DESIGN AND OPTIMIZATION OF ENGINEERED NUCLEASES FOR GENOME EDITING APPLICATIONS

A Dissertation
Presented to
The Academic Faculty

by

Yanni Lin

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Engineering

Georgia Institute of Technology
December 2014

DESIGN AND OPTIMIZATION OF ENGINEERED NUCLEASES FOR GENOME EDITING APPLICATIONS

Approved by:

Dr. Gang Bao, Advisor
Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. Paul Spearman
Department of Pediatrics
Emory University and Children's Healthcare of Atlanta

Dr. Mark Prausnitz
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

Dr. Wilbur Lam
Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. Manu Platt
Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Date Approved: August 05, 2014

To Lin HongJie, Zhao Xia, Lin Hao, and Chen Weixuan

ACKNOWLEDGEMENTS

I would like to first thank my advisor, Dr. Gang Bao, who decided to offer me a position in his lab five years ago, and generously provided me with funding, guidance, and support throughout my PhD study. I am grateful for the great suggestions I received from my committee members: Dr. Mark Prausnitz, Dr. Paul Spearman, Dr. Wilbur Lam, and Dr. Manu Platt. I would also like to thank my friendly and cheerful group of labmates, who have been my consultants for all the things I want to know in graduate school and science, for all the fun we had in casual chats, lab outings and lab Olympics. My thanks also goes to my collaborators for their help in my project. My research would not have been possible without their help. I am also thankful for my undergraduate research assistants for their hard work that accelerate my research.

I would like to acknowledge my mother and father, who have always been supportive through my life and provide unconditional love and care. I also thank my friends for providing support and friendship that I needed. I would most like to thank my beloved husband Weixuan Chen and my two cats, Simba and Little Monster, for their company. My husband has been the best man in the world, who cooks delicious food, takes care of everything, and always stands by my side through the good times and bad.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS AND ABBREVIATIONS

AAV	Adeno-Associated Virus
bp	Base Pair
BWA	Borrows-Wheeler Aligner
Cas	Crispr Associated
CCR5	C-C Chemokine Receptor Type 5
CF	Cystic Fibrosis
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DNA	Deoxyribonucleic Acid
DSB	Double Stranded Break
FACS	Fluorescence-Activated Cell Sorting
GAPDH	Glyceraldehyde 3-Phosphate Dehydrogenase
GFP	Green Fluorescent Protein
GG	Golden Gate
HAART	Highly Active Antiretroviral Therapy
HBB	Hemoglobin, Beta
HBD	Hemoglobin, Delta
HDR	Homology-Directed Repair
HEK293T	Human Embryonic Kidney 293 with SV40 Large T Antigen
HIV	Human Immunodeficiency Virus
HR	Homologous Recombination
indel	Insertions and Deletions
LTR	Long Terminal Repeat
NGS	Next-Generation Sequencing
NHEJ	Non-Homologous End Joining
NLS	Nuclease Locolization Signal
nt	Nucleotide
PAM	Protospacer-Adjacent Motif
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid
RFLP	Restriction Fragment Length Polymorphism
RVD	Repeat Variable Diresidue
s.e.m.	Standard Error of the Mean
SAPTA	Scoring Algorithm for Predicting TALEN Activity
SCD	Sickle Cell Disease
SMA	Spinal Muscular Atrophy
SMRT	Single Molecule Real Time
SSA	Single-Strand Annealing

T7E1	T7 Endonuclease I
TAL	Transcription Activator Like
TALE	Transcription Activator-Like Effectors
TALEN	Transcription-Activator-Like Effector Nuclease
ZFN	Zinc Finger Nuclease
ATCC	American Type Culture Collection
DMEM	Dulbecco's Modified Eagle Medium
EGFP	Enhanced Green Fluorescent Protein
ENCODE	The Encyclopedia Of DNA Elements
ORF	Open Reading Frame
PROGNOS	Predicted Report Of Genome-wide Nuclease Off-target Sites
UCSC	University of California, Santa Cruz

SUMMARY

Genome editing mediated by engineered nucleases, including Transcription Activator-Like Effector Nucleases (TALENs) and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) / CRISPR-associated (Cas) systems, holds great potential in a broad range of applications, including biomedical studies and disease treatment. In addition to creating cell lines and disease models, this technology allows generation of well-defined, genetically modified cells and organisms with novel characteristics that can be used to cure diseases, study gene functions, and facilitate drug development. However, achieving both high efficiency and high specificity remains a major challenge in nuclease-based genome editing. The objectives of this thesis were to optimize the design of TALENs to achieve high on-target cleavage activity, and analyze the off-target effect of CRISPR/Cas to help achieve high specificity. Based on experimental evaluation of >200 TALENs, we compared three different TALEN architectures, proposed new TALEN design rules, and developed a Scoring Algorithm for Predicting TALEN Activity (SAPTA) to identify optimal target sites with high activity. We also performed a systematic study to demonstrate the off-target cleavage by CRISPR/Cas9 when DNA sequences contain insertions or deletions compared to the RNA guide strand. Our results strongly indicate the need to perform comprehensive off-target analysis, and suggest specific guidelines for reducing potential off-target cleavage of CRISPR/Cas9 systems. The studies performed in this thesis work provide important insight and powerful tools for the optimization of engineered nucleases in genome editing, thus making a significant contribution to biomedical engineering and medical applications.

CHAPTER 1: INTRODUCTION

1.1 Genome editing approach to treat diseases

Single gene disorders account for over 10,000 diseases(8), including a large number of prevalent diseases, such as sickle-cell disease (SCD), cystic fibrosis (CF), and spinal muscular atrophy (SMA). These gene disorders impose a significant health and financial burden worldwide. Sickle cell disease (SCD) is a well-studied paradigm of single gene disorders. The majority of SCD patients carry an A to T mutation in both alleles of the β-globin gene (*HBB*) that changes a single amino acid from glutamate to valine. This small change leads to a malfunctioned form of adult hemoglobin, shortening the lifespan and rendering painful symptoms and complications for patients(9). Although blood and marrow stem cell transplants may cure a small number of patients, there is no widely available cure for this disease.

Genome editing techniques hold great potential in treating challenging diseases by precise manipulation of the genome, especially with the application of sequence-specific designer nucleases (Figure 1). With wild-type donor templates, the DNA double stranded break (DSB) generated by the designer nucleases can be repaired through the homologous recombination (HR) repair pathway, which will use the genetic information provided by the donor template to correct the SCD mutation.

In the absence of a donor DNA, the DSB will be repaired by the non-homologous end-joining (NHEJ) pathway. This error-prone pathway can then generate insertions and deletions at the site of the DSB to inactivate a gene, or a large targeted deletion in the range of kilobases if two DSBs are created at adjacent sites on a chromosome(10,11). The

NHEJ-mediated genome editing can be applied to inactivate essential genes of HIV proviruses integrated into HIV-infected cells or excise the entire HIV provirus DNA from human genome.

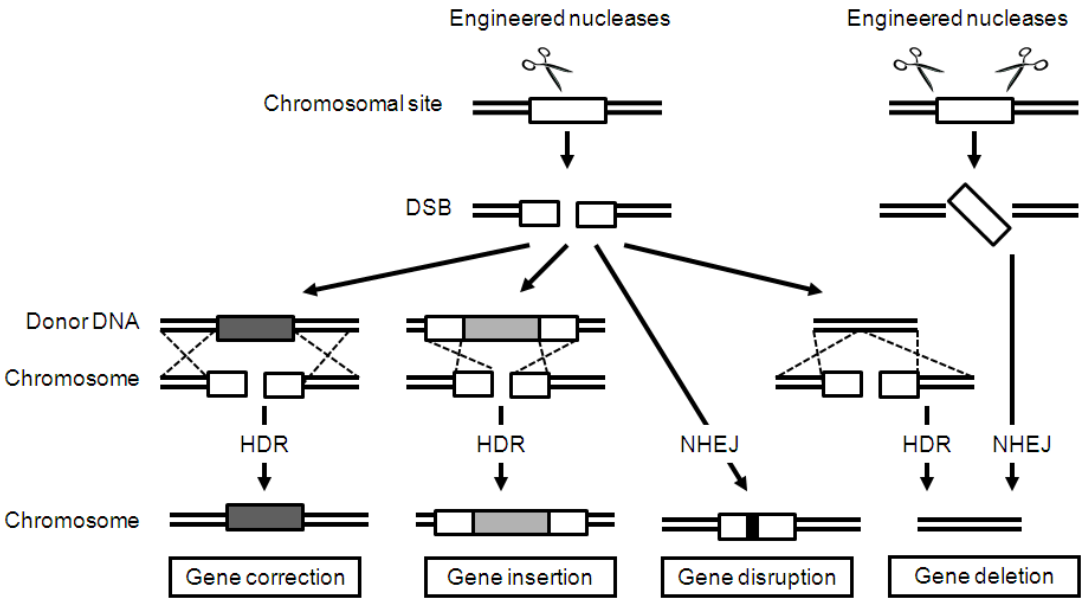


Figure 1: Nuclease-mediated genome editing.
A DNA double stranded break (DSB) generated by engineered nucleases, such as ZFNs, TALENs, or CRISPRs, can be repaired by one of the two major DNA repair pathways: homology-directed repair (HDR) or non-homologous end-joining (NHEJ). These repair mechanisms will lead to different modifications in the genome.

Gene targeting approach can potentially cure SCD at molecular levels by introducing the corrected DNA sequence to replace the sickle cell mutation using HR pathway. However, the application of this approach has been limited by its low

efficiency. It was found that the efficiency of gene targeting can be substantially improved by the use of designer nucleases to generate DSBs in target genes, which then stimulate the HR pathway(12). By correcting the "sickle" mutation in hematopoietic stem cells from patients ex vivo and infusing the corrected cells back into patients, we may be able to provide a permanent cure for this disease (Figure 2).

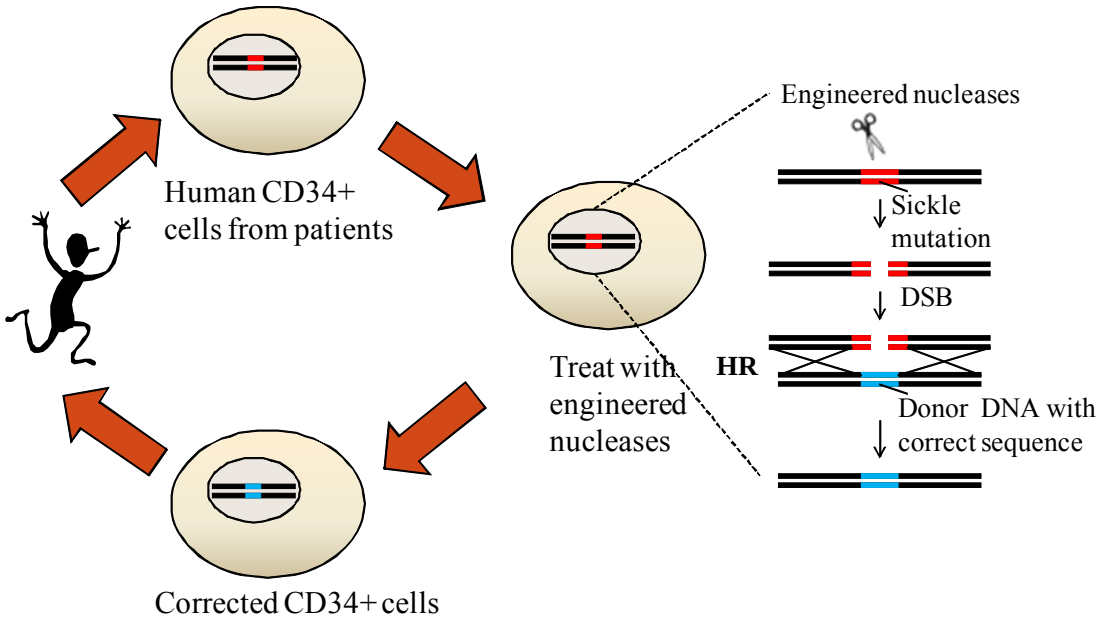


Figure 2: Schematic of genome editing approach to treat SCD.
Blood cells from patients can be enriched for hematopoietic stem cells by sorting for the CD34 marker on cell surface. Enriched cells are treated with engineered nucleases and donor DNA with correct, non-sickle sequence. Cells containing corrected stem cells can be infused back into patients to generate healthy red blood cells.

Designer nuclease can also be simultaneously targeted to two adjacent sites on the same chromosome to obtain targeted deletion of large DNA fragments. Our lab proposed to treat HIV-infected cells by excising the integrated HIV proviruses from the genome. Although progress has been made in prevention and treatment of HIV, the HIV-infected population continues to grow and remains a major concern of global health. The number of people living with HIV in 2008 was estimated to be 33.4 million(13). Highly active antiretroviral therapy (HAART) has been remarkably successful, often resulting in undetectable level of viral load in treated patients. However, complete eradication of HIV infection has thus far been unattainable, due to the protected reservoirs of latently infected cells (14). We proposed to design engineered nucleases to cleave the common sequences shared by 5' and 3' long terminal repeats (LTRs) located at both ends of the HIV genome, thus excising the integrated HIV provirus after two concurrent DSBs are generated in the chromosome (Figure 3). This unique strategy could truly remove virus from persistently-infected cells and provide a cure for infected individuals.

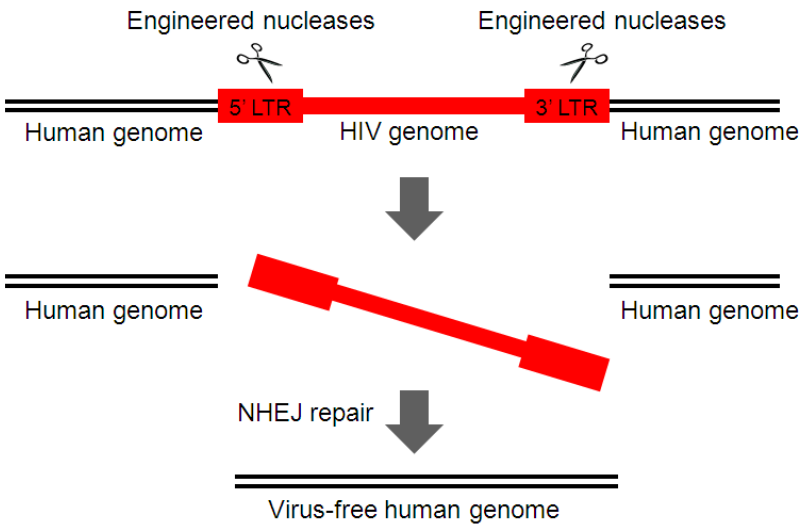


Figure 3: Schematic of removing the HIV provirus from infected human cells.

1.2 Engineered nucleases used in genome editing

The approaches described above that use genome editing approach to treat SCD and HIV infection require well-designed engineered nucleases which are highly sequence-specific and highly active. Several major classes of designer nucleases—meganucleases, Zinc Finger Nucleases (ZFNs), TALENs, and CRISPRs—have been successfully deployed in gene engineering(12,15-18) in different organisms for applications ranging from creating model organisms to treating genetic diseases. A barrier to the widespread application for meganucleases and ZFNs has been difficulty in obtaining new DNA sequence specificities through rationale design. This is largely due to the context dependent nature of their DNA-binding units. TALENs and CRISPRs have been developed rapidly into major tools for genome manipulation as it is much easier to

design than other designer nucleases due to their predictable and context-independent DNA-binding domains.

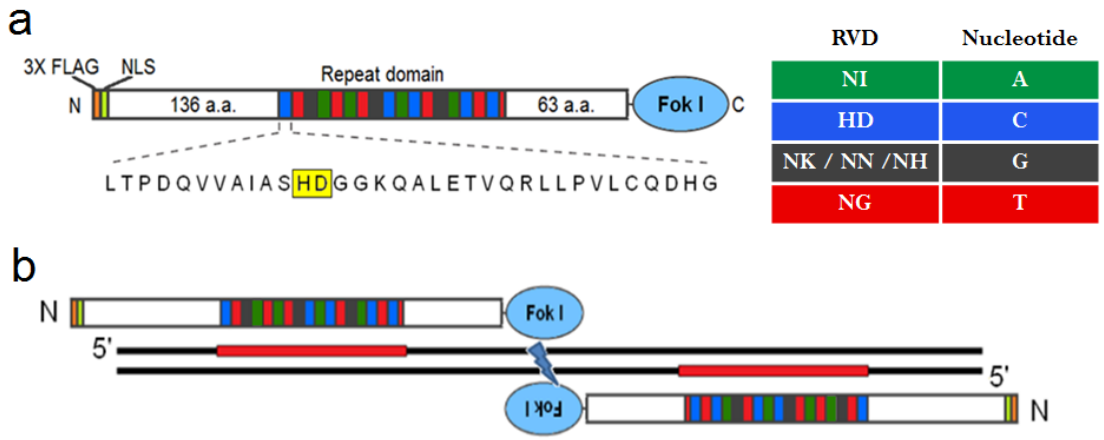


Figure 4: The architecture of TALEN.
(a) Structure of TALEN. The repeat domain which directly recognizes DNA molecules is flanked by an N-terminal domain with 136 amino acids and a C-terminal with 63 amino acids. The amino acid sequence of one repeat is shown with RVD boxed in yellow. The RVDs recognizing each nucleotide are shown on the right. The 3X FLAG epitope and nuclease localization signal (NLS) are also shown. (b) Schematic of a pair of TALENs generating a DSB in the target site.

TALENs are a family of DNA binding proteins, discovered in the plant pathogen Xanthomonas(19-22). Each DNA-binding domain contains a variable number of 33-35 amino-acid repeats that specify the DNA-binding sequence primarily through their 12th and 13th repeat-variable di-residues (RVDs)(19). Each RVD specifies one nucleotide with minimal context dependence(20,22,23). A TALEN targets a specific DNA sequence

by designing a set of repeats flanked by modified N- and C-termini(24,25) and linked to a FokI nuclease domain(15,26,27). When a pair of TALENs binds to their specific half-sites with the correct orientation and spacing to allow the nuclease domains to dimerize, the intervening sequence is cleaved (Figure 4). TALENs have been used to edit genomic DNA sequences in a variety of biological systems, including human cells, rats, zebrafish, nematodes, and plants(23-25,28-32).

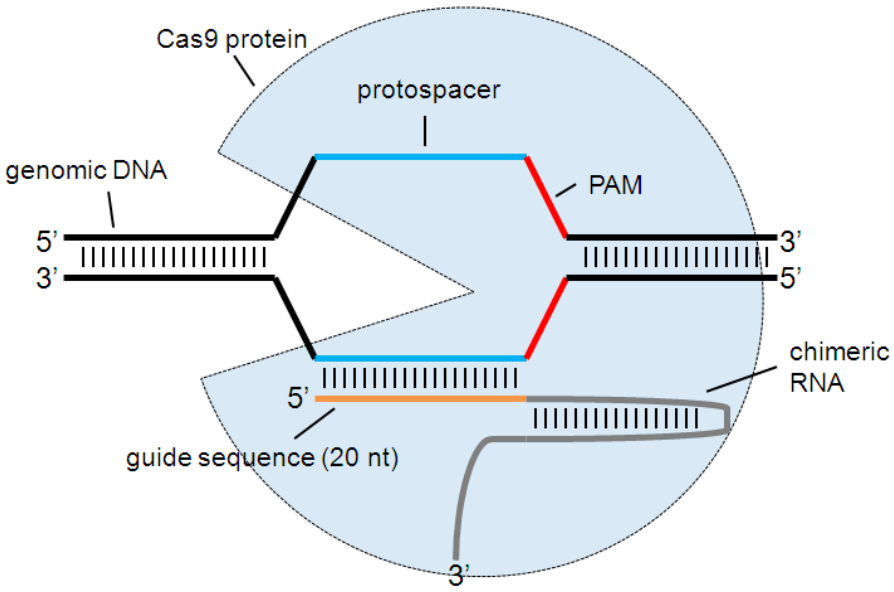


Figure 5: The architecture of the CRISPR/Cas9 system.
The guide RNA is a chimeric RNA with the first 20-nt sequence complimentary to the genomic DNA sequence at the target site. This 20-nt guide sequence direct where the Cas9 nuclease cleaves. The protospacer-adjacent motif (PAM) is a short sequence required right next to the target site (protospacer) for the CRISPR/Cas9 system to function.

CRISPR and CRISPR-associated (Cas) proteins constitute a bacterial defense system that cleaves invading foreign nucleic acids (33-40). Chimeric single-guided RNAs (sgRNAs) based on CRISPR (41) have been engineered to direct the Cas9 nuclease to cleave complementary genomic sequences when followed by a 5'-NGG protospacer-adjacent motif (PAM) in eukaryotic cells (17,42,43). Since gene targeting by CRISPR/Cas9 is directed by base pairing, such that only the short 20-nt sequence of the sgRNA needs to be changed for different target sites, CRISPR/Cas systems enable simultaneous targeting of multiple DNA sequences and robust gene modification (17,18,41,42,44-48).

1.3 Project overview

One challenge in TALEN design has been choosing an optimal RVD to recognize guanine. As shown in Figure 4, NN (Asn-Asn), NK (Asn-Lys), and NH (Asn-His) are three guanine-specific RVDs that can be used to construct TALENs for a specific DNA target sequence. NN RVD has been the most commonly used RVD for guanine. TALENs with NN RVDs (NN-TALENs) have a higher binding affinity than those with NK RVDs (NK-TALENs); however, their specificity is low since NN RVDs bind to both guanine and adenine (1,24,30,49,50). Recent studies investigated a new RVD NH using TAL effectors, which activate transcription once bound to a specific sequence. The results show that the NH RVD gives medium to high activity, together with a high specificity(1,50). While the NH RVD appears to be promising, it still remains to be seen whether high activity and specificity can be obtained when this RVD is built into TALENs. Specific Aim 1 addressed the question on which RVD to choose for optimized TALEN performance.

Another challenge in TALEN design is obtaining high nuclease activity while using the RVD NK that favors specificity rather than activity. The codes of nucleotide recognition by RVDs have been established(20,22), as shown in Figure 4a. If we choose NK as the G-specific RVD, for a given DNA sequence, we can readily assemble an array of repeats with corresponding RVDs to recognize the sequence. The design of TALEN is thus equivalent to choosing an optimal target DNA sequence. Due to the weaker RVD efficiency of NK, it has been difficult to obtain active NK-TALENs without careful selection of target sequences. Specific Aim 2 of this thesis addressed the question regarding selection of TALEN target sites for achieving high nuclease activity.

The major concern in targeting genes with CRISPR/Cas9 systems is their relatively low specificity. Recent studies reveal that CRISPR can even cut at genomic sites containing up to five base mismatches out of the total 20-nt target sequence(5,51-55). Carefully selection of target sites less similar to other genomic sequences becomes critical in designing CRISPRs. However, it is not definitive that mismatch is the only type of sequence variance that can be tolerated by CRISPRs, so other sequence variances also need to be interrogated to ensure specific genome editing by CRISPR/Cas9. Specific Aim 3 focused on the identification of CRISPR off-target effect at genomic sites containing missing bases (deletions) or extra bases (insertions) compared to the guide RNAs.

The goal of the thesis is to improve the performance of two new classes of engineered nucleases, TALENs and CRISPRs. The results will provide the community with a powerful and reliable tool for genetic engineering to treat and model diseases. Optimized engineered nucleases will also enable the development of novel paradigms in

curing single gene disorders and HIV infection using genome editing approaches. The goal of this thesis is fulfilled by performing studies with the following specific aims:

Specific aim 1

This aim is to optimize TALEN architecture by comparing the activity and specificity of TALENs constructed with different repeat-variable di-residues (RVDs) recognizing guanine. TALENs targeting sites in the different human genes were constructed with one of the three different RVDs, Asn-Asn (NN), Asn-Lys (NK), and Asn-His (NH). The on-target and off-target mutagenesis of TALEN pairs containing different RVDs was measured by a T7 endonuclease I (T7E1) mutation detection assay and deep sequencing with PCR products amplifying the genomic sites in HEK293T cells. The ability of these TALENs to trigger homology-directed repair (HDR) at the β-globin gene is quantified by the percentages of stably integrated GFP gene in K562 cells.

Specific aim 2

This aim is to develop a scoring algorithm to predict TALEN activities and validate its ability to predict optimal TALEN target sites. Around two hundred TALENs were designed to target sequences with different features, and constructed using a high-throughput protocol involving robot liquid handler. TALEN monomer activity was measured by the modified single-strand annealing (SSA) assay in HEK293T cells. A Scoring Algorithm for Predicting TALEN Activity (SAPTA), which gives a numerical score that predicts TALEN activity (a high score predicts a high activity), was developed using the measured monomer activities as training-set data. We then designed TALEN pairs targeting disease-related genes using SAPTA and measured the monomer and pair

activities in HEK293T cells. The results were used as test-set data to validate the effectiveness of the algorithm.

Specific aim 3

This aim is to discover and characterize another dimension of CRISPR off-target effect in addition to mismatch tolerance. Guide RNAs target to *HBB* and *CCR5* genes were modified by deleting bases from or inserting bases into the guide sequences, and the mutagenesis at these genes resulted from the guide RNA variants was quantified using the T7E1 assay in HEK293T cells. We then scanned the human genome for potential off-target sites containing insertions or deletions compared to randomly designed guide RNAs targeted to different genes. T7E1 assay and deep sequencing were used to detect any mutations at these potential off-target sites. An online tool that searches for genomic sites containing mismatches, insertions, and deletions was developed to help researchers in the community select CRISPRs with higher specificity.

CHAPTER 2: NUCLEASE ASSEMBLY AND TEST PIPELINE

2.1 Construction of TALENs

The ability to assemble TALENs in a high-throughput manner is essential for carrying out studies in this proposal. Construction of TALENs is not straight-forward due to the repetitive DNA sequence coding the TALEN DNA-binding domains. Various methods based on the Golden Gate (GG) cloning reaction were developed to overcome the difficulty(2,23,56,57).

We have successfully assembled TALENs using a hierarchical ligation-based strategy which involves two steps with one Golden Gate (GG) cloning reaction in each step(23). Plasmid tool kit was kindly provided by Dr. Daniel F. Voytas, University of Minnesota. These plasmids contain individual TALEN DNA-binding repeats, intermediate backbone vectors, and final destination backbone vectors. Repeat-encoding plasmids contain RVDs HD, NI, NG, and NN to recognize nucleotides C A, T, and G, respectively. There are two additional sets of repeat plasmids with RVDs NK and NH to recognize guanine.

The first GG reaction using BsaI restriction sites links individual DNA-binding repeats into an intermediate array of 10 or less repeats. Plasmids containing intermediate arrays are sequence confirmed. In the second GG reaction, these intermediate arrays and final repeats are further ligated together into a pcDNA3.1(-)-based backbone vector (3) using BsmBI restriction sites to replace a *lacZ* gene stuffer fragment for blue/white screening. The backbone vector was constructed by incorporating a Kozak sequence, a triple FLAG epitope tag, and a previously described TALEN framework (24) into the

pcDNA3.1(-) vector using NheI and AflII restriction sites (backbone vector kindly provided by Dr. Matthew H Porteus, Stanford University).

A detailed protocol of Golden Gate TALEN assembly published by Voytas Lab can be found on www.addgene.org. This assembly method was programmed by us into a protocol which can be processed by a Beckman Coulter Biomek 3000 liquid-handling robot. This allows high-throughput, error-free assembly of TALENs. After optimizing the protocol, >90% of colonies generated from each GG reaction are positive and contain the correct sequences. Typically a batch of TALEN-encoding plasmids can be constructed within two weeks, including the time for sequencing intermediate arrays and final constructs. Complete sequences of all TALEN plasmids can be generated using the TAL plasmid assembly website (bit.ly/assembleTALsequences).

2.2 Measurement of nuclease-mediated DNA modification

Measurement of NHEJ-mediated gene mutation

T7E1 assay was used to quantify mutation resulted from nuclease cleavage and subsequent NHEJ repair in the genome. This assay determines the percentage of mutated alleles in the population of cells treated with nucleases, which is an indication of the gene-modifying activity of the nuclease tested (Figure 6). This enzyme-based assay can be done in less than one day without expensive equipment, and do not require co-deliver of constructs other than nucleases into cells, so is widely used to estimate the activity of nucleases at endogenous genes in mammalian cells.

HEK293T cells were transfected with plasmids encoding both TALENs that form a pair, and pEGFP plasmid. Cells will be harvested 72 h after transfection and analyzed with an Accuri C6 flow cytometer to quantify GFP fluorescence, as a measurement of

transfection efficiency. Cell pellets were then collected and genomic DNAs were extracted. Transfected TALENs generated DSBs at the target sites or off-target sites in the genome. The DSBs were repaired mainly through the error-prone NHEJ pathway, generating insertions and deletions at the cleaved site. To perform the T7E1 assay, predicted TALEN cleavage sites were PCR-amplified from the genomic DNAs. DNA strands in the PCR product were denatured at a high temperature (95 °C) and allowed to cool down slowly, during which process DNA strands were randomly annealed. Re-annealed products contain mismatch "bubbles" due to the presence of mutations (Figure 6). T7E1 enzyme was then added to the re-annealed products, and cleaved at the mismatch "bubbles". The cleaved bands were visualized on 2% agarose gels, and used to quantify the percentage of insertions and deletions (% indels) following instructions(58). The gel images were analyzed using ImageJ, and the following equation: % indels = 100x (1 – (1-fraction cleaved)^0.5), assuming the mutations resulted from NHEJ will all be different from each other.

The negative control cells were transfected with a nonrelated filler plasmid (pUC19) and pEGFP plasmid. To ensure that the cleavage bands are due to NHEJ followed by TALEN-induced DSBs, and not due to polymorphism, we amplified each target locus in the genome using genomic DNA from negative control cells as template. Negative PCR products were also subjected to the T7E1 assay, and compared with the treated samples as shown in Figure 6.

T7E1 assay has been used routinely to measure endogenous gene-targeting efficiency of engineered nucleases(2,31). This assay can sometimes under-estimate mutation levels if a substantial portion of mutations have the same sequence (less chances

of forming hetero-duplexes with strands sharing the same mutation). Due to the limited sensitivity of DNA stain on gels, T7E1 assay has a detection limit of ~1%. To detect a very low percentage of mutation, such as off-target effect, a deep sequencing method is needed.

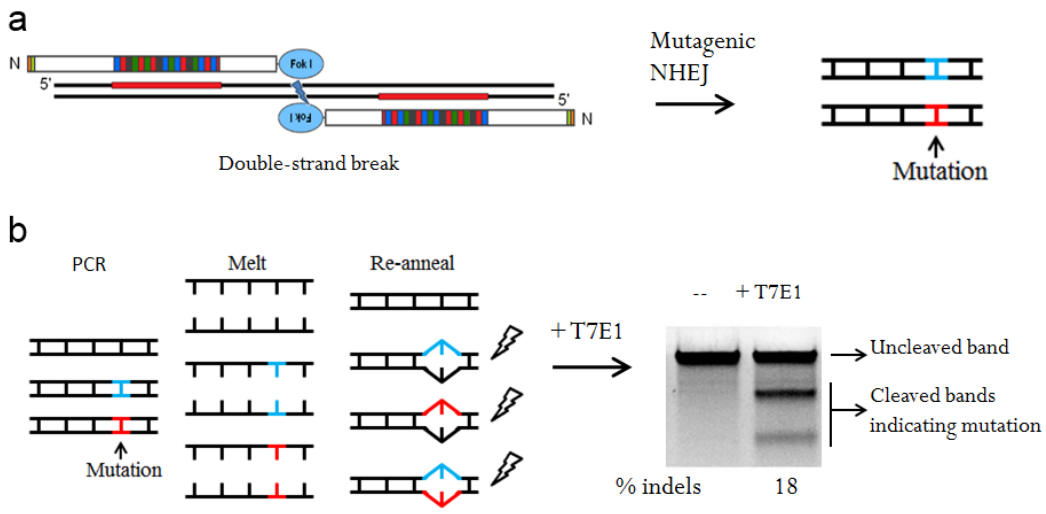


Figure 6: Schematic for T7E1 mutation detection assay.
(a) TALENs cleave genomic DNA and result in mutations mediated by NHEJ. (b) T7E1 assay measuring the percentages of mutated alleles in treated cells. T7E1, T7 endonuclease I enzyme. % indels, percentages of insertions and deletions. --, negative control sample treated without TALENs.

Measurement of HDR-mediated gene editing

HDR allows site-specific, user-designed gene editing / addition if providing cells with targeting DNA constructs (donor) containing a user-defined gene fragment (e.g.

transgene or edited sequence) flanked by substantial amount of homologous sequences to the target locus. Spontaneous DSBs randomly occurring at the target locus can lead to HDR using the homologous donor DNA provided. After repair, the chromosome will contain a copy of the genetic information we include between the two regions of homology. However, the frequency of gene editing for the spontaneous HDR is typically from 10^{-5} to 10^{-6} before selection (59). Generating a DSB at the target locus can dramatically increase the frequency of HDR (60). Site-specific engineered nucleases, ZFNs, TALENs, and CRISPRs, have successfully accelerated targeted gene engineering (Figure 7). Following the nuclease cleavage at the target site, the frequency of HDR can reach the range of 10^{-1} without selective pressure. To determine the frequency of successful gene editing, two methods of measurement are described below.

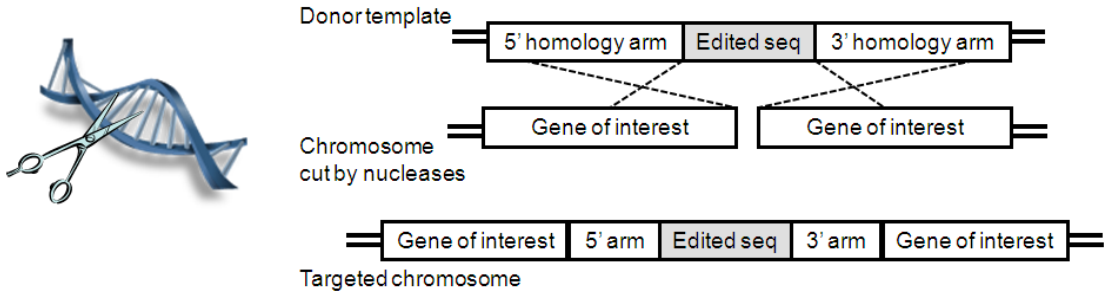


Figure 7: HDR-mediated targeted gene editing using a donor template with homology. Edited sequence (Edited seq) can be a new transgene to be integrated into genome, or specific nucleotide changes at an endogenous gene, such as correcting the "sickle" cell mutation.

Measurement of HDR using GFP signal

To test the ability of nucleases to stimulate HDR, we can include a GFP gene with a constitutive promoter between the two homology arms shown in Figure 7. Upon successful homologous recombination, this GFP cassette will be integrated into the genomic locus, and be stably maintained through cell passages. If not integrated into the genome, the GFP-containing donor plasmid will be diluted over time, resulting in a decrease of GFP signal. After continuous culture of transfected cells for several weeks, non-integrated GFP signal will be lost, the levels of stable GFP signal following the signal decrease thus indicate the successful gene targeting through HDR (Figure 8). When TALENs are co-transfected with donor plasmids, the percentages of stably integrated GFP will be substantially increased compared to cells transfected with donor plasmids only, provided that TALENs are active and generate DSBs efficiently (Figure 8).

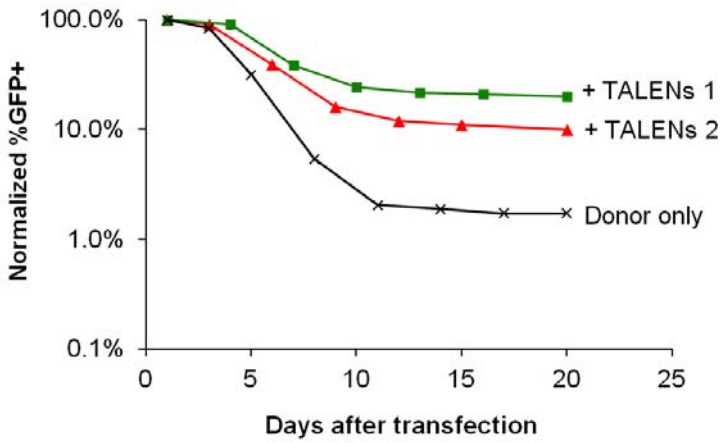


Figure 8: GFP signal indicating the HDR-mediated gene targeting efficiency at an endogenous locus.

This method is easy to perform and can detect a small percentage of gene targeting down to 0.1% if sorting through a large number of cells using flowcytometry. However, this method is time-consuming and often takes several weeks to perform. Moreover, if the nucleases have substantial activity at off-target sites, the GFP cassette can be stably integrated at the off-target sites, resulting in an overestimation of gene targeting at the intended target.

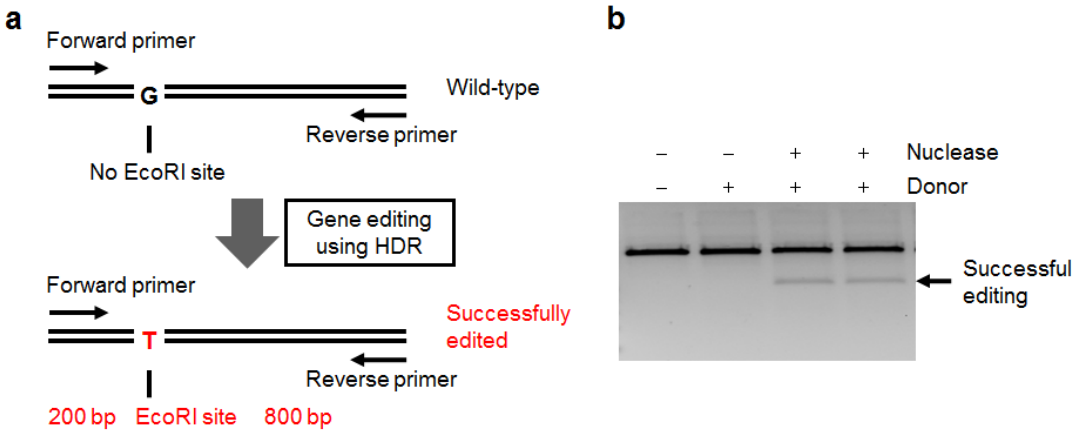


Figure 9: RFLP assay by introducing silent mutation(s) to form a new restriction site.
(a) Gene editing using HDR introduce specific mutations of nucleotides, resulting in a new restriction site at the genomic locus. (b) example of RFLP assay visualized by an agarose gel.

Measurement of HDR using restriction fragment length polymorphism (RFLP) assay

To measure the frequency of HDR repair occurred at a genomic locus without introducing a reporter gene, we can use the restriction fragment length polymorphism (RFLP) assay. By introducing a few silent mutations in the donor template that do not alter the amino-acid sequence of a protein-coding gene, the repaired sequence can be differentiated from wild-type or randomly mutated sequences. The silent mutations can often be designed so that the resulting sequence contains a new restriction site. After nuclease-mediated gene editing, the genomic locus edited to contain this restriction site can be PCR amplified. PCR products can then be digested by this restriction enzyme, so the amount of cleaved bands indicates the percentages of alleles successfully edited (Figure 9).

RFLP assay can be performed at day three to day five after transfection, so the results of HDR can be visualized quickly and conveniently using a restriction digest. However, due to the sensitivity of DNA stain on an agarose gel, this assay has a detection limit of ~1%. The frequency of successful gene targeting through homologous recombination in human cells rarely reaches 10%, especially in stem cells, where often <1% of HDR is observed (unpublished data). In this case, a deep sequencing method is needed instead of RFLP assay.

Sequencing analysis of gene modification

Regular Sanger sequencing and high-throughput next-generation sequencing (NGS) can also be used to visualize and quantify the genomic changes. Compared to the assays above where enzymes and flowcytometer are used, sequencing methods allow visualization of the actual sequence changes introduced by gene engineering. For

example, an "indel spectrum" can be seen by sequencing, which shows the occurrence of sequencing reads with different numbers of inserted bases or deleted bases (61). Furthermore, deep sequencing methods using platforms developed by Illumina and Pacific Biosciences are able to detect rare gene modification events around 0.1%. Below we describe two methods of sequence analysis for gene modification.

TOPO-cloning and Sanger sequencing

After PCR amplification of genomic regions containing predicted nuclease cut site, PCR products can be ligated into plasmid vectors using commercially available kits, such as the TOPO TA Cloning Kit (Life Technologies). The ligated library is transformed into competent cells, and plasmids from individual colonies are extracted and sequenced by regular Sanger method. Normally 20 to 100 clones can be sequenced for a genomic locus, giving a sensitivity of 5% to 1%. Sequences of individual colonies are aligned to the wild-type sequence. Any sequence containing inserted or deleted bases is called an "indel" (insertion or deletion) read, and the number of indel reads divided by the total number of sequences (reads) is named "% indels" (Figure 10). The "% indels" is widely used as an indication of nuclease activity. Purification of plasmids and sequencing a large number of plasmids can be time-consuming and costly. If a large number of cut sites need to be analyzed, high-throughput next-generation sequencing is preferred.

Next-generation sequencing / deep sequencing

If higher sensitivity is desired, next-generation sequencing (NGS) may be used to generate millions of reads per run. Using spatial separation methods, NGS methods can simultaneous sequence millions of reads in a single run, allowing cost-effective high-

throughput DNA sequencing. DNAs from hundreds of samples can be pooled together into a single tube and subjected to parallel sequencing.

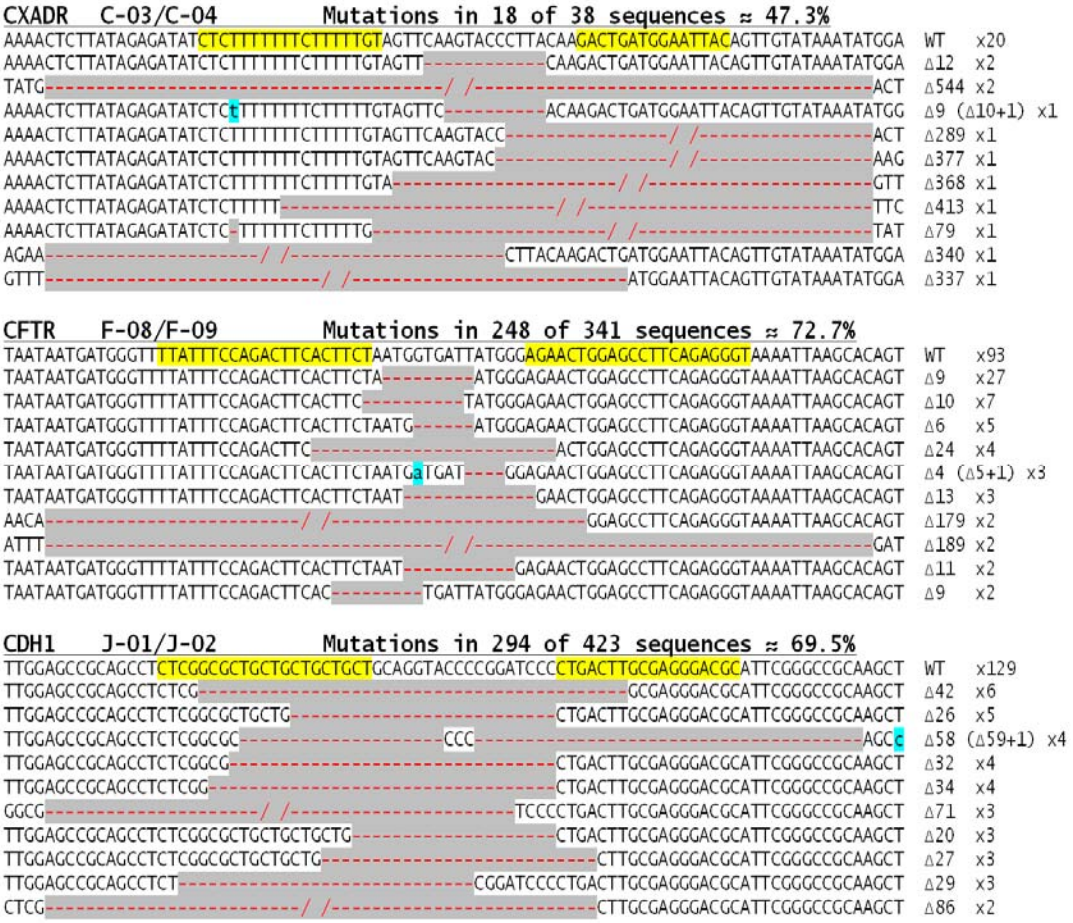


Figure 10: Sequencing results of the target loci aligned with ten most abundant mutated sequences. The target gene name, TALEN index and percentage of modified alleles (% indels) are indicated above each set of results. TALEN target sites are highlighted in yellow on the endogenous gene sequence, marked as wild type "WT". Deletions are shown as red dashes with a grey background. Insertions are highlighted in blue. The numbers to the right of each sequence read indicate the bases inserted (+) or deleted (Δ) and further right, the number of times each sequence was found.

Samples containing different sequences can be differentiated easily by aligning to their references respectively. If several samples share the same sequence (e.g. PCR from the same genomic locus), 8-bp barcode sequences were used to label these samples. Following sequencing, millions of reads were grouped into different barcodes, and were aligned to their reference sequences separately for each barcode. Within a group of reads corresponding to one specific sample identified by its barcode and sequence, we then counted the reads contain inserted bases or deleted bases compared to the reference sequence, and calculated the "% indels" as describe in the section above.

When we pool 200 genomic loci into a run, each locus is able to achieve around 10,000 reads, so theoretically a sensitivity of 0.01% mutation can be detected. However, background noise associated with the sample preparation steps and the sequencing run may obscure very small percentages of mutation. To characterize background noise, we included a negative control PCR amplifying a mock-treated cell sample with every PCR we performed. By comparing the sequencing results from the nuclease-treated samples and mock-treated samples, genetic changes associated with nucleases can be differentiated from background sequence changes possibly due to DNA polymorphism, PCR error, or sequencing error.

Background noise can be reduced by specifying a narrower window when counting insertions and deletions. It is often suitable to assume that only insertions and deletions located around the predicted nuclease cut site are caused by nuclease cleavage. In a sequence read of several hundred bases, we could ignore the indels outside a window of 10 bp before and after the cut site. This ±10 bp window allows us to filter out the

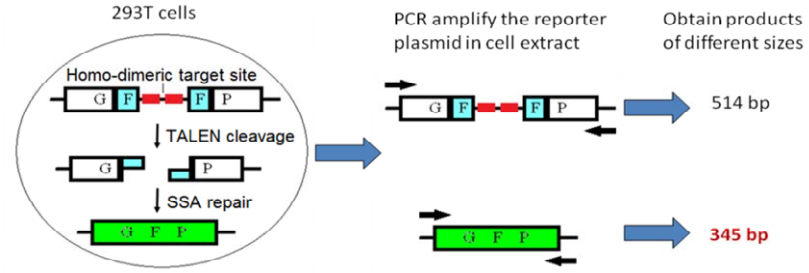
majority of irrelevant sequence changes. This window can be enlarged or narrowed depending on the levels of noise and the property of nucleases.

Measurement of nuclease-mediated DNA modification at a target plasmid

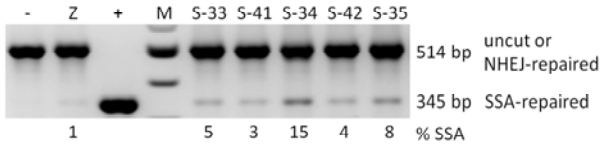
Nuclease cleavage at an endogenous gene can be affected by genomic context, including epigenetic factors, competing transcription factor binding sites and secondary structures. To bypass the effect of genomic context and focus on the intrinsic activity of nucleases, we also measure nuclease activity in HEK293T cells using a modified single-strand annealing (SSA) assay, which test the results of nuclease cutting at a target plasmid containing nuclease target sites. Nuclease-encoding plasmids and the corresponding target plasmids were co-transfected into HEK293T cells, and the percentages of plasmids repaired by the SSA pathway were quantified.

SSA assay measures the repaired plasmids following nuclease cleavage at the plasmid and SSA repair mechanism. SSA pathway is a type of homologous recombination that repairs a DSB between two adjacent repeat sequences in the same strand (Figure 11)(62). When the DSB is repaired by the SSA pathway, the sequence between the two repeats and one of the repeats are removed.

a



b



c

GCCCGAAGGCTACGTCCAGGAGCGCACAATTTTTTTCAAGGATGATGGAAACTACAAGTAAGGCGCGACCATCT

AscI EcoRI

TCTTCAAGGACGACGGCGCGCCTAGCAACCTCAAACAGACACCATGGTGCACCTATCGATGAATTCTTAAGAGG

SbfI

TGCACCATGGTGTCTGTTTGAGGTTGCTACCTGCAGGCAGGAGCGCACAATTTTTTTCAAGGATGATGGAAACT

ACAAGACCCGCGCCGAGGTGA

d

Expected sequence of SSA-repaired segment

GCCCGAAGGCTACGTCCAGGAGCGCACAATTTTTTTCAAGGATGATGGAAACTACAAGACCCGCGCCGAGGTGA

TOPO-cloned SSA-repaired sequences

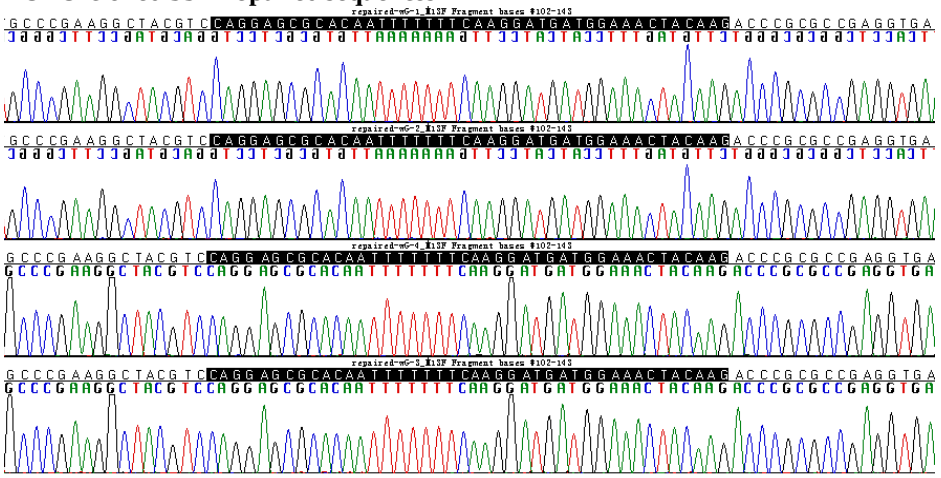


Figure 11: Modified SSA assay to test nuclease activity cutting a target plasmid.
(a) Schematic of the modified SSA assay. (b) Samples from a representative SSA assay used to quantitate the PCR products amplified from HEK293T cells. (c) An example of a homo-dimeric target plasmid segment. The target site in the sense strand and reverse complement of the antisense target site are highlighted in yellow. The GFP-ZFN target site is shaded in grey. The 42-bp repeats flanking the target site are shaded in black. The EcoRI site in the spacer and AscI and SbfI restriction sites are underlined. (d) Sequencing chromatograms of four TOPO-clones from gel-isolated SSA-repaired band, each matching the expected SSA-repaired sequence. The 42-bp repeat is shaded in black.

We made SSA target plasmids by cloning nuclease target sequences between two repeats to facilitate SSA assay (3). An SSA backbone plasmid should contain two repeat sequences separated by one or more stop codons, an optional control target site, and two unique restriction sites that are each present once plasmid. Our SSA backbone plasmid contains an EGFP gene, interrupted after 327 bp with a stop codon, the target site for a pair of GFP-targeted ZFNs(63), an AscI, and an SbfI cloning sites (plasmid kindly provided by Dr. Matthew Porteus, Stanford University). The downstream portion of the EGFP gene includes a 42-bp region repeating the sequence of the EGFP gene before the stop codon (Figure 11). The ZFN target site can be used as an activity control when the SSA plasmid is co-transfected with the GFP-ZFNs.

The two restriction sites, AscI and SbfI, are used to insert TALEN target sites that includes a left binding site, followed by a spacer and a right binding site (Figure 12). Three pairs of oligonucleotides that contain the left TALEN half-site, a spacer with an EcoRI site, and the right TALEN half-site are ligated into the vector. The ends of each oligonucleotide pair are compatible with its adjacent oligonucleotide pair(s) or the backbone vector. Various spacers from 11 bp to 30 bp with arbitrary sequences can be used in the ligation. Oligonucleotide pairs containing TALEN half-sites (red lines with arrows) can be used in assembling both homo-dimeric TALEN targets and hetero-dimeric TALEN targets. If the left binding sequence on the sense strand and the right binding

sequence on the anti-sense strand are the same, the target plasmids are “homo-dimeric”, otherwise called “hetero-dimeric”.

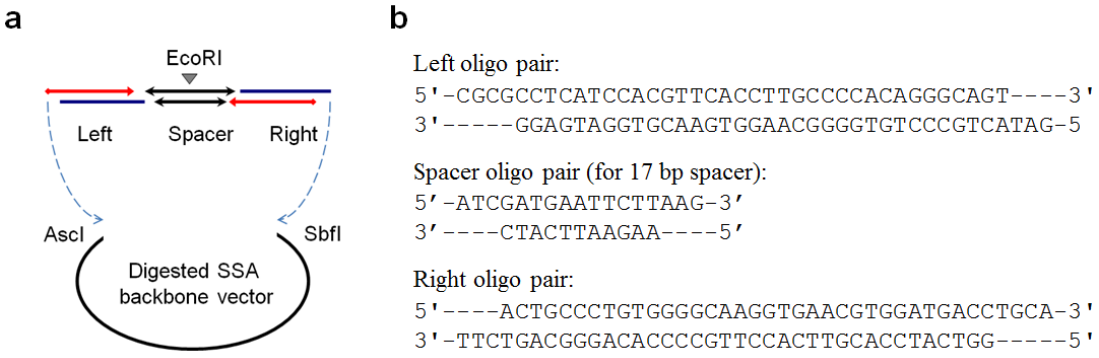


Figure 12: Assembly of TALEN target plasmid.
(a) Schematic of target plasmid assembly. (b) Examples of oligonucleotide pairs for making the target plasmid are shown with each sense and anti-sense oligonucleotide paired. The overhangs of these oligonucleotide pairs were designed to be complementary for annealing and ligation of the three oligo pairs into the backbone vector.

Plasmids encoding a TALEN monomer were co-transfected with the corresponding homo-dimeric target plasmid carrying the pair of TALEN binding half-sites separated by a 17-bp spacer. If the plasmid is cleaved by TALEN homo-dimers and repaired by the SSA pathway, the fragment is smaller from the loss of the target sites and bases between the 42-bp repeat sequences. Following PCR amplification of a region flanking the target site using primers specified in (3), the PCR products were analyzed by agarose gels which separated the 345-bp PCR fragments amplified from SSA-repaired

reporter plasmids and the 514-bp PCR fragments amplified from uncut or NEHJ repaired plasmids (Figure 11b) (3). The percentage of the SSA-repaired products relative to the total PCR products was determined using ImageJ. The negative controls cells were transfected with an empty TALEN backbone and an SSA reporter plasmid. The positive control cells were transfected with a pair of GFP-ZFNs(63) and an SSA reporter plasmid. Another control sample was transfected with pEGFP plasmid and an empty TALEN backbone to estimate the transfection efficiency by quantifying the percentages of EGFP-positive cells. The 345-bp band, as shown in Figure 11b, was gel-isolated, cloned into pCR4-TOPO vector (Life Technologies) and sequenced using T3 primer. The sequencing results of the band confirmed correct SSA repair of the target plasmid (Figure 11d).

CHAPTER 3: COMPARISON OF TALENS CONSTRUCTED WITH DIFFERENT GUANINE-TARGETING RVDS

The modular DNA binding of TALENs is enabled by the one-to-one interaction between repeat-variable di-residues (RVDs) and nucleotides. Specific recognition of guanine has been controversial because three RVDs, NN (Asn-Asn), NK (Asn-Lys) and NH (Asn-His), have been shown to target guanine with varied affinity and specificity, but the overall cleavage efficiency and genomic off-target effect of TALENs constructed with these three RVDs have yet to be well characterized. Here we constructed NK-, NH-, and NN-TALEN pairs for nine target sequences in the human genome. We quantified and compared the on- and off-target effect of TALENs with three different G-targeting RVDs at human genomic loci bearing one to 14 mismatches. NH- and NN-TALENs showed higher gene modification frequencies resulting from non-homologous end-joining (NHEJ) compared to NK-TALENs, whereas NN-TALENs stimulated highest frequencies of gene targeting by homologous recombination (HR). NK-TALENs were able to mitigate off-target cleavage at genomic sites harboring ≥3 mismatches, indicating an overall higher specificity compared to NH- and NN-TALENs. Our results will help researchers choose suitable G-targeting RVDs to construct TALENs where a balance between high on-target activity and minimal off-target activity is most critical, such as in clinical applications.

3.1 Introduction

The ease of TAL Effector Nucleases (TALENs) design has led to their effective use in genomic editing in a number of biological applications. The DNA-binding domain of TALENs contains a tandem array of 33-35 amino-acid repeats(15,26,27), with each repeat specifying one nucleotide through its 12th and 13th amino acids, the repeat-variable di-residues (RVDs)(20,22). TALENs are engineered to bind novel sequences by assembling the RVDs to correspond to the target sequence. There is a straightforward code linking RVDs in any context to their targeted nucleotide. RVDs NI (Asn-Ile), HD(His-Asp) and NG (Asn-Glu) specifically bind to nucleotides A, C, and T, respectfully(64). The exception is the nucleotides guanine (G), which has been targeted by engineered TAL effectors (TALEs) or TALENs using multiple RVDs, including NN (Asn-Asn), NK (Asn-Lys) and NH (Asn-His)(1,24,30,49,50). These RVDs bind with different affinity and specificity. The NN RVD binds to either G or A with relatively similar efficiency; the NK RVD has higher specificity but lower affinity in binding to G compared with NN. Studies of gene-regulating TALEs hinted that the NH RVD might combine these qualities and have high affinity and high specificity, though TALENs containing NH have not been characterized.

We have compared nine sets of TALENs that only differ in their G-targeting RVDs and have been constructed using NK, NH, or NN (KNH TALENs). Each of these sets of KNH TALENs was transfected and tested in parallel for their ability to cleave their unique target sequence. To identify the RVDs that confer the best specificity, off-target cleavage activities of KNH TALENs at genomic sites sharing similarities to the target loci were quantified by T7 endonuclease I (T7E1) mutation detection assay, Sanger sequencing, and Single Molecule Real Time (SMRT) sequencing. KNH TALENs were

further tested for their ability to stimulate gene targeting by homologous recombination. The activity and specificity results can be used to select the appropriate G-specifying RVDs to optimize cleavage activity and target specificity for ones genome engineering application.

3.2 Results

Comparing the activity of KNH TALENs at their intended targets

Each TALEN pair was constructed with one of the different G-targeting RVDs. These TALENs containing one of the most common three types of G-targeting RVDs (NK, NH, and NN) were tested in parallel to investigate in whether the change in activity resulted from switching G-targeting RVDs is universal (Figure 13). Besides the G-targeting RVDs, we use NI to target A, HD for C, and NG for T. We thus built three versions of TALEN pairs targeting the same genomic sequence, but constructed with three different G-targeting RVDs, NK, NH, and NN, respectively. Nine sets of target sites were targeted by the KNH TALENs, respectively. The target sequences include loci from six genes: *HBB*, *CXADR*, *CFTR*, *AAVS1*, *CCR5*, and *ATF4*. Target loci in the *HBB* gene were selected around the sickle cell mutation, and other target sites were chosen using the SAPTA program (3) from the highest ranking sites within these genes. Four combinations of target sequences were tested for the *HBB* gene: S-02/S-05, S-02/S-12, S-02/R-04, and S-116/S-120. These loci also contain various numbers of G in the left and right TALEN binding sites, ranging from zero to six, constituting 0% to 35% of the targeting half-sites (Figure 14).

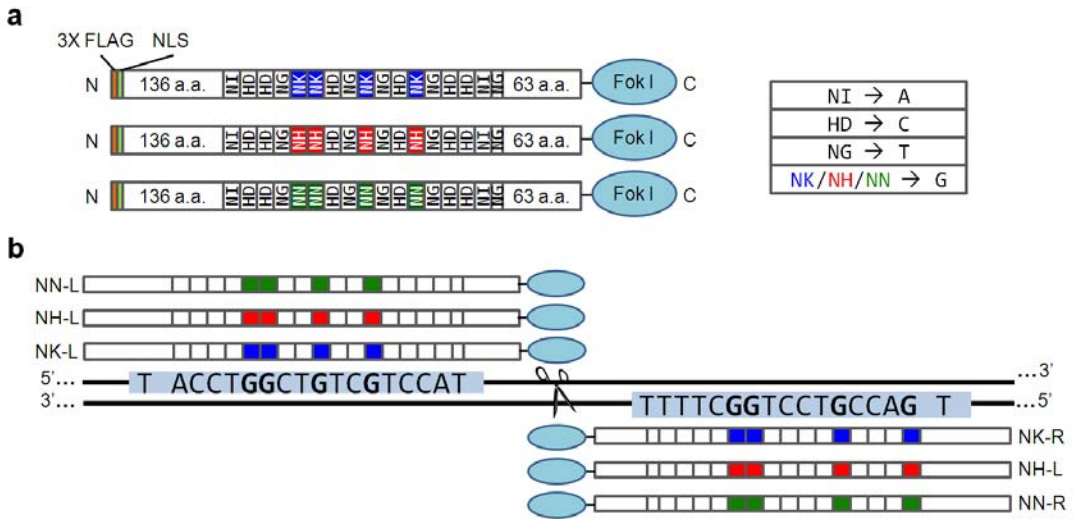


Figure 13: Schematic of TALENs constructed with different G-targeting RVDs.
(a) A set of three TALENs with the G-targeting RVDs switched to NK (blue), NH (red), and NN (green), respectively. The TALEN scaffold uses the +136/+63 architecture, with a triple FLAG tag (3X FLAG) and a nuclear localization signal (NLS)(3). Corresponding RVDs and their target nucleotides are listed on the right. (b) Three TALEN pairs with different G-targeting RVDs recognizing the same target site. Each pair of TALENs contains a single type of G-targeting RVD. NK-L, NH-L, and NN-L are left TALENs containing NK, NH, and NN, respectively. NK-R, NH-R, and NN-R are right TALENs containing NK, NH, and NN, respectively.

TALENs containing the same G-targeting RVD were paired and transfected into HEK293T cells, and the gene mutagenesis resulted from non-homologous end-joining (NHEJ) DNA repair following TALEN cleavage was quantified using a T7 endonuclease I (T7E1) assay (Figure 15). The percentages of insertions and deletions (% indels) at the intended target sites indicate the efficiency of TALEN cleavage (Figure 16).

Figure 14: TALEN target sites in six different genes.
Target sequences of nine target sites in six genes. The 5'-flanking T is shown in each nuclease target sequence. The G's are highlighted in orange.

We found that at their intended target sites, NH- and NN-TALENs show higher or comparable NHEJ-mediated mutation rates compared to NK-TALENs (Figure 16). The mutation rates resulted from NH-TALENs are comparable to that of NN-TALENs in most cases. Since the SAPTA program was optimized for the design of NK-TALENs, NK-TALENs designed by SAPTA (TALENs targeted to genes other than *HBB*) result in

mutation rates close to NH- and NN-TALENs. On the other hand, NK-TALENs targeted to *HBB*, which were not designed by SAPTA, sometimes result in less than half of the mutation rates from NH and NN. To determine if the different activities of KNH TALENs are dose-dependent, we tested activities of S-02/S-05 TALENs, but did not observe considerable difference at lower dosages (Figure 17). As expected, if there are fewer numbers of G's in the target sites, for example, in S-116/S-120, C-03/C-04, and F-08/F-09, the difference among NK-, NH-, and NN-TALENs is small.

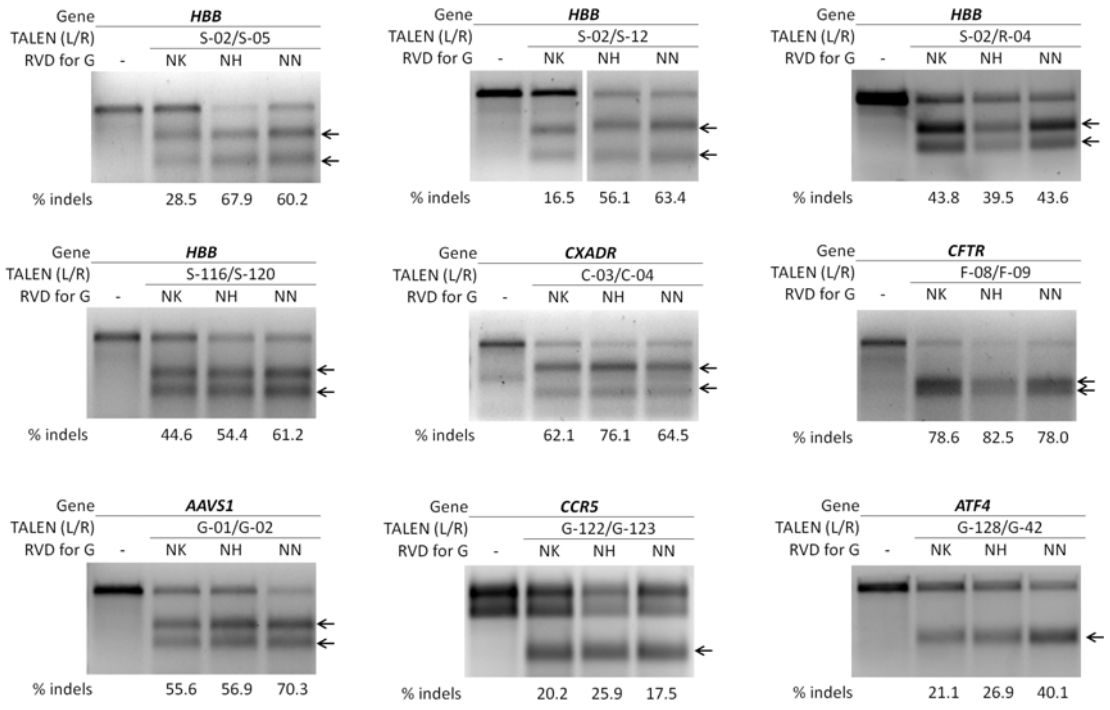


Figure 15: T7E1 assay of KNH TALENs at their intended target sites.
"-" denotes samples treated with an empty TALEN backbone. Numbers below each lane show the average percentage of modified alleles (n=3). Arrows indicate specific T7E1 cleavage products. Note that some cleavage products were too close in size to be separated.

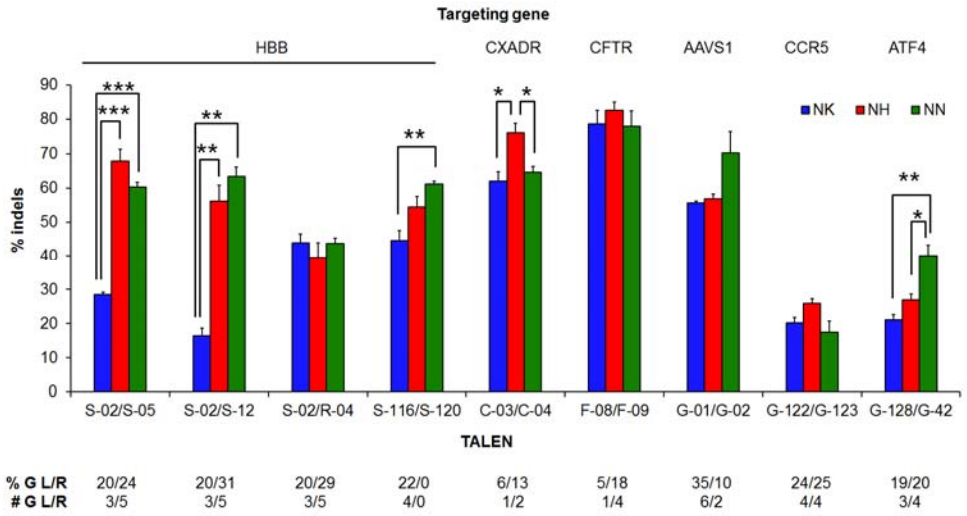


Figure 16: NHEJ-mediated mutation (% indels) of KNH TALENs at their intended target sites. NK (blue), NH (red), and NN (green) activity at target sites shown in (a) were measured in HEK293T cells. The percentage of G (%G) and number of G (#G) in left and right (L/R) target half-sites are indicated below each set of columns. Error bar, s.e.m. (n=3). Asterisks indicate P-values from a two-tailed independent two-sample t-test. *P < 0.05, **P < 0.01, ***P < 0.001.

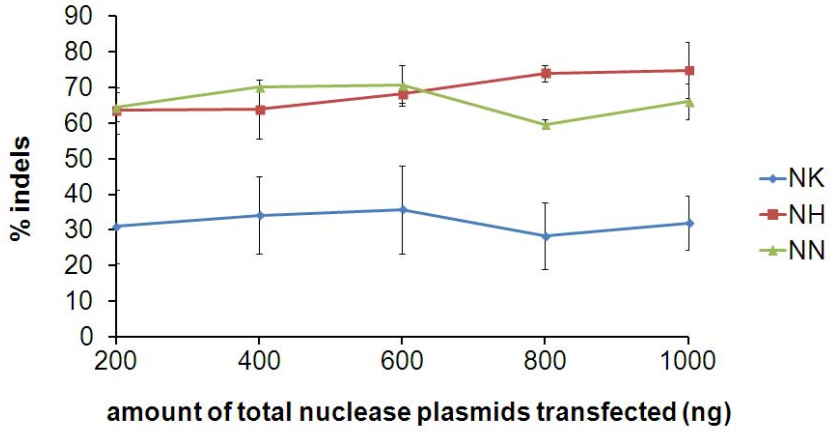


Figure 17: Activity of KNH TALENs with different doses of nuclease plasmids. Different total amount of nuclease plasmids were transfected into HEK293T cells, and T7E1 assay was performed to quantify the mutation frequencies (% indels). Error bars, s.e.m. (n=2).

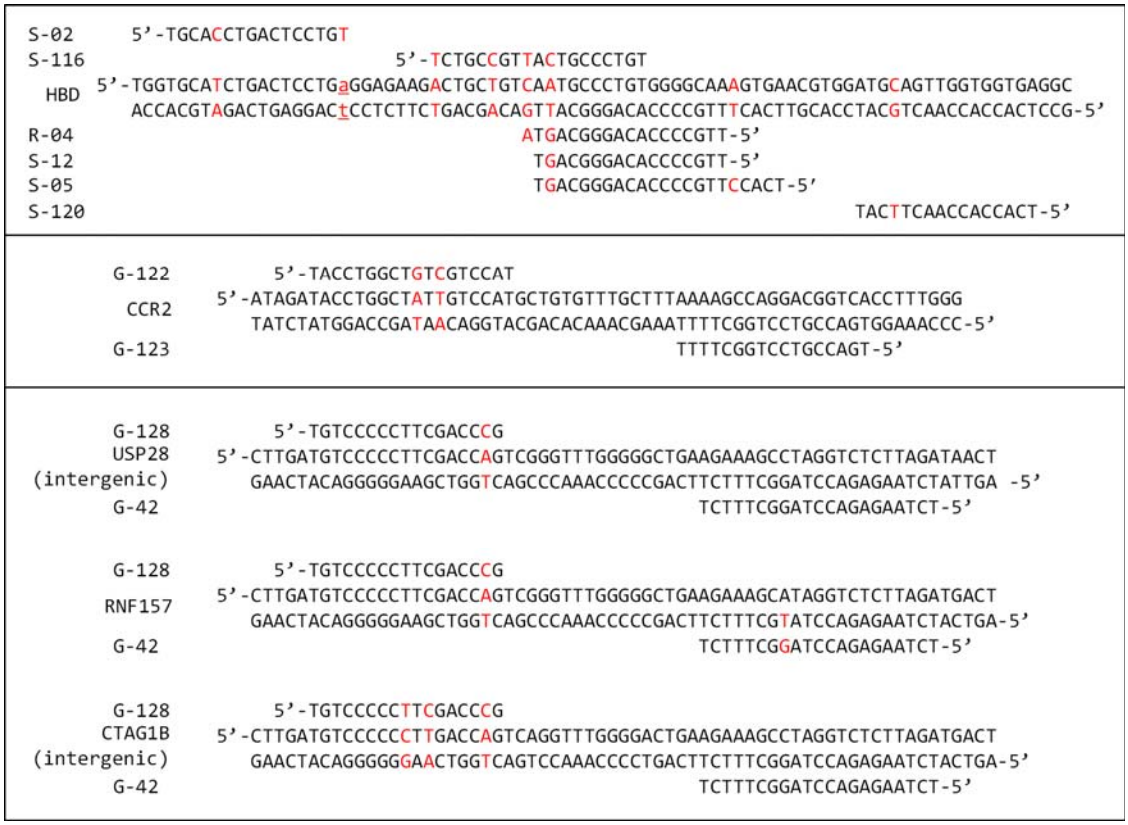


Figure 18: Putative off-target sites of TALEN pairs bearing various mismatches.
Off-target sites corresponding to six pairs of target sequences: S-02/S-05, S-02/S-12, S-02/R-04, S-116/S-120 (top box), G-122/G-123 (middle box), and G-128/G-42 (bottom box). The 5'-flanking T is shown in each target sequence. The mismatched nucleotides between intended targets and off-target sites are highlighted in red. The genes associated with these off-target sites are indicated on the left of the off-target sequences.

Comparing the off-target effect of KNH TALENs

We evaluated off-target effect of TALENs containing different G-targeting RVDs by investigating the putative off-target sites with sequence similarities to the intended

sequences. Previous studies compared the off-target activity of NK-, NH-, and NN-TAL effectors using artificial targets with A's in place of G's(1,50), in order to compare the ability of these TALENs to differentiate between nucleotides A and G. Here we compare the global, general off-target effect in the endogenous genome by investigating in putative sites bearing different mismatches not limited to "G to A mismatch" (Figure 18).

The putative off-target sites we identified contain one to five total mismatches compared to the intended targets. *HBD* gene is most similar to the *HBB* gene, so we used it to examine the off-target effect of the four sets of TALEN pairs targeted to *HBB*. For the G-128/G-42 pair designed to target the *ATF4* gene, we identified several genomic sites with one to three total mismatches using the PROGNOS search tool (65) (Figure 18). These sites allow us to compare how KNH TALENs tolerate different numbers of mismatches at endogenous genomic loci.

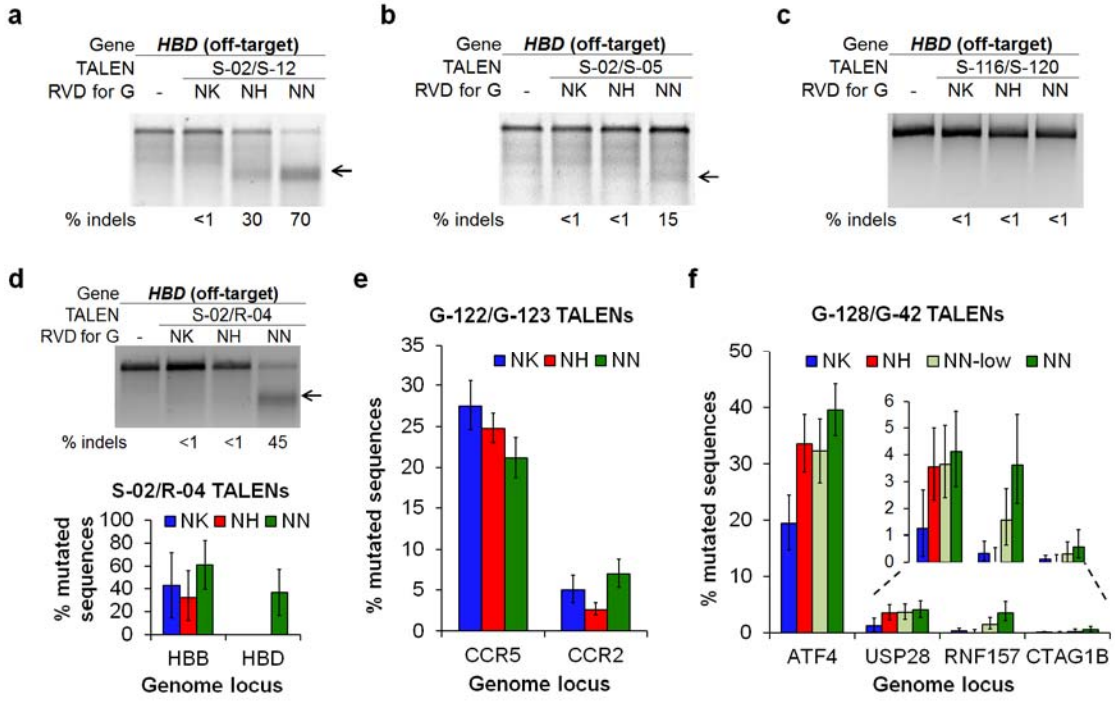


Figure 19: Off-target activities of NK-, NH-, and NN-TALEN pairs bearing various mismatches. (a-c) Off-target % indels of *HBB*-targeting KNH TALENs at the *HBD* gene measured by the T7E1 assay. "-" denotes samples treated with an empty TALEN backbone. Numbers below each lane show the average percentage of modified alleles (n=3). Arrows indicate specific T7E1 cleavage products. (d) Top, T7E1 assay showing the off-target % indels of the *HBB*-targeting KNH TALENs at the similar site in the *HBD* gene. Bottom, percentages of TALEN-induced mutated sequences at the *HBB* on-target site and *HBD* off-target site detected by Sanger sequencing. (e-f) Percentages of mutated sequences of NK- (blue), NH- (red), and NN-TALENs (green) at their intended target sites and off-target sites measured by SMRT sequencing. The on-target activities are the first set of columns from the left. The insert in (f) is a zoom-in view of the three off-target sites. NN-low (light green) in (f) indicates activities associated with a lower dose of plasmids transfected to match the on-target activity of NH-TALENs. Error bar, 95% confidence interval (Wilson method).

We first measured the off-target mutation rates at the *HBD* gene using T7E1 assay in HEK293T cells. Genomic DNAs used to quantify the on-target activity were analyzed again for the off-target effect. The four sets of TALEN pairs intended for *HBB* contain three to five total mismatches at the similar *HBD* gene (Figure 18). T7E1 assay for these

TALENs revealed different levels of off-target mutagenesis at the *HBD* locus (Figure 19a-d). The on- and off-target modification frequencies resulted from S-02/R-04 pairs were confirmed by Sanger sequencing of PCR products cloned into plasmid vectors (Figure 19d). Three out of four NN-TALENs resulted in substantial mutation rates (Figure 19a, b, and d); one NH-TALENs showed considerable off-target cleavage (Figure 19a); none of the four NK-TALENs was found to have off-target effect at the *HBD*. Mismatch of the 5' flanking T (position 0) may be a major contributor to the lack of off-target cleavage from S-116/S-120 TALENs, since 5'-T was speculated to be a potent discriminant between highly similar sites (66). Mismatches at both the N-terminus and C-terminus seem to contribute to the discrimination between similar sites. Compared to S-12, S-05 has one additional mismatch at the N-terminus, while R-04 has one extra mismatch at the C-terminus (Figure 18). When paired with the same left TALEN site S-02, NN- and NH-TALENs of both S-05 and R-04 showed reduced off-target mutagenesis compared to S-12 at the *HBD* locus (Figure 19). The additional mismatch at the N-terminus in S-05 resulted in more prominent decrease of the off-target effect, which is consistent with previous publications indicating that N-terminal repeats have more impact on TALE binding to DNA(67,68).

Additional off-target sites for TALENs targeted to the *CCR5* and *ATF4* genes were analyzed by Single molecule real time (SMRT) deep sequencing for a higher detection sensitivity (Figure 19e, f). To account for the variance in on-target activity, we calculated an "off-target factor" by normalizing the off-target activity with each TALEN pair at each site against the on-target activity of the TALEN pair (Table 1). NK-TALENs displayed overall less off-target effect: among the eight off-target sequences analyzed,

one NK-TALEN pair, three NH-TALEN pairs, and five NN-TALEN pairs have off-target factors >0.1, which indicate considerable off-target activity higher than 10% of the on-target activity. NH-TALENs have low off-target effect similar to NK-TALENs, except for the S-02/S-12 NH-TALENs (Figure 19a), which have significantly higher on-target activity compared to NK-TALENs. With similar on-target mutation rates, the S-02/S-12 NH-TALENs showed much lower off-target cleavage at *HBD* compared to NN-TALENs.

Table 1: Off-target levels of NK-, NH-, and NN-TALENs with different numbers of mismatches.

TALEN index	Off-target site	Nucleotide mismatch			Off-target factor[b]				Detection Method[d]
		Total	Left	Right	NK	NH	NN	NN-low[c]	
S-02/S-05	*HBD*	4	2	2	0	0	0.249		T7E1
S-02/S-12	*HBD*	3	2	1	0	0.534	1.103		T7E1
S-02/R-04	*HBD*	4	2	2	0	0	0.604		Sanger
S-116/S-120	*HBD*	5	4	1	0	0	0		T7E1
G-122/G-123	*CCR2*	2	2	0	0.182	0.107	0.329		SMRT
G-128/G-42	*USP28*[a]	1	1	0	0.064	0.106	0.104	0.113	SMRT
G-128/G-42	*RNF157*	2	1	1	0.017	0	0.091	0.048	SMRT
G-128/G-42	*CTAG1B*[a]	3	3	0	0.005	0	0.014	0.009	SMRT

a. The genes indicated are located closest to these off-target sites in intergenic regions.

b. Ratio of the off-target to on-target modification frequency. Non-detectable off-target activity is marked as a zero. Off-target factors above 0.1 are highlighted by gray background.

c. Samples with a lower dose of plasmids transfected to match the on-target activity of NH-TALENs.

d. Methods used to quantify the gene-modification frequencies include T7E1 assay, Sanger sequencing, and SMRT deep sequencing.

NN-TALENs constantly resulted in similar or higher off-target modification frequencies compared to NK and NH. The higher off-target effect associated with NN-TALENs cannot be explained by its higher on-target activity, since NN-TALENs used for off-target analysis always have similar on-target activity compared to NH- or NK-TALENs. Neither can this higher off-target effect be readily explained by less differentiation between "G" and "A" bases, as only one mismatch among these eight off-target sequences is a "G to A mismatch" (*CCR2* locus). Lower dose of G-128/G-42 NN-TALENs were transfected into HEK293T cells to reduce the on-target activity to match that of NH-TALENs, but the off-target effect from the NN-TALENs could not be eliminated using a lower dose (Figure 19f, Table 1).

We also investigated putative heterodimeric (left and right TALENs) and homodimeric (left paired with left TALENs, or right paired with right TALENs) off-target sites containing five to 14 mismatches to TALENs intended for *CXADR, CFTR, AAVS1* (*PPP1R12C*), and *ATF4*. For these putative sites, no off-target cleavage significantly higher than the mock-transfected sample was identified using SMRT deep sequencing or T7E1 assay (

Table 6, Table 7, and Table 8 in Appendix A, Figure 20).

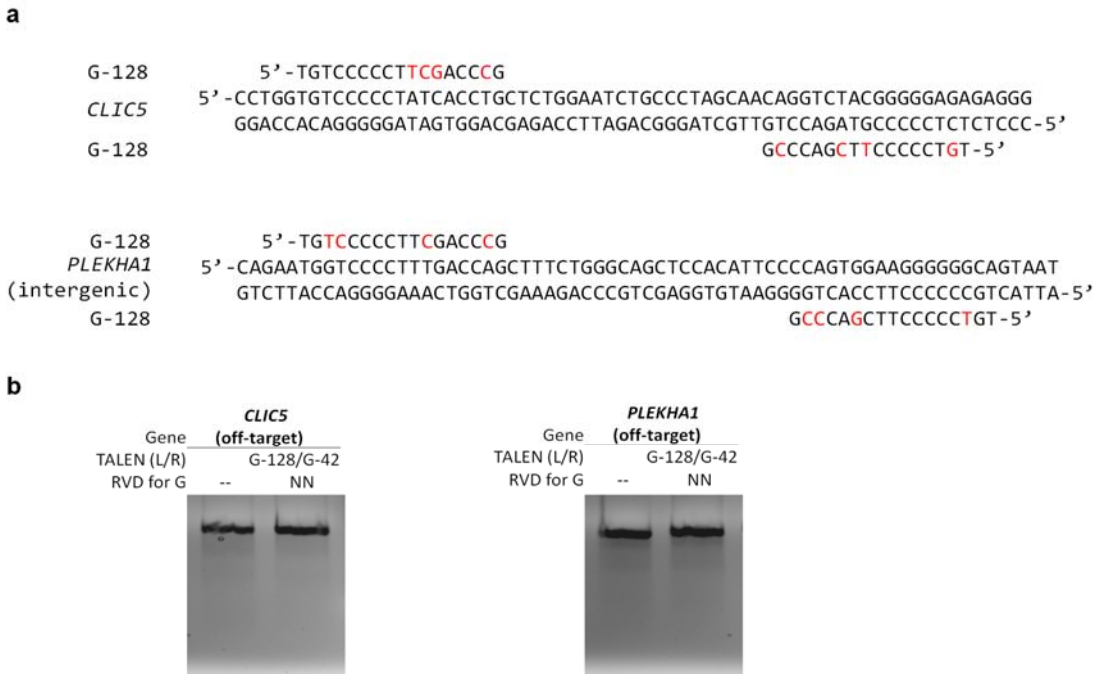


Figure 20: Off-target analysis for homo-dimeric sites of *ATF4*-targeting TALENs.
(a) Homo-dimeric off-target sites for *ATF4*-targeting TALENs. The mismatched nucleotides are highlighted in red. The 5'-flanking T is also shown. (b) T7E1 assay of *ATF4*-targeting NN-TALENs at its homodimeric off-target sites. No measurable activity was observed.

NN-TALENs achieve higher rates of homologous recombination (HR)

We measured gene-targeting frequencies of different TALENs by providing a β-Ubc-GFP donor vector with ~1 kb homology arms 5' and 3' of the TALEN target sites in the *HBB* gene. A Ubc-GFP cassette was included between the 5' and 3' homology arms, thus that GFP will be stably expressed upon successful HR(69). TALEN pairs targeted close to the sickle mutation were co-transfected with the β-Ubc-GFP donor vector into K562 cells, and the levels of HR were determined by quantifying stable integration of Ubc-GFP in cells (Figure 21).

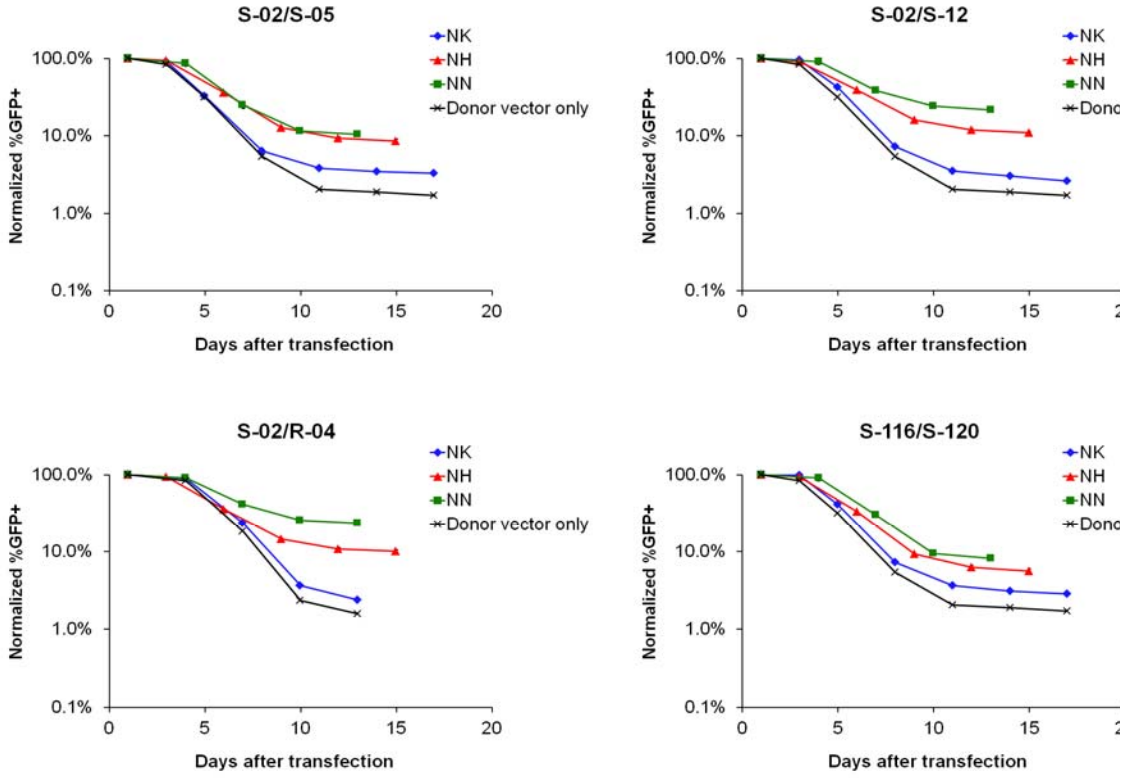


Figure 21: Percentages of GFP-positive cells after transfection of HBB-targeting TALENs and β-Ubc-GFP donor plasmid.
The percentages of GFP+ cells measured by flowcytometry were normalized to the GFP+ cells on day 1 after transfection to account for varying transfection efficiency.

All NN-TALENs achieved higher gene-targeting rates compared to NK- and NH-TALENs for the same target (Figure 22a). NN-TALENs of S-02/S-12 and S-02/R-04 both led to substantially higher gene-targeting efficiency compared to other TALENs. Surprisingly, the gene-targeting rates do not correlate well with the NHEJ-mediated mutagenesis from the same TALENs (Figure 22b). The NH- and NN-TALEN pairs with

NHEJ-mediated mutation rates around 60% (S-02/S-05, S-02/S-12, S-116/S-120) showed drastically different gene-targeting frequencies ranging from 6% to 22%. Even TALENs with the same target sequence and only different by the G-targeting RVD have poor correlation between NHEJ and HR. For example, the NK-, NH-, and NN-TALENs of S-02/R-04 have similar levels of NHEJ-mediated mutagenesis (Figure 16b), but very different gene-targeting rates (Figure 22a). Further study is needed to determine factors that affect the pathway choice between NHEJ and HR.

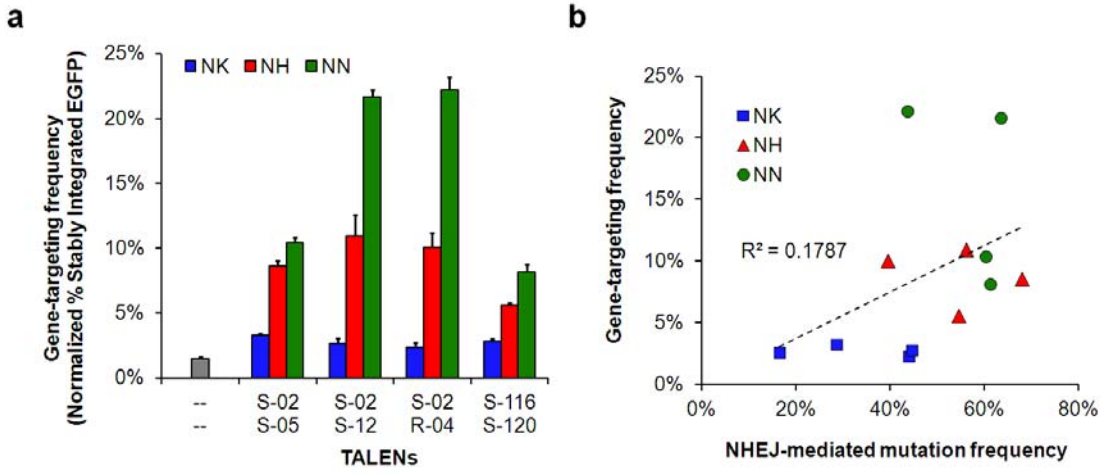


Figure 22: Gene targeting efficiency of NK-, NH-, and NN-TALEN pairs at the endogenous *HBB* locus.
(a) Gene targeting of β-Ubc-GFP to the *HBB* locus in K562 cells using different TALEN pairs. The percentages of stably integrated EGFP were normalized to transfection efficiency. "—" indicates cells transfected with the β-Ubc-GFP targeting vector alone. Error bar, s.e.m. (n=3). (b) Little correlation between gene-targeting frequency and NHEJ-mediated mutation frequency of TALEN pairs. Gene-targeting frequencies from (a) are plotted against the NHEJ-mediated mutation frequencies of TALEN pairs targeted to *HBB*. The R squared value of all TALENs in the plot is also indicated.

applications.

3.3 Discussion

The activity and specificity of TALENs directly correlate with the efficacy and safety of genome engineering applications, especially clinical use of nucleases, and thus need to be well characterized. Finding a highly efficient yet specific RVD to recognize the nucleotide G is a useful approach to improve TALEs and TALENs. The G-recognizing RVD NK employed by TALEs or TALENs was previously shown to target G specifically but with considerably less activity compared to the RVD NN(1,24,50,68). Studies of TALEs revealed that the RVD NH may yield increased specificity for recognizing G compared to NN, while maintain relatively high activity compared to NK (1,50). However, activity and specificity of the NH RVD has not been previously demonstrated in TALENs.

Although NN-TALEN specificity in endogenous genomes was evaluated by a number of studies (25,32,66,70-72), very few provided a comparison with NK-TALENs (30,65). There has not been any publication comparing activity or specificity of TALENs employing all three G-targeting RVDs, NK, NH, and NN. Other than a simple differentiation between G and A, a more complete and general specificity profile of TALENs constructed with different G-targeting RVDs needs to be defined and compared.

Here we compared the on- and off-target mutagenesis activity of nine sets of NK-, NH-, and NN-TALENs, and also evaluated the ability of four sets that recognize the *HBB* loci to trigger gene-targeting through the HR pathway. TALENs constructed with the RVD NH induced NHEJ-mediated gene modification comparable to or higher than those with NK, and the frequencies associated with NH RVD are similar to those with the

highly active NN RVD, except for the NH-TALEN pair targeted to the *ATF4* gene (Figure 16b). Despite the high NHEJ-mediated mutation rates resulted from NH-TALENs, their ability to trigger HR is still not as high as the NN-TALENs (Figure 22a). Interestingly, we found that the indel spectra of KNH TALENs are different, especially that the NH-TALENs seem to induce large deletions more frequently than NK- and NN-TALENs (Figure 23). Five out of the nine NK-TALENs tested have on-target mutation rates significantly lower compared to corresponding NH- or NN-TALENs, confirming that the NK RVD is less efficient in binding to guanine (Figure 16b). However, NK-TALENs still showed overall less off-target effect, and good discrimination against off-target sites with ≥3 total mismatches (Table 1). NH-TALENs are also able to eliminate off-target effect with similar sites containing ≥3 total mismatches (Table 1), except for the S-02/S-12 off-target site at *HBD* where two of the three mismatches are bound by the C-terminus of TALENs, which were previously shown to be less specific(67). NN-TALENs could not sufficiently differentiate similar off-target sites with up to four total mismatches (Table 1). We did not find any off-target cleavage above background signal for any of the TALENs at genomic sites with ≥5 total mismatches (Table 1**,** Figure 20, and

Table 6, Table 7, and Table 8 in Appendix A).

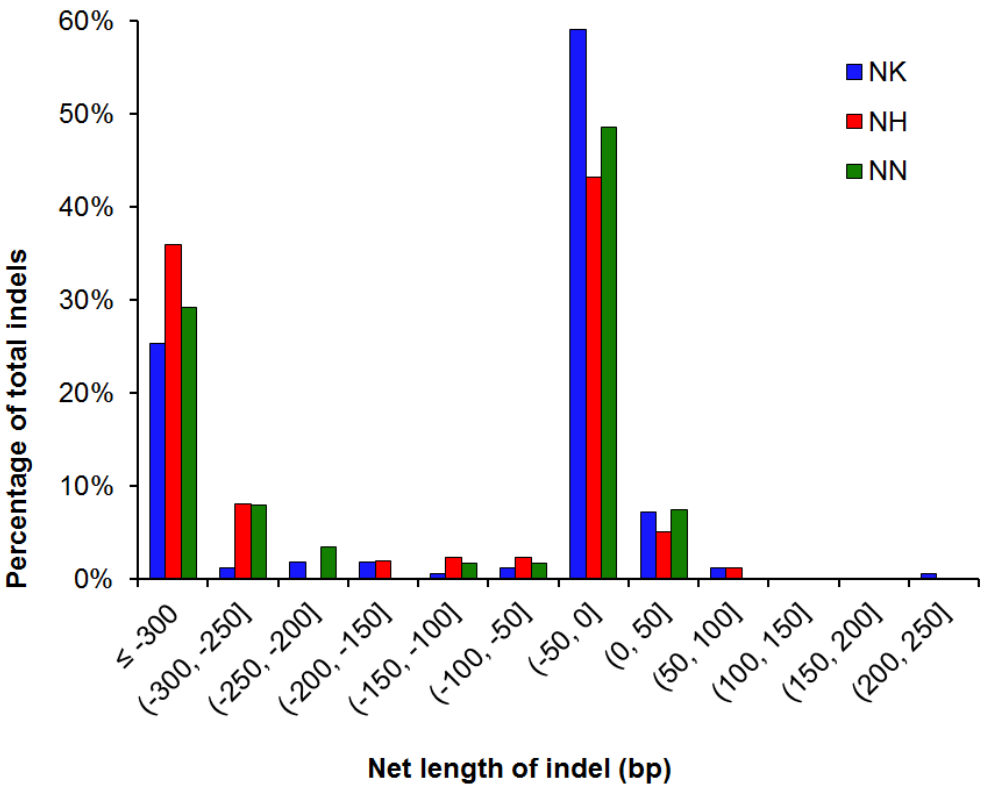


Figure 23: Mutation spectra of KNH TALENs targeted to *CXADR*.
The net lengths of indel (- deleted bp + inserted bp) were analyzed using SMRT sequencing with a 709-bp amplicon in the *CXADR* gene. Only sequencing reads showing insertion and/or deletion were included in this graph. When the net indel length is zero, there is still insertion and deletion in the read, but the number of deleted bases equals the number of inserted base. (a, b] means > a and ≤ b. The y axis (percentage of total indels) is the number of reads at a specific net length divided by the total number of indel reads for that specific pair of TALENs.

In summary, if a high HR rate is desired, NN-TALENs would be the first choice, but the off-target effect should be carefully avoided by selecting target sites with the most similar sites in the genome containing at least five total mismatches. If high specificity is desired, NK-TALENs are preferred, which can differentiate sites with ≥3 total mismatches. Mismatches in the off-target sites located close to the N-termini of TALENs

seemed to be better discriminated by TALENs. NH-TALENs can achieve high levels of NHEJ-mediated mutation, and adequately minimize off-target effect if mismatches are not located near the C-terminal end of TALENs. Our results also showed a disconnection between NHEJ-mediated mutation frequency and HR-mediated gene-targeting frequency. It will be interesting to see what elements contribute to the different pathway choice following TALEN cleavage of DNA.

3.4 Materials and methods

Assembly of TALENs

TALENs were assembled using a hierarchical Golden Gate method (23) with repeat plasmids containing RVDs HD, NI, and NG to recognize C, A, and T, respectively. To recognize G, RVDs NK, NH, or NN were constructed into NK-, NH-, and NN-TALENs, respectively. The process of assembly and sequence confirmation was as describe before (3). TALEN backbone vector was the same as in a previous study (3). Complete amino sequences of TALENs are included in Appendix A.

T7E1 mutation detection assay to determine mutation frequencies at endogenous gens

The NHEJ-mediated genetic mutation resulted from TALEN pairs was quantified by the T7E1 assay as previously described (3). Briefly, 40,000 HEK293T cells were seeded in a well of a 24-well plate, and transfected 24 hrs after seeding with 500 ng of each TALEN plasmid and 10 ng of pEGFP plasmid using FuGene HD (Promega, Madison, WI). 72 h after transfection, cells were harvested and genomic DNAs were extracted using QuickExtract DNA extraction solution (Epicentre, Charlotte, NC).

Genomic loci were PCR-amplified using AccuPrime Taq DNA Polymerase High Fidelity (Life Technologies, Carlsbad, CA) as described before (3). PCR products were purified using Ampure XP (Beckman Coulter, Pasadena, CA) following manufacture's protocol using a Biomek 3000 station (Beckman Coulter, Pasadena, CA). T7E1 assays were performed as described (3). Primers used for this assay are listed in Table 9 (Appendix A).

Sanger sequencing to determine genetic mutations resulted from TALENs

PCR products described in the T7E1 assay above were cloned into plasmid vectors using TOPO TA Cloning Kit for Sequencing (Life Technologies, Carlsbad, CA) or Zero Blunt TOPO PCR Cloning Kit (Life Technologies, Carlsbad, CA), following manufacturer's instructions. Plasmid DNAs were purified and subjected to Sanger sequencing using a M13F primer (5'-TGTAAAACGACGGCCAGT-3'). Plasmid sequences from a PCR product were aligned to the wild-type reference sequence. To calculate the percentages of indels, any plasmids with insertions or deletions compared to the reference sequence were counted, and divided by the total plasmids sequenced.

Identification of putative off-target sites and SMRT sequencing to quantify the mutation rates

Potential off-target sites for the TALENs were chosen from the PROGNOS RVD and Homology rankings to contain a mixture of the top-ranked sites from both algorithms(65). Genomic DNA was extracted from HEK293T cells 3 days post-transfection with TALENs or pUC plasmid. PCR primers designed by PROGNOS (Table 8 in Appendix A) were used to amplify the predicted off-target loci and amplicons were sequenced using the RS SMRT technology (Pacific Biosciences) exactly as previously

described(65). The results were analyzed by custom Perl scripts to identify reads containing evidence of NHEJ(65).

Measuring the frequencies of homologous recombination (HR) stimulated by TALENs

K562 cells were maintained in HyClone RPMI 1640 (GE Healthcare, Pittsburgh, PA) supplemented with 10% bovine growth serum, 100 U/ml penicillin, 100ug/ml streptomycin, and 2mM L-glutamine, filtered sterilized. K562s were transfected via nucleofection (Lonza, Basel, Switzerland) using program T-016 and nucleofection buffer containing 100mM KH_2PO_4, 15mM $NaHCO_3$, 12mM $MgCl_2$, 8mM ATP, 2mM glucose, pH 7.4.

10^6 K562 cells were transfected with 1 μg left TALEN plasmid and 1 μg right TALEN plasmid and 5 μg β-Ubc-GFP donor plasmid (69). Cells were split and analyzed regularly via FACS Accuri C6. Percentages of GFP-positive cells were measured over time until a plateau was reached (around two weeks). The amount of stably integrated GFP cassette was determined using the end-point GFP-positive cell percentages. Stable fluorescence achieved is then normalized to transfection efficiency (percent GFP-positive cells one day after transfection) to account for variation between nucleofections.

CHAPTER 4: DEVELOPMENT OF A NEW DESIGN TOOL FOR IMPROVING TALEN ACTIVITY

Transcription activator-like effector nucleases (TALENs) have become a powerful tool for genome editing due to the simple code linking the amino acid sequences of their DNA binding domains to TALEN nucleotide targets. While the initial TALEN design guidelines are very useful, user-friendly tools defining optimal TALEN designs for robust genome editing need to be developed. Here we evaluated existing guidelines and developed new design guidelines for TALENs based on 205 TALENs tested, and established the Scoring Algorithm for Predicting TALEN Activity (SAPTA) as a new online design tool. For any input gene of interest, SAPTA gives a ranked list of potential TALEN target sites, facilitating the selection of optimal TALEN pairs based on predicted activity. SAPTA-based TALEN designs increased the average intracellular TALEN monomer activity by >3 fold, and resulted in an average endogenous gene modification frequency of 39% for TALENs containing the repeat variable di-residue NK that favors specificity rather than activity. It is expected that SAPTA will become a useful and flexible tool for designing highly active TALENs for genome editing applications. SAPTA can be accessed via the website at http://baolab.bme.gatech.edu/Research/BioinformaticTools/TAL_targeter.html.

4.1 Introduction

Transcription activator-like effectors (TALE) are a family of DNA binding proteins, discovered in the plant pathogen Xanthomonas (19-22). Each DNA-binding

domain of TALE contains a variable number of 33-35 amino-acid repeats that specify the DNA-binding sequence primarily through their 12th and 13th repeat-variable di-residues (RVDs) (19). Each RVD specifies one nucleotide with minimal context dependence (20,22,23). A transcription activator-like effector nuclease (TALEN) targets a specific DNA sequence through designing a set of RVDs that are flanked by modified N- and C-termini (24,28) and linked to a FokI nuclease domain (15,26,27). When a pair of TALENs binds to their specific half-sites with the correct orientation and spacing to allow the nuclease domains to dimerize, the intervening sequence is cleaved. TALENs have been used to edit genomic DNA sequences in a variety of biological systems, including human cells, rats, zebrafish, nematodes, and plants (23-25,28-32).

Although the codes of nucleotide recognition by RVDs have been established and five design guidelines were derived from naturally occurring TALE target sites (23), these guidelines are not sufficient to provide discrimination against suboptimal target sites. Recent evaluation of the existing design guidelines using hetero-dimeric TALEN pairs (2) revealed that the activities of the TALEN pairs varied markedly; however, no significant correlation between guideline violations and TALEN activities was found, possibly because the assessment was on TALEN pairs rather than individual TALEN monomers. Existing design tools (such as TALE-NT 2.0) intend to help users to filter gene sequences based on simple qualitative criteria; however, they often result in a large number (hundreds to thousands) of potential TALEN target sites with activities varying over a wide range, clearly indicating the need for a new design tool that selects high-activity target sites for TALENs.

We have developed a new online design tool, Scoring Algorithm for Predicting TALEN Activity (SAPTA), to quantitatively evaluate target sites by assigning scores that reflect predicted TALEN activities, thus allowing end users to select the optimal target sites among many possible choices within a given gene segment through the use of SAPTA at http://bit.ly/SAPTA. The SAPTA prediction is based on experimentally measured activities of 130 TALEN monomers constructed with the guanine-targeting RVD NK (NK-TALENs). Although NK-TALENs generally have higher specificity compared with TALENs containing the guanine-targeting RVD NN (NN-TALENs) (1,24), their activity level is usually lower than NN-TALENs. Therefore, it is important to increase the activity level of NK-TALENs for effective gene editing.

To evaluate the performance of SAPTA, 75 additional NK-TALEN monomers were tested. We found that SAPTA-designed TALENs have significantly higher activity compared with TALENs designed following existing guidelines. Specifically, SAPTA-based TALEN designs increased the average intracellular TALEN monomer activity by >3 fold, achieved a larger percentage of highly active TALENs compared to that in previous studies (2,4,7), and resulted in an average endogenous gene modification frequency of 39% for TALENs containing the repeat variable di-residue NK that favors specificity rather than activity. Furthermore, we showed that SAPTA can also be used to design NN-TALENs that have improved activity.

4.2 Results

Design and test of TALENs in the Training Set

To establish the correlation between TALEN activity and design variables, we first performed PCR-based, modified single-strand annealing (SSA) assays to quantify

the activity of individual TALEN monomers in cleaving extrachromosomal plasmids with a homo-dimeric target site and a fixed 17-bp spacer. This assay bypassed the effect of genomic context at endogenous gene targets that may obscure the correlation between TALEN activity and design variables. We quantified each TALEN monomer's SSA activity or % SSA, defined as the percentage of SSA-repaired cleavage products in HEK293T cells co-transfected with plasmid encoding the TALEN monomer and the target plasmid (Figure 11). The validity range of this method was determined by establishing a standard curve that indicated a near-linear correlation between the percentage of SSA-repaired PCR products measured and the fraction of pEGFP (identical to a SSA-repaired plasmid) in the co-transfection mix up to ~50% (Figure 24).

We assembled and measured the SSA activities of 130 TALEN monomers targeting a variety of sequences (3). All of the 130 TALENs were constructed using RVDs NK for G, HD for C, NI for A and NG for T, and with a 5'-T preceding each TALEN target half-site. NN-TALENs generally have higher activity, but may lead to lower specificity compared with NK-TALENs (1,24). We therefore used 130 NK-TALENs as our Training Set to facilitate the design of highly active NK-TALENs. We compared the on- and off-target activities of a few sets of TALENs that differed only in the G-targeting RVDs (NN vs. NK). Transfecting cells with plasmids for NN-TALENs resulted in higher off-target cleavage than corresponding NK-TALENs (73).

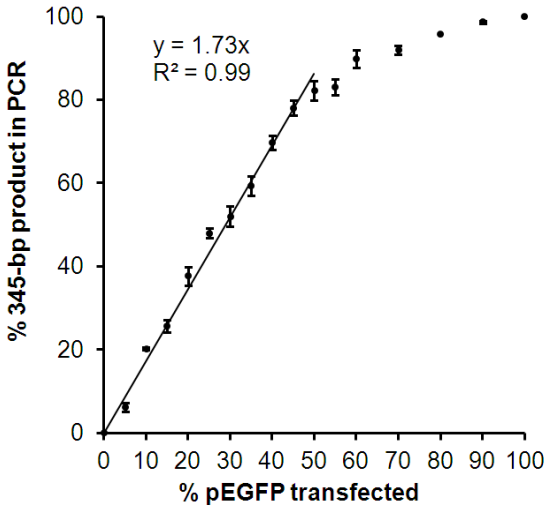


Figure 24: A standard curve validating the modified SSA assay measuring TALEN monomer activity in HEK293T cells.
Error bars, s.e.m. (n=3).

TALENs in the Training Set were designed by considering two major groups of TALEN target sequences. One group contains TALEN target sequences that were selected from output lists for several gene segments screened by the old version of TALEN Targeter (23). The selection of these TALEN target sites was somewhat arbitrary, except that they should be located near the site of interest. The other TALENs targeted artificial sequences that we specifically varied to test the effect of certain design variables, including mutating specific nucleotides at positions near the 5' and 3' of the target sequence, varying the numbers of maximum consecutive A's and G's, and increasing the percentages of certain nucleotides in the first and last five nucleotides of the target sequence. TALENs of the second group were labeled as "n/a" in the column of "Target gene" in Supplementary Table S4 of reference (3)**.** Together, the Training Set

includes 130 TALEN monomer target sequences, 74 for targeting *HBB*, 7 for *CFTR*, 2 for *CXADR*, 4 for *ERCC5*, and 43 with artificial sequences (n/a). The effects of neighboring nucleotides were not considered here since no neighboring effects have been reported in previous studies (22,23). If future studies identify any neighboring effect, additional variables (for example, product of two terms) will be introduced into the SAPTA function and the parameters re-optimize.

The Training Set also covers reasonably large ranges for each design variable we tested, including 14-30 repeat arrays, 0-56% A, 14-53% C, 0-45% G, 4-56% T; 0-80% A, C, G, and T in the first five nucleotides; 0-80% A and C in the last five nucleotides, and 0-100% G and T in the last five nucleotides. Due to practical considerations, it is not possible to test all possible combinations of nucleotides throughout the DNA binding domain of a TALEN. We chose the Training Set so that these TALENs cover a wide range of the design variables considered, with detectable and varying activities.

We tested the activities of 37 TALEN pairs from the Training Set at their endogenous gene targets using the T7 endonuclease I (T7E1) assay in HEK293T cells (Figure 25). These 37 TALEN pairs contain all combinations of active TALEN monomers in the Training Set separated by spacers ranging from 15 to 30 bp in targeted human genes. TALEN pairs with spacers smaller than 14 bp and higher than 19 bp showed no detectable gene-modification activity despite high SSA activities for some of the TALEN monomers. The activities of TALEN pairs with optimal spacers (defined as 14-19 bp) were compared with their "composite SSA activities" which integrated left and right TALEN monomer activities (see 'Materials and methods' section) (Figure 26). We

found that TALEN pairs from the Training Set having observable gene modification activities had composite SSA activities ≥30 (Figure 26).

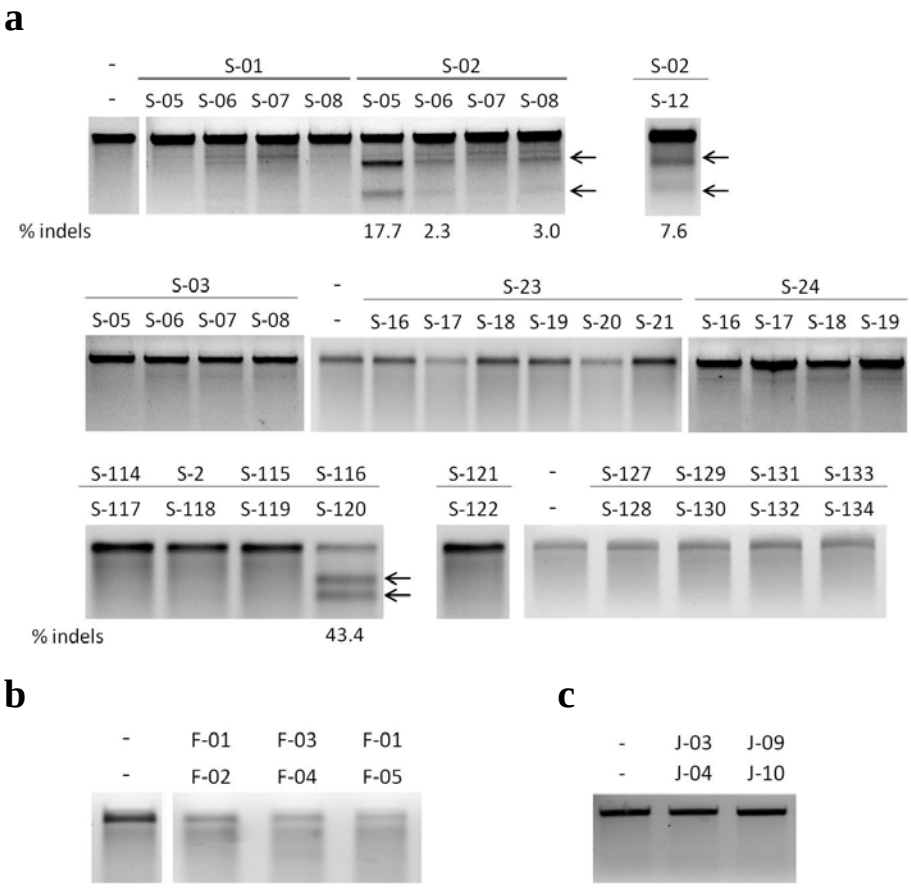


Figure 25: T7E1 assay results for 37 unselected, training- set TALEN pairs.
Detectable activity in the T7E1 mutation detection assay was observed for five of the 37 TALEN pairs in the training set. Activities greater than 1% are shown as percentage of indels below the corresponding gel lane. Each activity was calculated from 3 independent experiments. The 37 pairs include (a) 32 pairs targeting the β-globin (*HBB*) gene (b) 3 pairs targeting the *CFTR* gene and (c) 2 pairs targeting the *ERCC5* gene. Lane headings indicate the left and right TALEN index numbers. "-" denotes the control lane where samples were treated with an empty TALEN backbone. Arrows indicate the positions of specific T7E1 cleavage products.

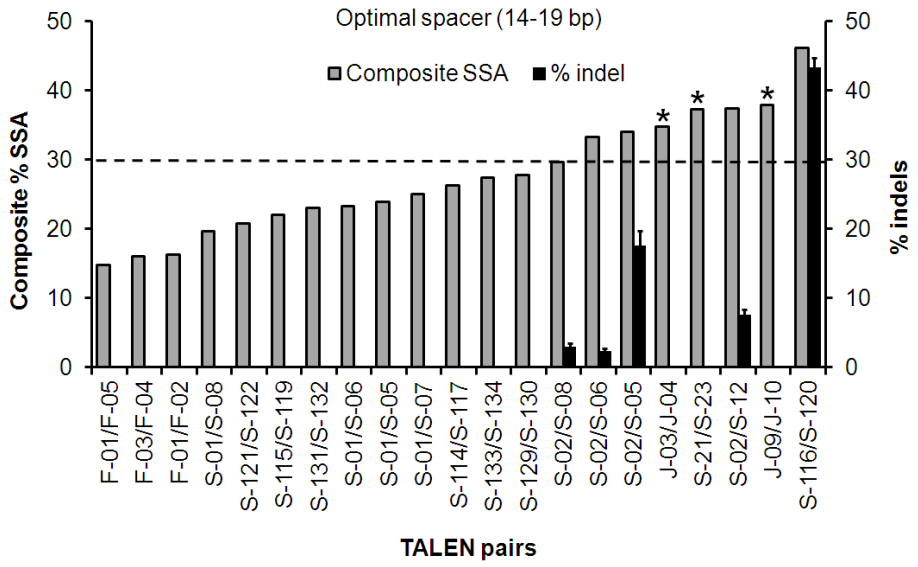


Figure 26: Comparison between composite SSA activity and the endogenous gene modification rates of TALEN pairs.
TALEN pairs from the training set with optimal range of spacer are ranked by their composite SSA activities (grey columns) from low to high (left to right). Endogenous gene modification rates of the corresponding TALEN pairs quantified by T7E1 assay are shown as black columns. Both SSA activity and endogenous gene-modification activity were measured in HEK293T cells. Dashed line indicates composite SSA activity of 30, which is achieved when both left and right TALEN SSA activities are ~10%. Asterisks indicate three TALEN pairs with >30 composite SSA activities but no gene-modification activity. Error bars, s.e.m. (n=3).

Three TALEN pairs (indicated with asterisks in Figure 26) showed no gene-modification activity, although they had >30 composite SSA activities. Sequencing results of the target sites showed no polymorphism, and the cellular expression levels of these TALENs were similar to other TALENs (with different lengths of repeat arrays) (Figure 27), suggesting that the lack of detectable gene-modification activity was not due to target variation or low protein expression. Further investigation showed that the two

ERCC5-targeting pairs, J-03/J-04 and J-09/J-10, both have high cellular activity with extrachromosomal plasmid targets. Bisulfite sequencing of this locus revealed methylated cytosines in all the CpG di-nucleotides in the target sequences of these two pairs, consistent with previous studies showing that methylated cytosines blocked the binding of TAL effectors (TALEs) or TALENs (7,74,75). Since no CpG site is present in the target sites of the TALEN pair S-21/S-23, undetectable gene-modification activity for this pair may be due to the long repeat array of S-21 (29 repeats). The large size of TALEN may prevent access to the target locus.

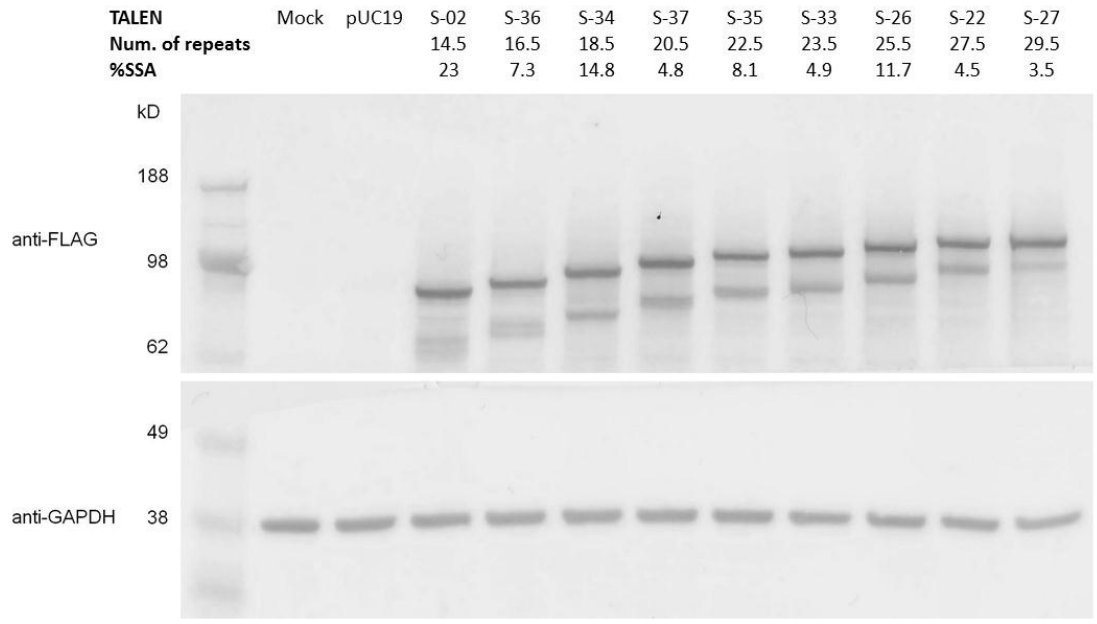


Figure 27: Western blot analysis of TALENs with 14.5 to 29.5 repeats.
These TALENs had a range of monomer activities, but similar expression levels when transfected into HEK293T cells, as for the Single Strand Annealing assay (Material and Methods). Cell lysates were collected two days after transfection, separated on NuPAGE Novex 4-12% Bis-Tris gels (Life Technologies) and blotted with antibodies. Expression levels were determined using antibodies to the FLAG tag (at the N-terminus of each TALEN). Antibodies to glyceraldehyde-3-phosphate dehydrogenase (GAPDH) were used as the internal control.

Table 2: Evaluation of existing design guidelines and development of new design guidelines

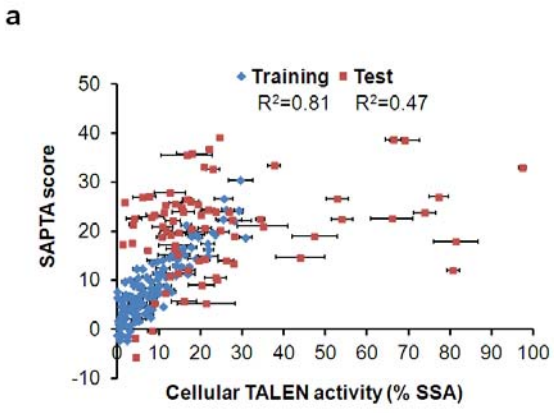
Feature in target site	Existing guidelines	Optimal value[a] from this study	Relative importance[b]	Recommendation
Comparison with previously published guidelines (2,23)				
Base identities at 5' (Pos. 1), 2nd nt. from 5' (Pos. 2), and 3' ends (23)	No T at Pos. 1 No A at Pos. 2 T at the 3' end	G at Pos. 1 T at Pos. 2 G at 3' end	+	Some specific nucleotides at these positions may have minimal effect
Overall base composition (23)	A 31 ± 16% C 37 ± 13% G 9 ± 8% T 22 ± 10%	0% 53% 10% 37%	+++	Choose target sequences with a large percentage of C[c]
Length of target sequence (2)	15-20 bp	15-25 bp	+	The length of a target sequence should be 15~25 bp[c]
Spacer length (2)	16-19 bp	14-19 bp[d]	n/a	Spacer length should be 14~19 bp
New design guidelines				
Base composition of the first 5 nt	A C G T	≤20% 60-80% Not decisive Not decisive	++	The first 5 nt of the target sequence should contain a large %C[c]
Base composition of the last 5 nt	A C G T	≤60% Not decisive ≤60% 80-100%	++	The last 5 nt of the target sequence should contain a large %T[c]
Max. num. of consecutive A's		≤3	+	Longer stretches of A's can lower the activity
Max. num. of consecutive G's		≤3	++	Longer stretches of G's can lower the activity

[a] Optimal value shows the value of a certain variable that maximizes its score contribution (defined as the sum of all contributions from the variable), with the constraint that the value of this variable should be within the range in the Training Set.

[b] Relative importance of each design feature was rated by its magnitude of contribution to the score.

[c] Target sequence refers to a half-site targeted by a TALEN monomer, excluding the 5'-T immediately before the 5' end of the half-site.

[d] Acceptable values for spacer length were observed from T7E1 assays of TALEN pairs (3).



b

Training set

	Low %SSA (≤10)	High %SSA (>10)
Low score (≤10)	94%(77/82)	6%(5/82)
High score (>10)	21%(10/48)	79%(38/48)

Test set

	Low %SSA (≤10)	High %SSA (>10)
Low score (≤10)	50%(4/8)	50%(4/8)
High score (>10)	15%(10/67)	85%(57/67)

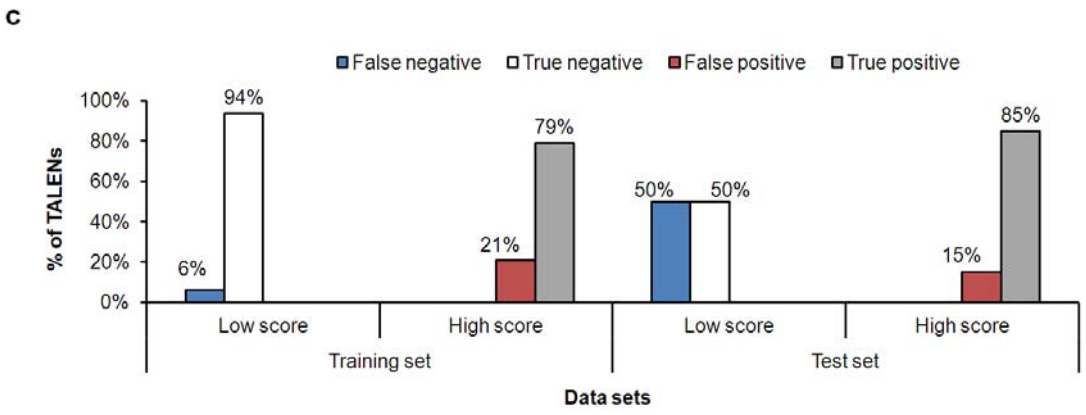


Figure 28: Development and evaluation of SAPTA using 205 NK-TALEN monomers.
(a) Correlation between predicted SAPTA scores (y-axis) and intracellular TALEN-monomer activity (x-axis) measured by the modified SSA assay in HEK293T cells. Error bars, s.e.m. (n=3). (b) Categorization of TALENs in the training set and test set using 10% SSA activity as a cut-off for high activity. TALENs with low scores (≤10) and high scores (>10) are evaluated separately. (c) The percentages of TALENs predicted by SAPTA scores being false negatives (score ≤10 but %SSA >10%) and true negatives in the low score group, and false positives (score >10 but %SSA ≤10%) and true positives in the high score group.

SAPTA: Scoring Algorithm for Predicting TALEN Activity

The SAPTA algorithm, which contains a total of 30 variables (see detailed in the 'Materials and methods' section), was established based on the measured SSA activities of 130 NK-TALENs in the Training Set. The SAPTA variables were chosen to evaluate the existing design guidelines (2,23) and to establish new design guidelines (Table 2). Without any pre-assumption, based on the measured SSA activities of TALENs in the Training Set, the parameters in the SAPTA algorithm were solved using linear modeling in the R statistical software (version 2.15.2) (76), which gave rise to an excellent correlation with the Training Set ($R^2 = 0.81$) (Figure 28a). Consequently, SAPTA is able to provide a numerical score that estimates TALEN activity (high scores indicate high activity).

Using SAPTA, we evaluated the TALEN design guidelines proposed by Cermak *et al*.(23), Reyon *et al*. (2), and Streubel *et al.* (1), and established new design guidelines (Table 2, Figure 29, and Figure 30). Specifically, the strong bias for C in the optimal nucleotide percentages calculated using the SAPTA agrees well with previous studies indicating that the RVD HD for C leads to high DNA-binding affinity (1,50). The optimal length of target sequences and spacer length identified using SAPTA are consistent with design guidelines proposed by Reyon *et al.* (2). In addition, our experimental results indicated that long stretches of A's or G's decrease the SSA activity, as reflected in SAPTA predictions (Figure 29**)**, especially with >3 consecutive G's. This conclusion derived by SAPTA is also consistent with the relatively low binding affinities of RVDs NI and NK that target A and G, respectively (1).

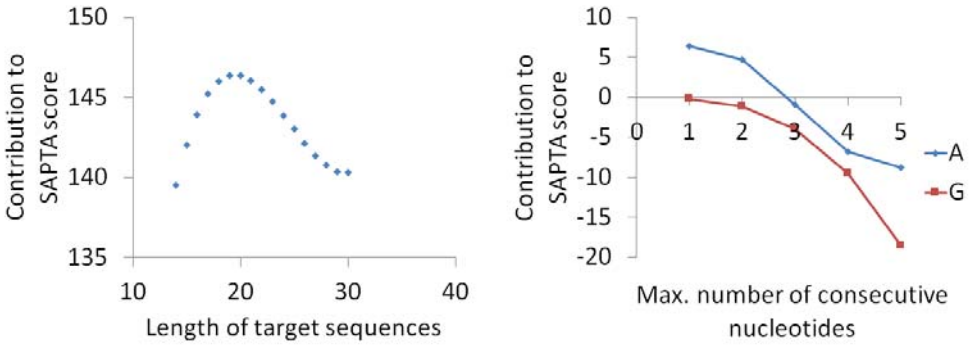


Figure 29: Contribution of target length (left) and long stretches of A's and G's (right) to SAPTA scores.
The graphs were plotted using functions corresponding to these variables from SAPTA algorithm (3) explained in the "materials and methods" section.

By examining the SAPTA function, we also found that the nucleotide percentages of the first and last five nucleotides of target sequences are important. We varied the first five nucleotides systematically to sample all possible combinations (4^5=1024), followed by a constant sequence "AACCTCTGGGTCCAA" to create a 20-nt sequence. SAPTA scores of these 1024 sequences that only differ in the first 5 nucleotides were recorded, except for those with variable values outside the ranges of Training Set. Box-and-whisker plots show the sets of scores recorded at 0-80% or 0-100% of each nucleotide (Figure 30). The upper limits of nucleotide percentages were determined by limits of the Training Set. Similarly, the last five nucleotides were varied to create 1024 sequences, preceded by a constant sequence "CACAGAACGTCAGGT". The score analysis was performed the

same as for the first five nucleotides. In general, high TALEN scores are achieved with a large percentage of C at the 5'-end and T at the 3'-end of the target sequence (Figure 30).

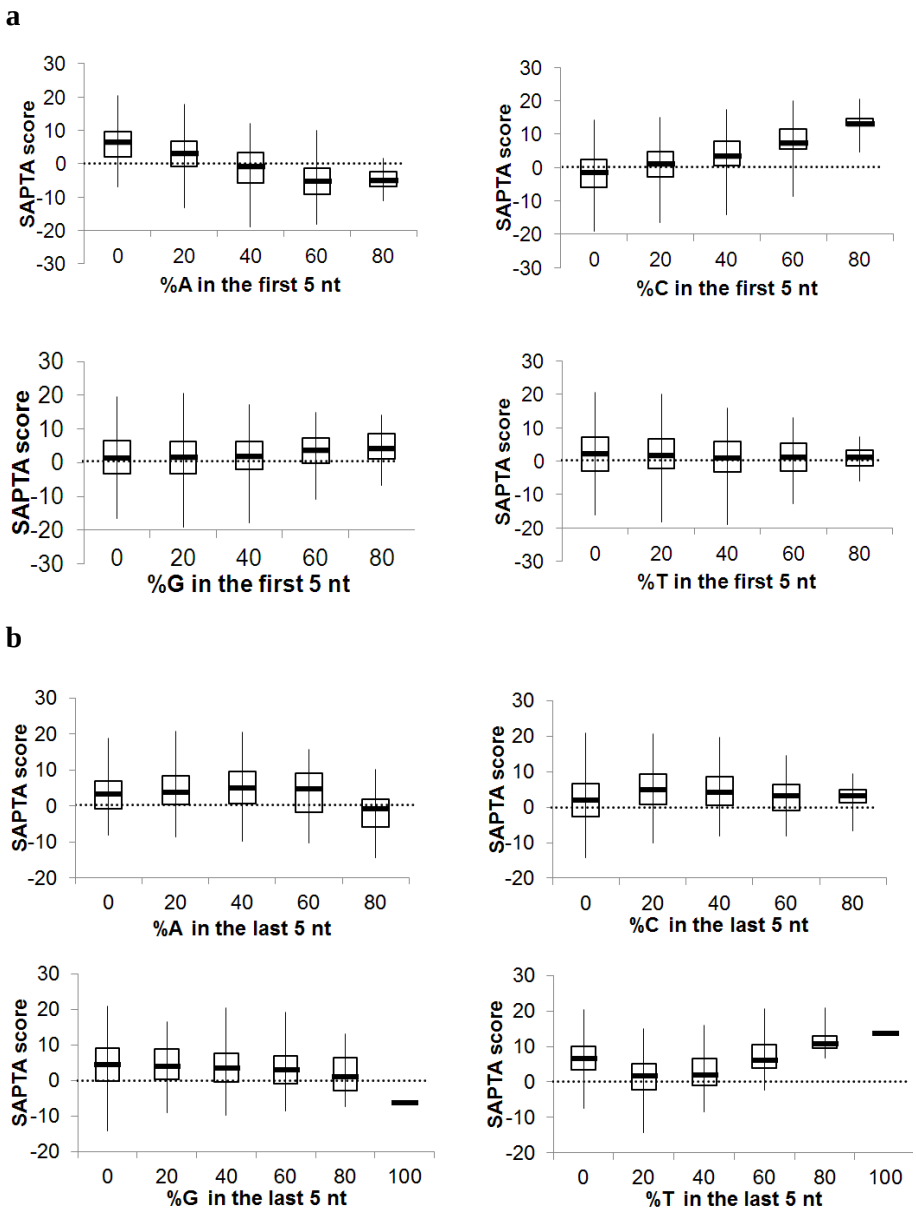
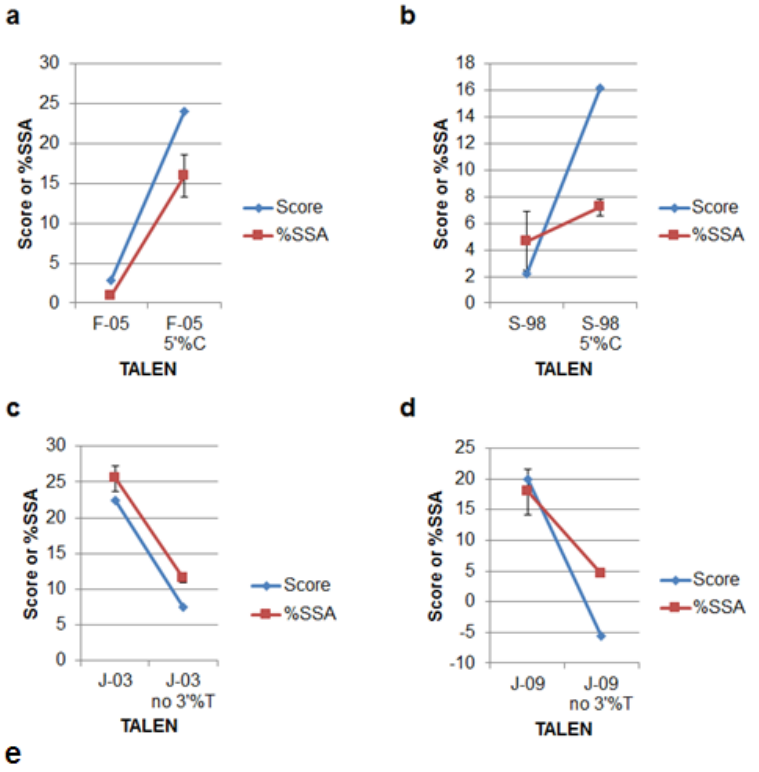


Figure 30: Contribution of base composition of the (a) first and (b) last five nucleotides to SAPTA scores.

This new design guideline was experimentally confirmed by specifically varying sequences at the 5' and 3' ends of TALENs in Test Set 1 (3), which contains 12 TALENs. To minimize the change of other variables, in Test Set 1, TALENs were designed by re-arranging nucleotide positions in the original target sequences without changing the overall base compositions and lengths. Five nucleotides containing 80% of C's from the 3' ends of TALENs labeled F-05 and S-98 were switched with nucleotides at their 5' ends (Figure 31). This switch resulted in large percentages of C's at their 5' ends and the removal of C's from the 3' ends, thus increasing their scores and SSA activities. Five nucleotides containing 80% of T's from the 3' ends of TALENs labeled J-03 and J-09 were switched with five nucleotides in the middle or their targets (Figure 31). This switch resulted in large percentages of C's at 3' ends and the removal of T's from the 3' ends, thus lowering the scores and SSA activities. When the five nucleotides at the 5' ends were replaced by 80% C's originally located at the 3' ends, the overall C composition remained the same, but the SSA activities increased. When the 3' T's were replaced with 60% or 80% of C's, while the overall base compositions stay unchanged, the SSA activities decreased (Figure 31). The increase in the percentage of C's at the 5' end is accompanied by changes at the 3' end due to the base swaps (Figure 31e), and both changes may cause alterations in TALEN activity. However, according to SAPTA predictions, changes at the 3' end from 80% C to 0% or 20% C would have minimal

effect on SAPTA scores (Figure 30). Therefore, our analysis suggests that the increase in TALEN activities is largely due to an increase in the percentage of C's at the 5' end.



e

Features	TALEN Index #	Target Sequence (without 5'-T)
F-05, 5' %C	F-05	GAAGGCTCCAGTTCTCCC
	F-16	CTCCCCTCCAGTTGAAGG
S-98, 5' %C	S-98	GGTGCACCTGACTCCC
	S-175	CTCCCACCTGAGGTGC
J-03, no 3' %T	J-03	TTTCGAATTCGTCCTATTT
	J-49	TTTCGAATTTATTTCGTCC
J-09, no 3' %T	J-09	CGGCTCTGCAAACTCTTATTTTT
	J-50	CGGCTCTGCAAACTCTTACCCCC

Figure 31: Experimental validation of new design rules on the base compositions in the first and last five nucleotides of NK-TALENs.

SAPTA, a user-friendly online search tool was established to help researchers identify optimal TALEN target sites within a selected DNA sequence. The DNA sequence of interest, together with the ranges of acceptable target and spacer lengths are entered into the web interface, which then outputs a ranked list of SAPTA scores for each TALEN pair, together with the corresponding target sequences and the nucleotide preceding each target half-site (2,23). For each pair of TALENs, a single numerical value—the composite SAPTA score—is defined based on the scores of the left and right TALENs in a way that favors pairs with balanced left and right scores (see 'Materials and methods' section).

To demonstrate that the use of SAPTA improves the design of NK-TALENs, we employed SAPTA to design 63 additional NK-TALEN monomers that form Test Set 2 (3), with SAPTA scores ranging from 5.3 to 39.1 and measured their SSA activities. Specifically, the target sites for TALENs in Test Set 2 were determined by using the SAPTA online tool to search 19 gene segments with TAL arrays (DNA binding domains) of 14 to 25 repeats, and spacer lengths from 14 to 19 bp. TALEN pairs with high composite scores ranked by the online tool were chosen. We mostly selected SAPTA-designed TALEN monomers with SAPTA scores >10.0, since TALENs with SSA activities >10% are likely to result in gene-modification activity (Figure 26). Based on SSA measurements, the average SSA activity of TALENs in Test Set 2 was 27.2%, compared with 8.6% in the Training Set. Furthermore, we found that 86% of NK-TALEN monomers in Test Set 2 had >10% SSA activity, compared with only 33% in the Training

Set (Figure 32). The resulting *P*-value of $2.70X10^{-12}$ suggests that the large frequency of highly active TALEN monomers (SSA activity >10%) is associated with the use of SAPTA.

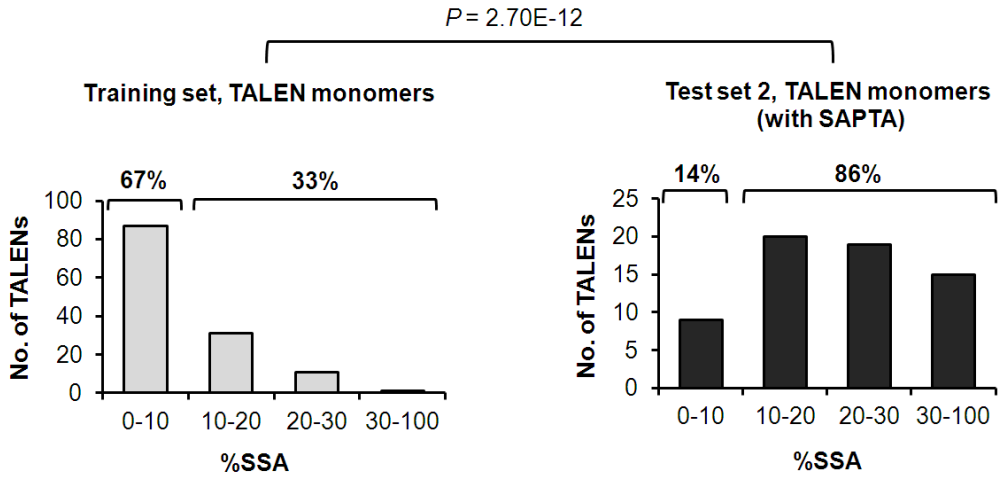


Figure 32: TALEN-monomer activity distribution was substantially improved in the Test Set 2 compared to the Training Set.
The percentages of TALENs with SSA activity ≤10 and >10 are shown above each graph. The two-sided P-value determined by Fisher's exact test is shown on top.

Most TALENs in the total Test Set (the combination of Test Set 1 and Test Set 2) had SSA activities that correlated with their SAPTA scores, especially those with cleavage activities within the range of the Training Set (Figure 28a). The relatively low R^2 (0.47) for the Test Set is largely due to the 12 TALENs that had SSA activities much higher than those in the Training Set (maximum 30% SSA activity), thus could not be

modeled accurately by SAPTA. However, as shown in Figure 28, SAPTA predictions for TALENs in the Test Set only had 15% "false positive" rate, defined as the ratio of the number of TALENs with low (≤10%) SSA activity in the high score group (score >10), divided by the total number of high score TALENs (Figure 28b, c). Similarly, the "false negative" rate is defined as the ratio of the number of TALENs with high (>10%) SSA activity in the low score group (score ≤10), divided by the total number of low score TALENs. The relatively high (50%) false negative rate for TALENs in the Test Set might be due to the small sample size (Figure 28b, c). However, this should not significantly affect SAPTA usage, since users typically select the highest scoring TALENs in a region for further testing, and having activities higher than predicted scores will not negatively impact effective TALEN designs.

Validation of SAPTA with NK- and NN-TALEN pairs targeting endogenous genes

To quantify the gene-modification activities of SAPTA-designed NK-TALENs, we examined 24 NK-TALEN pairs from Test Set 2 targeted to 15 human genes (see Supplementary Table S5 of reference (3) for details) by co-transfecting HEK293T cells with the two plasmids encoding each TALEN pair, and quantified the gene modification rates in the intended target regions using the T7E1 assay (Figure 33 and Supplementary Table S5 of reference (3)). The other TALEN pairs formed by TALEN monomers in the Test Set 2 were not tested either because the corresponding target loci could not be specifically amplified due to the presence of repetitive sequences, or due to the fact that they were targeted to non-human genes. Single Molecule Real Time (SMRT) sequencing (77) showed mutation frequencies comparable to those determined by the T7E1 assay (3).

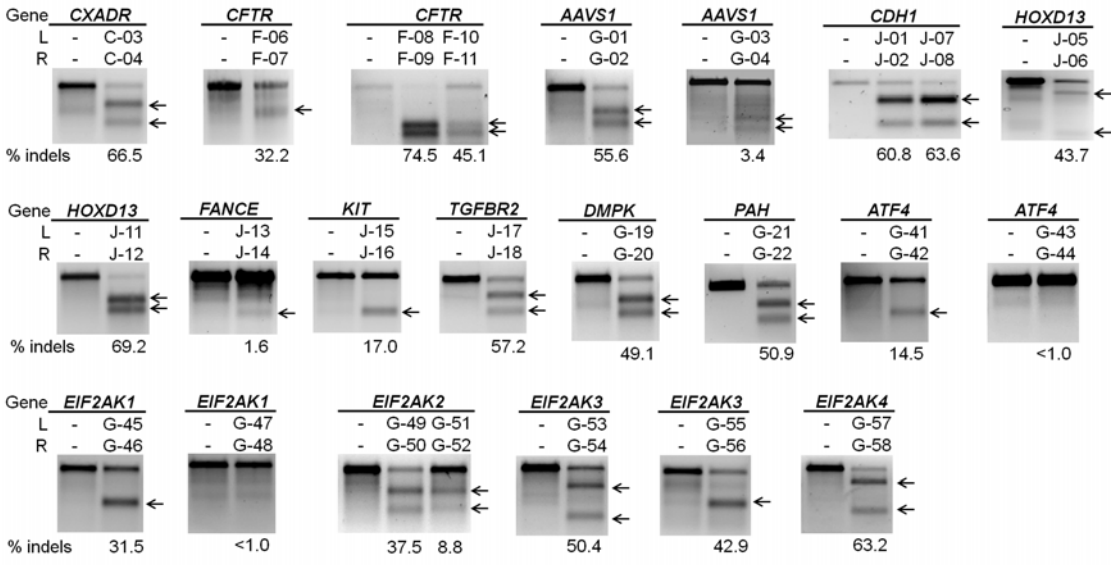


Figure 33: T7E1 assay measuring the endogenous gene modification efficiency of 24 NK-TALEN pairs designed by SAPTA.

We found that the majority of NK-TALEN pairs designed by SAPTA were highly active: the average mutation rate was 39.1%, and 71% (17/24) of TALEN pairs had mutation rates of >20% (Figure 34). However, for the 21 NK-TALEN pairs formed from the Training Set (designed without SAPTA), the average mutation rate was 3.5%, and only one TALEN pair had a mutation rate of >20%. Further, we found that qualitative design guidelines proposed in two previous studies (1,2) are not sufficient in designing active NK-TALENs. As illustrated in Figure 34 and Supplementary Table S6 of reference (3), of the 35 NK-TALEN pairs that followed the three design guidelines suggested by

Reyon *et al.* (2) and the first two guidelines suggested by Streubel *et al.* (1), 21 pairs (60%) showed non-detectable cleavage activity as measured by the T7E1 assay.

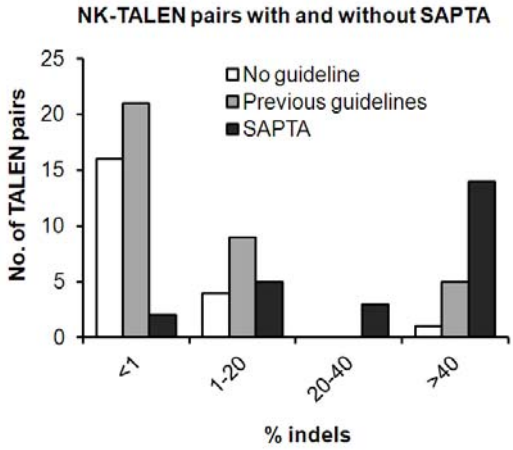


Figure 34: Activity distribution of SAPTA-designed NK-TALEN pairs targeting endogenous genes. The majority of NK-TALEN pairs designed by SAPTA have high activity targeting endogenous genomic loci (black columns), whereas unselected NK-TALENs built without design tools (white columns) and NK-TALENs that follow previous guidelines (1,2) (grey columns) generally have low or undetectable activities.

Previously, five genes *CDH1, HOXD13, FANCE, KIT* and *TGFBR2* were targeted by NN-TALENs without any success (2). The same five genes were targeted again by seven SAPTA-designed NK-TALEN pairs formed by TALENs in Test Set 2, resulted in gene modification rates of 1.6% - 69.2% (Figure 35 and Figure 33), further demonstrating the advantage of using SAPTA for TALEN designs.

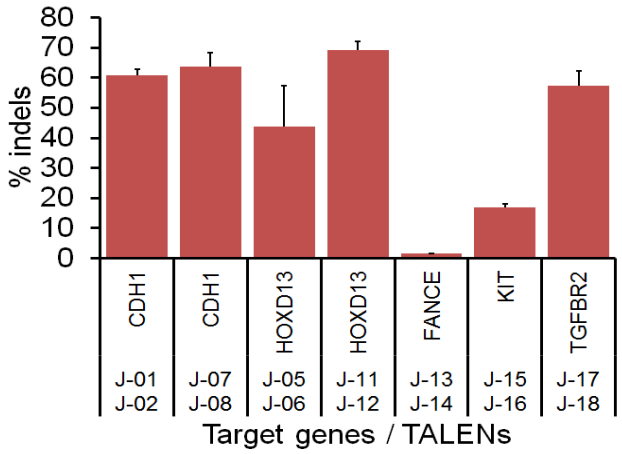


Figure 35: Activity of SAPTA-designed NK-TALENs targeted to five previously attempted genes. Error bars, s.e.m. (n=3). Gel images are shown in Figure 33.

Although SAPTA was established based on SSA measurements of NK-TALENs, we verified that it could also be used to identify target sites for highly active NN-TALENs. We designed 14 pairs of NN-TALENs using SAPTA (Figure 36), including 9 NN-TALEN pairs targeted to the same DNA sequences as the corresponding NK-TALEN pairs formed from Test Set 2, and 5 NN-TALEN pairs targeted respectively to five genes *AXIN2*, *EXT2*, *FANCA*, *FLT4* and *HMGA2* (see more details below). We found that 71% (10/14) of SAPTA-designed NN-TALEN pairs had mutation rates of >20% (Figure 37), with an average mutation rate of 35.2%, similar to the SAPTA-designed NK-TALEN pairs.

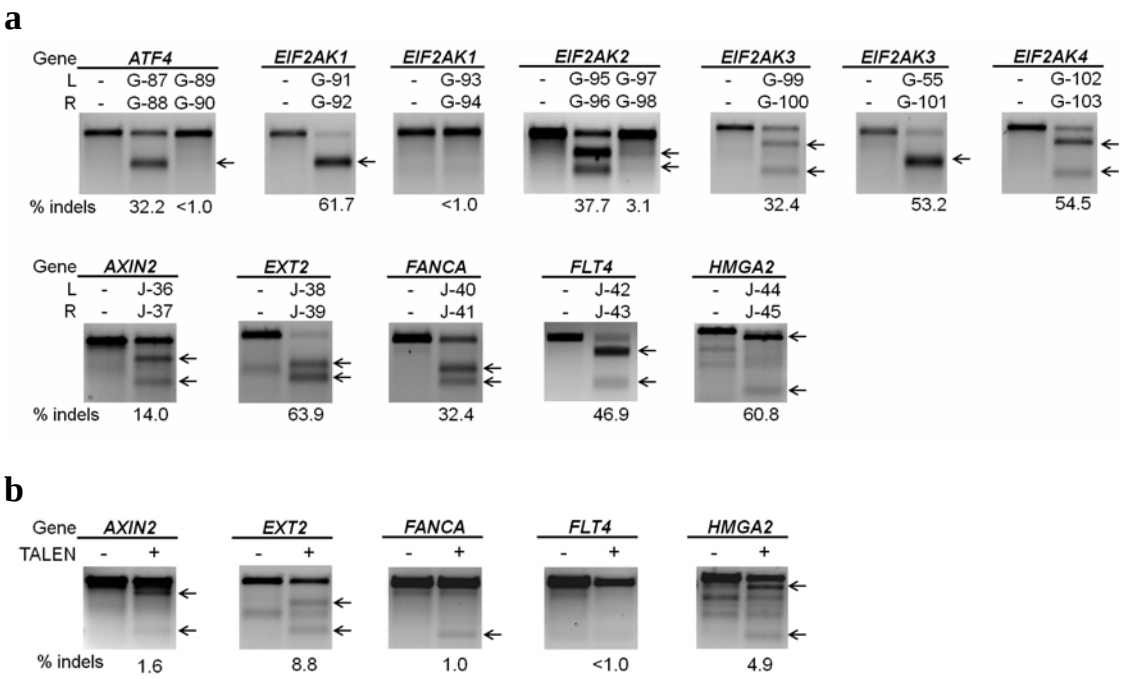


Figure 36: T7E1 assay measuring the endogenous gene modification efficiency.
(a) 14 pairs of NN-TALENs designed by SAPTA. (c) NN-TALENs obtained from a previous study(2) were tested using T7E1 assay in this study. Lane headings indicate the target genes and the left (L) and right (R) TALENs. "-" denotes samples treated with an empty TALEN backbone. Numbers below each lane show the average percentage of modified alleles (n=3). Arrows indicate specific T7E1 cleavage products.

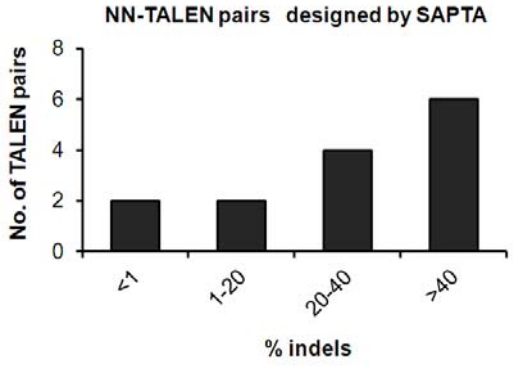


Figure 37: Activity distribution of NN-TALEN pairs designed by SAPTA.

As reported previously, NN-TALEN pairs conforming to existing guidelines (1,2) targeted five genes (*AXIN2*, *EXT2*, *FANCA*, *FLT4*, *HMGA2*) but showed mutation rates of <10% (2). We tested these NN-TALEN pairs (ordered from Addgene) in HEK293T cells (as with other TALENs we tested), and confirmed mutation rates of <10% (Figure 36). Note that the *EXT2*-targeting TALEN pair contains a left TALEN with a negative SAPTA score, therefore is predicted to have low activity. Further, four of these TALEN pairs (*AXIN2*, *FANCA*, *FLT4*, and *HMGA2*) contained TALEN half-sites with >45% G, which is outside (higher than) the range of the G content considered in SAPTA. Specifically, for nucleotide G, the RVD NK has a low binding affinity, thus a large percentage of this RVD in a TALEN can substantially weaken the overall binding of TALEN to the DNA target (1); on the other hand, the RVD NN was shown to recognize A in addition to G (1), so more than 45% of the NN RVD will likely lower the specificity of the TALEN. Therefore, SAPTA does not consider target sequence with >45% G content. We re-designed NN-TALEN pairs using SAPTA to target the same five genes <50 bp away from the previous targeted sites (2), and found that SAPTA-designed NN-TALEN pairs resulted in mutation rates of 14.0% - 63.9%. Each pair is >7-fold more active than the corresponding NN-TALEN pair reported previously (Figure 38).

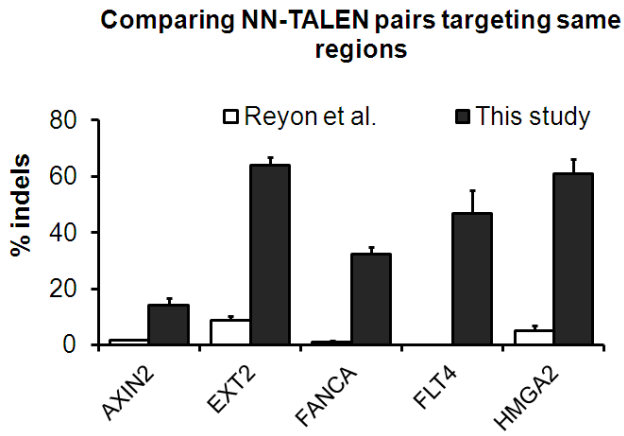


Figure 38: Gene modification frequencies (% indels) of NN-TALEN pairs designed by SAPTA compared to NN-TALEN pairs from a previous study(2) targeted to the same gene regions. The T7E1 assay was used to measure the activities of these TALEN pairs in HEK293T cells. Error bars, s.e.m. (n=3).

There are two NK-TALEN pairs from Test Set 2 that showed high monomer SSA activities, but undetectable endogenous gene-modification rates: G-43/G-44 and G-47/G-48 (Figure 33). Similar to the TALEN pair S-21/S-23 from the Training Set, these pairs contain large repeat arrays. The sum of the left and right repeats for each pair exceeds 45 repeats, suggesting that the large sizes of TALEN dimers may prevent access to their target loci. Possibly due to the same reason, two NN-TALEN pairs (G-89/G-90 and G-93/G-94) targeted to the same sites as G-43/G-44 and G-47/G-48, respectively, also failed to show any detectable gene-modification activity. There results suggest that caution should be taken in designing TALEN pairs containing large repeat arrays.

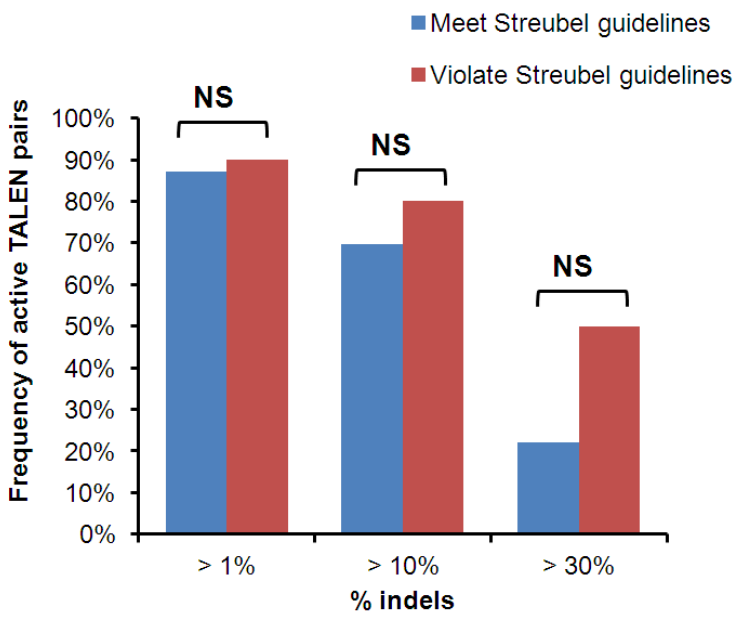


Figure 39: Comparing frequencies of active NN-TALEN pairs tested by Reyon *et al.* (2) that meet or violate guidelines proposed by Streubel *et al.*(1).
The 86 out of 96 pairs meet Streubel guidelines, and 10 out of 96 pairs violate the guidelines. Different mutation rates (% indels) are used as cut-offs for active TALEN pairs. Frequencies (%) of active TALEN pairs with mutation rates above these cut-offs are shown respectively. NS, not significant, as determined by Fisher's exact test, with the hypothesis that meeting the guidelines increases TALEN activity.

Evaluation of existing design guidelines

For effective TALEN designs, the existing guidelines are inadequate, since too many potential TALEN target sites conform to the current guidelines (1,2), even within a short 100-bp gene segment. Little guidance is given to identify the optimal target site for a gene of interest, especially with NK-TALENs. Reyon *et al.* and others (2,4,7,71) have shown that there is a large range of activities (and some with very low or non-detectible activity) among TALENs designed using the existing guidelines, even with NN-

TALENs. For example, of the 35 NK-TALEN pairs we tested that followed the guidelines proposed by Streubel *et al.* (1), only 14 pairs (40%) had detectable gene-modification activities (Figure 34). These guidelines did not seem to improve NN-TALEN activity either. Among the 96 TALEN pairs targeting endogenous human genes tested by Reyon *et al.* (2), 86 TALEN pairs that followed the guidelines showed activity levels comparable to the 10 pairs violated the guidelines (Figure 39).

We further used SAPTA to examine the target sites of these 96 NN-TALEN pairs tested by Reyon *et al.* (2). Within the constraints of all variables, SAPTA identified 21 TALEN pairs with composite scores above 30, which is our score threshold for active pairs. Only 2 of these 21 pairs have gene-modification rates below 10%, indicating a false positive rate of 9.5%. Without filtering by SAPTA, 28 out of 96 (29.2%) TALEN pairs have modification rates below 10%. The false positive rate of 29.2% for their original design is about three times higher compared to that using SAPTA. It should be noted that SAPTA parameters were optimized using NK-TALENs, thus the SAPTA scores may not correlate well with the activities of NN-TALENs. Nevertheless, our results suggest that SAPTA could be used to increase the chance of obtaining highly active NN-TALENs.

Comparison between SAPTA and other TALEN design web tools

A few TALEN design tools have been published recently: TALE-NT 2.0 (78), Mojo Hand(79), and E-TALEN (80). While these tools can facilitate TALEN designs, we strongly believe that SAPTA is a much better design tool in helping researchers choose highly active TALEN designs from hundreds from hundreds perhaps even thousands of potential designs by providing an experiment-validated scoring/ranking system for

TALEN target sites. A brief evaluation of the previously published TALEN design tools in comparison to SAPTA is given below.

TALE-NT 2.0 (78)

TALE-Nucleotide Targeter (NT) 2.0 tool was developed based on a minimal constraint (5'T) and recently proposed design guidelines (1). These guidelines have not been experimentally validated, nor has the correlation between the proposed guidelines and higher TALEN activity been established. In fact, these guidelines do not seem to correlate with high TALEN activity (Figure 39).

In contrast, SAPTA provides TALEN designers with a ranked list of high-activity TALEN target sites. The TALE-NT tool does not have a ranking system for TALEN activity, although it provides scores for potential off-target sites for a given TALEN pair or TALE (78). TALE-NT tool scans a gene segment and identifies all target sites that meet the following three criteria specified by users without further discrimination / ranking: (1) a T or C precedes the 5' end of each target half-site; (2) a spacer range; (3) a range for repeat array length. As shown in the study by Reyon *et al*. *et al*. (2), TALENs that meet all these criteria have activities varying over a wide range, from 0% to 55.8% in a NHEJ-mediated mutagenesis assay. Designing TALENs solely based on these criteria may not lead to high cleavage activity. Since there is no sufficient guidance in selecting target sites, the chances for picking up high-activity TALEN are equal to randomly choosing TALEN target sites. Unlike TALEN-NT, SAPTA outputs a ranked list of TALEN target sites predicted to give high activity, which has been experimentally validated. We have shown that TALENs designed by SAPTA have significantly higher activity compared to those designed by other methods.

TALE-NT typically provides hundreds of potential target sites that meet the design criteria without further discrimination. In contrast, SAPTA provides a ranked list of target sites with SAPTA scores, allowing researchers to choose the top-ranked TALENs for testing. For comparison, we used TALEN-NT 2.0 to search the 350-bp *HBB* gene fragment containing the sickle cell mutation site. We used the default criteria based on Miller *et al.* architecture, which specified spacer ranges from 15-20 bases and RVDs ranging from 15-20. TALEN-NT 2.0 output 3,612 un-ranked target sites for this region, but there is no guidance for deciding which TALENs to construct and test. There is the option to hide redundant TALENs, where TALEN-NT selected the pair with the shortest average length of repeat arrays and spacer and provided 240 unranked TALEN pairs. However, the hypothesis that shorter spacer and length of repeat arrays are better has not been experimentally validated.

Of the 3,612 potential target sites identified by TALEN-NT 2.0, we tested NK-TALEN pairs designed for eight target sites. Six out of the eight pairs showed no detectable gene modification rate (Table 3). The six pairs with no activity were ranked low by SAPTA, all with composite scores lower than 30 (the cut-off score for TALEN pairs considered to be active). Therefore, the ranking provided by SAPTA is a dramatic improvement over TALEN-NT. SAPTA also allows researchers to compare SAPTA scores and the position of the target sites relative to a mutation site to better evaluate the options for gene editing.

Table 3: SAPTA ranking results for eight target sites provided by a search using TALEN-NT 2.0 (78). Cellular activity measured as % indels using the T7E1 assay is shown for TALEN target sites tested in this study that were also found in the 3612 target sites in the TALEN-NT 2.0 output for this gene segment. The SAPTA composite score is shown for each pair and is used to select high scoring sites for TALEN targeting and to screen against using sub-optimal sites (composite score < 30), such as the third through eighth rows below. The TALEN pairs with SAPTA composite scores above 30 had detectable endogenous gene targeting, whereas those with lower scores, did not have detectable activity.

Gene	L-TALEN	L-score	R-TALEN	R-score	Composite score[a]	%indels±s.e.m.
HBB	S-116	30.5	S-120	19.5	44.7	43.4±1.4
HBB	S-02	19.9	S-12	4.8	31.6	7.6±0.7
HBB	S-133	8.0	S-134	5.4	25.7	0
HBB	S-127	9.7	S-128	1.9	25.5	0
HBB	S-115	3.5	S-119	8.6	24.2	0
HBB	S-131	11.1	S-132	1.4	23.0	0
HBB	S-129	2.6	S-130	7.3	22.2	0
HBB	S-114	17.6	S-117	-2.1	N/A[b]	0

[a] $Composite\ \ Score = 5 + 4 \times \sqrt{LS} + 4 \times \sqrt{RS}$, where LS is the L-score (left TALEN score), and RS is the R-score (right TALEN score).

[b] Ignored score due to a negative score for the right TALEN.

Mojo Hand (79)

Mojo Hand is a basic tool that applies a loose filter for possible TALEN target sites in a given gene segment. Besides user-defined ranges of spacer length and repeat array length, the only constraint applied is the requirement for a 5'T before each target site. As mentioned above, a loose filter results in hundreds of unranked potential TALEN target sites, making it difficult for users to identify the best TALENs to construct.

To illustrate, a 350-bp DNA sequence was given as input into Mojo Hand with 15-16 bp spacer lengths and 15-17 TALEN repeat array lengths, and Mojo Hand gave

325 possible target sites as output. Without further discrimination, users of this website would have no guidance on which TALEN pairs they should construct among the 325 possible choices. The Mojo Hand publication validated one single pair of TALENs, but it was not apparent how this pair was chosen (79).

E-TALEN (80)

E-TALEN is a new web-based tool that aims to help researchers with TALEN designs. It was difficult to evaluate E-TALEN, since it does not always provide functional output. For example, when a segment of the *HBB* gene (350 bp) was used as input with the default settings of the E-TALEN website to design TALEN pairs, E-TALEN gave an output of four TALEN pairs that have their target sites only differ by one or two bases. Further, alignment of these TALEN target sites against the input DNA sequence revealed that the orientation of these TALEN pairs was incorrect: Functional TALENs bind to DNA in the 5' to 3' orientation (N to C terminal of TALEN). However, the four TALEN pairs designed by E-TALEN have the opposite orientation, which resembles the binding orientation of zinc finger nucleases. When the default setting in E-TALEN was changed to allow for up to 50 pairs of TALENs in the output, most of the TALEN pairs designed by E-TALEN targeted the same position; the target sites only differ by a few bases. Again, all of the 50 TALEN pairs are in the ZFN-like (3'-to-5') orientation, which is incorrect according to published experimental results.

To further test E-TALEN, a 350-bp gene segment from the myocilin gene was used as input with the default setting changed to allow for up to 50 pairs of TALENs in the output. The TALEN pairs given by E-TALEN target the same position; the target sites only differ by a few bases. Of these 50 TALEN pairs, 27 TALEN pairs have the

wrong, ZFN-like orientation (3'-to-5'); only 23 pairs have the correct, 5'-to-3' orientation.

In summary, it appears that E-TALEN (Version 2.4) outputs TALEN target sites with both 3'-to-5' and 5'-to-3' orientations, and the range of different TALEN target sites is very limited.

Frequency of highly active TALENs with and without using SAPTA

The "success rates" claimed by previous studies categorized all detectable activities, whether 1% or 50% mutation rates measured by T7E1 assays, as being success (2,4,7). In contrast, SAPTA aims to increase the levels of activities, in addition to increasing the fraction of TALENs that have detectable activity. We compared the frequencies of "highly active" TALEN pairs designed using SAPTA to those reported in previous studies (2,4,7). Due to possible differences in experimental conditions across various studies, we categorized "highly active TALENs" by gauging TALEN activities within each study using the highest TALEN activity measured in the same study. Specifically, the "maximum mutation rate" was defined as the highest mutation rate observed in each individual study. The cut-offs for "high activity" were set at a range of activities depending on different maximum mutation rates, ranging from 20% to 80% of each maximum mutation rate (Figure 40). The number of highly active TALEN pairs with activities higher than each cut-off was counted. Using this sliding cut-off, we showed that frequencies of highly active TALEN pairs designed by SAPTA were at least twice the frequencies in other studies at higher cut-offs (40% - 80% of maximum mutation rate) (Figure 40). Therefore, SAPTA-designed TALENs showed a significant improvement in achieving high activity. Having TALEN pairs with high activity levels

(for example, ~40%) compared to those simply meet the relatively low bar of being "active" would be beneficial to many applications

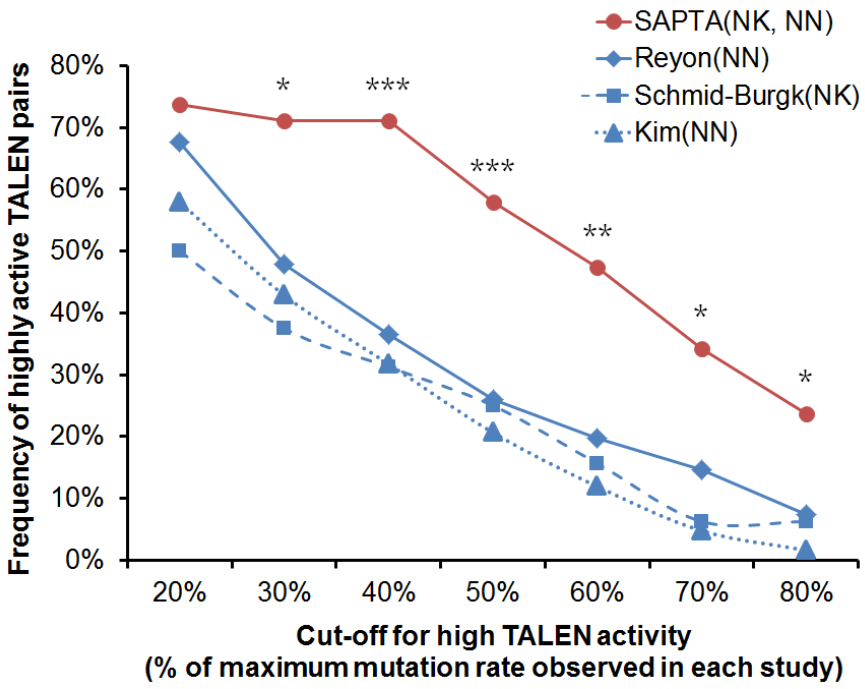


Figure 40: Frequencies of highly active TALEN pairs with and without SAPTA. "Highly active" is defined by a range of different cut-offs (x-axis) from >20% to >80% of the maximal TALEN-mediated mutation rates (% indels) observed in each study. Fractions of maximal % indels were used as cut-offs since the results of T7E1 assays may vary among different studies. SAPTA-designed NK- and NN-TALENs are shown comparing to TALENs from previous studies by Reyon *et al.* (2), Schmid-Burgk *et al.* (4), and Kim *et al.* (7). Asterisks indicate two-sided *P*-values associated with SAPTA-designed TALENs compared to NN-TALENs by Reyon *et al.*. *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$ as determined by Fisher's exact test.

4.3 Discussion

SAPTA is the first quantitative and experimentally-validated design tool for selecting TALEN target sites with high cleavage activity. This represents a significant advance over existing TALEN design tools that often output hundreds to thousands of unranked potential target sites (Table 4 and Table 3). Although the target sites selected by E-TALEN are ranked (81), the ranking is based on the guidelines by Cermak *et al.* (23), which were shown by Reyon *et al*. (2) to lack significant correlation with TALEN activity. In contrast, SAPTA has quantitatively incorporated a wide range of TALEN design guidelines (1,2,23) and provides a ranked-list of potential target sites, allowing users to design TALENs with high activity levels and the desired target sites.

Table 4: Comparison between SAPTA and other TALEN design tools.

Tool	SAPTA	TALE-NT 2.0 (30)	Mojo Hand (31)	E-TALEN (29)
Correct TALEN design	√	√	√	×[a]
Scoring system	√	×	×	√[b]
Experiment validation	√	×	√[c]	×
Number of hits for a 350-bp *HBB* sequence	32	3,612	325	1,902

[a] Some TALEN designs have the wrong, 3' to 5' binding direction.

[b] The scoring in E-TALEN (81) is based on the guidelines by Cermak *et al.* (23), which were shown by Reyon *et al*. (2) to lack significant correlation with TALEN activity.

[c] Only one TALEN pair was tested in this study.

We demonstrated that SAPTA-designed NK-TALENs improved the average monomer SSA activity from 8.6% to 27.2%, and SAPTA-designed NK- and NN-TALEN pairs have significantly higher endogenous gene-modification rates compared to those designed without SAPTA (2,4,7) (Figure 40), with 71% of SAPTA-designed NK- or NN-TALEN pairs showing gene-modification rates of >20%. Compared with the reported activity distributions of NK- (4) and NN-TALEN (2,7) pairs studied previously, SAPTA-designed NK- and NN-TALEN pairs showed better performance, since the majority of the TALEN pairs gave high mutation rates (Figure 34), whereas the activity distribution of previously tested NK- and NN-TALEN pairs peaks at much lower (1-20%) mutation rates (Figure 34 and Figure 41). Our results indicate that SAPTA-designed TALEN pairs with composite scores >30 generally resulted in high endogenous gene-modification rates (Supplementary Table S5 of reference (3)).

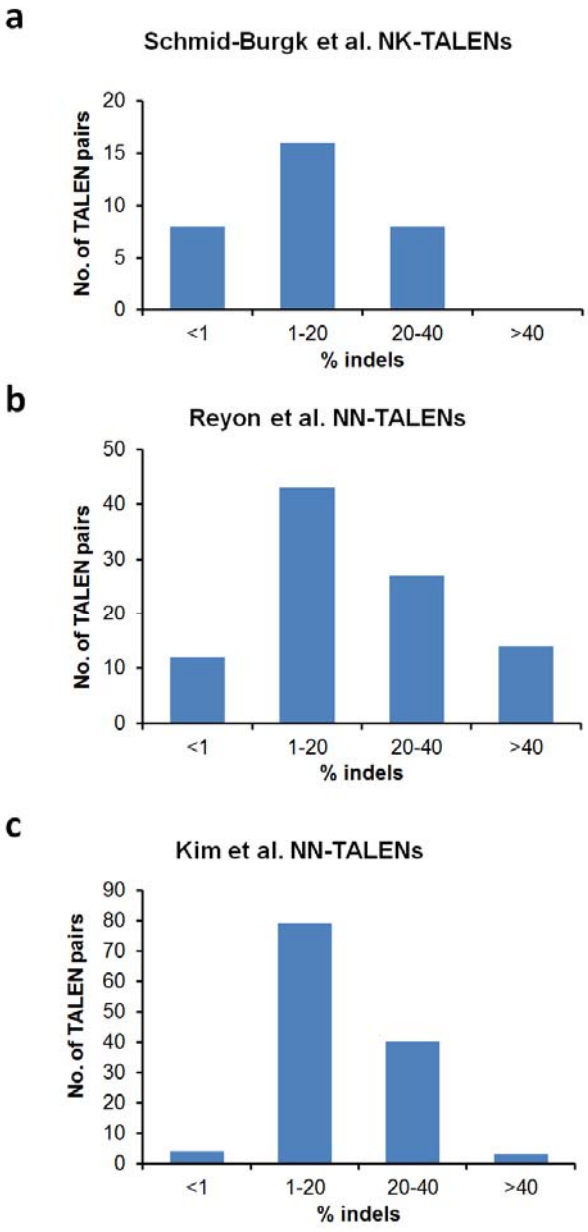


Figure 41: Distribution of TALEN-pair activities in previous publications.
(a) NK-TALEN pairs tested by Schmid-Burgk *et al.(4)*. (b) NN-TALEN tested by Reyon *et al.*(2). (c) NN-TALEN pairs tested by Kim *et al.*(7).

Further, our analysis suggests that SAPTA on average identifies high-scoring target sites within the first 24 bp in a search through the coding sequences of 48 human genes (Figure 42), thus allowing versatile gene editing using TALENs. To test the frequency of high-scoring TALENs in regular human genes, the first 500 bp of the mRNA sequence of the open reading frames of the first 48 genes listed in Reyon *et al.* (2) was obtained from the UCSC Genome Browser. The entire ORF was analyzed if the ORF was shorter than 500 bp. Each sequence was analyised using the SAPTA online targeter and the average distances from the start codon to the closest site scoring above 30, 35, 40, and 45 were noted. The average of these distances for scores 30 and higher was 24 bp with a standard deviation of 31 bp. While this is not as frequent as guidelines published by Reyon *et al*.(2) (every 1/3 bp), it is comparable to the frequency of sites using the guidelines published by Cermak *et al*. (23) (every 35 bp). We were able to find seven pairs of TALENs with scores above 40 in the 200-bp β-globin region centered by the sickle cell mutation. Some target sites in our search were excluded as restrictive constraints were used that limited the ranges of variables to within the values of the training data set. Restrictive constrains may be used as activities of TALENs outside these constraints may not be well predicted by SAPTA, although they can still have high activities. Removing these constraints would result in shorter distances to high-score TALEN pairs.

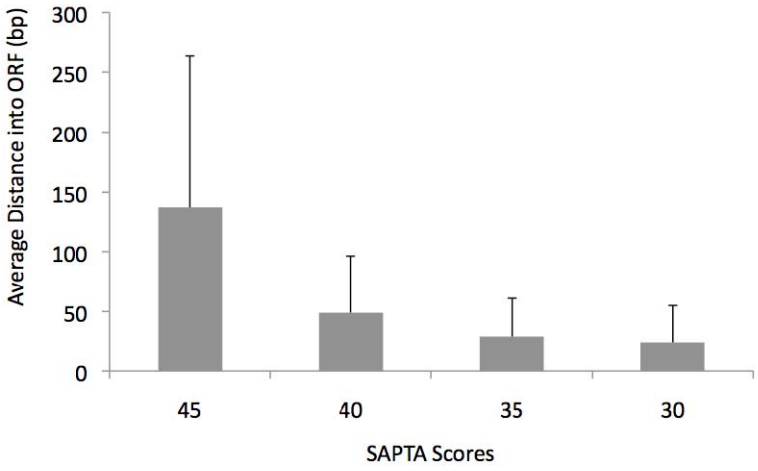


Figure 42: Average frequencies of high-scoring TALEN pair target sites identified by SAPTA.

Our results suggest that the activity of TALENs depends on numerous inter-related variables, and no single variable is able to fully explain the variation in activity levels. Therefore, quantitative integration of various design criteria is likely required for successful TALEN designs. For example, the target site for TALEN S-91 contains 43% C (recognized by the strong RVD HD), but its SSA activity was only 1.7% (3). We believe that multiple design variables act synergistically to affect TALEN activity, and attempted to model these variables collectively using SAPTA.

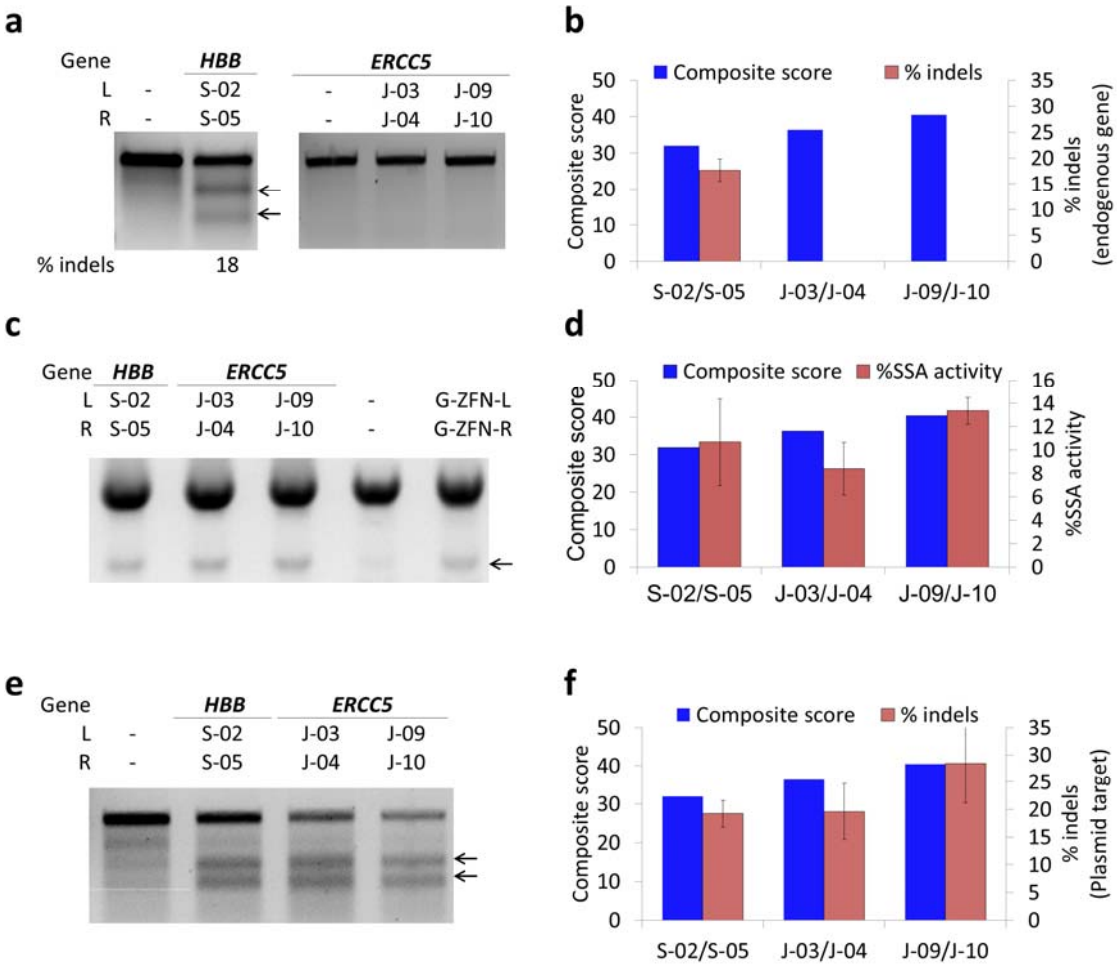


Figure 43: Two pairs of *ERCC5*-directed TALENs in the training set effectively cleaved their plasmid targets, but do not result in endogenous gene targeting.
(a) T7E1 assay at endogenous genes. (b) Comparison between percentages of indels at endogenous gene target sites (quantified from a) and composite SAPTA scores. Error bars, s.e.m. (n=3) (c) Hetero-dimeric SSA activity for two *ERCC5*-targeting TALEN pairs is comparable to the active β-globin (*HBB*) pair. The assay performed, as above, using a hetero-dimeric target plasmid co-transfected with both TALEN monomers. "-" represents a negative control; G-ZFN-L and G-ZFN-R represent the GFP-ZFN pair as in Supplementary Figure S3. Arrow indicates the position of SSA-repaired products. (d) Levels of hetero-dimer %SSA activity (quantified from c) compared with composite SAPTA scores. Error bars, s.e.m. (n=3). (e) T7E1 assay at plasmid targets showed similar activities of *HBB*-targeting and *ERCC5*-targeting pairs. The larger band as shown in b was gel-isolated and used to perform the T7E1 assay, detecting the NHEJ-induced mutagenesis at the plasmid target. (f) Comparison between percentages of indels at plasmid target sites (quantified from e) and composite SAPTA scores. Error bars, s.e.m. (n=2).

The SAPTA algorithm was trained using cleavage activity of TALEN monomers on plasmid substrates to avoid the complications induced by genomic context and other cellular factors, and minimize the variability associated with testing TALENs in pairs. However, the nuclease-induced endogenous gene modification efficiency is determined by the intrinsic activities of TALENs and the genomic context, including epigenetic factors, competing transcription factor binding sites and secondary structures. TALENs with similar activity levels in cleaving plasmid targets may have very different levels of activity when targeting different endogenous genes, as shown in Figure 43. Therefore, quantitative prediction of the rate of endogenous gene modification is challenging, and the SAPTA composite score for TALEN pairs calculated by combining the monomer TALEN sores is a semi-quantitative estimate of TALEN activity. However, TALENs targeting different sites within a short stretch of gene segment are likely to be influenced by similar genomic factors. Therefore, although the effect of TALENs in cleaving endogenous genes might not be accurately modeled by SAPTA due to genomic factors, SAPTA has the potential to rank different (nearby) target sites correctly and help researchers locate the optimal sites within a short gene segment.

SAPTA is based on the sum of scores corresponding to different design considerations. Therefore, it is flexible and will be able to incorporate more design variables into the function, as more information about factors affecting TALEN activity becomes available. The variables and parameters in the SAPTA algorithm can also be adaptively re-trained using new training sets of TALENs with different architectures or using alternate RVDs, such as NH. We anticipate that SAPTA will become a useful and flexible tool for designing highly active TALENs for genome editing applications.

4.4 Materials and methods

Assembly of TALENs

All TALENs were assembled using a modified two-step Golden Gate cloning method (23) to link DNA-binding repeats (plasmids kindly provided by Daniel F. Voytas, University of Minnesota) containing RVDs HD, NI, NG, and NK to recognize C, A, T, and G, respectively, to a wild-type FokI nuclease domain (3). Additional TALENs were assembled using the NN RVD to recognize G. The TALEN backbone vector was constructed by incorporating a Kozak sequence, a triple FLAG epitope tag, and a previously described TALEN framework (24) into the pcDNA3.1(-) vector using NheI and AflII restriction sites (3).

After the first Golden Gate ligation step, intermediate repeat arrays (pFUS) with ≤10 repeats were fully sequenced to confirm correct assembly. In the second Golden Gate step, two or three of these repeat arrays were ligated into the TALEN backbone vector using BsmBI restriction sites to replace a *lacZ* gene stuffer fragment for blue/white screening. The clones were then sequenced using flanking primers to confirm the outside cloning sites and arrays, though the reads do not allow re-sequencing of the middle arrays. The final TALEN plasmids were also validated by digestion with SacII and BamHI to confirm the total size, and digestion with BspEI, which cuts only in HD repeats.

All sequencing primers are described in the Golden Gate TAL Assembly protocol available online at http://www.addgene.org. Complete sequences of all TALEN plasmids and cloning intermediates can be generated using the TAL plasmid assembly website

(http://baolab.bme.gatech.edu/Research/BioinformaticTools/assembleTALSequences.html).

Assembly of single strand annealing (SSA) reporter plasmids

The SSA reporter plasmid backbone contains an EGFP gene, interrupted after 327 bp with a stop codon, the target site for a pair of GFP-targeted ZFNs (63), an AscI, and an SbfI cloning site (3). The downstream portion of the EGFP gene includes a 42-bp region repeating the sequence of the EGFP gene before the stop codon. SSA reporters were constructed using oligonucleotide pairs containing the left target site, the spacer, and the right target site ligated into the vector's AscI and SbfI sites (Figure 12).

SSA activity assay

Human embryonic kidney 293T (HEK293T) cells (ATCC) were cultured in Dulbecco's Modified Eagle Medium (Sigma), supplemented with 10% FBS and 2 mM L-glutamine. Cells were seeded 80,000 per well of a 24-well plate. After 4 h, cells were transfected with 200 ng of the TALEN plasmid (or 100 ng of each TALEN for hetero-dimeric pairs) and 10 ng of the corresponding SSA reporter plasmid using calcium phosphate transfection. Three control transfections were included: 1) 200 ng of an empty TALEN backbone and 10 ng of an SSA reporter plasmid, 2) 200 ng of an empty TALEN backbone and 10 ng of a pEGFP plasmid, and 3) 100 ng of each GFP-ZFN and 10 ng of an SSA reporter plasmid. Cells were harvested 48 h after transfection. The percentages of pEGFP-transfected samples expressing GFP were determined using an Accuri C6 flow cytometer, as an indication of transfection efficiency. Cells were lysed using QuickExtract DNA extraction solution (Epicentre) as described(58). Sample were PCR amplified for 35 cycles (95 °C, 30 s; 60 °C, 30 s; 72 °C, 60 s) in a 50 µl reaction that

contains 2 µl of the cell lysate, 2.5 µl of each 10 µM target region amplification primer (SSA-Cell-F4, 5'-TCGTGACCACCCTGACCTACGG; SSA-Cell-R4, 5'-TGCCGTCCTCGATGTTGTGGCG), and 25 µl of GoTaq green master mix (Promega). PCR reactions were then separated on 2% agarose gels and the percentages of SSA-repaired products relative to the total PCR products were quantified using ImageJ.

Standard curve for SSA assay

To generate the standard curve, EGFP plasmid (pEGFP), with a sequence identical to the SSA-repaired target plasmid, and a target plasmid were mixed at different ratios. HEK293T cells were transfected with the mixtures and an empty TALEN backbone, the genomic DNA harvested and the SSA assay performed, as above. The results from three transfections were averaged and plotted comparing the percentage of the EGFP plasmid versus the percentage of the smaller band (345 bp). Figure 24 shows that the near-linear relationship is valid up to ~50% of EGFP plasmid in the mixture.

SAPTA algorithm

The SAPTA algorithm is a fitted model containing an optimized set of dummy variables and continuous variables. Dummy variables were used to describe base identities of the first five and the last five nucleotides in the monomer target sequence, whereas cubic functions were used to characterize the effect of changes in other variables, including the length of the target sequence, the overall percentages of each nucleotide in the target sequence, percentages of each nucleotide in the first five or the last five nucleotides, and the maximum numbers of consecutive A's and G's (1,50) (Supplementary Table S1 of reference (3)). Cubic functions were chosen because third-

degree polynomials have the flexibility to approximate various curves, including linear, parabola, exponential, and asymmetric concave curves.

A SAPTA score predicts the activity of a TALEN monomer. As shown in Equation 1, the SAPTA score *S* of each monomer target sequence is calculated as the sum of seven terms:

$$S = S_{POS} + S_N + S_{PER} + S_{PER,F5} + S_{PER,L5} + S_{CONS} + C_0 \quad \text{(Eqn. 1)}$$

where C_0 is a constant.

$$S_{POS} = F(d_1)+F(d_2)+F(d_3)+F(d_4)+F(d_5)+F(d_{N-4})+F(d_{N-3})+F(d_{N-2})+F(d_{N-1})+F(d_N) \quad \text{(Eqn. 2)}$$

S_{POS} represents the effects of each nucleotide at the first five and last five positions of the target sequence (e.g. the impact of having a C as the first nucleotide in the target sequence), where

$$F(d_i) = \beta_{i,A}*d_{i,A}+\beta_{i,C}*d_{i,C}+\beta_{i,G}*d_{i,G}+\beta_{i,T}*d_{i,T} \quad \text{(Eqn. 3)}$$

The dummy variable $d_{i,x}$ in Equation 2 is either 1 (if the nucleotide at the position i is x) or 0 (otherwise). Positions are numbered starting from the first nucleotide after the 5'-T. N in Equation 2 denotes the total number of nucleotides in the sequence. For example, d_N corresponds to the last nucleotide at the 3' end of the sequence. Parameters (β_i) associated with the variables (d_i) are optimized as described below, and can be found in Supplementary Table S1 of reference (3).

$$S_N = Q(N) \quad \text{(Eqn. 4)}$$

S_N represents the effect of the length of target sequence (N);

$$S_{PER} = Q(\%A)+Q(\%C)+Q(\%G)+Q(\%T) \quad \text{(Eqn. 5)}$$

S_{PER} represents the effect of the overall base composition (percentages of A, C, G, T);

$$S_{PER,F5} = Q(\%A_{F5})+Q(\%C_{F5})+Q(\%G_{F5})+Q(\%T_{F5}) \quad \text{(Eqn. 6)}$$

$S_{PER,F5}$ represents the effect of the base composition of the first five nucleotides ($\%A_{F5}$, $\%C_{F5}$, $\%G_{F5}$, and $\%T_{F5}$);

$$S_{PER,L5} = Q(\%A_{L5})+Q(\%C_{L5})+Q(\%G_{L5})+Q(\%T_{L5}) \quad \text{(Eqn. 7)}$$

$S_{PER,L5}$ represents the effect of the base composition of the last five nucleotides ($\%A_{L5}$, $\%C_{L5}$, $\%G_{L5}$, and $\%T_{L5}$);

$$S_{CONS} = Q(A_{CONS})+Q(G_{CONS}) \quad \text{(Eqn. 8)}$$

S_{CONS} represents the effect of the maximum numbers of consecutive A's (A_{CONS}) and consecutive G's (G_{CONS}). In equations (4-8), Q(x) is a cubic function defined as

$$Q(x) = ax^3 + bx^2 + cx + d$$

where the values of *a*, *b*, *c* are listed in Supplementary Table S1 of reference (3). The constant terms *d* from each cubic function were combined and solved as the constant C_0 in Equation 1.

In the original functions that form the SAPTA algorithm, there are a total of 55 variables and 86 parameters. Specifically, the 55 variables include 4 x 10 (type of nucleotide x number of positions) dummy variables $d_{i,x}$ which have values of either 1 or 0, and 15 other variables (length of target, %A, %C, %G, %T, etc.) each associated with a cubic function. The 86 parameters include 40 parameters each associated with one dummy variable, 3 parameters for each cubic function (3 x15=45), and one constant the represents all the constants in the cubic functions.

To fully establish the algorithm in SAPTA showed above, 130 NK-TALENs were individually tested for their monomer SSA activity in cultured cells. The experimental results of the Training Set were used for linear regression to determine the parameters in

the SAPTA algorithm using the statistical software R. Since the number of explanatory variables is relatively large compared to the number of TALENs in the Training Set, we follow the standard statistical approach to use the step function in R to perform a model selection that eliminates non-essential variables whose parameters were set to zeros, as shown in Supplementary Table S1 of reference (3). The final functions in the SAPTA algorithm contain 30 variables and 43 parameters.

The fitted SAPTA functions were used to determine the overall optimal base composition shown in Table 2. Specifically, the score contribution of the overall base composition in the first 5 nucleotides was calculated using the following SAPTA function (non-essential variables were removed as described above):

Score contribution of the base composition =

$$1.66E\text{-}04\times(\%A)^3+ (-1.05E\text{-}00)\times(\%A)+ 1.36E\text{-}03\times(\%C)^3+ (-1.42E\text{-}01) \times(\%C)^2$$
$$+ 4.26E+00\times(\%C) +1.18E\text{-}03\times(\%G)^3 +(-6.49E\text{-}02) \times(\%G)^2 +3.01E\text{-}04\times(\%T)^3$$
$$+(-2.91E\text{-}02) \times(\%T)^2 \qquad \text{(Eqn. 9)}$$

The ranges of %A, %C, %G, and %T in the Training Set are 0-56.3%, 14.3-53.3%, 0-45.0%, and 4.2-55.6%, respectively. Thus the equation above was solved using these constraints to find the maximum score contribution. The optimal solutions were found as 0% A, 53% C, 10% G and 37% T.

Composite SSA activity and score

Composite SSA activity was calculated to estimate TALEN pair activity at endogenous genes. The following equation combines square roots of monomer activities measured by the SSA assay, allowing TALEN pairs with more balanced (closer) left and right TALEN activities to obtain a higher calculated composite activity.

$$Composite\ SSA\ activity = 5 + 4 \times \sqrt{left\ \%SSA} + 4 \times \sqrt{right\ \%SSA} \quad \text{(Eqn. 10)}$$

The numerical factors in Equation 10 were chosen so that the composite SSA activity is ~30 when both the left and right TALEN SSA activities are ~10%. Similar to the composite SSA activity, the composite score of a TALEN pair is calculated with the following equation to allow pairs with more balanced left and right TALEN scores to be ranked higher, since TALEN pairs with more balanced left and right monomer activities displayed better performance in pairs (Table 5).

$$Composite\ score = 5 + 4 \times \sqrt{left\ score} + 4 \times \sqrt{right\ score} \quad \text{(Eqn. 11)}$$

The numerical factors in Equation 11 were chosen to be the same as those in Equation 10.

Table 5: Comparison of TALEN pairs with balanced and unbalanced monomer activities.
To determine if TALEN pairs with balanced monomer activities outperform TALEN pairs with drastically different left (L) and right (R) TALEN monomer activities, pair activities were measured using hetero-dimeric target plasmids in an SSA assay. The use of target plasmid minimizes the effect of genomic context. Specifically, TALEN pairs with a similar sum of activities but different left and right monomer activities are compared (e.g. 77+1=78 vs. 53+25=78). Three comparisons with different spacer lengths are shown.

L	R	Spacer	L %SSA	R %SSA	Sum %SSA	Hetero-dimeric %SSA
J-11	J-15	18	54	1	55	6
S-116	S-120	18	30	24	54	18
S-16	S-22	17	31	4	35	6
G-02	J-02	17	18	17	35	19
J-16	J-15	15	77	1	78	3
J-08	C-03	15	53	25	78	7

SAPTA web interface and source code

The Web interface of the SAPTA online search tool can be found at http://baolab.bme.gatech.edu/Research/BioinformaticTools/TAL_targeter.html. The SAPTA web page allows for entering gene segments and basic parameters such as search type, maximum and minimum spacer lengths, and maximum and minimum TAL array lengths. The SAPTA output table contains the starting position of the left TALEN target half-sites, the left and right TALEN sequences, the sizes of the left TALEN, right TALEN and spacer, and the composite score of the TALEN pair. More details on the SAPTA web interface are provided on the website.

The source code for the online SAPTA search tool can be found at: http://baolab.bme.gatech.edu/Research/BioinformaticTools/ScoreTALEBindingSites2.js.

T7 endonuclease I (T7E1) mutation detection assay for measuring endogenous gene modification

The gene modification efficiency of hetero-dimeric TALEN pairs was quantified based on the level of imperfect repair of double-stranded breaks by NHEJ. HEK293T cells were seeded 40,000 per well of a 24-well plate. After 24 hrs, cells were transfected with 500 ng of each nuclease (TALEN or ZFN) plasmid and 10 ng of pEGFP plasmid using 3.4 µl FuGene HD (Promega), following manufacturer's instructions. Cells were harvested 72 h after transfection and analyzed with an Accuri C6 flow cytometer to quantify GFP fluorescence, as a measurement of transfection efficiency. Cell pellets were then collected and genomic DNA isolated using QuickExtract DNA extraction solution (Epicentre), as described in (58). T7E1 assays were performed, as described previously(2) and the digestions separated on 2% agarose gels. The cleavage bands were

quantified using ImageJ. The percentage of gene modification = 100x (1 – (1-fraction cleaved)$^{0.5}$), as described in (58). Primers used for this assay are listed in Supplementary Table S3 of reference (3). All PCR reactions were performed using AccuPrime Taq DNA Polymerase High Fidelity (Life Technologies) following manufacturer's instructions for 35 cycles (94 °C, 30 s; 60 °C, 30 s; 68 °C, 60 s) in a 50 µl reaction containing 2 µl of the cell lysate, 2.5 µl of each 10 µM target region amplification primer, and 5% DMSO. The PCR reactions for the *FANCE* locus gave non-specific bands under standard conditions and were amplified after addition of 1 M betaine.

Single molecule real time (SMRT) sequencing of NHEJ induced mutations

The same PCR products used for T7E1 assays were pooled for SMRT sequencing following the manufacturer's instructions (Pacific Biosciences). NHEJ mutations were detected and analyzed using algorithms recently developed (73) and compared to mock transfected cells.

Fisher's exact test

Fisher's exact test was used to determine the association of using SAPTA with achieving larger percentages of "highly active" TALEN monomers / pairs. For TALEN monomers, a 10% SSA activity was used as a cut-off for high activity. Therefore, TALENs with SSA activities >10% were considered to have "high SSA activity", and those with SSA activities ≤10% to have "low SSA activity". A two-tailed *P*-value was calculated using Fisher's exact test to compare the distribution of high and low SSA activities in the Test Set 2 and the Training Set. To evaluate the performance of TALEN pairs tested in this work, we took into account the variability of T7E1 assay in previously studies (2,4,7). In each case, we used fractions of the maximum indel percentage

observed in this particular study as the cut-off values. Furthermore, to avoid arbitrarily choosing cut-off, we applied sliding cut-offs at 20%, 30%, 40%, 50%, 60%, 70%, and 80% of the maximum indel percentage. TALENs with indel percentages higher than a cut-off value were considered to be "highly active". Two-tailed *P*-values were calculated for each cut-off value.

.

CHAPTER 5: CHARACTERIZATION OF CRISPR/CAS9 OFF-TARGET EFFECT AT GENOMIC SITES WITH INSERTIONS OR DELETIONS

CRISPR/Cas9 systems are a versatile tool for genome editing due to the highly efficient targeting of DNA sequences complementary to their RNA guide strands. However, it has been shown that RNA-guided Cas9 nuclease cleaves genomic DNA sequences containing mismatches to the guide strand. A better understanding of the CRISPR/Cas9 specificity is needed to minimize off-target cleavage in large mammalian genomes. Here we show that genomic sites could be cleaved by CRISPR/Cas9 systems when DNA sequences contain insertions ('DNA bulge') or deletions ('RNA bulge') compared to the RNA guide strand, and Cas9 nickases used for paired nicking can also tolerate bulges in one of the guide strands. Variants of single-guide RNAs (sgRNAs) for four endogenous loci were used as model systems, and their cleavage activities were quantified at different positions with 1-bp to 5-bp bulges. We further investigated 114 putative genomic off-target loci of 27 different sgRNAs and confirmed 15 off-target sites, each harboring a single-base bulge and one to three mismatches to the guide strand. Our results strongly indicate the need to perform comprehensive off-target analysis related to DNA and sgRNA bulges in addition to base mismatches, and suggest specific guidelines for reducing potential off-target cleavage.

5.1 Introduction

Advances with engineered nucleases allow high-efficiency, targeted gene editing in numerous organisms, primary cells and cell lines. Gene editing was used to create user-defined cells, model animals and gene-modified stem cells with novel characteristics that can be used for gene functional studies disease modeling and therapeutic applications. Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) proteins constitute a bacterial defense system that cleaves invading foreign nucleic acids (33-40). Chimeric single-guided RNAs (sgRNAs) based on CRISPR (41) have been engineered to direct the Cas9 nuclease to cleave complementary genomic sequences when followed by a 5'-NGG protospacer-adjacent motif (PAM) in eukaryotic cells (17,42,43). Since gene targeting by CRISPR/Cas9 is directed by base pairing, such that only the short 20-nt sequence of the sgRNA needs to be changed for different target sites, CRISPR/Cas systems enable simultaneous targeting of multiple DNA sequences and robust gene modification (17,18,41,42,44-48).

Endogenous DNA sequences followed by a PAM sequence can be targeted for cleavage by designing a ~19-nt sequence of the sgRNA complementary to the target. However, other sequences in the genome may also be cleaved non-specifically, and such off-target cleavage by CRISPR/Cas systems remains a major concern. Generally speaking, there is a partial match between the on- and off-target sites, and the differences between the on- and off-target sequences can be grouped into three cases: (a) same length but with base mismatches; (b) off-target site has one or more bases missing ('deletions'); (c) off-target site has one or more extra bases ('insertions'). Recent studies have shown that CRISPR/Cas9 systems non-specifically cleave genomic DNA sequences containing base-pair mismatches (case a) generating off-target mutations in mammalian cells with

considerable frequencies (5,51-55). Mismatches in the PAM sequence are less tolerated, although Cas9 also recognizes an alternative NAG PAM with low frequency (52,54,82). In addition, Cas9 off-target cleavage at a similar gene sequences with a base-pair mismatch may lead to gross chromosomal deletions with high frequencies, as demonstrated by the deletion of the 7-kb sequence between two cleavage sites in *HBB* and *HBD* respectively (5). These results indicate that, although Cas9 specificity extends past the 7-12 bp seed sequence (52,53), off-target effects may limit the applications of Cas9-mediated gene modification, especially in large mammalian genomes that contain multiple DNA sequences differing by only a few mismatches. A recent report revealed that 99.96% of the sites previously assumed to be unique Cas9 targets in human exons may have potential off-target sites containing a functional (NAG or NGG) PAM and one single-base mismatch compared with the on-target site (54).

In this work we investigated the above-mentioned cases (b) and (c) of potential CRISPR/Cas9 off-target cleavage in human cells by systematically varying sgRNAs at different positions throughout the guide sequence to mimic insertions or deletions between off-target sequences and RNA guide strand. To avoid confusion, for single-base insertions, we use a 'DNA bulge' to represent the extra, unpaired base in the DNA sequence compared with the guide sequence. Similarly, for single-base deletions, we use a 'RNA bulge' to represent the extra, unpaired base in the guide sequence compared with the DNA sequence (Figure 44). Therefore, adding a base into the guide RNA would result in an RNA bulge, while removing a base in the guide strand can be used to model a DNA bulge. The cleavage activity of RNA-guided Cas9 at endogenous loci in HEK293T cells transfected with plasmids encoding Cas9 and sgRNA variants was quantified as the

NHEJ-induced mutation rates. We found that off-target cleavage resulted from the sgRNA variants occurred with DNA bulge or sgRNA bulge at multiple positions in the guide strands, sometimes at levels comparable to or even higher than those of original sgRNAs. We further examined the Cas9-mediated mutagenesis at 80 potential off-target loci in the human genome carrying single-base DNA bulges or sgRNA bulges together with a range of base mismatches, and confirmed two off-target sites with mutation frequencies up to 45.5%. Our results clearly indicate the need to search for genomic sites with base-pair mismatches, insertions and deletions compared with the guide RNA sequence in analyzing CRISPR/Cas9 off-target activity and in designing RNA guide strands for targeting specific genomic sites.

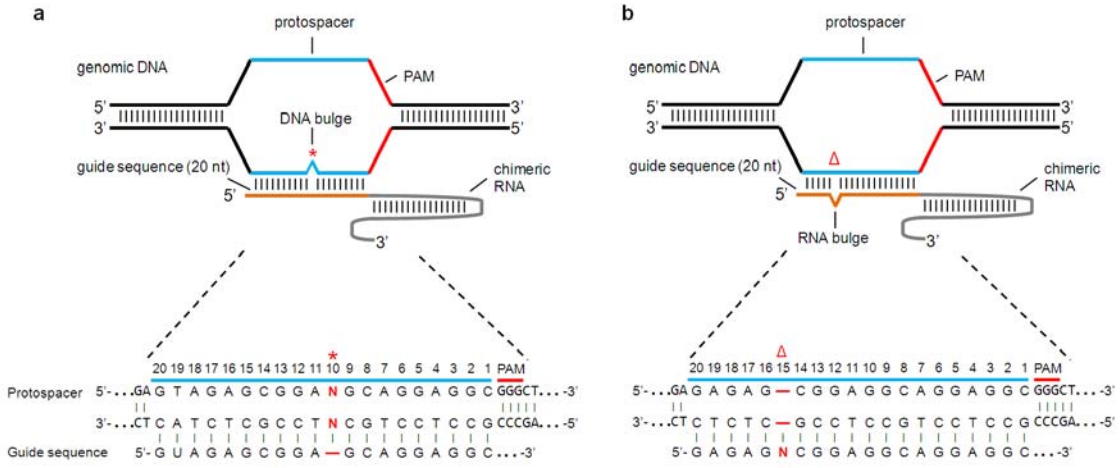


Figure 44: Schematic of CRISPR/Cas9 off-target sites with (a) 1-bp insertion (DNA bulge) or (b) 1-bp deletion (RNA bulge).
The 20-nt guide sequence (orange line) in the sgRNA is shown with genomic target sequence (protospacer) containing single-base DNA bulge (red asterisk) or single-base sgRNA bulge (red Δ). The zoom-in sequences of protospacer and PAM are shown above the sgRNA guide sequence. Positions of nucleotides in the target are numbered 3' to 5' starting from the nucleotide next to PAM.

5.2 Results

Cas9 cleavage with sgRNA variants containing single-base DNA bulges

To determine if CRISPR/Cas9 systems tolerate genomic target sites containing single-base DNA bulges (Figure 44a), we used the sgRNA-DNA interfaces of two sgRNAs, R-01 and R-30, targeting the *HBB* and *CCR5* genes, respectively as a model system (5). Systematically removing single nucleotides at all possible positions throughout the original 19-nt guide sequences of R-01 and R-30 resulted in single-base DNA bulges at their original *HBB* and *CCR5* target sites that model single-base insertion at potential off-target sites in the genome (Figure 45).

Cleavage of the genomic DNA in HEK293T cells was quantified using the T7E1 mutation detection assay. For both groups of sgRNA variants (generated from R-01 and R-30 respectively), single-base DNA bulges at certain positions in the DNA sequences were well tolerated (e.g., still had Cas9 induced cleavage), though variants of R-30 had higher cleavage activity at more locations (Figure 45).

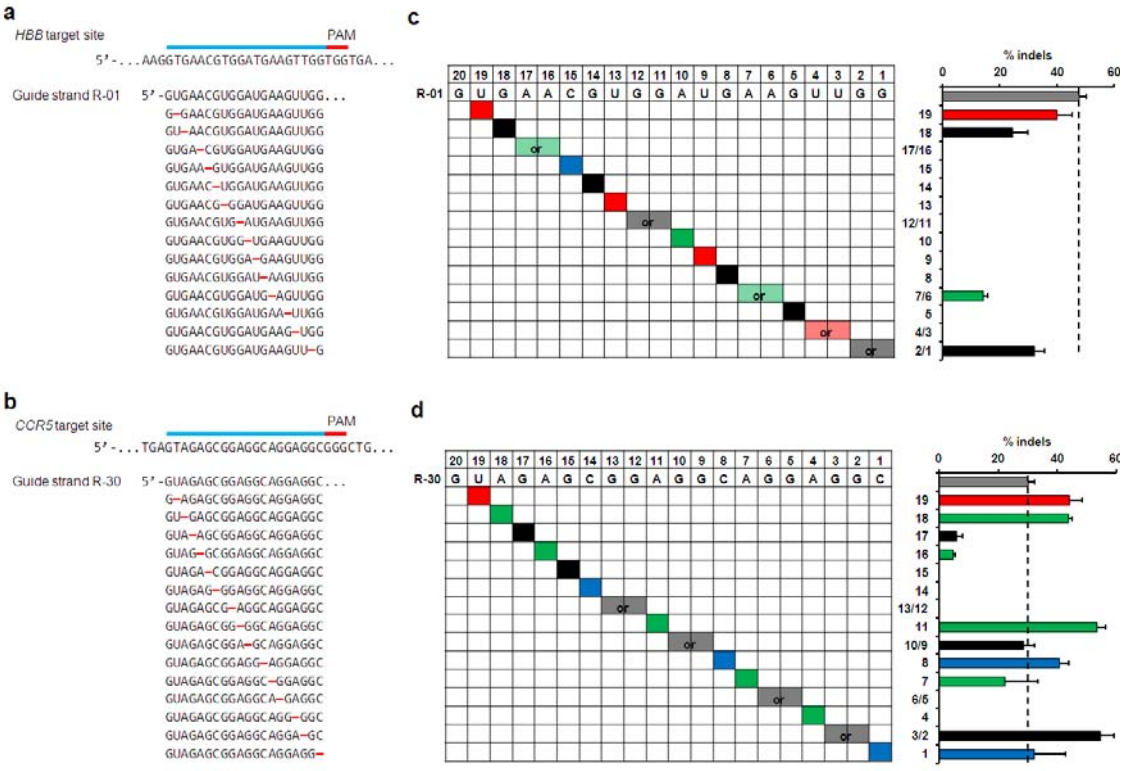


Figure 45: Activity of sgRNA variants targeted to genomic loci containing single-base DNA bulges. A single nucleotide was deleted from the original sgRNA at all possible positions (red dashes) throughout the guide sequence for (a) sgRNA R-01 targeting *HBB* or (b) sgRNA R-30 targeting *CCR5*. Cleavage activity for the corresponding sgRNA variants measured by T7E1 assay in HEK293T cells at (c) the *HBB* site or (d) *CCR5* site for the sgRNA variants in (a) and (b). Sequence of the original sgRNA is in the top row of the grid. Positions of the deleted nucleotides are highlighted for A (green), G (black), C (blue), or U (red) in the grid. Semi-transparent colors in two positions in the same sgRNA indicate that deletions can be interpreted at either of adjacent positions (also marked by "or") due to identical nucleotides at both positions. The bar graph on the right shows cleavage activity aligned to the corresponding sgRNA variants using the same color scheme. Positions relative to PAM are labeled on the y-axis. The vertical dashed lines mark the activity levels of the original sgRNAs. Error bar, s.e.m. (n=2).

For both groups, it was clear that Cas9 tolerated DNA bulges in target sites in three regions: 7 bases from PAM, the 5'-end (PAM-distal), and the 3'-end (PAM-proximal). Specifically, -1nt variants of R-01 induced Cas9 cleavage activity when a single-base DNA bulge is present at positions 1 or 2, 6 or 7, 18, and 19 of the target DNA

sequence from the PAM (Figure 45). Note that due to the presence of consecutive identical nucleotides at positions 1 and 2, 6 and 7, removing either one of the identical nucleotides in the sgRNA at these adjacent positions would give the same sequence and have the same sgRNA-DNA interface (their position is therefore marked as "or" in Figure 45). In contrast, -1nt variants of R-30 induced variable cleavage activity at more positions throughout the guide sequence: positions 1, 2 or 3, 7, 8, 9 or 10, 11, 16, 17, 18, and 19 from the PAM (Figure 45). Seven R-30 variants have activities comparable to or even higher than that of the original sgRNA. These variants correspond to DNA bulges at positions 1, 2 or 3, 8, 9 or 10, 11, 18, and 19 from the PAM (Figure 45). Consistent with previous studies showing that the specificity of CRISPR/Cas9 systems is guide-strand and target-site dependent (5,51,52), the positions in R-01 sgRNA variants where DNA-bulges were tolerated are different from that in R-30 sgRNA variants. However, these positions seem to group in the 5'-end, middle, and 3'-end regions of the target loci, as in both R-01 and R-30 sgRNA-DNA interfaces, single-base DNA-bulges at the following five positions seems to be tolerated: positions 1, 2, 7, 18, and 19. Although additional studies are needed to determine if these positions are common for different target sequences, single-base DNA-bulges at the target sites corresponding to these positions may be worth investigating when performing off-target analysis for CRISPR/Cas9 systems.

In certain cases, off-target sites with DNA bulges may also be interpreted as sequences having various base mismatches with guide sequence and/or PAM (6). For example, the sgRNA-DNA interfaces corresponding to removing 5'-end bases in the guide sequences (positions 18 and 19 of the R-01 interface, and 16-19 of the R-30

interface) can be viewed as having DNA bulges or having mismatches in the 5'-end region of sgRNA, which have been shown to be better tolerated compared to the 3'-end region (17,51,52). Therefore, the Cas9 cleavage activities induced by these guide strands may be interpreted as tolerance of base mismatches at the 5'-end of the guide RNA. In addition, the position -1 variant of R-30 results in a shift in the adjacent PAM from GGG to CGG (another canonical PAM), which could explain why the activity of this guide sequence variant was similar to the original R-30. However, off-target activities associated with most other DNA bulges for the R-01 and R-30 interfaces cannot be attributed to base mismatch tolerance, since a base removal in the sgRNAs (corresponding to a DNA bulge) could result in many base mismatches or mutation in the PAM sequence. For example, the cleavage activity induced by the R-01 variant at position 2/1 may be alternatively interpreted as Cas9 cleavage with a GTG PAM (Figure 45), which is highly unlikely according to previous studies (52,53). Further, a R-30 guide strand variant at position 11 would contain at least seven mismatches if modeled without a bulge. This guide strand resulted in a 1.8-fold higher cleavage activity compared to the original R-30 (Figure 45), which cannot be readily explained by the high level of base mismatches (which should prohibit cleavage), and thus should be attributed to the tolerance of DNA bulges.

Cas9 cleavage with small sgRNA truncations

We further investigated if sgRNAs with small truncations at the 5'-end retain cleavage activity. One to six nucleotides were deleted from the 5' end of R-01 except for the nucleotide at position 20, because the guanine here is required for the expression under the U6 promoter (Figure 46). For these guide sequence truncations, we found that

1- to 2-bp 5’ truncations could still induce cleavage activities similar to the full-length sgRNA (Figure 46).

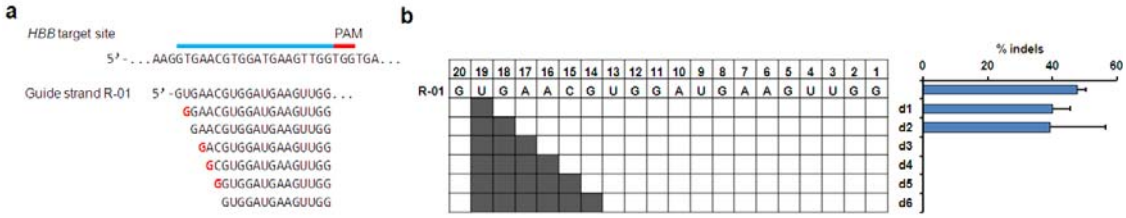


Figure 46: Activity for sgRNAs containing 5’-end truncations.
(a) 1-6 bp truncations at the 5’ end of the guide sequence R-01 targeted to the *HBB* gene. (b) Activity for truncated sgRNAs. Truncated positions are highlighted in gray in the grid. Bar graph shows corresponding cleavage activity measured by T7E1 assay in HEK293T cells. Error bar, s.e.m. (n=2).

Cas9 cleavage with sgRNA variants containing single-base sgRNA bulges

In addition to Cas9 induced cleave at off-target sites with single-base DNA bulges, we further investigated if single-base sgRNA bulges (that model single-base deletions in DNA sequence) could induce Cas9 cleavage (Figure 44b). Again, using sgRNA-DNA interfaces R-01 and R-30 as model systems, we systematically added single nucleotides at positions throughout the original guide sequences, so that the interfaces with target sequences in *HBB* or *CCR5* carries single-base sgRNA bulges (Figure 47 and Figure 48). For some positions, the addition of single nucleotide A, C, G and U respectively to the guide sequence was all tested to account for the effect of base identity.

As above, HEK293T cells were transfected with plasmids of the Cas9 and sgRNA variants, and the T7E1 mutation detection assay was used to measure the Cas9 cleavage activity.

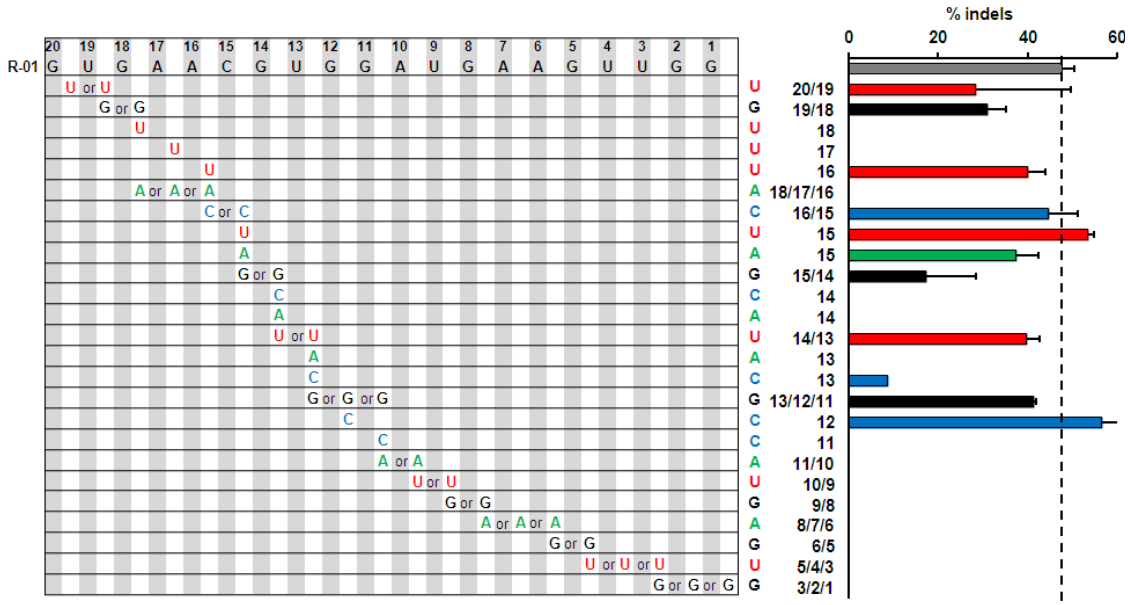


Figure 47: Activity of R-01 sgRNA variants targeted to genomic locus of *HBB* to make single-base sgRNA bulges.
Single nucleotide, A (green), G (black), C (blue), or U (red), was inserted into the original sgRNA throughout the guide sequence. Sequence of the original sgRNA is in the top row of the grid. Positions of the original guide sequence are shaded in gray, while the inserted positions are white. Due to identical nucleotides at adjacent positions, some inserted nucleotides can be in multiple positions (marked by "or"). Bar graphs on the right show corresponding cleavage activities quantified by T7E1 assay in HEK293T cells, with the same color scheme for different inserted nucleotides. Positions relative to PAM and the single nucleotides added are labeled on the y-axis. Error bar, s.e.m. (n=2).

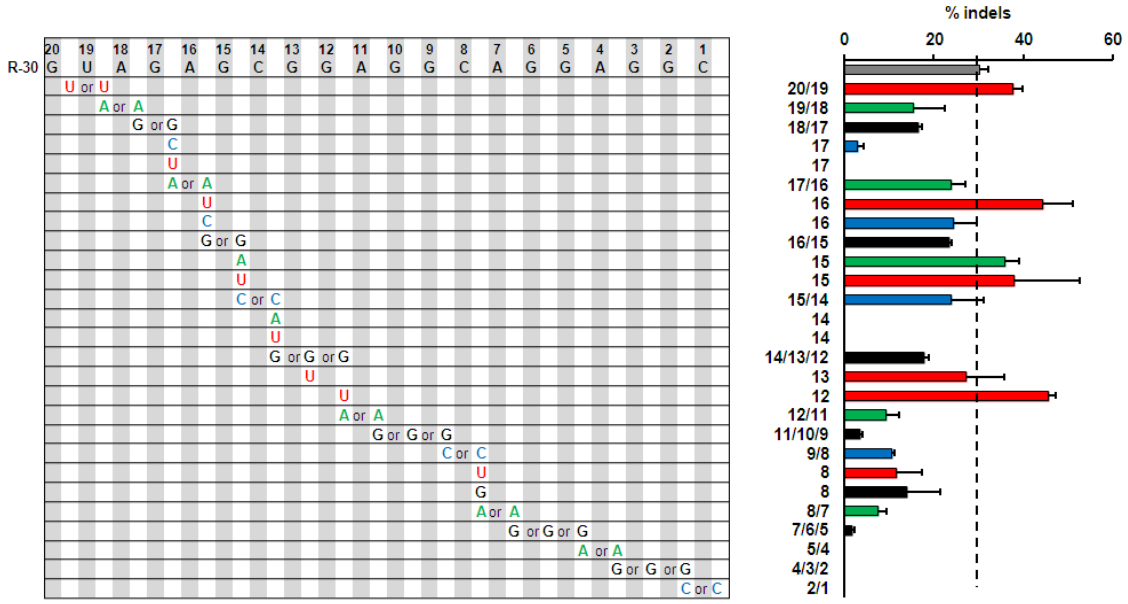


Figure 48: Activity of R-30 sgRNA variants targeted to genomic locus of *CCR5* to make single-base sgRNA bulges.
See the figure above for explanation.

We found that sgRNA bulges in the R-30 sgRNA-DNA interface were better tolerated compared to those of R-01. In contrast to the tolerances of DNA bulges adjacent to the PAM, sgRNA bulges close to the PAM prohibited cleavage. For the R-01 interface, single-base sgRNA bulges between each of the 11 PAM-proximal guide-strand nucleotides resulted in no detectable activity (Figure 47). Single-base sgRNA bulges of the four nucleotides closest to the PAM in R-30 also eliminated T7E1 activity (Figure 48). The sgRNA bulges 3' to the position 11 in R-30 resulted in reduced cleavage activities (Figure 48). The lack of activity with PAM-proximal sgRNA bulges in R-01 and low levels of activity with PAM-proximal sgRNA bulges in R-30 are consistent with the reduced mismatch tolerance in the "seed sequence" reported in previous studies

(17,41,83). Nucleotides additions in sgRNA sometimes created consecutive identical nucleotides, such as adding a G before or after position 14 of R-01 or before or after position 15 of R-30. These sgRNA variants model a G-bulge that can be at either position in the sgRNA (Figure 47). We found that in many cases sgRNA bulges with a single U gave rise to high nuclease activities. Among all sgRNA variants with activities higher than the original sgRNAs, ~71% (5/7) were targeted to the loci with a U-bulge. Overall, single-base sgRNA bulges induced higher Cas9 cleavage activities at many more positions than that with single-base DNA bulges. This is not surprising since RNA molecules are more flexible than DNA molecules, thus having smaller binding energy penalty with single-base RNA bulges, resulting in a higher tolerance (84).

RNA-DNA interfaces with single-base RNA bulges can also be viewed as sequences with various mismatches in the guide sequence and PAM (6). Specifically, sgRNA bulges at the 5'-end of guide RNA sequences (e.g. U+20/19 for R-01 and R-30 interfaces) can be alternatively viewed as having one to a few base mismatches with the 3'-end of DNA sequences (6), which are often tolerated, similar to deletions of 1-2 bp at the 5' end of guide strands (Figure 46). SgRNA bulges close to the 3'-end of guide sequence can be alternatively viewed as having base mismatches in the 3'-end region, including those at the third base of PAM (R-30 variants) (the last six variants in Supplementary Figure S2 of reference (6)). Among all sgRNA variants with considerable activities, most of them could not be explained by tolerance of base mismatches, since they would contain more than five mismatches or change in the third base of PAM, which was shown to abolish cleavage activity (52).

a

Gene	Index	Protospacer Target (5' to 3')	PAM	%indels at target	%GC
HBB	R-01	GTGAACGTGGATGAAGTTGG	TGG	45%	50%
HBB	R-08	GCTGTGGGGCAAGGTGAACG	TGG	25%	65%
CCR5	R-30	GTAGAGCGGAGGCAGGAGGC	GGG	30%	70%
CCR5	R-25	GTGTTCATCTTTGGTTTTGT	GGG	27%*	35%

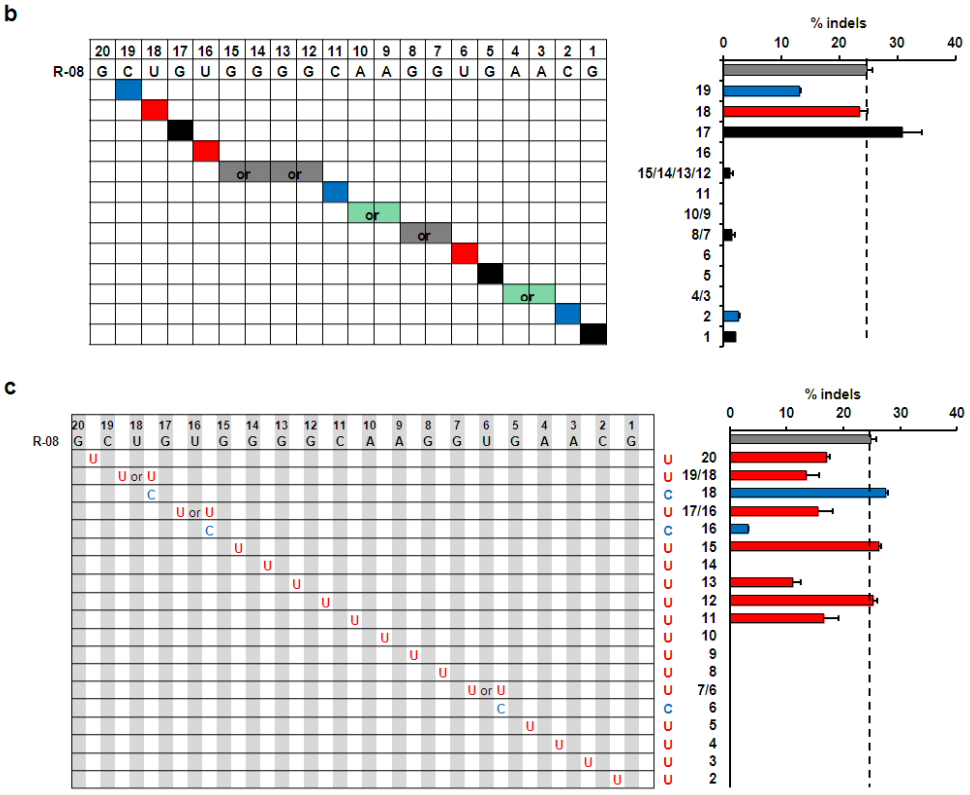


Figure 49: Activity of sgRNA variants with bulges targeted to genomic loci with different GC contents.
(a) Target sites, cleavage activities (% indels by T7E1 assay), and GC contents of different guide strands targeted to *HBB* and *CCR5* genes. Cleavage activity of R-25 is from reference (5). (b-c) T7E1 Activity of R-08 variants targeted to *HBB* genomic loci with (B) single-base DNA bulges or (c) single-base sgRNA bulges. Color schemes and labels are similar to the figures above. Error bar, s.e.m. (n=2).

The effect of GC content of sgRNAs on the tolerance of single-base sgRNA bulges

As revealed in our study, the specificity profile (location and level of off-target cleavage) of R-01 variants is substantially different from that of R-30 variants. R-30,

which showed a higher level of tolerance to DNA and RNA bulges than R-01, has a GC content of 70%, whereas R-01 has a GC content of 50%. We hypothesized that the GC content of guide strands R-01 and R-30 played a significant role in causing this difference. To investigate this hypothesis, we tested two additional sets of guide strands targeted to *HBB* and *CCR5* genes respectively, with different GC contents compared to R-01 and R-30 (Figure 49). Specifically, R-08 has a moderately higher GC content compared to R-01 (65% compared to 50%), whereas the GC content of R-25 is half of that of R-30 (35% compared to 70%). Cas9 induced cleavage with sgRNA variants of R-08 and R-25 was individually tested to quantify the bulge tolerance in HEK 293T cells.

For the guide strand R-25, which contains a low percentage of GC, we found that all R-25 variants tested showed non-detectable activities using the T7E1 assay (Supplementary Table S2 of reference (6)). In contrast, for R-08 variants with bulges throughout the guide sequence, we observed cleavage activities at more positions compared with R-01 (Figure 49). These results of bulge tolerance for variants of R-08 and R-25 support our GC dependence hypothesis.

Cas9 cleavage with sgRNA variants containing 2-bp to 5-bp bulges

In addition to single-base bulges between sgRNA and target sequence, it is important to determine if bulges longer than 1 bp can also be tolerated by the CRISPR/Cas9 systems. Consequently, the tolerance of 2-bp to 5-bp bulges was tested at locations where single-base bulges were well tolerated. For sgRNA bulges, we added two to five U's 15-bp or 12-bp upstream of PAM into the guide sequences of R-01 and R-30, respectively. To generate DNA bulges, we deleted two bases from the guide sequences of R-01 and R-30 (Figure 50a). Strikingly, we found that sgRNA variants forming 2-bp, 3-

bp, and 4-bp RNA bulges induced cleavage activities as determined by the T7E1 assay in HEK 293T cells (Figure 50b). Since sgRNA variants forming 2-bp DNA bulges did not show any detectable activity, we did not test longer DNA bulges. Our findings that sgRNA bulges of > 2-bp are better tolerated than DNA bulges of similar size are consistent with the higher cleavage activities by guide strands with 1-bp sgRNA bulges compared to those with 1-bp DNA bulges as shown in Figure 45, Figure 47, and Figure 48.

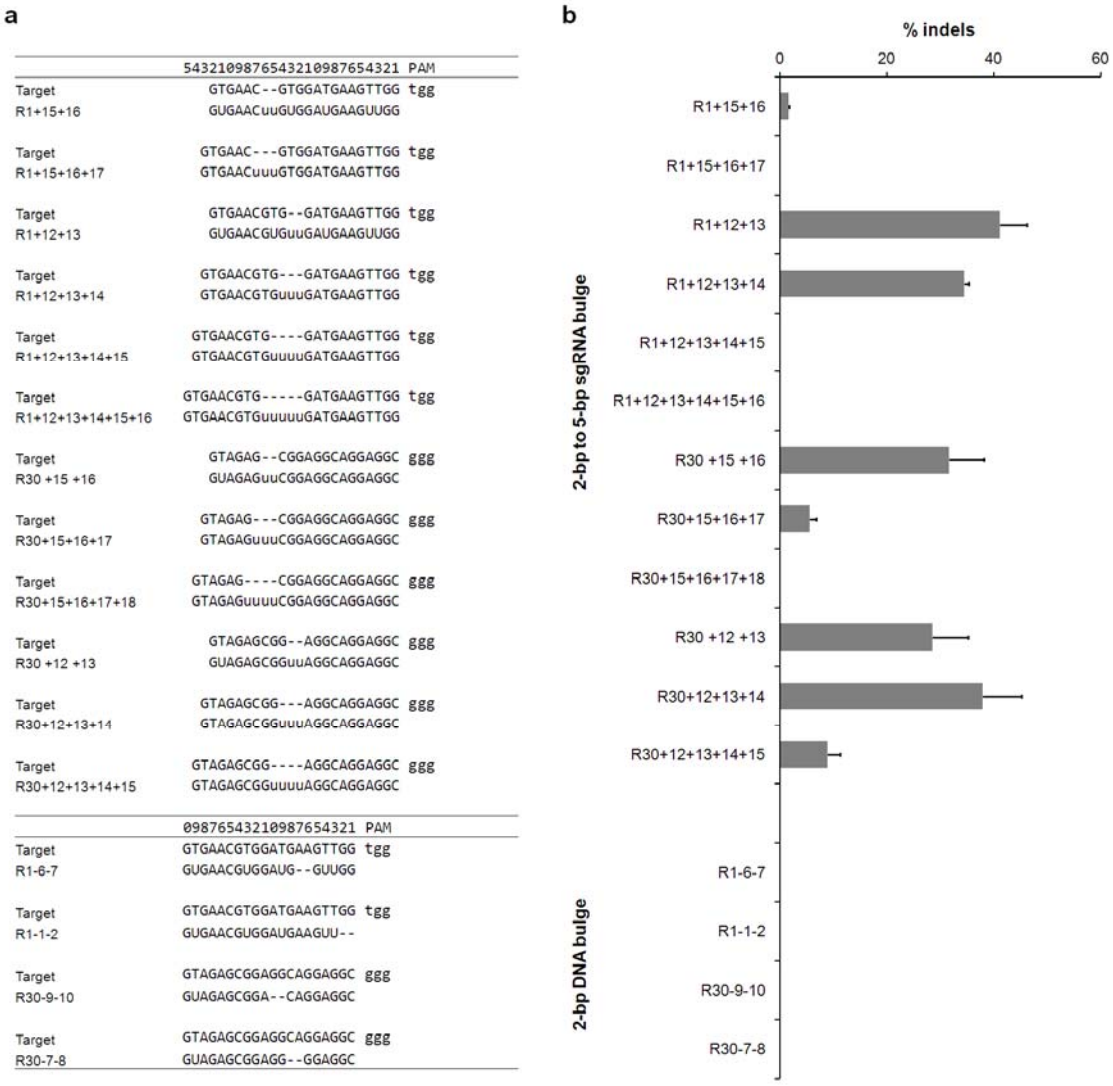


Figure 50: Activity of sgRNA variants with 2-bp DNA or 2-bp to 5-bp sgRNA bulges.
Guide strands with 2-bp to 5-bp addition are labeled with "+" and positions of the added bases, and guide strands with 2-bp deletion are labeled with "-" and positions of the deleted bases. (a) Sequences comparison of guide RNAs and target sites, with position numbers on top. (b) Bar graph showing cleavage activities of these sgRNA variants quantified by T7E1 assay in HEK293T cells. Error bar, s.e.m. (n=2).

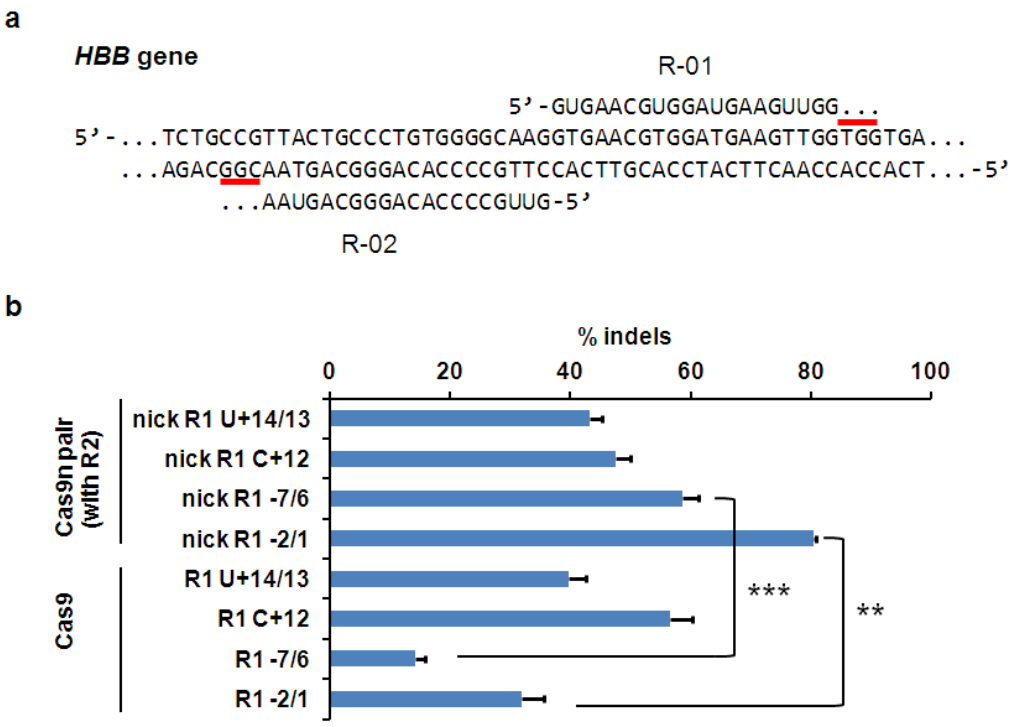


Figure 51: Paired Cas9 nickases with one bulge-containing sgRNA effectively cleave genomic DNA. (a) Human HBB gene targeted by Cas9 nickases (Cas9n) with paired guide strands R-01 and R-02. PAMs are indicated with red bars. (b) T7E1 activities of Cas9n with R-01 bulge-variants paired with R-02, compared with original Cas9 activities of the R-01 bulge-variants as in Figure 2 and Figure 4. Error bar, s.e.m. (n=2). Asterisks indicate P-values from a two-tailed independent two-sample t-test. *P < 0.05, **P < 0.01, ***P < 0.001.

Cleavage by paired Cas9 nickases with sgRNA variants containing single-base bulges

Paired Cas9 nickases (Cas9n) were recently developed to generate DNA double-strand breaks by inducing two closely spaced single-strand nicks using an appropriately designed pair of guide RNAs (54,85). This strategy may lower the off-target cleavage, as DSBs could occur only when both guide RNAs of the pair result in nearby nicks at the

same time. Here we tested if paired Cas9n systems can tolerate bulges by using one bulge-forming guide variant paired with a perfectly matched guide strand. Specifically, four variants of R-01 showing high activities with Cas9 were paired with R-02, including R1 U+14/13 and R1 C+12 to test sgRNA bulges, and R1 -7/6 and R1 -2/1 to test DNA bulges. Each paired sgRNAs created a 34-bp 5' overhang in the *HBB* gene (Figure 51) (5), and the Cas9n cleavage activities were determined by the T7E1 assay. We found that both sgRNA and DNA bulges were well also tolerated in the Cas9n system (Figure 51). The paired Cas9 nickases with single sgRNA bulges showed activities comparable to Cas9 system having one bulge in R0-1; however, for DNA bulges, the activities of paired Cas9 nickases were >2-fold higher than Cas9.

Cas9 cleavage at genomic loci with both base mismatches and DNA or sgRNA bulges

To gain a better understanding of CRISPR/Cas9 off-target activity, we examined 27 different sgRNAs targeting 6 different genes (Table 10), 7 targeted *HBB*, 2 for *EGFP*, 5 for *CCR5*, 7 for *ERCC5*, 4 for *TARDBP* and 2 for *HPRT1* respectively. We performed off-target analyses of these sgRNAs by searching the human genome for potential off-target sites and found that for the sgRNAs searched, single-base DNA or sgRNA bulges were not located without mismatches in the human genome. Therefore, for each sgRNA, we selected a subset of the potential sites with one to three mismatches and avoided mismatches close to the PAM as much as possible. All of these sgRNAs efficiently induced mutations at their intended target loci in human HEK293T cells, as measured by the T7E1 assay (Figure 52). Using the T7E1 assay, we initially investigated 18 potential

off-target sites containing target-site insertions and 62 containing deletions (Supplementary Table S4 of reference (6)).

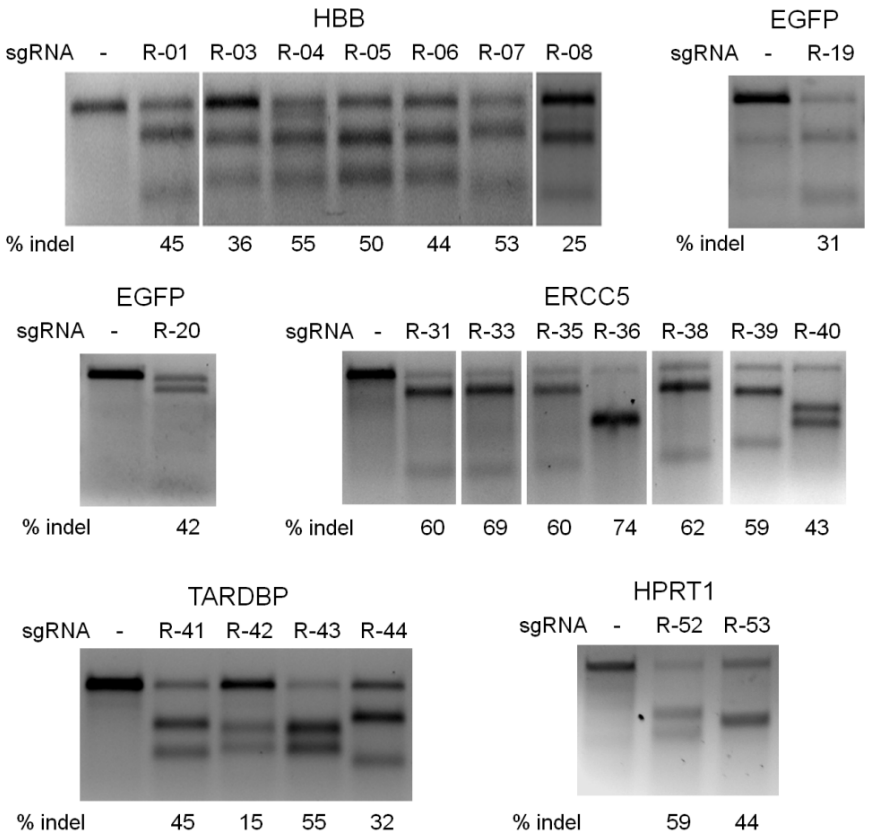


Figure 52: T7E1 assay measuring the on-target endogenous gene modification efficiency of sgRNAs in HEK293T cells.
Lane headings indicate the target genes and the sgRNA index names. "-" denotes samples transfected with a stuffer plasmid. Numbers below each lane having detectable activity show the percentage of modified alleles. Primers used for the PCR amplification are listed in Supplementary Table S3 of reference (6).

Cas9 induced cleavage activities in *CCR5* and *ERCC5* respectively were detected for two off-target sites each bearing one DNA bulge and one mismatch (Figure 53). For R-30, the identified off-target site R-30 Off-4 contains a single-base DNA bulge at position 5, 6, or 7 and a base mismatch at position 14. The off-target gene modification rate determined by T7E1 is 9%, almost one third of the 30% on-target activity at the *CCR5* gene (Figure 53). For an R-31 off-target site with a single-base DNA bulge at position 2 and a mismatch at position 20, the off-target gene modification rate determined by T7E1 was 3%, compared to 60% on-target activity at the *ERCC5* gene (Figure 53). Due to the high frequency of small indels that result from repair of Cas9 induced cleavage, which may be poorly detected by the T7E1 assay, we verified the mutagenesis at these off-target sites using Sanger sequencing (Figure 53). For both off-target sites, the mutation frequencies quantified by Sanger sequencing are higher than those by T7E1, which is consistent with a previous study (5). We did not observe any off-target cleavage for the 62 sites tested with both sgRNA bulge and base mismatch, although in our model systems with sgRNA bulges only, high cleavage activities were observed (Figure 47 and Figure 48). This discrepancy suggests that sites forming sgRNA bulges may be less tolerant to additional base mismatches, and vice versa.

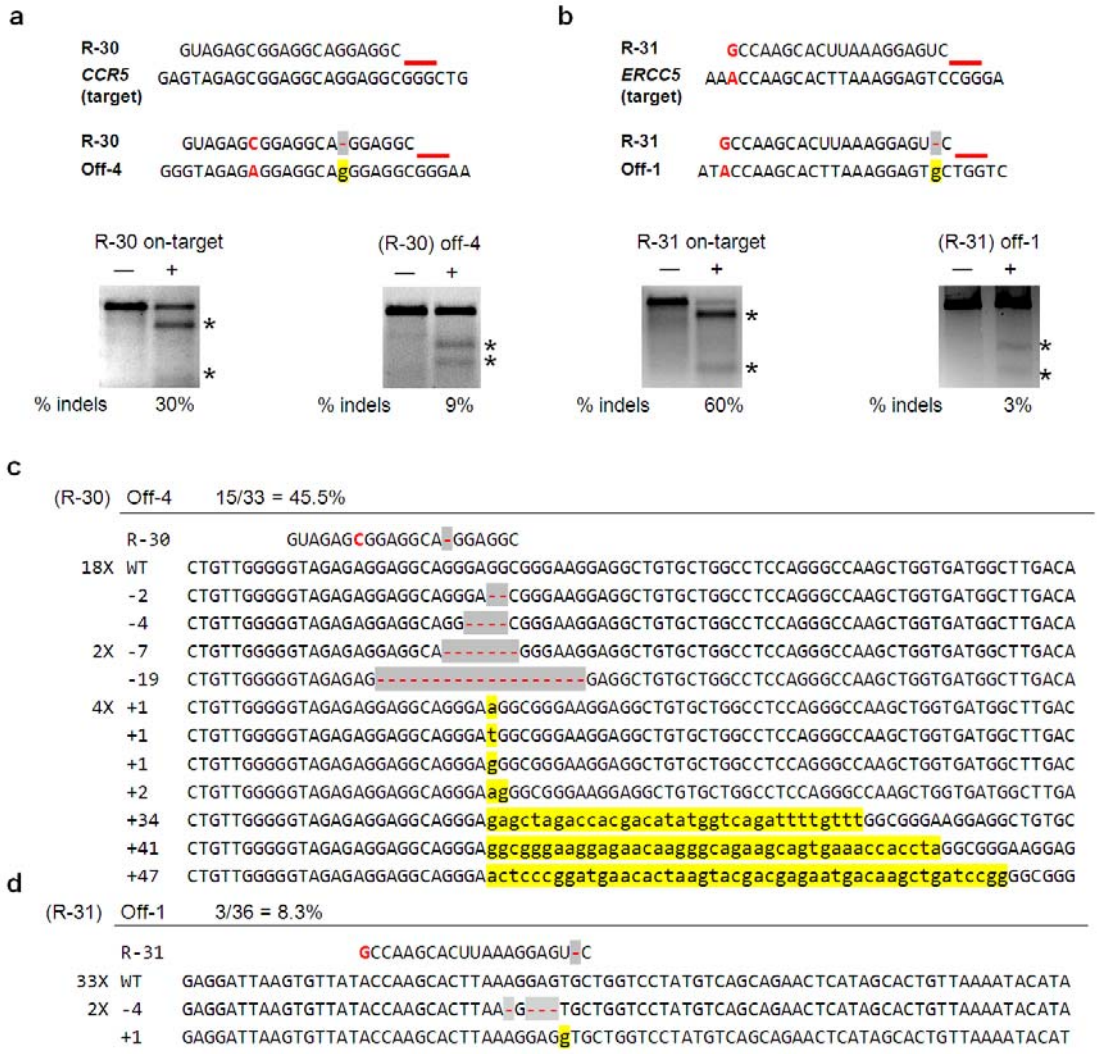


Figure 53: Activities of CRISPR/Cas9 nucleases for genomic target sites and for off-target sites with single-base DNA bulges coupled with mismatches.
(a-b) On-target and off-target cleavage activities for (a) sgRNAs R-30 targeted to *CCR5* gene, and (b) R-31 target to *ERCC5* gene. Upper: target sequences (*CCR5* and *ERCC5*) and off-target sequences (Off-4 and Off-1) with mismatch (red) and DNA bulge (shaded in yellow) are shown next to the sgRNA (R-30 and R-31) tested. Red lines indicate the PAM. Bottom: Cleavage activities at the target sites and off-target sites measured by T7E1 assay in HEK293T cells. Numbers below the lanes indicate average percentages of gene modification (n=2). Asterisks indicate specific T7E1 cleavage products. (c-d) Sanger sequencing reads of amplified off-target sites aligned to the wild-type genomic sequence and sgRNAs for (c) R-30 and (d) R-31. The occurrence of each sequence is indicated to the left of the alignment, if greater than one. Unmodified reads are indicated by "WT". Deletions are marked in gray, and insertions marked in yellow.

Two genomic off-target sites for guide strand R-30, Off-4 and Off-5, have identical target sequences (Figure 54), but were cleaved at different rates. Specifically, R-30 Off-4 had a cleavage rate of 9%, while the cleavage at Off-5 was undetectable with the T7E1 assay (Figure 54). Sanger sequencing revealed a 45.5% mutation rate at the R-30 Off-4 locus (Figure 53c), compared to a 4.2% mutation rate at R-30 Off-5 (Figure 54). Since R-30 Off-4 and R-30 Off-5 sites have identical sequences, our results clearly suggest that off-target cleavage of Cas9 nuclease is very dependent on genomic context (5). Further investigation of these two sites using the ENCODE annotation from UCSC genome browser (86,87) revealed that R-30 Off-4, which had high off-target activity, targeted a site within 400 bp of the 3' end of a long non-coding RNA (RP4-756H11.3) and 12kb of the protein-coding gene *RABGEF*. Analysis of the ENCODE data for chromatin structure in normal human embryonic kidney cells (NHEK) cells, the cell type of origin for the HEK293 cells used in this study shows Off-4 to be within 3kb of a strong enhancer (marked by H3K27Ac and H3K4me1) and a strong DNAse1 hypersensitive site, suggestive of an open chromatin structure. In contrast, R-30 Off-5, which had low activity, targeted a site in a 162-kb intergenic region between the *WBSCR28* and *ELN* genes that is marked by the more heterochromatic H3K27me3, and hence may be less accessible for Cas9 induced cleavage (Figure 55). Taken together, these data strongly suggest that differences in the local chromatin structure may underlie the observed differences in cleavage efficiency between Off-4 and Off-5.

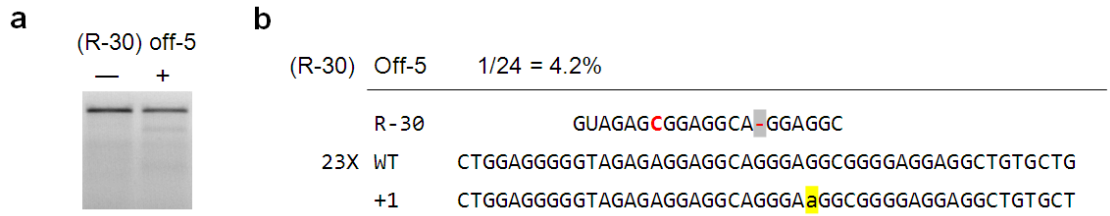


Figure 54: Off-target cleavage of R30 Off-5 quantified by (a) T7E1 assay and (b) Sanger sequencing. "-" and "+" denote samples treated without and with nuclease, respectively. The occurrence of each sequence is indicated to the left of the alignment, if greater than one. Unmodified reads are indicated by "WT". Deletions are marked in gray, and insertions marked in yellow.

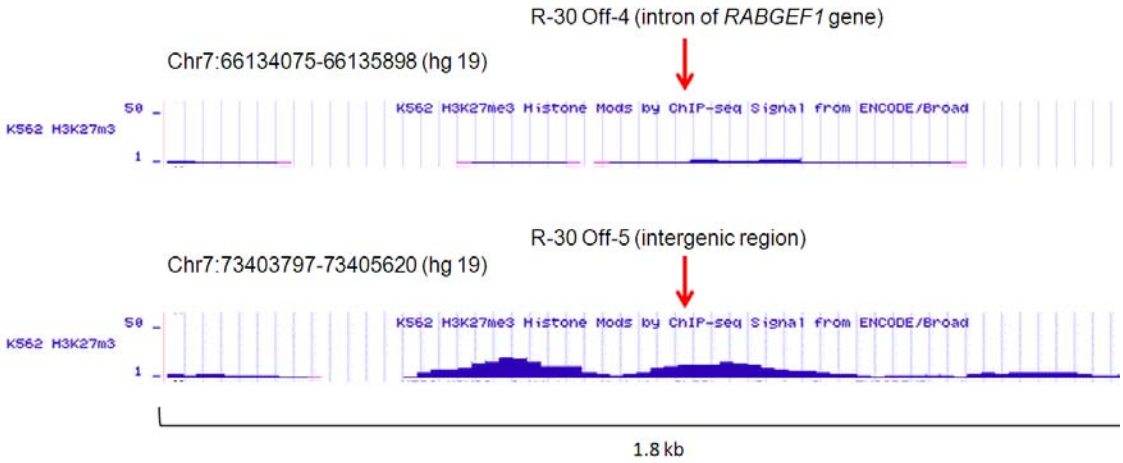


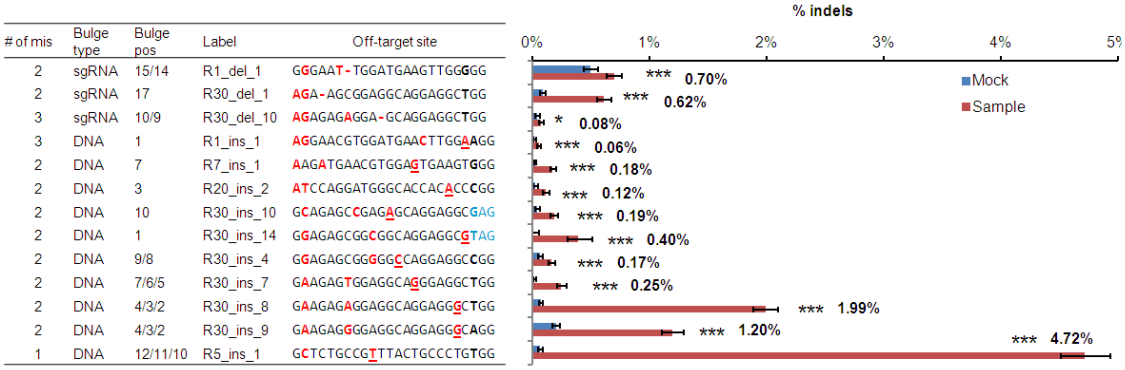
Figure 55: Histone modification status and annotation of R30 Off-4 and Off-5 loci obtained from the UCSC genome browser.

We further performed deep sequencing at 55 putative off-target sites corresponding to single-base sgRNA bulges and 21 sites corresponding to single-base

DNA bulges. The sites were amplified from genomic DNA harvested from HEK 293T cells transfected with Cas9 and sgRNAs (Supplementary Table S6 of reference (6)). The 55 sites with sgRNA bulges contain 35 sites tested in the preliminary T7E1 assay, and the 21 sites with DNA bulges include seven sites tested in the T7E1 assay. Putative bulge-forming loci containing one to three PAM-distal mismatches were chosen, since we did not find sites associated with a bulge without any base mismatch. We also selected some of the bulge-forming sites with a high level of sequence similarity, but containing an alternative NAG-PAM. For comparison, the deep sequencing also investigated 16 on-target sites of the sgRNAs tested. Each locus was sequenced from mock-transfected cells as control.

We identified additional 13 bulge-forming off-target sites with significant cleavage activities resulted from CRISPR/Cas9 systems compared to the mock-transfected samples (Figure 56). We found that the number of genomic off-target cleavage sites associated with sgRNA bulges was relatively small (some of these cases are indistinguishable from a few mismatches at 5' end), but there was considerable activity at genomic sites with DNA bulges coupled with one to three additional base mismatches, even with an alternative NAG-PAM. Similar results showing more off-target effect with DNA bulges plus mismatches compared to sgRNA bulges plus mismatches were observed in the preliminary T7E1 assay (Figure 53). The positions of these tolerated DNA bulges are 1-3 bp and 7-10 bp from PAM, consistent with the results from the model systems using sgRNA variants. The majority of the sites with off-target activities detected, as shown in Figure 53 and Figure 56 are associated with the sgRNA

R-30, which has a high GC content (70%). Other sgRNAs that resulted in off-target cleavage at bulge-forming loci have GC content ≥50%.



# of mis	Bulge type	Bulge pos	Label	Off-target site
2	sgRNA	15/14	R1_del_1	GGGAAT-TGGATGAAGTTGGGGG
2	sgRNA	17	R30_del_1	AGA-AGCGGAGGCAGGAGGCTGG
3	sgRNA	10/9	R30_del_10	AGAGAGAGGA-GCAGGAGGCTGG
3	DNA	1	R1_ins_1	AGGAACGTGGATGAACTTGGAAGG
2	DNA	7	R7_ins_1	AAGATGAACGTGGAGTGAAGTGGG
2	DNA	3	R20_ins_2	ATCCAGGATGGGCACCACACCCGG
2	DNA	10	R30_ins_10	GCAGAGCCGAGAGCAGGAGGCGAG
2	DNA	1	R30_ins_14	GGAGAGCGGCGGCAGGAGGCGTAG
2	DNA	9/8	R30_ins_4	GGAGAGCGGGGGCCAGGAGGCCGG
2	DNA	7/6/5	R30_ins_7	GAAGAGTGGAGGCAGGAGGCTGG
2	DNA	4/3/2	R30_ins_8	GAAGAGAGGAGGCAGGAGGGCTGG
2	DNA	4/3/2	R30_ins_9	GAAGAGGGGAGGCAGGAGGGCAGG
1	DNA	12/11/10	R5_ins_1	GCTCTGCCGTTTACTGCCCTGTGG

Figure 56: Significant activities analyzed by deep sequencing at genomic off-target loci containing bulges coupled with mismatches and alternative NAG-PAM.
Only bulge-containing off-target loci determined to have *P*-values less than 0.05 are shown. Table on the left shows numbers of mismatches at off-target loci in addition to bulge (# of mis), bulge types, positions of bulges from PAM (bulge pos), labels for the loci as in Supplementary Table S6 of reference (6), and sequences of off-target sites including PAMs. In these off-target genomic sequences, mismatches are marked by red, deleted base compared to sgRNA marked as "-" (sgRNA bulge), inserted base compared to sgRNA marked as underlined red letters (DNA bulge), NAG-PAMs are marked by blue. Bar graph on the right indicates indel percentages quantified for mock (blue) and treated samples (red) with sgRNAs at off-target loci shown in the table to the left. Error bars, Wilson intervals (Materials and Methods). *$P \leq 0.05$, ***$P \leq 0.001$ as determined by Fisher's exact test.
The % indel values of treated samples are also indicated.

5.3 Discussion

Although CRISPR/Cas9 systems can efficiently induce gene modification in many organisms, recent studies revealed that off-target cleavage may occur in mammalian cells with up to five-base mismatches between the short ~20-nt guide RNA

and DNA sequences (5,51-53). Here we show that CRISPR/Cas9 systems can have off-target cleavage when DNA sequences have an extra base (DNA bulge) or a missing base (sgRNA bulge) at various locations compared with the corresponding RNA guide strand. Importantly, our results revealed that, sgRNA bulges of up to 4-bp could be tolerated by CRISPR/Cas9 systems (Figure 50). The correlation between cleavage activity and the position of DNA bulge or sgRNA bulge relative to the PAM appears to be loci and sequence dependent when comparing the specificity profiles of guide sequences R-01 and R-30.

Our results suggest the need to perform comprehensive off-target analysis by considering cleavage due to DNA and sgRNA bulges in addition to base mismatches. We believe that the following design guidelines will help reduce potential off-target effects of CRISPR/Cas9 systems: (i) conservatively choose target sequences with relatively low GC contents (e.g. ≤35%), (ii) avoid target sequences (with either NGG- and NAG-PAM) with ≤3 mismatches that form DNA bulges at 5' end, 3' ends or around 7-10 bp from PAM, and (iii) if possible, avoid potential sgRNA bulges further than 12 bp from PAM. To aid the rational design of sgRNAs for an intended DNA cleavage site, as well as experimental determination of off-target activity, a robust bioinformatic tool that incorporating these design guidelines and ranking potential off-target sites is desired, and more extensive studies of off-target cleavage by CRISPR/Cas9 systems may be needed concerning the dependence of off-target activity on the type (base mismatch, DNA bulge, sgRNA bulge), location and length of sequence differences.

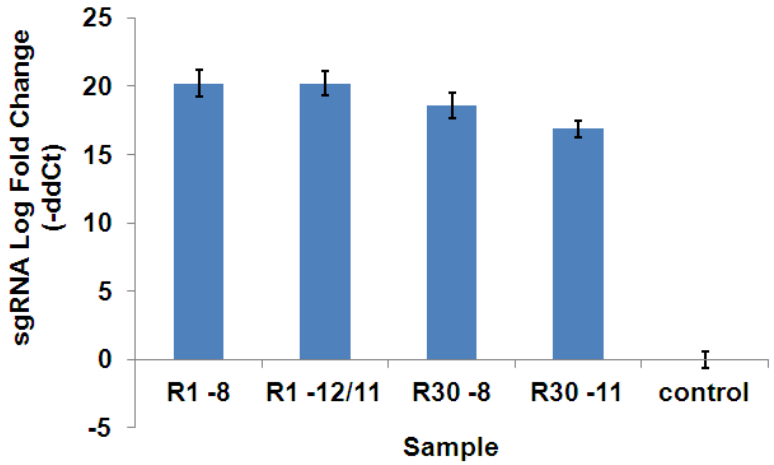


Figure 57: Quantitative PCR of sgRNA expression levels in HEK293T cells for R-01 and R-30 variant.
Relative expression of these sgRNAs was quantified using the ddCt method (see "Material and Methods").

We found that different specificity profiles of R-01 and R-30 guide sequences (and variants) are not due to different expression levels of the sgRNAs. Quantitative PCR of inactive R-01 variants and active R-30 variants indicated similar sgRNA expression levels (Figure 57). We believe that high GC-content, which makes the RNA/DNA hybrids more stable (88), may be responsible for increased tolerance of DNA bulges and sgRNA bulges. Consistent with our hypothesis, guide strand R-30 (70% GC) showed the highest tolerance to sgRNA and DNA bulges among the four guide strands we tested (R-01, R-08, R-25 and R-30), while guide strand R-25 (35% GC) does not seem to tolerate any bulges. Guide sequences showing bulge-related off-target activity in Figure 53 and Figure 56 all have GC contents ≥50%, which further confirms that it is important to

consider DNA-bulges for sgRNAs with high GC content, even with up to three base mismatches, when investigating off-target effects.

As shown in Supplementary Figure S1 and S2 of reference (6), bulges in the PAM distal or PAM proximal regions can reflect either mismatch tolerance or RNA/DNA bulge tolerance. Off-target sites containing these bulges may also be present in the output of a search considering mismatches only. In either scenario, these sites should be tested for off-target cleavage. A recent study reveals that a Cas9 ortholog from *Streptococcus thermophilus* has a PAM located 2 bps downstream of the protospacer (89). Instead of forming a DNA-bulge, the cleavage resulting from the variant R-01 -2/1 (Supplementary Figure S1 of reference (6)) may alternatively reflect the tolerance of a linker between the target sequence and PAM. Further studies are needed to distinguish between these two scenarios. On the other hand, bulges in the middle of the target sequence are likely to reflect only the tolerance of RNA or DNA bulges. Since there would be too many mismatches if bulges are not formed, potential off-target sites with bulges in the middle are not likely to be included in the output of a search allowing mismatches only. Though the mechanisms for bulge tolerance has yet to be determined, there is a need to investigate potential off-target sites from a search considering mismatches and bulges.

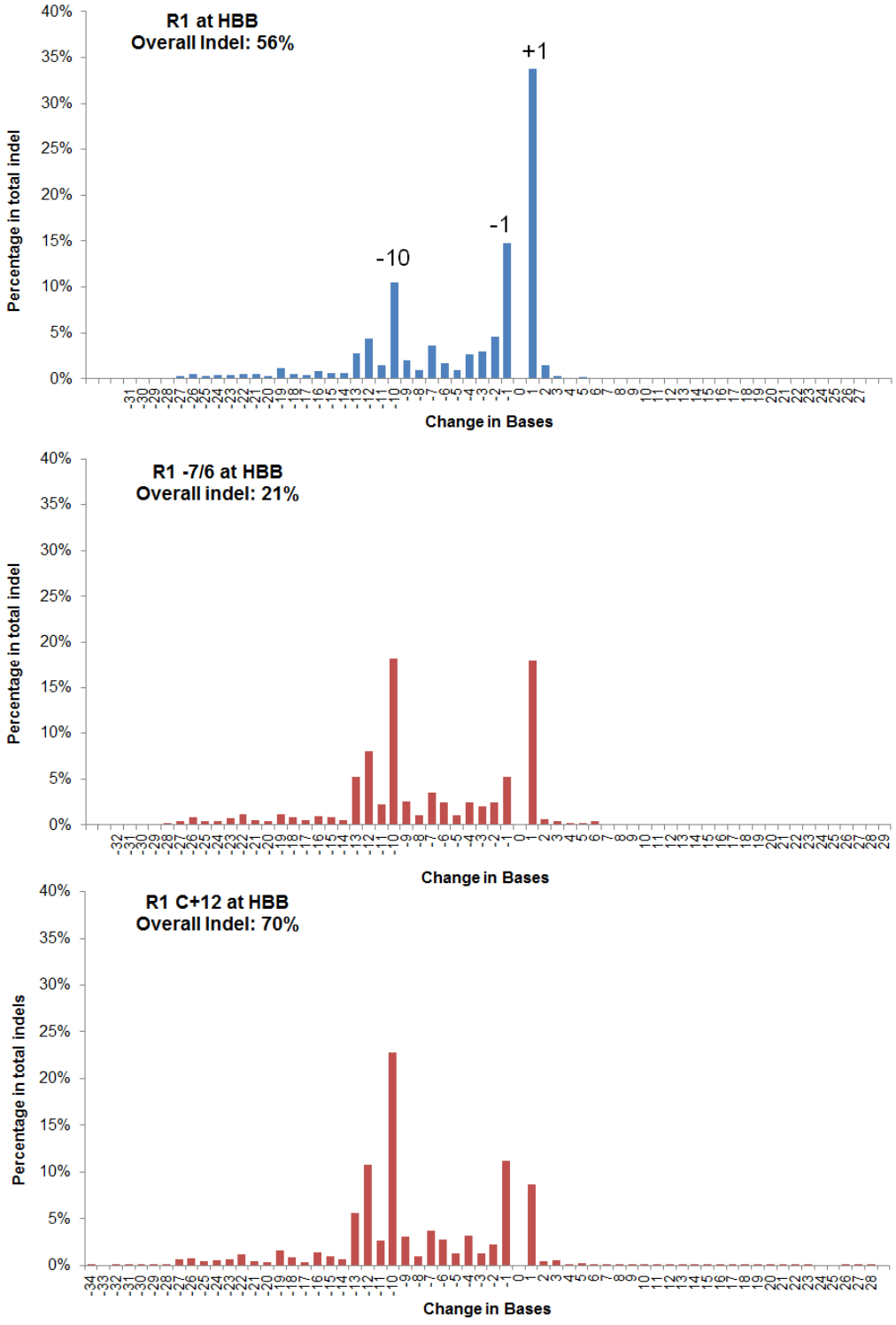


Figure 58: Indel spectra for original sgRNAs and sgRNA variants of R-01 determined using deep sequencing.
The change in bases at predicted cut sites resulting from indicated sgRNAs was calculated from ~10^4 reads per sample. The y-axis represents percentages in all indel-reads for that sgRNA. Overall % indel in total reads are indicated in each graph.

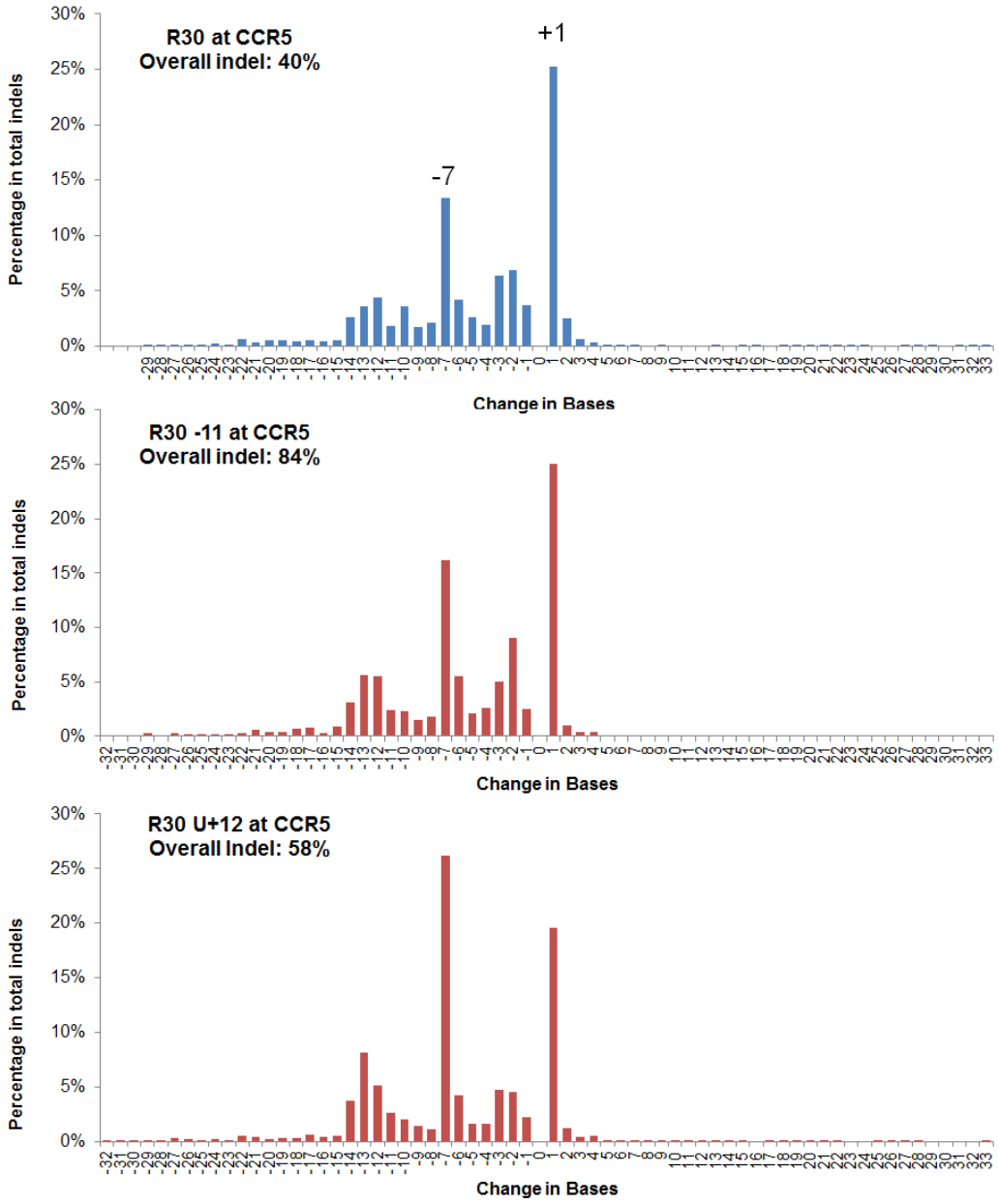


Figure 59: Indel spectra for original sgRNAs and sgRNA variants of R-30 determined using deep sequencing.
The change in bases at predicted cut sites resulting from indicated sgRNAs was calculated from ~10^4 reads per sample. The y-axis represents percentages in all indel-reads for that sgRNA. Overall % indel in total reads are indicated in each graph.

An interesting finding from this study is that sgRNA variants with bulges had different indel spectra than sgRNA without bulges (Figure 58 and Figure 59). We quantified indel spectra for original sgRNAs R-01 and R-30, as well as sgRNA variants R1 -7/6, R1 C+12, R30 -11, and R30 U+12, using deep sequencing with around 10^4 reads for each sample. Bulge-forming sgRNA variants showed higher ratios of larger deletions (-10 or -7), whereas the original sgRNAs without bulges generate mostly 1-bp insertions. This effect is more prominent for variants forming sgRNA bulges (R1 C+12 and R30 U+12). Bulge-forming sgRNA variants may be more effective than regular sgRNAs at creating larger deletions that may be preferred in certain applications , such as targeted disruption of genomic elements.

Recently, paired Cas9 nickases have been shown to increase target specificity of CRISPR/Cas9 systems. However, only off-target activity associated with single guide RNAs were investigated (54,85), and the effect of cooperative nicking at potential off-target sites with sequence similarity to a pair of guide RNAs has not been characterized. We showed that Cas9n is able to cleave efficiently at target sites despite a single-base bulge in one of the paired guide RNAs. The results of this work provide some insight into off-target cleavage of the paired Cas9 nickases, since nicking of opposite DNA strands is likely to be independent events, and the knowledge of bulge tolerance at the sgRNA-DNA interface would be applicable to off-target cleavage of Cas9 nickases.

Recent studies on the specificity of CRISPR/Cas9 systems revealed that a broad range of partial matches between sgRNA and DNA sequences could induce off-target cleavage (5,51-53), which may limit the choice of sgRNA designs. While the use of bioinformatic tools for predicting potential off-target loci based on sequence similarity

may be useful, our results suggest that partially matched sequences including base mismatches, deletions and insertions needs to be considered. Since there might be a large number of potential off-target sites due to the many partially matched sequences, and the effect of sgRNA-DNA sequence differences on off-target cleavage is target-site and genome-context dependent, experimentally determining true off-target activities is necessary, including the use of deep sequencing.

5.4 Materials and methods

CRISPR/Cas9 plasmid assembly

DNA oligonucleotides containing a G followed by a 19-nt guide sequence (Table 10) were kinased, annealed to create sticky ends, and ligated into the pX330 plasmid that contains the +85 chimeric RNA under the U6 promoter and a Cas9 expression cassette under the CBh promoter (kindly provided by Dr. Feng Zhang; it is also available at Addgene) (90). Variants of sgRNAs were constructed and tested with 1 or more nucleotides inserted or deleted (Supplementary Table S2 of reference (6)). The annealed oligonucleotides have 4-bp overhangs that are compatible with the ends of BbsI-digested pX330 plasmid. Constructed plasmids were sequenced to confirm the guide strand region using the primer CRISPR_seq (5'-CGATACAAGGCTGTTAGAGAGATAATTGG-3').

T7 endonuclease I (T7E1) mutation detection assay for measuring endogenous gene modification rates

The cleavage activity of RNA-guided Cas9 at endogenous loci was quantified based on the mutation rates resulting from the imperfect repair of double-stranded breaks by NHEJ. In a 24-well plate, 60,000 HEK293T cells per well were seeded and cultured in

DMEM media supplemented with 10% FBS and 2 mM fresh L-glutamine, 24 hours prior to transfection. Cells were transfected with 750 ng (sgRNA variants) or 1000 ng of CRISPR plasmids using 3.4 µl FuGene HD (Promega), following manufacturer's instructions. Each sgRNA plasmid was transfected as biological duplicates in two separate transfections. All subsequent steps, including the T7E1 assay were performed independently for the duplicates. A HEK293T-derived cell line containing stably integrated EGFP gene was used for sgRNAs targeted to the EGFP gene. This cell line was constructed by correcting the mutations in the EGFP gene in the cell line 293/A658 (91) (kindly provided by Dr. Francesca Storici). The genomic DNA was harvested after 3 days using QuickExtract DNA extraction solution (Epicentre), as described in (58). T7E1 mutation detection assays were performed, as described previously (2) and the digestions separated on 2% agarose gels. The cleavage bands were quantified using ImageJ. The percentage of gene modification = 100x (1 – (1-fraction cleaved)$^{0.5}$), as described (58). Unless otherwise stated, all PCR reactions were performed using AccuPrime Taq DNA Polymerase High Fidelity (Life Technologies) following manufacturer's instructions for 40 cycles (94 °C, 30 s; 60 °C, 30 s; 68 °C, 60 s) in a 50 µl reaction containing 1.5 µl of the cell lysate, 3% DMSO, and 1.5 µl of each 10 µM target region amplification primer (Supplementary Table S3 of reference (6)) or off-target region amplification primer (Supplementary Table S4 of reference (6)).

Sanger sequencing of gene modifications resulted from Cas9

To validate the mutation rates measured by T7E1 assay, the PCR products used in the T7E1 assays were cloned into plasmid vectors using TOPO TA Cloning Kit for Sequencing (Life Technologies) or Zero Blunt TOPO PCR Cloning Kit (Life

Technologies), following manufacturer's instructions. Plasmid DNAs were purified and Sanger sequenced using a M13F primer (5'-TGTAAAACGACGGCCAGT-3').

Identification of off-target sites

Potential off-target sites in the human genome (hg19) were identified using TagScan (http://www.isrec.isb-sib.ch/tagger), a web tool providing genome searches for short sequences (92). Guide sequences containing single-base insertions (represented with an "N" in the sequence) and single-base deletions at different positions were entered, followed by the PAM sequence "NGG". We alternatively searched for off-target sites using the recently developed bioinformatics program COSMID that can identify potential off-target sites due to insertions and deletions between target DNA and guide RNA sequences (Cradick, et al. submitted). Primers were individually designed to amplify the genomic loci identified in the output.

Quantitative PCR to measure the expression levels of different guide RNAs

HEK 293T cells were transfected with 750 ng sgRNA variants, as described above. Each sgRNA was transfected as biological triplicates in three separate wells, and processed independently. Total RNA was isolated from cells using the RNAeasy kit (Qiagen). Extracted RNA was reverse-transcribed using the iScript cDNA Synthesis (BioRad). The cDNA was amplified using the iTaq Universal SYBR Green Supermix (BioRad) and analyzed with quantitative PCR using specific primers that annealed at 60°C (Supplementary Table S3 of reference (6)). Quantitative PCR was performed in technical triplicates for each cDNA sample from single transfected well. Relative mRNA expression was analyzed using an MX3005P (Agilent) and normalized to glyceraldehyde-

3-phosphate dehydrogenase (GAPDH) expression. GAPDH expression remained relatively constant among treatments.

Relative mRNA expression of target genes was calculated with the ddCT method. All target genes were normalized to GAPDH in reactions performed in triplicate. Differences in CT values (ΔCT = CT gene of interest−CT GAPDH in experimental samples) were calculated for each target mRNA by subtracting the mean value of GAPDH. ΔCT values were subsequently normalized to the reference sample (mock transfected cells) to get ΔΔCT or ddCT (relative expression = $2^{-\Delta\Delta CT}$).

Deep sequencing to determine activities at genomic loci

Genomic DNAs from mock and nuclease-treated cells that were prepared for T7E1 assays were used as templates for the first round of PCR using locus-specific primers that contained overhang adapter sequences to be used in the second PCR (Supplementary Table S5 and S6 of reference (6)). PCR reactions for each locus were performed independently for 8 touchdown cycles in which annealing temperature was lowered by 1 °C each cycle from 65 °C to 57 °C, followed by 35 cycles with annealing temperature at 57 °C. PCR products were purified using Agencourt AmPure XP (Beckman Coulter) following manufacturer's protocol. The second PCR amplification was performed for each individual amplicon from first PCR using primers containing the adapter sequences from the first PCR, P5/P7 adapters, and sample barcodes in the reverse primers (Supplementary Table S5 of reference (6)). PCR products were purified as in first PCR, pooled in an equimolar ratio, and subjected to 2X250 paired-end sequencing with an Illumina MiSeq.

Paired-end reads from MiSeq were filtered by an average Phred quality (Q score) greater than 20, and merged into a longer single read from each pair with a minimum overlap of 10 nucleotides. Alignments were performed using Borrows-Wheeler Aligner (BWA) for each barcode (93), and percentage of insertions and deletions containing bases within a ±10-bp window of the predicted cut sites were quantified. Error bounds for indel percentages are Wilson score intervals calculated using binom package for R statistical software (version 3.0.3) with a confidence level of 95% (76). To determine if each off-target indel percentage from a CRISPR-treated sample is significant compared to a mock-treated sample, a two-tailed *P*-value was calculated using Fisher's exact test.

CHAPTER 6: CONCLUSIONS AND FUTURE PERSPECTIVES

Genome engineering can be used to create user-defined cells or organisms to serve our needs. For example, genetically modified cells can be used to produce biopharmaceutical reagents (94); engineered cells and organisms can be used as disease models(11); gene-modifying reagents can potentially be delivered into cells to correct diseases; gene function can by studied by knocking out genes. TALENs and CRISPRs are two powerful genome engineering tools developed in the past several years, and substantially accelerate genome modification and allow us to engineer a large variety of organisms even including pigs, cows, and monkeys. Despite The ease to design and make these nucleases, it is still challenging to simultaneously achieve both high efficiency and high specificity with these nucleases. This thesis studied the activity and specificity of TALENs and CRISPRs, and generated tools and guidelines to effectively improve these two aspects.

Architecture of TALEN was examined to help researchers choose the suitable architecture for balanced activity and specificity. A Scoring Algorithm for Predicting TALEN Activity (SAPTA) was developed to screen for active TALENs *in silico*. Finally, we identified the unexpected off-target effect of a recently developed nuclease, CRISPR/Cas9 system. This finding allows the selection of more specific CRISPRs for safer gene modification in cells. The optimization of engineered nucleases in this thesis provides powerful and reliable tools for genome engineering as a way to treat and model diseases.

While there is not a perfect nuclease with consistently high efficiency and specificity in genome engineering, we can choose different versions of nucleases to accommodate our needs in different applications. For example, if high specificity is needed in clinic, we can use TALENs which have less off-target effect. If a high-throughput genome-wide knock-out of a large number of genes is desired, CRISPRs can be used due to the capability to make CRISPRs in large-scale (95). By carefully selecting target sites to avoid similar sequences in the genome, we can also obtain more specific nucleases. This approach is relatively easy with TALENs since the target sequences of TALENs are in the range of 30-50 nt, so similar sequences are rare in the human genome.

A variety of nuclease architectures are also developed to overcome the off-target problem. Obligate hetero-dimeric TALENs(96) would not cut DNA when the same TALENs pair with itself (homo-dimer), and thus eliminate homo-dimeric off-target cleavage. Cas9 nickases cooperatively generate DSBs using two guide RNAs, thus increasing the target length and reducing potential off-target sites in the genome (54,85). Inactivated Cas9 was fused to FokI nuclease domain to further increase target specificity(97). The large size of Cas9 expression cassette (around 5.5 kb) hinders effective packaging of CRISPR/Cas9 system into Adeno-associated viruses (AAV). Smaller Cas9 and promoter, splicing systems of protein or RNA are being investigated to enable the AAV packaging of CRISPR/Cas9. Future optimization of CRISPR/Cas9 specificity and delivery will likely affect whether this tool can be used in clinical applications.

Besides engineered nucleases which cleave DNAs and change genomes, other effector domains can be fused to the DNA binding domains of Transcription-Activator-

Like (TAL) effectors or Cas9 to regulate gene expression or modify epigenetic markers. Activator and repressor domains have been linked with TAL domain and inactivated Cas9 to increase or decrease gene expression (50,98). Demethylation domain was linked to TAL domain to demethylate CpG regions(99). More genome engineering tool kits will likely be developed in the near future to facilitate fundamental research and clinical applications.

We have used these genome-engineering tools in numerous collaborations for disease-related research. The applications include genetic treatment of challenging diseases such as sickle-cell anemia, cystic fibrosis, spinal muscular atrophy, and HIV infection. We envision that these collaborations which combine the powerful genome-engineering tools and interesting research questions will result in potential solutions to various critical diseases, and serve as paradigms in gene therapy research.

Despite the effectiveness of the tools, a number of questions regarding to how the nucleases work still remain to be answered. These questions are discussed below.

Different DNA repair pathways lead to different outcomes of gene editing. Figure 22 shows that NH- and NN-TALENs with similar levels of NHEJ-mediated mutation show different HR-mediated gene targeting. The discrepancy between levels of the two repair pathways may imply that NH- and NN-TALENs cleave DNAs differently. Interestingly, Figure 23 displays that the ratios of large-size deletion are different for NH- and NN-TALENs, which may further indicate different cleavage from these two types of TALENs. It will be useful to investigate what characteristics cause different ratios of HR and NHEJ, and how to shift pathway choices to favor NHEJ or HR in different uses of nucleases.

Although some design guidelines of TALENs were deduced in Chapter 4, it is unclear how TALEN binding / DNA cleavage is affected by these design variables. Mechanistic study of TALEN-DNA binding *in vitro* is needed to further elucidate the importance of different design variables. Alternatively, molecular dynamics simulation of TALEN-DNA binding can also provide some insights and hypothesis that can be verified experimentally. In addition to current design variables considered in our scoring algorithm, there are likely more variables we need to take into account. The SAPTA scoring algorithm may be improved to better predict TALEN activity if these hidden variables are characterized in the future. A larger training data set and more sophisticated algorithm may also improve SAPTA. Machine learning and data mining approaches may be useful in analyzing the data set.

There is plenty of room for improving genome-engineering tool kit to make them more efficient, specific, and versatile, and using them to solve biological problems. Computational modeling and data analysis combined with experimental validation may be a viable approach for system improvement. Although numerous studies have been published following the development of TALENs and CRISPRs, the potential uses of the genome-engineering tools is yet to be fully explored. Collaboration between tool-developing labs and labs with interesting biological questions will likely lead to significant advances in both basic and translational research.

APPENDIX A: SUPPLEMENTARY INFORMATION

Chapter 3 supplementary information

Table 6: Off-target analysis of KNH TALENs targeted to *CXADR, CFTR*, and *AAVS1*(*PPP1R12C*) using SMRT sequencing.
The total number of mismatches in each binding site (T) as well as the number of mismatches in the left (L) and right (R) half-sites is provided as well as the closest annotated RefSeq gene, and the coordinates in the hg19 human genome build. The intended on-target gene is marked by "(on)" next to the gene name. Some sites failed to generate specific amplicons and are listed as "PCR Failure". One site did not obtain any sequencing reads passing quality metrics and is listed as "Seq Failure".

	Mismatches					Mock	NK	NH	NN
Nuclease	T	L	R	Closest Gene	hg19 Coordinates	% indel	% indel	% indel	% indel
C3/C4	0	0	0	CXADR (on)	chr21:18965449	0	53.55%	65.24%	49.86%
	5	3	2	PRDX2	chr19:12909463	1.15%	0	1.40%	0
	5	2	3	CDKL4	chr2:39450719	2.55%	1.56%	2.16%	1.24%
	5	3	2	AEBP2	chr12:19699361	0	0	0	0
	7	2	5	NFIA	chr1:61962413	0	0	0	0
	6	5	1	SP3	chr2:174563626	0	0	0	0
	6	4	2	ZNF91	chr19:23621161	Seq Failure			
	5	2	3	CDKL4	chr2:39450719	PCR Failure			
	5	3	2	CDKL4	chr2:39450721	PCR Failure			
	6	1	5	ABCA1	chr9:107696114	0	0	0	0
	8	5	3	CDH2	chr18:25842175	0	0	0	0
	6	2	4	GLCCI1	chr7:8088601	0	0	0	0
	6	3	3	DKFZp686O1327	chr2:145413307	0	0	0	0
	6	3	3	D21S2088E	chr21:24359673	0	0	0	0
	6	3	3	TMEM215	chr9:32800075	0	0	0	0
	6	1	5	ABCA1	chr9:107696114	PCR Failure			
	7	4	3	SLCO1A2	chr12:21506376	0.21%	0.07%	0.43%	0.28%
	5	3	2	PRDX2	chr19:12909465	0	0	0	0
	5	2	3	NFAT5	chr16:69657715	PCR Failure			
	6	3	3	LGALS8	chr1:236675094	0	0	0	0.06%
	8	4	4	EFHA2	chr8:16930149	0	0	0	0
	7	3	4	RAP1A	chr1:112167944	0	0	0	0
	7	2	5	TMEM2	chr9:74312571	0	0	0	0

Table 6 (continued).

	Mismatches					Mock	NK	NH	NN
Nuclease	T	L	R	Closest Gene	hg19 Coordinates	% indel	% indel	% indel	% indel
F8/F9	0	0	0	CFTR (on)	chr7:117199507	0	49.62%	48.17%	52.44%
	11	5	6	GRM7	chr3:7743292	0	0	0	0
	11	5	6	SORCS1	chr10:109376387	0	0	0	0
	11	6	5	MIR4532	chr20:56564815	0	0	0	0
	11	6	5	SEMA3A	chr7:83547837	0	0	0	0
	12	7	5	PHF20L1	chr8:133839511	0	0	0	0
	13	7	6	GHITM	chr10:85331718	0	0	0	0
	12	5	7	OSBPL6	chr2:179132155	0	0	0	0
	12	7	5	IKZF2	chr2:213921336	0	0	0	0
	12	7	5	AKT3	chr1:243699384	0	0	0	0
	12	5	7	TSPEAR	chr21:46083995	0	0	0	0
	12	5	7	ATP8B4	chr15:50345887	0.46%	0.24%	0.38%	0.36%
	12	5	7	FAM91A1	chr8:124770049	0	0	0	0
	12	5	7	ANKRD50	chr4:125700596	0	0	0	0
	12	5	7	MYB	chr6:135443049	0	0	0	0
	12	7	5	SPANXN1	chrX:143753106	0.23%	0.15%	0.15%	0.15%
	12	6	6	RNU6-53	chr10:111205307	0	0	0	0
	12	6	6	RORB	chr9:77272288	0	0	0	0
	12	6	6	PLA2G4C	chr19:48598753	0	0	0	0
	14	7	7	SYT17	chr16:19171179	0	0	0	0
	12	6	6	FPR3	chr19:52322997	0	0	0	0
	14	7	7	PHYH	chr10:13336769	0	0	0	0
	13	6	7	CDCA7L	chr7:21977297	0.00%	0	0	0
G1/G2	0	0	0	PPP1R12C (on)	chr19:55627036	0	48.53%	52.65%	58.31%
	8	6	2	PDLIM5	chr4:95416588	0	0	0	0
	8	4	4	CACNG4	chr17:65016227	0	0	0	0
	9	6	3	ARHGEF10L	chr1:17913943	0	0	0	0
	9	3	6	CPSF7	chr11:61171976	0	0	0	0
	9	3	6	EDC4	chr16:67912248	0	0	0	0
	9	6	3	STOM	chr9:124106256	0	0	0	0
	9	3	6	PLEKHG1	chr6:151106895	0	0	0	0
	9	6	3	MFSD6	chr2:191364289	0	0	0	0
	9	6	3	UMOD	chr16:20349508	0	0	0	0
	9	3	6	POC1B	chr12:89888350	0	0	0	0
	9	3	6	BAI1	chr8:143641530	PCR Failure			

Table 6 (continued).

	Mismatches					Mock	NK	NH	NN
Nuclease	T	L	R	Closest Gene	hg19 Coordinates	% indel	% indel	% indel	% indel
G1/G2	9	6	3	CHD6	chr20:40459509	0	0	0	0
	9	3	6	RASSF6	chr4:74533040	0	0	0	0
	9	3	6	TMEM135	chr11:87351794	0	0	0	0
	9	5	4	SCARA5	chr8:27847586	PCR Failure			
	9	4	5	UBE2V2	chr8:49105234	PCR Failure			
	10	6	4	DPY19L1P1	chr7:32690655	PCR Failure			
	10	4	6	NUP35	chr2:184257506	0	0	0	0
	10	6	4	ZEB1	chr10:31889755	0	0	0	0
	10	6	4	HIVEP3	chr1:42546490	0	0.10%	0	0
	12	6	6	FLJ31662	chr1:96504572	0	0	0	0
	11	6	5	DDIT4L	chr4:101002104	0	0	0	0
	9	5	4	ACVR2A	chr2:148573305	0	0	0	0

Table 7: Potential off-target sites listed in the table above for KNH TALENs targeted to *CXADR, CFTR*, and *AAVS1*(*PPP1R12C*).
TALEN half-sites are listed including the 5' base; mismatches relative to the intended binding site are given in lowercase. The right half-site is given as the 5'->3' sequence on the top DNA strand, so it is listed in the reverse anti-sense orientation to the sequence bound by the TALEN protein.

Closest Gene	Left Half-Site	Spacer	Right-Half Site
CXADR (on)	TCTCTTTTTTTCTTTTTGT	agttcaagtacccttacaa	GACTGATGGAATTACA
PRDX2	cCTCTTTTTTTtTTTTTtT	ttttttaaccaataccccatctctt	AaAAAAAaAAAAAAAGAGA
CDKL4	TCTCTTTTTTTtTTTTTtT	gaagacttctgtcttcaaa	AaAAAAAaAAAAAgAGAGA
AEBP2	TtTCTTTTTTTtTTTTTGa	agattgttttcttattagccttatta	gCAAAAAGAAAAAAAaAGA
NFIA	TtTCTTTTTTTCTTaTTGT	atcactttcagaatt	gCtAgAAGAtAAAAAGAGg
SP3	TCTCTcTgaTTCTTaTTGa	aatagcttttgcttcca	ACAAAAAaAAAAAAAGAGA
ZNF91	TCTCTTTcTcTCcaTTTGT	gacacccaggctggagtgtg	ACAAAAAaAAAAAAAGAaA
CDKL4	TCTCTTTTTTTtTTTTTtT	gaagacttctgtcttcaaa	AaAAAAAaAAAAAgAGAGA
CDKL4	TCTtTTTTTTTtTTTTTGa	agacttctgtcttca	AaAAAAAaAAAAAAAGAGA
ABCA1	TCTCTTTTTTTCTTTTTcT	ttataaaatattttta	ACAtAgAaAAAtAAAGAaA
CDH2	TCTCaTTTTTaaTgTTTGa	aaaaattctatttcttctagtcttcattt	ACAAgtAGAAAAAAAGAGg
GLCCI1	TtTCTTTTTTTgTTTTTGT	ttttactattctggtttttataa	ACAAAAAttgAAAAAGAtA

Table 7 (continued).

Closest Gene	Left Half-Site	Spacer	Right-Half Site
DKFZp686O1327	TCTCaccTTTTCTTTTTGT	aatcttttagaccagtaaagagaaa	ACAAAAAcAAAAAcAGAaA
D21S2088E	TgTCTTTaTgTCTTTTTGT	actgactatagt	ACAAAAAGAtAtAAAGAaA
TMEM215	TCTtTTTTTTaCTTTTTaT	catgtttccatcaaacacttattcatt	ACAAAAAGAAAAAtAGtaA
ABCA1	TCTCTTTTTTTCTTTTTcT	ttataaaatatttttaacat	AgAAAAAtAAAgAAAaAtA
SLCO1A2	cCTCTTTcTTTaTTTTTGg	gatggtgctcagaaaaa	AaAAAAAGAAAgAAAGAaA
PRDX2	TCTtTTTTTTTtTTTTTtT	ttttaaccaataccccatctctt	AaAAAAAaAAAAAAAGAGA
NFAT5	TCTtTTTTTTTCTTTTTtT	tgcccaaacaaaaaggaaatttga	AaAAAAAaAAAAAAAGAaA
LGALS8	TCTCTcTTTTaCTTTTTtT	aaagacttcttctcaaaa	AtAAAAAGAAAgAAAGAaA
EFHA2	cCcCTTTTTTTaTcTTTGT	aaaaaagattatggc	ttAAtAAGAAAAAAAtAGA
RAP1A	TCTtTTTTTTTtgTTTTGT	ttttgtctttagaaaag	ACAAAActAtAAAAAGAGg
TMEM2	TGTAAaTCCAaCAGTC	atattgcaacagaatcaaacatctg	AgAAAgAaAAAAAAtGAGg
CFTR (on)	TTTATTTCCAGACTTCACTTCT	aatggtgattatggg	AGAACTGGAGCCTTCAGAGGGTA
GRM7	TTTATTTCaAGtCaTCcCTTCa	ggcctcagta	AGAAtTGAAGgagGGtAATtAA
SORCS1	TccATTTCCAaAaTTCcCTTCT	gaattgccatgtcagcaaaatgg	AGAAGTactGTtTGGAAAatAA
MIR4532	TTTcTTTCCtttCcTCcCTTCT	ttcttcttggactct	gGAtGTGAtGTCTGGAgAcAAA
SEMA3A	aACCCTCTGcAGtaTCCAtTcCT	ctgcaacacgtggtcaggatatt	tGAAGaGcccTCTGGAAATAAA
PHF20L1	TTTgaTTaCcGAaTaCcCTTCT	gtgagtggtatgtcacaatacagaaatg	AGAAGctcAGgCTGcAAATAAA
GHITM	aTcATTTaCAGAgcTCAgTaCT	gccaaaagtacttaggggggaaaagttggga	AGAAtTGAttTCTGaAAgTtAA
OSBPL6	TTaccTTCCccACTTCACTTCT	cccatactcttaagctctattatagga	tGAAaaGAAaaCTGaAAATAAc
IKZF2	cTgAaTcCCAaACTTCttTTCT	gatatgctattc	AGAttTGAAGgCTGGAAgTcAA
AKT3	TTTtcTTgCtGAtTTgAgTTCT	ttataagattctgggtattagccctttttc	AGAAGTaAAGTtTGcAAATAtt
TSPEAR	TTTATTaCCAtACTggACTTCc	taaacagtgaagtaaa	AagcaTGcAGcCTGGAAAgAAA
ATP8B4	TaTATTTCCAGcaTTtACTaCT	tggtgggggggg	tGggGTGAAGgtgGGgAATAAA
FAM91A1	TTTtTTTtCAGACTcaACTTtT	tgtcaaggtttgacccagcataaaggagct	AGAAaTGcAGTaTGctAAaAAg
ANKRD50	TAggaTCTGAgGGCTtCAGTTCT	acaactgtag	gGAAaTGAAtTCTGccAAcAAc
MYB	TTTAacTCCAGACTTCAacTtT	ccactgagttttcgatatgtcacct	AGAAGatAAcaCTGGAggaAAA
SPANXN1	aTaAaTaCCAGAgTTCAaTTCa	tacaaaagagcagcatttgtg	tGAAGTGAAcTCTGGtAtTtAA
RNU6-53	cTgATgcCCAGACTcCACTTgT	agatattaaaattgccatttctggctt	tGcAGTGAAGTCTGGAcATgcg
RORB	gcTcTTTCgAaACcTCACTTCT	atcctcagtaaaggggagactggtcctaa	AGAAGTaAAGgtTGtAAAgAtA
PLA2G4C	gTTcTgaCCAGtCTTCcCTTCT	tgaatttgcttc	cGAAGTGggtTaTGGAAAcAAA
SYT17	caaATccCCAaACcTCACTTCT	aaatctactgcatcacaatctct	gGgtGaGgAGTCTGaAAATAAg
FPR3	TcTcTTTaCAGAaaaCACTTCT	cttaaagaaat	AGggtTGAAGctTGGAAAgAAA

Table 7 (continued).

Closest Gene	Left Half-Site	Spacer	Right-Half Site
PHYH	TccATTcCCAGAaTTaAaTTaT	tccctagtttagcatacc	AGtAGgGAAGTCTttttATAAt
CDCA7L	cacATTTCCtGACTTCAaTaCT	tactacaaagctatagtaatcaa	AacAGTGtgGTCTtGgcATAAA
PPP1R12C (on)	TCTGCCTAACAGGAGGTG	ggggttagacccaat	ATCAGGAGACTAGGAAGGAGGA
PDLIM5	aCTGtCTAtCAGatGtTG	accttaacaa	CAtCTCCTGTTAGGtAGA
CACNG4	gCTGCCTggCAGGAGcTG	cggctgttgggaagcaatgccacccctt	CACCTgCaGcTgGGCAGA
ARHGEF10L	TCTctCTtcCAGGAGaTc	agctggttccaggagtcccagg	CcCCTCCTcTgAGGCAGA
CPSF7	cCTGgCaAACAGGAGGTG	tgggattgggaggactgccaagcagtct	tcaaTCCTGTggGGCAGA
EDC4	TCTGCCTAcCtGcAGGTG	tctgcacgagtgga	aACCTCaTGaTgGGCgGc
STOM	TCTcCCTccCcGaAGGaG	ggcgaatcagtgga	ATttGGAGACTAGGtAGGAGGA
PLEKHG1	TCTGCagAgCAGGAGGTG	agagcaagtgaagctgagctc	CACCTCCTGTcAGatcag
MFSD6	TCTGCCTccaAGGAtaaG	aacccttatt	tACCTCCTGTTAGaCAaA
UMOD	TCTGCCTggCAaGAGaaa	agccctgaggttaagtggact	CAaCTCCTGaaAGGCAGA
POC1B	TCTGCtTcAaAGGAGGTG	gagtctgtttattt	CtCCTCCctTTgGGCtGg
BAI1	aCTGgCTAACAGGAGGgG	atgcagcacatggtccat	CACCggtTaaTAGGCAGg
CHD6	TCTGtCTAAaAacAGGaa	ctggaagtggcacatgccacttctgct	CACaTCCTGTTgGcCAGA
RASSF6	TCTaaCTAACAGcAGGTG	agtggttttgaca	CAaCTCCTGaTgGGtAtt
TMEM135	TgaGCCTAgCAGGAGGTG	ttgtgttgaacaag	gACtTCgTGTTtGttAGA
SCARA5	TCCaCCTTCCTccTCTtCTGgT	cagctcctgggacggggctggaatg	CACggCtTGTTAGcCAGA
UBE2V2	TCTtCCTAgCAGGAGagG	gaagataaggaggaagagg	AggAGGAGgagAGGAAGGAGGA
DPY19L1P1	TCCTCCTTgCTAcTgcCCaaAT	aagccagttatt	CtCCTgCTtTaAGGCAGA
NUP35	TCTGaaTAAgAGGAGGaG	gaggaagaagaggaagatgaagaaa	AgaAGGAGAagAaaAAGGAGGA
ZEB1	TCtTCCTTCCcAcTCTgCTtcT	gccttcatcctgcatcctct	CACCTCCTGcTtctCAGA
HIVEP3	TCTGaCTAcCAtataGTG	tgggaactct	ATCAtGAGAaaAGcAAGGAGGA
FLJ31662	TCccCCTAACAaGAaaTa	agtgcaaacaaaaagtgaccactctaagac	CAttTtgTGTTgaGCAGA
DDIT4L	cCaGCCTAAtAaGAaGTc	aagtaaacacctcaatacaatccgatggat	tAtgTaaTGTTAGGCAGA
ACVR2A	TCTtgCcAAgAGGAGcTG	ttccattacagacttagtcacaagacaaac	CAttTgCTGTTAGGgAGA

Table 8: Primers to PCR-amplify potential off-target sites in the table above for KNH TALENs targeted to *CXADR, CFTR*, and *AAVS1*(*PPP1R12C*).

Closest Gene	Forward Primer (5' to 3')	Reverse Primer (5' to 3')
CXADR (on)	CCCATAAGGAAGCAGTGTGATGAC	GCTCTAATCTCTGTGCCTTGTGC
PRDX2	AGGCCTCCCAAAATGCTGGGAT	GCAGAAGATTGCTTAGGCCCAG
CDKL4	CCCTGACTGTTGATAAGACTGGGT	TGAGCAGAGATCATGCCACTGCAC
AEBP2	GACATGCATGTTTGCTGGGCAGT	GCAACCGATGGCAATCCTAAAGC
NFIA	CACGTTTTGTGCTATCCAGCCTTG	ACTACTCCCTAGGAATGCTTGCC
SP3	CAGGCTATAACTCTGGCAACAGTC	CTGCACAATGTGCACATGTACCC
ZNF91	GCTTACAGCCAACAGCTGTCTC	CCCGGGACTACTCAGAACTTCTT
CDKL4	CCCTGACTGTTGATAAGACTGGGT	TGAGCAGAGATCATGCCACTGCAC
CDKL4	CCCTGACTGTTGATAAGACTGGGT	TGAGCAGAGATCATGCCACTGCAC
ABCA1	ACTTCCTTAGAGAGAGAGAGGCC	GGAAGGAAAGTGTGGTCAGTAGTG
CDH2	GCTCTGAATCTTGGGGAAAGTGAC	CCTGTCTCACTGACAATGCTAGGA
GLCCI1	TTACCCTTTGTCCCAGAGGGAAC	CTAATGAGAGGGCTTGCTGGTG
DKFZp686O1327	GCTACAGTCTGCTAGAGGATGACA	GCCAAGCAGACATTATGGAGGAC
D21S2088E	GGTCAGTACAGAGAAACGCACAG	GGGGGTGGGTGTCTTCTATTTC
TMEM215	CTTTCTGGCTTGCTCCTGTTTCTC	GTAGAATGTCCCCCAGTCTGGAT
ABCA1	ACTTCCTTAGAGAGAGAGAGGCC	GGAAGGAAAGTGTGGTCAGTAGTG
SLCO1A2	CCTCTCTGAGCAGTGCTACAAC	CACAGCAATCTGCTATCTCCTACC
PRDX2	GGCCTCCCAAAATGCTGGGAT	GGCAGAAGATTGCTTAGGCCC
NFAT5	CTCCCAAGGTGCTGGGATTAC	GCCAACATGGTGAGACCCTG
LGALS8	CGAGACTGAGCCACTGCCTT	CCTGGCTCCTAAAGAACTCCAGA
EFHA2	GGATGGGAGGATATGCATGCAAC	CTGGCGGGTGTCAAATAGTATCG
RAP1A	GCTGTACCCACTTATTCTGGAGG	CGTGGGTGATATAGCAAGACTCTG
TMEM2	GCATTTCGAACCAAAGTCAGTCGG	CCCCTGTGATACGTCAGTTTACC
CFTR (on)	CCCCTTCTCTGTGAACCTCTATC	GGCATGCTTTGATGACGCTTCTG
GRM7	TTAGCCAGGATGGTCTCGATCTC	CCACTGTCATGGAGCTTAGTGTC
SORCS1	TGAGCCCTTTGTGGTCTGGCA	GGACAGGCATCACTTTGGTATGAG
MIR4532	ACGTTCGCCTTCCCAGACCT	TGGCTCCACAGTGCTACTGAAAC
SEMA3A	CCTTGATTTGCTCTCGGTCTCTTG	CACCCAGCCTTCAGCTCACT
PHF20L1	ACCCCCCCATCACCTCTTCA	CCTGCACTGACAATGCAGGTCT
GHITM	GCTGAGCACTGCCTATGAGAAG	CCTCTCTTTCCTCTCCACTACC
OSBPL6	TTGCCACTGCACTCCAGCCT	CTCTGCCCCTGATGAAACCTG
IKZF2	CTGTGGTAGGATAGGCAGAGATCA	GACTGCCTGTGACATCTGATGTTG
AKT3	CCCATGTTTGTTGGCCATTTGCAC	CCGGAGATTGACCTAGGCAAAG
TSPEAR	CCCAAGAATGCAAGGTTGGTTCAG	GAGAGTCGACATATCCAGCAATCC
ATP8B4	TCTCTTGGCCCCTAACAAAGCTC	GAGGCAATCTCCCAAATGCTCC
FAM91A1	CTGGTGTTCTATTGCACAGTAGGG	CAGTGTCAGCTCCTATCGTTTGC
ANKRD50	AGTGGTAGACTGTCCTGCAGGT	CAGTTTCCACAGGTCAGAAGCC

Table 8 (continued).

Closest Gene	Forward Primer (5' to 3')	Reverse Primer (5' to 3')
MYB	GGGGCATCAGTCTTGTAGGG	TGGAAGACCTGTGTCCCCAAC
SPANXN1	CTGCACCTGGCTAACTTGGATTC	CTCGGAGGTGAAGGCAACAG
RNU6-53	CAGTGATCCTAGGTTACTCCGTG	TTGGCTTTCCAGGGAGAATGGTAG
RORB	TCGATCTCCTGACCTCGTGATCT	GGATGAGCACAGTTTGTCACATGG
PLA2G4C	GGCAAACTAGACCCTCAGAAGGA	CTGGTTCTGTGCTGGGACTC
SYT17	GTAGGAAGTGTGTGCACATCACC	GGTAACAGAGTGAGACCCTGTCAA
FPR3	GCATTAGCTGGAGTGCTAAGGAC	CCACTGTCTCTCGTCTTCCC
PHYH	GCAGACTTGAGGGGTGGTTTC	CCATGCCTCTGTTTCCTCATCTG
CDCA7L	GAGATTGCACCACTGCACCC	GGCTTAAGGTAAGGGTCCAACCTA
PPP1R12C (on)	CAGGATCCTCTCTGGCTCCAT	CCTTATATTCCCAGGGCCGG
PDLIM5	GACCAAATCAGGTTTCCCGTCAG	TGACCACAGAGCATACCATGCC
CACNG4	AACCCACAAGGGCTGGAGAG	TCCTCACTCCATCCTGCCCT
ARHGEF10L	ACCTGAGGTGCTCACCCAGT	AGTGGATCAGTGGGGGGTCT
CPSF7	CCCGGAGAATCAGAAACTCATACG	CATCTCCTCACCTTGATCAGGAC
EDC4	GCAGATCTACATTGAGGGGCAAG	CAAGGCAGTGAGTGACTGGC
STOM	AGAGAAGGCAGTGAAAGCTGCTAG	GCAGGACATGATGTGTGGCTTTG
PLEKHG1	CCCTTCCCAGCATATCTCTCC	GGGATGGTTTTGGGATGACTCAAG
MFSD6	CTGCTGGTGATGTTCTACTCATGG	TTTACTGAGCCCCCCAAGCC
UMOD	ATTTCGGAGGGAGGCTGAACAC	AAAGAGGAGACAAGTCGGGGAG
POC1B	CTCCGCTTCACAAAAGGTGAAGTC	GGGTTCCAGGAGAATCAAAGCAG
BAI1	TCCATCACCGGTTAACAGGAGG	GCCAGTGATGGACCATGTGC
CHD6	GCTGTGTCTGTGTTTGCTTCTCC	GACTCTTAGTTGGACACTTGGCC
RASSF6	CCTCCCTAGCACTTACCAAGAGA	CCCATTGTTACCCCAGAACATCC
TMEM135	GGCACCTGTCAAGATGTACGG	CTTCCCAGGCACAAAGTTTCCTC
SCARA5	CAAAGGCAATAGCCTCTGCTTCC	TGTGCACTCACATCTGGCCC
UBE2V2	GGTAGAAGCTGCCTCCAAACCA	CTTCTTCTCCTCCTCTCCTCC
DPY19L1P1	CTCTAGTGGCATTCCATCCACTC	TTTGGCCTGCTGTACTCCAGGT
NUP35	GACTGAGTGAGACTCTGCCTCAA	CTCTGTCTCTCTCTCTCTTTCTCC
ZEB1	CTTTCTCATTCCCTCCAGCCAC	CAGTGCTGGCATTTCAGGCACTGT
HIVEP3	GGCTGAGGAGGCCTTAGGAA	TGGTTTGGCTCTCTGTCACCATC
FLJ31662	CTGTCACTCAGAACAAGCCATGG	GTTTCTGAAGAACAGGGAGGCC
DDIT4L	GGGTGCAGCCTAGGTTATAGTC	TAAATGTCACTTCCTCCCAAGGCC
ACVR2A	GAGAGCAACACTTTTCACAGTGCC	CAAGATCACGCCACTCTGTCTC

Amino acid sequences of TALENs

S-02, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPEQVVAIASNKGGKQALETVQRLLPVLCQAHGL
TPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGGKQALET
VQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVV
AIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLP
VLCQAHGLTPEQVVAIASNKGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNI
GGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQA
HGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQA
LETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPE
QVVAIASNGGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNKGGKQALETVQ
RLLPVLCQAHGLTPEQVVAIASNGGGRPALESIVAQLSRPDPALAALTNDHLVAL
ACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSEL
RHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDG
AIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWW
KVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKA
GTLTLEEVRRKFNNGEINF

S-02, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET

VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
IGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLV
ALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKS
ELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKP
DGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNE
WWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEM
IKAGTLTLEEVRRKFNNGEINF

S-02, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
IGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETV

QRLLPVLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLV
ALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKS
ELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKP
DGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNE
WWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEM
IKAGTLTLEEVRRKFNNGEINF

S-05, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNKGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNI
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPA
LDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVP
HEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPI

DYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVT
EFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEV
RRKFNNGEINF

S-05, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNI
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPA
LDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVP
HEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPI
DYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVT
EFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEV
RRKFNNGEINF

S-05, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNI
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPA
LDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVP
HEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPI
DYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVT
EFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEV
RRKFNNGEINF

S-12, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGL
TPEQVVAIASNKGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALET
VQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVV
AIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLP
VLCQAHGLTPEQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHD
GGKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGGKQALETVQRLLPVLCQAH
GLTPEQVVAIASNKGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNKGGKQAL
ETVQRLLPVLCQAHGLTPEQVVAIASNKGGKQALETVQRLLPVLCQAHGLTPEQ
VVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGGKQALETVQRL
LPVLCQAHGLTPEQVVAIASNKGGKQALETVQRLLPVLCQAHGLTPEQVVAIAS
NGGGRPALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPA
LIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQ
DRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGG
YNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGN
YKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

S-12, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL

PVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAI
ASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPH
APALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARN
STQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAY
SGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFK
GNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

S-12, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAI
ASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPH

APALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARN
STQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAY
SGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFK
GNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

R-04, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGL
TPEQVVAIASNKGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALET
VQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVV
AIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLP
VLCQAHGLTPEQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHD
GGKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGGKQALETVQRLLPVLCQAH
GLTPEQVVAIASNKGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNKGGKQAL
ETVQRLLPVLCQAHGLTPEQVVAIASNKGGKQALETVQRLLPVLCQAHGLTPEQ
VVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGGKQALETVQRL
LPVLCQAHGLTPEQVVAIASNKGGKQALETVQRLLPVLCQAHGLTPEQVVAIAS
NGGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGGRPALESIVAQLSRPDPA
LAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQ
LVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYR
GKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQ
TRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLS
VEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

R-04, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAI
ASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALESIVAQLSRPD
PALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAG
SQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYG
YRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEE
NQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGA
VLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

R-04, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV

VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAI
ASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALESIVAQLSRPD
PALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAG
SQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYG
YRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEE
NQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGA
VLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

S-116, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
GGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETV

QRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACL
GGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHK
LKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYT
VGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVY
PSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTL
TLEEVRRKFNNGEINF

S-116, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
GGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACL
GGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHK
LKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYT

VGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVY
PSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTL
TLEEVRRKFNNGEINF

S-116, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
GGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACL
GGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHK
LKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYT
VGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVY
PSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTL
TLEEVRRKFNNGEINF

S-120 (no G in target)

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
IGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVA
LACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSE
LRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPD
GAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEW
WKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIK
AGTLTLEEVRRKFNNGEINF

C-03, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV

VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
GGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALAC
LGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRH
KLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIY
TVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKV
YPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGT
LTLEEVRRKFNNGEINF

C-03, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
GGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLT

PDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALAC
LGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRH
KLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIY
TVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKV
YPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGT
LTLEEVRRKFNNGEINF

C-03, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
GGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALAC
LGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRH

KLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIY
TVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKV
YPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGT
LTLEEVRRKFNNGEINF

C-04, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHD
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASNKGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASHDGGKQALESIVAQLSRPDPALAALTNDHLVA
LACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSE
LRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPD
GAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEW
WKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIK
AGTLTLEEVRRKFNNGEINF

C-04, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHD
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASHDGGKQALESIVAQLSRPDPALAALTNDHLVA
LACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSE
LRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPD
GAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEW
WKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIK
AGTLTLEEVRRKFNNGEINF

C-04, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLP

VLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHD
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASHDGGKQALESIVAQLSRPDPALAALTNDHLVA
LACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSE
LRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPD
GAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEW
WKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIK
AGTLTLEEVRRKFNNGEINF

F-08, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAI
ASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVL

CQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGG
GKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPAL
DAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPH
EYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPID
YGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTE
FKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVR
RKFNNGEINF

F-08, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAI
ASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVL
CQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGG
GKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPAL

DAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPH
EYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPID
YGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTE
FKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVR
RKFNNGEINF

F-08, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAI
ASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVL
CQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGG
GKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPAL
DAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPH
EYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPID
YGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTE

FKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVR
RKFNNGEINF

F-09, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
KGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAI
ASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVL
CQDHGLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGG
GKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQA
LESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNR
RIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMK
VMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQ
ADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTR
LNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

F-09, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
HGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAI
ASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVL
CQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGG
GKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQA
LESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNR
RIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMK
VMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQ
ADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTR
LNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

F-09, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA

LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
NGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAI
ASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVL
CQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGG
GKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQA
LESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNR
RIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMK
VMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQ
ADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTR
LNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

G-01, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALET

VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNI
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALETVQR
LLPVLCQDHGLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGLTPDQVVAIA
SNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALESIVAQLSRPD
PALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAG
SQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYG
YRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEE
NQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGA
VLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

G-01, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNI
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPD

QVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQR
LLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIA
SNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALESIVAQLSRPD
PALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAG
SQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYG
YRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEE
NQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGA
VLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

G-01, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNI
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQR
LLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIA
SNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALESIVAQLSRPD
PALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAG
SQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYG
YRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEE

NQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGA
VLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

G-02, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNK
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPAL
DAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPH
EYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPID
YGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTE
FKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVR
RKFNNGEINF

G-02, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNH
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPAL
DAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPH
EYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPID
YGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTE
FKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVR
RKFNNGEINF

G-02, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA

LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNN
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPAL
DAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPH
EYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPID
YGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTE
FKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVR
RKFNNGEINF

G-122, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPEQVVAIASNIGGKQALETVQRLLPVLCQAHGLT
PEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETV

QRLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPEQVVA
IASNKGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNKGGKQALETVQRLLPV
LCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNGG
GKQALETVQRLLPVLCQAHGLTPEQVVAIASNKGGKQALETVQRLLPVLCQAH
GLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQAL
ETVQRLLPVLCQAHGLTPEQVVAIASNKGGKQALETVQRLLPVLCQAHGLTPEQ
VVAIASNGGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRL
LPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIAS
NIGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNGGGRPALESIVAQLSRPDPA
LAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQ
LVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYR
GKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQ
TRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLS
VEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

G-122, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
GGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLT

PDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALESIVAQLS
RPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHR
VAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMK
VYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRY
VEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCN
GAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

G-122, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
GGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALESIVAQLS
RPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHR
VAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMK
VYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRY

VEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCN
GAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

G-123, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPEQVVAIASNKGGKQALETVQRLLPVLCQAHGL
TPEQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALET
VQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVV
AIASNKGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLP
VLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHD
GGKQALETVQRLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQA
HGLTPEQVVAIASNKGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNKGGKQA
LETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPE
QVVAIASNGGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQ
RLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPEQVVAI
ASNGGGRPALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHA
PALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNS
TQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYS
GGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFK
GNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

G-123, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT

AVEAVHAWRNALTGAPLNLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGL
PHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIA
RNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTK
AYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSG
HFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEI
NF

G-123, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASH
DGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQ

DHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGL
PHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIA
RNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTK
AYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSG
HFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEI
NF

G-128, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
GGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNKGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGL
PHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIA

RNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTK
AYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSG
HFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEI
NF

G-128, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
GGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNHGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGL
PHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIA
RNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTK
AYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSG
HFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEI
NF

G-128, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
GGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNNGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGL
PHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIA
RNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTK
AYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSG
HFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEI
NF

G-42, NK

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVV

AIASNKGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNKGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNI
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASNKGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNKGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHD
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPAL
AALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQL
VKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRG
KHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQT
RNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSV
EELLIGGEMIKAGTLTLEEVRRKFNNGEINF

G-42, NH

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNI
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQ

ALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASNHGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNHGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHD
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPAL
AALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQL
VKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRG
KHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQT
RNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSV
EELLIGGEMIKAGTLTLEEVRRKFNNGEINF

G-42, NN

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQ
QQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAA
LPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVT
AVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL
TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNI
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLP

VLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHD
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPAL
AALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQL
VKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRG
KHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQT
RNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSV
EELLIGGEMIKAGTLTLEEVRRKFNNGEINF

Table 6: Primers used for T7E1 assay and SMRT sequencing analysis.

Primers for mutation analysis by T7E1 assay			
Gene	**Forward primer name**	**Reverse primer name**	**PCR product length**
HBB	B-glo-Fwd	B-glo-Rev	404
CXADR	CAR-F5	CAR-R4	427
CFTR	CFTR-F	CFTR-R2	501
AAVS1	AAV-F	AAV-R	528
CCR5	CCR5_1_10_1_F	CCR5_1_10_1_R	477
ATF4	ATF4-F	ATF4-R	494
HBD	delta-6F	delta-6R	504
Primers for SMRT deep sequencing analysis[a]			
Gene	**Forward primer name**	**Reverse primer name**	**PCR product length**
CCR5	CCR5_1_10_1_F	CCR5_1_10_1_R	477
	CCR5_1_10_2_F	CCR5_1_10_2_R	477
CCR2	CCR2_1_10_1_F	CCR2_1_10_1_R	381
	CCR2_1_10_2_F	CCR2_1_10_2_R	381
ATF4	ATF4F_1	ATF4R_1	494
	ATF4F_2	ATF4R_2	494
USP28	USP28F_1	USP28R_1	342
	USP28F_2	USP28R_2	342
RNF157	RNF157F_1	RNF157R_1	463
	RNF157F_2	RNF157R_2	463
CTAG1B	CTAG1BF_1	CTAG1BR_1	347
	CTAG1BF_2	CTAG1BR_2	347

[a]Each locus may use barcode-containing primers if PCRs of the same locus from different samples were pooled. Barcodes are underlined in the sequences below.

Table 6 (continued).

Primer name	Sequence (5' to 3')*
B-glo-Fwd	CCAACTCCTAAGCCAGTGCCAGAAGAG
B-glo-Rev	AGTCAGTGCCTATCAGAAACCCAAGAG
CAR-F5	CATACTATTGAGCTATACCTTTATGATTTGTGCAG
CAR-R4	ACAGCTGCTCTAATCTCTGTGCCT
CFTR-F	TGTGCCCCTTCTCTGTGAACCTC
CFTR-R2	GGGTAGTGTGAAGGGTTCATATGC
AAV-F	CTCTCTAGTCTGTGCTAGCTCTTCCAG
AAV-R	CTCAGGTTCTGGGAGAGGGTAGC
ATF4-F	AGAAGTCCCGCCTCATAAGTGGAAG
ATF4-R	CATTTGAGTGATGGGGCCAAGTGAG
delta-6F	CCCATAACAGCATCAGGAGAGGACA
delta-6R	CAGCCAATCTCAGGGCAAGTTAAGG
CCR5_1_10_1_F	CTGTGTGCAG GCACAGGGTGGAACAAGATGG
CCR5_1_10_1_R	ACTACATATG ACCACCCCAAAGGTGACCGT
CCR5_1_10_2_F	AGTCGACACT GCACAGGGTGGAACAAGATGG
CCR5_1_10_2_R	GCATAGATCG ACCACCCCAAAGGTGACCGT
CCR2_1_10_1_F	CACGATACTC TTGAACAAGGACGCATTTCCCCAG
CCR2_1_10_1_R	ATCGCAGAGA CAAAGACCCACTCATTTGCAGCAG
CCR2_1_10_2_F	AGAGTCTACA TTGAACAAGGACGCATTTCCCCAG
CCR2_1_10_2_R	TATCTCATAC CAAAGACCCACTCATTTGCAGCAG
ATF4F_1	GGTAGCATAGATCGC AGAAGTCCCGCCTCATAAGTGGAAG
ATF4R_1	CCATCCTACATATGA CATTTGAGTGATGGGGCCAAGTGAG
ATF4F_2	GGTAGTCGTATACGC AGAAGTCCCGCCTCATAAGTGGAAG
ATF4R_2	CCATCGTCGACACTA CATTTGAGTGATGGGGCCAAGTGAG
USP28F_1	GGTAGCTACATATGA CTTAAGCCATGGCGCTTCTCAGG
USP28R_1	CCATCTGTGTGCAGC GAAGGCATCCTCCTTGCTGTTATTGG
USP28F_2	GGTAGCTCGACTGCA CTTAAGCCATGGCGCTTCTCAGG
USP28R_2	CCATCGTCGACACTA GAAGGCATCCTCCTTGCTGTTATTGG
RNF157F_1	GGTAGCATAGATCGC AAGTGTCATCCAACGTGGTCAAAAGG
RNF157R_1	CCATCCTACATATGA CTTAAGCCATGGCGCTTCCCAC
RNF157F_2	GGTAGTCGTATACGC AAGTGTCATCCAACGTGGTCAAAAGG
RNF157R_2	CCATCCTCGACTGCA CTTAAGCCATGGCGCTTCCCAC
CTAG1BF_1	GGTAGCTACATATGA GCTTAAGCCGTAGCACTTCTCACAG
CTAG1BR_1	CCATCCTACATATGA CCAAAGAAAGCATCCTCCTTGCCATC
CTAG1BF_2	GGTAGGTCGACACTA GCTTAAGCCGTAGCACTTCTCACAG
CTAG1BR_2	CCATCGTCGACACTA CCAAAGAAAGCATCCTCCTTGCCATC

Chapter 5 supplementary information

Table 7: Target sequences of CRISPRs (Chapter 5).
Target genes, target site sequences, and PAM sequences are listed.

Gene	Storage Index	Protospacer Target (5' to 3')	PAM
HBB	R-01	GTGAACGTGGATGAAGTTGG	TGG
HBB	R-03	GACGTTCACCTTGCCCCACA	GGG
HBB	R-04	GCACGTTCACCTTGCCCCAC	AGG
HBB	R-05	GGTCTGCCGTTACTGCCCTG	TGG
HBB	R-06	GGTTACTGCCCTGTGGGGCA	AGG
HBB	R-07	GAGGTGAACGTGGATGAAGT	TGG
HBB	R-08	GCTGTGGGGCAAGGTGAACG	TGG
EGFP	R-19	GGTGGTGCAGATGAACTTCA	GGG
EGFP	R-20	GACCAGGATGGGCACCACCC	CGG
CCR5	R-25	GTGTTCATCTTTGGTTTTGT	GGG
CCR5	R-26	GCTGCCGCCCAGTGGGACTT	TGG
CCR5	R-27	GGCAGCATAGTGAGCCCAGA	AGG
CCR5	R-29	GTGAGTAGAGCGGAGGCAGG	AGG
CCR5	R-30	GTAGAGCGGAGGCAGGAGGC	GGG
ERCC5	R-31	GCCAAGCACTTAAAGGAGTC	CGG
ERCC5	R-33	GCAAGCACTTAAAGGAGTCC	GGG
ERCC5	R-35	GTGAGTTCCCATGGCGATCC	CGG
ERCC5	R-36	GCTATTGAAGAAACAGACTT	TGG
ERCC5	R-38	GATTTTCTATTGAGTTCCCA	TGG
ERCC5	R-39	GGAAACAAAGTGAGAAGATG	AGG
ERCC5	R-40	GCCTATTTTTGTGTTTGATG	GGG
TARDBP	R-41	GCAGAGCAGTTGGGGTATGA	TGG
TARDBP	R-42	GGCAGCACTACAGAGCAGTT	GGG
TARDBP	R-43	GCAGCACTACAGAGCAGTTG	GGG
TARDBP	R-44	GCCTGACTGGTTCTGCTGGC	TGG
HPRT1	R-52	GTTTGTGTCATTAGTGAAAC	TGG
HPRT1	R-53	GCAACTTGAACTCTCATCTT	AGG

REFERENCES

1. Streubel, J., Blucher, C., Landgraf, A. and Boch, J. (2012) TAL effector RVD specificities and efficiencies. *Nature biotechnology*, **30**, 593-595.
2. Reyon, D., Tsai, S.Q., Khayter, C., Foden, J.A., Sander, J.D. and Joung, J.K. (2012) FLASH assembly of TALENs for high-throughput genome editing. *Nat Biotechnol*, **30**, 460-465.
3. Lin, Y., Fine, E.J., Zheng, Z., Antico, C.J., Voit, R.A., Porteus, M.H., Cradick, T.J. and Bao, G. (2014) SAPTA: a new design tool for improving TALE nuclease activity. *Nucleic Acids Res*.
4. Schmid-Burgk, J.L., Schmidt, T., Kaiser, V., Höning, K. and Hornung, V. (2013) A ligation-independent cloning technique for high-throughput assembly of transcription activator–like effector genes. *Nat Biotechnol*, **31**, 76-81.
5. Cradick, T.J., Fine, E.J., Antico, C.J. and Bao, G. (2013) CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res*, **41**, 9584-9592.
6. 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. (2014) CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res*, **42**, 7473-7485.
7. Kim, Y., Kweon, J., Kim, A., Chon, J.K., Yoo, J.Y., Kim, H.J., Kim, S., Lee, C., Jeong, E., Chung, E. *et al.* (2013) A library of TAL effector nucleases spanning the human genome. *Nat Biotechnol*, **31**, 251-258.
8. Roe, A.M. and Shur, N. (2007) From new screens to discovered genes: The successful past and promising present of single gene disorders. *Am J Med Genet C*, **145C**, 77-86.
9. Zou, J., Mali, P., Huang, X., Dowey, S.N. and Cheng, L. (2011) Site-specific gene correction of a point mutation in human iPS cells derived from an adult patient with sickle cell disease. *Blood*, **118**, 4599-4608.
10. Lee, H.J., Kim, E. and Kim, J.S. (2010) Targeted chromosomal deletions in human cells using zinc finger nucleases. *Genome research*, **20**, 81-89.
11. Carlson, D.F., Tan, W., Lillico, S.G., Stverakova, D., Proudfoot, C., Christian, M., Voytas, D.F., Long, C.R., Whitelaw, C.B. and Fahrenkrug, S.C. (2012) Efficient TALEN-mediated gene knockout in livestock. *Proc Natl Acad Sci U S A*, **109**, 17382-17387.
12. Porteus, M.H. and Carroll, D. (2005) Gene targeting using zinc finger nucleases. *Nat Biotechnol*, **23**, 967-973.
13. UNAIDS. (2009). UNAIDS/WHO.
14. Bowman, M.C., Archin, N.M. and Margolis, D.M. (2009) Pharmaceutical approaches to eradication of persistent HIV infection. *Expert reviews in molecular medicine*, **11**, e6.
15. Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A.J. and Voytas, D.F. (2010) Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics*, **186**, 757-761.

16. Galetto, R., Duchateau, P. and Paques, F. (2009) Targeted approaches for gene therapy and the emergence of engineered meganucleases. *Expert opinion on biological therapy*, **9**, 1289-1303.
17. Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819-823.
18. Li, D., Qiu, Z., Shao, Y., Chen, Y., Guan, Y., Liu, M., Li, Y., Gao, N., Wang, L., Lu, X. *et al.* (2013) Heritable gene targeting in the mouse and rat using a CRISPR-Cas system. *Nat Biotechnol*, **31**, 681-683.
19. Schornack, S., Meyer, A., Römer, P., Jordan, T. and Lahaye, T. (2006) Gene-for-gene-mediated recognition of nuclear-targeted AvrBs3-like bacterial effector proteins. *J Plant Physiol*, **163**, 256-272.
20. Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A. and Bonas, U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509-1512.
21. Boch, J. and Bonas, U. (2009) Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu Rev Phytopathol*, **48**, 419-436.
22. Moscou, M.J. and Bogdanove, A.J. (2009) A simple cipher governs DNA recognition by TAL effectors. *Science*, **326**, 1501.
23. 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. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res*.
24. 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. *et al.* (2010) A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol*, **29**, 143-148.
25. Tesson, L., Usal, C., Ménoret, S., Leung, E., Niles, B.J., Remy, S., Santiago, Y., Vincent, A.I., Meng, X., Zhang, L. *et al.* (2011) Knockout rats generated by embryo microinjection of TALENs. *Nat Biotechnol*, **29**, 695-696.
26. Li, T., Huang, S., Jiang, W.Z., Wright, D., Spalding, M.H., Weeks, D.P. and Yang, B. (2010) TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res*, **39**, 359-372.
27. Mahfouz, M.M., Li, L., Shamimuzzaman, M., Wibowo, A., Fang, X. and Zhu, J.K. (2011) De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. *Proc Natl Acad Sci U S A*, **108**, 2623-2628.
28. Wood, A.J., Lo, T.W., Zeitler, B., Pickle, C.S., Ralston, E.J., Lee, A.H., Amora, R., Miller, J.C., Leung, E., Meng, X. *et al.* (2011) Targeted genome editing across species using ZFNs and TALENs. *Science*, **333**, 307.
29. Sander, J.D., Cade, L., Khayter, C., Reyon, D., Peterson, R.T., Joung, J.K. and Yeh, J.R. (2011) Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nat Biotechnol*, **29**, 697-698.
30. Huang, P., Xiao, A., Zhou, M., Zhu, Z., Lin, S. and Zhang, B. (2011) Heritable gene targeting in zebrafish using customized TALENs. *Nat Biotechnol*, **29**, 699-700.

31. Mussolino, C., Morbitzer, R., Lütge, F., Dannemann, N., Lahaye, T. and Cathomen, T. (2011) A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res*, **39**, 9283-9293.
32. Hockemeyer, D., Wang, H., Kiani, S., Lai, C.S., Gao, Q., Cassady, J.P., Cost, G.J., Zhang, L., Santiago, Y., Miller, J.C. *et al.* (2011) Genetic engineering of human pluripotent cells using TALE nucleases. *Nat Biotechnol*, **29**, 731-734.
33. Bolotin, A., Quinquis, B., Sorokin, A. and Ehrlich, S.D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, **151**, 2551-2561.
34. Horvath, P. and Barrangou, R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science*, **327**, 167-170.
35. Marraffini, L.A. and Sontheimer, E.J. (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet*, **11**, 181-190.
36. Garneau, J.E., Dupuis, M., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H. and Moineau, S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67-71.
37. Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M. and Terns, M.P. (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*, **139**, 945-956.
38. Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. and Koonin, E.V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct*, **1**, 7.
39. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. and Horvath, P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709-1712.
40. Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V. and van der Oost, J. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960-964.
41. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816-821.
42. Mali, P., Esvelt, K.M. and Church, G.M. (2013) Cas9 as a versatile tool for engineering biology. *Nat Methods*, **10**, 957-963.
43. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823-826.
44. Yang, H., Wang, H., Shivalila, C.S., Cheng, A.W., Shi, L. and Jaenisch, R. (2013) One-Step Generation of Mice Carrying Reporter and Conditional Alleles by CRISPR/Cas-Mediated Genome Engineering. *Cell*, **154**, 1370-1379.
45. Xie, K. and Yang, Y. (2013) RNA-Guided Genome Editing in Plants Using a CRISPR-Cas System. *Mol Plant*.

46. Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.R. and Joung, J.K. (2013) Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat Biotechnol*, **31**, 227-229.
47. Cho, S.W., Kim, S., Kim, J.M. and Kim, J.S. (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat Biotechnol*, **31**, 230-232.
48. Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., Zhang, K., Liu, J., Xi, J.J., Qiu, J.L. *et al.* (2013) Targeted genome modification of crop plants using a CRISPR-Cas system. *Nat Biotechnol*, **31**, 686-688.
49. Morbitzer, R., Römer, P., Boch, J. and Lahaye, T. (2010) Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors. *Proceedings of the National Academy of Sciences*, **107**, 21617-21622.
50. Cong, L., Zhou, R., Kuo, Y.C., Cunniff, M. and Zhang, F. (2012) Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature communications*, **3**, 968.
51. Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K. and Sander, J.D. (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol*, **31**, 822-826.
52. 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. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol*, **31**, 827-832.
53. Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A. and Liu, D.R. (2013) High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat Biotechnol*, **31**, 839-843.
54. Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L. and Church, G.M. (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol*, **31**, 833-838.
55. Cho, S.W., Kim, S., Kim, Y., Kweon, J., Kim, H.S., Bae, S. and Kim, J.S. (2014) Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res*, **24**, 132-141.
56. Zhang, F., Cong, L., Lodato, S., Kosuri, S., Church, G.M. and Arlotta, P. (2011) Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat Biotechnol*, **29**, 149-153.
57. Briggs, A.W., Rios, X., Chari, R., Yang, L., Zhang, F., Mali, P. and Church, G.M. (2012) Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. *Nucleic Acids Res*, **40**, e117.
58. Guschin, D.Y., Waite, A.J., Katibah, G.E., Miller, J.C., Holmes, M.C. and Rebar, E.J. (2010) A rapid and general assay for monitoring endogenous gene modification. *Methods Mol Biol*, **649**, 247-256.
59. 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. *et al.* Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology. *Nucleic Acids Res*, **38**, e152.

60. Rouet, P., Smih, F. and Jasin, M. (1994) Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proc Natl Acad Sci U S A*, **91**, 6064-6068.
61. Cradick, T.J., Fine, E.J., Antico, C.J. and Bao, G. (2013) CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res*.
62. Helleday, T., Lo, J., van Gent, D.C. and Engelward, B.P. (2007) DNA double-strand break repair: from mechanistic understanding to cancer treatment. *DNA Repair (Amst)*, **6**, 923-935.
63. Porteus, M.H. (2006) Mammalian gene targeting with designed zinc finger nucleases. *Molecular therapy : the journal of the American Society of Gene Therapy*, **13**, 438-446.
64. 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. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res*, **39**, e82.
65. Fine, E.J., Cradick, T.J., Zhao, C.L., Lin, Y. and Bao, G. (2014) An online bioinformatics tool predicts zinc finger and TALE nuclease off-target cleavage. *Nucleic Acids Res*, **42**, e42.
66. Mussolino, C., Alzubi, J., Fine, E.J., Morbitzer, R., Cradick, T.J., Lahaye, T., Bao, G. and Cathomen, T. (2014) TALENs facilitate targeted genome editing in human cells with high specificity and low cytotoxicity. *Nucleic Acids Res*.
67. Meckler, J.F., Bhakta, M.S., Kim, M.S., Ovadia, R., Habrian, C.H., Zykovich, A., Yu, A., Lockwood, S.H., Morbitzer, R., Elsäesser, J. *et al.* (2013) Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Res*, **41**, 4118-4128.
68. 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. (2012) Targeting G with TAL Effectors: A Comparison of Activities of TALENs Constructed with NN and NK Repeat Variable Di-Residues. *PLoS One*, **7**, e45383.
69. Voit, R.A., Hendel, A., Pruett-Miller, S.M. and Porteus, M.H. (2014) Nuclease-mediated gene editing by homologous recombination of the human globin locus. *Nucleic Acids Res*, **42**, 1365-1378.
70. Lei, Y., Guo, X., Liu, Y., Cao, Y., Deng, Y., Chen, X., Cheng, C.H., Dawid, I.B., Chen, Y. and Zhao, H. (2012) Efficient targeted gene disruption in Xenopus embryos using engineered transcription activator-like effector nucleases (TALENs). *Proc Natl Acad Sci U S A*, **109**, 17484-17489.
71. Ding, Q., Lee, Y.K., Schaefer, E.A., Peters, D.T., Veres, A., Kim, K., Kuperwasser, N., Motola, D.L., Meissner, T.B., Hendriks, W.T. *et al.* (2013) A TALEN genome-editing system for generating human stem cell-based disease models. *Cell Stem Cell*, **12**, 238-251.
72. 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. *et al.* (2013) TALEN-based gene correction for epidermolysis bullosa. *Mol Ther*, **21**, 1151-1159.

73. Fine, E.J., Cradick, T.J., Zhao, C.L., Lin, Y. and Bao, G. (2013) An online bioinformatics tool predicts zinc finger and TALE nuclease off-target cleavage. *Nucleic Acids Res*, in press.
74. Deng, D., Yin, P., Yan, C., Pan, X., Gong, X., Qi, S., Xie, T., Mahfouz, M., Zhu, J.K., Yan, N. *et al.* (2012) Recognition of methylated DNA by TAL effectors. *Cell Res*.
75. Valton, J., Dupuy, A., Daboussi, F., Thomas, S., Maréchal, A., Macmaster, R., Melliand, K., Juillerat, A. and Duchateau, P. (2012) Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. *J Biol Chem*, **287**, 38427-38432.
76. R Core Team. (2013). R Foundation for Statistical Computing, Vienna, Austria.
77. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133-138.
78. Doyle, E.L., Booher, N.J., Standage, D.S., Voytas, D.F., Brendel, V.P., Vandyk, J.K. and Bogdanove, A.J. (2012) TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Res*, **40**, W117-122.
79. Neff, K.L., Argue, D.P., Ma, A.C., Lee, H.B., Clark, K.J. and Ekker, S.C. (2013) Mojo Hand, a TALEN design tool for genome editing applications. *BMC bioinformatics*, **14**, 1.
80. Heigwer, F., Kerr, G., Walther, N., Glaeser, K., Pelz, O., Breinig, M. and Boutros, M. (2013) E-TALEN: a web tool to design TALENs for genome engineering. *Nucleic Acids Res*.
81. Heigwer, F., Kerr, G., Walther, N., Glaeser, K., Pelz, O., Breinig, M. and Boutros, M. (2013) E-TALEN: a web tool to design TALENs for genome engineering. *Nucleic Acids Research* published online September 3, 2013.
82. Jiang, W., Bikard, D., Cox, D., Zhang, F. and Marraffini, L.A. (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol*, **31**, 233-239.
83. Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P. and Siksnys, V. (2011) The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res*, **39**, 9275-9282.
84. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2007) *Molecular Biology of the Cell*. Garland Science, New York.
85. 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. *et al.* (2013) Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell*, **154**, 1380-1389.
86. Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R., Heitner, S.G. *et al.* (2013) ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.*, **41**, D56-63.
87. Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P. *et al.* (2012) ChIP-seq

guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*, **22**, 1813-1831.
88. Sugimoto, N., Nakano, S., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., Yoneyama, M. and Sasaki, M. (1995) Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry*, **34**, 11211-11216.
89. Chen, H., Choi, J. and Bailey, S. (2014) Cut Site Selection by the Two Nuclease Domains of the Cas9 RNA-guided Endonuclease. *J Biol Chem*.
90. 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. *et al.* (2013) DNA Targeting Specificity of the RNA-guided Cas9 Nuclease. *Nat Biotech*.
91. Porteus, M.H. and Baltimore, D. (2003) Chimeric nucleases stimulate gene targeting in human cells. *Science*, **300**, 763.
92. Iseli, C., Ambrosini, G., Bucher, P. and Jongeneel, C.V. (2007) Indexing strategies for rapid searches of short words in genome sequences. *PLoS One*, **2**, e579.
93. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589-595.
94. Fan, L., Kadura, I., Krebs, L.E., Hatfield, C.C., Shaw, M.M. and Frye, C.C. (2012) Improving the efficiency of CHO cell line generation using glutamine synthetase gene knockout cells. *Biotechnol Bioeng*, **109**, 1007-1015.
95. Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G. *et al.* (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**, 84-87.
96. Cade, L., Reyon, D., Hwang, W.Y., Tsai, S.Q., Patel, S., Khayter, C., Joung, J.K., Sander, J.D., Peterson, R.T. and Yeh, J.R. (2012) Highly efficient generation of heritable zebrafish gene mutations using homo- and heterodimeric TALENs. *Nucleic Acids Res*, **40**, 8001-8010.
97. Tsai, S.Q., Wyvekens, N., Khayter, C., Foden, J.A., Thapar, V., Reyon, D., Goodwin, M.J., Aryee, M.J. and Joung, J.K. (2014) Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat Biotechnol*, **32**, 569-576.
98. Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F. and Marraffini, L.A. (2013) Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res*, **41**, 7429-7437.
99. Maeder, M.L., Angstman, J.F., Richardson, M.E., Linder, S.J., Cascio, V.M., Tsai, S.Q., Ho, Q.H., Sander, J.D., Reyon, D., Bernstein, B.E. *et al.* (2013) Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat Biotechnol*, **31**, 1137-1142.

VITA

Yanni Lin

Yanni was born in Haikou, Hainan Province, China. Her hometown is a tropical island with beautiful oceans and tasty tropical fruits. She attended public schools in Haikou, Hainan Province, received a B.S. in Biological Science and Biotechnology from Tsinghua University, Beijing, China in 2007 and a M.S. in Biochemistry from University of Illinois at Urbana-Champaign in 2009 before coming to Georgia Tech / Emory University to pursue a doctorate in Biomedical Engineering. In colleague, she played the clarinet in the Tsinghua University Military Band (THUMB), while her husband played the trumpet. When she is not working on her research, Yanni enjoys watching random movies while eating hot pot meals with her beloved husband Weixuan Chen and her two cats, Simba and Little Monster.