesis addresses two issues inhibiting clinical translation of engineered nucleases: off-target effects and gene editing efficiency. The impact of this work is significant because it will optimize nuclease treatment conditions for the repair of sickle cell anemia and, through the development of easier methods, allow other laboratories to optimize nuclease-based therapies for other single gene disorders.

CHAPTER II

BACKGROUND

2.1 Significance

This research was the first to **accurately predict novel nuclease off-target sites using a purely bioinformatics-based approach**. While previous methods that used experimental characterization of the nucleases to guide a bioinformatic search of the genome have successfully found off-target sites [88, 44, 89, 37, 116, 51], previous purely bioinformatics-based approaches have failed to locate any novel valid off-target sites [54, 129, 26, 103, 66, 134]. Current methods for experimental characterization of nucleases for off-target prediction are extremely technically challenging and their use has therefore been effectively limited to the originating laboratory. By accurately predicting potential off-target sites purely *in silico*, the ability to investigate off-target activity is now extended to every laboratory.

Limitations in next generation sequencing technologies and enzymatic assays have thus far prevented direct molecular analysis of endogenous gene editing events in a high-throughput manner. Although a fluorescent reporter system has been useful in examining general aspects of the gene editing process [15], that system is unable to provide data for editing endogenous gene targets and—while theoretically possible—has yet to be implemented in clinically relevant cell lines due to the difficulty involved. The method developed in this thesis using SMRT sequencing was the first to provide high-throughput molecular data that **can be used to optimize gene editing conditions using the exact reagents that would be translated clinically**.

At the onset of this work, only two TALENs [116, 51] (See section 2.3.2) and three ZFNs [88, 44] (See section 2.3.1) had been successfully interrogated for off-target effects,

uncovering a total of three and ~120 off-target sites for TALENs and ZFNs respectively. In order to more thoroughly understand nuclease off-target effects—and how to diminish them for clinical applications—many more examples, especially for TALENs, needed to be found in order to search for trends and build meaningful hypotheses. Multiple *in silico* prediction attempts were made to discover additional TALEN and ZFN off-target sites [54, 129, 26, 103, 66, 134], but none were successful—possibly due to their use of only simplistic bioinformatics models to predict potential off-target sites. This thesis **dramatically increased the total number of ZFNs and TALENs which have known off-target sites** which provided a data set to better train off-target predictive models. Better understanding of off-target effects will allow for selection of nucleases in pre-clinical stages with minimal chances of inducing adverse events in patients.

As a whole, this work establish methods to easily interrogate nucleases for off-target effects and to optimize gene editing conditions. These methods were used in this thesis specifically to optimize nucleases that target the hemoglobin beta (*HBB*) gene for correction of sickle cell anemia. While many high-throughput sequencing and bioinformatics analysis packages require specialized software or hardware systems that only highly trained researchers can operate, this thesis developed **an analysis toolkit consisting entirely of user-friendly components that can be run on a standard Windows® machine**. Insights gained through the use of these new methods will lead to improved gene therapy applications for single gene disorders.

2.2 Sickle Cell Anemia

Sickle cell anemia was the first disease to have its molecular basis uncovered; an A→T transversion in the sixth codon of the *HBB* gene creates a missense mutation substituting valine for glutamate[110]. This mutation causes polymerization of the hemoglobin molecules under low oxygen tension which causes the red blood cells to adopt a “sickle” conformation which can lead to vasooclusion. Acute and chronic pain, increased risk of

stroke, and a drastically shorter life expectancy are hallmarks of the illness. Bone marrow stem cell transplants are currently the only cure, but suitable donors are only available for ~10% of patients and complications (fatal in some cases) arise during the transplant process. Because of the simple molecular basis of the disease and the knowledge that allogeneic transplants are curative, sickle cell anemia has been a major focus of gene therapy attempts [103]; by correcting the patient's own hematopoietic stem cells *ex vivo* and performing an autologous transplant, the limitations of current allogeneic transplants could be overcome.

2.3 Engineered Nucleases

Restriction enzymes are common tools in molecular biology due to their ability to recognize and cleave a specific sequence of DNA. Since DNA is a nucleic acid, these enzymes fall into the broad class of "nucleases". While restriction enzymes commonly recognize a short 6 bp sequence, engineered nucleases expand the concept by targeting an 18-50 bp sequence. The longer recognition sequence ensures that, even in large genomes, there is only a single site that perfectly matches the target sequence. There are a few naturally occurring restriction enzymes that have long ($\geq$ 18 bp) recognition sequences, but they generally do not target near any region of interest in the human genome. The ability to re-engineer the sequence that the nuclease targets so that it will recognize a location near a gene of interest has driven a revolution in genetic studies and formed the basis of the new field of "genome engineering".

2.3.1 Zinc Finger Nucleases

Nearly two decades ago, the discovery was made that a DNA binding domain from a "zinc finger" transcription factor could be fused to the cleavage domain from a restriction enzyme to create a nuclease—which later became known as a zinc finger nuclease (ZFN)—with a novel target sequence [60]. Three aspects of this work were critical to the subsequent proliferation of ZFNs. Firstly, zinc fingers are a modular family of naturally occurring

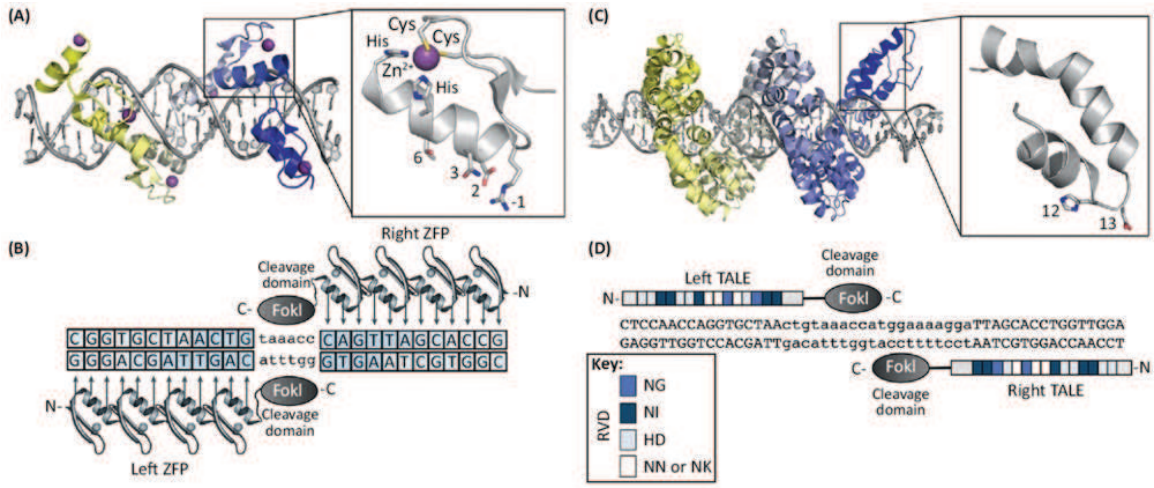

Figure 2-1: Protein-DNA interactions of ZFNs (a, b) and TALENs (c, d). Adapted from Gaj et al.[38]

transcription factors with each subunit—the "finger"—specifying three nucleotides (Figure 2-1a) and these interchangeable subunits are able to be linked together to recognize longer sequences.

2.3.2 Transcription Activator-Like Effector Nucleases

In 2009, it was discovered that the transcription activator-like effector (TALE) family of proteins in pathogenic plant bacteria had a surprisingly simple pattern in its DNA binding domains to determine what sequence it targeted [78]. The proteins consist of tandem repeats of 34 amino acid blocks where all amino acids except the 12^{th} and 13^{th} are constant in each repeat. The 12^{th} and 13^{th} residues, termed repeat variable di-residues (RVDs), specify the single nucleotide targeted by that repeat unit (Figure 2-1c). The most common RVDs create a very simple system: Asparagine-Isoleucine (NI) to target adenosine, His-Asp (HD) to target cytidine, Asn-Asn (NN) to target guanosine, and Asn-Gly (NG) to target thymidine. Soon after the binding pattern was discovered, designed blocks of TALE repeats were attached to the FokI cleavage domain to create a new class of engineered nucleases that became known as TALE nucleases (TALENs) [77] (Figure 2-1d). TALENs have been used to edit the genomes of rats [116], grasshoppers [129], human embryonic and induced

pluripotent stem cells [51], and many other animals and cell lines.

Compared to ZFNs, TALENs have certain advantages and disadvantages. Due to the simple RVD code, it is easy to design TALENs to target a new DNA sequence, and the majority of TALEN pairs meeting minimal design criteria are able to cut the intended genomic target [97]. While ZFNs can theoretically target any sequence, using triplets with a motif other than GNN (where "N" stands for any nucleotide) greatly reduce the chances that the ZFNs will be able to function correctly in cells. Although studies of TALEN off-target activity have so far been limited, initial findings suggest that they are more specific than ZFNs in preferentially cutting their intended target sequence as opposed to other locations in the genome [51, 116]. The major disadvantage of TALENs is their size—they are ~4 times larger than ZFNs—which makes delivery into primary cells more difficult because packaging them into viral delivery vectors has proved troublesome [52].

2.3.3 CRISPR/Cas9 Nucleases

In 2012, it was discovered that the clustered regularly interspaced short palindromic repeat (CRISPR) bacterial defense system—that operates by cleaving invading DNA sequences—could be co-opted to easily direct cleavage to any DNA sequence of interest [56]. The combination of a short RNA sequence and the Cas9 protein is all that is required to cause cleavage at the intended genomic locus. The first 20 bp of the single guide RNA (sgRNA) strand can be customized to any desired sequence (Figure 2-2) to direct Cas9 to cleave at the complementary matching site in the genome.

Although the sgRNA fragments are easy to design and construct, they do not confer a high level of specificity. Many off-target sites with relatively high nuclease-induced mutation rates were easily found for CRISPR/Cas9 systems using very simple search methods [53, 21] but, counterintuitively, substantially less cytotoxicity is observed than with TALENs or ZFNs. The Cas9 from *Streptococus pyogenes*, the version of Cas9 used in nearly all studies to date, is only one of many different proteins that can be paired with sgRNAs

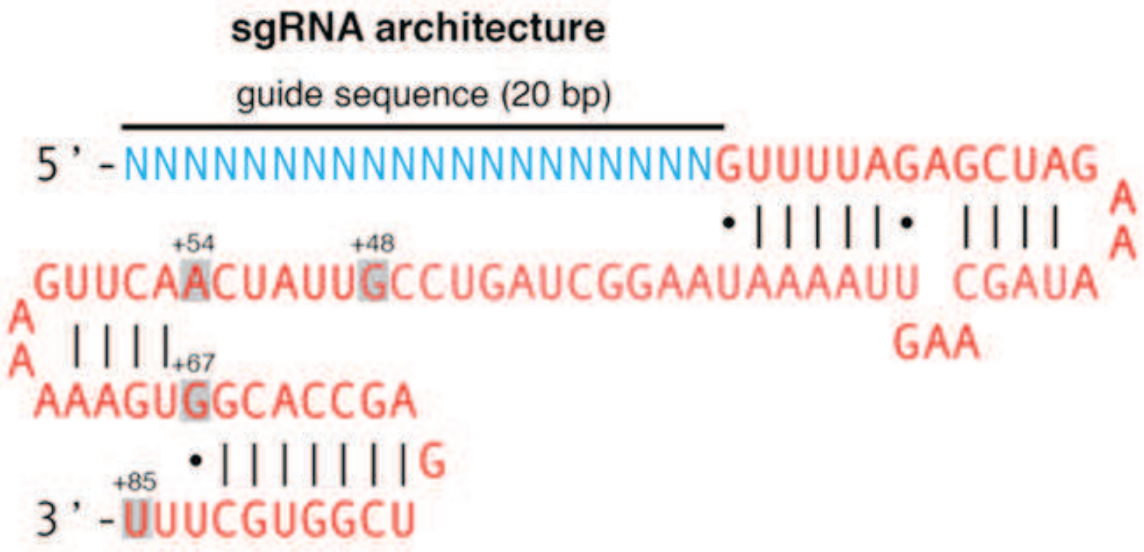

Figure 2-2: sgRNA architecture for the CRISPR/Cas9 nuclease system. The sgRNA consists of a 20-nt guide sequence (blue) and scaffold (red). The scaffold was truncated at various positions as indicated and the +85 bp scaffold was found to be optimal for achieving NHEJ. Taken from Hsu, ...**Fine** et al., *Nature Biotechnology* 2013 [53].

to act as an engineered nuclease and as of yet Cas9 itself has not been studied for ways to improve specificity. Because CRISPR systems are so efficient at inducing cleavage and so simple to design, it is expected that attempts to re-engineer Cas9 to be more specific or to adapt a different Cas protein will be forthcoming shortly in an effort to determine if CRISPR could be a viable gene therapy option.

2.3.4 Paired CRISPR/Cas9 Nickases

A recent development to reduce CRISPR off-target effects are paired CRISPR nickases [95]. While Cas9 normally causes a double strand break in the DNA, these nickases have one catalytic domain inactivated (commonly through the D10A mutation but also through H840A) which causes them to only cleave one of the phospho-diester bonds of the DNA backbone, thereby creating a 'nick' rather than a double strand break. Because nicks are repaired in an error-free manner much more frequently than double strand breaks, individual CRISPR guide strands causing nicking at off-target sites in the genome is of less

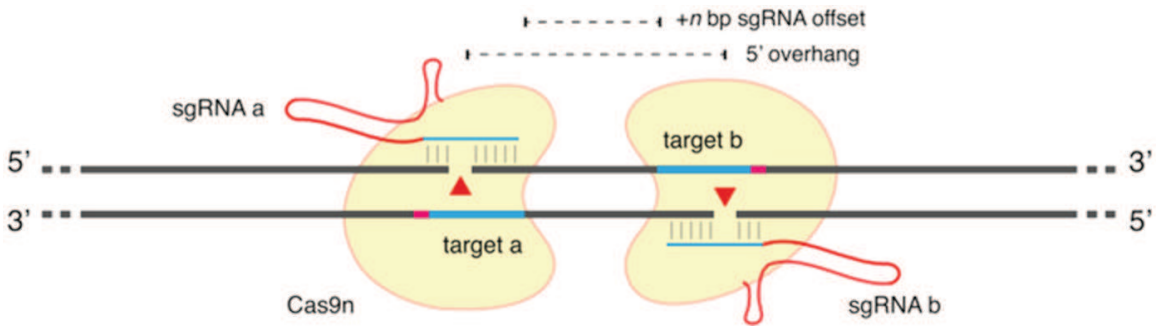

Figure 2-3: Paired CRISPR Nickases. A schematic detailing the double offset nicks generated by a pair of CRISPR nickases. Taken from Ran et al. [95].

concern—reductions in mutation levels at off-target sites exceeded 1400 fold in some cases for nickases compared to nucleases [95]. In order to promote on-target activity, two guide strands are deployed that cause offset nicks in nearby sequences in order to create an effect similar to a double strand break (Figure 2-3). Somewhat surprisingly, studies have found that these paired nickases can cause on-target mutations at rates even higher than single CRISPR nucleases [95].

2.4 DNA Repair Pathways

Of all types of DNA damage, breaks in the phosphodiester backbones of both strands of the double helix in close proximity to each other—termed a double strand break (DSB)—are the most dangerous to cells [55]. Inattention to these breaks will rapidly lead to genome instability which can cause programmed cell death or transformation into a cancerous phenotype. To prevent this, cells have several different mechanisms to repair DSBs.

2.4.1 Non-Homologous End-Joining

The non-homologous end-joining (NHEJ) repair pathway is the most common DSB repair pathway in higher eukaryotes [102]. In the case of a clean break—such as the type made by engineered nucleases—where no DNA bases have been lost and compatible overhangs are present, NHEJ can often result in an error free re-ligation of the two ends of DNA. However, if the two strands have begun to diffuse away from each other, end-processing

enzymes are not able to bind quickly enough, or other circumstances have interfered, NHEJ can result in small (typically < 50 bp) insertions and deletions (indels) at the site of the break in the process of re-joining the ends together. In rare cases, exacerbated if there are a large number of DSBs being generated, NHEJ can re-join two pieces of DNA that were not originally together, causing chromosomal translocations, gross deletions, or inversions.

2.4.2 Homology Directed Repair

The homology directed repair (HDR) pathway is a slower process than NHEJ, but typically results in "error-free" repair whereas NHEJ is "error-prone". Instead of simply re-ligating the broken DNA ends together, complexes form in HDR to find a section of DNA with high homology to the regions surrounding the DSB. Typically, the homologous template is the sister chromatid and therefore HDR is restricted to the S and G_2 phases of the cell cycle (after chromosome replication has taken place), whereas NHEJ can occur at any stage of the cell cycle. If a homolgous template is found, the broken strands are resected a short distance, the broken strands then invade the region of homologous DNA, and DNA polymerases use the homologous DNA as a template to fill in the gap in the original strands where the DSB occurred.

2.5 Gene Therapy using Engineered Nucleases

While viral gene therapy is useful in many applications, there are several shortcomings. Viruses cannot knock-out a specified gene, alter the status of the diseased allele, or regulate the inserted healthy gene in the same manner as the full endogenous promoter. Additionally, viruses integrate at random into the genomes of dividing cells, which can cause the cells to become cancerous if a sensitive area of the genome is altered. By directly editing the genetic defect at the *in situ* location in the genome, engineered nucleases offer the promise of curing genetic disorders at the most fundamental level: repairing the mutation causing the disease.

Some potential gene therapy treatments using engineered nucleases involve the NHEJ

repair pathway. The NHEJ pathway creates small indels during repair of the DSB. If the DSB induced by the nucleases occurs in the coding sequence of a gene and the length of the resulting indel is not a multiple of three (roughly 66% of the time), the reading frame of the downstream sequence will be shifted. In most cases, this frameshift will cause a premature stop codon which effectively knocks-out the gene; this strategy is used to treat HIV/AIDS by knocking-out the CCR5 receptor that HIV requires for cell entry [89]. Alternatively, if the original mutation that causes the disease is a short deletion or insertion that creates a frameshift, the indels resulting from NHEJ repair can restore the correct reading frame; this strategy is being developed to treat Duchenne muscular dystrophy by restoring the reading frame of the dystrophin gene to allow cells to properly express the full length protein [85].

However, since the NHEJ pathway does not allow the exact nature of the DNA modification to be precisely specified, most potential gene therapy treatments utilize the HDR repair pathway. While HDR normally uses the sister chromatid as a template to repair the DSB by restoring the original sequence, if another DNA template with homology to the region surrounding the DSB is introduced, the cell can use that as a template instead. For precise genome editing, a plasmid can be constructed containing homologous sequences stretching ~400-800 bp upstream and downstream of the site of the DSB—typically called "arms of homology"—and the desired DNA modifications placed in the region immediately adjacent (within ~200 bp) to the DSB between the two arms of homology (Figure 2-4). This plasmid can then act as a template for HDR repair, "donating" the desired DNA sequence into the genome at the location of the DSB to replace the endogenous sequence.

HDR-based engineered nuclease gene therapies have been validated in cell lines for several diseases. The single point mutations causing sickle cell anemia [103], epidermolysis bullosa [84], α_1-antitrypsin deficiency [133], and inactive versions of the p53 tumor suppressor protein [49] as well as the three bp deletion causing most forms of cystic fibrosis (Δ508) [64] have all been repaired using ZFNs or TALENs. For diseases where a range of mutations throughout the gene exists in the patient population, a healthy version of the

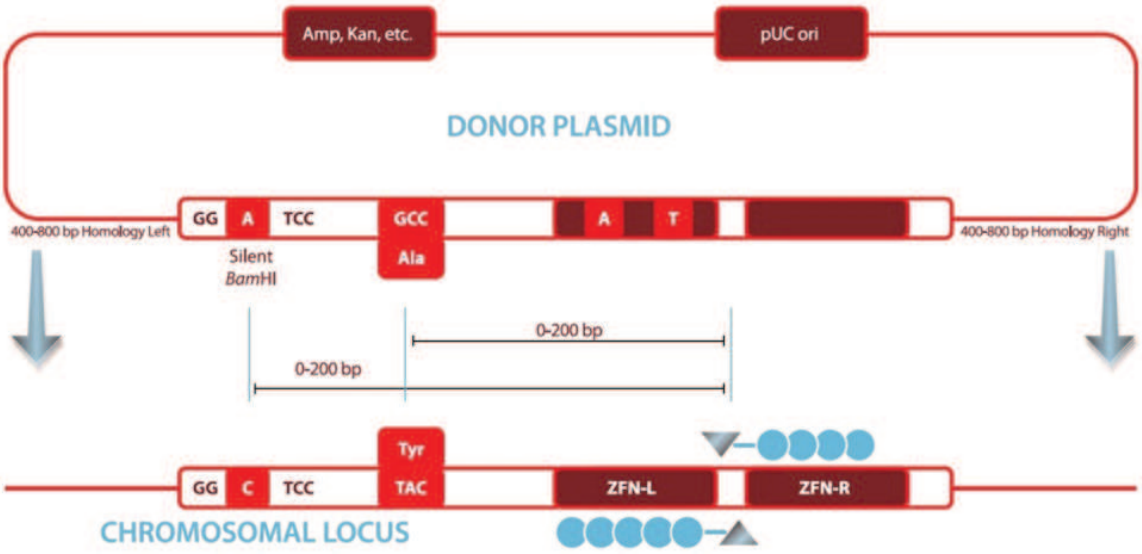

Figure 2-4: Donor plasmid repair template for precise genome editing using HDR. Adapted from Sigma Aldrich *Biowire* Fall 2010.

full length cDNA for the gene can be placed between the arms of homology and targeted to the start codon of the gene; this approach has been used to insert healthy copies of the IL2Rγ gene to correct severe combined immunodeficiency (SCID) [70]. The advantage of the cDNA approach using engineered nucleases over viruses is that since the cDNA is targeted to the endogenous location of the gene instead of randomly integrated, it will be under the control of the entire endogenous promoter, including any distal elements, which will encourage a more natural gene expression pattern.

2.6 Experimental-based Off-Target Prediction Methods[1]

Most previous studies of nuclease off-target activity have used experimental characterization of the specific nuclease in order to predict potential off-target sites. Although experimental-based prediction methods are generally quite effective (nearly all publications employing them have located at least one bona fide nuclease off-target site), they are very

[1]Modified from: Fine et al. (in press). Strategies to Determine Off-Target Effects of Engineered Nucleases. Springer.

technically challenging, costly, and time intensive. Because of the difficulty of implementing these techniques, most have never been replicated outside of the original laboratories in which they were developed.

2.6.1 SELEX

Systematic Evolution of Ligands by Exponential Enrichment (SELEX) is an established technique for determining nucleic acid sequences that have high affinity for target molecules. This approach has been used to ascertain nuclease specificity in works from the laboratories of Sangamo Biosciences. SELEX has been used to determine the binding preference of ZFNs [50, 89] and TALENs [77, 116, 51] and subsequently to guide bioinformatics searches for potential genomic off-target sites. The general approach is to (i) genetically tag the nuclease with an affinity molecule, such as hemagglutinin (HA), (ii) express the nuclease in vitro, (iii) incubate it with a semi-randomized library of oligonucleotides (usually biased towards the expected binding site of the nuclease), (iv) capture the nuclease protein using antibodies, and (v) PCR the bound DNA fragments to amplify them. Steps (iii)-(v) are then repeated for multiple rounds of enrichment with the PCR products from step (v) replacing the initial semi-randomized library in step (iii). After the desired number of selection rounds, the PCR amplicons are sequenced to determine the identity of the selected DNA. SELEX typically yields 20-50 unique sequences that were bound by the nucleases or DNA binding domains [77]; if too few or too many unique sequences are found, amplicons from prior rounds can be sequenced or additional rounds of selection can be carried out. These sequences can be compiled to form position weight matrices (PWMs) indicating the binding preferences of the nuclease at each position (Figure 2-5).

Once PWMs for each nuclease half-site have been established, the genome can be searched bioinformatically and scores calculated for each position. Each potential bi-partite nuclease off-target siteone half-site, separated by an appropriate length spacer sequence, and the other half-sitecan be given a score by calculating the product of the values of the

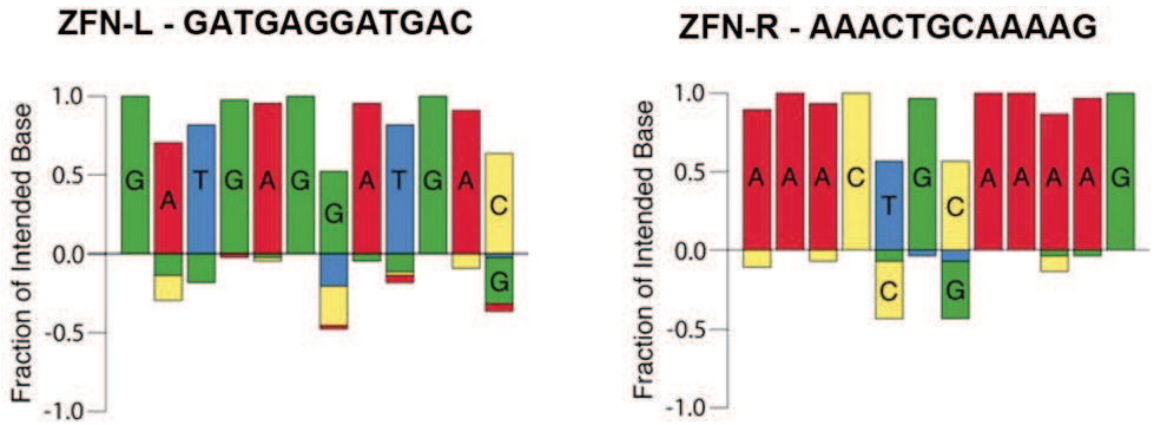

Figure 2-5: SELEX characterization of zinc finger binding domains. The positive y-axis represents a preference for the intended base at that binding position. The negative y-axis represents preferences for alternative bases at that position. Adapted from Perez et al. [89]

PWM for each nucleotide comprising the potential off-target site (Note: the authors from Sangamo did not publish their formula for generating a score from the PWMs, but calculating the product of all positions provides a close approximation [20]). All sites in the genome can then be ranked and a subset can be chosen for further investigation.

This technique has faced several criticisms, but has proven remarkably robust at finding off-target sites for both ZFNs and TALENs. Drawbacks of this technique include the fact that it only provides information about the binding preferences of each nuclease half-site, therefore ignoring interactions between the two half-sites required for nuclease cleavage. Another limitation is that it is performed completely *in vitro*, therefore ignoring changes that may occur to the protein in the cellular environment as well as ignoring any factors that may affect the genomic DNA at the potential off-target sites in the cells, such as chromatin structure, accessibility, and methylation status. Finally, because the starting oligonucleotide library is only semi-randomized, this method is biased towards finding sites with relatively high homology to the intended nuclease target. Nevertheless, this has been the most published experimental characterization technique for successfully finding nuclease off-target sites and is one of only two experimental-based prediction technique thus far published that

has found bona fide TALEN off-target sites [116, 51].

2.6.2 Bacterial One-Hybrid

The bacterial one-hybrid (B1H) approach is similar to SELEX in that it analyzes the binding preferences of nuclease monomers. To begin, a library of reporter plasmids is generated with a semi-randomized (biased towards the intended nuclease target) region upstream of the reported gene. This library is co-transformed into bacteria along with a plasmid encoding the nuclease DNA binding domain fused to a transcriptional activator [44]. *E. coli* colonies expressing the reporter gene, due to the nuclease DNA binding domain having sufficient affinity for the sequence in the plasmid to be able to activate the gene, are selected and the plasmid is sequenced to determine the sequence of the semi-randomized binding site. All sequences recovered from the *E. coli* colonies are compiled to create a PWM that can be used to screen the genome for potential off-target sites in the same way as the SELEX method. Using B1H for nuclease off-target prediction was developed by Scot Wolfe's laboratory which has been the only group to employ this method so far, and only for predicting off-target activity of ZFNs [44]. This approach faces many of the same criticisms as SELEX relating to the analysis of single monomers, but it has the advantage of being performed in a cellular (albeit bacterial and not eukaryotic) environment which may better model protein-DNA interactions than a completely *in vitro* analysis.

2.6.3 *In vitro* cleavage

Unlike the previous two prediction methods that separately characterize the DNA binding abilities of each monomer, *in vitro* cleavage assays explicitly investigate which DNA sequences in a random pool can be cut by a nuclease [88]. This approach has been applied to ZFNs [88], TALENs [43], and CRISPRs [87], but has only been in studies published by David Lius laboratory. In this approach, a semi-randomized oligonucleotide library is synthesized that consists of the full nuclease recognition site. For paired nucleases (such

as ZFNs, TALENs, or paired CRISPR nickases), the half-sites of both monomers are included, separated by appropriate length spacer sequences. The nuclease is then expressed *in vitro*, and incubated with the oligonucleotide library. Several enzymatic and gel isolation steps allow separation of sequences cleaved by the nuclease. Libraries are deep sequenced before and after nuclease incubation to identity the sequences cleaved by the nuclease. A bioinformatics search is then performed to determine if any of the sites that were cleaved *in vitro* also exist in the genome of interest. These sites can then be assayed for off-target activity.

There are several advantages and limitations of this technique. By examining nuclease cleavage instead of merely binding, insights were gained in the original study [88] that led to the hypothesis of an "energy compensation" model of dimeric ZFN interactions where larger numbers of mismatches in one half-site can be compensated by few or no mismatches in the other half-site. However, as this technique is performed entirely *in vitro*, effects of the cellular environment on the nuclease and genomic DNA are not accounted for. Furthermore, since the oligonucleotide library is semi-randomized, the analysis is biased towards finding sites with higher levels of homology to the intended nuclease target site.

An extension to this approach was recently developed to make better use of the large amount of data generated [101]. The original applications of this method searched through genomes to find *exact* matches to sequences that had been cleaved [88, 87], but those sequences that matched the genome were only a small fraction of the total sequences that the nuclease was shown to be able to cleave. By applying a Bayesian machine learning algorithm to the full list of sequences that the *CCR5* and *VEGF* ZFNs were confirmed to cleave in the original study, classifiers were developed for each nuclease that could generate a score for any given sequence predicting the likelihood of cleavage. The full genome was then screened bioinformatically—in a similar manner to the PWM screening in the SELEX and B1H methods—for sites that scored highly by the classifier. The analysis of off-target activity at the sites predicted by this method demonstrated that it could locate bona fide

off-target sites with relatively low sequence homology and sites that have low activity. Impressively, this method also appears to have a fairly low false discovery rate that resulted in the analysis locating a large number of new bona fide off-target sites for both the *CCR5* and *VEGF* ZFNs [101] and two TALENs targeting *CCR5* and *ATM* [43]. Unfortunately, the incredibly difficult and time consuming nature of this approach that must be performed for each nuclease to be studied—conducting the *in vitro* cleavage experiments and then subsequently building a Bayesian classifier using machine learning—will likely severely limit the number of nucleases that are studied using this method.

2.6.4 IDLV LAM-PCR

Integrase-Deficient Lentiviral Vector Linear Amplification Mediated Polymerase Chain Reaction (IDLV LAM-PCR) is one of the two off-target prediction methods that is performed in the full intracellular environment. This approach was developed by Christof von Kalle's laboratory [37] and thus far, they have the only publications using this method. In this approach, the cells are transduced by an IDLV encoding a selectable marker, such as green fluorescent protein (GFP). Because the virus is integrase deficient, its ability to integrate into the genome is severely limited. Therefore, cultured dividing cells would rapidly dilute the IDLV gene sequence after several weeks, as it is not replicated during cell division. If nucleases are added, the resulting DSB can lead to a much higher efficiency of IDLV integration into the cellular genome. In this case then, after culturing dividing cells for several weeks, a larger fraction of nuclease treated cells express the selectable marker compared to control cells. These cells are then selected and viral integration site analysis is performed. Briefly, their approach was to use LAM-PCR on the genomic DNA using primers that bind to the long terminal repeat (LTR) regions of the IDLV. The amplicons resulting from LAM-PCR include a portion of the genomic sequence flanking the LTR, and therefore the location of the integration site of the IDLV can be deduced by high-throughput sequencing of the amplicons. Clustered integration site (CLIS) analysis is performed to

filter out much of the random integration by imposing a criteria that two independent integration sites must be observed within 500 bp of each other in order for that locus to be considered a potential site of nuclease activity (although fragile sites in the genome are also prone to clustered integration sites). The next step is to search the sequence space surrounding the CLIS locus for a region with homology to the intended nuclease target that might be a location of bona fide nuclease cleavage activity; random sequence space has an expected level of ~45% homology to the nuclease target [37], so sequences with > 60% homology to the nuclease target are likely candidates. The predicted off-target sites can then be interrogated in cells treated with nucleases without the IDLV.

The major limitation of this approach is its lack of sensitivity. This drawback is an inherent part of the method since the underlying process relies on the relatively rare event of the IDLV being captured during the repair of a DSB. Consequently, many off-target sites, especially those with lower activity, are overlooked; the IDLV LAM-PCR analysis of the hetero-obligate CCR5 ZFNs [37] predicted only four of the 38 known off-target sites [101, 34]. This method has thus far only been successfully used to locate ZFN off-target sites; an attempt was made using it to locate TALEN off-target sites, but very few cases of CLIS were observed and no attempt was made to validate off-target activity at the predicted loci [84]. However, this approach does provide an unbiased survey of any highly active off-target sites in the full intracellular environment with the cells genome in its native structure. Because it is not biased by oligonucleotide selection libraries, this method was able to locate bona fide off-target sites with extremely low (66%) homology to the intended nuclease target [37]. As this method lacks sensitivity, it may not be optimal for testing nucleases for potential use as human therapeutics—or other applications where even rare off-target cleavage could cause adverse events—but it remains a highly useful research tool because its unbiased nature allows it to uncover sites that might not fit standard models of nuclease specificity used to guide the generation of oligonucleotide libraries or *in silico* searches.

2.6.5 ChIP-Seq

Chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq) is a well-established method for determining what sequences in a genome a certain protein binds. ChIP-Seq involves genetically tagging the nuclease with an affinity epitope (commonly hemagglutinin) and catalytically inactivating the nuclease (so that it binds but does not cleave DNA), expressing the modified nuclease in cells, cross-linking the protein and DNA together, shearing the genomic DNA into smaller fragments, purifying the nuclease (and the DNA fragments to which it is cross-linked) using antibodies (immunoprecipitation), sequencing the DNA fragments bound to the nuclease, and then mapping those sequences to the genome. In early 2013 it was noted how well the idea of ChIP-Seq seemed to be suited to facilitating an unbiased genome-wide survey of nuclease off-target activity in living cells [104], but the results of recent studies thus far have not been as promising as initially hoped.

Dimeric nucleasessuch as ZFNs, TALENs, RFNs, and paired Cas9 nickasespresent special challenges for ChIP-Seq. As noted in an unsuccessful attempt in late 2013 to use ChIP-Seq to identify off-target sites of the *CCR5* ZFNs: "thousands of high affinity monomeric target sites may exist in the genome, however a monomer is not sufficient to generate a lesion. Alternatively, dimeric ZFN sites that are bound weakly by both monomers may be sufficient to cleave DNA at a low frequency but may not bind stably enough to be detected reliably via ChIP" [101].

Because Cas9 can act as a monomeric nuclease, three groups attempted to use ChIP-Seq to locate CRISPR/Cas9 off-target sites in early 2014. While Cas9 binding was observed at many (up to thousands, depending on the gRNA used) sites throughout the genome other than the intended target site, off-target nuclease activity (NHEJ) was only found at a tiny fraction of the sites interrogated by two of the groups [130, 82], indicating that this method has a very high false positive rate as a method of discovery of off-target nuclease activity.

In summary, ChIP-Seq has not yet become a reliable method to predict off-target nuclease activity. Although it identifies the on-target site with reasonably accuracy , ChIP-Seq

may be fundamentally ill-suited to identifying off-target sites of nuclease activity. Emerging evidence into the mechanism of Cas9 shows that it uses a multi-step approach to DNA cleavage [130, 111]: binding to many points along a chromosome as it pauses in the search for a target matching the gRNA but only cleaving when a good match is found. ChIP-Seq detects all binding events which leads to very high false positive prediction rates of nuclease activity. Without a method to discriminate between binding and cleavage events, ChIP-Seq may be incapable of detecting sites of low frequency off-target cleavage because the ChIP-Seq signal at those sites may not rise above the background noise.

2.7 SMRT Sequencing

Single molecule real-time (SMRT) sequencing is a recently developed "third-generation" sequencing platform with different strengths and weaknesses compared to the "next-generation" sequencing (NGS) platforms such as Illumina and Ion Torrent. Chief among these tradeoffs is throughput versus read length. While NGS platforms can sequence millions of pieces of DNA in parallel, they cannot provide reliable information about sequences longer than ~400 bp [93]. In contrast, SMRT sequencing has lower throughput—providing information about tens of thousands of pieces of DNA instead of millions—but can provide highly accurate information about much longer pieces of DNA; when sequencing 900 bp samples, nearly half of the reads have an average QV score of > 40, meaning that the error rate is less than one in ten thousand bases. Examining longer pieces of DNA (> 600 bp) is critical to gaining a full understanding of the different types of DNA repair events occurring after a nuclease induced DSB. SMRT sequencing achieves this high accuracy by having the DNA polymerase iteratively loop around the amplicon multiple times (Figure 2-6).

Beyond the technical differences between SMRT and NGS sequencing, there are several aspects that make SMRT a more favorable choice for a system designed to be user-friendly and accessible to any lab. The average cost per sequenced DNA base is substantially cheaper for NGS systems than for SMRT [93], however the cost *per sequencing*

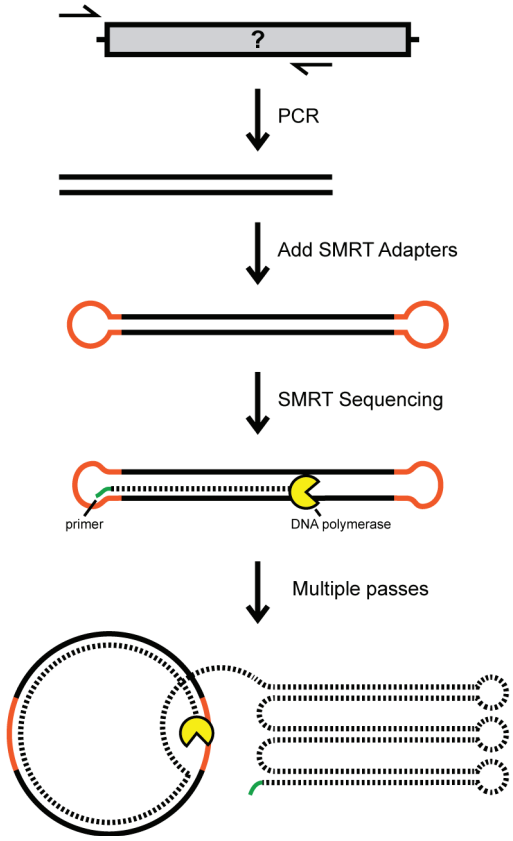

Figure 2-6: SMRT sequencing process. Adapted from Hendel*, Kildebeck*, **Fine*** et al. [48]

reaction is much lower for SMRT. This is because each NGS run provides a much larger amount of data than each SMRT run. For a core facility analyzing dozens of different nucleases on a regular basis, NGS platforms would hold a strong price advantage. But for a single lab that wants to analyze just a few experiments, an NGS run would cost much more than a SMRT run and would generate an amount of sequencing data far in excess of what would be needed to answer their questions. Additionally, there are various restrictions and requirements for the sizes of amplicons and adapter sequences that must be used for NGS platforms that can make the process difficult for novice users, whereas SMRT only requires that within each sequencing run the amplicons are of roughly the same length (±100 bp). For this proposal, the only effect of that SMRT requirement is that off-target and gene editing investigations are sequenced separately, which makes SMRT much more user-friendly for a lab without extensive sequencing experience.

2.8 Machine Learning

Machine learning involves looking for predictive patterns in large datasets. Each item in the dataset is accompanied by a list of attributes—also known as 'features'—which can be readily observed which the machine uses an inputs into its algorithm in order to predict an unknown attribute. How each feature is weighted and other features of the algorithm are customized to the question at hand by 'training' the algorithm on a dataset where the desired attribute is known. In this manner, the machine can select parameters which result in a close fit between the predictions of the algorithm and the known results.

2.8.1 Feature Extraction

Most machine learning algorithms work best when the 'features' of the dataset are numerical values, however many experimental attributes do not naturally exist in that state. While percentages, concentrations, and temperatures can easily be formatted into machine learning features, other types of data (such as diverse lengths of DNA sequences representing potential nuclease off-target sites) present more of a challenge. The process of creating

numerical representations of complex data that can be used as machine learning input is known as 'feature extraction' and is as much an art as a science, analogous to determining how best to quantify microscopy images.

2.8.2 Learning Algorithms

Most machine learning algorithms fall into the following three categories:

Classification Predicting an item's category based on other available attributes

Regression Predicting a quantitative value for an item based on other available attributes

Clustering Determining groups of items which are similar to each other based on available attributes

For the purposes of predicting off-target activity, clustering algorithms are not as relevant because it is known (for the sites in the algorithm development dataset) whether a site is or is not a bona fide off-target site. Additionally, while using regression algorithms to accurately predict exact levels of off-target activity would be extremely useful, with current available datasets it is not yet feasible (for reasons discussed further in Section 5.3.3).

Understanding how classification algorithms work can be facilitated using the following model. Consider a case where there are only two 'features' of a datapoint; all datapoints can then be plotted on a simple plot with one feature on the x and y axes respectively. The goal of the classification algorithm is to find a line which can divide the data such that the two different classes fall on opposite sides of the line while the relative emphasis placed on false positives and false negatives is given through tuning the algorithm parameters (Figure 2-7). As the model becomes more complex, the dataset is plotted in multi-dimensional space and the dividing 'line' can become a curved hyper-surface.

2.8.3 Performance Metrics[2]

[2]Modified from: `http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/`

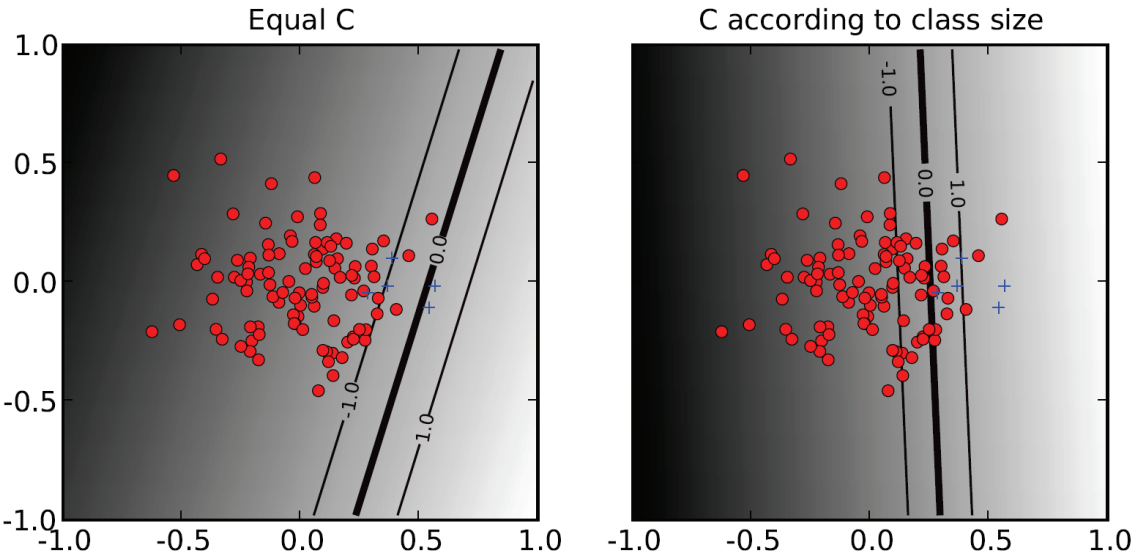

Figure 2-7: Classification Algorithm Model. Two different dataset classes (red circles and blue pluses) are plotted according to their two features (x- and y-axes). A line separating the two classes is determined by the classification algorithm. Different algorithm parameters can result in lower false positive (left) or false negative (right) rates. Figure modified from Ben-Hur and Weston [6]

.

In a binary classification problem, performance metrics are based on the values of the 'confusion matrix' (Table 2-1). Three commonly used attributes of the confusion matrix are:

'True Positive Rate' or 'Recall': When it's actually yes, how often does it predict yes?

- True Positives divided by the number of positives in the dataset ('Actual Yes')
  - Method 1: $\frac{9}{10} = 0.9$
  - Method 2: $\frac{9}{10} = 0.9$

False Positive Rate: When it's actually no, how often does it predict yes?

- False Positives divided by the number of negatives in the dataset ('Actual No')
  - Method 1: $\frac{1}{99990} = \quad 1 * 10^{-5}$
  - Method 2: $\frac{191}{99990} = 191 * 10^{-5}$

Precision: When it predicts yes, how often is it correct?

- True Positives divided by the number of predicted positives ('Predicted Yes')
  - Method 1: $\frac{9}{10} = 0.900$
  - Method 2: $\frac{9}{200} = 0.045$

The most frequently used performance metric is known as the "Receiver Operating Characteristic" (ROC). ROC is based on the True Positive Rate and the False Positive Rate. While ROC is a informative metric for many cases, it is not as useful in datasets where the positives are greatly outnumbered by the negatives (as in the case of a small number of bona fide nuclease off-target sites mixed in with a very large number of additional sites in the genome with some homology to the intended nuclease target). For example, compare the two methods in Table 2-1; both classifiers retrieve the same number of true positives (and therefore have the same 'True Positive Rate' / 'Recall'), but method 2 has a much higher number of false positives and is therefore clearly inferior. Using the False Positive Rate component of the ROC metric, the two rates have only a small difference of 0.0019 (see above definition of False Positive Rate) which masks the performance difference.

An alternative to the ROC that is particularly well-suited to cases with large numbers of negatives in the dataset is "Precision-Recall". The Precision measurements of these two methods show a large difference of 0.855 (see above definition of Precision), accurately capturing the considerable difference in the performance of the two classifiers.

The examples above calculated precision and recall for a specific data point, but in evaluating algorithm performance, precision is evaluated for all possible recalls to form a two dimensional curve (Figure 2-8). Precision-Recall curves typically trend from the upper left (high precision, low recall) to the bottom right (low precision, high recall), signifying that an algorithm becomes less precise as it is required to recovers more of the true positives. The overall performance of the algorithm is quantified by calculating the area under the curve (AUC) and is termed the 'average precision'. Precision-Recall curves often exhibit a

Table 2-1: Sample Confusion Matrix. Categorization of classifier predictions for a sample dataset. TN–True Negative. FN–False Negative. FP–False Positive. TP–True Positive.

	Method 1		Method 2	
n=100,000	Predicted: NO	Predicted: YES	Predicted: NO	Predicted: YES
Actual: NO	99,989 (TN)	1 (FP)	99,799	191
Actual: YES	1 (FN)	9 (TP)	1	9
	99,990	10	99,800	200

jagged 'sawtooth' like appearance due to the fact that finding several true positives in quick succession can actually cause an increase in the measured precision even at a higher recall value.

2.8.4 Cross-validation

In order to mitigate the possibility of over-fitting the model to the training data, cross-validation approaches are typically employed. Cross-validation (Figure 2-9) consists of randomly dividing a training set into several equal parts, using all but one subset of those parts as a training set, and then testing the model's predictive abilites on the "hold-out" data set (also known as the "test-set"). The process is then repeated choosing different subsets as the "hold-out" data set and the performance is averaged across all of the tested subsets. Additional cross-validation can be performed by iteratively repeating the whole previously described process with different randomized divisions of the dataset.

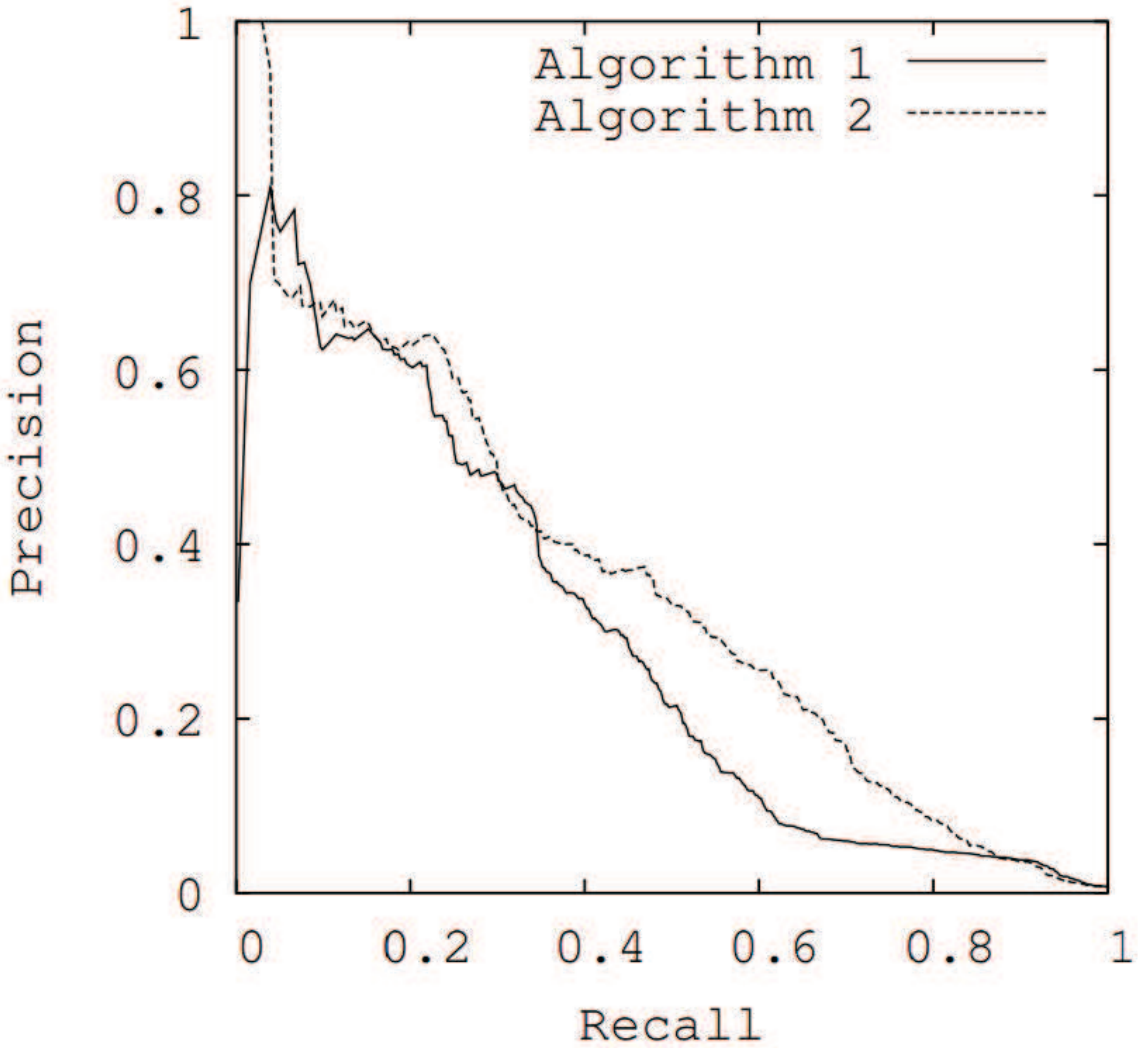

Figure 2-8: Precision-Recall Curve. Figure obtained from Davis and Goadrich [25].

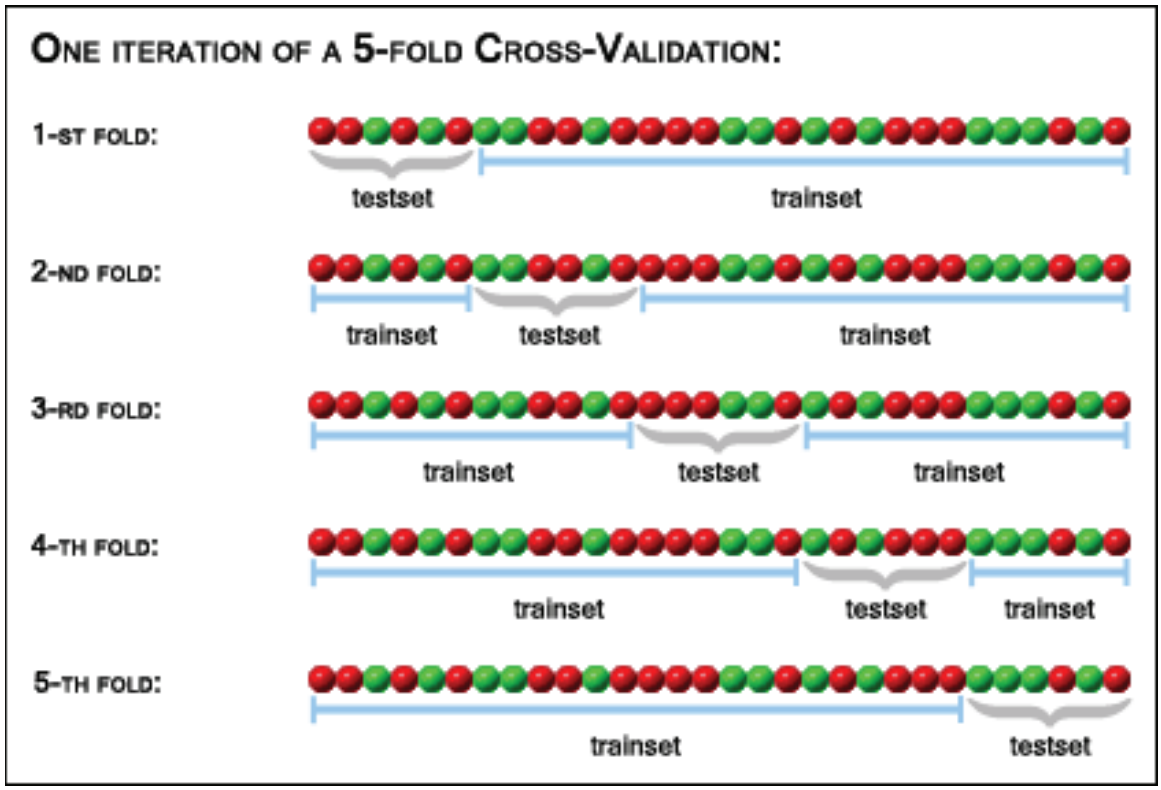

Figure 2-9: Cross Validation. A dataset consisting of two different classes (red circles and green circles) is broken up into a test set and a training set for each fold of the cross-validation. Image obtained from `http://genome.tugraz.at/proclassify/help/pages/XV.html`.

CHAPTER III

DEVELOPMENT OF AN ONLINE BIOINFORMATICS TOOL TO PREDICT ZFN AND TALEN OFF-TARGET SITES[1]

3.1 Abstract

Although engineered nucleases can efficiently cleave intracellular DNA at desired target sites, major concerns remain on potential off-target cleavage that may occur throughout the genome. We developed an online tool: Predicted Report Of Genome-wide Nuclease Off-target Sites (PROGNOS) that effectively identifies off-target sites. The initial bioinformatics algorithms in PROGNOS were validated by predicting 44 of 65 previously confirmed off-target sites, and by uncovering a new off-target site for the extensively studied zinc finger nucleases (ZFNs) targeting C-C chemokine receptor type 5. Using PROGNOS, we rapidly interrogated 128 potential off-target sites for newly designed transcription activator-like effector nucleases containing either Asn-Asn (NN) or Asn-Lys (NK) repeat variable di-residues (RVDs) and 3- and 4-finger ZFNs, and validated 13 bona fide off-target sites for these nucleases by DNA sequencing. The PROGNOS algorithms were further refined by incorporating additional features of nuclease-DNA interactions and the newly confirmed off-target sites into the training set, which increased the percentage of bona fide off-target sites found within the top PROGNOS rankings. By identifying potential off-target sites *in silico*, PROGNOS allows the selection of more specific target sites and aids the identification of bona fide off-target sites, significantly facilitating the design of engineered nucleases for genome editing applications.

[1]Modified from: **Fine EJ** et al. (2013). An online bioinformatics tool predicts zinc finger and TALE nuclease off-target cleavage. *Nucleic Acids Research* [34]

3.2 Introduction

The efficiency of genome editing in cells is greatly increased by specific DNA cleavage with zinc finger nucleases (ZFNs) or transcription activator-like (TAL) effector nucleases (TALENs), which have been used to create new model organisms [54, 66, 134, 129, 116, 44], correct disease-causing mutations [103], and genetically engineer stem cells [51]. However, both ZFNs [44, 88, 37, 101] and TALENs [116, 51] have off-target cleavage that can lead to genomic instability, chromosomal rearrangement, and disruption of the function of other genes. It is vitally important to identify the locations and frequency of off-target cleavage to reduce these adverse events, and ensure the specificity and safety of nuclease-based genome editing. Although the emerging systems utilizing clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated (Cas) proteins are highly active at their intended target sites, recent publications indicate that they likely have much greater levels of off-target cleavage than ZFNs or TALENs [53, 21, 35].

Experimental identification of ZFN and TALEN off-target sites is a daunting task because of the size of genome and the large number of potential cleavage sites to assay. Previous attempts to identify new off-target sites based entirely on bioinformatics search methods have all failed to locate any off-target cleavage sites [54, 129, 26, 103, 66, 134], which has led to the belief that identifying off-target activity based on sequence homology alone would not be fruitful [37]. In contrast, efforts using experimental methods to characterize the specificity of nucleases have successfully identified several off-target cleavage sites for ZFNs [44, 88, 37, 101] and TALENs [116, 51]. While most of these characterization methods incorporate a bioinformatics component to search through the genome, the final decision of what sites to investigate is dictated by the experimental data; for example, Perez et al. applied a classifier based on their characterization of the nucleases to narrow the full list of 136 genomic sites with two or fewer mismatches in each ZFN down to the top 15 sites they chose to interrogate [89].However, these experimental characterization methods, including SELEX [89, 50, 51, 116], bacterial one-hybrid [44], in vitro cleavage [88],

or IDLV trapping [37], can be very time consuming, costly, and technically challenging. This has severely limited the number of laboratories undertaking these experiments and the number of nucleases characterized for off-target effects. There is a clear unmet need for a rapid and scalable online method that can predict nuclease off-target sites with reasonable accuracy without requiring the user to have specialized computational skills, especially for application of nucleases in disease treatment.

3.3 Methods

3.3.1 Major Features of PROGNOS Ranking Algorithms

All PROGNOS algorithms only require the DNA target sequence as input; prior construction and experimental characterization of the specific nucleases are not necessary. Based on the differences between the sequence of a potential off-target site in the genome and the intended target sequence, each algorithm generates a score that is used to rank potential off-target sites. If two (or more) potential off-target sites have equal scores, they are further ranked by the type of genomic region annotated for each site with the following order: Exon > Promoter > Intron > Intergenic. A final ranking by chromosomal location is employed as a tie-breaker to ensure consistency in the ranking order.

The average 5' base and RVD-nucleotide frequencies for engineered TALEs were calculated by compiling previously published SELEX results of nine engineered TALEs [77, 51, 116] and calculating frequency matrices (Table 3-1).

The PROGNOS algorithm operates in three stages. First, all potential off-target sites that meet the search criteria (spacing distances, number of mismatches, etc.) are located. Second, these sites are all assigned a score according to the different algorithms. Finally, these sites are ranked by their scores and PCR primers are designed for top ranking sites.

3.3.2 PROGNOS Homology, RVDs and Conserved G's Algorithms

The 'Homology', 'RVDs', and 'Conserved G's' algorithms in PROGNOS all apply the "energy compensation" model of dimeric nuclease cleavage [88] to account for the interactions

Table 3-1: Comparisons of SELEX studies of engineered TALs to TALE-NT frequency matrix. The frequencies of RVD-nucleotide binding derived from natural TAL Effectors in Supplementary Table 1 from Doyle et al. [28] were subtracted from the averages of SELEX data from engineered TALs. A positive number implies that the RVD is predicted to be more likely to associate with that nucleotide in engineered TALs compared to naturally occurring TAL Effectors.

SELEX Averages from all TAL domains					
	Count	A	C	G	T
5' Base	9	0.053333	0.037942	0.008889	0.900947
NI	23	0.825696	0.077411	0.072375	0.024517
HD	52	0.075316	0.882704	0.009886	0.031709
NK	6	0.058025	0.012346	0.9	0.02963
NN	21	0.225663	0.04575	0.661518	0.068498
NG	37	0.092757	0.060536	0.043682	0.803835
Comparison to TALE-NT Frequency Matrix					
	A	C	G	T	
NI	-0.0153	-0.03159	0.047375	-0.00048	
HD	-0.00868	0.024704	-0.01511	-0.00129	
NK	0.033025	-0.01265	-0.025	0.00463	
NN	-0.13134	-0.10525	0.225518	0.011498	
NG	-0.00924	-0.04146	0.005682	0.045835	

between the two half-sites, but the scores for each half-site are calculated in different ways. The Homology algorithm can be applied to both ZFNs and TALENs and is based largely on the number of mismatches relative to the intended target sequence. The RVDs algorithm is designed for use with TALENs and utilizes the RVD-nucleotide binding frequencies of natural TAL Effectors [28]; alternate "5T"and "5TC" versions require either a thymidine or a pyrimidine to be in the 5' position of each half-site. The Conserved G's algorithm is designed for use with ZFNs and applies a weighting factor to the Homology algorithm that biases the rankings towards sites where intended guanosine contacts are maintained.

3.3.2.1 Homology Algorithm

If M represents the maximum number of mismatches allowed per nuclease half-site in a given PROGNOS search, a homology score for a potential off-target site is calculated by Equation 1:

$$Score_H = (M + 1 - L)^2 + (M + 1 - R)^2 \tag{1}$$

where L and R are the number of mismatches in the left and right half-sites, respectively. A higher score indicates a more likely off-target site. The squared factor captures some of the energy compensation effects observed in previous work [88]. L and R must be less than or equal to M as defined by the original search criteria.

3.3.2.2 Conserved G's Algorithm

Ranking ZFN off-target sites by counting the number of guanine residues—the "G's"—has proven useful because many ZFNs, especially those using canonical frameworks, bind to guanosine residues more strongly than other nucleic acids. The Conserved Gs ranking system adds a weighting factor to the homology score based on the number of guanosine residues in the intended target sequence (G_total) and the number of guanosine residues

matching the target sequence at potential off-target sites ($G_conserved$) according to Equation 2:

$$Score_G = Score_H * (\frac{G_{conserved}}{G_{total}} * 10)^{2.5} \quad (2)$$

A higher score indicates a site more likely off-target cleavage site. The weighting factor of 2.5 was developed by varying the weighting factor in order to optimize the number of previously published off-target sites for the CCR5, VEGF, and kdrl ZFNs identified in the top rankings [44, 88].

3.3.2.3 RVD Algorithm

The repeat variable di-residue (RVD) ranking system is intended for ranking TALEN off-target sites and uses the previously published method for ranking single TAL effector sites based on RVD nucleotide preferences observed in natural TAL effectors [28, 78]. Given RVD_{Lmin}, and RVD_{Rmin} as the scores for the left and right TALs binding to their intended target sites, and RVD_L and RVD_R as the scores for the TALs binding to a potential off-target sequence the score is calculated according to Equation 3:

$$Score_{RVD} = \left(\frac{RVD_L}{RVD_{Lmin}}\right)^{0.5} + \left(\frac{RVD_R}{RVD_{Rmin}}\right)^{0.5} \quad (3)$$

A lower score indicates a more likely off-target site. The exponent is an attempt to capture the "energy compensation" effects observed for the interaction between ZFN dimers [88]. Although there has been no direct testing of the energy compensation model with TALENs, the similarities between the nucleases (both use FokI cleavage domain, both bind in pairs) led us to include this model in our original algorithm. However, the optimization of the parameter in the TALEN v2.0 algorithm found a slight preference for an "energy distribution" model for TALENs instead of energy compensation. If no RVDs are specified by the user in the PROGNOS online input form, RVDs are assumed to follow the standard code based on the intended target sequence: NI→A, HD→C, NN→G, NG→T.

3.3.3 PROGNOS "ZFN v2.0" and "TALEN v2.0" Algorithms

The weightings of the parameters for the refined PROGNOS algorithms "ZFN v2.0" and "TALEN v2.0" were developed by training the algorithms to maximize recovery of previously confirmed off-target sites, as well as the novel off-target sites found using the initial algorithms developed in this study. For each algorithm, approximately 10^5 randomly assigned parameter sets (within a constrained range) were analyzed for their performance using the Perl off-target ranking script. The top performing parameter sets were further optimized by running further analyses allowing each parameter to vary slightly from the original value.

3.3.3.1 ZFN v2.0 Algorithm

The ZFN v2.0 algorithm was constructed based on the binding of individual zinc finger subunits rather than treating all mismatches equally. Previously, it had been hypothesized that each zinc finger subunit contributed a certain amount of overall binding energy and that if one nucleotide contact in the zinc finger were disrupted, further disruptions in that zinc finger were less detrimental. To model this, we analyzed each zinc finger subunit separately rather than considering all mismatches as equal (as was the case with the original "Homology" algorithm). To model the increased affinity most zinc fingers have for guanosine residues, a parameter for binding a guanosine at the intended position was included. Because many zinc fingers bind especially well to the "GNN" motif, if a guanosine was at the 5' position of a finger's triplet, the parameter was doubled.

Each finger was therefore analyzed as follows:

i) An initial score of 100 was given as a starting point.

ii) If there was at least one mismatch, "First_Penalty" was subtracted

iii) If there were additional mismatches, "Additional_Penalty" was subtracted for each additional mismatch

iv) If a guanosine was the intended base at position 2 or 3 and it matched, "G_Bonus" was added

v) If a guanosine was the intended base at position 1 and it matched, "G_Bonus"$*2$ was added

vi) If the resulting score was less than zero, it was set to zero

Because mismatches further from the FokI domain were better tolerated by zinc fingers in previous bona fide off-target sites, we introduced parameters weighting the impact of each of the 2nd-4th nucleotide triplets away from the FokI domain to model these polarity effects. The 1st triplet weighting was arbitrarily set to 1 and there were no published bona fide off-target sites at the time for ZFNs containing more than four zinc fingers to use to establish weightings of additional fingers. In the online implementation, all fingers past the fourth use the weighting parameter of the fourth finger.

The score for each zinc finger subunit is multiplied by the corresponding polarity parameter and all scores for the half-site are summed together. The sum is then divided by the score of a perfect match to the intended target sequence of that half-site and multiplied by 100 to generate a score from 0-100 (100 being a perfect match).

To allow for compensation effects between the two ZFN dimers, the score for each half-site was raised to the power of "Dimer_Exponent" before being summed together, divided by two, and multiplied by 100 to generate a score from 0-100 (100 being a perfect match).

The final optimized parameter values are given in Table 3-2. The formal definition of how to arrive at the scores of each ZFN half-site is as follows:

> The score of a ZFN/DNA alignment of l consecutive DNA bases in a 5'$\rightarrow$3' orientation starting at the FokI junction, with b_1 being the first base in the DNA helix and z_1 being the first intended base of the zinc finger helix, is calculated

using the following formula:

$$ZFN\_Half\_Site\_Score = \sum_{k=0}^{\frac{l}{3}-1} PC(3*k+1)*NN(100-MS(3*k+1)+GS(3*k+1)) \tag{4}$$

The score of the "Intended" half-site is calculated by setting b to equal z at all positions.

The formula utilizes the following functions:

- $NN(x)$ ensures that the value is not negative by returning x if $x \geq 0$ and returning 0 if $x < 0$
- $MC(start)$ returns the mismatch count (an integer from zero to three), for the zinc finger beginning at *start*, between the intended target sequence and the DNA bases according to the formula:

$$MC(start) = 3 - \sum_{k=start}^{start+2} (z_k == b_k) \tag{5}$$

- $PC(position)$ returns the polarity coefficient algorithm parameter for that position according to the following:

  $1 \leq position \leq 3 : 1.0$

  $4 \leq position \leq 6 : Polarity\_2$

  $7 \leq position \leq 9 : Polarity\_3$

  $10 \leq position : Polarity\_4$

- $MS(start)$ returns the mismatch score for the zinc finger beginning at *start* according to the rules:

  If $MC(start) == 0$, return 0

  If $MC(start) == 1$, return $First\_Penalty$

  If $MC(start) == 2$, return $(First\_Penalty + Additional\_Penalty)$

  If $MC(start) == 3$, return $(First\_Penalty + Additional\_Penalty * 2)$

Table 3-2: ZFN v2.0 Algorithm Parameters

Parameter Name	Training Range	Optimized Value
First_Penalty	30–90	70
Additional_Penalty	10–80	65
G_Bonus	0–35	17.5
Polarity_2	0.4–1	0.85
Polarity_3	0.4–1	0.8
Polarity_4	0.4–1	0.7
Dimer_Exponent	0.2–2.8	1.75

- $GS(start)$ returns the guanosine score for the zinc finger beginning at *start* according to the rules:

  Initialize $gSum = 0$

  If (z_{start} == "G" AND b_{start} == "G"): $gSum = gSum + G\_Bonus * 2$

  If ($z_{start+1}$ == "G" AND $b_{start+1}$ == "G"): $gSum = gSum + G\_Bonus$

  If ($z_{start+2}$ == "G" AND $b_{start+2}$ == "G"): $gSum = gSum + G\_Bonus$

  return $gSum$

Once the scores for each ZFN half-site have been calculated, then full off-target score can be calculated according to Equation 6:

$$ZFN\_v2.0\_Score = \frac{\left(\frac{Off\_Target\_Left\_Score}{Intended\_Left\_Score}\right)^{Dimer\_Exponent} + \left(\frac{Off\_Target\_Right\_Score}{Intended\_Right\_Score}\right)^{Dimer\_Exponent}}{2} * 100 \quad (6)$$

Several interesting points emerged from the optimization of the ZFN algorithm:

a) The polarity parameters did form a decreasing trend away from the FokI domain as hypothesized despite the possibility of remaining flat given their training range.

b) The dimer exponent was optimized at a value > 1, supporting the "energy compensation" model of dimeric nuclease interactions [88], despite the possibility of being equal

to or less than 1 given its training range.

c) The *First_Penalty* parameter optimized at a relatively high value (70% of the default starting energy of 100), implying that most of the binding energy of a zinc finger can be disrupted by a single mismatch.

d) The *Additional_Penalty* parameter also optimized at a relatively high value which brings the score for that finger close to zero in the case of two mismatches even with a high guanosine composition in the binding site, implying that nearly all remaining binding energy is disrupted by the second mismatch and that a third mismatch has negligible effect. However, there were limited examples of fingers with all three nucleotides disrupted in the training set of bona fide off-target sites, decreasing the significance of this observation.

3.3.3.2 TALEN v2.0 Algorithm

A score for each RVD-nucleotide interaction is calculated using the same formula as in TALE-NT [28] (and the original RVDs algorithm) except that the RVD-nucleotide frequencies used were derived from engineered TAL domains instead of naturally occurring TAL Effectors (Table 3-1). A value for the 5' base is also able to be calculated using the values derived from engineered TAL domains. Although there is no SELEX data for other RVDs, a user may enter any of the RVDs allowed by TALE-NT and the RVD-DNA binding frequency from TALE-NT will be utilized for that interaction; however, we make no claims as to the accuracy of predictions employing these alternate RVDs and it should be noted that substituting the TALE-NT frequencies into the TALEN v2.0 algorithm resulted in worse overall performance (Figure 3-9). If no RVDs are specified by the user in the PROGNOS online input form, RVDs are assumed to follow the standard code based on the intended target sequence: NI→A, HD→C, NN→G, NG→T.

Streubel et al. found that the presence of the "strong" RVDs NN and HD are key to TAL binding [112]. We hypothesized that these RVDs may impart excess binding energy that

could compensate for local effects of adjacent RVD-nucleotide mismatches. Accordingly, we developed two parameters, "Single_Strong" and "Double_Strong" that were applied to the score of RVDs that were flanked on one or both sides by NNs or HDs correctly bound to their respective intended bases (guanosines or cytidines). If these criteria were met, a fraction (defined by the parameter) of the difference between the mismatched RVD binding to its intended base and the base at the potential off-target site was subtracted from the score for that RVD-nucleotide interaction; subtraction is used because the base formula [28] uses a negative logarithm indicating lower scores as more likely binding sites.

In accordance with the findings that a polarity effect exists in TAL-DNA binding where mismatches further from the N-terminus have a less disruptive effect [75], the scores for the 14^{th} RVD (including score modifications relating to strong RVDs) and any RVDs further towards the C-terminus are all multiplied by the "Polarity" parameter.

The scores of all positions in each half-site are summed together to create the "Off_Target" score for that half-site and the full score for the potential off-target sites is computed using the "Dimer_Exponent" parameter and the score for a complete match between the RVDs and their intended target bases to yield a score from 0 to 100 (a perfect match) according to Equation 7.

$$TALEN\_v2.0\_Score = \frac{\left(\frac{Intended\_Left\_Score}{Off\_Target\_Left\_Score}\right)^{Dimer\_Exponent} + \left(\frac{Intended\_Right\_Score}{Off\_Target\_Right\_Score}\right)^{Dimer\_Exponent}}{2} * 100 \tag{7}$$

The formal definition to calculate the score of a TALEN half-site is as follows:

> The score of a TALEN/DNA alignment of l consecutive RVDs (r_l at each position with r_0 representing the N-terminal sequence that binds to the 5' base) aligned to DNA bases (b_l at each position with b_0 representing the 5' base) is

calculated using the following formula:

$$TALEN\_Half\_Site\_Score = P_{rn}(r_0, b_0) * PC(0) + \sum_{k=1}^{l}(PC(k) * (P_{rn}(r_k, b_k) - ((SFS(k) + DFS(k)) * (P_{rn}(r_k, b_k) - P_{rn}(r_k, I(r_k))))))) \quad (8)$$

The score of the "Intended" half-site is calculated by setting b to equal $I(r)$ at all positions.

The formula utilizes the following functions:

- A formula for the probability of an RVD-nucleotide interaction of the identical form as used in Doyle et al. [28]:
  $P_{rn}(RVD, base) = -\log(0.9 * SELEX(RVD, base) + (1 - 0.9) * 0.25)$
- $SELEX(RVD, base)$ returns the frequency of that RVD (or the N-terminal sequence that binds to the 5' base) binding to that DNA base according to the SELEX data in Table 3-1.
- $PC(position)$ determines the polarity coefficient by returning 1.0 for all $position \leq 14$ and returning the *Polarity* algorithm parameter for all $position \geq 15$.
- $I(RVD)$ returns the intended DNA base most frequently associated with RVD according to the SELEX analysis: NI→A, HD→C, NN→G, NK→G, NG→T
- $SFS(position)$ determines if that position has a single flanking strong RVD by returning the $Single\_Strong$ algorithm parameter if the following statement is true, and 0 if it is false:
  (($r_{position-1}$ ==“NN” OR $r_{position-1}$ == “HD”) AND $b_{position-1} == I(r_{position-1})$) XOR (($r_{position+1}$ ==“NN” OR $r_{position+1}$ == “HD”) AND $b_{position+1} == I(r_{position+1})$)

Table 3-3: TALEN v2.0 Algorithm Parameters

Parameter Name	Training Range	Optimized Value
Single_Strong	0–0.4	0.10
Double_Strong	0–0.5	0.15
Polarity	0.4–1.9	0.9
Dimer_Exponent	0.2–2.8	0.6

- $DFS(position)$ determines if that position has double flanking strong RVDs by returning the $Double\_Strong$ algorithm parameter if the following statement is true, and 0 if it is false:
  $((r_{position-1} ==$"NN" OR $r_{position-1} ==$ "HD") AND $b_{position-1} == I(r_{position-1}))$ AND $((r_{position+1} ==$"NN" OR $r_{position+1} ==$ "HD") AND $b_{position+1} == I(r_{position+1}))$

Several interesting points emerged from the optimization of the TALEN algorithm:

a) The polarity parameter was optimized to a value less than one, supporting the hypothesis that TAL binding in off-target locations exhibits polarity effects.

b) The dimer exponent was optimized at a value < 1, which is contrary to the "energy compensation" model of dimeric interactions for that was proposed for ZFNs and approximates a more "distributed energy" model where mismatches are preferred to be in equal numbers in each half-site rather than concentrated in one half-site. However, this finding is based on a relatively small number of bona fide off-target sites and it should be noted that several parameter sets with exponents > 1 performed fairly well, but not quite as well as the top performing sets with exponents < 1.

c) The $Double\_Strong$ and $Single\_Strong$ parameters were optimized at values greater than 0, indicating that there are some compensatory effects of flanking strong RVDs, supporting our hypothesis. Moreover, the double flanking strong RVD optimized to a

higher value than a single flanking strong RVD supporting the hypothesis that a second strong RVD has an additive effect over only a single flanking strong RVD.

3.3.4 Nuclease Construction

Four novel TALEN pairs and two novel ZFN pairs were designed to target sequences near the A→T mutation that causes sickle-cell anemia in the human beta-globin gene. TALENs were assembled using the Golden Gate method [14] and cloned into a mammalian expression destination vector containing the wild-type FokI domain. ZFNs were rationally designed to target overlapping sites. As these ZFNs target the same site, the activity and specificity of the 3-finger (3F) and 4-finger (4F) ZFNs can be directly compared. ZFN1-4F contains an additional finger added to ZFN1-3F, extending the target site from 9 bp to 12bp. ZFN2-4F shares two proximal fingers with ZFN2-3F, and uses a long linker between fingers two and three, extending the target site from 9bp to 13bp (Figure 3-1). The coding sequences for the ZFNs were ordered (IDT) and cloned into a wild-type FokI expression vector.

3.3.5 Cellular Transfection of Nucleases

HEK-293T cells were cultured under standard conditions (37°C, 5% CO2) in Dulbecco's Modified Eagle's Medium (Sigma Aldrich), supplemented with 10% FBS. Plates were coated with 0.1% gelatin. Passaging was performed with 0.25% Trypsin-EDTA. For TALENs, $2 * 10^5$ cells/well were seeded in 6-well plates 24 hours prior to transfection with FuGene HD (Promega). 3.3 μg of each nuclease plasmid along with 80 ng of an eGFP plasmid were transfected with 19.8 μL of FuGene reagent. Media was changed 24 and 48 hours after transfection. 72 hours after transfection, cells were trypsinized and the genomic DNA extracted using the DNeasy Kit (Qiagen). A small fraction of the cells were analyzed with the Accuri C6 flow cytometer to determine transfection efficiency by GFP fluorescence. For ZFNs, $8 * 10^4$ cells/well were seeded in 24-well plates and 100 ng of each ZFN was transfected using 3.4 μL of FuGene HD along with 10 ng of eGFP and 340 ng of

a

ZFNs:

```
                                    GCAGTAACGGCA – ZFN1-4F
                                    GCAGTAACG – ZFN1-3F
5′-CAACTTCATCCACGTTCACCTTGCCCCACAGGGCAGTAACGGCAGACTTCTCCTCAGGAGTCAGGTGCACCATGGTGTCTG-3′
3′-GTTGAAGTAGGTGCAAGTGGAACGGGGTGTCCCGTCATTGCCGTCTGAAGAGGAGTCCTCAGTCCACGTGGTACCACAGAC-5′
          ZFN2-3F – GGAACGGGG
      ZFN2-4F – AAGTGGAACGGGG
```

b

TALENs:

```
            S7 – TCACCTTGCCCCACAGGGCAGTAAC
            S5 – TCACCTTGCCCCACAGGGCAGT
5′-CAACTTCATCCACGTTCACCTTGCCCCACAGGGCAGTAACGGCAGACTTCTCCTCAGGAGTCAGGTGCACCATGGTGTCTG-3′
3′-GTTGAAGTAGGTGCAAGTGGAACGGGGTGTCCCGTCATTGCCGTCTGAAGAGGAGTCCTCAGTCCACGTGGTACCACAGAC-5′
                                                    TGTCCTCAGTCCACGT – S2
                                                      TCCTCAGTCCACGTGGT – S1
```

c

TALEN RVDs:

```
S1-NK: NK NK NG NK HD NI HD HD NG NK NI HD NG HD HD NG
S1-NN: NN NN NG NN HD NI HD HD NG NN NI HD NG HD HD NG

S2-NK: NK HD NI HD HD NG NK NI HD NG HD HD NG NK NG
S2-NN: NN HD NI HD HD NG NN NI HD NG HD HD NG NN NG

S5-NK: HD NI HD HD NG NG NK HD HD HD HD NI HD NI NK NK NK HD NI NK NG
S5-NN: HD NI HD HD NG NG NN HD HD HD HD NI HD NI NN NN NN HD NI NN NG

S7-NK: HD NI HD HD NG NG NK HD HD HD HD NI HD NI NK NK NK HD NI NK NG NI NI HD
S7-NN: HD NI HD HD NG NG NN HD HD HD HD NI HD NI NN NN NN HD NI NN NG NI NI HD
```

Figure 3-1: Binding sites of novel nucleases in the beta-globin gene. The binding sites of the novel ZFNs (**a**) and TALENs (**b**) designed to target the beta-globin gene mutation that causes sickle cell anemia. The start codon of beta-globin is underlined in bold. HEK-293T cells contain a T→C SNP relative to the reference genome in the middle of the S1 and S2 binding sites, which is underlined. The S2 TALEN is designed to target the sickle allele, which has an A→T mutation (underlined) at the 3' position of the TALEN binding site. This mutation is not present in HEK-293T cells. (**c**) The RVDs of the TALENs are shown. In "NK" versions of the TALENs, all guanosines are targeted by the NK RVD. In "NN" versions of the TALENs, all guanosines are targeted by the NN RVD: the other RVDs remain the same. The differences between the TALEN versions are underlined.

a Mock vector containing FokI but no DNA binding domain. 72 hours after transfection, cells were harvested and the genomic DNA extracted using 100 μL of QuickExtract (Epi-Centre). Mock transfections were performed similarly to the TALEN transfections, except that 6.6 μg of the mock FokI vector was transfected instead of TALEN plasmid.

3.3.6 Implementation of PROGNOS Search Algorithm

PROGNOS exhaustively searches for matches to queries by moving the query mask iteratively across the sequence of an entire genome, base by base. To optimize search time, the sequence comprising the length of the 5' binding site is first examined to determine if the number of mismatches does not exceed the query maximum. If that requirement is met, the sequences comprising potential 3' binding sites (separated by allowed spacing distances) are examined. PROGNOS was implemented in Strawberry Perl 5.12 and can be run in parallel on different processors. We found that scale up to 8 parallel processors with minimal efficiency losses was possible for most genomes.

3.3.7 PCR Amplification of Regions of Interest

3.3.7.1 Automated Design of PROGNOS PCR Primer Sequences[2]

An automated primer pair design process was included in PROGNOS to design primers appropriate for off-target validation assays, matching user input criteria. This greatly simplifies the standard method for primer design that requires iterative steps of primer design and verification of the resulting fragment sizes. In addition to speeding the primer design throughput, this automated design process allows the primers to be custom designed for the downstream assays or sequencing, and to be matched for high-throughput, full-plate PCR amplification.

For Surveyor assays, the primer design parameters can be specified to ensure that the nuclease site is placed in an optimal position within the amplicon to yield cleavage bands

[2]This automated primer design algorithm was also incorporated into the COSMID tool. Cradick TJ, Qiu P, Lee CM, **Fine EJ**, Bao G. (2014). COSMID: A Web-based Tool for Identifying and Validating CRISPR/Cas Off-target Sites. *Molecular Therapy—Nucleic Acids* [22]

that can be easily distinguished on gels from the parental band and each other. For resolution on a 2% agarose gel, the recommended parameters are: Minimum Distance Between Cleavage Bands—100, Minimum Separation Between Uncleaved and Cleaved Products—150.

To optimize amplicons for different sequencing platforms, the minimum distance from the edge of the amplicon to the nuclease site can be specified. For SMRT sequencing, the recommended parameters are: Minimum Distance Between Cleavage Bands—0, Minimum Separation Between Uncleaved and Cleaved Products—125.

The design process implemented by PROGNOS uses the following steps and considerations to yield primer pairs suitable for high-throughput PCR:

> Each possible position in the sequence 5' of the nuclease binding sites is considered as a possible 5' base for a primer (beyond the minimum distance specified between the edge of the amplicon and the nuclease site). For a given 5' starting position, the first 18 bases in the 3' direction are taken as an initial sequence for the primer. Then the following design loop begins:
>
> 1) Check for potential secondary structure that could result from the 3 end folding back.
>
>   Check that the sequence of the primer up to the 4th most 3' base does not contain any exact matches to the reverse complement of the three most 3' bases.
>
>   Example: For the potential primer sequence 5'-ACATTGAGGCACTACTTG-3', check that the sequence 'CAA' does not appear in 'ACATTGAGGCACTA'.
>
>   If there is a match, lengthen the primer by one base in the 3' direction and repeat the loop.
>
> 2) Check the predicted melting temperature of the primer and GC content.

Let *Length* be the number of nucleotides in the primer sequence

Let *%GC* be the percentage of guanosine and cytidine residues in the primer sequence

Calculate the melting temperature of the primer according to the equation:

$$T_m = 56.7 + 44.67 * \%GC - \frac{479.7}{Length} \quad (9)$$

If $T_m < 58$ or $\%GC < 0.33$ or $\%GC > 0.63$ then lengthen the primer by one base in the 3' direction and repeat from Step 1.

3) If the primer length is > 30 bp, then exist the design loop unsuccessfully—no primer for this position.

4) Check for high self-complementarity by comparing all 7 bp sequences within the primer to the reverse-complement sequence of the primer. If any exact matches are found, then exit the design loop unsuccessfully—no primer for this position.

5) If all requirements are met, then exit the design loop successfully and record the primer for this position.

After attempts to generate primers for all forward positions and all reverse positions are complete, pairs are made with each forward pair to each possible reverse pair. This list of pairs is then pruned to remove any that would result in products where the distances between nuclease sites and the ends of the amplicon fall outside of the specified ranges. This list is further pruned to remove any primer pairs that could potentially form primer dimers as defined by having the final 3' bases of one primer match the reverse complement of the final 3' bases of the other primer.

All primer pairs are then sorted by how close their melting temperature is to the target melting temperature (the default is 60°C) by computing $T_{diff} =$

$(T_{m_{forward}} - 60)^2 + (T_{m_{reverse}} - 60)^2$

All pairs where $T_{diff} < 2$ are then further sorted according to the following criteria (in order of priority):

1. Prefer shorter amplicon length
2. Prefer a shorter length of the longer primer sequence in the pair
3. As a final tie-breaker, sort the primer sequences alphabetically

If no primer pairs are found acceptable under this optimal set of criteria, the primer design constraints are iteratively relaxed and the primer search repeated. The most lenient set of criteria still require a minimum *%GC* of 0.25, a maximum *%GC* of 0.70, a maximum *Length* of 38, and a minimum melting temperature of 55°C.

3.3.7.2 PCR Experimental Conditions

The primers designed by PROGNOS (ordered from Eurofins-MWG-Operon) were used in a high-throughput manner to amplify genomic regions of interest in a single-plate PCR reaction. Each 25 μL reaction contained 0.5 units of AccuPrime Taq DNA Polymerase High Fidelity (Invitrogen) in AccuPrime Buffer 2 along with 150 ng of genomic DNA or 0.5 μL of QuickExtract, 0.2 μM of each primer, and 5% DMSO vol/vol. Touchdown PCR reactions were found to yield the highest rate of specific amplification. Following an initial 2-minute denaturing at 94°C, 15 cycles of touchdown were performed by lowering the annealing temperature 0.5°C per cycle from 63.5°C to 56°C (94°C for 30 seconds, anneal for 30 seconds, extend at 68°C for 90 seconds). After the touchdown, an additional 29 cycles of amplification were performed with the annealing temperature at 56°C before a final extension at 68°C for 10 minutes. Reactions were purified using MagBind EZ-Pure (Omega), quantified using a Take3 Plate and SynergyH4 Reader (Biotek) and normalized to 10 ng/μL.

3.3.8 High-Throughput Sequencing

3.3.8.1 Sequencing Chemistry

Amplicons from each transfection were pooled in roughly equimolar ratios and SMRT sequenced using the C2/C2 Chemistry and Consensus Sequencing options, according to the manufacturer's protocol (Pacific Biosciences).

3.3.8.2 Sequencing Analysis

There are three main processing steps of the raw SMRT sequencing reads to detect nuclease-induced non-homologous end joining (NHEJ). First, because many amplicons are pooled into a single SMRT sequencing cell, sequencing reads must be mapped to the amplicon from which they were generated. Second, because the processivity of the polymerase used in SMRT sequencing is a stochastic factor, the quality of the sequencing reads ranges over a distribution. However, for detecting the small insertions and deletions characteristic of NHEJ, sequencing artifacts that would yield false positives must be eliminated. Therefore, the sequencing reads must be filtered to obtain only the higher quality sequencing reads. Third, the high quality sequencing reads need to be analyzed to determine if they show mutations consistent with nuclease-induced NHEJ.

To address these issues, we created a sequencing processing pipeline based in Perl and also utilizing the BLAST and Needle software packages for sequence alignment. All aspects of the pipeline were implemented on a Windows machine and the source code is available on the PROGNOS website `http://bit.ly/PROGNOS`.

Sequence Mapping

1. Create a BLAST database of all expected amplicons obtained from the reference genome

2. BLAST each consensus SMRT sequencing read against the BLAST database using the parameters: `gapopen 2, gapextend 1, reward 1, penalty -1`.

3. Remove from further processing any reads that failed to make a significant BLAST alignment to any sequence in the database.

Pairwise Alignment

1. Use the Needleman-Wunsch algorithm [81] to align each sequence read with the expected amplicon to which it was mapped. Needle parameters: `gapopen 10, gapextend 1`.

2. If the alignment of the sequencing read extends more than 65 bp past the end of the reference sequence, remove it from further processing.

Sequence Quality Filtering

1. Calculate the average Phred score of each consensus SMRT read from the FASTQ data.

2. Remove from further processing any reads that have an average Phred score lower than 40.

3. Scan the region of the pairwise alignment extending 100 bp out from the edge of the nuclease binding sites for "indels". Define "indels" as a stretch of deleted, inserted, or mismatched bases in the sequencing read relative to the reference sequence.

4. If an indel is found that does not overlap the nuclease target site (the region encompassing the binding site of the left nuclease, the spacer region, and the right nuclease in the reference sequence), add the square of its length to a running total *errorCount*.

5. If $\frac{errorCount}{Length\_of\_scanned\_sequence} > 0.005$, remove that sequencing read from further processing.

Identifying Events of Non-Homologous End-Joining (NHEJ)

1. Scan the pairwise alignment extending 100 bp out from the edge of the nuclease binding site for indels.

2. Check if the observed indel overlaps the spacer region in the reference sequence.

3. If the indel overlaps the spacer and is of length 5 or greater, classify as NHEJ.

4. If the indel overlaps the spacer, is 4 or 4 bp, and is either composed entirely of a deletion or entirely of a tandem repeat of flanking sequence, classify as NHEJ.

5. Manually verify suspected NHEJ events by hand to confirm true cases of NHEJ.

3.3.9 Statistical Analysis

P values for off-target cleavage in Table 3-4 were calculated exactly as previously described [88]. Briefly, the t-statistic was calculated based on the fraction of mutated reads in the nuclease-treated sample compared to the fraction of mutated reads in the mock-treated sample and the number of sequencing reads was given as the degrees of freedom. In a similar manner, 90% confidence intervals were calculated by determining the upper and lower bounds of the fractions of mutated sequences that would yield *P* values of 0.05.

3.4 Results

3.4.1 Construction of initial bioinformatics ranking algorithms

The initial PROGNOS algorithms codified several established factors influencing nuclease specificity, including sequence homology, zinc fingers' preference for binding guanine residues [44], and RVD-nucleotide binding frequencies of natural TAL effectors [78]. To improve upon simple "mismatch counting", we incorporated the recently proposed "energy compensation" model of dimeric nuclease interactions [88]. Using these factors, three different algorithms were initially developed. The "Homology" algorithm, which could be used for both ZFNs and TALENs, generates a score based primarily on sequence divergence from the intended target site, including the number of mismatches in the left and right nuclease half-sites, and the maximum number of mismatches allowed per half-site. The "Conserved G's" algorithm (for ZFNs only) ranks ZFN target sites by counting

the number of guanine bases and adding a weighting factor to the homology score accordingly. The "RVDs" algorithm (for TALENs only) weighs mismatches based on RVD nucleotide preferences observed in natural TAL effectors and then applies the energy compensation model. Since all three of the TALEN off-target sites discovered previously using experimental-based off-target prediction methods contained a pyrimidine at the 5' position, a "5TC" version of the "Homology" and "RVDs" algorithms was also applied to TALEN rankings that required a thymidine or cytidine in the preceding 5' position of each half-site. For any given potential off-target site, these algorithms generate a score that allows ranking of all potential off-target sites in a genome for a specific nuclease target site. Search parameters, such as target sites, maximum mismatches per half-site and allowed spacer lengths are entered as inputs using the online interface (Figure 3-2a) and ranked lists of potential cleavage sites in the selected genome are given as PROGNOS outputs for further analysis. Although two online tools—ZFN Site [20] and TALE-NT [28]—exist to help search genomes for cleavage sites with homology to intended nuclease on-target sites, neither automatically ranks the potential off-target sites, nor has led to a report of any new experimentally verified off-target cleavage sites. In a direct comparison, we found that TALE-NT was only able to predict two of the seven bona fide TALEN off-target sites in unrelated gene families—three sites from previous work [116, 51] and four from this work—while PROGNOS could predict six. Recently, a new tool for identifying TALEN off-target sites, TALENoffer, was published [42]. Although it performs better than TALE-NT and does provide a rank-order for the potential off-target sites, it is outperformed by the refined TALEN v2.0 algorithm (Table 3-5).

3.4.2 Validation of PROGNOS Algorithms with Previously Confirmed Off-target Sites

To validate the initial PROGNOS ranking algorithms, we compared PROGNOS predictions with the off-target sites of ZFN and TALEN pairs identified by others using experimental characterization methods. If the same number of sites (1X) were interrogated as in the

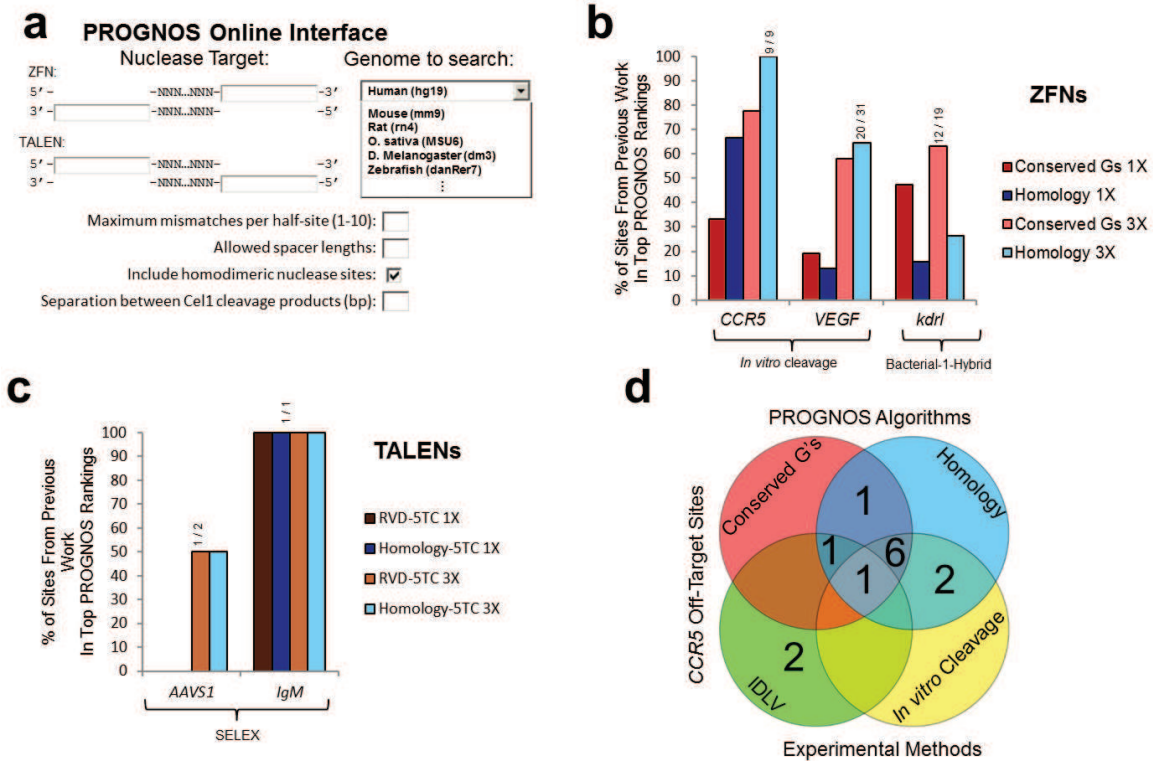

Figure 3-2: PROGNOS search interface and comparison to previous prediction methods. (**a**) The PROGNOS online interface allows users to enter the target site of their nuclease pair and specify search parameters and primer design considerations. (**b**) A comparison of PROGNOS predictions to previously reported methods identifying off-target sites for different ZFNs [44, 88]. The Homology and Conserved Gs algorithms were used to determine what percentage of the sites with previously identified off-target activity fell within the top fractions of PROGNOS rankings. The "1X" top fraction corresponds to searching the same number of top PROGNOS sites as were investigated in the original paper and "3X" corresponds to searching three times as many PROGNOS sites as were investigated in the original paper. (**c**) A comparison of the PROGNOS search algorithms to previously reported methods identifying off-target sites for TALENs [116, 51]. The top PROGNOS rankings using the Homology-5TC and RVD-5TC algorithms were searched to determine what percentage of off-target sites found to have activity fell within the top fractions of PROGNOS rankings. (**d**) Venn diagram displaying the 13 known off-target sites identified for the heterodimeric CCR5 ZFNs during development and testing of the original PROGNOS algorithms [88, 37]. The sites ranked at the top of the PROGNOS Homology and Conserved G's in silico algorithms (allowing 3X the number of sites searched by Pattanayak et al. [88]) are compared to the 12 sites identified previously and 1 site uncovered in this study.

original studies, but the sites were chosen by taking the top-ranked PROGNOS predictions, 33 ± 21% (mean ± std) of the off-target sites previously found in studies of ZFNs targeting *CCR5* [88], *VEGF* [88], and *kdrl* [44] could be located. Since off-target searches using the in silico PROGNOS predictions can be scaled up readily, we tripled (3X) the number of sites interrogated from PROGNOS top-ranked lists, and found that PROGNOS could identify 65 ± 24% of the off-target sites previously confirmed experimentally (Figure 3-2b). Excluding sites in highly homologous gene pairs such as *CCR5*/*CCR2*, only three bona fide TALEN off-target sites had previously been experimentally identified to date [116, 51], making a rigorous analysis of the predictive power of PROGNOS for ranking TALEN off-target sites more difficult. Nevertheless, we found that the "Homology-5TC" and "RVD-5TC" algorithms in PROGNOS could predict several off-target sites confirmed previously for TALEN pairs targeting the *AAVS1* [51] and *IgM* [116] loci (Figure 3-2c). Since no single off-target analysis method has yet been able to provide a comprehensive list of all off-target sites of a nuclease (Figure 3-2d) [37, 88], the comparison of PROGNOS predictions with previously published results may underestimate the power of PROGNOS. Specifically, these comparisons are limited by the small number of off-target sites experimentally validated previously, and do not reflect the ability of PROGNOS to predict new off-target sites.

3.4.3 Validation of Novel CCR5 ZFN Off-target Site Predicted by PROGNOS

To date, the only nuclease pair to have its off-target sites experimentally interrogated using two independent methods is a ZFN pair targeting *CCR5* (analyzed using *in vitro* cleavage [88] and IDLV [37]). These two studies located a total of 12 hetero-dimeric bona fide off-target sites, verified by sequencing the resulting mutations. A comparison between PROGNOS predictions using the "Homology" and "Conserved G's" algorithms and those 12 sites identified experimentally shows that PROGNOS (analyzing the top 3X number of sites interrogated by Pattanayak et al. [88]) was able to predict 10 out of the 12 off-target

sites (Figure 3-2d.

3.4.4 PROGNOS Search Output

PROGNOS provides ranked lists of potential nuclease cleavage sites that can be used to guide experimental evaluation of ZFN and TALEN off-target activities (Figure 3-3a). Specifically, for each pair of ZFNs or TALENs, the user-friendly online interface of PROGNOS (`http://bit.ly/PROGNOS` or `http://baolab.bme.gatech.edu/bao/Research/BioinformaticTools/prognos.html`) allows entry of the nuclease search parameters and returns lists of the top-ranked off-target sites according to the PROGNOS algorithms, as well as a full list of un-ranked potential off-target sites meeting the search parameters (Figure 3-3b). While the top-ranked sites provide a list of likely locations in a genome where off-target cleavage may occur, neither the PROGNOS rankings nor any published method can yet directly correlate the ranking with the precise level of observed off-target mutagenesis at a given site (Figure 3-4). Furthermore, to aid experimental analysis, PROGNOS also provides PCR primer sequences that can be used to amplify the potential nuclease cleavage sites in a high-throughput manner, a unique feature not present in other online search tools. Automated design of PCR primers significantly facilitates the analysis of off-target sites, since an initial experimental study of off-target cleavage by a single pair of nucleases typically requires at least 40 primers [54, 51], and an in-depth investigation of nuclease off-target effects may require >250 primers [44, 88]. Although tools such as Primer3 [100] can assist in primer design, they require a large amount of effort to generate primers optimal for off-target analysis due to specific requirements of where the nuclease site must be positioned within the amplicon. Although PCR amplification is an essential step in examining a potential off-target site, in previous investigations the success rates of amplifying off-target loci varied from 31% [54] to 95% [51]. In contrast, the primers automatically designed by PROGNOS had a robust 95% success rate across the 116 potential off-target loci interrogated in this study (Figure 3-3c). PROGNOS also provides the sequences, the

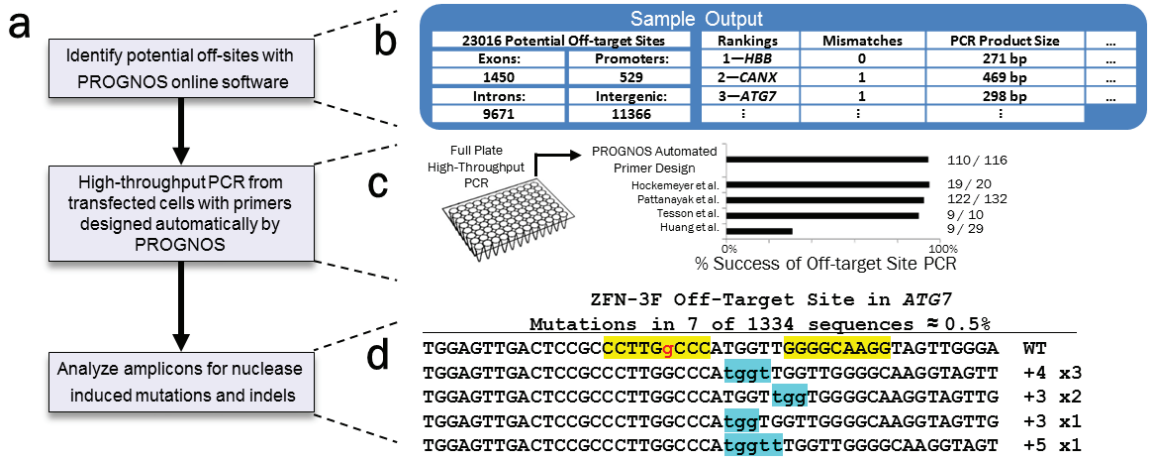

Figure 3-3: Using PROGNOS to identify nuclease off-target sites. (**a**) Outline of the procedure to identify nuclease off-target activity. (**b**) Sample outputs of the PROGNOS online software showing all sites found and what types of genomic regions they are located in as well as rankings of the top potential off-target sites. The rankings include the closest gene, the number of mismatches, the size of PCR product from the automatically designed primers, and other helpful information. (**c**) Comparison of the success of the automatically designed PROGNOS primers used in high-throughput full plate PCR of off-target sites to primers designed in other off-target publications. (**d**) Sequencing reads of an off-target location for the 3-finger ZFN pair that show evidence of NHEJ. In the wild-type (WT) sequence, the ZFN binding sites are highlighted in yellow and mismatches to the intended target sequence are lowercase red. In the sequencing reads, inserted bases are lowercase and highlighted in blue. The size of the indel is displayed to the right of the sequence, along with the number of times that mutation was observed.

sizes of expected cleavage products of the amplicons, and site of expected cleavage. This information is used when testing for nuclease-induced mutations—typically short insertions and deletions (indels) resulting from error-prone resolution of the DNA double strand break through the non-homologous end-joining (NHEJ) repair pathway—using methods such as the Surveyor Nuclease assay, high-throughput sequencing, or Sanger sequencing of TOPO-cloned fragments (Figure 3-3d).

3.4.5 Determination of NHEJ-mediated Indels Using high-throughput SMRT sequencing

To experimentally measure nuclease activity at on-target and potential off-target sites identified by PROGNOS, we used Single Molecule Real-Time (SMRT) sequencing of the PCR

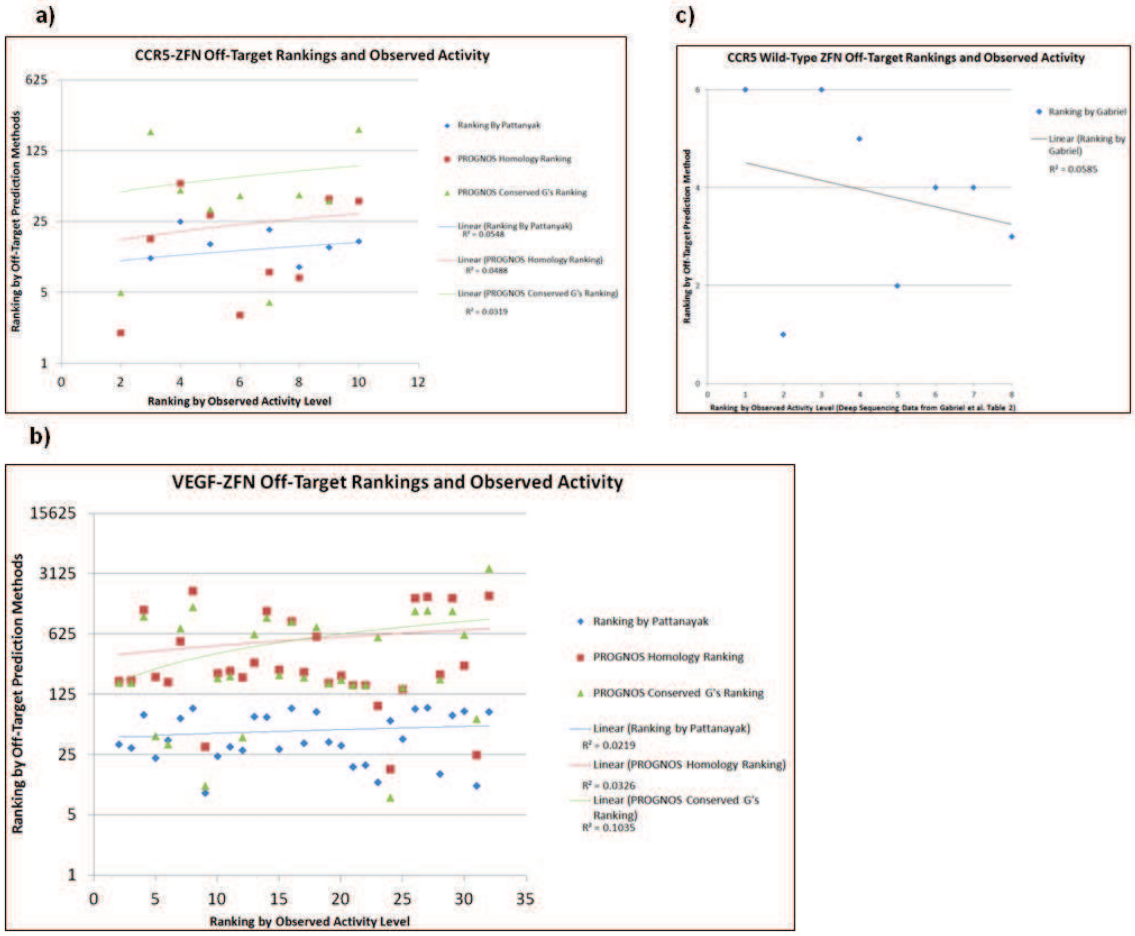

Figure 3-4: Low correlation between predicted off-target ranking and observed activity. The current goal of off-target prediction methods is to refine a global list of all possible off-target sites in the genome into a list of the top ranked sites that are likely to have off-target activity that can then be tested. An ideal prediction method would additionally be able to correlate local rankings (i.e. the 7th ranked site vs. the 12th ranked site) with observed levels of off-target activity. The off-target prediction method described by Gupta et al. [44] did not associate any rankings with the predicted off-target sites, however Pattanayak et al. [88] did provide rankings, as well as detailed information about observed off-target activity, and Gabriel et al. [37] provided the number of IDLV-CLIS events at each predicted site which can be used as a ranking. Using the results from Pattanayak et al. and the PROGNOS predictions for the same ZFNs, we constructed linear correlations measuring the ability of the different off-target prediction methods to rank order sites for the *CCR5* (**a**) and *VEGF* (**b**) ZFNs by their precise level of observed off-target activity. Similarly, we constructed linear correlations for the Gabriel et al. prediction of the wild-type ZFNs (the hetero-obligate ZFNs have only four points, making a linear correlation not statistically meaningful) and the corresponding activity (**c**); several of the sites found by Gabriel et al. are not found by PROGNOS so a meaningful comparison is not possible. The extremely low correlation coefficients indicate that neither the current experimental-based prediction methods, nor the PROGNOS algorithms can yet accurately model precise off-target activity levels. This is likely due to variability in the cells and poorly understood factors such as the genomic accessibility of a given off-target locus.

amplicons. The consensus sequencing mode of the SMRT platform provides highly accurate long length reads [117] that allowed determination of nuclease activity and specificity with reasonable sensitivity, and at a lower cost per run than other deep sequencing platforms. The good agreement between SMRT sequencing results and Sanger sequencing of TOPO-cloned samples further confirmed the accuracy of the SMRT-based analysis of nuclease cleavage (Figure 3-5a). Further, the high quality of the SMRT consensus sequence reads allowed us to achieve a much better signal to noise ratio for the mutation analysis than other sequencing methods [54]. We found that only three sequencing reads from mock treated control cells (~0.003% of the total) contained indels flagged by the analysis and all three were from the same genomic site, which in retrospect should have been excluded from sequencing analysis due to several long adjacent homopolymer stretches known to be error-prone during the sequencing process.

Although the spectrums of indels induced by ZFNs [44] or TALENs [54, 62] have been investigated previously, the long SMRT read lengths provided a more comprehensive analysis (Figure 3-5b). We found that ZFNs induced predominately 3, 4, and 5 bp insertions or deletions, with just a small number of large deletions. In contrast, TALENs induced indels over a much broader range, centered around 5 bp to 20 bp deletions, possibly due to the flexibility of the +63 C-terminal TAL domain [18].

3.4.6 Prediction and Validation of Off-target Sites for Novel Nucleases

To demonstrate the application of PROGNOS in analyzing newly designed nucleases, we investigated the off-target cleavage of four pairs of TALENs and two pairs of ZFNs (Table 3-4). TALENs containing the Asn-Asn (NN) RVD have been shown to be less specific than corresponding TALENs containing the Asn-Lys (NK) RVD [18]; however the difference in off-target activity of NN-TALENs and NK-TALENs has not been demonstrated in a genome-wide context. For ZFNs, although both 3-finger (3F) and 4-finger (4F) ZFNs have been shown to have off-target cleavage [44, 88, 37], there has been no direct comparison

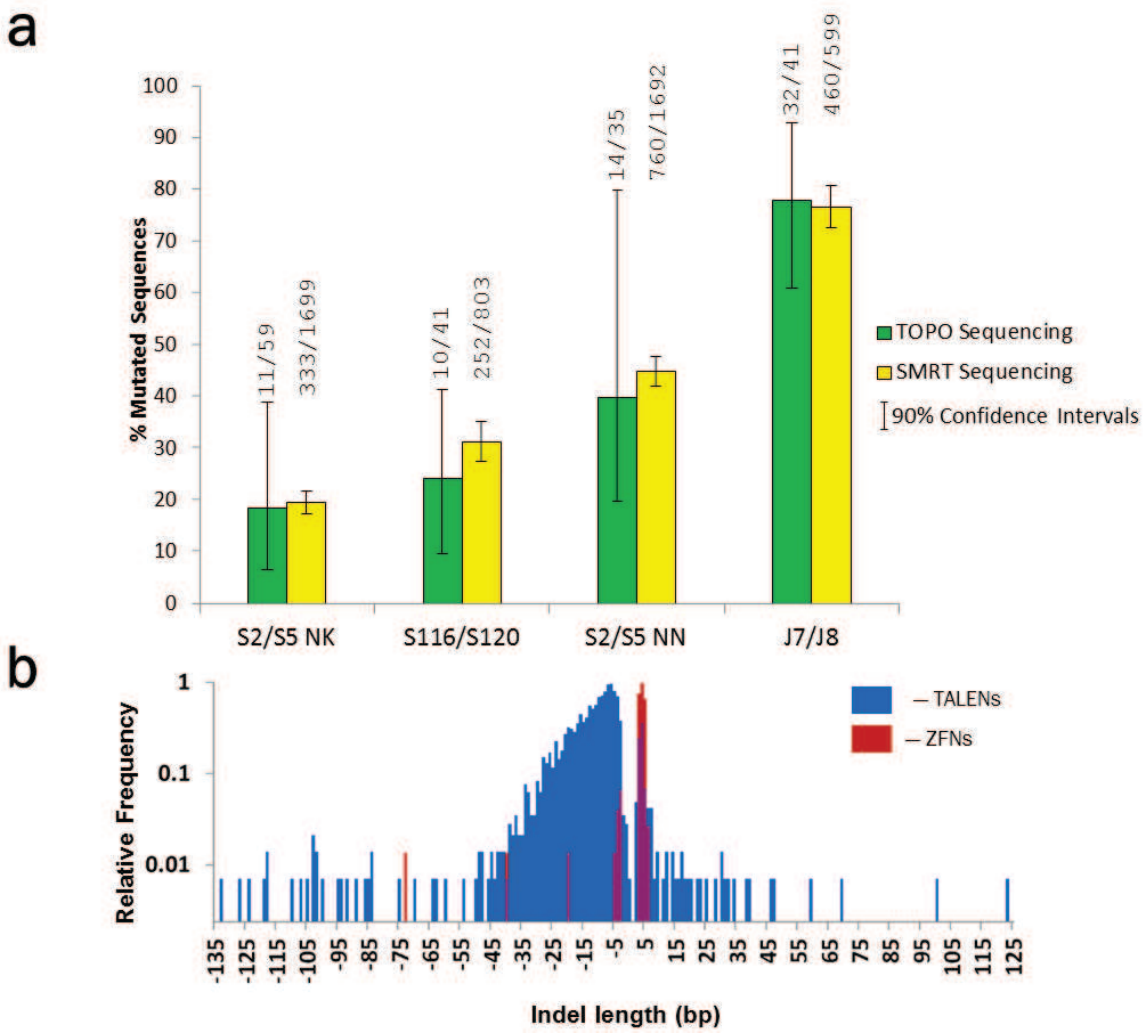

Figure 3-5: Using SMRT Sequencing to analyze nuclease activity. (**a**) SMRT sequencing produced very similar results to standard TOPO sequencing over a range of mutation rates from ~20% to ~76%. Error bars are 90% confidence intervals. S2/S5 NK and S2/S5 NN are the TALENs targeting beta-globin compared in this study. S116/S120 and J7/J8 are NK-TALENs targeting beta-globin and CDH1, respectively from Lin et al. [69]. (**b**) Comparison of the range and frequency of different sizes of indels observed in cells treated with TALENs or ZFNs. The observed frequencies of the different sizes are normalized to the frequency of the most common indel size for each nuclease type.

of off-target cleavage induced by 3F- and 4F-ZFNs that target the same DNA sequence. We expressed the TALENs and ZFNs in HEK-293T cells, and analyzed the PROGNOS top-ranked off-target sites (Table 3-4a,b). We found that TALENs exclusively using the NN RVD to target all of the guanosine nucleotides in the target sequence imparted higher activity level than TALENs exclusively using the NK RVD at corresponding positions, in agreement with previous reports [54, 18]. However, the NN-TALENs tested in this study had higher off-target cleavage activity than the corresponding NK-TALENs. For the first time, off-target cleavage by NK-TALENs was uncovered, as well as bona fide TALEN off-target sites with substantial (>5%) sequence divergence from the intended target that lacked a 5' pyrimidine and a site with a spacer longer than 24 bp (Table 3-4a). For ZFNs, we found that the 4F-ZFNs had higher on-target activity (consistent with previous reports that additional fingers increased activity [9]) and much lower off-target activity compared with the corresponding 3F-ZFNs targeting the same DNA site. Specifically, all six of the off-target sites found for the 3F-ZFNs had equal or greater activity than the off-target site of the 4F-ZFNs (a single site with 0.2% activity), with three sites having activity >1% (Table 3-4b).

3.4.7 Refinement of PROGNOS Ranking Algorithms

Although the set of initial PROGNOS algorithms (2 for ZFNs and 4 for TALENs) performed well in locating bona fide off-target sites for newly-designed nucleases based solely on in silico prediction, a user would still need to choose a specific algorithm or use all the available algorithms without knowing a priori which one would be most predictive for their nuclease. Using the expanded set of bona fide off-target sites including those found in this study (Table 3-4) as well as new insights into TALEN-DNA binding [112, 75], we refined the PROGNOS algorithms so that they are more sensitive, efficient, and user friendly compared with the initial algorithms. Although the "Homology", "Conserved G's" and

Table 3-4: SMRT Sequencing confirms on-target and off-target activity at sites ranked by PROGNOS We interrogated 138 highly ranked genomic loci for the novel TALENs (**a**) and ZFNs (**b**) using SMRT, and observed off-target activity in 13 cases, 9 of which were outside the globin gene family. The "match type" indicates the orientation of the left (L) and right (R) nucleases at the site and the length of the spacer sequence. In sequences, lower-case letters indicate mutations compared to the target site. Site sequences are listed as 5'–(+) half-site–spacer–(-) half-site–3'. Therefore, the (-) half-site for TALENs and the (+) half-site for ZFNs are listed in the reverse anti-sense orientation compared to the DNA sequence that the nuclease binds. Rankings by the initial PROGNOS algorithms Homology (column H), RVDs for NK (RK), RVDs for NN (RN), and Conserved G's (C) are displayed as well as the rankings by the refined "TALEN v2.0" algorithm for NK (TK) and for NN (TN) and the "ZFN v2.0" algorithm (Z). 293T Modification Frequency is the percentage of observed sequences showing evidence of non-homologous end-joining repair. (**c**) 15 off-target sites for the CCR5 ZFNs ranked by PROGNOS that had not been previously investigated were interrogated using SMRT, validating a novel off-target site. Additionally, 6 known highly active off-target sites were sequenced as positive controls. The PROGNOS rankings for the site near KDM2A are listed as "N/A" because the site was not found by PRGONOS due to the high number of mismatches. * indicates $P < 0.05$ in cells expressing active nuclease compared to cells expressing empty vector. ^ indicates $P < 0.05$ for the difference in activity between NK and NN at that site.

a) Novel TALENs

Nucleases	Closest Gene	Match Type	Mutations per half-site (+)	Mutations per half-site (-)	(+) half-site	(-) half-site	PROGNOS Rankings H	RK	RN	TK	TN	293T Cell Line Modification Frequency RVD Targeting Guanine NK	NN
S2/S5 TALENs	HBB	L-16-R	0	1	TCACCTTGCCCCACAGGGCAGT	tCAGGAGTCAGGTGCA	1	1	1	1	1	23.0% *	48.7% *^
	FAM3D	R-17-R	3	3	TGCcCCTGACTCCTta	AaAtGAGgCAGGTGCA	4	15	25	5	6	0.1% *	0.05%
	HBD	L-16-R	2	2	TCACtTTGCCCCACAGGGCAtT	tCAGGAGTCAGaTGCA	2	2	2	2	2	0%	5.0% *^
	GPR6	R-30-R	2	5	TcCACCTGgCTCCTGT	gCAGGAGTtAaGgGtA	21	241	16	11	5	0%	0.09% *^
							Total Sites Interrogated:					21	20
S1/S7 TALENs	HBB	L-15-R	0	0	TCACCTTGCCCCACAGGGCAGTAAC	AGGAGTCAGGTGCACCA	1	1	1	1	1	0.3% *	42.8% *^
	LINC00299	R-23-R	3	5	TGGaGCACCTGACcCCa	AGGAGaaAaGgGCACCt	17	8	50	5	10	0.2% *	0.1%
	HBD	L-15-R	3	1	TCACtTTGCCCCACAGGGCAtTgAC	AGGAGTCAGaTGCACCA	2	2	3	2	2	0%	4.9% *^
	FAM3D	R-21-R	3	5	ctGTGCcCCTGACTCCT	AtGAGgCAGGTGCAttt	8	4	2	13	6	0%	0.1% *^
							Total Sites Interrogated:					24	25

b) Novel ZFNs

							H	C	Z	ZFN Activity
4F ZFNs	HBB	L-5-R	0	0	TCACCTTGCCCC	GCAGTAACGGCA	1	1	1	6.3% *
	PLG	R-5-R	3	1	TGCCaTTgaTGC	GCAGTAACtGCA	24	41	28	0.2% *
							Total Sites Interrogated:			23
3F ZFNs	HBB	L-5-R	0	0	CCTTGCCCC	GCAGTAACG	1	1	1	1.9% *
	ATG7	L-6-L	1	0	CCTTGgCCC	GGGGCAAGG	3	7	11	0.5% *
	TMEM132C	R-6-L	1	0	aGTTACTGC	GGGGCAAGG	4	35	9	0.2% *
	PARD3B	L-5-L	0	1	CCTTGCCCC	GGGGCAAGc	5	8	7	1.3% *
	GLIS2	L-6-L	1	0	CCTgGCCCC	GGGGCAAGG	9	6	4	0.6% *
	AFF3	L-6-L	2	0	CCTaGgCCC	GGGGCAAGG	16	37	20	2.9% *
	RGS10	L-6-L	0	2	CCTTGCCCC	GGGGCAgaG	22	39	15	5.2% *
							Total Sites Interrogated:			23

c) *CCR5* ZFNs

PROGNOS Investigation	CCR5	L-5-R	0	0	GTCATCCTCATC	AAACTGCAAAAG	1	1	1	31% *
	CSNK1G3	L-5-R	3	1	GcCtTCCcCATC	AAAgTGCAAAAG	33	13	18	0.09% *
							Total Sites Interrogated:			16
Known Off-Target Sites	CCR2	L-5-R	1	1	GTCgTCCTCATC	AAACTGCAAAAa	2	5	2	11% *
	KDM2A	R-5-L	2	5	CTaTTaCAGTTT	GATGAGGtctca	N/A	N/A	N/A	2.6% *
	BTBD10	L-5-R	2	1	GTttTCCTCATC	AAACTGCAAAAt	3	45	6	2.6% *
	KCNB2	L-5-R	3	1	aTgtTCCTCATC	AAACTGCAAAtG	29	33	8	1.3% *
	WBSCR17	R-6-L	2	2	CTgTTcCAGTTT	GcTGAGGATaAC	60	51	95	1.4% *
	TACR3	L-5-R	1	3	GTCATCtTCATC	AAACTGtAAAgt	17	197	26	8.6% *

"RVDs" algorithms (including the "5TC" version for TALENs) all located bona fide off-target sites, no algorithm was consistently superior across all ZFNs or all TALENs studied (Figure 3-2b,c and Table 3-4). In developing the refined algorithms, we were able to unify the different algorithms for each type of nuclease into a single algorithm (ZFN v2.0 for ZFNs, TALEN v2.0 for TALENs). Compared with the original PROGNOS algorithms, ZFN v2.0 and TALEN v2.0 achieved higher precision within the Top 16 rankings (representing the minimum recommended size of a small-scale off-target analysis), located higher mean percentages of known off-target sites per nuclease across all nucleases tested (within the top 3X rankings for previously investigated nucleases and within the same number of sites as in the PROGNOS-based investigations, and had lower standard deviations of the mean percentages, demonstrating that the refined algorithms performed more consistently across all nucleases tested.

In developing the refined and unified ZFN algorithm, we added factors weighing a model of the binding energy of each zinc finger subunit [88] and polarity effects reflecting the distance of a mismatch from the FokI domain and allowed more flexible models of the previous concepts of energy compensation between the two half-sites of a nuclease pair and a stronger affinity for guanosine residues. This new "ZFN v2.0" algorithm outperforms the initial "Homology" and "Conserved G's" algorithms for ZFNs in terms of both identifying a larger set of bona fide off-target sites for the nucleases tested and having superior precision within the Top 16 rankings (Figure 3-6a). The Top 16 ranked sites were chosen as a cutoff because by necessity nearly all of the novel off-target sites found were within the Top 24 rankings of one of the original algorithms since that was their initial criteria for being selected for investigation. Therefore, a stricter cutoff was required in order to observe differential performances between the algorithms for these new sites.

Recently, Sander et al. [101] used Bayesian machine learning to re-analyze the original results of the in vitro cleavage experiments for *CCR5* and *VEGF* ZFNs [88] and subsequently developed two separate classifiers that ranked all sequences in the human genome

for their potential as off-target sites of either the *CCR5* or *VEGF* ZFNs respectively. Their work validated 25 new bona fide off-target sites for the *CCR5* and 26 new sites for the *VEGF* ZFNs, but did not locate—among any of the 15,882 possible off-target sites predicted for the *CCR5* ZFNs by their classifier system—the novel off-target site for the *CCR5* ZFNs predicted by the PROGNOS algorithms near *CSNK1G3* that was validated in this study (Table 3-4c).

Since the 51 new sites found by Sander et al. [101] were not part of the training set for the "ZFN v2.0" algorithm, this provided an opportunity to test the new algorithm for its ability to locate additional off-target sites. By extending the standard PROGNOS search limit recommendations for the *CCR5* ZFNs to allow for a larger number of possible off-target sites (3X the number of possible off-target sites considered by Sander et al.), we found that the refined ZFN algorithm successfully identified more than half (13 of 25 = 52%) of the new off-target sites for those ZFNs (Figure 3-6b). For the *VEGF* ZFNs, the standard PROGNOS search provided enough potential off-target sites to make an appropriate 3X comparison to Sander et al. [101], and the refined algorithm again located more than half (18 of 26 = 69%) of the new off-target sites for those ZFNs. Three additional pairs of ZFNs (a 3-finger pair, a 4-finger pair, and a 5-finger CompoZr pair from Sigma-Aldrich) which had previously been investigated using the Homology and Conserved G's PROGNOS algorithms [79, 1] were also re-analyzed using the refined algorithm and all six of the previously located bona fide off-target sites were highly ranked by ZFN v2.0. Taken together, these results provide significant evidence that the refined ZFN algorithm was not overtrained to existing sites during its development and is able to robustly predict additional bona fide off-target sites. An analysis of each of the components of the ZFN v2.0 algorithm showed that while all play a part in the improved performance, some parameters are more critical to the algorithm than others (Figure 3-7).

In developing the refined and unified TALEN algorithm, we added new parameters based on compensatory effects of strong RVDs (NN and HD) [112] on adjacent mismatches

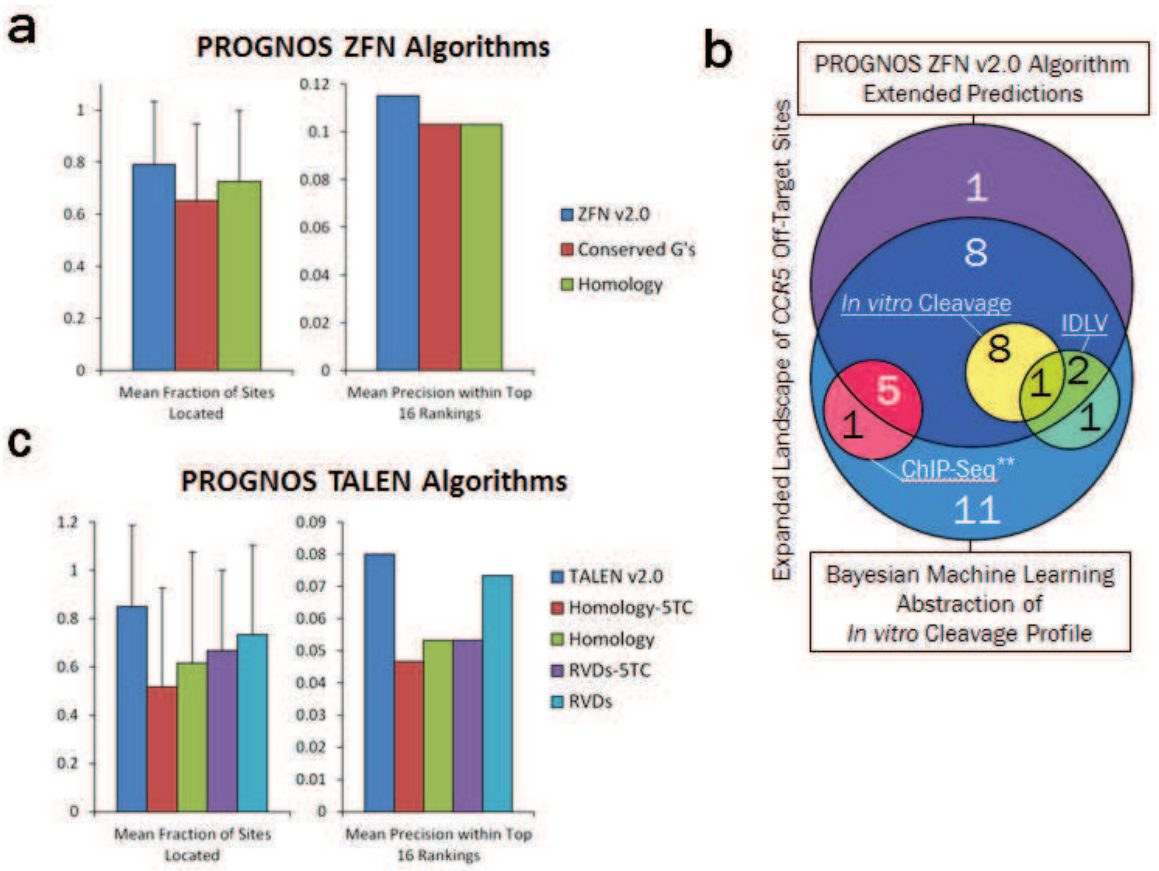

Figure 3-6: Improved performance of the refined PROGNOS algorithms. (**a**) The performance of the two initial ZFN algorithms and the refined "ZFN v2.0" algorithm are compared for their ability to predict off-target sites for the all ZFNs in the training and validation sets. Percentages of off-target sites located were calculated according to 3X limits for previous studies and within the number of sites interrogated for PROGNOS-based studies (typically the top 24 ranked sites). Error bars represent standard deviation. (**b**) The expanded landscape of 38 total heterodimeric off-target sites for the *CCR5* ZFNs found by four different experimental-based prediction methods and the refined "ZFN v2.0" PROGNOS algorithm. The PROGNOS sites are drawn from the top rankings spanning 3X the number of predictions by the Bayesian abstraction of the *in vitro* cleavage profile. (**) Note that only six of the sites found using ChIP-Seq were provided by Sander et al. [101], so the full degree of overlap of all ChIP-Seq sites with sites found by other methods is unclear. (**c**) The performance of the four original TALEN algorithms and the refined "TALEN v2.0" algorithm are compared for their ability to predict off-target sites for all TALENs in the training and validation sets.

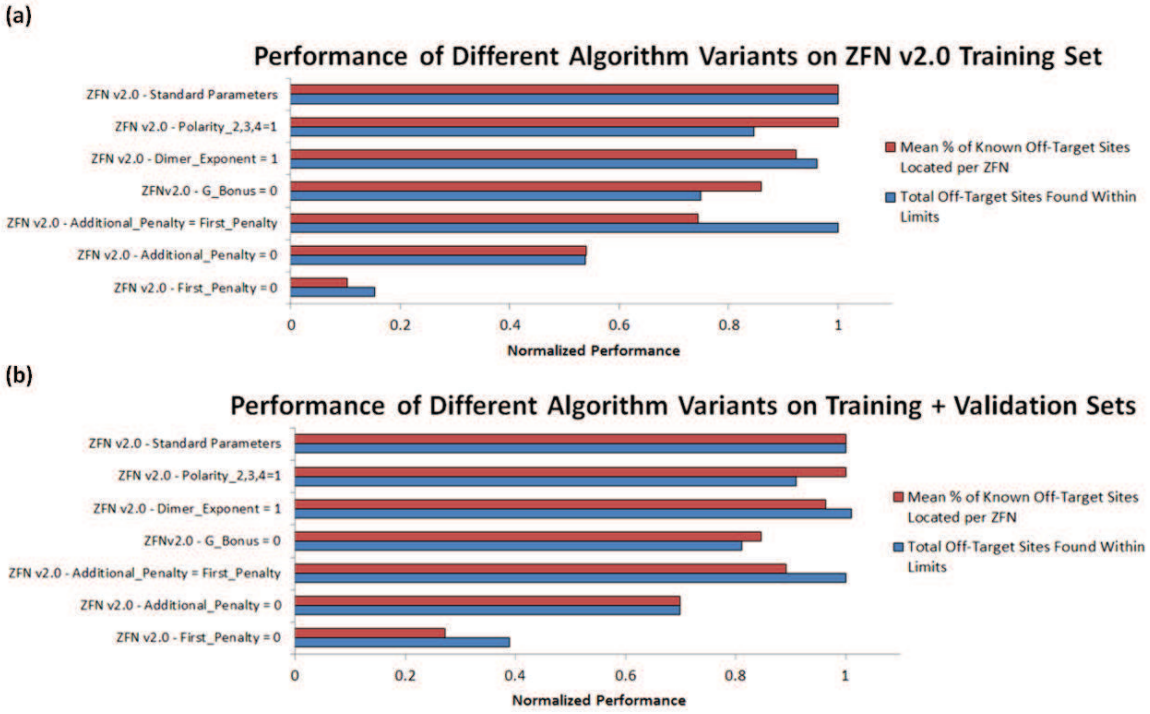

Figure 3-7: Performance of ZFN v2.0 Algorithm Variants. To determine the impact of each of the individual parameters of the ZFN v2.0 algorithm, each parameter was systematically removed from the analysis of both the training set (**a**) and the full list of ZFN off-target sites comprising the training set and the validation set (**b**). The specific alterations to the parameters in the algorithm formula are supplied in the graph. To measure aggregate performance of the algorithms across a wide range of ZFNs, we measured the percentage of known off-target sites for each ZFN pair that the algorithms were able to identify (within 3X limits for previously studied ZFNs, and within the limit of the original Homology and Conserved G's PROGNOS experimental searches for newly studied ZFNs), and then took the mean of these percentages across all ZFNs. In some cases, algorithms tied in performance by this metric, so we further measured the ability of the algorithms for the total number of off-target sites that they identified within the defined limits. The ordering of the algorithm variants was chosen based on the results for the training set data (**a**) and maintained in the same ordering for the full off-target set (**b**). From this analysis, we observed that while removing any one of the individual parameters does not cause a severe decrease in effectiveness of the algorithm (with the exception of setting either First_Penalty or Additional_Penalty to 0), it is only when all of the parameters are incorporated that optimal results—both in the training set and the full set—are achieved.

and polarity effects indicating that mismatches further from the N-terminus are less disruptive [75]. These new considerations were combined with a model of dimeric nuclease interactions, as well as RVD-nucleotide association frequencies. To improve upon the RVD-nucleotide association frequencies derived from natural TAL effectors [28], as were used in the initial "RVDs" algorithm and the TALE-NT online tool [28], we calculated association frequencies based on SELEX data from engineered TAL domains [77, 116, 51] (Figure 3-8 and Table 3-1). Importantly, this generated an association frequency for the 5' "Position 0" in the TALEN binding site that allowed us to use this parameter to unify the "5TC" and standard versions of the "RVDs" algorithm. Further, we found that while the nucleotide frequencies for the RVDs NI, HD, NK, and NG did not appreciably vary between engineered TALEs and natural TALEs, the results for NN were substantially different. Although the NN RVD is still the least specific of all the standard RVDs, in engineered TALEs it showed a stronger preference for its intended base (guanosine) and a reduced preference for adenosines and cytidines compared with that of naturally occurring TALEs (Table 3-1). We found that the new unified "TALEN v2.0" algorithm outperforms the four initial algorithms for TALENs in terms of both having higher precision within the Top 16 rankings and locating a higher mean percentage of known off-target sites per nuclease across all nucleases tested (Figure 3-6c). The refined TALEN algorithm was additionally able to predict several bona fide TALEN off-target sites not in its training set that were found using the initial PROGNOS algorithms [79], demonstrating that the refined algorithm was not overtrained during development and retains robust predictive capabilities. An analysis of each of the components of the TALEN v2.0 algorithm showed that while all play a part in the improved performance, some parameters are more critical to the algorithm than others (Figure 3-9).

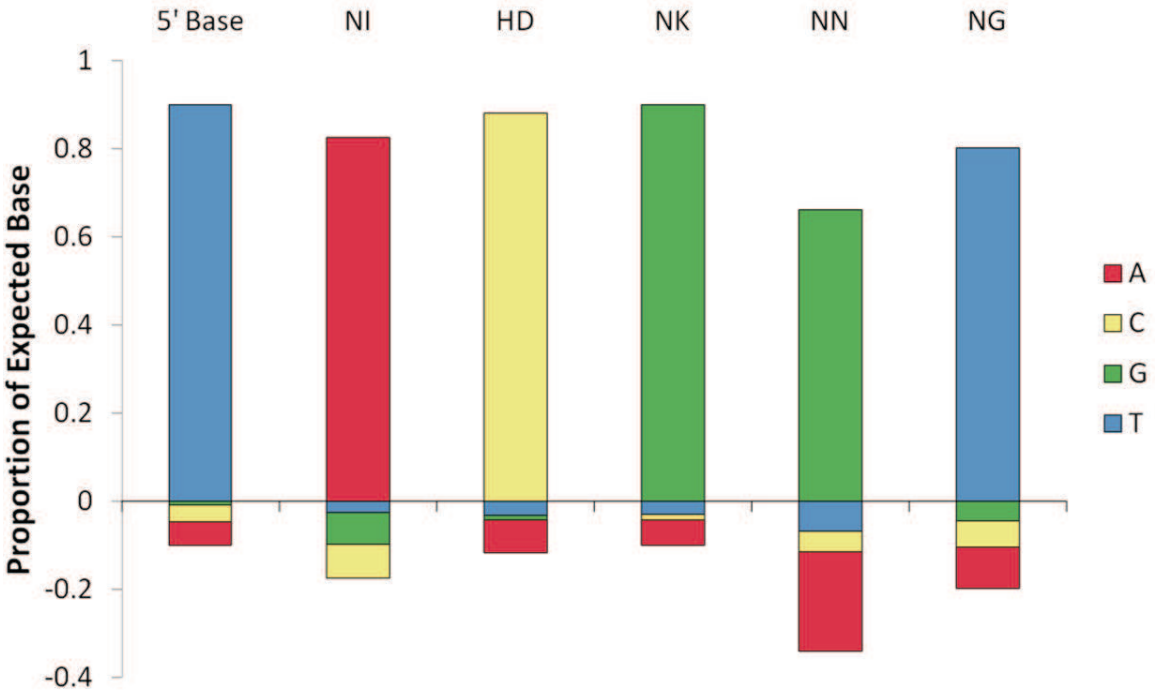

Figure 3-8: Average RVD-nucleotide frequencies of engineered TAL domains. Published SELEX data from nine engineered TAL domains [77, 116, 51] were compiled to derive the average nucleotide association frequencies for the 5' base and the 5 common RVDs (Table 3-1). These frequencies are employed in the "TALEN v2.0" prediction algorithm.

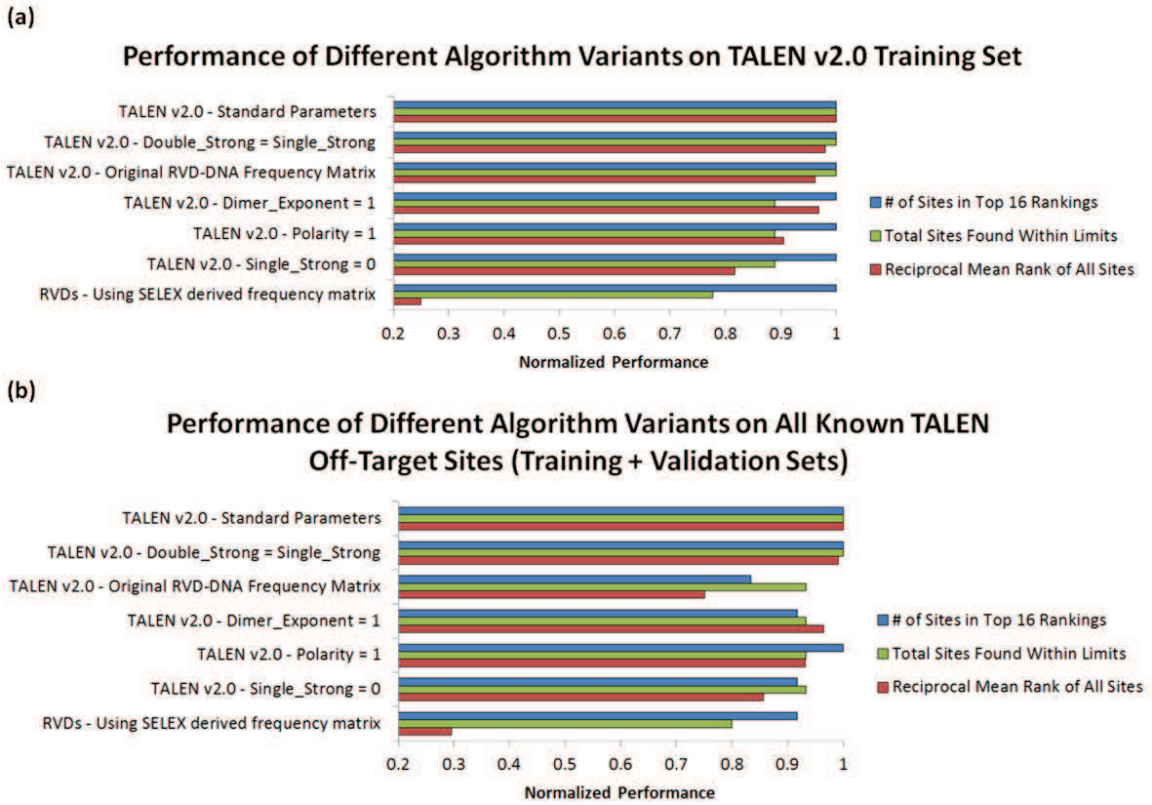

Figure 3-9: Performance of TALEN v2.0 Algorithm Variants. To determine the impact of each of the individual parameters of the TALEN v2.0 algorithm, each parameter was systematically removed from the analysis of both the training set (**a**) and the full list of known TALEN off-target sites comprising the training set and the validation set (**b**). The specific alterations to the parameters in the algorithm formula are supplied in the graph. Because investigations of TALEN off-target activity have been much more limited than ZFNstypically investigating ~20 locations and finding 1-2 bona fide off-target sites as opposed to investigating 30-100 potential sites out of up to thousands of predicted locationswe were able to use more precise measurements of the algorithms' performances. Our initial chief metric for this analysis was number of bona fide sites found in the Top 16 rankings, but as the total number of bona fide sites within those rankings was so low (6 sites was the highest result in the training set) we found that there were not any differences in this metric for many of the variants. We therefore further included secondary and tertiary measurements of the total number of bona fide off-target sites found within the search limits (3X for previous investigations and 24 sites for novel nucleases tested) and the mean rank of all off-target sites located (taking the reciprocal in this case so that larger values implied a superior performance). The ordering of the algorithm variants was chosen based on the results for the training set data (**a**) and maintained in the same ordering for the full off-target set (**b**). While removing some of the individual parameters does not cause a marked decrease in effectiveness of the algorithm, it is only when all of the parameters are incorporated that optimal results—both in the training set and the full set—are achieved.

3.4.8 Sensitivity and Specificity of PROGNOS Search Algorithms

When applying the initial PROGNOS algorithms to identify off-target sites for newly constructed NN-TALENs and 3- and 4-finger ZFNs, we obtained a very manageable average false positive ratiodefined as the number of interrogated sites with no detectable activity compared to the number with detectable activityof only ~11:1, which is less than 2-fold higher than current experimental prediction methods (Figure 3-10a). When interrogating three additional pairs of NN-TALENs with the initial algorithms, we observed a similarly low false positive ratio of 11:1 [79]. For NK-TALENs, the false positive ratio was higher (~21:1); however, since no previously published method has identified any off-target sites for NK-TALENs, we were not able to make a meaningful comparison of the false positive ratio with experimental-based prediction methods. As the new "ZFN v2.0" and "TALEN v2.0" algorithms have higher precision among the Top 16 rankings, we would expect that their false positive ratios would be even lower than the initial algorithms when used as the basis for investigations of novel nucleases.

As mentioned above, to date only a single nuclease pair (the heterodimeric sites of the *CCR5* ZFNs) has had its off-target cleavage investigated by independent experimental prediction methods [88, 101, 37], and it is therefore the only pair for which a false negative rate analysis can be conducted. Defining the false negative rate as the percentage of all known off-target sites that are not predicted by the particular method within a top portion of the rankings, the PROGNOS algorithms had false negative rates equal or superior to the IDLV and *in vitro* cleavage experimental prediction methods (Figure 3-10b). A Precision-Recall analysis of the different predictive methods for the *CCR5* ZFNs using the false discovery and true positive rates also demonstrates that the PROGNOS algorithms perform comparably to experimental based prediction methods.

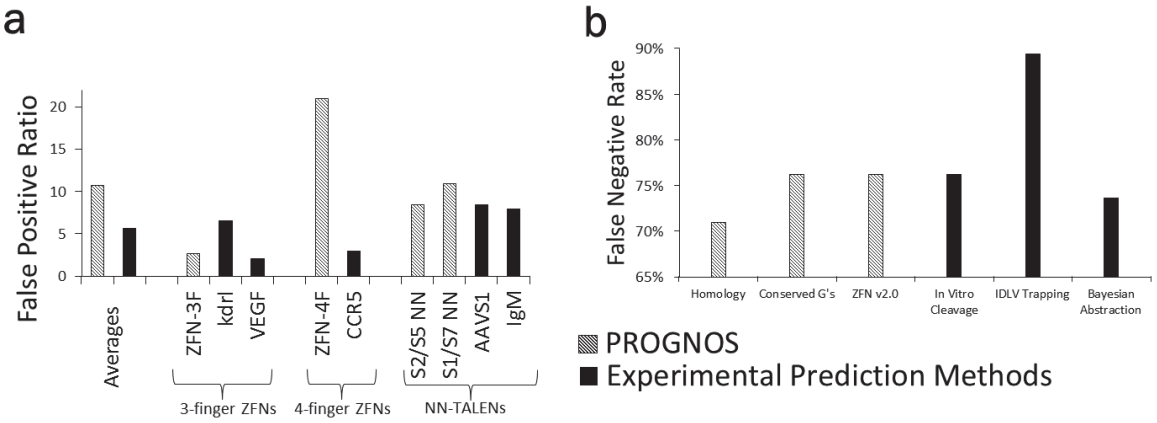

Figure 3-10: Sensitivity and specificity analysis of PROGNOS algorithms. (**a**) Average false positive ratios are shown for the PROGNOS investigation of novel nucleases using the initial algorithms, and for previous experimental prediction methods. Ratios are also shown for individual nucleases in the three different categories of nuclease that have been investigated previously by experimental prediction methods. (**b**) The false negative rates of the different PROGNOS algorithms and previous experimental prediction methods are shown. These were determined by each methods ability to identify the 38 known heterodimeric off-target sites of the *CCR5* ZFNs in their top ranking predictions.

Table 3-5: Comparison of TALENOffer to PROGNOS TALEN v2.0 Algorithm For the full set of TALEN off-target sites, as well as different relevant subsets, the PROGNOS TALEN v2.0 algorithm consistently outperforms the TALENoffer search program in terms of both having lower (better) mean rankings for all off-target sites and a higher (better) number of sites in the Top 16 rankings of the different TALENs. For all TALENs and for both algorithms, the off-target sites in highly homologous, closely related genes were ranked #2.

All Known TALEN Off-Target Sites:		
	Training TALEN v2.0 Ranking	TALENoffer Ranking
Mean Rank of Off-Target Sites	7.8	45.8
# of sites in Top 16 Rankings	12 out of 14	9 out of 14
Excluding Highly Homologous Sites In Closely Related Genes:		
Mean Rank of Off-Target Sites	10.1	65.1
# of sites in Top 16 Rankings	8 out of 10	5 out of 10
Excluding Nucleases in TALEN v2.0 Training Set and Sites in Highly Homologous Genes:		
Mean Rank of Off-Target Sites	12.8	110.4
# of sites in Top 16 Rankings	4 out of 5	2 out of 5

3.5 Discussion

Engineered nucleases can readily be designed and optimized to target specific endogenous sequences in a genome. However, to reach their potential for generating research model systems and treating human diseases, the specificity of engineered nucleases must be better understood. However, the analysis of the location and frequency of TALEN and ZFN off-target effects has been beyond the reach of most laboratories due to the limitations of the existing methods. We created PROGNOS, an online search tool solely based on bioinformatics and the current understanding of nuclease-DNA interactions, which allows users to predict potential nuclease off-target sites by following a simple set of instructions, and to evaluate the sites using standard molecular biology techniques if so desired (Figure 3-11). The novel bioinformatics ranking algorithms in PROGNOS predict many of the off-target sites of the *CCR5* ZFNs that were identified previously using experimental methods and also identified a novel off-target site that was missed in those studies. However, there are several highly active (>5% mutation rate) off-target sites for these ZFNs that PROGNOS did not rank highly, suggesting that there are still unknown factors influencing ZFN off-target activity that are not accounted for in our current models. Future unbiased genome-wide analysis of off-target activity (such as the IDLV method [37]) will be critical to build a larger data base of sites with low sequence homology from which further insight into the factors affecting off-target activity can be gained. Nevertheless, PROGNOS is able to successfully predict many off-target sites and overcomes the drawbacks of the current experimental-based prediction methods that limit the number of nucleases tested, as evidenced by the fact that no bona fide off-target sites for new ZFNs or TALENs have been reported over the last two years (Sept 2011 through Dec 2013) [116, 51]. The improved performance of the refined "ZFN v2.0" and "TALEN v2.0 algorithms over the initial algorithms highlights a key advantage of bioinformatics-based predictions: as more bona fide off-target sites are discovered, increasingly better predictive models can be incorporated.

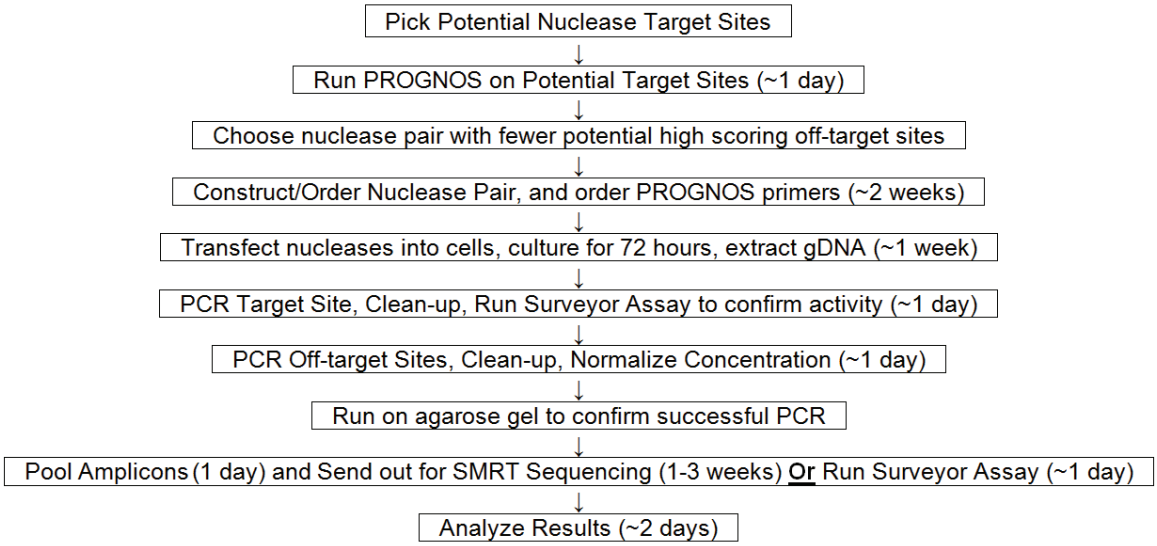

Figure 3-11: Flowchart of PROGNOS-aided search of off-target sites. The entire process of designing a nuclease, confirming on-target activity, and testing off-target activity can be completed with minimal "hands-on" time when aided by PROGNOS. After transfection and genomic DNA harvesting, the off-target search process takes several days of preparation before sequencing, and several days after to process the results.

PROGNOS allowed interrogation and comparison of the off-target activities of several novel nucleases targeting the beta-globin gene. We directly compared 3-finger vs. 4-finger ZFNs that targeted the same site, and compared NK-TALENs vs. NN-TALENs that shared target sites. We found that these NN-TALENs and 3-finger ZFNs had more off-target activity than the corresponding NK-TALENs and 4-finger ZFNs, respectively. While NN-TALENs generally have high on-target cleavage, this may be accompanied by decreased specificity leading to high off-target activity. To confirm the conclusion that the 4F-ZFNs targeting this site are more specific than the 3F versions, we interrogated several of the validated 3F-ZFN off-target sites in cells expressing 4F-ZFNs and found no statistically significant off-target activity. Our comparison of the specificity of NN-TALENs vs. NK-TALENs is somewhat limited by the fact that the NN-TALENs had higher on-target activity than the corresponding NK-TALENs, but the dramatic difference in off-target activity at *HBD* for the S2/S5 NN- and NK-TALENs (Table 3-4) strongly supports the notion that NK-TALENs have improved specificity over NN-TALENs.

Although TALENs seem to be much easier to design and appear less cytotoxic than ZFNs, there still remain concerns about off-target effects. The three previously reported cases of TALEN off-target sites shared only 78% [116], 74% [51], and 72% [51] sequence homology to the intended target site. We observed off-target activity at four additional sites sharing only 81%, 78%, and two cases of 76% homology to the intended target. Given the abundance of sites in a genome that share that level of homology with a TALEN target site, these findings strongly reinforce the need to interrogate these types of genomic loci for possible off-target cleavage.

The lack of discrimination of NN RVDs between guanosine and adenosine is a major concern. We report here, at the *GPR6* off-target site for the S2/S5 TALENs, a case where there are two substitutions of G→A in one half-site and the NN TALENs show off-target activity while the NK TALENs do not. In the *HBD* site for both TALEN pairs, even though there are no G→A substitutions, the NN TALENs show high off-target activity while the NK TALENs show none at all. It is likely then, that NN may increase off-target activity of the TALENs through alternative means; perhaps the strong affinity of the NN RVD can compensate for mismatched bases in other positions in the TALEN binding site. Although NN imparts higher activity than NK, we caution its use due to increased off-target effects.

Previously, TALENs had been shown to have strong preference for a pyrimidine in the 5' position of the binding site. We report here, the first cases of off-target activity at sites lacking a 5' pyrimidine. Both of these were homodimeric sites of the S1 TALEN where an adenosine was in the 5' position of one of the half-sites. It remains to be determined whether this is a general feature of all TALENs, or if somehow specific to a characteristic of the S1 TALEN.

We employed wild-type FokI domains that allow homodimerization of nucleases in order to see a broader picture of nuclease off-target activity. We found that in several cases, homodimers of one nuclease showed nearly equal, or greater activity at an off-target site compared to the activity of the heterodimer at the target site. While research into the

fundamental basis of nuclease off-target effects should continue to look at homodimeric interactions, we would strongly recommend use of hetero-obligate FokI architectures [30] for use in nucleases for targeted genome editing in order to reduce the off-target effects.

TALENs using the +63 C-terminal truncation (as were used here) have been shown to cleave over a wide range of spacers [18]. This makes design of TALENs easier and increases the number of potential sequences that can be targeted, but it also increases the number of potential regions of the genome that could be cleaved through off-target activity. At the *GPR6* off-target site for the S2 homodimer, we observed off-target activity with a 30 bp spacer—the longest reported for an off-target site. Reducing the range of spacers that TALENs can cleave at—potentially through use of shorter C-terminal truncations [18]—has the potential to greatly reduce TALEN off-target effects.

3.6 Conclusion

In summary, PROGNOS provides a user-friendly, web-based tool for rapid identification of potential nuclease off-target cleavage sites that can be evaluated using standard molecular biology techniques. The bioinformatics-based ranking algorithms in PROGNOS identify most nuclease off-target cleavage sites found by existing experimental methods. PROGNOS has relatively low false positive ratios and comparable false negative rates to experimental-based predictions, making it a robust method that can be readily implemented by most laboratories. Screening potential target sites using PROGNOS can facilitate the selection of better nuclease target sites that minimize the number of likely genomic off-target sites. PROGNOS allows nuclease off-target analysis to become a routine component of nuclease design and testing, facilitating the discovery of new off-target sites for ZFNs and TALENs, which expand the off-target database and may improve the PROGNOS algorithms. These capabilities give PROGNOS the potential to help expand and expedite the application of engineered nucleases for a wide range of biological and medical applications.

CHAPTER IV

USING PROGNOS TO EXPAND THE DATASET OF NUCLEASES WITH KNOWN OFF-TARGET SITES

4.1 Introduction

A major advantage of *in silico* predictive models is that they can be iteratively refined and improved as more data become available. Although the algorithms developed in Chapter 3 successfully located several bona fide off-target sites, the data available at the time were fairly limited; off-target sites were known for only three different ZFNs and two different TALENs (for a total of only 3 TALEN off-target sites) in 2011 when the study was initiated. By applying the PROGNOS algorithms to many different ZFNs and TALENs, the training set of known off-target sites was greatly expanded to allow for more sophisticated machine learning techniques that reveal underlying patterns and trends that can better differentiate between bona fide off-target sites and other sites with homology to the intended target of the nuclease. This work moves the field closer to the ultimate goal of off-target prediction algorithms that allow for screening of potential nuclease target sites in order to design, with reasonable confidence, a nuclease with no likely off-target sites in the genome of interest. In addition to locating new off-target sites, the following studies also revealed information about off-target activity related to different delivery methods, nuclease types, cell types, and doses.

4.2 Methods

Genomic DNA samples were provided to me by various collaborators. The genome of interest was scanned using the PROGNOS algorithm [34] and potential off-target sites were rank-ordered. Using primers designed by the PROGNOS algorithm, PCR reactions were

performed and amplicons were sequenced and analyzed exactly as previously described (Section 3.3.7.2, Section 3.3.8.1, and Section 3.3.8.2).

4.3 Results

4.3.1 Lentiviral delivery of ZFNs into mouse SC-1 cells[1]

Delivery of nucleases into primary cells is a challenge in the field. Packaging nucleases into lentiviral vectors allows for efficient delivery. We tested the on-target and off-target activity of ZFNs delivered into murine SC-1 embryonic fibroblasts targeting the "safe-harbor" *Rosa26* locus. After analyzing the top 24 sites predicted by the PROGNOS ZFNv2.0 algorithm [34], robust activity was observed at the target site and modest off-target activity at the locus near *Lgals3* (Table 4-1).

In addition, collaborators also had lentivirus encoding the CCR5-ZFNs [89] available. Although the mouse homolog to *CCR5* contains several mismatches within the binding site of the ZFNs designed for the human gene, these ZFNs were also interrogated for off-target activity after lentiviral transduction (Table 4-1).

4.3.2 Lentiviral delivery of ZFNs into rat cells[2]

4.3.2.1 Abstract

ZFNs are promising tools for genome editing for biotechnological as well as therapeutic purposes, however delivery remains a major issue impeding targeted genome modification. Lentiviral vectors are highly efficient for delivering transgenes into cell lines, primary cells and into organs, such as the liver. However, the reverse transcription of lentiviral vectors leads to recombination of homologous sequences, as found between and within ZFN monomers. We used a codon swapping strategy to both drastically disrupt sequence identity between ZFN monomers and to reduce sequence repeats within a monomer sequence. We

[1]Collaboration with Kerstin Schmidt and Dr. Dorothee von Laer, Division of Biology, Innsbruck Medical University, Austria

[2]Modified from: Abarrategui-Pontes C, Creneguy A, Thinard R, **Fine EJ** et al. (2014). Codon Swapping of Zinc Finger Nucleases Confers Expression in Primary Cells and In Vivo from a Single Lentiviral Vector. *Human Gene Therapy* [1]

Table 4-1: *Rosa26* and *CCR5* ZFN off-target activity at sites ranked by PROGNOS. 23 sites from throughout the top-ranked cleavage sites of the *Rosa26* and *CCR5* ZFNs, as determined by the PROGNOS 'ZFN v2.0' algorithm, were evaluated using SMRT sequencing in transduced SC-1 cells. The site in the *Eif4g3* locus was chosen based on a 'rational design' analysis of the zinc finger binding helices. The 'match type' indicates the orientation of the left (L) and right (R) nucleases at the site and the length of the spacer sequence. Site sequences are listed as 5'–(+) half-site–spacer–(-) half-site–3'. Therefore, the (+) positive half-site is listed in the reverse anti-sense orientation as compared to the DNA sequence that the ZFN binds. In sequences, lower-case letters indicate mismatches as compared to the target site. "SC-1 cells modification frequency" is the percentage of observed sequences showing evidence of NHEJ events. For all sites shown, significantly higher ($p < 0.05$) frequencies of indels were observed in transduced cells as compared to mock treated cells.

ZFN	Closest Gene	Match Type	Mutations per half-site (+)	Mutations per half-site (-)	(+) half-site	(-) half-site	ZFN v2.0 Rankings	SC-1 Cells Modification Frequency
Rosa26	Gt(ROSA)26Sor	L-6-R	0	0	GACTCCCGCCCA	AGAAAGACTGGAGTTGCA	1	29.05%
	Lgals3	L-6-R	5	4	ctCcCCCcaCCA	ccAAAGACTGttGTTGCA	20	0.07%
						Total Sites Interrogated:		23
CCR5	Gm2176	L-6-R	1	2	GTCAgCCTCATC	AAACTGaAAAtG	2	0.46%
	Syt17	L-5-R	3	1	GagATCCaCATC	AAACTGCAAgAG	7	3.87%
	Ccr5	L-5-R	1	2	GTCtTCCTCATC	AAgCTGCAAAAa	10	0.06%
	Ppfia2	L-5-R	2	2	tTCAgCCTCATC	AAAacGCAAAAG	23	0.45%
	C530044C16Rik	L-5-R	3	1	GTCAggCTCAgC	AAACTGCAAAgG	26	0.72%
	BC031181	L-5-R	3	2	aagATCCTCATC	AAACaGgAAAAG	46	13.29%
	Mamdc2	L-5-R	3	1	tTCAgtCTCATC	tAACTGCAAAAG	54	0.95%
	Eif4g3	L-5-R	5	1	caCtTaCTCcTC	AAACTGgAAAAG	3152	18.47%
						Total Sites Interrogated:		23

constructed lentiviral vectors encoding codon-swapped ZFNs or unmodified ZFNs from a single mRNA transcript. We reduced total identity between ZFN monomers from 90.9% to 61.4% and showed that a single lentivirus allowed efficient expression of functional ZFNs targeting the rat *UGT1A1* gene after codon-swapping, leading to much higher ZFN activity in cell lines (up to 7-fold increase compared to unmodified ZFNs and 60% activity in C6 cells), as compared to plasmid transfection or a single lentivirus encoding unmodified ZFN monomers. Off-target analysis located several active sites for the 5-finger *UGT1A1*-ZFNs.

4.3.2.2 Methods

C6 cells ($5 * 10^4$) were seeded in each well of a 6-well plate one day before transduction. On the day of transduction, the medium was changed before addition of the viral particles. Cells were kept at 37°C for four days before genomic DNA was harvested and subjected to off-target analysis.

4.3.2.3 Results

The high levels of targeted DSBs in the *UGT1A1* gene observed after C6 cell transduction with codon-swapped lentivirus gave the opportunity to look for potential off-target effects, which may not be easily detectable with a low ZFN activity, as observed after transfection. Off-target sites were chosen for investigation based on a mixture of the top-ranked sites by the PROGNOS method [34]. Using SMRT sequencing, we interrogated 22 genomic loci that were highly ranked for cleavage in the rat genome by the *Ugt1A1* ZFNs (the intended target site and 21 potential "off-target" sites) in transduced C6 cells. High levels of modification (~64.4%) at the *UGT1A1* intended target was confirmed and we also detected off-target activity at 6 additional loci (Table 4-2). Off-target sites were found in an intronic region of the *Tarsl2* gene at a frequency of ~10% and in five other sites at low frequency. Of note, *Tarsl2* and *Gbas* off-target sites share only 80% sequence identity with the target site.

Table 4-2: *UGT1A1* ZFN off-target activity at sites ranked by PROGNOS. The 22 top-ranked cleavage sites of the *UGT1A1* ZFNs, as determined by the PROGNOS 'homology' and 'Conserved G's' algorithms, were evaluated using SMRT sequencing in transduced C6 cells. The 'match type' indicates the orientation of the left (L) and right (R) nucleases at the site and the length of the spacer sequence. Site sequences are listed as 5'–(+) half-site–spacer–(-) half-site–3'. Therefore, the (+) positive half-site is listed in the reverse anti-sense orientation as compared to the DNA sequence that the ZFN binds. In sequences, lower-case letters indicate mismatches as compared to the target site. "C6 cell line modification frequency" is the percentage of observed sequences showing evidence of NHEJ events. For all sites shown, significantly higher ($p < 0.05$) frequencies of indels were observed in transduced cells as compared to mock treated cells. Our analysis revealed that the locus near the *Ftcd* gene contains a homozygous C→T SNP at the second position of the (-) half-site, introducing an additional mutation in that ZFN binding site relative to the reference genome.

		Mutations per half-site				PROGNOS Rankings		C6 Cell Line Modification
Closest Gene	Match Type	(+)	(-)	(+) half-site	(-) half-site	Homology	Conserved G's	Frequency
Ugt1a1	L-5-R	0	0	CTCCGGTTCCCATGG	ATGAAGGAATATGCA	1	1	64.469%
Tarsl2	L-5-R	3	3	CTCCtaTcCCCATGG	ATGAAGGAAgcTGCt	9	2	10.464%
Gbas	L-6-R	2	4	CTCaGGTTCCCgTGG	ATGgAGGAATATctt	4	22	1.282%
Il1rapl1	L-5-R	5	2	aTCCGGTgCCCtTtt	cTGcAGGAATATGCA	29	9	0.229%
Myo16	R-6-L	5	3	TtCAatgcCCTTCAT	CCtTGGGtACtGGAG	205	12	0.157%
Ssr1	R-6-L	2	4	TGCATATTCCTTgcT	gCATGGGAcCaGaAG	3	20	0.117%
Ftcd	R-6-L	5	4	gtCAcAaTCCaTCAT	atgTGGGAACaGGAG	240	13	0.051%
							Total Sites Interrogated:	22

4.3.2.4 Discussion

When nucleases cut at other locations in the genome other than their intended target, they can potentially induce unwanted gene disruption, destabilization of the cell's genome, or transformation of the cell into a cancerous pheno- type. Therefore, it is important to reveal potential off-target sites of a given pair of ZFNs. A pre-requisite for detecting off-target sites is to induce sufficient on-target activity. Because of high ZFN activity in C6 cells, we could investigate off-target effects using the PROGNOS algorithm in these cells [34]. Given that only 21 potential off-target sites were interrogated, finding six bona fide locations of off-target activity, validated as having a statistically significant mutation frequency greater than untreated cells, was higher than expected. Based on off-target studies of other 3- and 4-finger ZFNs, an analysis of this size would be expected to yield perhaps four bona fide off-target sites [34, 37, 44, 88, 101]. Moreover, this study is the first report of bona fide off-target activity discovered for 5-finger ZFNs with heterodimeric FokI domains. In our study, the levels of on-target activity after lentiviral transduction were much higher than in previous studies [34]. Thus, it is probable that conducting these studies by delivering ZFNs using a single lentiviral vector and our codon swapping approach revealed off-target sites that wouldn't have been detected otherwise with an approach that yielded lower on-target activity. Our data highlight the impact of the nuclease design and of the efficacy of delivery methods on off-target effects and thus on the biosafety of artificial nucleases for gene therapy purposes.

4.3.3 Dosing experiments with GFP ZFNs[3]

The pair of ZFNs targeting a sequence in the green fluorescent protein (GFP) have been used in many previous studies [90], but their off-target activity had never been characterized. The GFP-ZFNs were transfected into a special line of HEK-293T cells with an integrated copy of the GFP gene. Several different doses were transfected to examine the

[3]Collaboration with Dr. Zhong Chen and Dr. Steffen Meiler, Medical College of Georgia

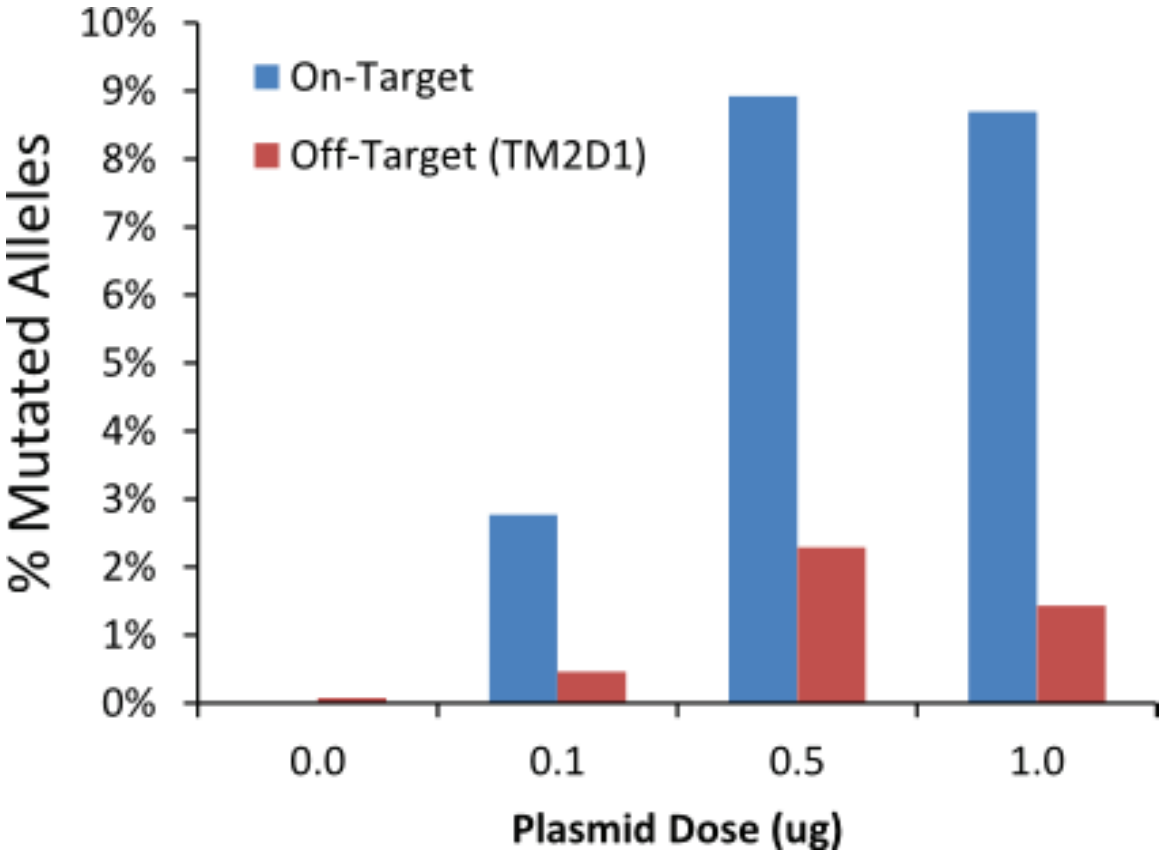

Figure 4-1: GFP-ZFN Dose Response

dose-response and to determine if a "sweet-spot" of the on-target:off-target activity ratio could be found. We interrogated the top 23 sites ranked by the PROGNOS 'Homology' and 'Conserved G's' algorithms [34] and found one bona fide off-target site near the *TM2D1* locus resulting from homodimeric binding of the left zinc finger to a sequence with a single A→G mismatch in the 2nd position of the right half-site. Although a dose-dependent response for both the on-target and off-target sites was observed, the on-target:off-target specificity ratio did not appreciably vary (Figure 4-1).

4.3.4 TALENs facilitate targeted genome editing in human cells with high specificity and low cytotoxicity[4]

4.3.4.1 Abstract

Designer nucleases have been successfully employed to modify the genomes of various model organisms and human cell types. While the specificity of ZFNs and RGENs has been assessed to some extent, little data are available for TALENs. Here, we have engineered TALEN pairs targeting two human loci (*CCR5* and *AAVS1*) and performed a detailed analysis of their activity, toxicity and specificity. The TALENs showed comparable activity to benchmark ZFNs, with allelic gene disruption frequencies of 15-30% in human cells. Notably, TALEN expression was overall marked by a low cytotoxicity and the absence of cell cycle aberrations. Bioinformatics-based analysis of designer nuclease specificity confirmed fairly substantial off-target activity of ZFNs targeting *CCR5* and *AAVS1* at six known and five novel sites, respectively. In contrast, only marginal off-target cleavage activity was detected at four out of 49 predicted off-target sites for *CCR5*- and *AAVS1*-specific TALENs. The rational design of a *CCR5*-specific TALEN pair decreased off-target activity at the closely related CCR2 locus considerably, consistent with fewer genomic rearrangements between the two loci. In conclusion, our results link nuclease-associated toxicity to off-target cleavage activity and corroborate TALENs as a highly specific platform for future clinical translation.

4.3.4.2 Methods

Target sites at the two chosen human loci where identified in proximity of benchmark ZFN targets without using previously published algorithms. TALE-based DNA binding domains were assembled using a previously described Golden Gate assembly kit [14] that was modified with four different Level 3 destination vectors to express functional nuclease monomers

[4]Modified from: Mussolino C, Alzubi J, **Fine EJ** et al. (2014). TALENs facilitate targeted genome editing in human cells with high specificity and low cytotoxicity. *Nucleic Acids Research* [79]

based on our previously optimized TALEN scaffold [80] (Δ135/+17). These vectors included the 17.5th repeat and the wild-type FokI cleavage domain (pVAX_CMV_TALshuttle(xx); 'xx' stands for the four different 17.5th RVDs used, NI, NG, HD and NN). The ZFN expression plasmids were generated by subcloning previously published zinc-finger arrays (codon-optimized and synthesized by GeneArt/Life Technologies, Regensburg) into the pRK5.N backbone [3], which includes an N-terminal HA tag and the SV40 nuclear localization domain and either of the obligate heterodimeric FokI variant KV or EA [114] at the C-terminus.

HEK-293T cells were cultured in Dulbeccos modified Eagles medium (DMEM, PAA) supplemented with 10% Fetal Bovine Serum (FBS, PAA), 100 U/ml penicillin (PAA) and 100 µg/ml streptomycin (PAA). Cells were transfected with polyethylenimin (PEI) as previously described [80]. Genomic DNA was extracted from HEK-293T cells 3 days post-transfection with ZFNC, ZFNA, TC06, TC-NC, TA04 or empty vector. Potential off-target sites for the TALENs were chosen from the PROGNOS RVD-5TC and Homology-5TC rankings to contain a mixture of the top-ranked sites from both algorithms [34]. In addition, *CCR2* was also interrogated for the TC-NC TALENs to facilitate comparison at that off-target site, even though this site did not appear in the 5TC PROGNOS rankings because of the lack of the 5' pyrimidine in one of the half-sites. The 'specificity factor' in Tables 4-3 and 4-4 is calculated as the ratio of on-target to off-target mutagenesis frequency based on the detected indels events.

4.3.4.3 Results

We investigated the specificity of our *CCR5*- and *AAVS1*-specific designer nucleases by high-throughput sequencing at potential off-target sites predicted *in silico* using the recently developed PROGNOS software [34]. When applying the bioinformatics tool, four of seven previously validated off-target sites (Table 4-3; [37, 88]) were listed in the top 32 off-target sites for the *CCR5*-specific ZFN For the *AAVS1*-specific ZFN, few specificity data are

available in the literature. PROGNOS confirmed the prediction of two out of five previously identified but not validated [50] off-target sites in the top 23 list of predicted off-target sites. We also used PROGNOS to predict the top 15 off-target sites for TC06 and TC-NC *CCR5*-specific TALENs, respectively, and the top 22 off-target sites for the *AAVS1*-specific TALEN TA04.

For the *CCR5*-specific ZFN, we confined high-throughput sequencing analysis to those sites that were previously identified and we confirmed off-target cleavage activity at six out of seven sites analyzed, sometimes with disruption activities over 12%, i.e. almost as high as at the *CCR5* locus. Moreover, PROGNOS was able to predict a novel off-target site on chromosome 5 [34]. Analysis of off-target cleavage of the *AAVS1*-specific ZFN pair revealed activity of up to 4.4% at 5 out of the 23 predicted off-target sites (Table 4-4). Hence, we validated three previously predicted off-target sites and identified and validated two additional off-target sites for the *AAVS1*-specific ZFN. In addition to *CCR2*, the two *CCR5*-specific TALEN pairs exhibited minor off-target cleavage activity (0.12%) at a total of three off-target sites. The *AAVS1*-specific TALEN pair revealed statistically significant activity (0.13%) at 1 out of the 22 predicted off-target sites (Table 4-4).

The majority of mutations retrieved at the analyzed loci consisted of short deletions in the respective spacer sequences, as previously reported [62], and DNA repair seemed to be driven mostly by microhomology-mediated repair.

In summary, the high-throughput sequencing results exposed important information regarding the specificity of the assessed nucleases. In particular, compared to the benchmark ZFN, our *CCR5*-specific TALENs are better tolerated when expressed in human cells and can be designed to discriminate between highly identical sequences, such as *CCR5* and *CCR2*. The *CCR5*:*CCR2* targeting ratio was determined to be in the range of 3:1 for the ZFN, and 130:1 or 7:1 for the TALEN pairs TC-NC and TC06, respectively. The calculated 'specificity factor', indicating the ratio of the frequency of on-target mutagenesis to the overall frequency of off-target mutagenesis, was 1:2 for the *CCR5*-specific ZFN, 60:1

Table 4-3: Off-target sites of *CCR5*-specific designer nucleases. In addition to *CCR5* and *CCR2*, the top 14 PROGNOS off-target loci for each of the TALEN pairs and several previously identified loci for the *CCR5* ZFNs were interrogated with SMRT sequencing. The 'match type' indicates the orientation of the left (L) and right (R) nucleases at the site and the length of the spacer sequence. Site sequences are listed as 5'–(+) half-site–spacer–(-) half-site–3'. Therefore, the (-) half-site for TALENs and the (+) half-site for ZFNs are listed in the reverse anti-sense orientation compared to the DNA sequence that the nuclease binds. Lowercase red letters indicate mismatches as compared to the target site. PCR amplification failed for sites listed as "N/A". For some nucleases, two biological replicates ('Rep #1' and 'Rep #2') were analyzed. The 'specificity factor' is calculated as the ratio of on-target to total observed off-target mutation frequencies. * indicates that significantly higher ($p < 0.05$) frequencies of indels were observed in nuclease-treated cells as compared to mock treated cells.

Nuclease ID	Closest Gene	Match Type	Mutations per half-site (+)	Mutations per half-site (-)	(+) half-site	(-) half-site	Modification Frequency Rep #1	Modification Frequency Rep #2	Specificity Factor
ZFNC	CCR5	L-5-R	0	0	GTCATCCTCATC	AAACTGCAAAAG	27.89% *	14.40% *	0.48
	CCR2	L-5-R	1	1	GTCgTCCTCATC	AAACTGCAAAAa	9.88% *	7.95% *	
	KRR1	L-5-R	2	2	GgCcTCCTCATC	AAACTGgAAAtG		N/A	
	KCNB2	L-5-R	3	1	aTgtTCCTCATC	AAACTGCAAAtG		1.10% *	
	BTBD10	L-5-R	2	1	GTttTCCTCATC	AAACTGCAAAAt		6.33% *	
	TACR3	L-5-R	1	3	GTCATCtTCATC	AAACTGtAAAgt		12.64% *	
	KDM2A	R-5-L	5	2	tgagaCCTCATC	AAACTGtAAtAG		8.58% *	
	WBSCR17	R-6-L	2	2	GTtATCCTCAgC	AAACTGgAAcAG		2.70% *	
						Total Sites Interrogated:	2	8	
TC-NC	CCR5	L-15-R	0	0	TGTGGGCAACATGCTGGTC	AACTGCAAAAGGCTGAAGA	11.43% *	7.85% *	59.6
	CCR2	L-15-R	0	2	TGTGGGCAACATGCTGGTC	AACTGCAAAAaGCTGAAGt	0.07% *	0.07% *	
	NRXN1	R-15-R	4	5	TaTTCAGCaaaTTGCAGTT	cACTGtActAGGtTGAAGA	0.05%	0.12% *	
						Total Sites Interrogated:	16	16	
TC06	CCR5	L-14-R	0	0	TTTGTGGGCAACATGCTGG	ATAAACTGCAAAAGGCTGA	21.53% *		6.51
	CCR2	L-14-R	0	1	TTTGTGGGCAACATGCTGG	ATAAACTGCAAAAaGCTGA	3.08% *		
	TBL1X	L-18-L	5	5	TaaGTaGGCAACATcCTGt	CtAtCtTcTTGCCCAaAAA	0.12% *		
	TRMT44	L-22-L	5	3	TTTGaGGGggAaATGCTtG	CCAGCgTGgTGCCCACAcA	0.12% *		
						Total Sites Interrogated:	16		

for TALEN TC-NC and 7:1 for TALEN TC06 (Table 1). A similar trend was observed for the *AAVS1*- specific nucleases. While we identified and validated five novel off-target sites for the *AAVS1*-specific benchmark ZFN, with a specificity factor of 1:2, we detected off-target cleavage activity at a single predicted site for TALEN TA04, with a calculated specificity factor of 27:1 (Table 4-4).

4.3.4.4 Discussion

Nuclease specificity is a key factor for advancing targeted genome engineering into the clinic. Here we show that expression of TALENs is generally well tolerated by human cells. In contrast, expression of our ZFN pairs reduced cell survival, was associated with

Table 4-4: Off-target sites of *AAVS1*-specific designer nucleases. In addition to several previously predicted (but unconfirmed) off-target sites for the *AAVS1* ZFNs [50], the top PROGNOS off-target loci for the *AAVS1* TALENs and ZFNs were interrogated with SMRT sequencing. The 'match type' indicates the orientation of the left (L) and right (R) nucleases at the site and the length of the spacer sequence. Site sequences are listed as 5'–(+) half-site–spacer–(-) half-site–3'. Therefore, the (-) half-site for TALENs and the (+) half-site for ZFNs are listed in the reverse anti-sense orientation compared to the DNA sequence that the nuclease binds. Lowercase red letters indicate mismatches as compared to the target site. The 'specificity factor' is calculated as the ratio of on-target to total observed off-target mutation frequencies. All sites shown had significantly higher ($p < 0.05$) frequencies of indels were observed in nuclease-treated cells as compared to mock treated cells.

Nuclease ID	Closest Gene	Match Type	Mutations per half-site (+)	Mutations per half-site (-)	(+) half-site	(-) half-site	Modification Frequency	Specificity Factor
ZFNA	AAVS1	L-6-R	0	0	ACCCCACAGTGG	TAGGGACAGGAT	9.09%	0.55
	CHRAC1	R-6-L	2	3	caCCCACAGTGG	ctGGGACAGGAg	2.95%	
	ATRNL1	L-5-R	5	0	caCCCACAGatt	TAGGGACAGGAT	3.71%	
	BEGAIN	R-6-L	2	2	cCCCCACtGTGG	cAGGGACAGGAc	4.40%	
	LINC00548	L-6-R	1	1	ACCCCACAGTaG	TAGGGACAGGAa	4.15%	
	H19	L-5-R	1	3	tCCCCACAGTGG	gAGGGcCAGGAg	1.28%	
						Total Sites Interrogated:	24	
TA04	AAVS1	R-15-L	0	0	TCTGTCCCCTCCACCCCAC	GACAGGATTGGTGACAGAA	12.57%	26.9
	CPN1	L-11-R	5	5	cCacTCCCtcCCACCCCAC	aAtAGGATTGGgGgCAGgA	0.13%	
						Total Sites Interrogated:	23	

cell cycle arrest and increased cell death, all of which suggest surplus DNA cleavage at off-target sites. Using the newly developed PROGNOS bioinformatics tool [34], we predicted potential ZFN and TALEN off-target sites in the human genome and screened them by high-throughput sequencing. Importantly, the top 32 ranked potential off-target sites of the *CCR5*-specific ZFN included four off-target sites in the top seven that were previously verified experimentally [37, 88] and one novel site [34]. The top 23 predicted off-target sites for the *AAVS1*-specific ZFN included two previously predicted sites and allowed us to verify a total of five novel off-target sites. PROGNOS was also successfully employed to predict four novel off-target sites for TALENs targeting beta-globin [34]. These results clearly validate and underline the high predictive value of PROGNOS that can be reliably applied to predict both ZFN and TALEN off-target sites.

In our study we have profiled TALEN off-target activity in a non-clonal cell population

with high-throughput sequencing. Many previous reports have either focused on identifying the integration site of non-integrating viral vectors trapped in TALEN-induced DSBs [84] or using Surveyor/T7E1 assays to monitor off-target activity at genomic loci predicted by bioinformatics [113]. However, the small number of off-target sites identified and analyzed or the use of genomic DNA extracted from clonally derived cell populations or live animals [113, 85] may likely not provide enough depth to identify rare mutagenic events at off-target sites as those identified here. The five novel off-target sites of the widely used *AAVS1*-specific ZFN are certainly of importance to those researchers currently using or planning to use this nuclease pair for targeted integration at the *AAVS1* 'safe harbor' locus. Given the superior toxicity profile and the absence of significant off-target activity, our *AAVS1*-specific TALEN pair TA04 might be a valid alternative for targeted gene addition into the AAVS1 'safe harbor' site, in particular in primary cells.

Due to the high sequence similarity with *CCR5*, assessment of nuclease specificity confirmed that the major off-target sites of our *CCR5*-specific ZFN and TALEN pairs were found in the *CCR2* locus. As reported before, concomitant cleavage at CCR2 and CCR5 results in deletion of the intervening sequence [21, 65]. We show that nucleases, which are not specific enough to discriminate between these two highly similar loci (i.e. ZFNC and TC06), induce deleterious genomic rearrangements, including deletions and inversions. Even though not investigated further, we envision that simultaneous off-target activity on two different chromosomes can induce translocations between two major off-target sites, as previously shown [12]. Importantly, we show that such deleterious genotoxic events can be restrained by using rationally designed nucleases, such as TALEN TC-NC, which shows an unprecedented *CCR5*:*CCR2* targeting ratio of ~130:1. We speculate that the ability of TALEN TC-NC in discriminating between *CCR5* and *CCR2* is based on the fact that one of the two TC-NC subunits targets a T in position 0, which is not present in *CCR2*. Thus, the 5'-T can be used as a major discriminant between highly similar off-target sites to overcome cytotoxic side effects.

Examination of the few off-target sites shows that TALENs tolerate up to five mismatches in their 19-bp target half-sites. Whether the position of these mismatches has an impact on the overall binding affinity of TALEN monomers has not been addressed here. In any case, a systematic analysis of TALEN off-target activity may provide novel design guidelines that can be taken into consideration in the future to avoid potential off-target activity at sites with more than 74% nucleotide identity to the intended target site. Our sequencing data revealed some off-target mutagenic events at predicted homodimeric TALEN off-target sites for our *CCR5*-specific TALENs. Thus, TALEN specificity can be further increased by coupling the TALE DNA binding domains to obligate heterodimeric FokI endonuclease variants [30, 108] that have been shown to abolish cleavage of ZFNs at homodimeric off-target sites [37]. Even though most of our findings are based on TALENs with the Δ135/+17 scaffold developed in the Cathomen lab [80], we believe that they can be extended to alternative TALEN designs with longer linkers [77]. Importantly, as shown for ZFNs [47], our short 17-residue linker may improve specificity by limiting spacer tolerability. Additionally, intranuclear concentration of the designer nucleases and duration of the expression are important parameters affecting cleavage specificity. For instance, delivery of a *CCR5*-specific ZFN pair in the form of proteins resulted in short persistence in the transduced cells and was associated with reduced off-target activity at *CCR2* [39].

We systematically measured higher off-target activity of the *CCR5*-specific ZFNs than previously published [37, 88]. We reason that this is most likely due to the cell line (HEK-293T versus K562 cells) and the transfection method (PEI versus nucleofection) that we used, which may result in higher intranuclear nuclease expression levels. Also, our ZFNs contain a slightly different obligate heterodimeric FokI domain [114] compared to the most commonly used one [76] and a different epitope tag (HA versus FLAG). Given that we were able to identify previously reported off-target sites at much higher frequencies demonstrates that PEI-mediated transfection of HEK-293T cells represents an ideal 'mutation-prone system' to evaluate nuclease specificity and further emphasizes the high specificity of the

TALENs tested here.

Taken together, our results establish TALENs as a safe and specific platform to edit the human genome, paving the way for their use for clinical translation. As mentioned previously, RGENs have been established as a versatile tool to modify the human genome [73, 19]. Although there has been some concern regarding their specificity [35, 87, 53, 21], it will be interesting to see whether this novel class of designer nucleases can match the high specificity set by TALENs, e.g. as recently shown by truncating the gRNA molecule [36].

4.3.5 Functional Gene Correction of IL2RG by TALEN-mediated Genome Editing[5]

4.3.5.1 Abstract

Gene editing with engineered nucleases allows for precise and complex changes to be made to the human genome, but translation of this technology for the treatment of human disease is complicated by the diversity of disease-causing mutations in patient populations. Here we show that nuclease-mediated targeting of *IL2RG* cDNA to *IL2RG* Exon 1 provides endogenous control over transgene expression, functionally correcting the *IL2RG* gene. Using TALENs, we achieved high frequencies of *IL2RG* gene replacement in cell lines (up to 20%) and $CD34^+$ hematopoietic stem/progenitor cells (up to 2.9%). Targeted addition of wild-type *IL2RG* cDNA to the endogenous *IL2RG* promoter showed decreased levels of IL2Rγ activity, but modification of the cDNA with codon-optimization and the introduction of an artificial intron increased IL2Rγ activity to wild-type levels. Off-target activity was analyzed in 293T, K562, and CD34 cells. This strategy has potential clinical application for multiple genetic diseases by providing a functional gene in patients with essentially any disease-causing mutation while maintaining endogenous gene expression patterns.

[5]Modified from: Kildebeck EJ, Clark JT, Hendel A, **Fine EJ**, Bao G, Porteus MH (manuscript in preparation). Functional Gene Correction of IL2RG by TALEN-mediated Genome Editing

Table 4-5: Off-target site identification in 293T cells. The top PROGNOS off-target loci for the *IL2RG* TALENs were interrogated with SMRT sequencing. The 'match type' indicates the orientation of the left (L) and right (R) nucleases at the site and the length of the spacer sequence. Site sequences are listed as 5'–(+) half-site–spacer–(-) half-site–3'. Therefore, the (-) half-site for TALENs is listed in the reverse anti-sense orientation compared to the DNA sequence that the nuclease binds. Lowercase red letters indicate mismatches as compared to the target site. All sites shown had significantly higher ($p < 0.05$) frequencies of indels were observed in nuclease-treated cells as compared to mock treated cells.

		Mutations per half-site					
Closest Gene	Match Type	(+)	(-)	(+) half-site	(-) half-site	Modification Frequency	TALEN v2.0 Ranking
IL2RG	R-14-L	0	0	TAATGATGGCTTCAACAT	ATTCCCTGGGTGTA	18.18%	1
TNN	L-15-L	2	3	TACACCCAtGaAAT	ATgCCtTGGtTGTA	9.08%	2
HDDC2	L-25-L	1	4	TACACtCAGGGAAT	ATTCtCTGGagGcA	0.62%	4
TMEM182	L-14-L	1	4	TACACCCAGGtAAT	AcTCCCTGaaTtTA	0.71%	9
DEK	L-11-L	4	1	TACAgCtgGGGAAa	ATTgCCTGGGTGTA	0.90%	12
GPD2	L-13-L	3	4	TACACCaAaGaAAT	ATTttCTGGtTGTt	2.52%	23
SLC19A1	L-13-L	1	4	aACACCCAGGGAAT	ATTCCtTGGGTcgc	0.70%	24
CCDC171	L-30-L	3	3	acCACCCAGaGAAT	ATgCCtTGGtTGTA	0.58%	25
CSNK1G3	L-17-R	3	2	TAggCCCAGGGAgT	gaGTTGAAGCCATCATTA	3.73%	26
					Total Sites Interrogated:		24

4.3.5.2 Results

Because HEK-293T cells have previously been shown to be a mutation-prone cell line that can be used as a model system to screen for off-target activity [79], we first analyzed genomic DNA from HEK-293T cells transfected with TALEN plasmids. The top 24 sites predicted by the PROGNOS TALEN v2.0 algorithm [34] were analyzed and bona fide off-target activity was observed at 8 sites (Table 4-5). Having determined that the TALENs did induce off-target activity in the 293T model system, we began to investigate methods to reduce off-target activity and the relative off-target mutagenesis tendencies of different cell lines.

Since gene repair through homologous recombination was the ultimate goal of our approach, we next examined off-target activity in the K562 cell line—a model system for homologous recombination. Furthermore, because delivery of nuclease protein has been previously linked to lower off-target activity compared to expression from plasmids [39], we sought to also examine the effects of delivery as mRNA transcripts compared to DNA

Table 4-6: Off-target activity in K562 cells. The top PROGNOS off-target loci for the *IL2RG* TALENs were interrogated with SMRT sequencing. Consistent activity across all replicates was observed at the locus near *TNN*. On-target activity was quantified as the percentage of sequences modified through either NHEJ or HDR. 'NHEJ' sequencing reads are the number of observed sequences containing evidence of indels in the TALEN target site.

Nucleic Acid	Replicate	On-Target:Off-Target Ratio	*TNN* Locus Modification Frequency	Sequencing Reads	
				NHEJ	Total
Plasmid	1	57:1	0.36%	13	3614
Plasmid	2	52:1	0.32%	7	2161
Plasmid	3	38:1	0.41%	10	2443
mRNA	1	36:1	1.33%	18	1350
mRNA	2	22:1	1.46%	18	1230
mRNA	3	47:1	1.08%	60	5540

plasmids. We therefore nucleofected K562 cells with donor DNA to act as a template for homologous recombination, and either mRNA or plasmids encoding the TALENs. In all samples tested, the off-target site near the *TNN* locus, which was found to have the highest level of off-target activity in HEK-293T cells (Table 4-5), also had the highest off-target activity in K562 cells. However, while the ratio of on-target activity to off-target activity near *TNN* was ~2:1 in HEK-293T cells, the ratios were much higher (i.e. less off-target activity) in K562 cells (Table 4-6), affirming the notion that HEK-293T cells are highly susceptible to off-target nuclease activity. While we observed consistent activity at the *TNN* locus across all replicates, providing evidence that nuclease off-target activity is a highly repeatable phenomenon, we did not observe any statistically significant differences in the on-target:off-target ratios between plasmid and mRNA delivery of the TALENs (although the ratio was 1.4-fold greater for plasmid delivery).

4.3.6 TALEN pairs with one inactive FokI domain have increased specificity[6]

4.3.6.1 Abstract

Transcription Activator-Like Effector Nucleases (TALENs) have the potential to become a powerful tool for genome editing in a wide range of biomedical applications. However, TALEN pairs that target an intended gene locus (on-target) may also cleave similar sequences in the genome (off-target), inducing genomic mutagenesis and instability. Here we show that TALEN pairs targeting the human beta-globin gene with one inactive FokI domain had similar levels of on-target gene modification, but lower levels of off-target cleavage compared with the TALEN pair having two active FokI domains (normal TALEN pair), in contrast to ZFN nickase pairs. TALEN pairs with one inactive FokI domain may thus prove a general method for improving the specificity of TALENs for genome editing applications.

4.3.6.2 Results

As a potentially effective way to reduce TALEN off-target mutagenesis, we constructed and evaluated TALEN pairs that consist of one inactive and one active FokI domain, similar to the "nickase" Zinc Finger Nucleases (ZFNs) [94, 128]. Specifically, four TALEN pairs were constructed and characterized based a previously validated TALEN pair targeting the β-globin mutation responsible for sickle cell anemia, with 'L4' and 'R4' representing the 'left' and 'right' TALENs respectively [125]:

i) a pair with both FokI domains active (L4+/R4+, referred to as normal pair of TALENs)

ii) a pair with the right FokI domain inactivated (L4+/R4-)

iii) a pair with the left FokI domain inactivated (L4-/R4+)

iv) a pair with both FokI domains inactive (L4-/R4-).

[6]Modified from: Cradick TJ, Lindsay C, Kumaran RI, **Fine EJ**, et al. (manuscript in revision). TALEN pairs with one inactive FokI domain have increased specificity

Table 4-7: L4+R4+ TALEN off-target activity at sites ranked by PROGNOS. 49 sites were chosen for interrogation based on the top-ranked sites of the L4R4 TALENs, as determined by a mixture of the 'Homology' and 'RVDs' PROGNOS algorithms (preference given to potential heterodimeric sites), were evaluated using SMRT sequencing in transfected HEK-293T cells. The 'match type' indicates the orientation of the left (L) and right (R) nucleases at the site and the length of the spacer sequence. Site sequences are listed as 5'–(+) half-site–spacer–(-) half-site–3'. Therefore, the (-) negative half-site is listed in the reverse anti-sense orientation as compared to the DNA sequence that the TALEN binds. In sequences, lower-case letters indicate mismatches as compared to the target site. "293T cells modification frequency" is the percentage of observed sequences showing evidence of NHEJ events. The target site in the *HBB* gene contains a mismatch in the left binding site because the TALEN is designed to target the sickle allele, not the wild-type allele found in 293T cells. For all sites shown, significantly higher ($p < 0.05$) frequencies of indels were observed in nuclease-treated cells as compared to mock-treated cells.

Closest Gene	Match Type	Mutations per half-site (+)	Mutations per half-site (-)	(+) half-site	(-) half-site	293T Cells Modification Frequency
HBB	L-15-R	1	0	TGCACCTGACTCCTGa	TACTGCCCTGTGGGGCAA	39.57%
HBD	L-15-R	2	2	TGCAtCTGACTCCTGa	cAaTGCCCTGTGGGGCAA	12.66%
GREB1L	L-17-R	5	5	TaCAgCTaACaCaTGT	TACTtgCtTGTGGGGtAg	0.39%
TMED10	L-14-R	5	5	TGtACCccACTCCTcc	TACTGtttTGTtGGGtAA	0.09%
					Total Sites Interrogated:	49

Here, '+' or '-' represent respectively TALEN monomers with an active or inactive FokI domain. The on- and off-target cleavage activities were systematically quantified for all four pairs of TALENs. We found that TALEN pairs with one inactivated FokI domain had the same on-target activity as the normal TALEN pair in HEK-293T cells, but with lower off-target activity.

Initially, an off-target screen was conducted using the L4+R4+ 'normal' TALEN pair at potential off-target sites predicted by the PROGNOS algorithm [34]. From this search, three heterodimeric off-target sites were discovered (Table 4-7)—heterodimeric sites were of particular interest for this study because they would allow analysis of how the inactivation of one of the two FokI domains affected activity.

Following the initial off-target screen, we determined that the sites at the *HBD* and *GREB1L* loci had off-target activity sufficiently above the detection limit of the assay to warrant investigation using the inactive FokI domains. Using SMRT sequencing, the *HBB*,

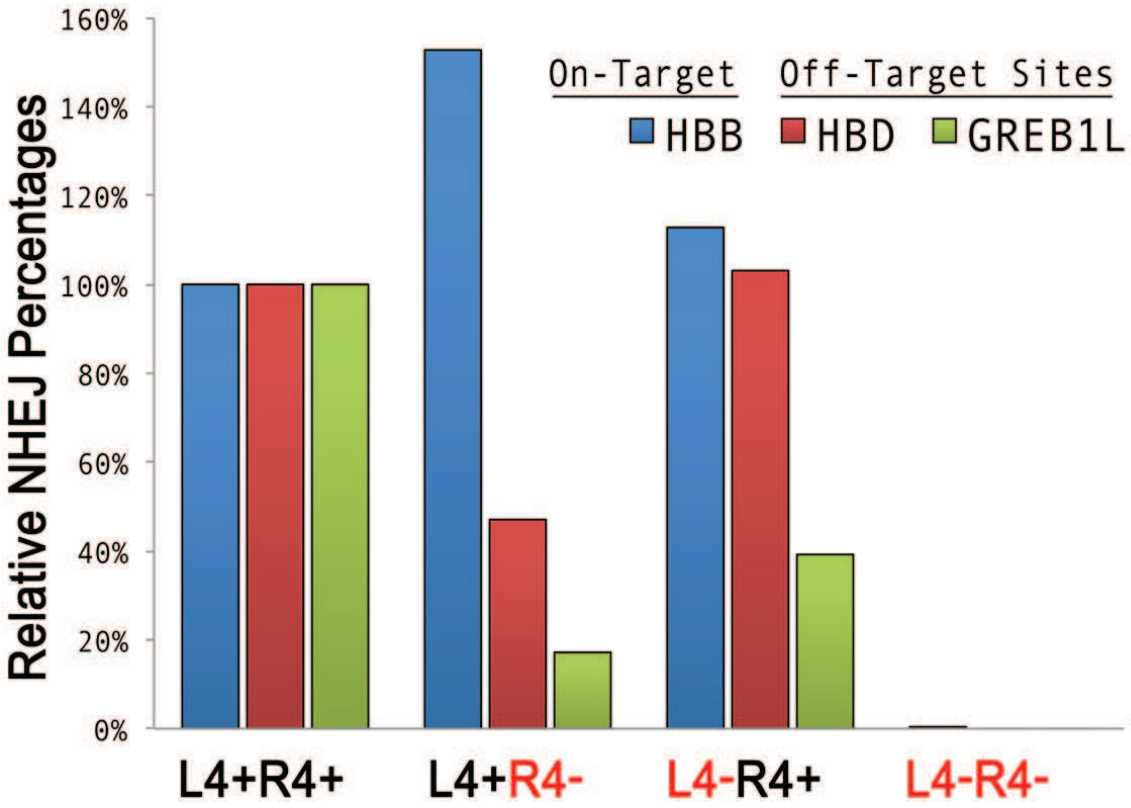

Figure 4-2: TALEN pairs with one inactivate cleavage domain have lower off-target NHEJ. Using SMRT sequencing, the on- (*HBB*) and off-target (*HBD*, *GREB1L*) indel percentages were calculated and compared for the variants to the level of the L4+/R4+ TALEN pair.

HBD, and *GREB1L* loci were analyzed from HEK-293T cells transfected with each of the four different TALEN pairs (Figure 4-2). As expected, the L4-R4- control TALENs had no NHEJ activity at any loci. Surprisingly, the L4+R4- TALENs had higher on-target activity and substantially lower off-target activity at both *HBD* and *GREB1L* than the L4+R4+ TALENs. The L4-R4+ TALENs had reduced activity only at the off-target site near *GREB1L*. At this time, the reason for the differences in performance observed by inactivating the left vs. the right FokI domain are unknown.

4.3.7 Total increase in off-target dataset expansion

Due to the difficulty of previous off-target prediction methods, as of 2011, only 7 TALENs and ZFNs had identified off-target sites. By utilizing the solely bioinformatics-based

Table 4-8: All nucleases investigated using PROGNOS. 9 total TALENs and 7 total ZFNs were investigated using PROGNOS in various cell types and species. *a modified HEK-293T cell line with a single copy of the *GFP* gene integrated into its genome.

Nuclease Name	ZFN/TALEN	Species	Target Gene	Cell Type	Bona Fide Sites	Total Sites	Delivery Method	Year	Publication?	Thesis Section
S2/S5 NK	TALEN	Human	HBB	HEK-293T	1	21	FuGene Transfection	2012	**Fine EJ**, et al. [34]	Table 3-4
S2/S5 NN	TALEN	Human	HBB	HEK-293T	3	20	FuGene Transfection	2012		
S1/S7 NK	TALEN	Human	HBB	HEK-293T	1	24	FuGene Transfection	2012		
S1/S7 NN	TALEN	Human	HBB	HEK-293T	3	25	FuGene Transfection	2012		
ZFN_3F	ZFN	Human	HBB	HEK-293T	6	23	FuGene Transfection	2012		
ZFN_4F	ZFN	Human	HBB	HEK-293T	1	23	FuGene Transfection	2012		
L4/R4	TALEN	Human	HBB	HEK-293T	3	49	FuGene Transfection	2013	Cradick et al. (in revision)	Table 4-7
TC-NC	TALEN	Human	CCR5	HEK-293T	2	16	PEI Transfection	2013	Mussolino et al. [79]	Table 4-3
TC06	TALEN	Human	CCR5	HEK-293T	3	16	PEI Transfection	2013		
TA04	TALEN	Human	AAVS1	HEK-293T	1	23	PEI Transfection	2013		Table 4-4
ZFNA	ZFN	Human	AAVS1	HEK-293T	3	24	PEI Transfection	2013		
GFP-ZFNs	ZFN	Human	GFP	HEK-293T*	1	23	Lipofectamine Transfection	2013		Figure 4-1
Ugt1a1-ZFNs	ZFN	Rat	UGT1A1	C6	6	22	Lentiviral Vector	2013	Abarrategui-Pontes et al. [1]	Table 4-2
L3/R3	TALEN	Human	IL2RG	HEK-293T/K562	8/1	24	FuGene / Nucleofection	2014	Kildebeck et al. (in preparation)	Table 4-5
Rosa26-ZFNs	ZFN	Mouse	Rosa26	SC-1	1	23	Lentiviral Vector	2014		Table 4-1
CCR5-ZFNs	ZFN	Mouse	CCR5	SC-1	7	23	Lentiviral Vector	2014		

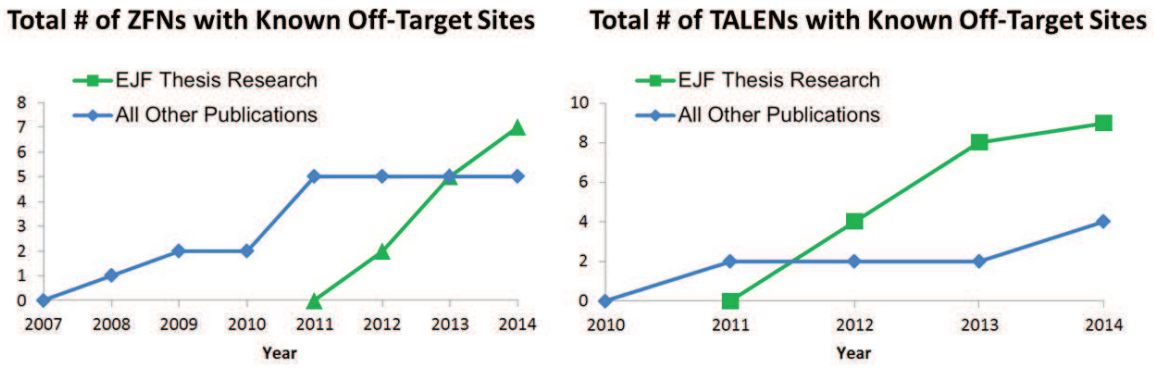

Figure 4-3: Nuclease Off-Target Studies Over Time.

PROGNOS algorithms to predict off-target sites, we were able to study 16 different nucleases and identify bona fide off-target sites for all of them (Table 4-8). This represents an increase of ~230% over the 2011 level. Further underscoring the contributions of PROGNOS to the field is the fact that only one study was published between 2011 and 2014 identifying off-target sites for a new nuclease (Figure 4-3). In total, the nucleases studied by PROGNOS represent over half (64%) of all published ZFNs and TALENs with known off-target sites.

4.3.8 Engineered nucleases and lentiviral vectors have different off-target profiles[7]

4.3.8.1 Introduction

Gene therapy approaches continue to gain momentum as methods employing integrating viral vectors [2] or engineered nucleases [41] demonstrate great success in improving efficacy. However, "off-target" effects remain a major safety concern as the negative outcomes of the X-SCID clinical trials [46] continue to haunt the field. Potentially harmful modifications to the genome can be made by viral vectors randomly integrating their cargo or by nucleases—including zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and RNA-guided endonucleases (RGENs) such as CRISPR/Cas9 systems—creating DNA breaks at locations other than their intended target site, which are then repaired in an error-prone manner by non-homologous end-joining (NHEJ). While many factors influence whether the off-target effects will result in an adverse clinical outcome, one metric that can be used to compare gene therapy methods is the genomic context of the off-target event. For example, disruptions within genes are more likely to be detrimental to a patient than those in regions with no currently known function.

We used the QuickMap tool [4] to compare the genomic contexts of the off-target events between two prominent gene therapy platforms: lentiviral vectors (LVs) and engineered nucleases. Our LV dataset consisted of 806 integration sites gathered from three independent publications. By combining nearly all bona fide off-target sites identified for ZFNs, TALENs and RGENs in the human genome to date (totaling 282 sites from 35 different nucleases), we created an engineered nuclease dataset of sufficient size to permit statistical comparisons. As a control, QuickMap provides a dataset of 10^6 random locations in the human genome.

[7]Modified from: **Fine EJ** & Bao G (manuscript submitted). Engineered nucleases and lentiviral vectors have different off-target profiles

4.3.8.2 Datasets

All sites were converted to coordinates in the hg19 build of the human genome prior to QuickMap analysis.

Our engineered nuclease dataset is drawn from a diverse set of observations in order to ensure that the results are truly representative. The results were obtained from the experiments described in this thesis and 12 additional manuscripts [53, 35, 119, 88, 87, 21, 22, 68, 37, 101, 57, 43, 51], from 7 independent laboratories which employed 9 different methods to locate off-target activity. The set of 282 off-target sites is diverse, consisting of 35 different nucleases with a mean of 8 sites per nuclease and only a modest non-parametric skew of 0.43 due to the presence of some nucleases with larger numbers of identified off-target sites.

Our lentiviral vector (LV) integration site dataset is drawn from 6 separate transductions of 4 different cell types by 3 independent laboratories [121, 5, 7]. Although there are many additional publications analyzing LV integrations, only a small fraction are in human cells (many are performed in mice) and listed the genomic coordinates of the integration events; indeed, two of these datasets were only obtained through personal correspondence with the authors (many thanks to Dr. Duran Üstek and Dr. Nirav Malani).

4.3.8.3 Results

In comparing the two platforms, we observed marked differences in the preference of where off-target activities occurred. Compared to the random dataset, LVs showed a strong preference (3.2-fold) for integration within cancer-associated genes, which was not observed in the off-target activity of nucleases (Figure 4-4a). For off-target events within genes, LVs also showed a clear preference compared to the random control (1.7-fold), whereas nucleases only had a modest preference (1.3-fold) (Figure 4-4b). In contrast, nucleases modify exons much more frequently than LVs (Figure 4-4c). Finally, we compared the three different major classes of engineered nucleases (ZFNs, TALENs, and RGENs) and found no

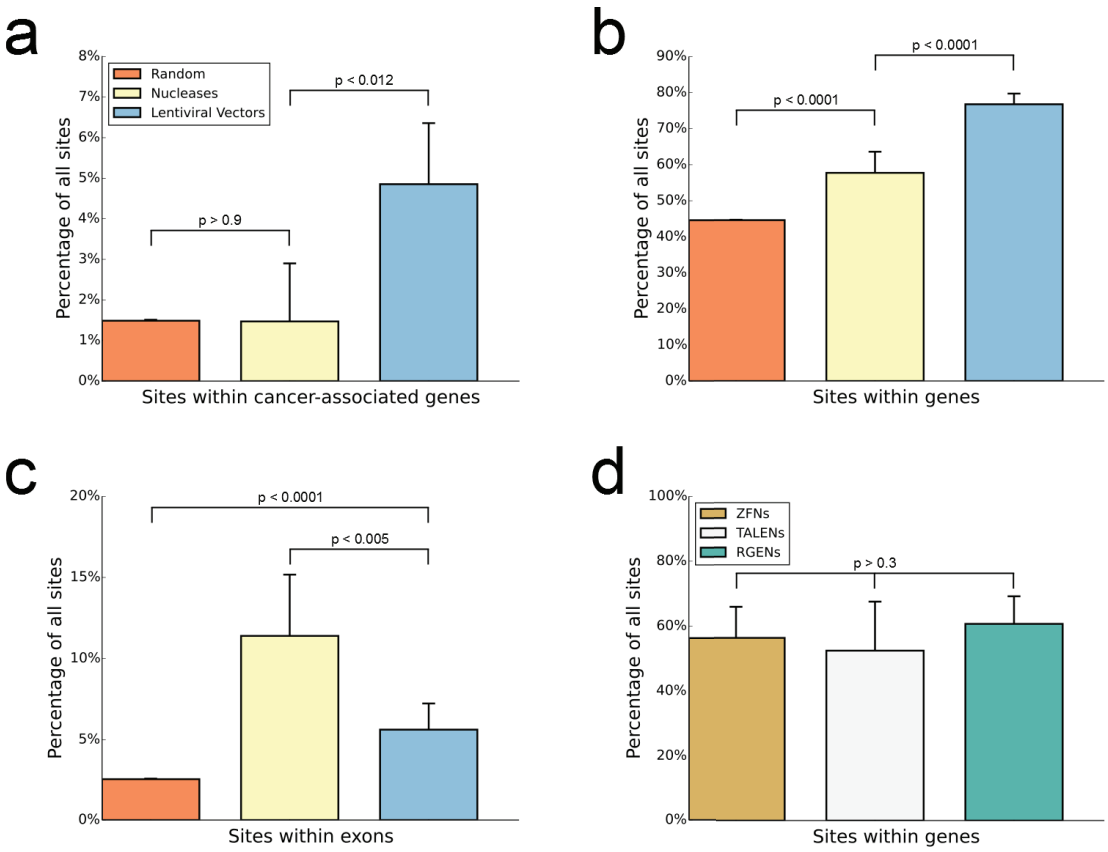

Figure 4-4: Comparison of engineered nuclease and lentiviral vector off-target profiles. Error bars are 95% confidence intervals. P-values calculated using two-tailed Fisher's exact test. (**a**) Percentage of sites located within cancer-associated genes. (**b**) Percentage of sites located within all genes. (**c**) Percentage of sites located within exons. (**d**) Comparison of the percentage of sites located within genes for different types of engineered nucleases.

statistically significant difference between the percentage of off-target sites located within genes (Figure 4-4d), suggesting a similar underlying mechanism in how off-target sites are accessed. The limited sample size did not permit statistically meaningful comparisons for other metrics such as cancer-associated genes or exons for the three classes of engineered nucleases.

4.3.8.4 Discussion

There are some limitations to our study. One limitation is that most engineered off-target studies take place in immortalized cell lines. Indeed, in our dataset, all sites were discovered in K562, HEK-293, or U2OS cells except for two sites found in hESCs. In contrast,

many LV integration site analyses take place in primary cells. To compensate for this, we obtained LV data from HEK-293 and K562 cells to make our datasets more comparable. However, as we lack substantial data on engineered nuclease off-target activity in primary cells, it remains possible that the location pattern reported here are different than what might be observed in primary cells.

In conclusion, our study shows that engineered nucleases and LVs can produce different profiles of off-target modifications. Whether disruptions within exons or within cancer-associated genes are more prone to trigger adverse clinical outcomes will require future investigations.

4.4 Conclusion

The large number of additional nucleases that were able to be successfully analyzed by the PROGNOS algorithms [34] to locate new bona fide off-target sites demonstrates that the two major goals of the original approach were achieved:

- A method so easy to apply that it could be readily used on large numbers of nucleases
- Reasonably accurate predictive algorithms that successfully locate sites of bona fide off-target activity

The substantial (~230%) increase in the total number of ZFNs and TALENs with known bona fide off-target sites created a dataset large enough to attempt to use for machine learning and also allowed insight into the nature of the genomic context of where nuclease off-target activity occurs. Throughout the course of identifying off-target sites for the various ZFNs and TALENs, several insights into the general nature of nucleases and off-target activity were gained:

- Lentiviruses can delivery ZFNs into a variety of cell types that are refractory to traditional transfection techniques and achieve robust nuclease activity.

- ZFNs with additional numbers of zinc finger units do not necessarily confer higher specificity.
- While varying the delivered dose of nuclease does affect the activity level accordingly, there is not necessarily an optimal zone which can achieve a high on-target:off-target ratio.
- TALENs are generally more specific than ZFNs.
- TALENs can be rationally designed to discriminate between highly similar sequences.
- The locations of off-target activity are consistent across different cell types and the level of off-target activity is consistent across experimental replicates.
- mRNA delivery of nucleases does not necessarily confer an improved on-target:off-target ratio compared to plasmid delivery.
- Nucleases have a distinct tendency (compared to a random control) for off-target activity to occur in exons.
- While nucleases do not have a tendency for off-target activity to occur in cancer-associated genes, this tendency is observed in lentiviral vectors.

CHAPTER V

USING MACHINE-LEARNING TECHNIQUES TO IMPROVE OFF-TARGET PREDICTION

5.1 Abstract

5.2 Introduction

Machine learning is a powerful approach for building predictive models which can greatly facilitate the analysis of large datasets. While the field of genome engineering has historically been a relatively low-throughput environment that did not generate sufficient data on which to train machine learning algorithms, the development of transcription activator-like repeat nucleases (TALENs) and RNA-guided endonucleases (RGENs) based on the CRISPR/Cas9 system have led to a rapid increase in the amount of raw data available. This has enabled several recent efforts to incorporate machine learning techniques into genome engineering research. Doench et al. [27] trained a logistic regression classifier on a dataset of >1800 RGENs to better predict highly active target sites within genes. In the realm of off-target analysis, Sander et al. [101] applied a Naïve Bayesian algorithm to the full set of sequences cleaved by ZFNs *in vitro* in order to build a model that could rank any sequence in the genome for its likelihood as an off-target site for that pair of zinc finger nucleases (ZFNs); this approach was later extended as a method to analyze pairs of TALENs as well [43]. While that approach is highly effective at predicting off-target activity for a particular nuclease, the reliance on extensive experimental work to obtain the necessary *in vitro* cleavage data for that nuclease limits its broader applicability.

The extreme ease with which the newer generation TALENs and CRISPRs can be designed and constructed—compared to earlier technologies such as ZFNs or meganucleases—has triggered a paradigm shift in the genome engineering field. It is now possible

to rapidly create dozens of nucleases in a matter of a few weeks that are highly efficient at modifying their intended target site. As a result, off-target prediction methods that rely on experimental characterization of the nuclease create a bottleneck in the development process. There is therefore a great need for methods that can predict off-target activity for a nuclease purely *in silico* based solely on the sequence of the intended target site.

To address this need, we trained classifiers on a set of all known off-target sites in order to develop two generalized models, one for ZFNs and one for TALENs, which reliably predict off-target sites in a genome for any given nuclease target sequence. This approach will allow for judicious target site selection in the initial design of a nuclease in order to choose an approach that has the minimal potential for off-target activity.

5.3 Methods

The `scikit-learn` package for Python was used to implement all machine learning approaches.

5.3.1 Feature Extraction

All features were scaled to values between 0 and 1 for input into learning algorithms since many algorithms are not scale invariant. Because ZFNs and TALENs can target different lengths of sequence as well as different bases at each position, special care was taken to ensure that all features could be generalized to any ZFN or any TALEN.

Because ZFNs and TALENs operate as dimers, most of the features were extracted separately for each monomeric half-site. During training, the ordering of the 'left half-site' vs. 'right half-site' designation was randomized in order to eliminate any biases.

5.3.1.1 ZFNs

Because ZFNs consist of individual zinc fingers that bind to a triplet of bases, many features were created relating to the division of the intended binding site into base triplets. There is well-documented evidence that many zinc finger subunits exhibit context-dependent effects

due to the bases surrounding the triplet, but a generalized set of rules regarding these effects has not been established [72].

The following were extracted as one-hot encoded features (either `1` if the statement was true, or `0` if it was false) separately for each half-site:

- The closest base to the FokI domain is matched to its target sequence
- The 2nd closest base to the FokI domain is matched to its target sequence
- The 3rd closest base to the FokI domain is matched to its target sequence
- The closest triplet of bases to the FokI domain has 0 mismatches
- The closest triplet of bases to the FokI domain has 1 mismatch
- The closest triplet of bases to the FokI domain has 2 mismatches
- The closest triplet of bases to the FokI domain has 3 mismatches
- The furthest triplet of bases from the FokI domain has 0 mismatches
- The furthest triplet of bases from the FokI domain has 1 mismatch
- The furthest triplet of bases from the FokI domain has 2 mismatches
- The furthest triplet of bases from the FokI domain has 3 mismatches
- The 3rd further base from the FokI domain is matched to its target sequence
- The 2nd further base from the FokI domain is matched to its target sequence
- The furthest base from the FokI domain is matched to its target sequence

The following were extracted as decimal percentages (i.e. ranging from 0–1) regarding the triplets lying between the most proximal triplet to the FokI domain and the most distal triplet from the FokI domain:

- The % of triplets that contain 0 mismatches
- The % of triplets that contain 1 mismatch
- The % of triplets that contain 2 mismatches
- The % of triplets that contain 3 mismatches
- The % of triplets that contain a matched base at the position proximal to the FokI domain

The following were extracted as decimal percentages (i.e. ranging from 0–1) regarding the full sequence of each half-site:

- The % of all positions where the base matches the intended target (Homology)
- The % of intended guanosine targets that remain matched (Conserved G's)

The following were extracted as decimal percentages (i.e. ranging from 0–1) regarding the relationship between the two half-sites:

- The ratio of %Homologies of [less well-matched half-site]:[more well-matched half-site] (so that the resulting value is ≤ 1)
- The ratio of %Conserved G's of [less well-matched half-site]:[more well-matched half-site] (so that the resulting value is ≤ 1)

5.3.1.2 TALENs

Many of the extracted features were based on the approach of the TALEN v2.0 algorithm previously developed (Section 3.3.3.2). Interactions between individual RVDs and the corresponding DNA bases in a target site were scored using binding frequencies derived from SELEX experiments as previously described (Table 3-1). As in the TALEN v2.0 algorithm, the concepts of TALEN polarity [75] and the effects of strong RVDs [112] were also taken into account.

The following features were extracted as decimal percentages (i.e. ranging from 0–1) by normalizing the RVD-DNA interaction scores of the potential binding site to the RVD-DNA interaction scores of the intended binding site:

- The normalized score of position 0 (5' to where the RVDs begin)
- The average normalized score of positions 1 to 5 (starting at the 5' end of the binding site)
- The average normalized score of positions 6 to N-5 (where 'N' is the 3' distal RVD)
- The average normalized score of positions N-4 to N-1
- The normalized score of position N
- The average normalized score of all positions (0 to N)

The following were extracted as decimal percentages (i.e. ranging from 0–1):

- The percentage of strong RVDs (NN or HD) which matched their intended base (calculated separately for each half-site)
- The ratio of the average normalized score of all positions of of [less well-matched half-site]:[more well-matched half-site] (so that the resulting value is ≤ 1)

5.3.2 Cross-Validation

In order to mitigate the possibility of over-fitting the model to the training data, we utilized a nested (5 iterations) 4-fold cross-validation strategy. Because the ratios of the two classes in our datasets are highly unbalanced, we used the `StratifiedKFold` method which ensures that each fold contains the same percentage of each class (i.e. bona fide active vs. inactive/untested) as the overall dataset.

The average precision (the area under the Precision-Recall curve) was used as a metric to score the performance of the classifiers. To construct the Precision-Recall curve, the

decision function of the classifier (which provides a confidence estimate of the classification assignment, based on the signed distance to the hyperplane) was used to rank-order all samples. The mean of the best score from each of the 5 nested cross-validation iterations was used to evaluate the performance of a classifier.

5.3.3 Datasets

By aggregating nearly all published bona fide off-target sites for ZFNs and TALENs, we were able to create datasets of sufficient size to permit training machine learning algorithms. Because of the wide variety of experimental methods utilized in the various manuscripts documenting off-target activity from which we drew our dataset, it is difficult to make quantitative comparisons as to the exact level of off-target activity at each different off-target site. Therefore rather than train an estimator to predict exact rates of off-target activity, we strove to create a model which would classify a potential site as either 'active' or 'inactive'. For our datasets, the 'active' class consisted of all bona fide off-target sites as well as on-target sites. To generate the 'inactive' class, we used the PROGNOS algorithm [34] to search the genome for all sites with >66% homology to the intended target sequence. Since the resulting 'inactive' class contained ~10^6 as many sites as the 'active' class, we selected a random subset of ~2.5% of the 'inactive' sites for each nuclease in order to reduce the computational resources required for the machine learning process. Since a nuclease typically had several bona fide off-target sites, the number of off-target sites in the 'active' class outnumbered the on-target sites. In order to help guide the machine towards recognizing the baseline biological assumption that all on-target sites should be classified as 'active', we added in additional 'spike-in' examples of on-target sites into the datasets.

5.3.3.1 ZFNs

The 'active' class of our ZFN dataset consisted of 134 bona fide off-target sites gathered from 11 different ZFNs, the on-target sites for those ZFNs, and 50 additional 'spiked-in' on-target sites (Figure 5-1). ~$35,000$ 'inactive' sites are drawn from genomic sequences

with homology to the intended target sites. The off-target sites were uncovered using a variety of methods and the number of off-target sites per nuclease shows only a moderate positive skew (non-parametric skew of 0.384). Taken together, this should result in a model that can be broadly applied to accurately predict off-target sites for any given ZFN.

5.3.3.2 TALENs

Several publications have sought to investigate off-target activity by designing TALENs to sites that have additional highly homologous sites in the genome. While this acts as an effective model system to examine what may affect levels of off-target activity, these sites are trivial to locate by any off-target prediction method. Therefore, while they are included in the 'active' class of TALEN sites in our dataset, any site with >87% homology to the intended target is marked as a 'trivial' off-target site to denote the fact that it is not appreciably contributing to the diversity of off-target sites in our dataset.

The 'active' class of our TALEN dataset consisted of 43 bona fide off-target sites gathered from 14 different TALENs, the on-target sites for those TALENs, 33 'trivial' off-target sites, and 50 additional 'spiked-in' on-target sites (Figure 5-1). ~35,000 'inactive' sites are drawn from genomic sequences with homology to the intended target sites. The off-target sites were uncovered using a variety of methods and the number of bona fide off-target sites per nuclease shows only a moderate positive skew (non-parametric skew of 0.381). Taken together, this should result in a model that can be broadly applied to accurately predict off-target sites for any given TALEN.

5.4 Results

5.4.1 Hyperparameter Tuning

After testing several different machine learning algorithms, we determined that a support vector classifer (SVC) employing a radial basis function (RBF) kernal achieved the best results on the ZFN dataset while a stochastic gradient descent (SGD) logarithmic regression classifer employing an elastic net penalty achived the best results on the TALEN dataset. In

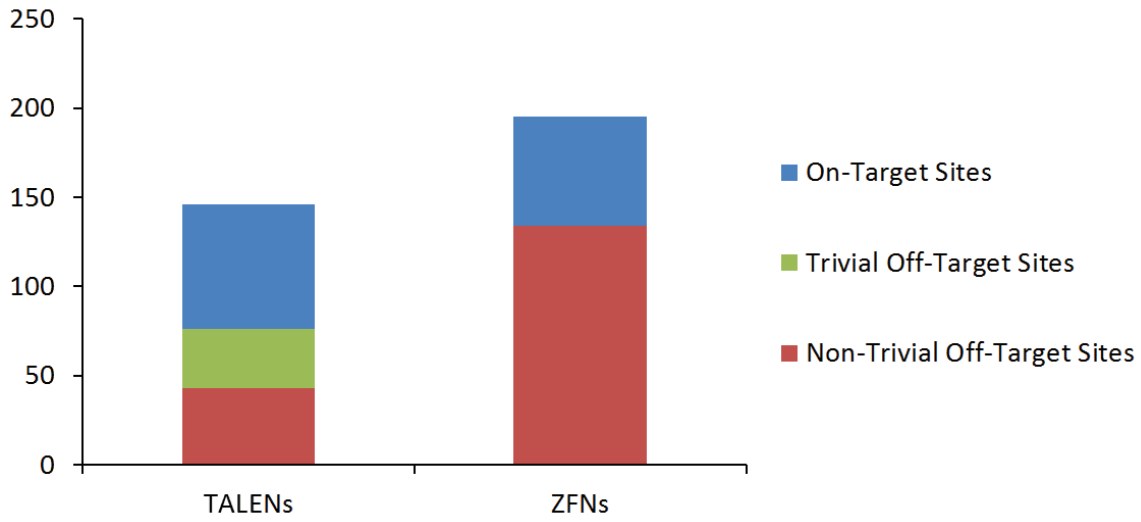

Figure 5-1: Machine learning datasets

order to further refine the classifiers, we performed a grid search over a range of hyperparameters. For ZFNs, we found that setting $C = 10^5$ and $\gamma = 10^{-4}$ achieved the best results (Figure 5-2a). For TALENs, we found that setting $\alpha = 10^{-6}$ and the L1/L2 ratio at 0.01 achieved the best results (Figure 5-2b).

5.4.2 Training Results

After optimizing the choice of machine hyperparameters, we trained the models on the ZFN and TALEN datasets using 4-fold cross-validation. By plotting the Precision-Recall curves for each of the folds, we observed that each fold was clustered relatively closely around the average of all of the folds, providing evidence that the model was not overtraining to the dataset (Figure 5-3).

5.4.3 Performance on Hold-Out Test Set

Prior to the onset of any experiments, 10% of each dataset (ZFNs and TALENs) was set aside to use for a hold-out test set on the final classifiers to asses the performance of the off-target prediction models. Due to the relatively small size of the overall training sets, the hold-outs were also limited in size (13 and 6 bona fide for ZFNs and TALENs respectively),

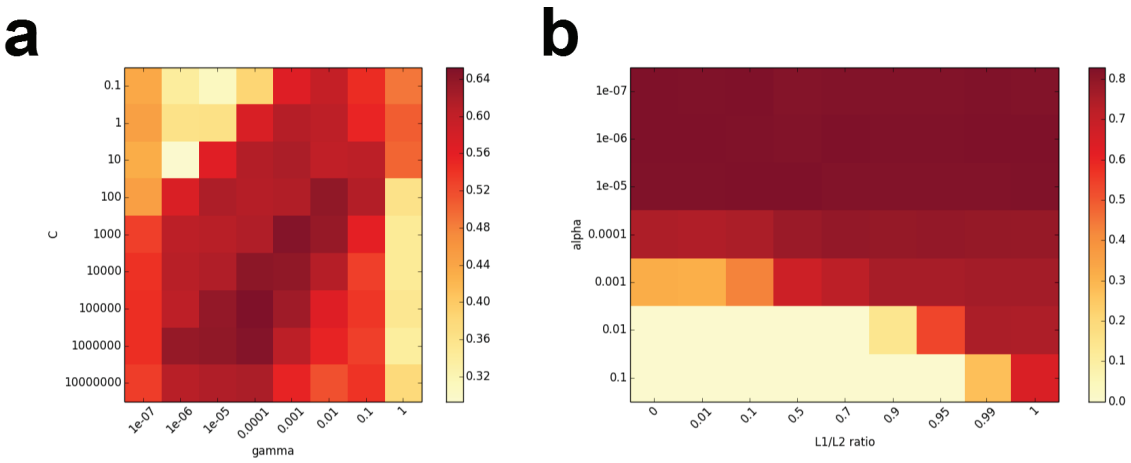

Figure 5-2: Hyperparameter Tuning. Results of a grid search over a range of hyperparameters. (**a**) Tuning a support vector classifier using a radial basis function kernal on the ZFN dataset. (**b**) Tuning a stochastic gradient descent logarithmic regression classifer with an elastic net penalty on the TALEN dataset.

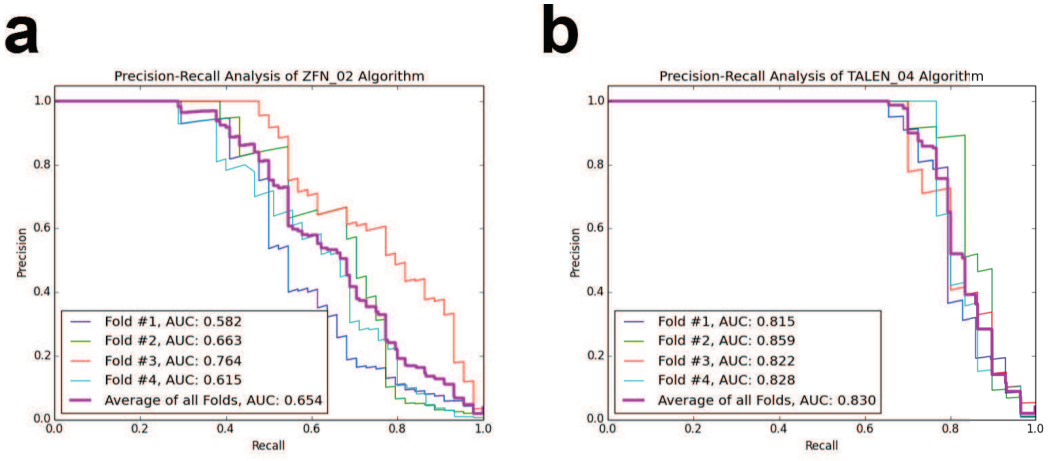

Figure 5-3: Cross-validation on training set. Performance of individual cross-validation folds compared to the average of all folds for the ZFN (**a**) and TALEN (**b**) classifiers.

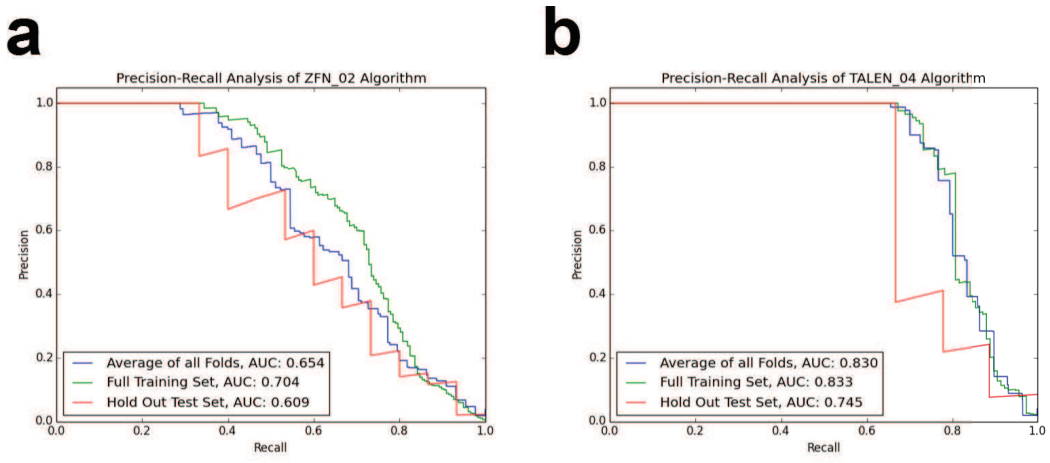

Figure 5-4: Testing performance of classifiers on hold-out datasets. Performance on the full training set, the hold-out test set, and the average performance across folds of the cross-validation for the ZFN (**a**) and TALEN (**b**) classifiers.

resulting in a somewhat coarse-grained Precision-Recall plot. Nevertheless, we observed relatively high concordance between the performance of the models on the training and test sets (Figure 5-4), increasing our confidence that the classifiers were not overtrained and that these results can be used to estimate their performance on future off-target studies of nucleases.

5.4.4 Cross-Validation by Nuclease

In addition to performing standard cross-validation on the whole dataset, we performed an additional type of cross-validation based on the off-target sites of the individual nucleases. Rather than dividing the training set into random folds for cross-validation, we trained the classifier on all sites *except* for those of a particular nuclease, and then assessed the performance of the classifier on that nuclease (Figure 5-5). As expected, the precision-recall curves for any individual nuclease are less smooth than the average performance, but overall the classifier seemed to have similar performance across a wide variety of nucleases. To quantify the performances, we calculated the coefficient of variation (CV)—the standard deviation divided by the mean—of the AUC. For ZFNs, the CV across all nucleases was 50%, but if two extreme outliers (the GFP ZFNs and the Li_Mouse_Albumin ZFNs) are

excluded, then the CV drops to 35%. The CV across all TALENs was 29%.

5.4.5 Comparison to Other Off-Target Prediction Algorithms

After validating that the classifiers were not over-trained, we compared their performance to the previous algorithms developed in Chapter 3. Specifically, we used three metrics to measure algorithm performance. As in Chapter 3, we calculated the precision within the top 16 rankings (taking the mean across all nucleases) and also the percentage of all bona fide sites found within the search limit for each nuclease (the search limit is defined by the number of sites interrogated in the original study that uncovered the bona fide sites). Additionally, we calculated the average precision (the area under the Precision-Recall curve) of the classifier for each nuclease.

The 'ZFN_02' classifier developed through machine learning represented a 14% improvement over the best previously developed algorithm (Figure 5-6a). Delving deeper into the individual metrics, the 'ZFN_02' classifier outperformed all previous algorithms in all three measurements (Figure 5-6b).

The 'TALEN_04' classifier developed through machine learning represented a more modest, but still readily observed, 7% improvement over the best previously developed algorithm (Figure 5-6c). Looking at the individual metrics, the 'TALEN_04' classifier outperformed all previous algorithms in terms of average precision and the percentage of bona fide sites within the search limit, but several other algorithms had slightly better performance in terms of the precision within the top 16 rankings (Figure 5-6d).

5.4.6 Insights from Feature Weightings

Because the TALEN classifier was constructed using a logarithmic regression algorithm, each feature is assigned a discrete value during training. These weighting values provide insight into what the most influential features are—what has the most impact on TALEN specificity. Previous experimental data had provided evidence that there can be a 'polarity' effect where mismatches in the 5' region of the TALEN binding site less well tolerated

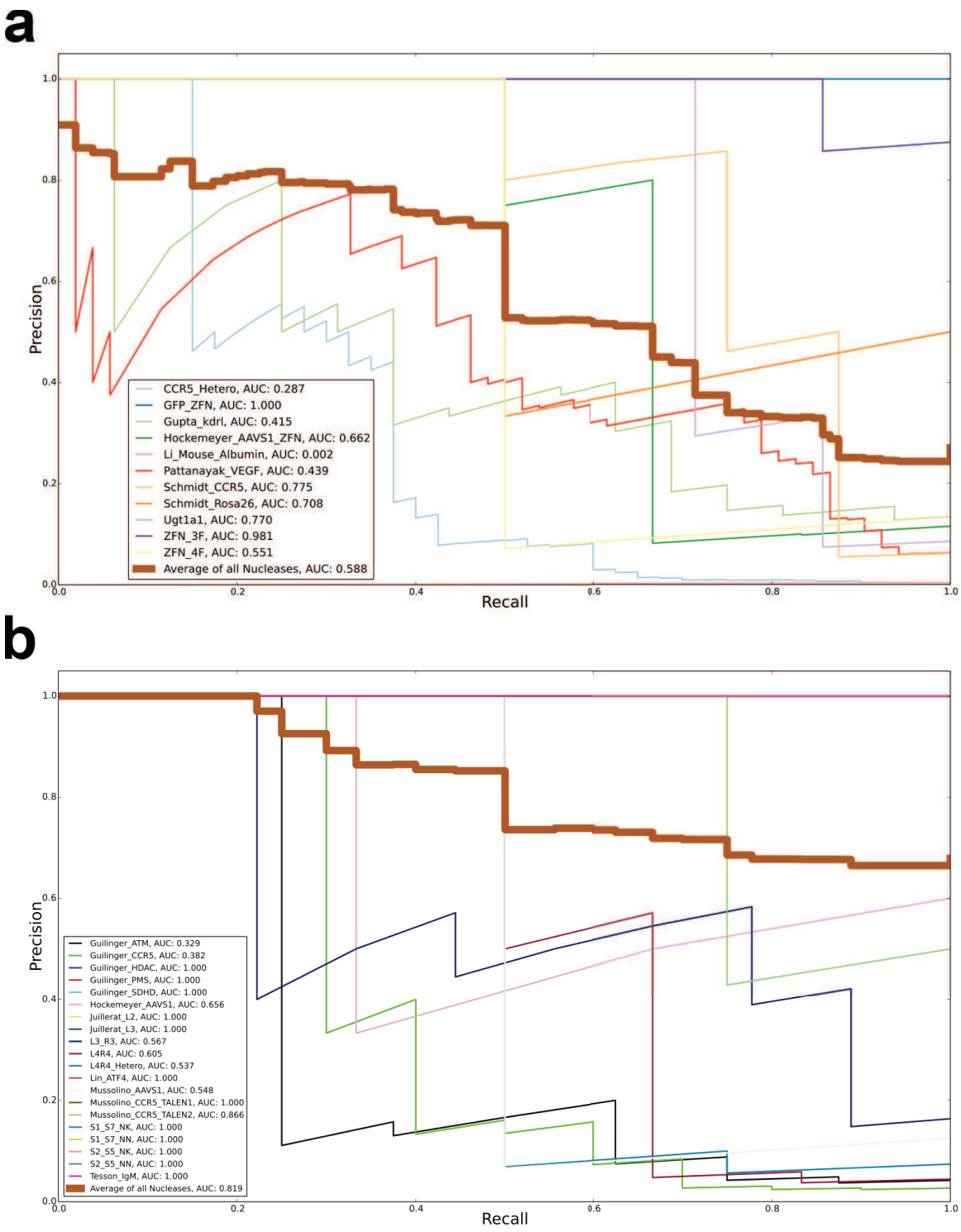

Figure 5-5: Cross-Validation by Nuclease. Performance of classifiers trained on all data except the indicated nuclease and then tested on that nuclease for ZFNs (**a**) and TALENs (**b**). Average performance of all classifiers is shown by the thicker brown line.

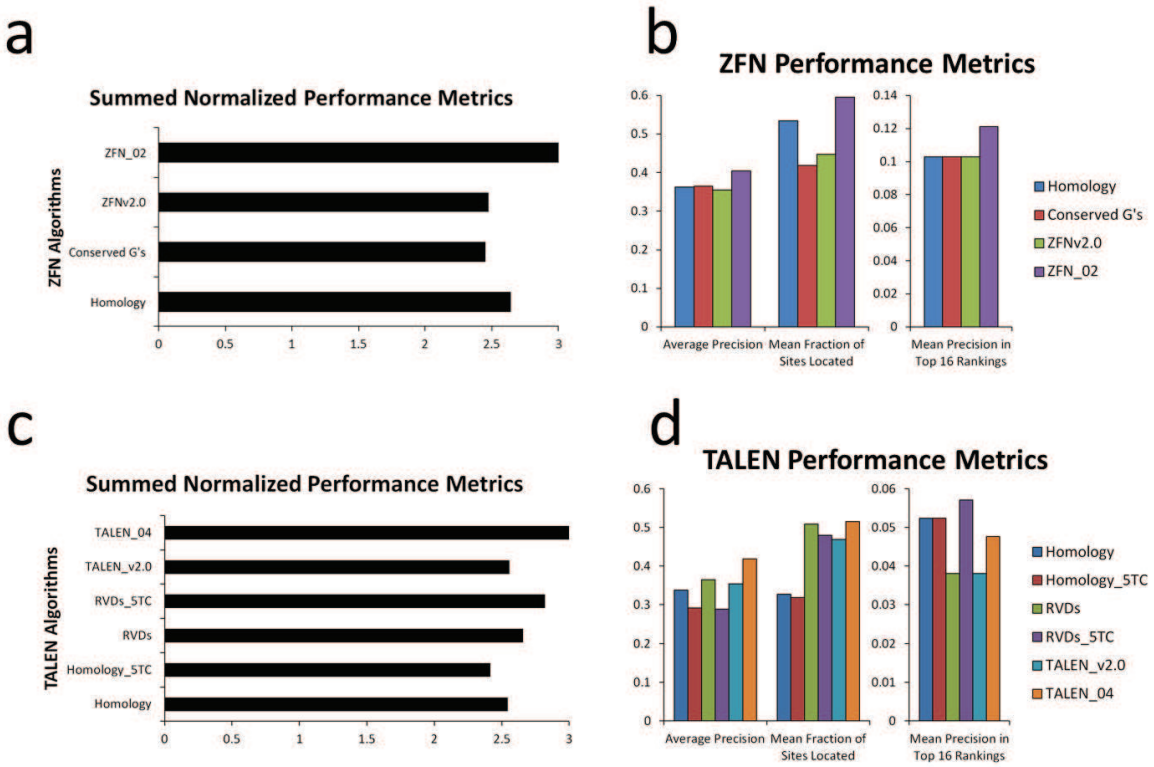

Figure 5-6: Performance of different off-target prediction algorithms. Comparing the overall performance across all nucleases of different predictive algorithms for ZFNs (**a**, **b**) and TALENs (**c**, **d**).

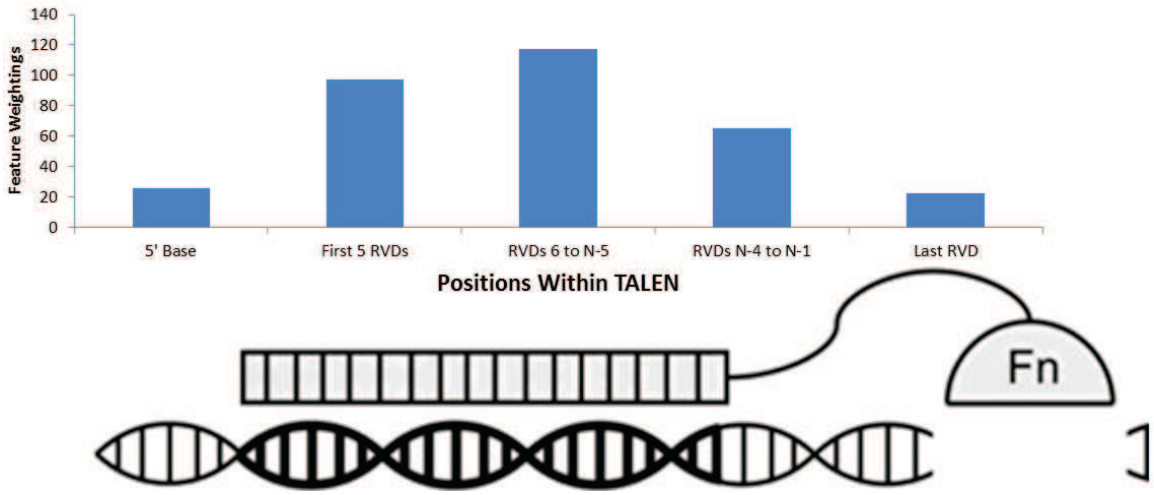

Figure 5-7: Positional feature weightings of the TALEN classifier. Weightings assigned to different features of the 'TALEN_04' classifier along the axis of the target site beginning with the 5' base and ending with the RVD closest to the FokI nuclease (Fn) domain.

than mismatches in the 3' region of the binding site [75]. Looking at the features in this classifier, we were able to see a clear trend matching this experimental observation in that the RVDs in positions 1 to N-5 were weighted very highly, RVDs in positions N-4 to N-1 less so, and the last RVD (furthest in the 3' direction) substantially lower (Figure 5-7). However, we also observed a low weighting assigned to the 5' base by the classifier.

5.5 Discussion

Evidence from the cross-validation analyses indicated that these algorithms do not appear to be over-trained, and that their performance on these test-sets provides a close approximation to how the algorithms will perform on future ZFNs and TALENs. The internal cross-validation analysis found relatively low variance between the different folds using the optimized hyperparameter settings (Figure 5-3). Furthermore, testing the algorithms on the 'hold-out' datasets showed only a slight reduction in performance compared to the average performance of the internal cross-validation (Figure 5-4). However, the small size of the 'hold-out' datasets (an unavoidable limitation given the current number of total known ZFN and TALEN off-target sites) somewhat limits the interpretation of those findings.

Since the overall goal of the algorithms is to be able to predict the off-target behavior of a novel ZFN or TALEN solely based in information gathered about other nucleases, we sought to also cross-validate the algorithms by testing their performance using each different nuclease as a test-set. Although the performance differences between individual nucleases were certainly greater than the randomized subdivisions used for the standard cross-validation procedures, the CVs for the two algorithms were both < 50%. The small number of off-target sites for some individual nucleases likely contributed to the larger variability observed in this cross-validation approach. The overall lower average precision values (the area under the curve) for the 'by nuclease' cross-validation approach—compared to the standard cross-validation approach—are due to the fact that additional 'spike-in' on-target sites were not present in the test-sets of the 'by nuclease' approach.

The algorithms derived from the machine learning approach outperformed the previously developed algorithms (Figure 5-6), however the gain in predictive power was more modest than originally anticipated. These findings may reflect the limitation imposed by the number of known off-target sites available as training data. While the two previously developed performance metrics (the percentage of total bona fide off-target sites found within the search limits and the precision within the top 16 rankings) are more intuitive, we believe that average precision is a more appropriate performance metric as it is well established in the machine learning community and better reflects the aggregate performance of the predictive algorithm. We note that while some TALEN algorithms outperformed the new 'TALEN_04' algorithm developed through machine learning in terms of the precision within the top 16 rankings, none of the previous algorithms achieved as high of an average precision score (Figure 5-6d).

In our analysis here of an expanded dataset of off-target sites, the 'v2.0' algorithms did not retain the performance advantage over the first-generation algorithms that was previously observed (Figure 3-6). Although the ZFNv2.0 and TALENv2.0 algorithms still performed comparably or better than the first-generation algorithms (Figure 5-6), they was

not clearly superior. We speculate that the less sophisticated validation approaches used in developing the 'v2.0' algorithms may have led them to be somewhat over-trained.

The weightings assigned by machine learning algorithms to different features can provide a variety of useful information. On one hand, the weightings can offer insight into potential new mechanisms of how the phenomenon in question operates, which can guide researchers towards testable hypotheses. On the other hand, if certain factors have already been experimentally validated, then seeing whether the feature weightings recapitulate those findings can be used as an assessment of how well the algorithm is capturing the underpinnings of the problem in question. In our case, we observed consensus with previous publications [75] that mismatches towards the 3' end of TALENs are better tolerated than mismatches towards the 5' end (Figure 5-7), indicating that the algorithm appropriately captured that aspect. However, the weighting of the 5' base was substantially lower than the weightings of the RVDs in the 5' region; most published evidence [79] suggests that the 5' base has a very high impact on TALEN binding. This could reflect an issue with our choice of extracted features or could simply be that the weighting of the impact of a single base and the weighting of 5 grouped RVDs are not directly comparable.

5.6 Conclusion

Using the most comprehensive list of ZFN and TALEN off-target sites compiled to date, we successfully applied machine learning techniques to create algorithms with abilities to predict bona fide off-target sites *de novo* solely based on the intended target site that are superior to any algorithms previously developed. Although the overall improvement in algorithm performance gained through the machine learning process was incremental rather than transformative, we have developed a robust framework that should continue to provide performance improvements as more bona fide off-target sites are added to the training datasets through future investigations.

CHAPTER VI

QUANTIFYING GENOME-EDITING OUTCOMES AT ENDOGENOUS LOCI WITH SMRT SEQUENCING[1]

6.1 Abstract

Targeted genome editing with engineered nucleases has transformed the ability to introduce precise sequence modifications at almost any site within the genome. A major obstacle to probing the efficiency and consequences of genome editing is that no existing method enables the frequency of different editing events to be simultaneously measured across a cell population at any endogenous genomic locus. We have developed a novel method for quantifying individual genome editing outcomes at any site of interest using single molecule real time (SMRT) DNA sequencing. We show that this approach can be applied at various loci, using multiple engineered nuclease platforms including TALENs, RNA guided endonucleases (CRISPR/Cas9), and ZFNs, and in different cell lines to identify conditions and strategies in which the desired engineering outcome has occurred. This approach facilitates the evaluation of new gene editing technologies and permits sensitive quantification of editing outcomes in almost every experimental system used (Figure 6-1).

6.2 Introduction

Genome editing with engineered nucleases is a transformative technology for efficiently modifying essentially any genomic sequence of interest [74]. This technology utilizes engineered nucleases to generate site-specific double-strand breaks (DSB) at desired genomic

[1]Modified from: Hendel A*, Kildebeck EJ*, **Fine EJ*** et al. (2014). Quantifying Genome-Editing Outcomes at Endogenous Loci with SMRT Sequencing. *Cell Reports* [48]. * These authors contributed equally to the work.

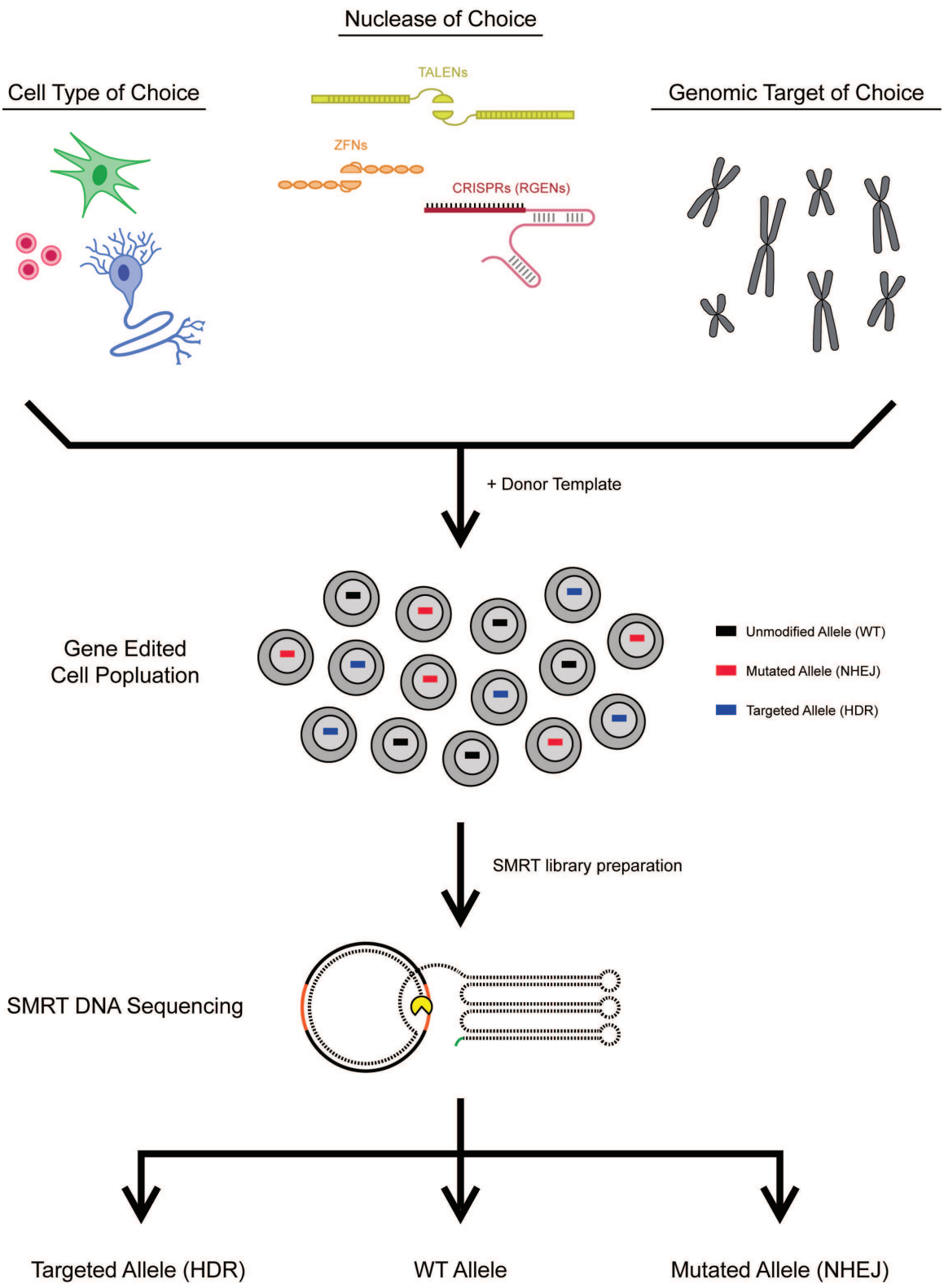

Figure 6-1: Graphical Abstract

locations followed by resolution of DSBs using the endogenous cellular repair mechanisms of nonhomologous end-joining (NHEJ) and homology directed repair (HDR) [91]. A variety of desired genetic modifications can be achieved with this approach, including mutation of a specific site through mutagenic NHEJ and precise change of a genomic sequence to a new sequence through HDR. There are currently four principal families of engineered nucleases used for gene editing: Zinc Finger Nucleases (ZFNs) [91], Transcription Activator-Like Effector Nucleases (TALENs) [10], Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR/Cas9) or RNA-guided endonucleases (hereafter called "RGENs") [38, 73], and engineered meganucleases [106]. The rapid development of these technologies is allowing for the precise alteration of genomes for numerous applications, including plant engineering [67], generation of cell lines for basic science [107], human gene therapy [120], and industrial applications [33].

When a new set of gene editing reagents is developed for a custom application, the activity levels of nucleases and the frequency of the desired gene editing event at the target locus must be determined and often need to be optimized for the specific cell type and system being used. This need has previously been met by a variety of methods including gel-based assays to measure mutagenic NHEJ [45], gene addition of fluorescent reporters to measure HDR [90, 109], analysis of large numbers of single cell clones, and the use of optimization assays to measure NHEJ and HDR at engineered reporter loci [15]. While each of these assays have their utility, each have important limitations including a lack of sensitivity required for difficult applications (gel based assays), the use of an indirect rather than a direct measure of genome editing (targeted gene addition of fluorescent reporters), and the need to generate reporter cell lines (Traffic Light Reporter system). The Traffic Light Reporter (TLR) system [15] is the only one of these assays that allows simultaneous measurement of NHEJ and HDR by expressing GFP in cells that undergo HDR-mediated correction of a GFP gene and expressing mCherry in cells with NHEJ-induced frameshift mutations. While the TLR is a very sensitive assay for measuring DSB repair (DSBR)

pathway choice, the need to generate a fluorescent reporter locus precludes measurement at endogenous target loci and thus far has prevented the use of the TLR in human primary cells. High-throughput sequencing of endogenous loci overcomes these limitations, but the range of outcomes that can be measured is limited by sequencing read-lengths. Illumina [132] and 454 [92] sequencing have recently been used to measure HDR and NHEJ outcomes when single-stranded oligodeoxynucleotides (ssODNs) or plasmids with short homology arms are used as donor templates, but the read-length limitations of these platforms do not allow analysis of longer arms of homology that drive more efficient HDR and provide the flexibility to target long gene cassettes.

Here, we present a new method for measuring genome editing outcomes at endogenous loci using single molecule real time (SMRT) DNA sequencing, which provides read-lengths approaching 15 kb and is an affordable approach that can be widely used. This technique allows for analysis of gene editing frequencies when donor templates with long arms of homology are used, which is a common strategy to increase HDR efficiency in primary cells and for the addition of large gene inserts. Using this method, we were able to measure simultaneous frequencies of NHEJ and HDR in primary cells and cell lines with greatly improved detection sensitivity. We describe the use of SMRT sequencing analysis to measure genome editing outcomes and rare large insertions generated by TALENs, ZFNs, and RGENs at the endogenous *IL2RG*, *HBB*, and *CCR5* loci. In addition, we use this system to quantify the effect of varying different parameters on the frequency of different gene editing outcomes.

6.3 Methods

6.3.1 Plasmid Construction

IL2RG TALENs were synthesized (Genscript) using the Δ152 N-terminal domain and the +63 C terminal domain as previously described [77] and fused to the FokI nuclease domain and cloned into pcDNA3.1 (Invitrogen). *HBB* TALENs were previously described in Voit

et al. [125]. *CCR5* TALENs containing the same RVDs as previously described in Mussolino et al. [80] were made using a Golden Gate cloning strategy [14] and cloned with the same N- and C- termini and nuclease domain into pcDNA3.1. CCR5 ZFNs were previously described in Perez et al. [89]. For generating RGEN expression vectors, the bicistronic expression vector (pX330, provided by Dr. Feng Zhang, and also available through Addgene #42230) expressing Cas9 and sgRNA were digested, and the linearized vector was gel purified. Oligo pairs for the *IL2RG* and *HBB* sites were annealed, phosphorylated, and ligated to linearized vectors.

The *IL2RG*, *HBB*, and *CCR5* targeting vectors were constructed by PCR amplifying arms of homology from the corresponding loci using genomic DNA isolated from K562 cells. The point mutations that, upon successful homologous recombination, would be stably integrated into the genome and prevent binding and cleavage by the engineered nucleases were added as part of the PCR primers used to generate the arms of homology. The homology arms were then cloned into a ~2900 base pair vector based on pBlueScript SK+ using standard cloning methods.

6.3.2 Cell Culture

K562 cells (ATCC) were maintained in RPMI 1640 (Hyclone) supplemented with 10% bovine growth serum, 100 units/ml penicillin, 100 μg/ml streptomycin and 2mM L-glutamine. Human $CD34^+$ hematopoietic stem/progenitor cells (HSPCs) were purchased from Lonza (2M-101B) and thawed per the manufacturers instructions. $CD34^+$ HSPCs were maintained in X-VIVO15 (Lonza) supplemented with SCF (100 ng/ml), TPO (100 ng/ml), Flt3-Ligand (100 ng/ml), IL-6 (100 ng/ml), and StemRegenin1 (0.75 μM). hESC line H1 (WiCell) was maintained in feeder-free culture conditions in mTeSR1 (Stem Cell Technologies) on a thin layer of Matrigel (BD). Cultures were passaged every 3-5 days enzymatically with Accutase (Innovative Cell technologies). Cells were transfected between passage 45 and 47.

6.3.3 Transient Transfection for Genome Editing

$1 * 10^6$ K562 cells were transfected with 2 μg TALEN-encoding plasmid and 5 μg donor plasmid (unless otherwise indicated) by nucleofection (Lonza) using program T-016 and a nucleofection buffer containing 100 mM KH_2PO_4, 15mM $NaHCO_3$, 12 mM $MgCl_2 \cdot 6\,H_2O$, 8mM ATP, 2mM glucose, pH 7.4. $4 * 10^5$ $CD34^+$ HSPCs were nucleofected with an Amaxa 4D Nucleofector with the P3 Primary Cell Nucleofector Kit (V4XP-3032) and program EO-100 per the manufacturers instructions. $1 * 10^6$ H1 cells were transfected with 0.5 μg or 2.5 μg of each TALEN-encoding plasmid and 4 μg donor plasmid (unless otherwise indicated) by nucleofection (Lonza) using an Amaxa 4D Nucleofector (program B-105) with the P3 Primary Cell Nucleofector Kit (V4XP-3032) and following manufacturers instructions.

6.3.4 Flow Cytometry

Samples were collected 72h post-nucleofection and analyzed for fluorescence using an Accuri C6 flow cytometer. GFP expression was measured using a 488-nm laser for excitation and a 530/30 bandpass filter for detection.

6.3.5 Restriction Fragment Length Polymorphism Assay

Restriction fragment length polymorphism assay was performed as previously described [16]. Briefly, Genomic DNA was extracted from transfected cells with DNeasy Blood & Tissue Kit (Qiagen). Genomic DNA was then PCR amplified with primers flanking the donor target region. The amplification was carried out with Accuprime polymerase (Invitrogen), using the following cycling condition: 95°C for 5 min for initial denaturation; 30 cycles of 95°C for 30 s, 67°C for 45 s and 68°C for 120 s; and a final extension at 68°C for 5 min. PCR products were digested with 20 U of AfIII at 37°C for ~2h and resolved with PAGE.

6.3.6 Single cell clone analysis

Single-cell cloning was performed by flow cytometry cell sorting on a BD FACSAria. Genomic DNA was isolated from single clones using the Qiagen DNeasy kit (Qiagen). The *IL2RG* target region was PCR amplified with Accuprime polymerase (Invitrogen) and the following cycling condition: 95°C for 5 min for initial denaturation; 30 cycles of 95°C for 30 s, 67°C for 45 s and 68°C for 120 s; and a final extension at 68°C for 5 min. PCR amplicons were sequenced using standard Sanger sequencing. Sequences were analyzed using the ApE plasmid editor by M. Wayne Davis.

6.3.7 SMRT Sequencing

Genomic DNA containing *IL2RG* alleles was harvested from cultured K562, $CD34^+$ HSPC, and hESC samples using the DNeasy Blood & Tissue Kit (Qiagen). *IL2RG* alleles were amplified with Accuprime polymerase (Invitrogen) and the following cycling condition: 95°C for 5 min for initial denaturation; 30 cycles of 95°C for 30 s, 67°C for 45 s and 68°C for 60 s; and a final extension at 68°C for 5 min for the K562 samples and 95°C for 5 min for initial denaturation; 30 cycles of 95°C for 30 s, 67°C for 45 s and 68°C for 90 s; and a final extension at 68°C for 5 min for the HSPC, and hESC samples. Sequencing libraries were constructed, as previously described [117], using the DNA Template Prep Kit 1.0 (Pacific Biosciences, Menlo Park, CA). SMRTbell libraries contained amplicons that were pooled together, with different barcodes appended to allow multiplex analysis. Purified, closed circular SMRTbell libraries were annealed with a sequencing primer complementary to a portion of the single-stranded region of the hairpin. For all SMRTbell libraries, annealing was performed at a final template concentration between 30 and 60 nM, with a 20-fold molar excess of sequencing primer. All annealing reactions were carried out at 80°C for 2 min, with a slow cool to 25°C at a rate of 0.1°C/second. Annealed templates were stored at -20°C until polymerase binding. DNA polymerase enzymes were stably bound to the primed sites of the annealed SMRTbell templates using the DNA Polymerase Binding Kit

2.0 (Pacific Biosciences). SMRTbell templates (3 nM) were incubated with 6 nM of polymerase in the presence of phospholinked nucleotides at 30°C for 2 h. Following incubation, samples were stored at 4°C. Sequencing was performed within 72 h of binding using final on plate concentration of 0.3 nM. Each sample was sequenced as previously described [96] using DNA Sequencing Kit 2.0 (Pacific Biosciences). Sequencing data collection was performed on the PacBio® RS (Pacific Biosciences) using C2/C2 chemistry and movies of 55 min in each case.

6.3.8 SMRT Analysis Pipeline

A SMRT library can contain amplicons from different genomic loci, with different barcodes appended to allow multiplex analysis, with indels resulting from NHEJ, or with SNPs introduced through dHDR. To separate out all of these components, as well as deal with variable read quality and other artifacts introduced by the sequencing process, we implemented an analysis pipeline using Perl (Figure 6-2).

6.3.8.1 Determining amplicon source

To begin, we take the FASTQ output files (Figure 6-2a) of the SMRT consensus sequence generation algorithm that is provided by the PacBio RS instrument. These files are labeled with the extension ".ccs.fastq" and contain the assigned DNA bases for each sequencing read as well as Phred quality values (QV) indicating the degree of confidence in the assignment of each base. Using the wild-type amplicon sequences (excluding barcodes) that would be expected to be contained in the SMRT library (Figure 6-2b), a BLAST database is constructed (Figure 6-2c). All sequence reads are BLASTed against this database to determine the amplicon from which they originated (BLAST Parameters: `gapOpen 2, gapExtend 1, reward 1, penalty -1`) and also to align the read in the proper orientation (because the SMRT polymerase can initially bind to either strand of DNA, some reads are in the "sense" orientation while others are "anti-sense"). Due to trace levels of contamination or non-specific PCR amplification, a small fraction (typically <0.5%) of the

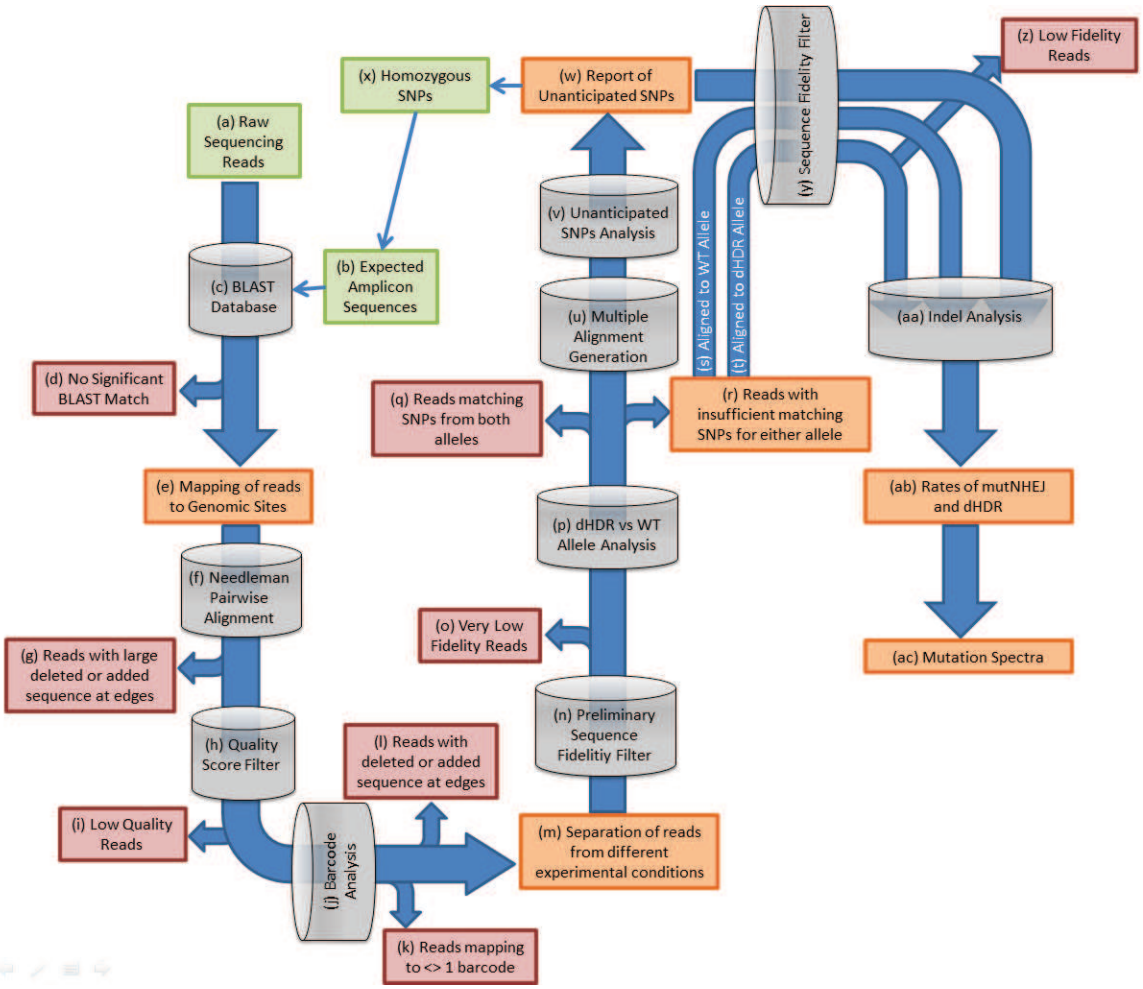

Figure 6-2: SMRT Analysis Pipeline

raw sequence reads do not have a significant BLAST match to any of the expected amplicon sequences (Figure 6-2d) and are subsequently discarded. After all BLAST queries are completed, each read has been mapped to a genomic site (Figure 6-2e).

6.3.8.2 Initial quality filtering

After reads have been mapped to a genomic site, reads are pairwise aligned to the expected amplicon sequence for further analysis (Figure 6-2f). The alignment is performed using the Needleman-Wunsch algorithm obtained from the mEmboss server (all pairwise alignments are performed with the parameters: `gapOpen 10, gapExtend 1`). PCR and the sequencing process sometimes generate reads that are missing large segments of the 5' or 3' ends of the amplicon or have concatemerized amplicons together. To remove these artifacts, the 40 proximal bases of the 5' and 3' ends of the pairwise alignments are scanned for "gaps" between the sequencing read and the expected amplicon. If more than 30 gaps are observed at either end of the alignment, the read is discarded (Figure 6-2g); typical loss at this step is 4%-15% of the sequencing reads. To decrease the computing requirements of the pipeline, a simple quality filter is applied to the reads (Figure 6-2h) and all those with a mean Phred QV of less than 40 are discarded (Figure 6-2i); typical loss at this step is 20%-40% of the remaining sequencing reads. Skipping this initial QV filter would be expected to have negligible effect on the final results because low quality reads would be removed in downstream quality filters (Figure 6-2n,y). The threshold of 40 was found through empirical testing to be optimal for decreasing total analysis time while minimizing the number of reads discarded using this simple preliminary filter that would have passed through later, more complex, quality filters (Figure 6-2n,y).

6.3.8.3 De-multiplexing experimental conditions

For investigation of multiple experimental conditions at a single genomic locus, amplicons can be barcoded and pooled together in the same SMRT library for multiplex analysis. To determine the barcode of each sequencing read (Figure 6-2j), the 5' and 3' ends of the

read are pairwise aligned against all barcode sequences for that amplicon in the library. An additional two bases of the sequence read are included in the alignment to allow for small errors generated during sequencing (i.e. for a 10 bp barcode, 10 + 2 = 12 bp of the edges of the sequencing read would be aligned to each 10 bp barcode). The alignments are analyzed for matches between a given barcode and the sequencing read. If more than three gapped bases occur in the alignment, it is determined that that barcode does not match the sequence. If ≥80% of the bases align in both the left and right barcodes, that sequence is counted as a match for that barcode set. Due to sequence similarity between barcodes and small sequencing errors, there are rare cases (0%-0.01% of sequencing reads) where a sequencing read can match (at the ≥80% threshold) more than one barcode. Any reads that match more than one barcode or are not matched to any barcode are discarded (Figure 6-2k); typical loss due to reads not matching any barcode is 4-15% of the remaining sequencing reads. The 80% threshold was found through empirical testing to be optimal for reducing the number of reads matching no barcode while keeping the number of reads matching to more than one barcode very close to zero. A further filter is applied to remove any reads that have an initial gap of more than six bases at the 5' or 3' ends of the alignment (Figure 6-2l); removal of reads by this filter is very rare (0%-0.1% of sequencing reads). After all these steps are completed, the sequencing reads have been separated using the barcodes to determine what experimental condition each read originated from (Figure 6-2m).

6.3.8.4 Secondary quality filtering

The processivity of the polymerase used in the SMRT sequencing process is highly stochastic, which leads to a wide distribution of the number of passes around the SMRTbell and therefore a wide distribution of the quality of the consensus sequence reads. Low quality reads containing small insertions, deletions, and substitutions must be removed so that they do not produce false positive reports of a sequence read showing evidence of mutNHEJ

or dHDR. Several approaches to filter low quality sequences using the Phred QV scores were attempted, but background noise still persisted, perhaps due to the lack of a 1:1 correspondence between calculated QV scores and measured empirical accuracy (Travers et al. Figure 5 [117]). Consequently, a filter to remove low fidelity sequences (Figure 6-2n) was constructed based on empirical comparison of the sequencing read to the expected sequence. The sequencing read is expected to match the wild-type genomic sequence in the positions flanking the nuclease cleavage site and excluding any SNPs introduced by the DNA Donor Template. Therefore, the accuracy of the sequencing read in these constant regions was used as a proxy for the accuracy of the sequence in the variable regions of the amplicons. Specifically, the reads were scanned for substitutions, insertions, and deletions (indels), relative to the wild type sequence that did not overlap any regions within ten bases of DNA Donor SNPs or nuclease binding sites. The sum of the squares of the lengths of the indels was divided by the length of the total search sequence and if the quotient was greater than 0.03, the read was discarded (Figure 6-2o). For example, a read containing a one bp indel and a 3 bp indel out of a total sequence length of 500 would be calculated as: $\frac{1^2+3^2}{500} = 0.02$; since $0.02 \leq 0.03$, this read would meet acceptable quality standards for this filter and would remain in the pipeline. Typical loss at this step is 4%-10% of the remaining sequencing reads.

6.3.8.5 Allele assignment

The next step in the pipeline is to determine if a sequencing read originated from an allele that had undergone homology directed repair using the DNA Donor as a template (dHDR) or from an allele that was the wild-type (WT) genomic sequence (Figure 6-2p). To perform this comparison, the read is pairwise aligned to both the expected WT and dHDR amplicon sequences (in separate steps). For each alignment, the positions of the SNPs introduced by the DNA Donor Template are examined. If the position of the SNP, in addition to the two flanking bases on each side (a total of five bases), contains no gaps or substitutions in the

pairwise alignment, that SNP is counted as a match. If the percentage of matching positions is ≥80%, then a read is declared as a match for that allele. Similar to the barcode analysis (Figure 6-2j), in rare cases (0-0.01%), a read will be declared as matching both alleles due to small sequencing errors; these reads are discarded from further analysis (Figure 6-2q). Another subset of reads will not meet the threshold to match either allele (Figure 6-2r); these are typically 0.1%-1.5% of the remaining sequencing reads. While sometimes these reads simply do not have the necessary sequence fidelity to be confidently matched, a substantial fraction consists of reads with a large deletion that has removed the entire region where the SNPs exist. As these large deletions are typically resulting from a mutNHEJ event, these reads are saved for further analysis (Figure 6-2s,t).

6.3.8.6 Detecting unanticipated SNPs

When testing different cell lines, there may often be unknown SNPs present in the cell line that are not in the reference genome which could interfere with accurate sequence fidelity, mutNHEJ, and dHDR analysis. To identify any unanticipated SNPs, all the sequencing reads in each barcode set are compiled together in a multiple sequence alignment (Figure 6-2u) which allows analysis of the base in each sequencing read that was pairwise aligned to a position in the expected amplicon sequence. In the unanticipated SNP analysis (Figure 6-2v), any positions where 40% or more of the bases in the reads are different than the expected base are flagged for manual inspection. In some cases, a homozygous SNP relative to the reference genome is identified (Figure 6-2x) and the expected amplicon sequence can be adjusted accordingly and the analysis re-run. In other cases, the SNP is heterozygous and is simply noted for downstream steps in the pipeline. Bases flagged as SNPs that occur near the nuclease cleavage site should be inspected very carefully, especially if the SNP analysis showed that most reads had a deletion instead of the expected base, to determine if this "SNP" may actually be the result of high levels of mutNHEJ modifying that position.

6.3.8.7 Final quality filtering

Once all SNPs have been identified, a final sequence fidelity filter can be applied to prune any remaining low quality sequences (Figure 6-2y). This filter works similarly to (Figure 6-2n) in that it searches for indels in the pairwise alignment and calculates an "error score". It differs in a few aspects though: any positions where a SNP was identified are not counted as an indel if the base matches the alternate SNP base, only the 100 bases flanking each side of the nuclease binding site are searched instead of the whole sequence, indels overlapping any region except the nuclease binding site and the DNA Donor SNPs are counted instead of allowing ten bases of "buffer" surrounding those regions in which indels would also not be counted, and the quality threshold for excluding low fidelity sequences is set at the more stringent 0.01 level (instead of 0.03). The 100 flanking bases are used for this filter instead of the whole sequencing read because different portions of a longer read may have different levels of quality and the regions flanking the nuclease binding site are presumed to be a better proxy for the sequence quality in the binding site than regions further away. Any reads not meeting the 0.01 error threshold are discarded (Figure 6-2z); typical loss at this step is 5%-20% of the remaining sequencing reads.

6.3.8.8 NHEJ analysis

Once the final set of high fidelity sequencing reads has been filtered, they can be analyzed for the presence of indels caused by mutNHEJ (Figure 6-2aa). Some indels in the sequencing reads may be due to errors in the SMRT sequencing process, so a set of guidelines was developed to characterize indels that would result from mutNHEJ but not from sequencing errors. For an indel to be considered to be caused by mutNHEJ, the initial requirements are that the indel must overlap the spacer region between the nuclease binding sites and be at least three bases long. If those criteria are met, then the following characteristics are considered:

a) If the indel is larger than four bases, it is considered mutNHEJ

b) If the indel consists entirely of deleted sequence, it is considered mutNHEJ

c) If the indel consists entirely of inserted sequence AND is a tandem repeat of either the immediately adjacent 5' or 3' flanking sequence or allowing up to a one base separation. Examples:

```
Wild-type Sequence:   ACACCCAGGGAATGAAGAGCAAGCGCCATGTTGAAGCCATCATT
3 bp tandem repeat:   ACACCCAGGGAATGAAGagcAGCAAGCGCCATGTTGAAGCCATC
4 bp tandem repeat:   ACACCCAGGGAATGAAGagcaAGCaAGCGCCATGTTGAAGCCAT
 1 base separation:   ACACCCAGGGAATGAAGgcaAGCAAGCGCCATGTTGAAGCCATC
```

Sequencing reads previously identified in step (Figure 6-2p) as originating from dHDR alleles theoretically should not have any mutations indicative of mutNHEJ due to the SNPs in the nuclease binding site that should inhibit binding and cleavage. However, these reads are still analyzed for evidence of mutNHEJ by the pipeline to ensure that this is the case. To date, we have not observed any reads that were identified as dHDR that showed evidence of mutNHEJ.

Once all sequencing reads have been analyzed for indels, the rates of mutNHEJ and dHDR for each barcoded experimental condition can be determined (Figure 6-2ab). Reads that had previously been unable to be definitively assigned as either originating from a dHDR or WT allele (Figure 6-2r) are analyzed for indels in two configurations: pairwise aligned to the dHDR allele (Figure 6-2s) and pairwise aligned to the WT allele (Figure 6-2t). Only sequencing reads that show evidence of mutNHEJ in BOTH pairwise alignments are considered to have been originally WT alleles in which mutNHEJ occurred. If indels indicative of mutNHEJ only occur in one of the pairwise alignments, then it is possible that it is due to small sequencing errors combined with alignments with the incorrect allele. If indels indicative of mutNHEJ do not occur in either pairwise alignment, it is still unknown which allele the read originated from and so it cannot be appropriately assigned. The rates of mutNHEJ and dHDR for a given barcode set are calculated as follows:

Let R be the number of reads from (Figure 6-2r) with mutNHEJ indels in both alignments.

Let W be the total number of reads assigned to the WT allele.

Let N be the number of reads in W with mutNHEJ indels.

Let D be the total number of reads assigned to the dHDR allele.

$\%mutNHEJ = \frac{R+N}{W+R+D}$

$\%dHDR = \frac{D}{W+R+D}$

6.3.8.9 Indel spectrum analysis

After the sequencing reads with mutNHEJ have been identified, the sizes of the different indels are analyzed to produce a mutation spectra showing the distribution of indels created (Figure 6-2ac). For indels consisting of both insertions and deletions, the resulting total change in the amplicon size is used for the distribution (i.e. an indel consisting of a +7 insertion and a -3 deletion would be recorded as a +4 overall insertion). Mutation spectra are displayed as cumulative distribution functions representing the fraction of all mutation sizes observed that are more negative than the given value.

6.3.8.10 Systemic biases

Certain steps in the analysis pipeline introduce small amounts of systemic bias in the measurements of mutNHEJ and dHDR. From the comparison of the SMRT pipeline to single cell clone analysis, it appears that these small biases largely cancel each other out to yield highly accurate results (Figure 6-4). However, for special applications of this pipeline, certain biases may become more prominent and should be kept in mind. Any biases introduced by the different steps are denoted at the end of the paragraphs as either increases (↑) or decreases (↓) to the measurements of mutNHEJ or dHDR.

SMRT sequencing process:

The diffusion mechanics of the SMRT sequencing process cause shorter amplicons to be slightly overrepresented and longer amplicons to be slightly underrepresented in the raw

sequencing reads (Figure 6-2a). Therefore, amplicons with a large deletion caused by mutNHEJ will be slightly overrepresented compared to wild-type amplicons while those with large insertions caused by mutNHEJ will be slightly underrepresented. Similarly, longer amplicons tend to have lower average sequence quality than shorter amplicons. This is due to the reduced number of passes around the SMRTbell that can be achieved with a given read length for longer compared to shorter amplicons. This results in reads with large insertions tending to be discarded at a slightly higher frequency than wild-type amplicons and reads with large deletions discarded at a slightly lower frequency (Figure 6-2 i,o,z).

↑mutNHEJ, ↓mutNHEJ

Allele analysis:

In determining whether a sequencing read originated from a dHDR allele or a WT allele (Figure 6-2p), there are a small fraction of reads that do not contain enough SNPs for definitive matching to either allele (Figure 6-2r). A large fraction of these reads contain large deletions resulting from mutNHEJ that removed the entire sequence containing the SNPs from the allele. These reads with mutNHEJ are later recovered in the indel analysis step (Figure 6-2aa), but the only way to determine if a read from (Figure 6-2r) was originally WT or dHDR is by making the assumption that mutNHEJ can only occur in WT alleles. Therefore, no reads can be recovered from (Figure 6-2r) that did not have mutNHEJ which means that in the final determination of the rate of mutNHEJ in (Figure 6-2ab), the numerator and denominator are always increased simultaneously; there are no cases where only the denominator is increased. This results in a slight overestimation of the rate of mutNHEJ. Excluding these reads would result in an underestimation of the rate of mutNHEJ and would also make the mutation spectra inaccurate. Since the mutation spectra is an analysis of the types of mutations and not the overall rate of mutation, including these reads means that the mutation spectra is accurate.

↑mutNHEJ

Indel analysis:

In determining whether or not a sequencing read has indels indicative of mutNHEJ, some small types of indels are ignored. Because the predominant error mode of SMRT sequencing is single inserted and deleted bases, these are discounted in the indel analysis to remove background noise. However, this reduction in background noise comes at the cost of a slight underestimation in the rate of mutNHEJ since it is known that mutNHEJ can produce small one or two base indels.

↓mutNHEJ

6.3.9 Statistical analysis

To calculate confidence intervals, t-statistics were calculated as previously described [88]. 90% confidence intervals were calculated by determining the upper and lower bounds of the mutation rates that would yield *P* values of 0.05. 66% confidence intervals were calculated similarly using a target *P* value of 0.32.

6.4 Results

6.4.1 Measurement of Gene Editing Outcomes at the Endogenous *IL2RG* Locus

To develop a method for quantitatively and rapidly measuring the different gene alterations occurring at an endogenous locus of interest, we used a highly active TALEN pair to stimulate DSBs at the endogenous IL-2 receptor common γ-chain gene (*IL2RG*), mutations in which are responsible for the congenital primary immunodeficiency SCID-X1 [58, 105]. For the introduction of precise sequence alterations at this locus, we designed a donor template with approximately 400bp arms of homology 5' and 3' of the TALEN cut site (Figure 6-3a). Within the 3' arm of homology we introduced seven point mutations that, upon successful HDR, are stably integrated into the *IL2RG* gene and prevent binding and cleavage by the TALEN pair (Figure 6-3a,b). To measure the frequency of mutagenic NHEJ and HDR with this system, we developed a strategy based on single molecule real time (SMRT) DNA sequencing, a high-throughput sequencing technology capable of analyzing long DNA fragments. First, the *IL2RG* locus was amplified using a forward primer that is

5' and outside the start of the 5' homology arm and a reverse primer that is downstream of the TALEN pair target site (Figure 6-3c). With this approach, non-integrated and randomly integrated donor templates are not amplified, removing common sources of background noise. The SMRT DNA sequencing technology allows for the determination of DNA sequence from individual DNA templates [31, 98]. Single-molecule read lengths approaching 15 kb were reached in this study, with an average read length approaching 3 kb. For DNA fragments shorter than the read limit of the polymerase, improved sequence accuracy (frequently reaching an average Phred QV score of 40, denoting 99.99% accuracy) is achieved by iteratively sequencing the same circular DNA template [117] (Figure 6-3c).

To induce sequence alterations in *IL2RG* we expressed the *IL2RG* TALENs from plasmid DNA with or without introduction of donor DNA, and then analyzed cell populations by SMRT DNA sequencing. Following transfection with TALENs alone we detected unmodified alleles and alleles with deletions or insertions indicative of mutagenic NHEJ (Figure 6-3d). When cells were transfected with both the TALENs and donor DNA, we detected unmodified alleles, alleles with deletions or insertions, and alleles with the 7 point mutations precisely integrated into *IL2RG* by HDR. Notably, no alleles were detected with both the 7 point mutations and indels indicative of NHEJ, validating the ability of the point mutations to prevent TALEN cleavage of HDR-modified alleles. High frequencies of 'on-target' *IL2RG* modification were observed in K562 cells under these conditions, with 18% of alleles mutated by NHEJ and 17% of alleles precisely modified by HDR (Figure 6-3d). Due to the PCR strategy being used, where one primer is outside of the donor template arm of homology, essentially no background signal was detected from amplification of non-integrated or randomly integrated donor template. Control experiments (either mock or donor only transfections) had low background frequencies of NHEJ and HDR reads resulting from PCR or DNA sequencing errors, which ranged from 0.00% to 0.03% for individual samples with average frequencies of 0.007% NHEJ and 0.001% HDR. This low

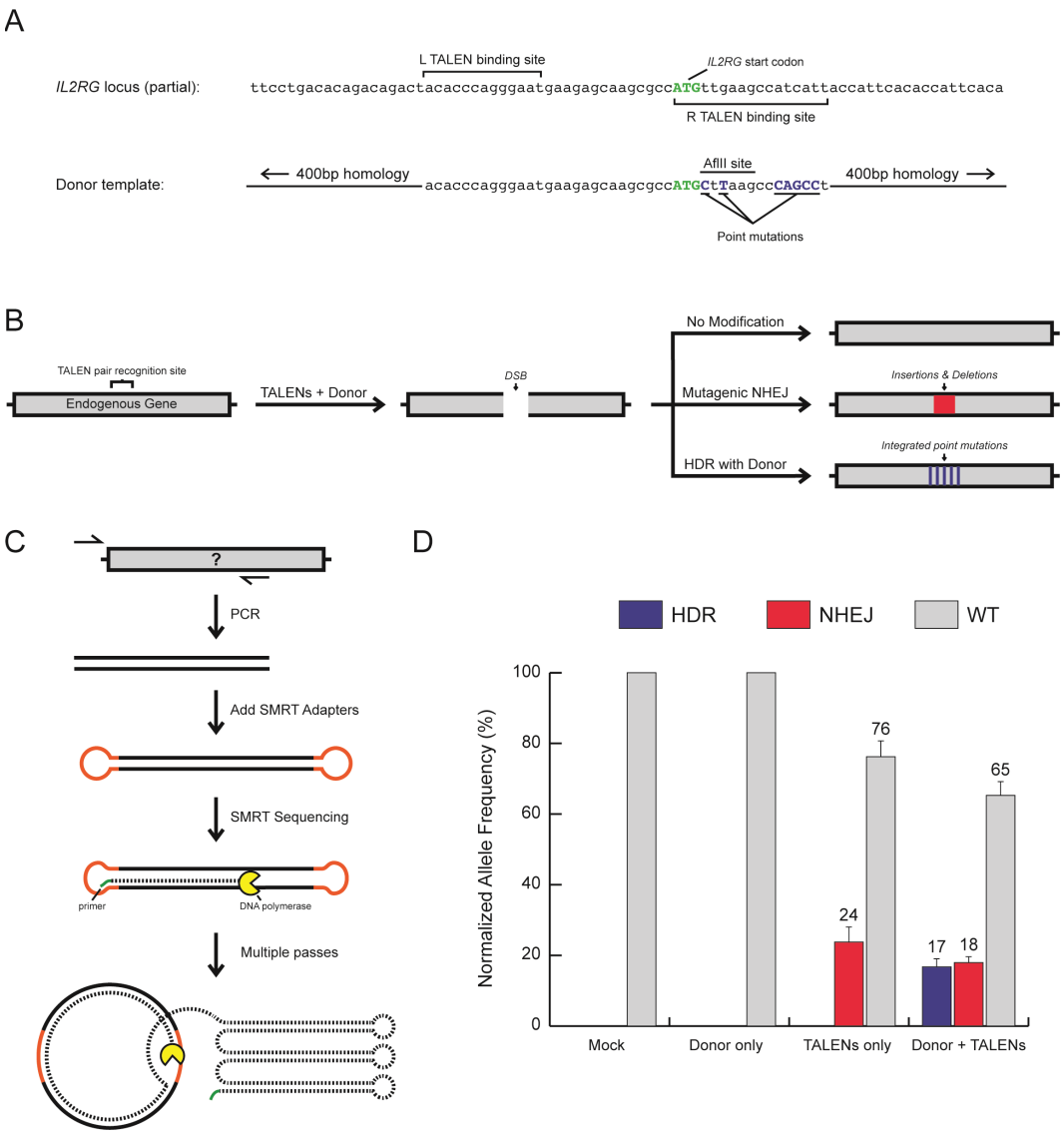

Figure 6-3: Measuring gene editing at an endogenous locus with SMRT sequencing. (**a**) Sequence of the TALEN target site at the *IL2RG* locus and the *IL2RG* donor template. The donor template harbors seven point mutations that, when integrated into *IL2RG*, create silent mutations and a novel AflII restriction site. These substitutions alter the right TALEN binding site and provide a signature for alleles precisely modified by HDR. (**b**) Diagram of gene editing at an endogenous locus. TALENs create a double strand break (DSB), which can lead to no modification, insertion or deletion mutations, or integration of point mutations from the donor template. (**c**) Schematic of SMRT DNA sequencing analysis. The endogenous locus is amplified by PCR, with at least one primer outside the arms of homology of the donor template, and SMRT adapters are added to PCR amplicons. Individual DNA molecules are sequenced by SMRT sequencing, with read lengths averaging ~3kb in length and approaching ~15kb. (**d**) Measurement of gene editing outcomes at the *IL2RG* locus in K562 cells. Modification frequencies are normalized to transfection efficiency. Bars represent three independent biological replicates; error bars, s.d.

level of background noise, coupled with the high-throughput nature of this approach, produces a high level of sensitivity and creates new possibilities for studying rare DNA repair events.

6.4.2 Reliability of SMRT Sequencing Analysis at a Single Endogenous Locus

To validate the accuracy of the SMRT DNA sequencing strategy, we compared our high-throughput results with standard gel-based assays and single cell clone analysis of K562 cells. First, we used a restriction fragment length polymorphism (RFLP) assay to measure the frequency of HDR by measuring the presence of an AflII restriction site that is created when the 7 point mutations within the donor template are precisely incorporated into the target locus (Figure 6-4a). Using the RFLP assay, the AflII restriction site was detected in an average of 14.3% of alleles normalized for transfection efficiency compared to 16.8% of alleles by SMRT sequencing analysis of the same populations. The most commonly used methods for determining the frequency of NHEJ measure any small deletion or insertion events, which is confounded by sequence alterations introduced by HDR. To independently determine the true frequency of alleles modified by NHEJ and HDR, we grew single cell clones from a representative sample. Analysis of these clones showed that 11.3% of alleles had undergone mutagenic NHEJ and 11.1% of alleles had been precisely modified by HDR, compared to frequencies of 11.2% and 11.0% respectively as measured by SMRT sequencing analysis of the same population (Figure 6-4b); SMRT frequencies represent the total population frequency, not normalized to transfection efficiency, in order to directly compare to the clonal analysis. To confirm the reproducibility of SMRT sequencing analysis, we analyzed a single targeted population eight times and found standard deviations for NHEJ and HDR of 0.66% and 0.79% respectively (Figure 6-4c). This experimental variance between samples was only slightly higher than the expected statistical variance for the number of sequences analyzed, demonstrating the reliability of this approach (Figure 6-4d).

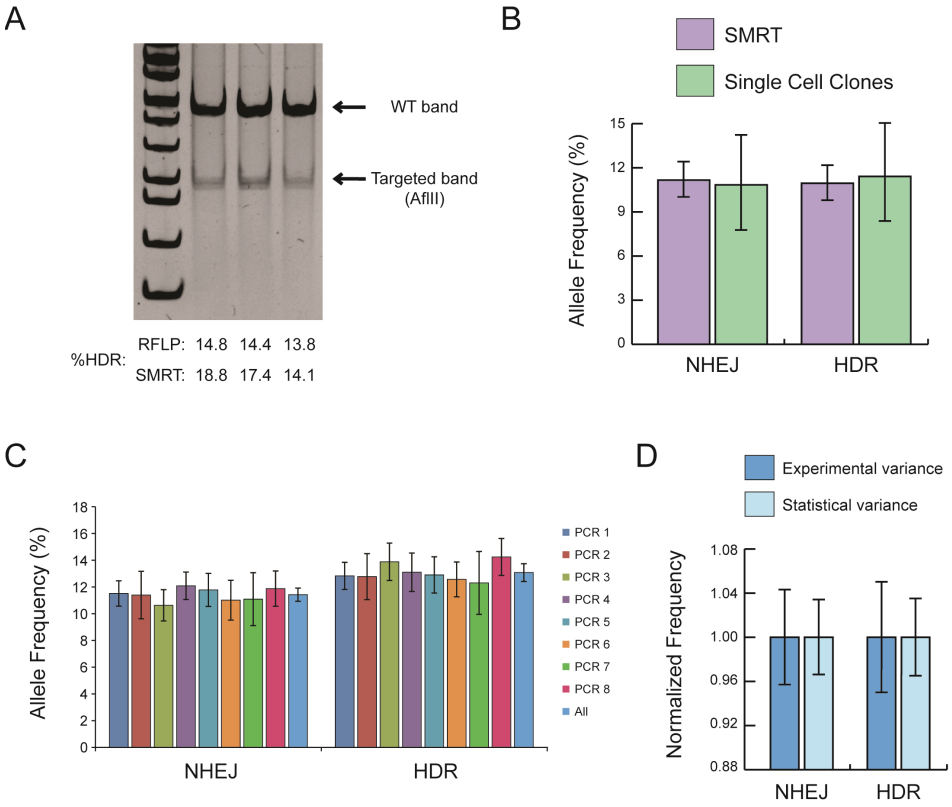

Figure 6-4: Reliability of SMRT sequencing analysis for measuring gene editing outcomes at an endogenous locus. (**a**) RFLP analysis of K562 cells targeted with 1 μg of each TALEN and 5 μg donor in triplicate. The frequency of HDR in each sample as measured by RFLP and SMRT sequencing analysis is shown. (**b**) Quantification of NHEJ and HDR frequencies in single cell clones grown from a representative population of K562 cells. Error bars represent 90% confidence intervals. (**c**) A representative sample of targeted K562 cells was analyzed by SMRT sequencing 8 separate times to determine the variability introduced by PCR, SMRT library synthesis, and sequencing. Error bars represent 90% confidence intervals. (**d**) Quantification of the observed experimental variation compared to the expected statistical variation for the number of sequences analyzed for the 8 replicates. Error bars for experimental variation represent standard deviation. Error bars for statistical variation represent 68% confidence intervals (corresponding to the fraction of the normal distribution covered by ±1 standard deviation).

6.4.3 Quantification of Gene Editing at the *IL2RG* Locus in Primary Cells

For novel gene editing applications, moving from known conditions in commonly-used cell lines to more difficult experimental platforms, such as induced pluripotent stem cells or primary cells, poses a significant challenge. Using the gene editing tools previously described, we next tested our ability to measure gene editing events in $CD34^+$ hematopoietic stem/progenitor cells (HSPCs) and human embryonic stem cells (hESCs), both of which are difficult to target, but important cell types for basic research and gene therapy. After introducing TALENs and donor DNA into $CD34^+$ HSPCs, SMRT sequencing analysis showed frequencies of mutagenic NHEJ and HDR of 7% and 1% respectively at the endogenous IL2RG locus (Figure 6-5a). In hESCs, which commonly require enrichment of targeted clones due to low gene editing efficiencies, addition of our gene editing reagents resulted in mutagenic NHEJ and HDR frequencies of 0.10% and 0.14% respectively (Figure 6-5b). Control hESC samples transfected with only donor DNA showed background frequencies of 0.02% NHEJ and no HDR, illustrating the very low level of background noise for this technique. Our transfection efficiency for these hESC populations was approximately 20%, suggesting that with enrichment for transfected cells we would generate modification frequencies of 0.5-0.7%. These numbers are consistent with the numbers published by Soldner et al. [107], who showed that after sorting for highly transfected hESCs targeting frequencies of 0.4-0.8% were obtained. Importantly, since *IL2RG* is silent in both of these cell types, these results demonstrate the ability of this approach to provide quantitative and sensitive measures of gene editing and DNA damage repair at a silent endogenous locus in primary cells.

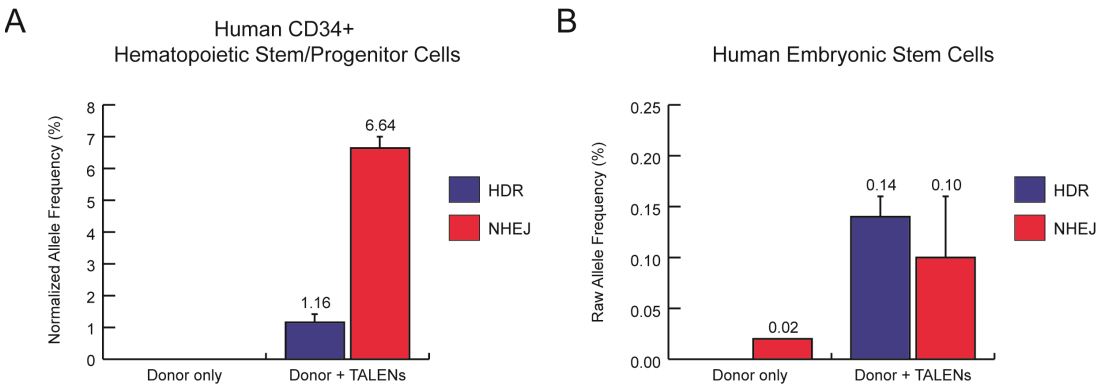

Figure 6-5: Measurement of genome editing at an endogenous locus in human primary cells. (**a**) Measurement of gene editing outcomes at *IL2RG* in CD34$^+$ HSPCs using the high-expression TALEN plasmids. (**b**) Measurement of gene editing outcomes at *IL2RG* in hESCs using the high-expression TALEN plasmids.
Bars represent three independent biological replicates; error bars, s.d.

6.4.4 Analysis of Gene Editing with TALENs, RGENs, and ZFNs at the Endogenous *IL2RG*, *HBB*, and *CCR5* Loci

The extraordinary expansion of gene editing technologies over recent years has created a plethora of opportunities for researchers attempting to modify genomes. With the introduction of TALENs and RGENs, it is now possible to generate tens to hundreds of candidate nucleases in a matter of weeks, or even days, and target multiple genomic sites simultaneously [11, 19, 61, 97, 131]. Via simultaneous analysis of different genomic sites and conditions in a single SMRT sequencing run, this approach has the potential to rapidly expedite the process of characterizing nuclease activities and optimizing targeting parameters. To determine the ability of this method to measure the activities of different classes of nucleases at multiple genomic sites, we treated cells with TALENs, RGENs, and ZFNs designed to target the *IL2RG*, *HBB*, and *CCR5* genes and analyzed gene modification.

To test the relative activity of TALENs and RGENs at the *IL2RG* locus we first constructed an RGEN with a target site overlapping the target site of the *IL2RG* TALENs (Figure 6-6a). Expression of the RGEN in K562 cells generated targeted mutations in 37% of *IL2RG* alleles compared to only 13% of alleles using TALENs. Introduction of donor

template DNA with either the RGEN or TALENs produced alleles modified by mutagenic NHEJ and alleles precisely modified by HDR. As expected, the more active RGEN stimulated a higher level of HDR than the TALENs with 33% and 22% of alleles harboring the integrated SNPs respectively. Moving to a different genomic locus, we used a TALEN pair and an RGEN targeting the *HBB* gene, mutations in which are responsible for sickle cell anemia and thalassemia [40] (Figure 6-6b). At *HBB*, the RGEN again produced significantly higher frequencies of gene disruption than the TALENs and stimulated higher frequencies of HDR upon the introduction of donor template. When expressed with donor template, the *HBB* RGEN mutated 41% of *HBB* alleles while the *IL2RG* RGEN mutated 21% of *IL2RG* alleles, suggesting that more DSBs were being created at the *HBB* locus. Despite this increase in mutagenesis, the simultaneous level of HDR was only 14% at *HBB* compared to 33% at *IL2RG*. Thus, total modification levels at *HBB* and *IL2RG* were highly similar, 55% and 54% respectively, but the ratio of HDR to NHEJ was markedly lower at *HBB* (0.34:1) than at *IL2RG* (1.6:1) (Figure 6-6a,b). This large difference in the efficiency of precise gene targeting suggests that there could be intrinsic differences between these loci affecting their ability to participate in plasmid-mediated gene targeting by HDR.

To further confirm the utility of SMRT sequencing analysis to measure targeted genomic modifications for multiple classes of nucleases, we compared the activity of previously reported ZFNs and TALENs designed to target the *CCR5* locus [80, 89]. As seen at *IL2RG* and *HBB*, addition of nucleases led to targeted disruption of the endogenous gene by NHEJ and the further addition of donor template DNA stimulated targeted integration through HDR (Figure 6-6c). The *CCR5*-specific ZFNs, variants of which are currently being used in clinical trials for HIV [89, 115], produced higher levels of modification at the endogenous *CCR5* locus than TALENs designed to an overlapping target site. It is important to note that in this study, stable modifications in K562 cells were measured for TALENs, RGENs, and ZFNs 14 days post-transfection. The absolute genome editing frequencies reported here are thus somewhat different than published results for previously

described nucleases where nuclease activity was measured at different time points, in different cell types, and with different nuclease levels [80, 89, 126]. Nuclease-induced NHEJ is typically measured with high nuclease levels 3 days post-transfection to detect maximal NHEJ levels, but these modifications decrease over time due to toxicity [29, 30, 59, 70]. Instead of measuring NHEJ and HDR separately or with different transfection conditions, SMRT DNA sequencing provides a simple alternative for comparing stable NHEJ and HDR frequencies simultaneously.

6.4.5 Optimization of Gene Targeting Parameters

In addition to comparing different nuclease platforms, we have also used the SMRT DNA sequencing approach to study different variables that might affect genome editing outcomes. To explore how varying the dose of TALENs effects gene editing frequencies we measured frequencies of NHEJ and HDR at *IL2RG* while progressively decreasing the amount of TALENs transfected in K562 cells. Keeping the amount of donor DNA constant and titrating down the amount of TALENs by 100-fold, we saw a progressive decrease in both mutagenic NHEJ and HDR events while their relative frequencies remained largely unchanged (Figure 6-7a). Using this approach we were able to reliably detect gene editing outcomes at frequencies ranging from >20% at high TALEN levels to ≤0.1% at very low TALEN levels. Even at modification frequencies of 0.1-0.4%, relative activity levels were easily distinguished with this approach. Next, to test conditions for maximizing the frequency of HDR at *IL2RG*, we investigated the effect of the amount of donor template DNA on gene modification. Keeping the amount of TALENs constant and titrating the amount of donor DNA, we saw the overall level of modification at *IL2RG* remain relatively constant while the total level of HDR rose from 1.6% to 17.8% at optimal levels (Figure 6-7b). With this rise in the contribution of HDR, the ratio of HDR to NHEJ increased from 0.12 to 1.37 with increasing abundance of donor DNA (Figure 6-7c).

Another important variable for efficient HDR-mediated DNA repair is the length of

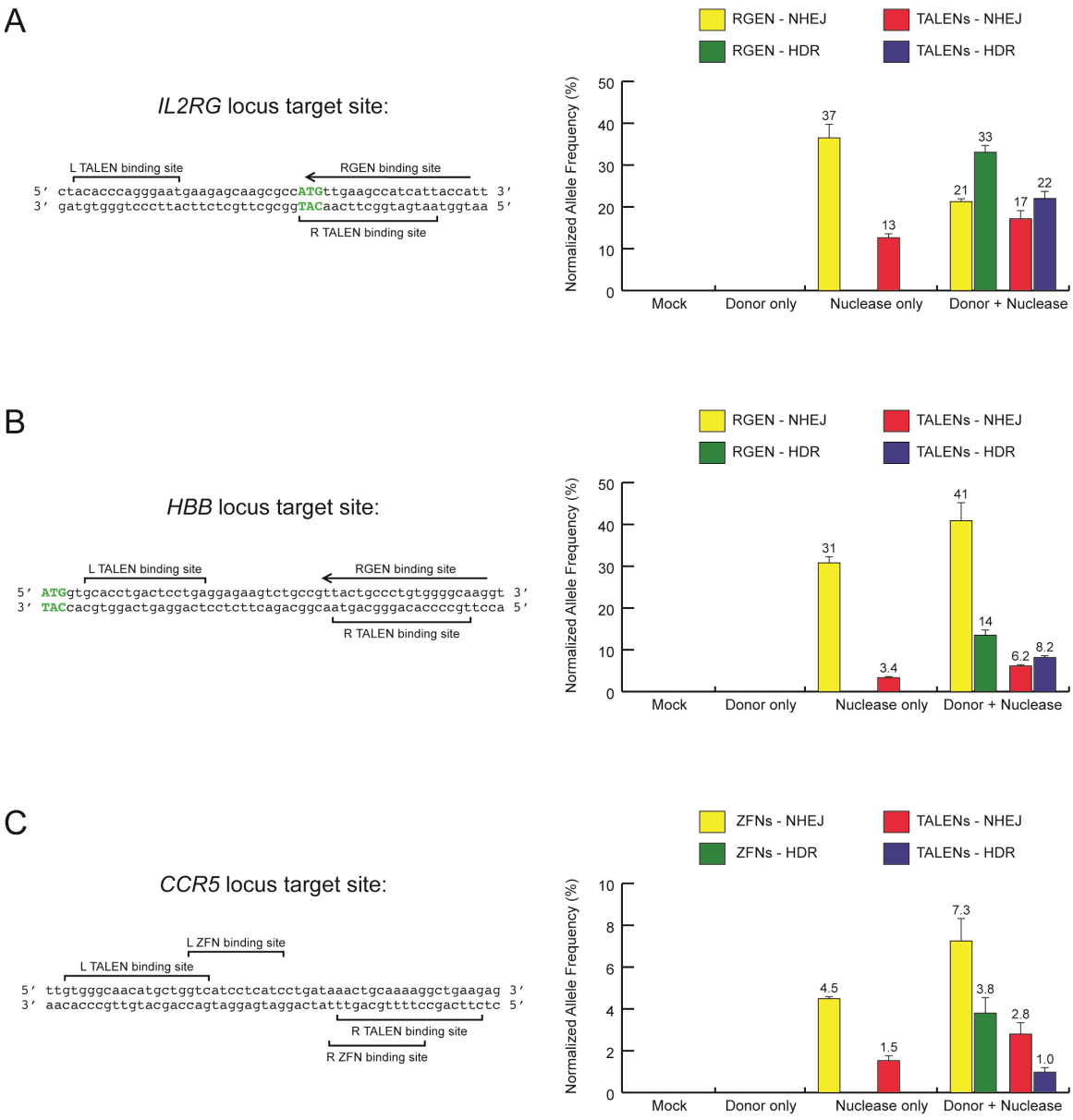

Figure 6-6: Measuring gene editing with different engineered nuclease platforms at different genomic targets. (**a**) Left; *IL2RG* target site for TALENs and RGEN guide sequence. The *IL2RG* start codon is shown in cyan. Right; Modification of the *IL2RG* locus in K562 cells. (**b**) Left; *HBB* target site for TALENs and RGEN guide sequence. The *HBB* start codon is shown in cyan. Right; Modification of the HBB locus in K562 cells. (**c**) Left; *CCR5* target site for TALENs and ZFNs in exon 3. Right; Modification of the *CCR5* locus in K562 cells.
Bars represent three independent biological replicates; error bars, s.d.

the homologous regions in donor DNA, which has been shown to vary between species and in different cell types of the same species [8, 83, 127]. To determine the effect of homology arm length on HDR efficiency with plasmid donors, we constructed a series of donor templates with a range of homology arm lengths from 800bp to 100bp (Figure 6-7d). At the *IL2RG* locus in K562 cells, homology arms 100bp or 200bp in length were found to be significantly less effective for HDR than plasmid donor templates with 400bp or 800bp homology arms. As would be predicted, the homology arm length did not change the frequency of mutagenic NHEJ. In this cell type, 400bp arms of homology actually resulted in the same levels of HDR as more commonly used 800bp arms, suggesting that for gene targeting in some human cell types maximal levels of HDR may be achieved with relatively short 400bp arms of homology. For all areas of optimization, however, the specific setting of the cell type and the chromosomal locus under investigation should be taken into consideration.

6.4.6 SMRT Sequencing of Genome Editing Outcomes Reveals Genomic and Plasmid DNA Sequences Captured into Targeted Sites

A unique feature of the SMRT sequencing method is the combination of high-throughput with long sequence read-lengths. This combination allowed us to see rare mutations including large insertions and deletions hundreds of base pairs (bp) in length. Analysis of mutations at the *IL2RG*, *HBB*, and *CCR5* loci showed a wide range of insertions and deletions ranging from +334 bp to -412 bp. Interestingly, when we BLASTed the inserted sequences that were >30 bp against NCBI non-redundant nucleotide databases, we were able to identify sequences originating from the same chromosome as the target site, nonhomologous chromosomes, plasmid DNA, and the *E. coli* genome. Representative samples of such insertion events can be seen in Figure 6-8. At each end of an insertion event, there is a junction with the chromosomal segment on one side and the inserted segment on the other side. Processing of the chromosomal sides of the junctions can be tracked by examining the sequences at these sites. NHEJ-mediated DSBR commonly involves deletion of a few

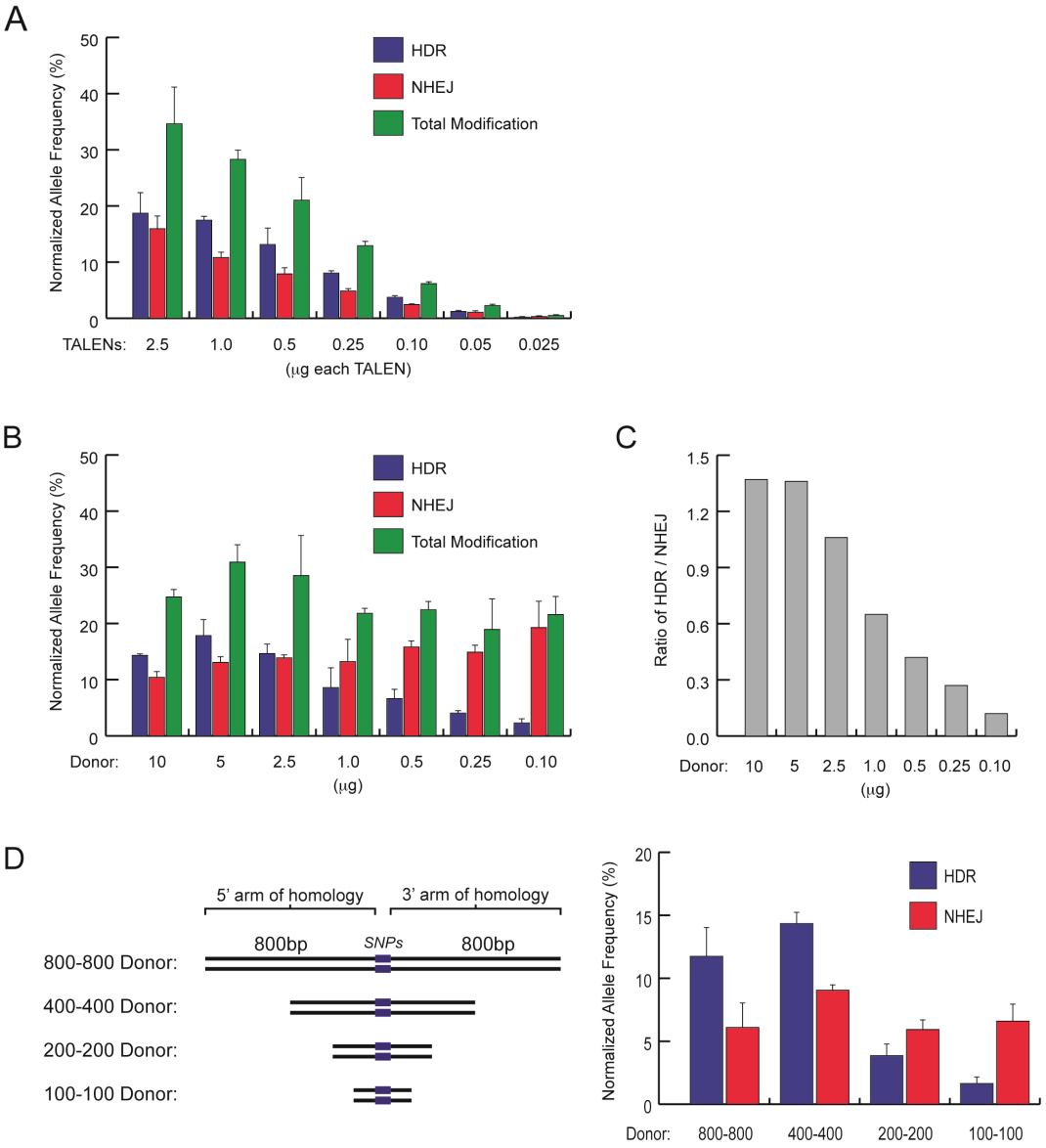

Figure 6-7: Optimization of gene editing parameters at *IL2RG* with SMRT sequencing. (**a**) Titration of TALEN amount in K562 cells with amount of donor DNA held constant at 5 μg. (**b**) Titration of donor DNA amount in K562 cells with TALEN DNA amount held constant at 1 μg each TALEN. (**c**) Ratio of HDR to NHEJ for samples in (b). (**d**) Left; Schematic of donor templates with varying arm of homology lengths. Right; Quantification of effect of homology arm length on gene editing frequencies in K562 cells.
Bars represent three independent biological replicates; error bars, s.d.

nucleotides. Surprisingly, of the 42 long insertions we found only 7 that involved deletions of nucleotides. The remaining 35 did not involve even single nucleotide deletions. This finding suggests that the insertion event may have a role in protecting chromosomes from harmful deletions during DSBR. Microhomology is known to drive some small insertion and deletion events during NHEJ. But analysis of the sequences flanking the external sources of these long insertion events did not reveal any trends suggesting that these events were driven by flanking microhomologous sequences.

6.5 Discussion

While the simplicity and flexibility of engineering TALENs and RGENs has transformed gene editing over recent years, many questions remain about DSBR processes and gene targeting in different cell types and at different genomic sites. Here, we present a rapid, accurate, and sensitive strategy for analyzing gene editing outcomes and DSBR pathway choice at endogenous loci in potentially any cell type using any type of engineered nuclease. The SMRT DNA sequencing strategy offers three principle advantages over currently available techniques:

1) Sensitive measurement of genome editing in any cell type, including primary stem cells, without the need to make a stable reporter cell line

2) Measurement of modifications at endogenous loci regardless of transcriptional status

3) Long sequencing read-lengths that allow insight into a wide range of DNA repair outcomes when donor templates with long arms of homology are used

Without generating reporter cell lines, we used the SMRT DNA sequencing strategy to measure gene editing outcomes in $CD34^+$ HSPCs, hESCs, and K562 cells. Measurement at the IL2RG locus was not inhibited by the lack of transcription of this gene in $CD34^+$ HSPCs and hESCs, demonstrating the ability of this method to provide quantitative and sensitive analysis of silent endogenous loci. Epigenetic status is known to affect all

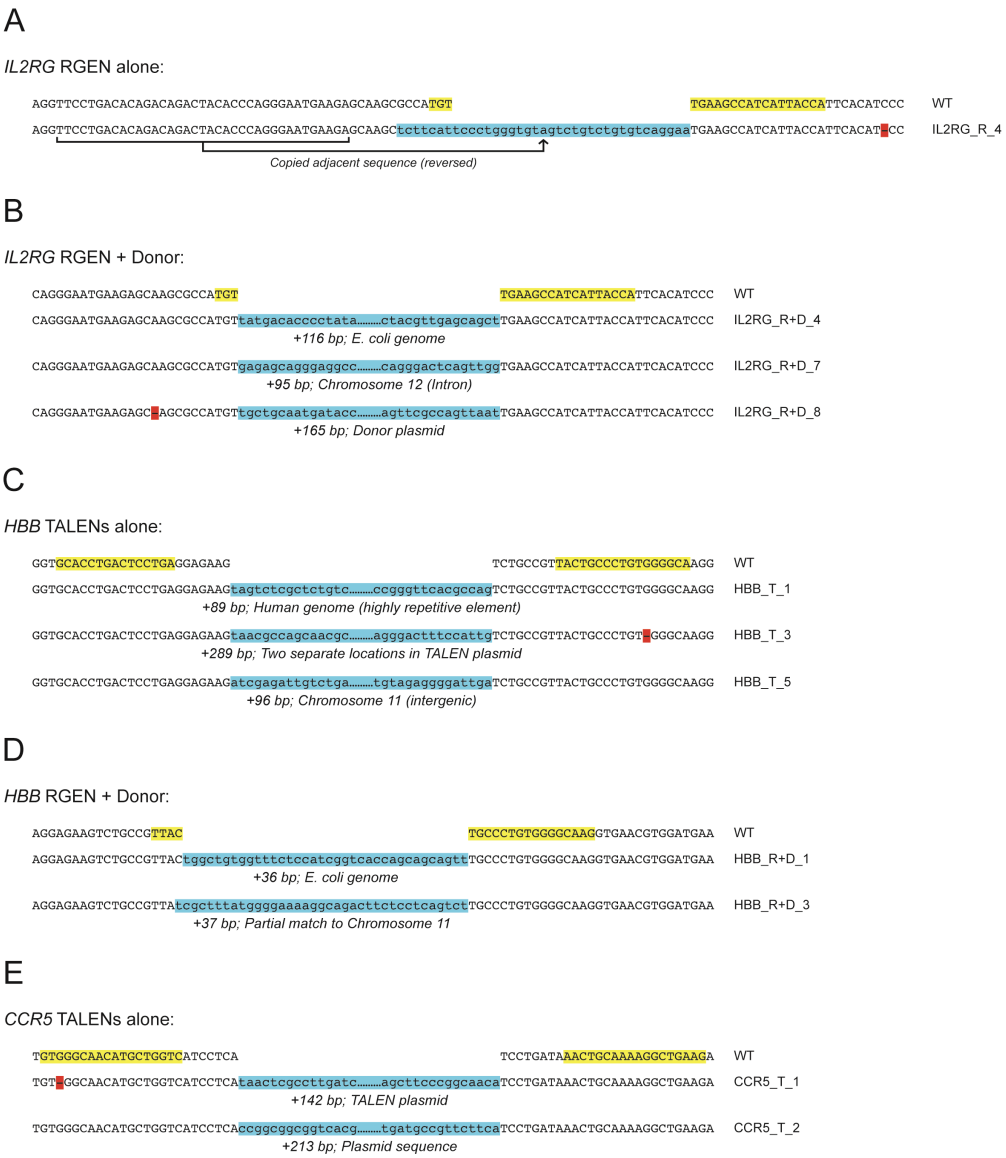

Figure 6-8: DNA repair by insertion of large sequences from various sources. (**a**) An event from the *IL2RG* RGEN-only transfections where the insert is an exact repeat—in the inverse orientation—of a sequence near the cleavage site. (**b**) Reads from the *IL2RG* RGEN and Donor transfections containing inserts derived from the *E. coli* genome, an intronic sequence in chromosome 12, and a Donor plasmid. (**c**) Several representative SMRT reads from the *HBB* TALENs-only transfections were recovered containing inserts derived from a repetitive genomic element, one of the TALEN plasmids, and a region in chromosome 11 on the opposite arm from the *HBB* gene. (**d**) Reads from the *HBB* RGEN and Donor transfections containing inserts derived from the *E. coli* genome and a region in chromosome 11 on the opposite arm from the *HBB* gene. (**e**) Reads from the *CCR5* TALENs-only transfections containing inserts derived from a TALEN plasmid and an unknown plasmid. Wild-type (WT) sequences are shown with nuclease binding sites highlighted in yellow. Inserted sequences in the SMRT reads are lowercase and highlighted in blue with the size and source listed below. Deleted bases, likely resulting from sequencing errors, are represented by hyphens highlighted in red.

DNA-metabolism processes including transcription, replication, and repair [13, 86]. The importance of transcriptional activation and epigenetic status for gene editing efficiency is still largely unknown, but epigenetic modification was recently shown to impact DSBR pathway choice [24, 63, 122, 123, 124]. The SMRT DNA sequencing strategy could be used to further study how chromatin status influences DSBR pathway choice and gene editing efficiency by providing analysis in a broader range of cell types in which the chromatin state of the targeted site is known. One other potentially important variable for gene editing efficiency, particularly when working between different cell types, is the method of delivery and this strategy could be used to quantitatively measure the impact of using different delivery methods including electroporation-based techniques, viral-based strategies, and lipid or nanoparticle-based methods.

Genome editing with engineered nucleases can be used to create many types of changes to a genome, and any site within an organisms genome is now a potential target. The versatility of this approach, combined with the ease of synthesizing new nucleases, creates a need for a method to evaluate different types of nucleases at different genomic locations. In this study we used SMRT DNA sequencing analysis to measure genome editing at sites within the *IL2RG*, *HBB*, and *CCR5* genes using the three most widely used classes of engineered nucleases. For the specific nucleases we investigated in K562 cells, we found that at both the *IL2RG* and *HBB* loci the RGEN generated significantly higher frequencies of mutagenic NHEJ than TALENs designed to overlapping sites. When transfected alone, the *IL2RG*-specific and *HBB*-specific RGENs created very similar levels of mutagenic NHEJ, suggesting that a similar amount of DSBs are being created at the two loci. Despite this, the *HBB*-specific RGEN stimulated a significantly lower frequency of HDR than that seen at *IL2RG*. Whether this difference in repair pathway utilization is the result of different chromatin status or the sequence composition of the target sites and corresponding donor DNA is still unclear, but this technique could be applied to further elucidate how spatial parameters affect DNA repair. Additionally, moving between genomic loci we have encountered

single nucleotide polymorphisms (SNPs) that confound measurement of NHEJ using standard mismatch detection assays [45], including one at *HBB* in K562 cells. The ability to use SMRT DNA sequencing to quantify mutation frequencies even in the presence of SNPs is another advantage of this system.

In addition to the gamut of new nucleases and target sites, molecular and genetic strategies to influence DSBR pathway choice can play a significant role in achieving a desired outcome or minimizing unwanted outcomes. By titrating the amount of donor template in K562 cells we were able to optimize conditions for generating HDR events and alter the ratio of HDR to NHEJ significantly. Furthermore, the long read-lengths of SMRT DNA sequencing allowed us to measure gene editing outcomes using donor templates with 800bp arms of homology. These data demonstrate the advantage of using long arms of homology to stimulate higher frequencies of HDR with plasmid donors, and this technique could further be used to directly compare gene editing outcomes with different donor architectures including plasmids, minicircle DNA, viral vectors, and ssODNs [16, 17, 70]. Beyond targeted gene editing, this method offers a new experimental system for studying DNA repair pathway utilization when DSBs occur at endogenous genomic loci following manipulation of DNA repair genes. As was shown in this study, and previous studies, parameters like nuclease properties, donor template architecture, cell type, and the site being modified can influence DSBR. The SMRT DNA sequencing method thus opens up new possibilities for studying DSBR with engineered nuclease-induced breaks, where previous work has focused significantly on breaks induced by I-SceI at defined sites within reporter cell lines.

One area where application of SMRT DNA sequencing is challenging is the quantification of gene modifications that result in differently sized alleles, such as when entire linear donor templates are captured by NHEJ at 'on-target' and 'off-target' DSBs [23, 37]. PCR bias when amplifying WT and modified loci with significantly different sizes can favor shorter alleles and confound quantification, which is further effected by loading bias for shorter molecules in the SMRT sequencing cells. For analysis of large gene inserts

mediated by HDR, we have overcome this obstacle using embedded primers that distinguish between targeted and WT allele sequences while producing similar PCR amplicon sizes [125]. Simultaneous measurement of amplicons with different lengths has also been achieved by adding a size standard ladder to the SMRT sequencing reaction, and a similar strategy could be used for quantification of large gene additions or NHEJ-mediated integrations of the donor template [71].

By analyzing thousands of alleles within cell populations modified by TALENs and CRISPRs, this technique revealed the presence of rare insertional events where large stretches of DNA from other sources were integrated at nuclease cleavage sites. Choosing a cutoff of >30bp to exclude sequences generated by DNA polymerase, we analyzed these inserted sequences and found matches to the donor template and nuclease expression plasmids introduced for gene targeting, sequences from nearby chromosomal sites, sites on other chromosomes, and sites within the *E. coli* genome that may have originated from trace impurities from the plasmid purification process. The presence of these events highlights the importance of minimizing the amount of exogenous DNA added for gene targeting and illustrates the potential for SMRT DNA sequencing to measure large, rare sequence alterations at sites throughout the genome.

6.6 Conclusion

The recent explosion in custom gene editing technologies is ushering in a new age of genome engineering where scientists across fields of study and using different organisms and cell types can precisely modify essentially any locus they desire. Here we show that SMRT DNA sequencing provides a simple, rapid, quantitative, and sensitive strategy for measuring genome editing outcomes with different cell lines, at any endogenous loci, including transcriptionally silent loci, and using multiple nuclease platforms. Moreover, our strategy offers a new approach for studying DNA repair pathway utilization when DNA

breaks occur within genomic sites that have been difficult to study using previous methodologies. With the flexibility to evaluate new engineered nucleases and targeting constructs directly at desired loci without the development of reporter systems, SMRT DNA sequencing can streamline the development of genome editing projects and hasten the expansion of these technologies to a wider range of applications.

CHAPTER VII

FURTHER ANALYSIS OF NHEJ VS HDR

7.1 Introduction

With the new tool developed in Chapter 6 to simultaneously analyze the NHEJ and HDR repair pathways at endogenous loci, we conducted a variety of different experiments aimed at further understanding how to optimize gene repair therapeutic strategies.

7.2 Comparing mRNA vs Plasmid Delivery of TALENs[1]

There have been previous comparisons of the performance of nucleases encoded as mRNA vs plasmids (DNA) in microinjection experiments [116], but very limited data on the performance in bulk cell populations. mRNA tends to be better tolerated by cells than plasmid DNA and can therefore be given in higher doses, which would lead to higher expected nuclease activity. On the other hand, expression from mRNA is much more transient than from DNA plasmids, which would lead to lower expected nuclease activity. We sought to determine the overall effect on nuclease activity as well as to observe if any differences in the HDR:NHEJ ratio occurred.

We nucleofected K562 cells with mRNA (synthesized by TriLink Biotechnologies, San Diego CA) or plasmid DNA encoding TALENs targeting the IL2RG gene along with a donor plasmid (as described previously in Figure 6-3a). SMRT analysis was used to determine the rates of HDR and NHEJ repair.

We found that the was no significant difference in HDR:NHEJ ratios between the two different methods, however mRNA delivery resulted in substantially higher levels of both NHEJ and HDR (Figure 7-1).

[1] In collaboration with Dr. Eric Kildebeck and Dr. Matthew Porteus, Stanford University.

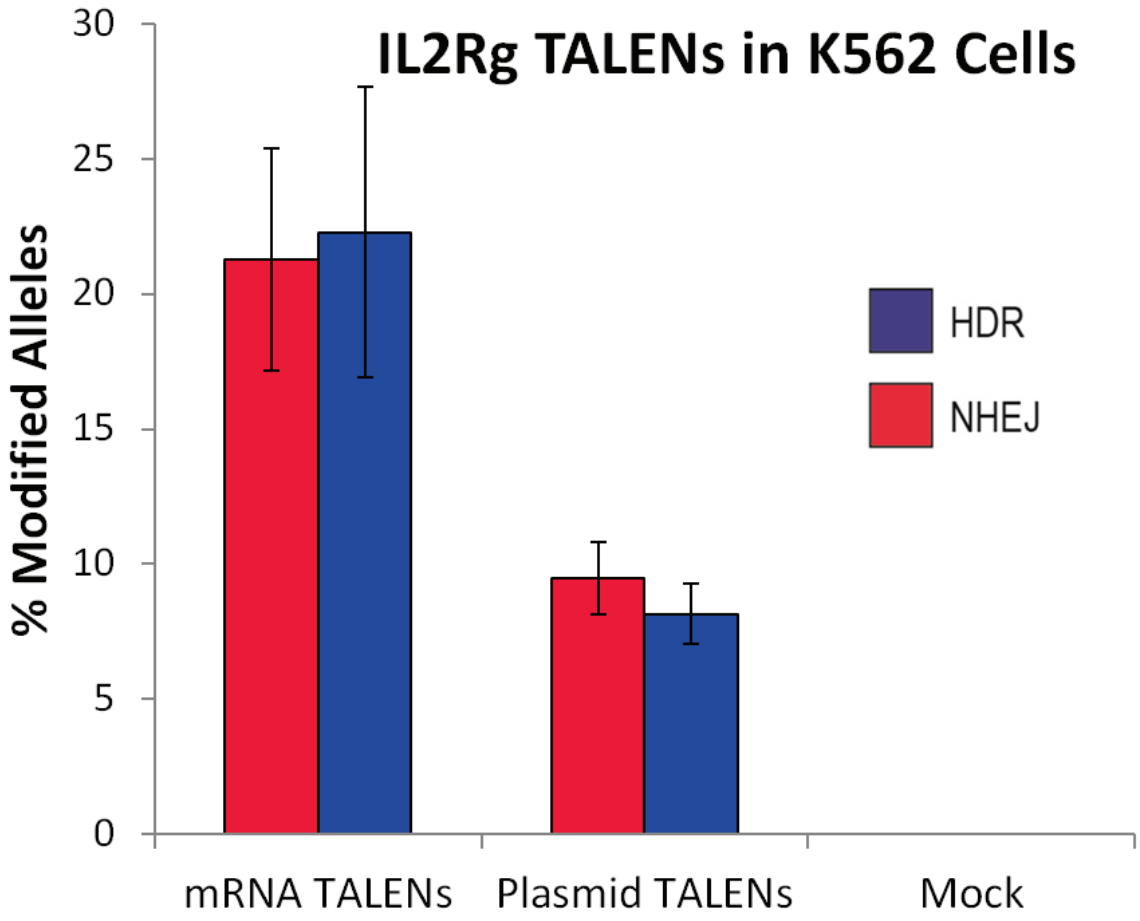

Figure 7-1: mRNA vs Plasmid Delivery of TALENs. n=3, error bars are s.t.d.

While synthesized mRNA remains much more expensive than plasmid DNA, the marked increase in nuclease activity makes mRNA a promising option for more critical experiments, such as those in primary $CD34^+$ cells.

7.3 Enrichment of Gene Modified CD34⁺ Cells through FACS[2]

Rates of HDR in primary cells are typically much lower than in immortalized cell lines such as the K562 model system. Therefore, an area of growing interest is the ability to enrich the subpopulation of HDR-modified primary cells.

We constructed a donor plasmid which would fuse the gene for the mCitrine fluorescent protein to the *HBB* sequence such that expression of mCitrine would only occur from the endogenous *HBB* locus (and not from the donor plasmid itself). $CD34^+$ cells were nucleofected with the TALENs previously described (Figure 6-6b) with or without the donor plasmid. The cells were then put through a two-week erythroid differentiation protocol so that the *HBB* locus would be actively expressed. At that point, a portion of the bulk population was set aside, and the remaining cells were sorted using FACS into $mCitrine^+$ and $mCitrine^-$ sub-populations. Genomic DNA was harvested from the cells and cDNA libraries were created from the RNA. Both the DNA and the RNA were analyzed using the SMRT analysis pipeline previously described [48].

Modified alleles and RNA transcripts were observed at low rates in the bulk cell population (Figure 7-2). After FACS selection, the percentage of HDR-modified alleles increased ~40-fold while the percentage of NHEJ-modified alleles stayed roughly equal; selective enrichment of only HDR-modified alleles was expected because mCitrine is only integrated in the case of HDR. While the percentage of HDR-modified alleles in the $mCitrine^+$ population was ~25%, the percentage of HDR-modified transcripts was much higher—nearly 90%. This was an unexpected result since our original hypothesis was that the percentage

[2] In collaboration with Dr. Adam Hartigan (Harvard), Dr. Amy Wagers (Harvard), Gabriel Washington (Stanford), and Dr. Matthew Porteus (Stanford).

of modified transcripts would roughly mirror the percentage of modified alleles. Possible reasons for this observation include preferential gene modification of the more highly expressed allele and/or bias during the FACS process for selecting cells where one allele (containing mCitrine) is expressed at much higher levels than the other allele (therefore giving a stronger fluorescent signal). These possibilities will require further investigation.

7.4 Analysis of Many HBB Nucleases

In order to further advance gene therapy approaches for correcting sickle cell anemia, we analyzed the HDR:NHEJ ratios of many different classes of nucleases all targeting *HBB*. In addition to determining an optimal nuclease to move forward with for further sickle cell experiments, this study also provides a direct comparison different types of nucleases all at the same genetic locus to control for location-dependent effects.

7.4.1 Methods

7.4.1.1 Preliminary SMRT Investigation

200,000 K562 cells were nucleofected with 1000 ng of total DNA, consisting of 200 ng of total nuclease plasmid and 800 ng of donor plasmid (or pUC control). Nucleofections were performed in simultaneous triplicate (separate nucleofections of cells taken from the same cultured population). After 72 hours, genomic DNA was harvested and the *HBB* locus was analyzed using SMRT sequencing as described in Chapter 6.

7.4.1.2 Follow-up Illumina Investigation

200,000 K562 cells were nucleofected with 1400 ng of total DNA, consisting of 400 ng of total nuclease plasmid and 1000 ng of donor plasmid (or pUC control). Nucleofections were performed in biological triplicate with each replicate performed on different weeks. After 72 hours, genomic DNA was harvested and the same sequence of the *HBB* locus as described in Chapter 6 was amplified; however, the primers contained additional 5' sequences to facilitate binding to the Illumina flow cell. Dual i5/i7 Illumina barcodes were

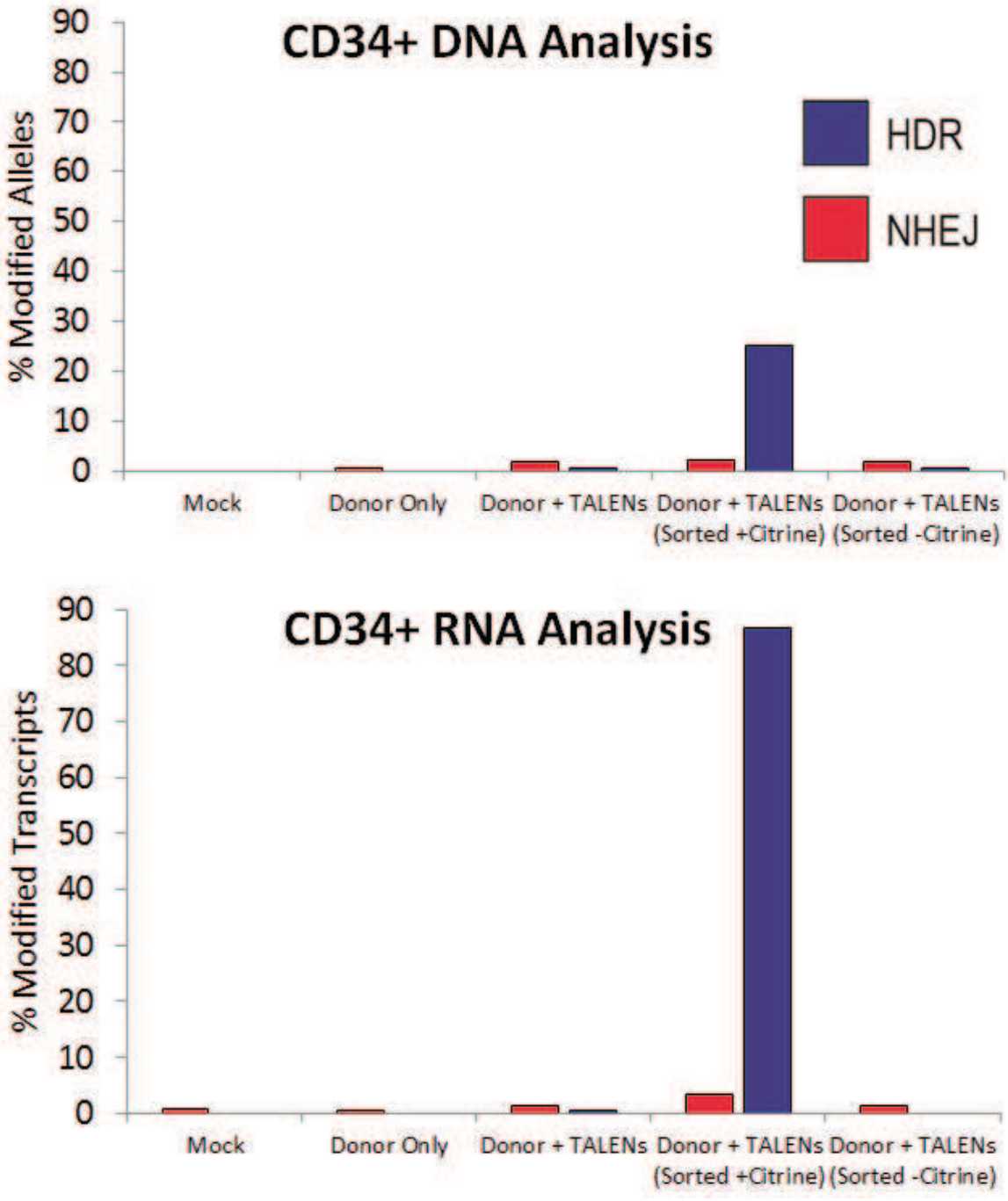

Figure 7-2: Analysis of CD34^{+} DNA and RNA.

added to each amplicon to facilitate multiplexing 128 samples into a single MiSeq run. 250-bp paired end (PE) sequencing was used.

Because Illumina sequencing only provides the sequences at the edges of the amplicons, as opposed to SMRT sequencing which provides the full sequence of the amplicon, several alterations needed to be made to the previously described analysis pipeline (Figure 6-2). Since the total *HBB* amplicon length is ~600 bp, the two 250 bp Illumina sequencing reads do not overlap and are therefore handled separately by the modified pipeline. The following modified initial steps were changed to accommodate the Illumina data:

1) Reads were de-multiplexed according to their i5/i7 barcodes by the internal Illumina analysis software.

2) If the average FASTQ score of either read is below 25, the reads are discarded.

3) The 20 bp at the 5' end of each read must exactly match the expected template sequence.

4) The read distal from the nuclease target site was pairwise-aligned to the expected *HBB* sequence and the 5' 100 bp were checked to ensure ≥80% match to the template sequence.

5) The 3' 50 bp of the read proximal to the nuclease target site were also checked by pairwise alignment to ensure ≥80% match to the template sequence.

6) If all those requirements were met, the remaining expected template sequence 3' of the end of the proximal read was computationally appended to the sequencing read to make the data more closely resemble a SMRT read (which consists of the full amplicon) and then the reads were fed into the previously developed pipeline at the allele analysis step (Figure 6-2p) for subsequent analysis.

7.4.1.3 Donors and Nucleases

Several different repair donors were tested (Figure 7-3). In the preliminary SMRT investigation, the 'EcoRI Donor' (which introduces an EcoRI restriction site that can be assayed

Figure 7-3: *HBB* Donors and Nucleases. (**upper panel**) The wild-type *HBB* sequence beginning with the start codon in green, followed by the corresponding sequence of different donor plasmids with mismatches highlighted in red—all three donor plasmids had ~400 bp flanking arms of homology to the endogenous *HBB* sequence on either side. (**middle panel**) CRISPR/RGEN binding sites within *HBB* are given. PAM sequences are underlined and a mismatched 5' base in the guide strand is signified by a lowercase 'g'. (**lower panel**) ZFN and TALEN binding sites within *HBB* are given. The L4 TALEN targets a mismatched base relative to the wild-type *HBB* sequence, signified by a lowercase 't'—it is designed to target the sickle mutation.

using RFLP) was used. While this donor blocked the binding of all of the nucleases (or at least one element of each nuclease pair) tested in the preliminary study, its SNPs did not block all of the nucleases planned for the followup study; therefore, a donor was made with synonymous SNPs spread throughout exon 1 ("M4" Donor). In parallel to the development of "M4", our collaborators created the "AfeI Donor" which also contains a larger number of synonymous SNPs as well as an AfeI restriction enzyme site.

Several different nucleases were tested (Figure 7-3). The R7, R1, R3, and R2 CRISPR / RGEN guide sequences have been previously described [21] while the R2d3 guide strand is a truncated version of R2 following the findings of Fu et al. [36]. The L4/R4 TALEN pair is the same as was used in Chapter 6. The 4F-ZFN pair is the same as was used in Chapter 3. The S136/S120 TALEN pair was previously described by Lin et al. [69].

7.4.2 Results

7.4.2.1 Preliminary SMRT Investigation

A wide range of performance was observed across the different types of nucleases (Figure 7-4). The L4/R4 TALENs outperformed the ZFNs, the S136/S120 TALENs, and another version of the L4/R4 TALENs utilizing a bi-partite nuclear localization signal (NLS) generated by a collaborator in terms of NHEJ, but none of those nucleases had appreciable HDR activity above background levels. The "RH" CRISPR paired nickases (employing the H840A mutation) performed poorly, consistent with previous results [95], in contrast to the "RN" CRISPR paired nickases (employing the D10A mutation) which had extremely high NHEJ and activity and moderate HDR activity as well. The single CRISPR nucleases (R2 and R3) had the highest rates of HDR as well as the most favorable HDR:NHEJ ratios. The truncated guide strand (R2d3) showed slightly reduced NHEJ activity compared to the full length guide strand (R2) but much lower HDR activity.

7.4.2.2 Follow-up Illumina Investigation

In order to examine some of the trends observed in the preliminary SMRT investigation in more detail, a more comprehensive set of experiments was conducted. In order to reduce sequencing costs for the 128 total samples, we modified the approach to be compatible with the Illumina sequencing approach.

Because analyzing the Illumina sequencing data required modifications to the previously developed bioinformatics pipeline, we performed side-by-side comparisons of the two different analysis methods on the same genomic DNA (Figure 7-5). Overall, the same trends were observed between the two different analysis methods; both reported that the RN1/RN2 paired CRISPR nickases induced high levels of NHEJ and lower levels of HDR and that the R3 CRISPR nuclease had higher levels of HDR than NHEJ. However, the exact reported values different slightly between the two methods and the Illumina method was

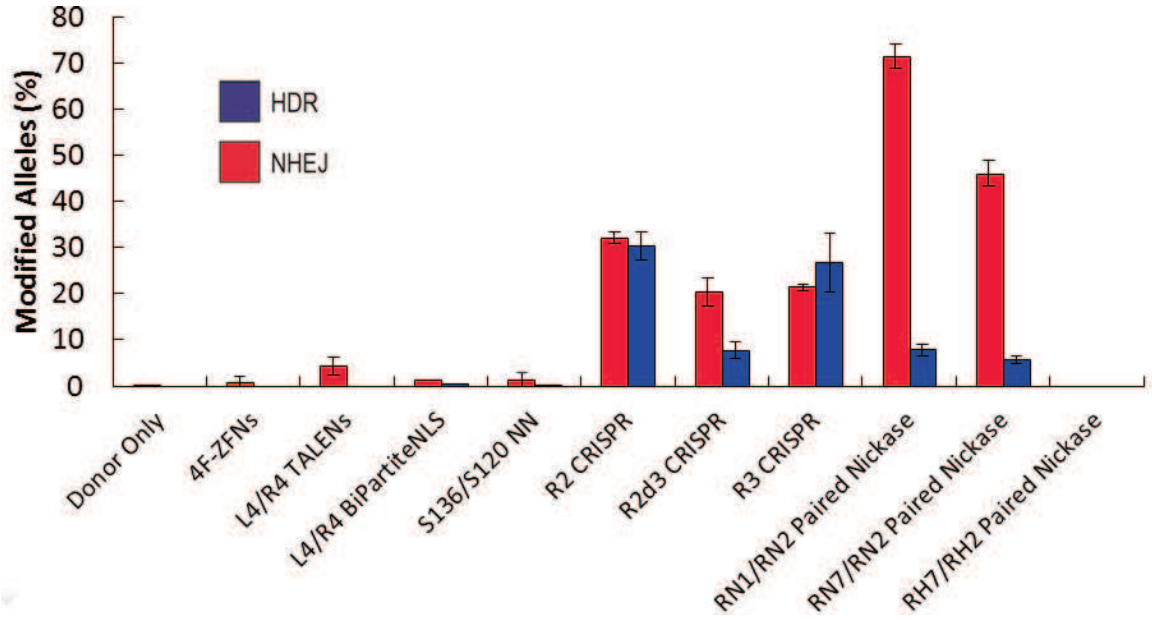

Figure 7-4: Preliminary SMRT Investigation of *HBB* nucleases. Several different types of nucleases were nucleofected into K562 cells, including ZFNs, TALENs, CRISPR nucleases (R2 and R3), CRISPR nucleases with truncated guide strands (R2d3), D10A paired CRISPR nickases (RN1/RN2 and RN7/RN2), and H840A paired CRISPR nickases (RH7/RH2). n=3, error bars are s.t.d.

characterized by much higher variability in the measurements (indicated by higher standard deviations of the replicates). Overall, while the current Illumina analysis pipeline has clearly not reached the same level of refinement and precision as the SMRT pipeline, it appears to provide reasonably accurate results and warrants some, but not complete, confidence in the results obtained for the following experiments.

Two pairs of TALENs and a pair of ZFNs were analyzed. As in the preliminary SMRT analysis, only moderate levels of activity were observed (Figure 7-6). In contrast to the robust NHEJ activity observed for the S136/S120 TALENs in 293T cells (~60% modified alleles [69]), extremely low activity was seen under these conditions in K562 cells. The levels of NHEJ for the L4/R4 TALENs were relatively constant in the presence or absence of a donor. However, the HDR activity using the AfeI donor was much higher than when using the "M4" donor.

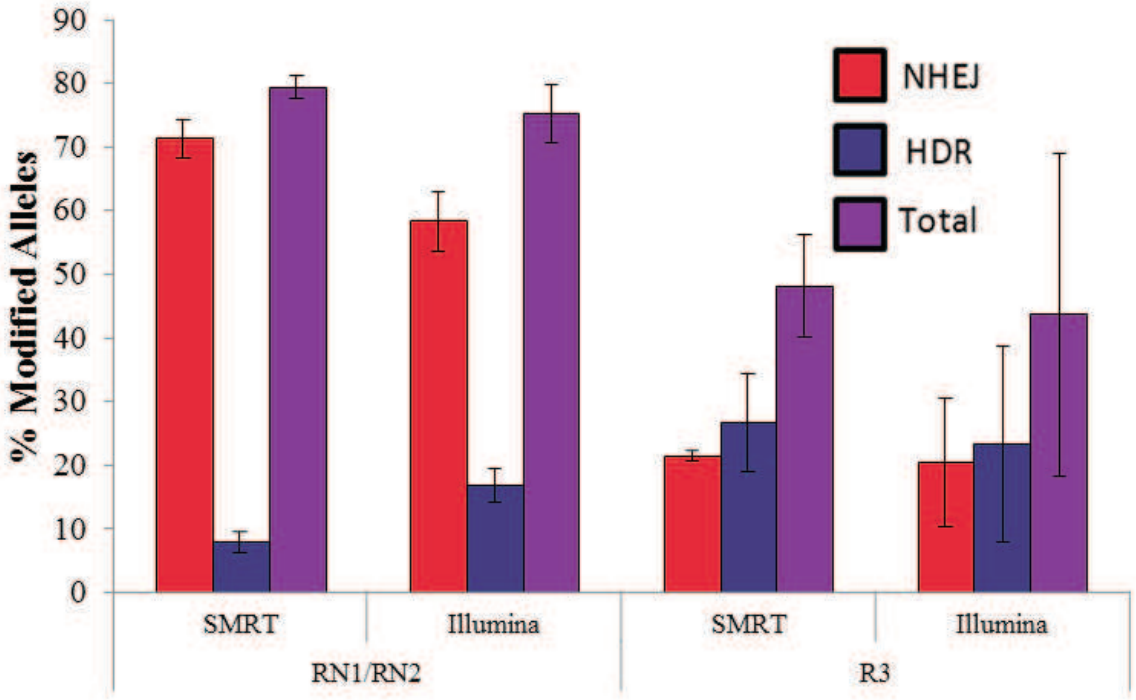

Figure 7-5: Comparing SMRT and Illumina Analysis Methods. The same 6 genomic DNA samples (3 replicates of each experiment) were analyzed by SMRT and by Illumina for HDR and NHEJ repair. n=3, error bars are s.t.d.

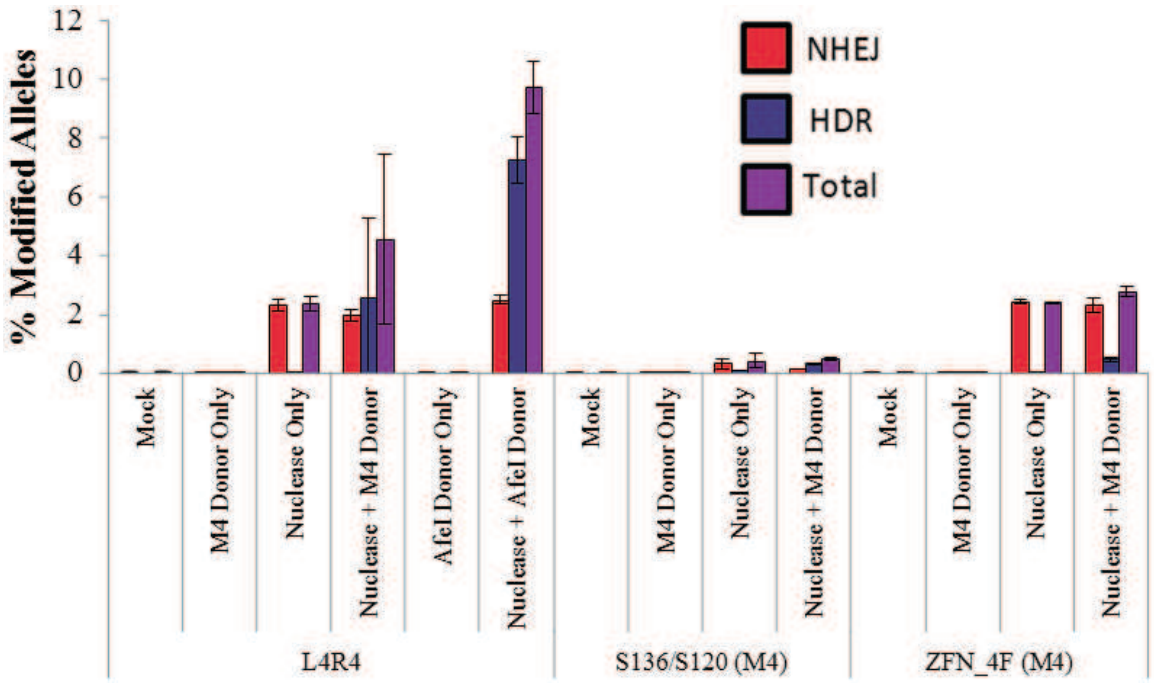

Figure 7-6: TALEN and ZFN Analysis. Two pairs of TALENs and a pair of ZFNs were analyzed, in the presence or absence of either the "M4" or AfeI donors. n=3, error bars are s.t.d.

Working with collaborators[3], we tested the performance of a "mini-circle" version of the AfeI donor. Mini-circles are plasmids that are processed to remove elements that are essential for production in *E. coli* but are not relevant for the final purpose (such as antibiotic resistance gene or origins of replication). For genome engineering applications, this results in a plasmid that consists of just the homology arms and the internal region with the desired SNPs to be integrated. We hypothesized that the higher numbers of donor molecules (since total donor mass is held constant and mini-circles are smaller than plasmids) combined with the longer persistence of the mini-circles in cells [17] might lead to higher levels of HDR. While the total level of gene modification was similar between the experiments with the three different donors, the mini-circle experiments had markedly lower levels of HDR (Figure 7-7). Surprisingly, the mini-circle donor performed worse than the parental

[3]In collaboration with Carol Dickerson and Dr. Steffen Meiler, Georgia Regents University, Augusta GA.

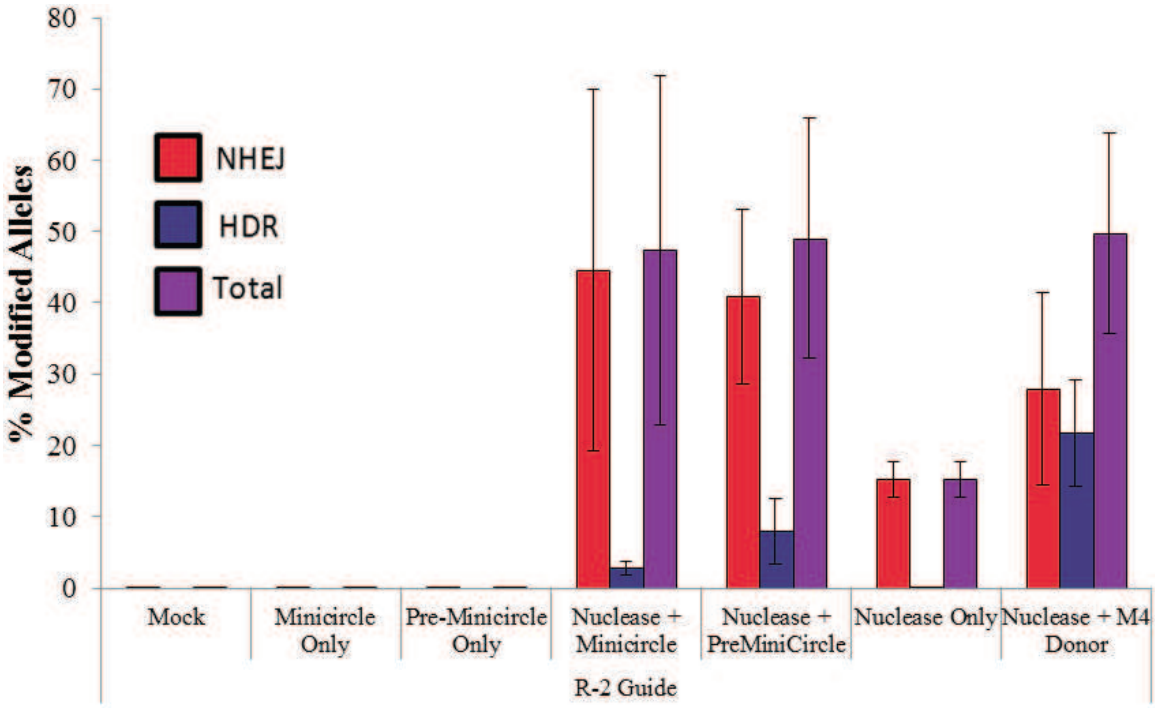

Figure 7-7: Mini-circle Donor Analysis. n=3 for most samples, n=2 for Nuclease+Mini-Circle and Nuclease+PreMiniCircle. error bars are s.t.d.

"pre"-mini-circle plasmid.

Truncated guide RNAs have been shown in previous studies to have reduced off-target effects, but sometimes at a cost of reduced on-target activity. We tested a guide RNA with a 3 bp truncation compared to the full length guide RNA (Figure 7-8). In contrast to the substantial decrease in HDR efficiency we observed in the preliminary SMRT analysis (Figure 7-4), the full-length and truncated guide RNAs performed at statistically indistinguishable levels in these experiments (although the mean activities were slightly lower with the truncated guide RNA for both NHEJ and HDR).

Two different configurations of paired CRISPR nickases were analyzed. In order to better understand the contributions of each component of the system to the levels of NHEJ and HDR, each individual guide was analyzed separately as a nickase and as a nuclease in addition to analyzing the full paired nickase systems (Figure 7-9). Consistent with previous reports [118], individual nickases can, but do not always, trigger NHEJ and HDR (i.e. the

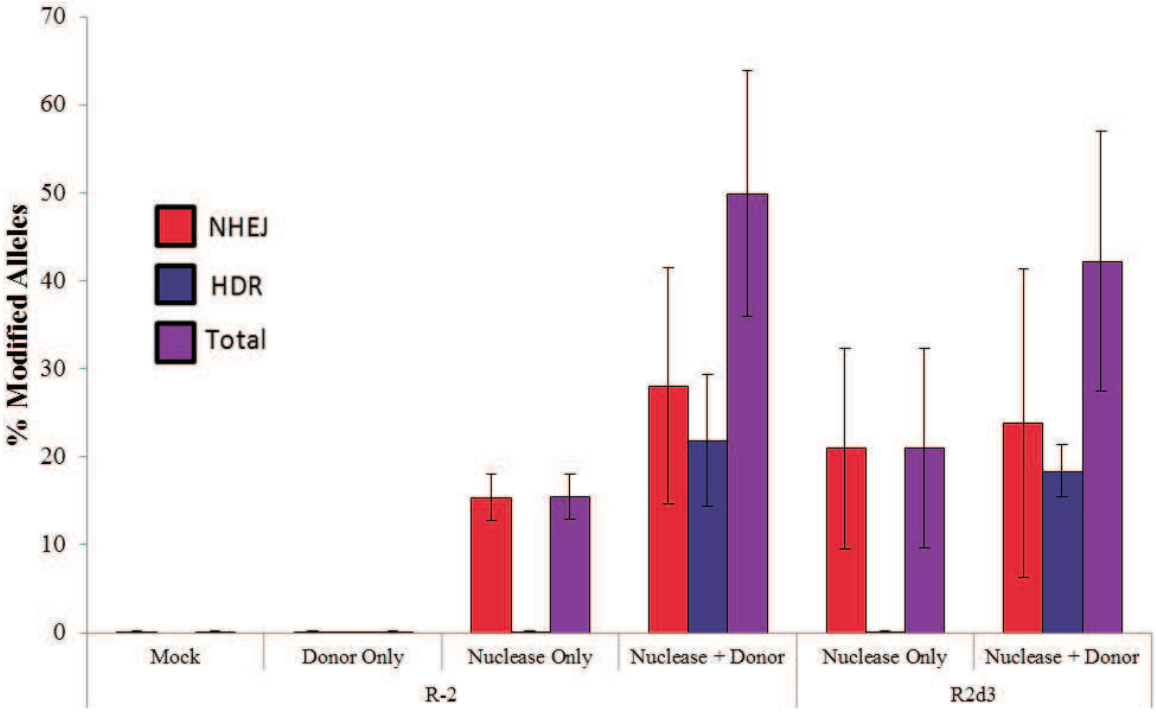

Figure 7-8: Truncated Guide RNA Analysis. n=3, error bars are s.t.d.

R-1 guide) even without their paired partner. This phenomenon seems to be highly guide-strand dependent with no known features being yet able to predict this behavior *a priori*. Apart from the aberrantly low levels of NHEJ activity observed for the "R-2 Nuclease Only" sample, the paired nickases had similar levels of total gene modification activity as the single nucleases. However, in contrast to the results from the preliminary SMRT analysis (Figure 7-4), these experiments showed the paired nickases having equal or greater rates of HDR as the single CRISPR nucleases.

7.4.3 Discussion

High-throughput sequencing analysis allowed a detailed look at the rates of NHEJ and HDR for a wide variety of nucleases at the *HBB* locus. While the newer Illumina analysis pipeline provided similar results as the well-validated SMRT pipeline, there was increased variability and slight discrepancies between the measurements. Given the marked cost advantage of Illumina, further refinement of the analysis pipeline should provide a valuable

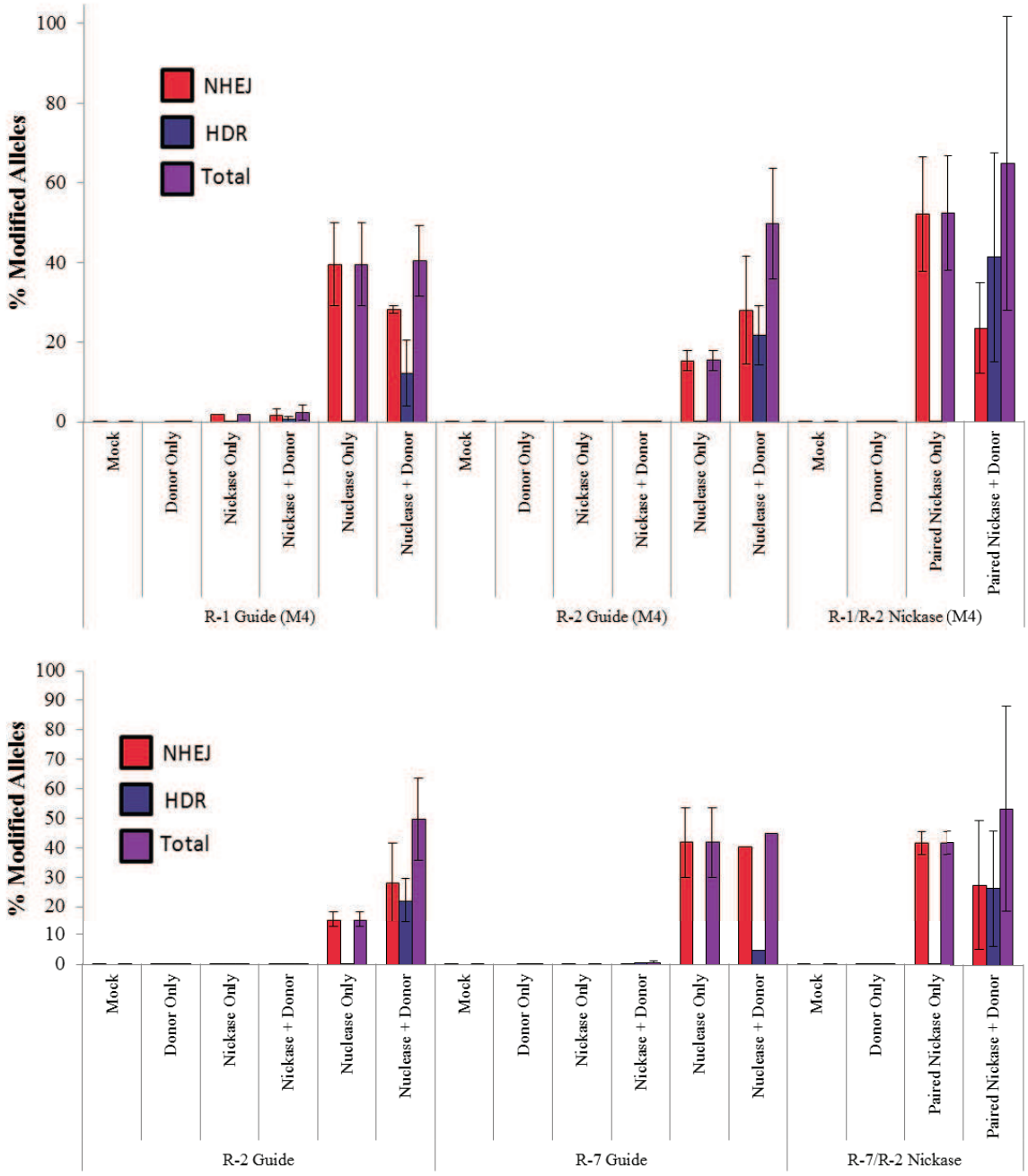

Figure 7-9: Paired CRISPR Nickase Analysis. Two different paired CRISPR nickase configurations were analyzed; the R-2 guide paired with the R-1 guide (**upper**) and the R-2 guide paired with the R-7 guide (**lower**). n=3 for most samples, error bars are s.t.d.

tool for use in cases where it is acceptable to place one of the PCR primers inside the homology arms near the nuclease cut site; in cases which require both primers to be outside the homology arms, SMRT sequencing remains the only available method.

Although the preliminary SMRT analysis pointed to several major performance differences—particularly related to HDR rates and NHEJ:HDR ratios—between the different types of nucleases, these trends were largely not observed in the follow-up Illumina study. While it is clear that CRISPR systems outperform TALENs and ZFNs, the Illumina analysis showed no clear differences between CRISPR nucleases, truncated guide strands, and paired CRISPR nickases. The exact reason for these differences is unclear, but it may be related to the fact that all SMRT experiments were performed from a single passage of the K562 cells whereas the Illumina experiments were performed over three different passages.

CHAPTER VIII

FUTURE CONSIDERATIONS

The tools created in this thesis represented 'first-in-class' developments of new methods of off-target prediction and analysis of DNA repair outcomes. However, there are several areas in which these tools could be improved with future work.

8.1 Accurately Predicting CRISPR Off-Target Sites

Although several online tools are now available to search a genome for CRISPR sites with homology to the intended target sequence [53, 22], none of the tools provides a quantitative basis for their rankings nor their overall predictive power. Indeed, recent studies have shown that current online tools routinely do not locate most known CRISPR off-target sites [119, 22]. The poor performance of current tools is likely due to their simplistic models, but the current understanding in the field as a whole on the underlying biology of what dictates the observed performance differences between CRISPR guide strands is still fairly limited, although single molecule studies have provided some useful insights [111]. However, the increasing numbers of researchers involved with CRISPR and the tendency of CRISPRs to cause readily observable off-target effects has led to an accelerating pace of the discovery of bona fide CRISPR off-target sites; a recent study validated over 100 off-target sites [119], although it was still unable to divine any broad principles governing CRISPR off-target activity. The rapid rise in the number of CRISPR off-target sites provides hope that in the near future there may be a large enough dataset on which to attempt machine learning, even in the absence of preconceived models of the most important features that distinguish bona fide off-target sites from other sites with homology to the intended target.

8.2 Cell-Type-Specific Off-Target Prediction

The intracellular genomic context of a particular off-target location can have a large influence on what level (if any) of off-target activity occurs. One study found that at two identical off-target DNA sequences, one site had off-target activity at nearly the same frequency as the on-target site while the other site had no detectable off-target mutagenesis [21]. Chromatin status, methylation, histone modifications, or other bound proteins are all factors that can influence the level of nuclease activity at an off-target site, but these properties are often cell-type dependent. As most nucleases would only need to be applied clinically in one certain cell type, accurate predictions could greatly enhance the ability to rationally design nucleases to be less prone to off-target activity in that particular cell as well as to narrow the focus of off-target assessments on the most appropriate regions.

Databases such as the ENCODE project [32] now provide large amounts of information about the genomic context in particular cell types. Unfortunately, this is only one side of the coin—sufficiently large datasets of comparisons of nuclease off-target activity across different cell types must still be performed in order to better understand how different genomic contexts affect the propensity of a nuclease to cleave that location. However, intermediate steps in this process could include annotations in the off-target prediction summaries about known attributes such as whether a site was within a DNAse I hypersensitivity region. Such information might still provide some utility to end-users even without formal quantitative algorithms to assess the impact of such factors on off-target activity levels.

8.3 Quantitative Prediction of Off-Target Frequencies

Correlating the rank order of off-target predictions with the observed off-target mutagenesis has long been a goal in this field. However, even experimental-based off-target characterization methods have not had much success in this area (Figure 3-4). In fact, it was only very recently that any published experimental-based off-target prediction method was able to achieve any correlation between its predictive ranking and actual observed off-target

activity [119]. However, there are several challenges in implementing this as a generalized predictive bioinformatic algorithm. As mentioned above, cell-type-specific factors can play a large confounding role in the exact frequency of mutagenesis at a given off-target site. Furthermore, compiling data from multiple studies in order to assemble an appropriate training set would be complicated by different doses, incubation times, and nuclease delivery methods used in different studies—all of which are factors that affect the on-target:off-target activity ratios. Given the different downstream applications of nucleases, it seems unlikely that the field as a whole will adopt more standardized off-target testing conditions since those may not be as relevant for their particular area of interest. Therefore, unless one institution or consortium dedicates itself to conducting a large number of off-target investigations of different nucleases all under similar experimental conditions, it seems unlikely that computational quantitative prediction of off-target frequencies will occur in the near future.

8.4 Optimizing Illumina Sequencing of NHEJ vs HDR

Although the short readlengths of Illumina sequencing (compared to SMRT) impose certain restrictions, the cost-savings available for large scale studies of NHEJ vs HDR makes it an attractive option to pursue. Current reliable Illumina readlengths are ~250 bp (although newer 300 bp chemistries are being further developed). While this is not sufficient to observe integration of large pieces of exogenous DNA (Figure 6-8) nor in cases where both PCR primers must be placed outside of the homology arms in order to completely eliminate any non-specific donor amplification—we have not found this to be an issue in K562 cells, but we observed some non-specific amplification of the donor plasmid in recent experiments in hESCs—but can be sufficient in some applications. Specifically, if one PCR primer is inside the donor homology arms (while the other primer is outside the homology arms), then Illumina can provide sequence information about the ~100 bp flanking the nuclease target site. In our preliminary investigation, we found that Illumina sequencing

could provide reasonably accurate readouts (Figure 7-5) at less than 20% of the cost of SMRT sequencing for large scale analyses. However, the Illumina readout is currently not as accurate or precise as the SMRT measurements. Because Illumina sequencing reads are formatted differently than SMRT (uni-directional paired reads from the edges of the amplicon rather than providing the full amplicon sequence) and because the Illumina chemistry is prone to different types of sequencing errors and artifacts than SMRT chemistry, several additional changes to the bioinformatics analysis pipeline in order to accomodate these differences will be needed in order to allow the Illumina approach to provide comparable results as SMRT. Further optimizations of the experimental protocol are also possible including more advanced size selection of the PCR amplicons and variation in the number of PCR cycles in each of the steps.

APPENDIX A

EFFECT OF SERUM STARVATION ON PAIRED CRISPR/CAS9 NICKASE ACTIVITY

A.1 Introduction

Although paired Cas9 nickases create indels at very high rates, most measurements have been made in rapidly dividing cells. The nicks in these cases are typically separated by ~35 bp, a much larger distance than the typical 4 bp that separates the breaks made by FokI. One hypothesis was that the indels observed after paired nickase treatment might be generated by a different method than cannonical NHEJ repair; the nicks might be causing interference during DNA replication which would lead to deletion of the intervening sequence. In order to test this hypothesis, we reduced the rate of cell division using serum starvation and measured the activity of CRISPR nucleases and paired nickases.

A.2 Methods

On Day 0, K562 cells were passaged into media containing 10% FBS (standard treatment) or no FBS (serum starvation). On Day 1, the K562 cells were nucleofected with Cas9 plasmids and plated in fresh media (+/- serum). On Day 4, cells were harvested and genomic DNA was analyzed by the T7E1 assay to test for indel activity.

A.3 Discussion

Serum starvation had no effect on the activity of Cas9 nucleases but did reduce the activity of Cas9 nickases (Figure A-1). However, the activity of the Cas9 nickases under serum starved conditions was still substantially greater than the activity of the Cas9 nucleases.

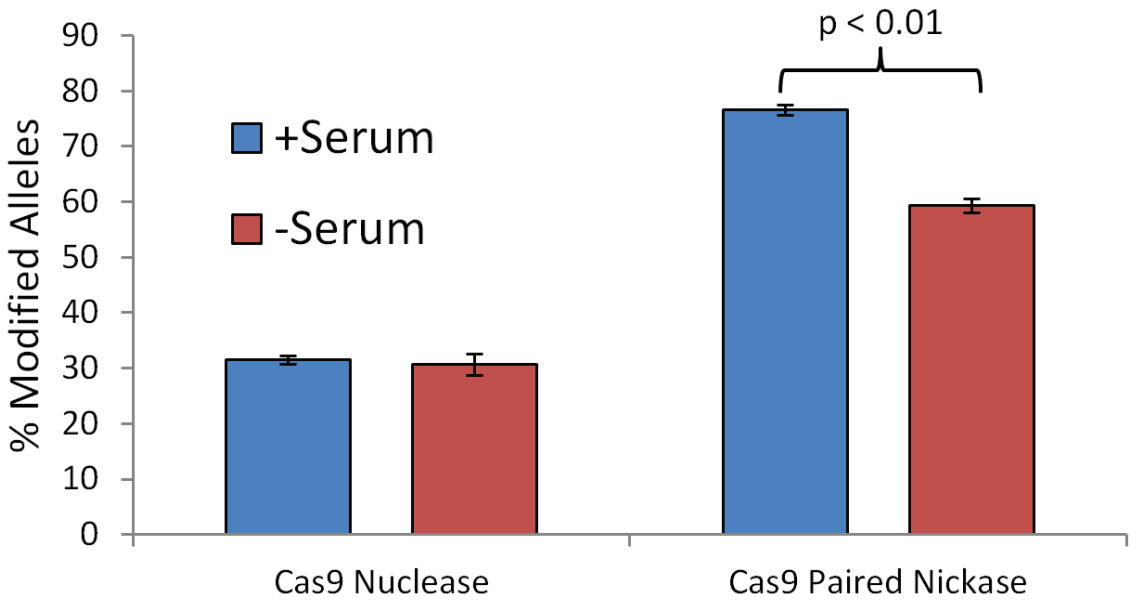

Figure A-1: Effect of Serum Starvation on Cas9 Activity. Error bars are s.t.d., n=3. P value derived from Students' paired t-test (two-tailed).

Clearly, paired Cas9 nickases are still active in cells undergoing limited cell division, however the reason for the observed reduction in activity requires further investigation.

REFERENCES

[1] Abarrategui-Pontes, C., Alison, C., Reynald, T., Fine, E. J., Virginie, T., Le Ray, L., Cradick, T. J., Bao, G., Laurent, T., Guillaume, P., and others, "Codon swapping of zinc finger nucleases confers expression in primary cells and in vivo from a single lentiviral vector," *Current Gene Therapy*, vol. 14, no. 5, pp. 365–376, 2014.

[2] Aiuti, A., Biasco, L., Scaramuzza, S., Ferrua, F., Cicalese, M. P., Baricordi, C., Dionisio, F., Calabria, A., Giannelli, S., Castiello, M. C., Bosticardo, M., Evangelio, C., Assanelli, A., Casiraghi, M., Di Nunzio, S., Callegaro, L., Benati, C., Rizzardi, P., Pellin, D., Di Serio, C., Schmidt, M., Von Kalle, C., Gardner, J., Mehta, N., Neduva, V., Dow, D. J., Galy, A., Miniero, R., Finocchi, A., Metin, A., Banerjee, P. P., Orange, J. S., Galimberti, S., Valsecchi, M. G., Biffi, A., Montini, E., Villa, A., Ciceri, F., Roncarolo, M. G., and Naldini, L., "Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome," *Science*, vol. 341, no. 6148, 2013. 10.1126/science.1233151.

[3] Alwin, S., Gere, M. B., Guhl, E., Effertz, K., Barbas, C. F., Segal, D. J., Weitzman, M. D., and Cathomen, T., "Custom zinc-finger nucleases for use in human cells," *Molecular Therapy*, vol. 12, no. 4, pp. 610–617, 2005.

[4] Appelt, J. U., Giordano, F. A., Ecker, M., Roeder, I., Grund, N., Hotz-Wagenblatt, A., Opelz, G., Zeller, W. J., Allgayer, H., Fruehauf, S., and Laufs, S., "QuickMap: a public tool for large-scale gene therapy vector insertion site mapping and analysis," *Gene Therapy*, vol. 16, no. 7, pp. 885–93, 2009.

[5] Arens, A., Appelt, J.-U., Bartholomae, C. C., Gabriel, R., Paruzynski, A., Gustafson, D., Cartier, N., Aubourg, P., Deichmann, A., Glimm, H., von Kalle, C., and Schmidt, M., "Bioinformatic clonality analysis of next-generation sequencing-derived viral vector integration sites," *Human Gene Therapy Methods*, vol. 23, no. 2, pp. 111–118, 2012.

[6] Ben-Hur, A. and Weston, J., "A user's guide to support vector machines," in *Data mining techniques for the life sciences*, pp. 223–239, Humana Press, 2010.

[7] Berry, C. C., Ocwieja, K. E., Malani, N., and Bushman, F. D., "Comparing DNA integration site clusters with scan statistics," *Bioinformatics*, vol. 30, no. 11, pp. 1493–1500, 2014. 10.1093/bioinformatics/btu035.

[8] Beumer, K. J., Trautman, J. K., Mukherjee, K., and Carroll, D., "Donor DNA utilization during gene targeting with zinc-finger nucleases," *G3: Genes— Genomes— Genetics*, vol. 3, no. 4, pp. 657–664, 2013.

[9] Bhakta, M. S., Henry, I. M., Ousterout, D. G., Das, K. T., Lockwood, S. H., Meckler, J. F., Wallen, M. C., Zykovich, A., Yu, Y., Leo, H., Xu, L., Gersbach, C. A., and Segal, D. J., "Highly active zinc finger nucleases by extended modular assembly," *Genome Research*, 2012.

[10] Bogdanove, A. J. and Voytas, D. F., "TAL effectors: Customizable proteins for DNA targeting," *Science*, vol. 333, no. 6051, pp. 1843–1846, 2011.

[11] Briggs, A. W., Rios, X., Chari, R., Yang, L., Zhang, F., Mali, P., and Church, G. M., "Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers," *Nucleic Acids Research*, p. gks624, 2012.

[12] Brunet, E., Simsek, D., Tomishima, M., DeKelver, R., Choi, V. M., Gregory, P., Urnov, F., Weinstock, D. M., and Jasin, M., "Chromosomal translocations induced at specified loci in human stem cells," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10620–10625, 2009.

[13] Cavalli, G. and Misteli, T., "Functional implications of genome topology," *Nature Structural & Molecular Biology*, vol. 20, no. 3, pp. 290–299, 2013.

[14] Cermak, T., Doyle, E. L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J. A., Somia, N. V., Bogdanove, A. J., and Voytas, D. F., "Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting," *Nucleic Acids Research*, vol. 39, no. 12, 2011.

[15] Certo, M. T., Ryu, B. Y., Annis, J. E., Garibov, M., Jarjour, J., Rawlings, D. J., and Scharenberg, A. M., "Tracking genome engineering outcome at individual DNA breakpoints," *Nature Methods*, vol. 8, no. 8, pp. 671–676, 2011.

[16] Chen, F., Pruett-Miller, S. M., Huang, Y., Gjoka, M., Duda, K., Taunton, J., Collingwood, T. N., Frodin, M., and Davis, G. D., "High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases," *Nature Methods*, vol. 8, no. 9, pp. 753–755, 2011.

[17] Chen, Z.-Y., He, C.-Y., Ehrhardt, A., and Kay, M. A., "Minicircle DNA vectors devoid of bacterial DNA result in persistent and high-level transgene expression in vivo," *Molecular Therapy*, vol. 8, no. 3, pp. 495–500, 2003.

[18] Christian, M. L., Demorest, Z. L., Starker, C. G., Osborn, M. J., Nyquist, M. D., Zhang, Y., Carlson, D. F., Bradley, P., Bogdanove, A. J., and Voytas, D. F., "Targeting G with TAL effectors: A comparison of activities of TALENs constructed with NN and NK repeat variable di-residues," *PLoS ONE*, vol. 7, no. 9, 2012.

[19] Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., and others, "Multiplex genome engineering using CRISPR/Cas systems," *Science*, vol. 339, no. 6121, pp. 819–823, 2013.

[20] Cradick, T. J., Ambrosini, G., Iseli, C., Bucher, P., and McCaffrey, A. P., "ZFN-site searches genomes for zinc finger nuclease target sites and off-target sites," *BMC Bioinformatics*, vol. 12, pp. 152–152, 2011.

[21] Cradick, T. J., Fine, E. J., Antico, C. J., and Bao, G., "CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity," *Nucleic Acids Research*, vol. 41, no. 20, pp. 9584–9592, 2013.

[22] Cradick, T. J., Qiu, P., Lee, C. M., Fine, E. J., and Bao, G., "COSMID: A web-based tool for identifying and validating CRISPR/Cas off-target sites," *Molecular Therapy-Nucleic Acids*, vol. 3, no. 12, p. e214, 2014.

[23] Cristea, S., Freyvert, Y., Santiago, Y., Holmes, M. C., Urnov, F. D., Gregory, P. D., and Cost, G. J., "In vivo cleavage of transgene donors promotes nuclease-mediated targeted integration," *Biotechnology and bioengineering*, vol. 110, no. 3, pp. 871–880, 2013.

[24] Daboussi, F., Zaslavskiy, M., Poirot, L., Loperfido, M., Gouble, A., Guyot, V., Leduc, S., Galetto, R., Grizot, S., Oficjalska, D., and others, "Chromosomal context and epigenetic mechanisms control the efficacy of genome editing by rare-cutting designer endonucleases," *Nucleic Acids Research*, vol. 40, no. 13, pp. 6367–6379, 2012.

[25] Davis, J. and Goadrich, M., "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, ACM, 2006.

[26] Ding, Q., Lee, Y.-K., Schaefer, E. K., Peters, D., Veres, A., Kim, K., Kuperwasser, N., Motola, D., Meissner, T., Hendriks, W., Trevisan, M., Gupta, R., Moisan, A., Banks, E., Friesen, M., Schinzel, R., Xia, F., Tang, A., Xia, Y., Figueroa, E., Wann, A., Ahfeldt, T., Daheron, L., Zhang, F., Rubin, L., Peng, L., Chung, R., Musunuru, K., and Cowan, C., "A TALEN genome-editing system for generating human stem cell-based disease models," *Cell Stem Cell*, 2013.

[27] Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J., and Root, D. E., "Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation," *Nature Biotechnology*, vol. advance online publication, 2014.

[28] Doyle, E. L., Booher, N. J., Standage, D. S., Voytas, D. F., Brendel, V. P., VanDyk, J. K., and Bogdanove, A. J., "TAL effector-nucleotide targeter (TALE-NT) 2.0: Tools for TAL effector design and target prediction," *Nucleic Acids Research*, 2012.

[29] Doyon, Y., Choi, V. M., Xia, D. F., Vo, T. D., Gregory, P. D., and Holmes, M. C., "Transient cold shock enhances zinc-finger nuclease–mediated gene disruption," *Nature Methods*, vol. 7, no. 6, pp. 459–460, 2010.

[30] Doyon, Y., Vo, T. D., Mendel, M. C., Greenberg, S. G., Wang, J., Xia, D. F., Miller, J. C., Urnov, F. D., Gregory, P. D., and Holmes, M. C., "Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures," *Nature Methods*, vol. 8, no. 1, pp. 74–79, 2011.

[31] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., and others, "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.

[32] ENCODE-Project-Consortium and others, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[33] Fan, L., Kadura, I., Krebs, L. E., Hatfield, C. C., Shaw, M. M., and Frye, C. C., "Improving the efficiency of CHO cell line generation using glutamine synthetase gene knockout cells," *Biotechnology and Bioengineering*, vol. 109, no. 4, pp. 1007–1015, 2012.

[34] Fine, E. J., Cradick, T. J., Zhao, C. L., Lin, Y., and Bao, G., "An online bioinformatics tool predicts zinc finger and TALE nuclease off-target cleavage," *Nucleic Acids Research*, 2013.

[35] Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., and Sander, J. D., "High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells," *Nature Biotechnology*, vol. 31, no. 9, pp. 822–826, 2013.

[36] Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M., and Joung, J. K., "Improving CRISPR-Cas nuclease specificity using truncated guide RNAs," *Nature Biotechnology*, vol. 32, no. 3, pp. 279–284, 2014.

[37] Gabriel, R., Lombardo, A., Arens, A., Miller, J. C., Genovese, P., Kaeppel, C., Nowrouzi, A., Bartholomae, C. C., Wang, J., Friedman, G., Holmes, M. C., Gregory, P. D., Glimm, H., Schmidt, M., Naldini, L., and von Kalle, C., "An unbiased genome-wide analysis of zinc-finger nuclease specificity," *Nature Biotechnology*, vol. 29, no. 9, pp. 816–823, 2011.

[38] Gaj, T., Gersbach, C. A., and Barbas III, C. F., "ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering," *Trends in Biotechnology*, vol. 31, no. 7, pp. 397–405, 2013.

[39] Gaj, T., Guo, J., Kato, Y., Sirk, S. J., and Barbas III, C. F., "Targeted gene knockout by direct delivery of zinc-finger nuclease proteins," *Nature Methods*, vol. 9, no. 8, pp. 805–807, 2012.

[40] Gallagher, P. G., "Disorders of red cell volume regulation," *Current Opinion in Hematology*, vol. 20, no. 3, pp. 201–207, 2013.

[41] Genovese, P., Schiroli, G., Escobar, G., Di Tomaso, T., Firrito, C., Calabria, A., Moi, D., Mazzieri, R., Bonini, C., Holmes, M. C., Gregory, P. D., van der Burg, M.,

Gentner, B., Montini, E., Lombardo, A., and Naldini, L., "Targeted genome editing in human repopulating haematopoietic stem cells," *Nature*, vol. 510, no. 7504, pp. 235–240, 2014.

[42] Grau, J., Boch, J., and Posch, S., "TALENoffer: genome-wide TALEN off-target prediction," *Bioinformatics*, 2013.

[43] Guilinger, J. P., Pattanayak, V., Reyon, D., Tsai, S. Q., Sander, J. D., Joung, J. K., and Liu, D. R., "Broad specificity profiling of TALENs results in engineered nucleases with improved dna-cleavage specificity," *Nature Methods*, vol. 11, no. 4, pp. 429–435, 2014.

[44] Gupta, A., Meng, X., Zhu, L. J., Lawson, N. D., and Wolfe, S. A., "Zinc finger protein-dependent and -independent contributions to the in vivo off-target activity of zinc finger nucleases," *Nucleic Acids Research*, vol. 39, no. 1, pp. 381–392, 2011.

[45] Guschin, D. Y., Waite, A. J., Katibah, G. E., Miller, J. C., Holmes, M. C., and Rebar, E. J., "A rapid and general assay for monitoring endogenous gene modification," in *Engineered zinc finger proteins*, pp. 247–256, Springer, 2010.

[46] Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M. P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C. S., Pawliuk, R., Morillon, E., Sorensen, R., Forster, A., Fraser, P., Cohen, J. I., De Saint Basile, G., Alexander, I., Wintergerst, U., Frebourg, T., Aurias, A., Stoppa-Lyonnet, D., Romana, S., Radford-Weiss, I., Gross, F., Valensi, F., Delabesse, E., Macintyre, E., Sigaux, F., Soulier, J., Leiva, L. E., Wissler, M., Prinz, C., Rabbitts, T. H., Le Deist, F., Fischer, A., and Cavazzana-Calvo, M., "LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1," *Science*, vol. 302, no. 5644, pp. 415–419, 2003.

[47] Händel, E.-M., Alwin, S., and Cathomen, T., "Expanding or restricting the target site repertoire of zinc-finger nucleases: the inter-domain linker as a major determinant of target site selectivity," *Molecular Therapy*, vol. 17, no. 1, pp. 104–111, 2009.

[48] Hendel, A., Kildebeck, E. J., Fine, E. J., Clark, J. T., Punjya, N., Sebastiano, V., Bao, G., and Porteus, M. H., "Quantifying genome-editing outcomes at endogenous loci with SMRT sequencing," *Cell Rep*, vol. 7, no. 1, pp. 293–305, 2014.

[49] Herrmann, F., Garriga-Canut, M., Baumstark, R., Fajardo-Sanchez, E., Cotterell, J., Minoche, A., Himmelbauer, H., and Isalan, M., "p53 gene repair with zinc finger nucleases optimised by yeast 1-hybrid and validated by solexa sequencing," *PLoS ONE*, vol. 6, no. 6, 2011.

[50] Hockemeyer, D., Soldner, F., Beard, C., Gao, Q., Mitalipova, M., DeKelver, R. C., Katibah, G. E., Amora, R., Boydston, E. A., Zeitler, B., Meng, X., Miller, J. C., Zhang, L., Rebar, E. J., Gregory, P. D., Urnov, F. D., and Jaenisch, R., "Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases," *Nature Biotechnology*, vol. 27, no. 9, pp. 851–857, 2009.

[51] Hockemeyer, D., Wang, H., Kiani, S., Lai, C. S., Gao, Q., Cassady, J. P., Cost, G. J., Zhang, L., Santiago, Y., Miller, J. C., Zeitler, B., Cherone, J. M., Meng, X., Hinkley, S. J., Rebar, E. J., Gregory, P. D., Urnov, F. D., and Jaenisch, R., "Genetic engineering of human pluripotent cells using TALE nucleases," *Nature Biotechnology*, vol. 29, no. 8, pp. 731–734, 2011.

[52] Holkers, M., Maggio, I., Liu, J., Janssen, J. M., Miselli, F., Mussolino, C., Recchia, A., Cathomen, T., and Gonçalves, M. A. F. V., "Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells," *Nucleic Acids Research*, vol. 41, no. 5, pp. e63–e63, 2013.

[53] Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G., and Zhang, F., "DNA targeting specificity of RNA-guided Cas9 nucleases," *Nature Biotechnology*, vol. 31, no. 9, pp. 827–832, 2013.

[54] Huang, P., Xiao, A., Zhou, M., Zhu, Z., Lin, S., and Zhang, B., "Heritable gene targeting in zebrafish using customized TALENs," *Nature Biotechnology*, vol. 29, no. 8, pp. 699–700, 2011.

[55] Jackson, S. P. and Bartek, J., "The DNA-damage response in human biology and disease," *Nature*, vol. 461, no. 7267, pp. 1071–1078, 2009.

[56] Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E., "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity," *Science (New York, N.Y.)*, vol. 337, no. 6096, pp. 816–821, 2012.

[57] Juillerat, A., Dubois, G., Valton, J., Thomas, S., Stella, S., Marchal, A., Langevin, S., Benomari, N., Bertonati, C., Silva, G. H., Daboussi, F., Epinat, J.-C., Montoya, G., Duclert, A., and Duchateau, P., "Comprehensive analysis of the specificity of transcription activator-like effector nucleases," *Nucleic Acids Research*, vol. 42, no. 8, pp. 5390–5402, 2014. 10.1093/nar/gku155.

[58] Kildebeck, E., Checketts, J., and Porteus, M., "Gene therapy for primary immunodeficiencies," *Current Opinion in Pediatrics*, vol. 24, no. 6, pp. 731–738, 2012.

[59] Kim, H. J., Lee, H. J., Kim, H., Cho, S. W., and Kim, J.-S., "Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly," *Genome Research*, vol. 19, no. 7, pp. 1279–1288, 2009.

[60] Kim, Y. G., Cha, J., and Chandrasegaran, S., "Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 3, pp. 1156–1160, 1996.

[61] Kim, Y., Kweon, J., Kim, A., Chon, J. K., Yoo, J. Y., Kim, H. J., Kim, S., Lee, C., Jeong, E., Chung, E., and others, "A library of TAL effector nucleases spanning the human genome," *Nature Biotechnology*, vol. 31, no. 3, pp. 251–258, 2013.

[62] Kim, Y., Kweon, J., and Kim, J.-S., "TALENs and ZFNs are associated with different mutation signatures," *Nature Methods*, vol. 10, no. 3, pp. 185–185, 2013.

[63] Kuhar, R., Gwiazda, K. S., Humbert, O., Mandt, T., Pangallo, J., Brault, M., Khan, I., Maizels, N., Rawlings, D. J., Scharenberg, A. M., and others, "Novel fluorescent genome editing reporters for monitoring DNA repair pathway utilization at endonuclease-induced breaks," *Nucleic Acids Research*, p. gkt872, 2013.

[64] Lee, C. M., Flynn, R., Hollywood, J. A., Scallan, M. F., and Harrison, P. T., "Correction of the ΔF508 mutation in the cystic fibrosis transmembrane conductance regulator gene by zinc-finger nuclease homology-directed repair," *BioResearch Open Access*, vol. 1, no. 3, pp. 99–108, 2012.

[65] Lee, H. J., Kim, E., and Kim, J.-S., "Targeted chromosomal deletions in human cells using zinc finger nucleases," *Genome Research*, vol. 20, no. 1, pp. 81–89, 2010.

[66] Lei, Y., Guo, X., Liu, Y., Cao, Y., Deng, Y., Chen, X., Cheng, C. H. K., Dawid, I. B., Chen, Y., and Zhao, H., "Efficient targeted gene disruption in Xenopus embryos using engineered transcription activator-like effector nucleases (TALENs)," *Proceedings of the National Academy of Sciences*, vol. 109, no. 43, pp. 17484–17489, 2012.

[67] Li, T., Liu, B., Spalding, M. H., Weeks, D. P., and Yang, B., "High-efficiency TALEN-based gene editing produces disease-resistant rice," *Nature Biotechnology*, vol. 30, no. 5, pp. 390–392, 2012.

[68] Lin, Y., Cradick, T. J., Brown, M. T., Deshmukh, H., Ranjan, P., Sarode, N., Wile, B. M., Vertino, P. M., Stewart, F. J., and Bao, G., "CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences," *Nucleic Acids Research*, vol. 42, no. 11, pp. 7473–7485, 2014. 10.1093/nar/gku402.

[69] Lin, Y., Fine, E. J., Zheng, Z., Antico, C. J., Voit, R. A., Porteus, M. H., Cradick, T. J., and Bao, G., "SAPTA: a new design tool for improving TALE nuclease activity," *Nucleic Acids Research*, vol. 42, no. 6, pp. e47–e47, 2014. 10.1093/nar/gkt1363.

[70] Lombardo, A., Genovese, P., Beausejour, C. M., Colleoni, S., Lee, Y.-L., Kim, K. A., Ando, D., Urnov, F. D., Galli, C., Gregory, P. D., Holmes, M. C., and Naldini, L., "Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery," *Nature Biotechnology*, vol. 25, no. 11, pp. 1298–1306, 2007.

[71] Loomis, E. W., Eid, J. S., Peluso, P., Yin, J., Hickey, L., Rank, D., McCalmon, S., Hagerman, R. J., Tassone, F., and Hagerman, P. J., "Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene," *Genome research*, vol. 23, no. 1, pp. 121–128, 2013.

[72] Maeder, M. L., Thibodeau-Beganny, S., Osiak, A., Wright, D. A., Anthony, R. M., Eichtinger, M., Jiang, T., Foley, J. E., Winfrey, R. J., Townsend, J. A., Unger-Wallace, E., Sander, J. D., Mller-Lerch, F., Fu, F., Pearlberg, J., Gbel, C., Dassie, J. P., Pruett-Miller, S. M., Porteus, M. H., Sgroi, D. C., Iafrate, A. J., Dobbs, D., McCray, P. B., Cathomen, T., Voytas, D. F., and Joung, J. K., "Rapid open-source engineering of customized zinc-finger nucleases for highly efficient gene modification," *Molecular cell*, vol. 31, no. 2, pp. 294–301, 2008.

[73] Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., and Church, G. M., "RNA-guided human genome engineering via Cas9," *Science*, vol. 339, no. 6121, pp. 823–826, 2013.

[74] McMahon, M. A., Rahdar, M., and Porteus, M., "Gene editing: not just for translation anymore," *Nature Methods*, vol. 9, no. 1, pp. 28–31, 2012.

[75] Meckler, J. F., Bhakta, M. S., Kim, M.-S., Ovadia, R., Habrian, C. H., Zykovich, A., Yu, A., Lockwood, S. H., Morbitzer, R., Elsesser, J., Lahaye, T., Segal, D. J., and Baldwin, E. P., "Quantitative analysis of TALEDNA interactions suggests polarity effects," *Nucleic Acids Research*, vol. 41, no. 7, pp. 4118–4128, 2013.

[76] Miller, J. C., Holmes, M. C., Wang, J., Guschin, D. Y., Lee, Y.-L., Rupniewski, I., Beausejour, C. M., Waite, A. J., Wang, N. S., Kim, K. A., Gregory, P. D., Pabo, C. O., and Rebar, E. J., "An improved zinc-finger nuclease architecture for highly specific genome editing," *Nature Biotechnology*, vol. 25, no. 7, pp. 778–785, 2007.

[77] Miller, J. C., Tan, S., Qiao, G., Barlow, K. A., Wang, J., Xia, D. F., Meng, X., Paschon, D. E., Leung, E., Hinkley, S. J., Dulay, G. P., Hua, K. L., Ankoudinova, I., Cost, G. J., Urnov, F. D., Zhang, H. S., Holmes, M. C., Zhang, L., Gregory, P. D., and Rebar, E. J., "A TALE nuclease architecture for efficient genome editing," *Nature Biotechnology*, vol. 29, no. 2, pp. 143–148, 2011.

[78] Moscou, M. J. and Bogdanove, A. J., "A simple cipher governs dna recognition by TAL effectors," *Science (New York, N.Y.)*, vol. 326, no. 5959, 2009.

[79] Mussolino, C., Alzubi, J., Fine, E. J., Morbitzer, R., Cradick, T. J., Lahaye, T., Bao, G., and Cathomen, T., "TALENs facilitate targeted genome editing in human cells with high specificity and low cytotoxicity," *Nucleic Acids Research*, vol. 42, no. 10, pp. 6762–6773, 2014.

[80] Mussolino, C., Morbitzer, R., Lütge, F., Dannemann, N., Lahaye, T., and Cathomen, T., "A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity," *Nucleic Acids Research*, vol. 39, no. 21, pp. 9283–9293, 2011.

[81] Needleman, S. B. and Wunsch, C. D., "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[82] O'Geen, H., Henry, I. M., Bhakta, M. S., Meckler, J. F., and Segal, D. J., "A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture," *bioRxiv*, 2014.

[83] Orlando, S. J., Santiago, Y., DeKelver, R. C., Freyvert, Y., Boydston, E. A., Moehle, E. A., Choi, V. M., Gopalan, S. M., Lou, J. F., Li, J., and others, "Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology," *Nucleic Acids Research*, vol. 38, no. 15, pp. e152–e152, 2010.

[84] Osborn, M. J., Starker, C. G., McElroy, A. N., Webber, B. R., Riddle, M. J., Xia, L., DeFeo, A. P., Gabriel, R., Schmidt, M., Von Kalle, C., Carlson, D. F., Maeder, M. L., Joung, J. K., Wagner, J. E., Voytas, D. F., Blazar, B. R., and Tolar, J., "TALEN-based gene correction for epidermolysis bullosa," *Molecular therapy: the journal of the American Society of Gene Therapy*, vol. 21, no. 6, pp. 1151–1159, 2013.

[85] Ousterout, D. G., Perez-Pinera, P., Thakore, P. I., Kabadi, A. M., Brown, M. T., Qin, X., Fedrigo, O., Mouly, V., Tremblay, J. P., and Gersbach, C. A., "Reading frame correction by targeted genome editing restores dystrophin expression in cells from Duchenne muscular dystrophy patients," *Molecular Therapy*, 2013.

[86] Papamichos-Chronakis, M. and Peterson, C. L., "Chromatin and the genome integrity network," *Nature Reviews Genetics*, vol. 14, no. 1, pp. 62–75, 2013.

[87] Pattanayak, V., Lin, S., Guilinger, J. P., Ma, E., Doudna, J. A., and Liu, D. R., "High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity," *Nature Biotechnology*, vol. 31, no. 9, pp. 839–843, 2013.

[88] Pattanayak, V., Ramirez, C. L., Joung, J. K., and Liu, D. R., "Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection," *Nature Methods*, vol. 8, no. 9, pp. 765–770, 2011.

[89] Perez, E. E., Wang, J., Miller, J. C., Jouvenot, Y., Kim, K. A., Liu, O., Wang, N., Lee, G., Bartsevich, V. V., Lee, Y.-L., Guschin, D. Y., Rupniewski, I., Waite, A. J., Carpenito, C., Carroll, R. G., S Orange, J., Urnov, F. D., Rebar, E. J., Ando, D., Gregory, P. D., Riley, J. L., Holmes, M. C., and June, C. H., "Establishment of HIV-1 resistance in $CD4^+$ T cells by genome editing using zinc-finger nucleases," *Nature Biotechnology*, vol. 26, no. 7, pp. 808–816, 2008.

[90] Porteus, M. H. and Baltimore, D., "Chimeric nucleases stimulate gene targeting in human cells," *Science*, vol. 300, no. 5620, pp. 763–763, 2003.

[91] Porteus, M. H. and Carroll, D., "Gene targeting using zinc finger nucleases," *Nature Biotechnology*, vol. 23, no. 8, pp. 967–973, 2005.

[92] Qi, Y., Zhang, Y., Zhang, F., Baller, J. A., Cleland, S. C., Ryu, Y., Starker, C. G., and Voytas, D. F., "Increasing frequencies of site-specific mutagenesis and gene

targeting in Arabidopsis by manipulating DNA repair pathways," *Genome Research*, vol. 23, no. 3, pp. 547–554, 2013.

[93] Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y., "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers," *BMC Genomics*, vol. 13, no. 1, 2012.

[94] Ramirez, C. L., Certo, M. T., Mussolino, C., Goodwin, M. J., Cradick, T. J., McCaffrey, A. P., Cathomen, T., Scharenberg, A. M., and Joung, J. K., "Engineered zinc finger nickases induce homology-directed repair with reduced mutagenic effects," *Nucleic Acids Research*, vol. 40, no. 12, pp. 5560–5568, 2012.

[95] Ran, F. A., Hsu, P. D., Lin, C.-Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., Scott, D. A., Inoue, A., Matoba, S., Zhang, Y., and Zhang, F., "Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity," *Cell*, vol. 154, no. 6, pp. 1380–1389, 2013.

[96] Rasko, D. A., Webster, D. R., Sahl, J. W., Bashir, A., Boisen, N., Scheutz, F., Paxinos, E. E., Sebra, R., Chin, C.-S., Iliopoulos, D., and others, "Origins of the E. coli strain causing an outbreak of hemolytic–uremic syndrome in germany," *New England Journal of Medicine*, vol. 365, no. 8, pp. 709–717, 2011.

[97] Reyon, D., Tsai, S. Q., Khayter, C., Foden, J. A., Sander, J. D., and Joung, J. K., "FLASH assembly of TALENs for high-throughput genome editing," *Nature Biotechnology*, vol. 30, no. 5, pp. 460–465, 2012.

[98] Roberts, R. J., Carneiro, M. O., and Schatz, M. C., "The advantages of SMRT sequencing," *Genome Biology*, vol. 14, no. 6, p. 405, 2013.

[99] Ropers, H.-H., "Single gene disorders come into focus - again," *Dialogues in Clinical Neuroscience*, vol. 12, no. 1, pp. 95–102, 2010.

[100] Rozen, S. and Skaletsky, H., *Primer3 on the WWW for General Users and for Biologist Programmers*, pp. 365–386. Totowa, NJ: Humana Press, 1999.

[101] Sander, J. D., Ramirez, C. L., Linder, S. J., Pattanayak, V., Shoresh, N., Ku, M., Foden, J. A., Reyon, D., Bernstein, B. E., Liu, D. R., and Joung, J. K., "In silico abstraction of zinc finger nuclease cleavage profiles reveals an expanded landscape of off-target sites," *Nucleic Acids Research*, 2013.

[102] Schipler, A. and Iliakis, G., "DNA double-strandbreak complexity levels and their possible contributions to the probability for error-prone processing and repair pathway choice," *Nucleic Acids Research*, 2013.

[103] Sebastiano, V., Maeder, M. L., Angstman, J. F., Haddad, B., Khayter, C., Yeo, D. T., Goodwin, M. J., Hawkins, J. S., Ramirez, C. L., Batista, L. F. Z., Artandi,

S. E., Wernig, M., and Joung, J. K., "In situ genetic correction of the sickle cell anemia mutation in human induced pluripotent stem cells using engineered zinc finger nucleases," *Stem Cells*, vol. 29, no. 11, pp. 1717–1726, 2011.

[104] Segal, D. J. and Meckler, J. F., "Genome engineering at the dawn of the golden age," *Annual Review of Genomics and Human Genetics*, vol. 14, no. 1, pp. 135–158, 2013.

[105] Shaw, K. L. and Kohn, D. B., "A tale of two SCIDs," *Science Translational Medicine*, vol. 3, no. 97, pp. 97ps36–97ps36, 2011.

[106] Silva, G., Poirot, L., Galetto, R., Smith, J., Montoya, G., Duchateau, P., and others, "Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy," *Current Gene Therapy*, vol. 11, no. 1, p. 11, 2011.

[107] Soldner, F., Laganière, J., Cheng, A. W., Hockemeyer, D., Gao, Q., Alagappan, R., Khurana, V., Golbe, L. I., Myers, R. H., Lindquist, S., and others, "Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations," *Cell*, vol. 146, no. 2, pp. 318–331, 2011.

[108] Şöllü, C., Pars, K., Cornu, T. I., Thibodeau-Beganny, S., Maeder, M. L., Joung, J. K., Heilbronn, R., and Cathomen, T., "Autonomous zinc-finger nuclease pairs for targeted chromosomal deletion," *Nucleic Acids Research*, p. gkq720, 2010.

[109] Stark, J. M., Pierce, A. J., Oh, J., Pastink, A., and Jasin, M., "Genetic steps of mammalian homologous repair with distinct mutagenic consequences," *Molecular and Cellular Biology*, vol. 24, no. 21, pp. 9305–9316, 2004.

[110] Steinberg, M. H., "Sickle cell anemia, the first molecular disease: overview of molecular etiology, pathophysiology, and therapeutic approaches," *TheScientificWorldJournal*, vol. 8, pp. 1295–1324, 2008.

[111] Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C., and Doudna, J. A., "DNA interrogation by the CRISPR RNA-guided endonuclease Cas9," *Nature*, vol. 507, no. 7490, pp. 62–67, 2014.

[112] Streubel, J., Blcher, C., Landgraf, A., and Boch, J., "TAL effector RVD specificities and efficiencies," *Nature Biotechnology*, vol. 30, no. 7, pp. 593–595, 2012.

[113] Sun, N. and Zhao, H., "Seamless correction of the sickle cell disease mutation of the HBB gene in human induced pluripotent stem cells using TALENs," *Biotechnology and Bioengineering*, vol. 111, no. 5, pp. 1048–1053, 2014.

[114] Szczepek, M., Brondani, V., Büchel, J., Serrano, L., Segal, D. J., and Cathomen, T., "Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases," *Nature Biotechnology*, vol. 25, no. 7, pp. 786–793, 2007.

[115] Tebas, P., Stein, D., Tang, W. W., Frank, I., Wang, S. Q., Lee, G., Spratt, S. K., Surosky, R. T., Giedlin, M. A., Nichol, G., and others, "Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV," *New England Journal of Medicine*, vol. 370, no. 10, pp. 901–910, 2014.

[116] Tesson, L., Usal, C., Menoret, S., Leung, E., Niles, B. J., Remy, S., Santiago, Y., Vincent, A. I., Meng, X., Zhang, L., Gregory, P. D., Anegon, I., and Cost, G. J., "Knockout rats generated by embryo microinjection of TALENs," *Nature Biotechnology*, vol. 29, no. 8, pp. 695–696, 2011.

[117] Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S., and Turner, S. W., "A flexible and efficient template format for circular consensus sequencing and SNP detection," *Nucleic Acids Research*, vol. 38, no. 15, pp. e159–e159, 2010.

[118] Tsai, S. Q., Wyvekens, N., Khayter, C., Foden, J. A., Thapar, V., Reyon, D., Goodwin, M. J., Aryee, M. J., and Joung, J. K., "Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing," *Nature Biotechnology*, vol. 32, no. 6, pp. 569–576, 2014.

[119] Tsai, S. Q., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A. J., Le, L. P., Aryee, M. J., and Joung, J. K., "GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases," *Nature Biotechnology*, vol. advance online publication, 2014.

[120] Urnov, F. D., Miller, J. C., Lee, Y.-L., Beausejour, C. M., Rock, J. M., Augustus, S., Jamieson, A. C., Porteus, M. H., Gregory, P. D., and Holmes, M. C., "Highly efficient endogenous human gene correction using designed zinc-finger nucleases," *Nature*, vol. 435, no. 7042, pp. 646–651, 2005.

[121] Üstek, D., Sirma, S., Gumus, E., Arikan, M., Cakiris, A., Abaci, N., Mathew, J., Emrence, Z., Azakli, H., Cosan, F., Cakar, A., Parlak, M., and Kursun, O., "A genome-wide analysis of lentivector integration sites using targeted sequence capture and next generation sequencing technology," *Infection, Genetics and Evolution*, vol. 12, no. 7, pp. 1349–54, 2012.

[122] Valton, J., Daboussi, F., Leduc, S., Molina, R., Redondo, P., Macmaster, R., Montoya, G., and Duchateau, P., "5'-cytosine-phosphoguanine (CpG) methylation impacts the activity of natural and engineered meganucleases," *Journal of Biological Chemistry*, vol. 287, no. 36, pp. 30139–30150, 2012.

[123] Valton, J., Dupuy, A., Daboussi, F., Thomas, S., Maréchal, A., Macmaster, R., Melliand, K., Juillerat, A., and Duchateau, P., "Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation," *Journal of Biological Chemistry*, vol. 287, no. 46, pp. 38427–38432, 2012.

[124] van Rensburg, R., Beyer, I., Yao, X.-Y., Wang, H., Denisenko, O., Li, Z.-Y., Russell, D. W., Miller, D. G., Gregory, P., Holmes, M., and others, "Chromatin structure of