이학박사학위논문

A genome-wide library of TAL effector nucleases

2014년 2월

서울대학교 대학원

화학부 생화학 전공

김 용 섭

A genome-wide library of TAL effector nucleases

지도교수 김진수

이 논문을 이학박사학위논문으로 제출함
2013년 11월

서울대학교 대학원
화학과 생화학전공
김 용 섭

김용섭의 박사학위논문을 인준함
2013년 11월

위 원 장 _____ (인)
부 위 원 장 _____ (인)
위 원 _____ (인)
위 원 _____ (인)
위 원 _____ (인)

Abstract

A genome-wide library of TAL effector nucleases

Yongsub Kim

Department of Chemistry

The Graduate School

Seoul National University

Transcription activator–like (TAL) effector nucleases (TALENs) are newly developed programmable nucleases which use a simple 'protein–DNA code' that relates modular DNA-binding TALE repeat domains to individual bases in a target binding site. Unlike homing endonucleases and zinc figer nucleases (ZFNs), they can be readily engineered to bind specific genomic loci, enabling the introduction of precise genetic modifications such as gene disruptions, additions,

corrections and genome rearrangements. In this thesis, I developed improved TALEN architectures to avoid unwanted mutations in genome. Then, I carefully chose genome-wide TALEN target sites that did not have highly similar sequences elsewhere in the genome and assembled TALEN pairs using high throughput Golden Gate cloning system. A pilot test including over a hundred pairs of TALENs showed that all TALENs were active and disrupted their target genes at high frequencies, although two of these TALENs became active only after their target sites were partially demethylated using a DNA methyltransferase inhibitor. I used the TALEN library to generate single- and double-gene knockout cells in which NF-kB signaling pathways were disrupted. Compared with cells treated with short interfering RNAs (siRNA), these cells showed unambiguous suppression of the signal transduction. Furthermore, I developed the TALEN library for targeting every exon in protein coding genes of several organisms including human. The TALEN library reported here will be broadly useful for research and drug discovery.

Keywords : TAL effector nuclease (TALEN), Double-strand breaks (DSB), Non-homologous end-joining (NHEJ), Genome engineering.

Student Number : 2008-22719

Table of Contents

List of Figures

List of Tables

List of Abbreviations

DSB	Double-strand break
FACS	Fluorescence-activated cell sorting
HEK	Human embryonic kidney
HR	Homologous recombination
KO	Knockout
NHEJ	Non-homologous end-joining
PCR	Polymerase chain reaction
T7E1	T7 Endonuclease 1
ZFN	Zinc-finger nuclease
TALEN	Transcription activator-like effector nuclease
TALE	TAL Effector
RNAi	RNA interference
siRNA	short interfering RNA
RGEN	RNA-Guided Endonuclease

I. Introduction

In the post-genome era, although human genome has been sequenced (Lander et al. 2001; Venter et al. 2001), the precise function of most genes is still unknown. Hence, one of the most challenging tasks in biological research is investigation the function of these genes. Important contributions to our understanding of gene functions often derive from knockout studies that a specific gene is silenced in cells or organisms achieved by gene targeting through homologous recombination (HR) (Smithies et al. 1985), an endogenous DNA double strand break (DSB) repair mechanism. Although gene targeting is widely used in mouse embryonic stem cells to create gene knockout animals, the efficiency of HR is extremely low in mammalian and other higher eukaryotic cells, ranging from 10^{-6} to 10^{-7} (Deng and Capecchi 1992).

In a decade ago, attractive alternative to the above described techniques was established, called RNA interference (RNAi). RNAi, discovered in the nematode *Caenorhabditis elegans,* is used to suppress the expression of a gene of interest by artificial short interfering RNAs (siRNAs) (Fire et al. 1998). Despite of the broad utility of this approach in basic research and drug discovery, siRNAs are limited by many factors: 1) siRNAs induce gene knockdown rather than gene knockout. Thus, small fraction of the activities of target genes remains after siRNA treatment (Krueger et al. 2007). 2) siRNAs are not specific, displaying sequence dependent off target effects (Jackson et al.

2003). Only several nucleotide matches between an siRNA seed sequence and the 3' untranslated region (UTR) of an off target gene can trigger suppression of gene expression (Birmingham et al. 2006). 3) siRNAs can broadly affect cellular physiology by activating an innate immune response and competing with endogenous microRNAs (miRNAs) for proteins such as Dicer and the RNA-induced silencing complex (RISC), which are critical for miRNA function (Sledz et al. 2003;Khan et al. 2009). 4) Furthermore, many genes are refractory to inhibition via siRNAs (Krueger et al. 2007).

Engineered nucleases such as zinc finger nucleases (ZFNs) (Bibikova et al. 2003; Urnov et al. 2005; Doyon et al. 2008; Maeder et al. 2008), TAL effector nucleases (TALENs) (Cermak et al. 2011; Miller et al. 2011; Sander et al. 2011) and CRISPR/Cas9–based RNA-guided endonucleases (RGENs) (Cho et al. 2013; Cong et al. 2013; Hwang et al. 2013; Mali et al. 2013), enable the targeted alteration precisely at a predetermined locus. These programmable nucleases induce site specific DSBs in a genome, which are then repaired by endogenous mechanisms that can be exploited to create sequence alterations at the cleavage site. ZFNs and TALENs, although they share the same FokI-derived nuclease domain, differ in that they employ distinctive DNA binding arrays: ZFNs use zinc finger arrays (Kim et al. 1996) and TALENs use TAL effector (TALE) repeat arrays (Miller et al. 2011). Because these arrays recognize target DNA sequences in a modular fashion, tailor-made DNA-binding arrays with desired specificities can be constructed by mixing-and-matching pre-characterized

modules (Kim et al. 2011).

In this thesis, I described novel TALEN architectures for genome engineering using improved fluorescent reporter system (Kim et al. 2013b). This TALEN architecture showed high activity at target site with 12- to 14-bp spacers comparable with TALEN constructs in previous study. And then, to prepare a collection of TALENs that is designed to target every protein coding gene in the human genome, I developed one-step Golden Gate cloning system to assemble TALEN plasmids and computational strategy to search for unique TALEN target sites in each gene to avoid potential off target mutations. Nearly 100% of TALENs induced site specific mutations at high frequencies. Only two TALENs failed to show any measurable mutagenesis activities. Each of two TALEN sites was turned out to be heavily methylated. With this TALEN library, I generated single- and double-gene knockout cells in which NF-κB signaling pathways were disrupted. Furthermore, I designed TALEN library for several organisms targeting every exon of protein coding genes and showed that almost TALENs were highly active. Finally, I investigated difference of mutation signature between TALENs and ZFNs. I found that ZFN-induced insertion frequency was much higher than the TALEN-induced insertion frequency. From these result, the TALEN library resources will be broadly useful for research and drug discovery.

II. Materials and Methods

1. Dual fluorescent reporter plasmid construction

The amplified product of encoding EGFP sequence from pEGFP-N1 using Primer A and B was cloned into the NotI site of the pRGS plasmid (Kim et al. 2011). This vector was named pRG2S. Oligonucleotides that contained each target sites were synthesized (Macrogen, Seoul, South Korea) and annealed *in vitro*. The annealed oligonucleotides were ligated into the vector (pRG2S) digested with EcoR1 and BamH1. As previous pRGS plasmid, note that an in-frame stop codon should be avoided at the target site either by altering the frame or by changing the orientation of the target site.

2. Cell culture and transfection

HEK293T/17 (ATCC, CRL-11268) and HeLa cells (ATCC, CCL-2) were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 100 units/mL penicillin, 100 μg/mL streptomycin, 0.1 mM nonessential amino acids, and 10% FBS (FBS). We transfected 200,000 HEK293 cells using 3 μl of polyethylenimine and 1 μg of plasmid DNA in 24-well plates. We transfected 100,000 HeLa cells with Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol.

3. High-throughput assembly of TALENs

All steps in TALEN assembly were performed in 96-well plates. In each plate, 47 pairs of TALENs were assembled and one pair of negative control (FokI vector alone) was included. Our one-step Golden-Gate system consisted of 424 TALE array plasmids (6 × 64 tripartite arrays, 2 × 16 bipartite arrays, and 2 × 4 monopartite arrays). For convenience, we numbered each TALE array as follows. We used these numbers to choose appropriate arrays to assemble TALEN plasmids.

For example, the half-site sequence, "5'-TGGGGGAGGTGGCG AGGAAC", can be divided into 8 parts (the first T, GGG, GGA, GGT, GGC, GAC, GAA, and the last C). The first T and last C are not recognized by TALE arrays. To assemble a TALEN subunit specific to this sequence, we chose the following arrays: position1-GGG + position2-GGA + position3-GGT + position4-GGC + position5-GAC + position6-GAA + the FokI expression vector that contains C-specific half-repeat. A detailed protocol is shown below:

1) Six TALE array plasmids and a FokI expression vector are mixed in each well as follows:

1.0 μl	TALE array vectors (50ng/μl each)
0.5 μl	FokI expressing vector (50ng/μl)
0.5 μl	BsaI (New England BioLabs, 10U/μl)
2.0 μl	10× T4 DNA Ligase Reaction Buffer
0.1 μl	T4 DNA Ligase (New England BioLabs, 2000U/μl)
10.9 μl	ddH_2O

In a 20μl restriction-ligation reaction.

2) The restriction-ligation reaction is carried out in a thermocycler as follows:

37℃ 5 min] → 20 cycles

16℃ 5 min

50℃ 15 min

80℃ 5 min

3) After the thermocycling reaction, the reaction mixture (6 μl) from each well is used for transformation of chemically-competent DH5 α cells (30 μl). Subsequently, cells are inoculated in Flat-Bottom Blocks (Qiagen) that are filled with LB medium (800 μl) containing ampicillin (50μg/ml). The transformants in 96-well blocks are incubated overnight at 37℃ with vigorous shaking.

4) Two sets of E. coli stocks are prepared by mixing the LB culture (50 μl) with 60% glycerol (150 μl); they are separately stored at -80℃.

4. Flow cytometry (FACS)

Transfected HEK293T/17 and HeLa cells were trypsinized and resus-pended in growth media. Single-cell suspensions were analyzed and sorted using the FACSCanto (BD Biosciences). Untransfected cells and cells transfected with reporters alone were used as controls.

5. T7E1 assay for mutation detection

HEK293T/17 cells ($2x10^5$) pre-cultured in a 24 well plate were

transfected with two plasmids encoding a TALEN pair (500ngeach). After 72h of incubation, genomicDNA was extracted from the transfected cells using the G-dex™ Genomic DNA Extraction Kit (iNtRON BIOTECH NOLOGY, Seongnam, S.Korea). Purified genomic DNA samples were subjected to the T7E1 assay as described previously (Kim et al. 2009). Briefly, genomic region around TALEN target site was amplified, melted, and annealed to form heterodupelx DNA. The annealed DNA was treated with 5 units of T7E1 endonuclease I (T7E1) for 20 min at 37℃. DNA was analyzed by agarose gel electrophoresis. PCR primers used for the T7E1 assay in previous study (Kim et al. 2013b).

6. PCR analysis to detect genomic mutations and sequencing

Genomic DNA (50 ng per reaction) was subjected to PCR analysis using Taq DNA polymerase (GeneAll Biotech, Seoul, Korea) and appropriate primers as described in previous study (Kim et al 2013b). For sequencing analysis, PCR products corresponding to genomic rearrangements were purified using the QIAquick Gel Extraction Kit (Quiagen, Valencia, CA) and cloned into the T-Blunt vector using the T-Blunt PCR Cloning Kit (SolGent, Daejeon, S. Korea). Cloned PCR products were sequenced using the M13 primer or primers used for PCR amplification.

7. Analysis and rescue of genome-inactive TALENs

One day before TALEN plasmid transfection, HEK293 cells (two million) were pretreated with 0.2 μM 5-aza-dC or 100 ng/ml trichostatin A in 24-well plates. After 3 days of incubation, genomic DNA was isolated from transfected cells and subjected to the T7E1 assay. To determine whether the TALEN sites were methylated, genomic DNA was treated with bisulfite using the EpiTect Bisulfite kit (QIAGEN) according to the manufacturer's protocol. The bisulfite-converted DNA was then amplified using bisulfite-specific primers, and the amplified products were subcloned into the T-Blunt vector and sequenced.

8. Digital PCR to estimate mutation frequency

Digital PCR analysis was performed based on previous study (Lee et al. 2010). Genomic DNA samples were quantified and serially diluted with Tris-EDTA buffer. The serially diluted genomic DNA were amplified using appropriate primers. The frequencies of mutation were calcuated as previous study.

9. Gene-knockout cell lines

HEK293T/17 and HeLa cells were co-transfected with TALEN plasmids and surrogate reporter plasmids that contain the TALEN target site. Gene-knockout cells were enriched by selection as described (Kim

et al. 2013a) and cloned by limiting dilution in 96-well plates. Typically, cells were maintained for 2 weeks in 96-well plates to isolate single clones. These clones were analyzed using T7E1, fPCR and dideoxy sequencing.

10. Fluorescent PCR analysis

Genomic DNA was extracted from each single clones and subjected to fluorescence PCR using 5'-carboxyfluorescein-labeled primers. Each PCR amplicons were analyzed using an ABI 3730xl DNA analyzer. The positions of peaks indicate the lengths of PCR products.

11. Episomal reporter assay to detect NF-κB signaling

The NF-κB–dependent firefly luciferase reporter was constructed by placing three tandem copies of the NF-κB recognition element (TGGGGACTTTCCGC) (Duan et al. 2005) in front of a synthetic promoter that consists of the TATA-box and the initiator element. Gene-knockout or wild-type cells were co-transfected with the luciferase reporter plasmid and the *Renilla* luciferase plasmid. After 24 h of incubation, cells were treated with TNFα (1 ng/ml) or IL-1β (25 ng/ml) and incubated for 15 h. Cells were lysed in 1× lysis buffer (100 μl) (Promega), and the dual luciferase assays were done according to the manufacturer's protocol.

12. Western blotting

HEK293 knockout clones were lysed and the lysates were eletrophoresed on a 7% SDS-PAGE gel. Primary antibodies specific for *TNFR1* (1:200, Santa cruz biotechnology) or *GAPDH* (1:200, Santa cruz biotechnology) and anti-mouse secondary antibody (1:1000, Santa cruz biotechnology) were used. Immunoreactive bands were visualized using the ECL method.

III. Results

A. Optimization of TAL Effector Nucleases (TALENs)

1. Dual fluorescent reporter system

First of all, to measure genome-editing activity of engineered nucleases, I developed dual fluorescent reporter system (Figure 1). In the previous study, surrogate reporter systems were developed for detection of engineered nuclease's activity and enrichment of mutant cells (Kim et al. 2011). The reporter plasmid contains nuclease's target site between the RFP- and GFP- encoding DNA sequences. Because the GFP sequence fused to the RFP sequence is out of frame, the cells transfected with the reporter plasmid and inactive nuclease-encoding plasmid express RFP only. In contrast, functional GFP is expressed only when nucleases induce DSBs at target site, whose repair via error-prone NHEJ lead to indels that often result in frameshifting mutations. Although this system is valuable resource for nuclease mediated genome editing, it is limited that only one third of the reporter plasmids, which introduced +1 or +2 frameshift, can produce in-frame frameshifting. Actually, some active TALENs did not generate functional GFP proteins in this reporter system (Figure 2). In this regard, I developed improved fluorescence reporter for detecting nuclease's activity by fusing another GFP sequence at the behind of

GFP sequence as other frame. Nuclease-mediated indels formation generate two different frame shift. The +1 frameshifting mutation makes the former GFP protein functional, The +2 frameshifting mutation makes latter GFP protein functional. In this dual fluorescent reporter assay, TALENs which were inactive in the previous reporter assay were highly active (Figure 2). This result showed that dual fluorescent reporter system is valuable for detecting engineered nucleases activity.

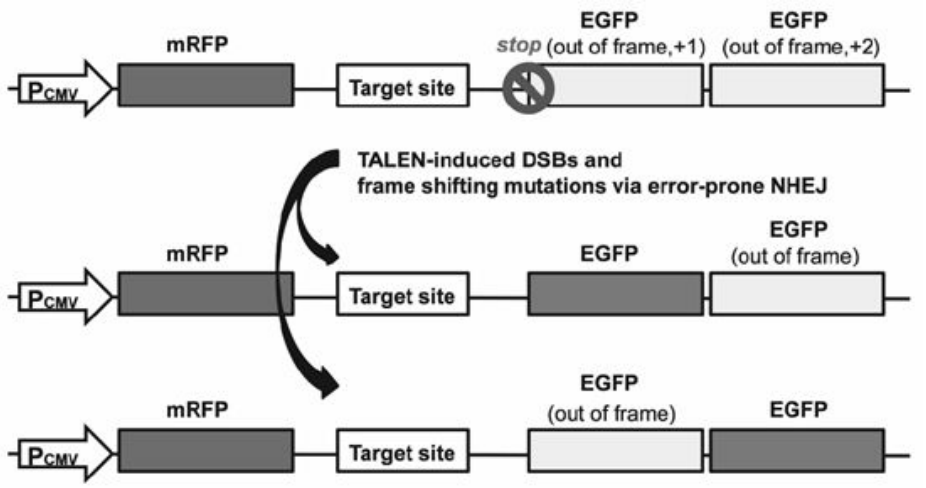
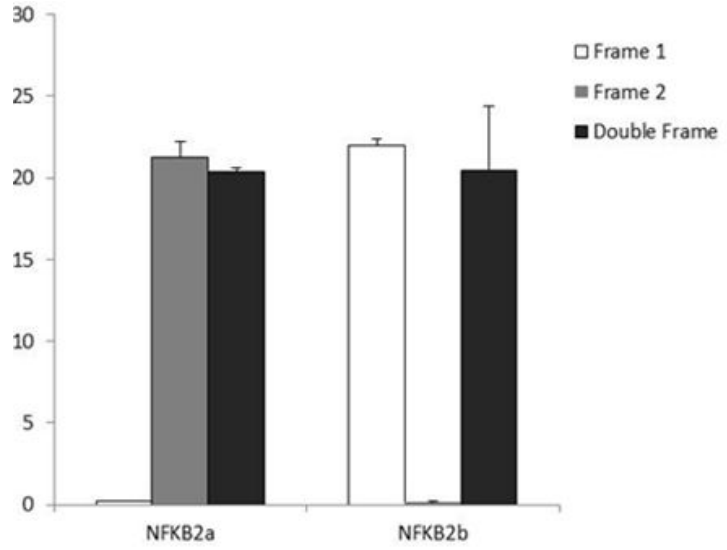


Figure 1. Scheme of dual fluorecent reporter-based assay. The reporter consists of the mRFP gene, the engineered nuclease's target sequence and the two EGFP genes. mRFP is constitutively expressed by the CMV promoter, wheras functional EGFP is not expressed because it is out of frame. Nuclease-mediated indel formation in target sites generate two different frame shift. The +1 frameshifting mutation makes the former GFP protein functional, The +2 frameshifting mutation makes the latter GFP protein linked with fused non-functional amino acid functional.

a



b

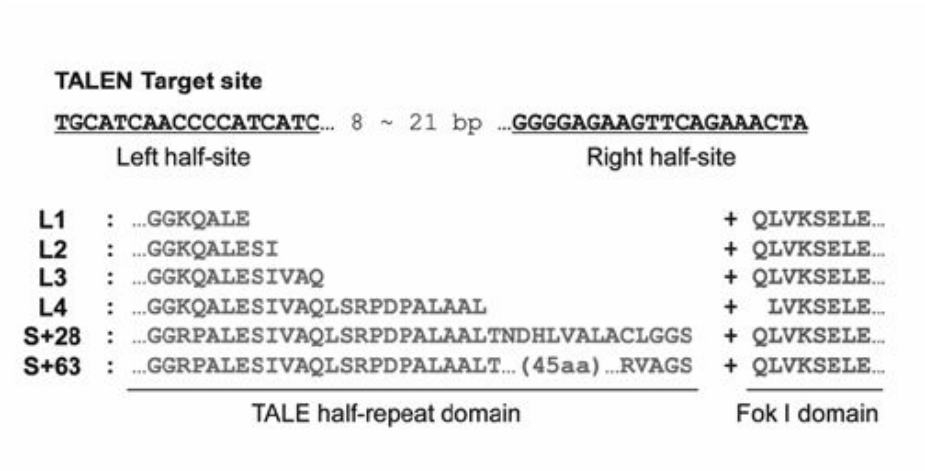
Label	Target sequence
NFkB2a	tcgggggtggctcccacatgggtggaggctctgggggtgcagccgggggcta
NFkB2b	tcgactacggcgtcaccgcggacgcgcgcgcgctgctggcgggacagcgcca

Figure 2. Improved results of dual flourescent reporter assay for measuring activities of engineered nucleases. (a) Two TALEN pairs were not active in RGS reporter assay (Kim et al. 2011). But in other frame reporters and RG2S reporters, these TALENs we highly active. (b) Target seuquence of TALEN pairs in this assay.

2. Design of prototype TALENs

In the previous study, TAL effectors fused FokI nuclease domain can induce DSBs in human cells and leads to gene disruption as ZFNs. Although TALENs are highly active than ZFNs, their broad spacer between two TALEN monomer which could induce unwanted mutation in genome is barrier for gene therapy or other applications (Christian et al. 2010; Li et al. 2011; Miller et al. 2011). To optimize the fusion junction between a TALE domain and the FokI nuclease domain, I prepared a series of TALE-FokI fusions with different junctions by linking each TALE to various amino acid residues in the appropriate region of the FokI nuclease domain (Figure 3a). I tested these TALE-FokI fusions using the dual reporter system with variable spacers between two TALEN monomer binding sites. As the optimized TALENs, I chose L4 that showed high activity at target sites with 12- to 14-bp spacers but no or negligible activity at sites with < 12-bp or > 14-bp spacers (Figure 3b). Compared to the two original TALEN constructs that contain additional amino acid residues between the TAL effector array and the FokI sequence (S+28 and S+63) (Miller et al. 2011), L4 TALEN constructs showed nuclease activities at sites with a narrow range of spacers, a desirable property for high specificity.

a



b

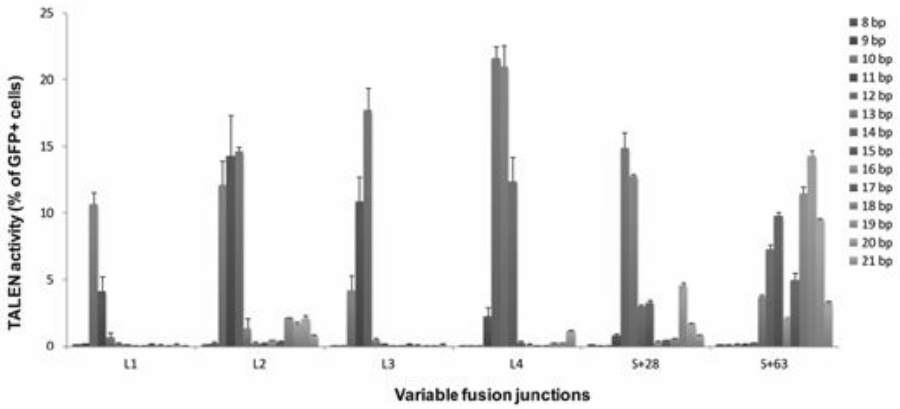


Figure 3. Optimization of TALEN architectures. (a) TALEN target site and amino-acid sequences in the fusion junctions that connect the TAL effector array to the FokI cleavage domain. (b) Comparison of TALEN gene-editing activities. Reporter plasmids that contain target sites with variable spacers and TALEN plasmids were co-transfected into HEK293 cells, and GFP+ cells were counted by flow cytometry. S+28 and S+63

are two prototype TALEN architectures previously reported (Miller et al. 2011). Error bars indicated s.e.m. from at least three independent experiments.

B. Human genome-wide TALEN library

1. One-step Golden-Gate cloning system

I developed a one-step Golden-Gate cloning system to assemble TALEN plasmids with variable lengths in a high-throughput manner (Figure 4). Although Golden-Gate cloning methods have been used for the assembly of TALEN plasmids previously (Cermak et al. 2011; Li et al. 2011; Morbitzer et al. 2011; Weber et al. 2011; Zhang et al. 2011), these methods either rely on the use of PCR, or require gel isolation of DNA segments, or require at least two rounds of subcloning steps. First of all, I used four TAL effector repeat domains termed NI, NN, NG and HD repeat variable di-residues (RVDs), each specific to one of the four bases (A, G, T and C, respectively), to make the cloning system (Boch et al. 2009; Moscou and Bogdanove 2009). These repeat domains consisted of 34 amino acid residues with similar sequences; the RVDs at positions 12 and 13 determine the base specificities. Next, I designed 64 tripartite-, 16 bipartite- and 4 monopartite- TAL effector repeat domain arrays with minimum sequence similarity as following criteria; i) removed human's rare codon (< 10%) and generated ii) 192 (64 × 3) monopartite with minimum similarity (max. 81.25%), iii) 64 tripartite, 16 bipartite with minimum similarity in single arrays (max. 72%). iv) Then, these arrays were subcloned to 6 subgroups via PCR-based mutagenesis. Therefore, I prepared a total of 424 TAL effector array plasmids (6 × 64 tripartite arrays, 2 × 16 bipartite arrays,

and 2 × 4 monopartite arrays). Then, to make engineered nucleases, I prepared 8 obligatory heterodimeric FokI-encoding plasmids (DAS/RR) with L4 fusion junction in Figure 3 (Guo et al. 2010).

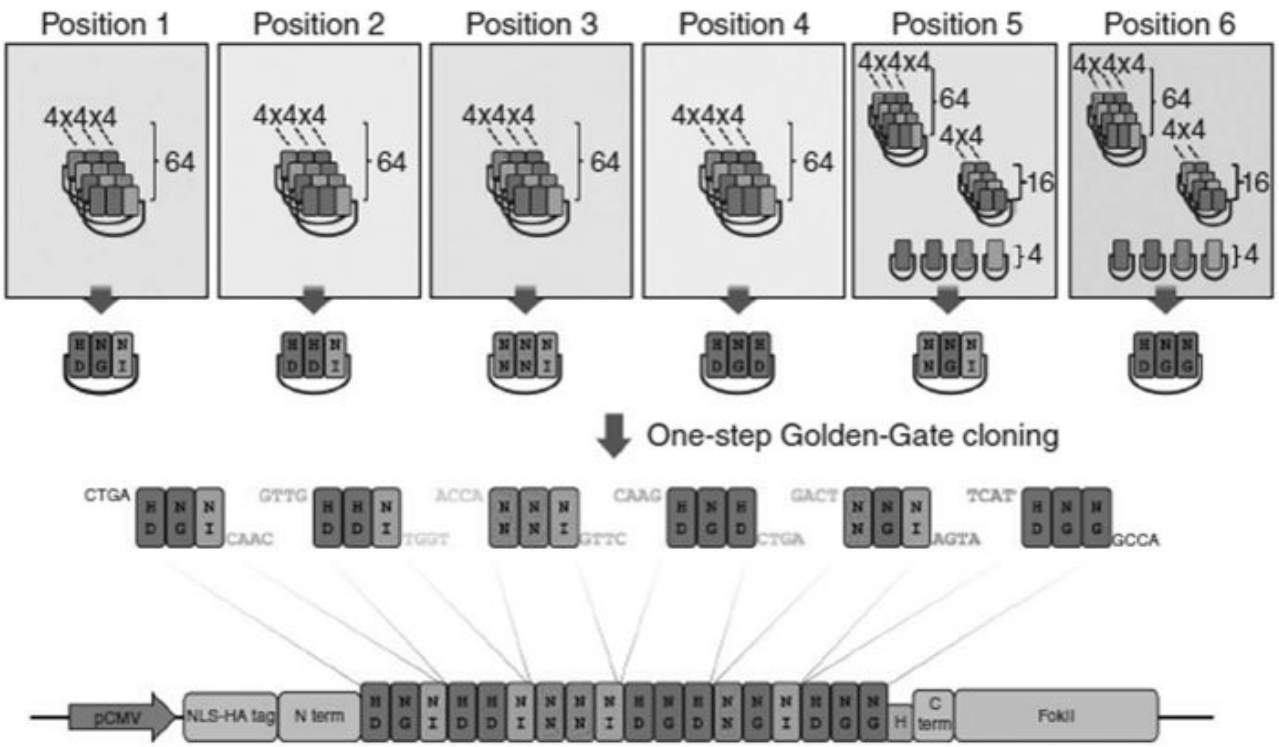

Position 1
Position 2
Position 3
Position 4
Position 5
Position 6
4x4x4
64
4x4
16
4
One-step Golden-Gate cloning
CTGA
CAAC
GTTG
TGGT
ACCA
GTTC
CAAG
CTGA
GACT
AGTA
TCAT
GCCA
pCMV
NLS-HA tag
N term
H
C term
FokI

Figure 4. Scheme of one-step Golden-Gate cloning system. A total of 424 TAL effector array plasmids (64 × 6 + 16 × 2 + 4 × 2) (Kan^R) and 8 FokI plasmids (Amp^R) are used. One member in each of the six positions is chosen; the six arrays are combined with each other for subcloning into one of the FokI expression plasmids. This system allows construction of TALEN plasmids that contain at least 14.5 (4 tripartite arrays + 2 monopartite arrays) and up to 18.5 (6 tripartite arrays) RVD modules in a single Golden-Gate reaction.

The TAL effector array plasmids are separated into six subgroups in accordance with their positions (Figure 4). When digested with BsaI restriction enzyme, each array in a given position generates the same four-base pair overhang sequences. TAL effector arrays in different subgroups produce different four-base overhang sequences. One member in each of the six subgroups is chosen; the six arrays are combined with each other for subcloning into one of the FokI expression plasmids. This system allows construction of TALEN plasmids that contain at least 14.5 (4 tripartite arrays + 2 monopartite arrays) and up to 18.5 (6 tripartite arrays) RVDs in a single Golden-Gate reaction. The last half-repeat (0.5) is encoded in the FokI plasmids. These TALENs recognize DNA sequences of 16–20 bps in length, including a conserved base T at the 5′ end. Because TALENs function as dimers, these TALEN pairs recognize 32- to 40-bp DNA sequences, which consist of two half-sites separated by 12- to 14-bp spacers because of L4 fusion junction. This one-step Golden-Gate cloning system enables us to TALEN synthesis high-throughput manner (Figure 5).

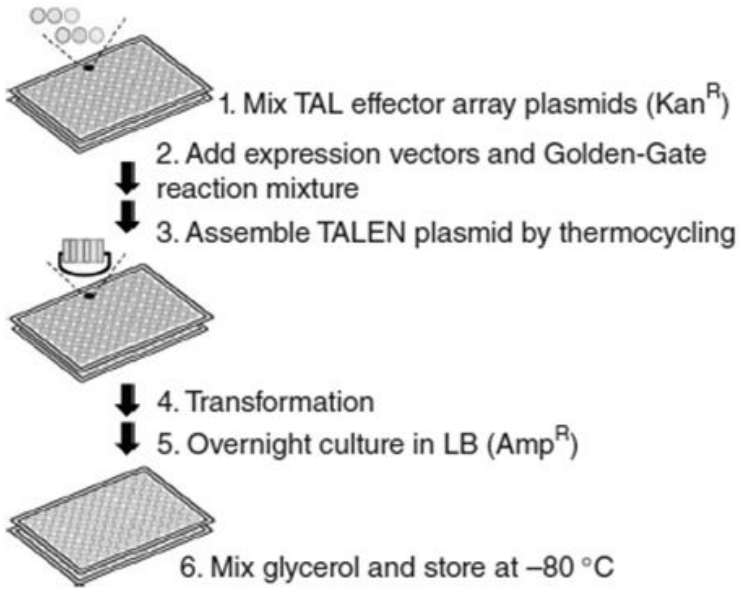
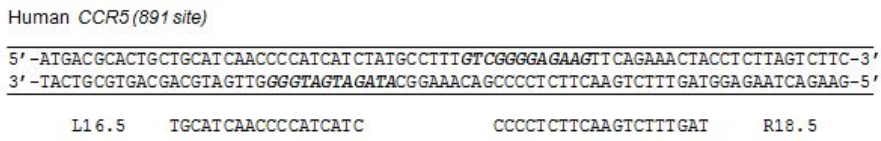


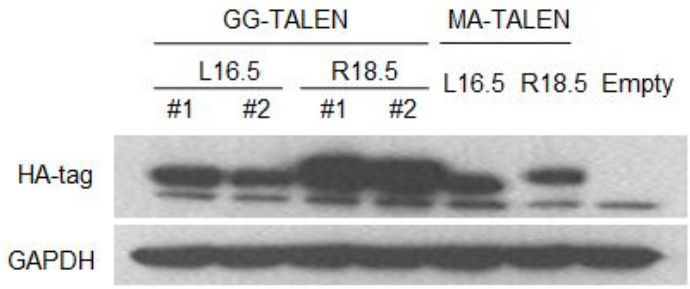
Figure 5. High-throughput synthesis of TALENs. High-throughput Golden-Gate cloning could process in 96-well plates. Six TAL effector array plasmids and one FokI plasmid are mixed in each well. BsaI releases the TAL effector arrays and allows an ordered assembly of six TAL effector arrays into the FokI plasmid.

I investigated whether TALEN pairs made by one-step Golden-Gate system are functional as TALEN pairs generated by modular assembly method. I generated TALEN pair targeting *CCR5* gene as previous study (Kim H.J. 2011) (Figure 6a), and confirmed protein expression levels and mutagenesis activity. TALEN pair that synthesized by Golden-Gate cloning method (GG-TALEN) was highly expressed and induced similarly mutagenesis with TALENs synthesized by modular assembly method (MA-TALEN) (Figure 6b, c). And then, I constructed 15 TALEN pairs targeting each human genes (Table 1). Because longer TALENs are more active (Reyon et al. 2012), I designed these TALEN pairs with 18.5 RVD modules. I measured the mutagenesis activities of these TALEN pairs in HEK293 cells using T7E1 assay. HEK293 cells were transfected with TALEN-encoding plasmids, and PCR amplicons from genomic DNA were analyzed by T7E1. Mutation frequencies were detected from the intensities of cleaved bands relative to intact bands. Mutations were estimated at all 15 sites at frequencies of 3.9-43% (Figure 7). These pilot tests demonstrate that both the new TALEN construct and the one-step Golden-Gate cloning system are robust enough to allow genome-scale construction of TALENs.

a



b



c

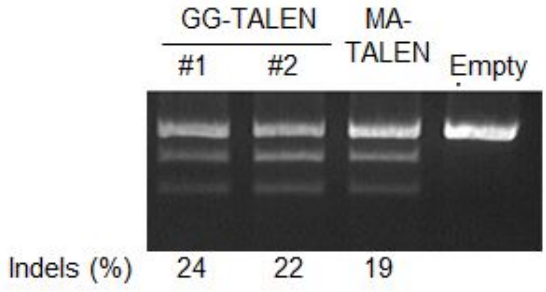


Figure 6. Validation of TALENs constructed by Golden-Gate assembly. (a) TALENs targeted CCR5 in previous study (Kim H.J. 2011) were synthesized. (b) Protein expression levels were detected by Western blotting. Protein expression of TALEN pairs synthesized by Golden-Gate cloning method (GG-TALEN) were comparable with TALEN pairs constructed by modular assembly (MA-TALEN). Each numbers are

independently synthesized TALENs. (c) T7E1 assay were detected endogenous mutagenesis activities of TALEN pairs. GG-TALENs could induce mutations in endogenous target sites with high frequency. The numbers at the bottom of the gels indicate mutation frequencies measured by band intensities

Table 1. List of 15 TALEN target sequences in pilot test

	Gene name	TALEN binding site	Number of CpG sites			Mutation frequency (%)
			Left site	Right site	Total	
Pilot study	*ACAT1*	TGCCTTGAAGCAAGGAGAATacggtcttgccaGTATTTGCAATGGAGGAGGA	0	0	0	9.5
	ANGPT1	TGGGATAAAGTGGCACTACTtcaaagggcccaGTTACTCCTTACGTTCCACA	0	1	1	24
	AIRE	TCTACAAGCACCTGCCGGCTccgccttctgcaGCCCCGCTGCCAGGGCTGGA	1	1	2	43
	CYP11A1	TGGTGCAAGTGGCCATCTATgctctgggccgaGAGCCCACCTTCTTCTTCGA	0	1	1	11
	F8	TTACTGCTTCATCCTACTTTaccaatatgtttGCCACCTGGTCTCCTTCAAA	0	0	0	26
	FOXF1	TCAACGCCATGGCGTCCTCTtccatgcactcgGCCGGCGGGGGCTCCTACTA	2	2	4	18
	GFAP	TGGAGGAAGAGGGGCAGAGCctcaaggacgagATGGCCCGCCACTTGCAGGA	0	1	1	11
	AGT	TCGCCCAGGCTGAGCTGCCCgccattctgcacACCGAGCTGAACCTGCAAAA	2	1	3	4.2
	KRT23	TGTTATCCGAGACCCAGTCTcggtactcctgcAAGCTCCAGGACATGCAAGA	1	0	1	3.9
	BBS9	TATTTTGAAAAACAGGGAGTcaaagattttgcATGTTCTTTTTCGGGATCTA	0	1	1	13
	TRIM45	TCACCCTGCTTTGTAAGGATgccgcaggagaaATCATGGGCAGGGGAGGAGA	0	0	0	15
	FCRL3	TCCAGGAACAGAACAGGCCTtaccgctgcgggAATCACGGGGCTGGTGCTCA	0	1	1	32
	BEND2	TGTCGGCTAGATACCTTATTcagaaactcttcACAAAAGATGTCCTGGTCCA	1	0	1	21
	EME1	TACACACAGAAGCCCAGGCTcaaattgtgcagAGCTGGAAAGAGCTGGCCGA	0	1	1	12
	MOSPD2	TACGTTAAAAGACAATGCTTtcaatatgtcagATAAAACCAGTGAAGATATA	1	0	1	7.9

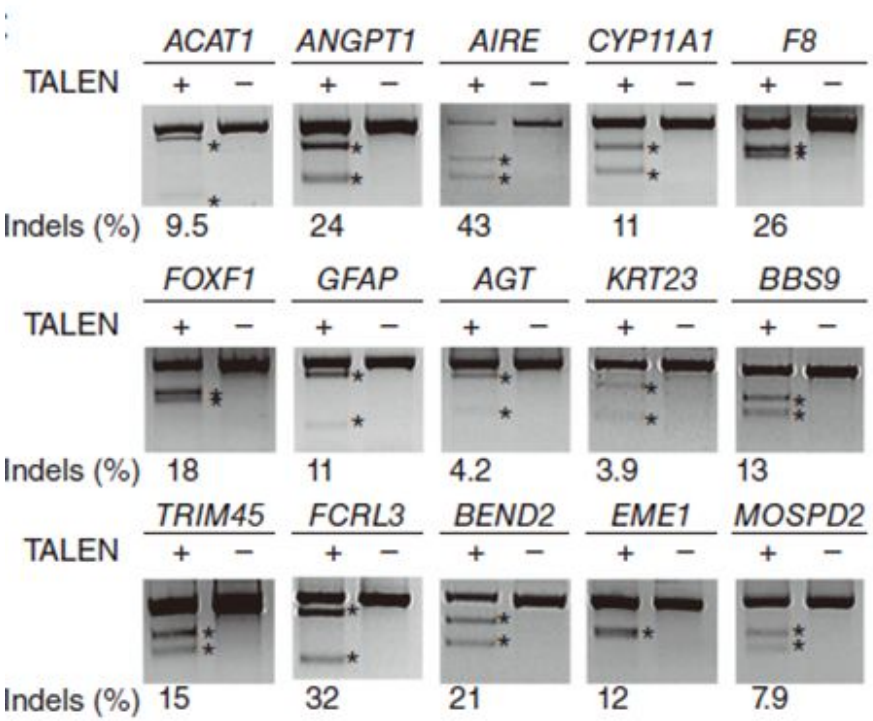


Figure 7. Pilot test of 15 TALENs. TALEN-mediated mutation frequencies were detected using the T7E1 assay. Asterisks indicate the expected positions of DNA bands cleaved by T7E1. The numbers at the bottom of the gels indicate mutation frequencies measured by band intensities.

2. Design of human genome-wide TALENs

To construct a human TALEN library - a genome-scale collection of human gene-targeting TALENs - I obtained the sequences of 18,742 protein-coding genes from HUGO Gene Nomenclature Committee(Seal et al. 2011) (www.genenames.org) on March 2011 and Refseq mRNA database from the National Center for Biotechnology Information (Pruitt et al. 2012) (www.ncbi.nlm.nih.gov) on November 2011. From the coding sequences (CDS) in Refseq mRNA database, we searched for candidate TALEN target sites that consist of two half-sites of 20-bp in length (18.5RVDs, total 40-bp) separated by 12- or 13-bp spacers and that start with the base T and end with A. To identify appropriate TALEN target sites for functionally gene disruption, I sequentially filtered these sites in the following steps:

1) I removed TALEN target sites that correspond to exon-exon junctions.
2) I selected target sites that reside in common exons in cases in which genes are expressed in two or more splicing variants
3) I scored the relative positions of target sites in CDS from the longest mRNA variant and excluded sites that resided within the downstream 30% of each coding sequence.
4) I searched for unique target sequences with the minimum number of potential off-target sites in the following manner:
    i) I screened for potential off-target sites in the human genome sequence (hg19) by allowing 3-bp mismatches to

the left and right half-target sites using Bowtie, an ultrafast memory-efficient short read aligner (Langmead et al. 2009).

ii) All sequences were aligned at chromosome loci positions for calculating distance between two half-target sites

iii) I excluded candidate sites that are associated with potential off-target sites that carry 6 or fewer mismatches. I defined potential off-target sites as any homodimeric or heterodimeric half-sites separated by 12- to 14-bp spacers. As a result, the most homologous potential off-target sites carried at least 7-base mismatches with the target site of choice.

These criteria were stringent enough to avoid poor sites that were not appropriate for site-specific gene knockout but were flexible enough to identify multiple sites in most genes (Table 2).

Table 2. Pilot test of computational strategy for TALENs design

Gene ID	Symbol	Initial found sites	off-target filtering		Position in CDS (<70%) filtering		Common exon filtering	
			numbers	percents	numbers	percents	numbers	percents
672	*BRCA1*	944	777	82.3	628	66.5	56	5.9
1029	*CDKN2A*	38	27	71.1	20	52.6	0	0.0
1234	*CCR5*	140	83	59.3	54	38.6	54	38.6
3043	*HBB*	39	7	17.9	4	10.3	4	10.3
3651	*PDX1*	34	25	73.5	23	67.6	23	67.6
4791	*NFKB2*	195	91	46.7	82	42.1	82	42.1
5728	*PTEN*	222	59	26.6	56	25.2	56	25.2
5888	*RAD51*	167	84	50.3	59	35.3	40	24.0
6657	*SOX2*	56	25	44.6	16	28.6	16	28.6
7157	*TP53*	121	77	63.6	59	48.8	32	26.4
10661	*KLF1*	48	40	83.3	32	66.7	32	66.7
29102	*DROSHA*	548	353	64.4	245	44.7	228	41.6
Total		2552	1648	64.6	1278	50.1	623	24.4

Genome-wide search identified at least one TALEN site that satisfied all of the criteria described in the text in 17,120 out of 18,742 genes (91%) (Group A in Table 3). To identify additional target sites, I loosened the criteria as follows: 1) I considered TALEN pairs that consisted of at least one monomer with 14.5 to 17.5 RVDs (16- to 19-bp, total 32- to 39-bp) and found 2,842 such sites (Group B) in 1,361 genes including 162 additional genes. 2) I searched for target sites whose potential off-target sites elsewhere in the genome carry at least 2-bp mismatches at each of the two half-sites: I identified 706 sites (Group C) in 393 genes including 338 additional genes. 3) In cases in which I failed to find sites that resided within the upstream 70% of a coding sequence, I searched for TALEN sites that resided within the upstream 90%: A total of 270 sites in 154 genes were identified (Group D). 4) 60 Genes encode splicing variants that do not share a common protein-coding exon. For these genes, I searched for sites in exons shared by at least two variants (Group E). 5) Finally, I identified 4,582 sites in the remaining 922 genes after I further loosened the off-target criterion (Group F). I designed to choose all of these additional sites (Groups B to F) preferentially recognized by 18.5 RVD TALENs with minimum off-target effects. As a result, I identified a total of 169,362 TALEN target sites in 18,740 protein-coding genes. I couldn't identify target sites in the rest 2 genes because potential TALEN target sites were not exist in their coding sequences. The TALEN target sites were designed 9.0 target sites per gene on average. The vast majority (98%) of these sites are targeted by 18.5/18.5 RVD

TALENs. Of these sites 95% (160,712/169,362) do not have any homologous sites with >85% sequence identity (that is, 6-base mismatches/ ≤ 40-bp sequence) in the human genome (Table 4).

Table 3. Summary of human genome-wide TALEN Library

Group	Target-site length (bp)[a]	TALEN length (RVD)	Position in coding sequences	Number of genes	Number of genes (running total)	Number of target sites	Number of target sites (running total)	Minimum number of mismatches at potential off-target sites
A	40	18.5	upstream 70%	17,120	17,120	160,712	160,712	>6 (=3 + 3)[b]
B	32~39	14.5~18.5	upstream 70%	1,361	17,282	2,842	163,554	>6 (=3 + 3)[b]
C	32~40	14.5~18.5	upstream 70%	393	17,620	706	164,260	>4 (=2 + 2)[c]
D	32~40	14.5~18.5	upstream 90%	154	17,758	270	164,530	>6 (=3 + 3)[b] → >4 (=2 + 2)[c]
E	32~40	14.5~18.5	upstream 70%	60	17,818	250	165,780	>6 (=3 + 3)[b] → >4 (=2 + 2)[c]
F	32~40	14.5~18.5	upstream 70%	922	18,740	4,582	169,362	≤4 (=2 + 2)[b]

[a] TALEN target sequence excluding 12- or 13-bp spacers

[b] At least 3-base mismatches at each half-site

[c] At least 2-base mismatches at each half-site

Table 4. Analysis of TALEN target sites in the human genome.

Number of target sites	Numer of genes	Position in CDS	Total sites	Number of CpG sites	Total sites
1	41	upstream 10%	41,276	0	60,266
2	685	upstream 10% to 20%	33,323	1	48,353
3	245	upstream 20% to 30%	26,808	2	27,525
4	202	upstream 30% to 40%	21,877	3	15,102
5	1,112	upstream 40% to 50%	18,699	4	8,445
6	258	upstream 50% to 60%	14,610	5	4,785
7	329	upstream 60% to 70%	12,529	6	2,692
8	491	upstream 70% to 80%	104	7	1,306
9	701	upstream 80% to 90%	166	8	616
10	14,676	upstream 90% to 100%	0	>8	272
Total genes	18,740	Total	169,392	Total	169,362
Total target sites	169,362				

3. TALEN-mediated genome editing activity

To validate success rate of TAL effector nucleases, I randomly chose 478 TALEN pairs that targeted different genes to validate their mutagenesis activities in HEK293 cells using dual fluorescent reporter assay. As previous study, I measured GFP positive ratio in each TALENs by FACS (Kim et al. 2011). Although all of cells transfected with reporter plasmid were only <1% GFP positive populations, 472 of 478 cells (98.7%) transfected with reporter and TALEN-encoding plasmids were active (detection limit, ~1%) (Figure 8 and Appendix 1).

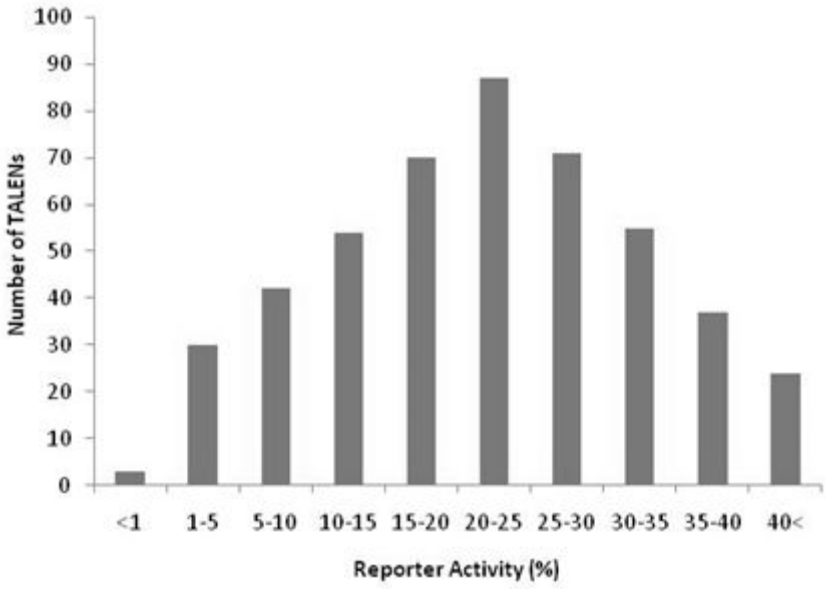
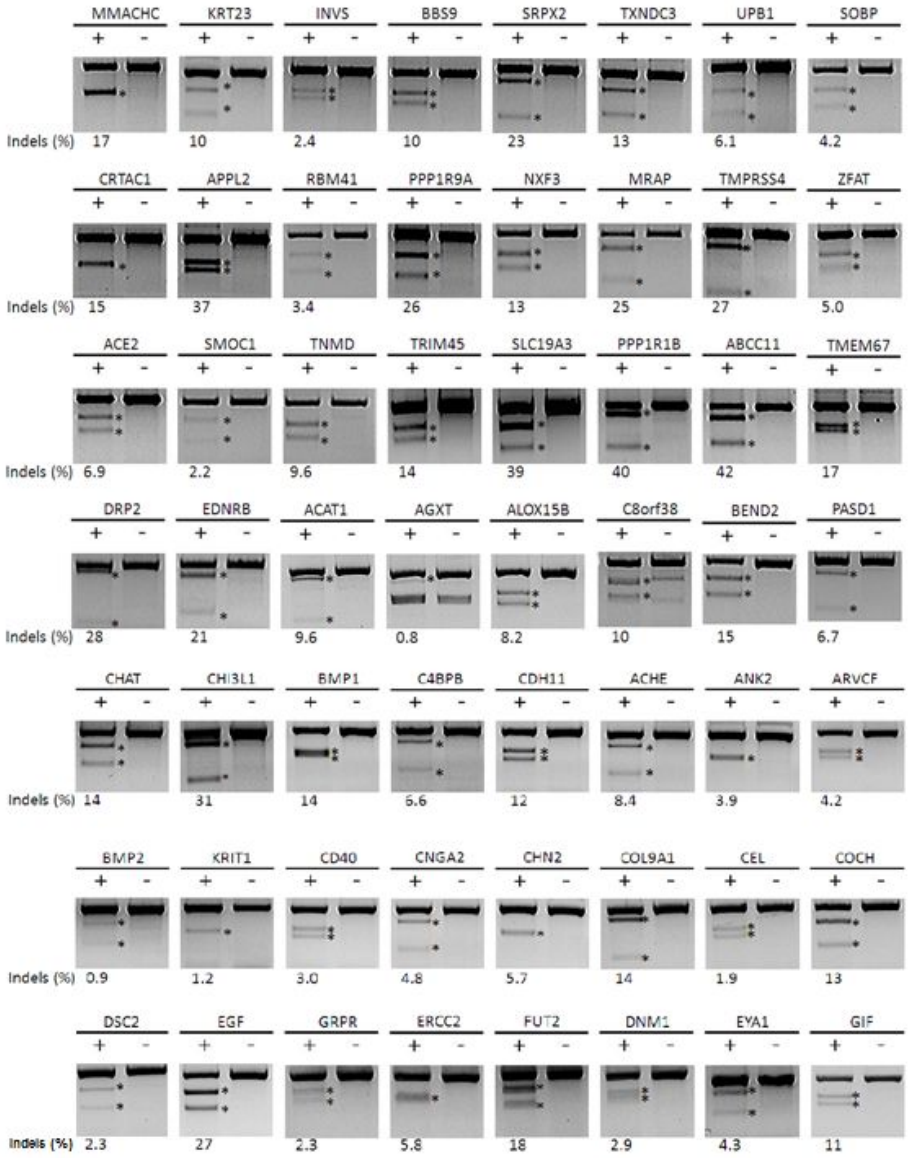


Figure 8. Reporter-based assay for detecting TALEN activities. HEK293 cells were co-transfected with the reporter plasmid including appropriate target site and TALEN-encoding plasmis. Two days after transfection, GFP+ cells were counted by FACS. Almost TALENs were active in this reporter assay.

I investigated whether these active TALENs could, indeed, induce mutagenesis in endogenous target sites highly successful as reporter assay in HEK293 cells. I selected 104 TALEN pairs in 479 TALENs and validated their genome editing activities in HEK293 cells using T7E1 assay (Figure 9 and Appendix 2). Mutations were detected at 101 out of 103 target sites that were successfully PCR-amplified (assay sensitivity, ~0.5%). Accordingly, the success rate of our TALENs was 98.1%. These TALENs were highly acitive: 78/103 (76%) of TALENs were associated with mutation frequencies of >5%, and 57/103 (55%) of TALENs showed frequencies of >10% (Figure 10). The average mutation frequency was 16%. The two best-performing TALENs, which were specific to the *FKTN* and *OPN1SW* genes, induced mutations at frequencies of 54% and 46%, respectively. Mutations induced by ten different TALENs were confirmed by DNA sequencing (Figure 11). Indels, signatures of error-prone NHEJ repair of DSBs, were detected at the target sites.


MMACHC KRT23 INVS BBS9 SRPX2 TXNDC3 UPB1 SOBP
+ - + - + - + - + - + - + - + -
Indels (%) 17 10 2.4 10 23 13 6.1 4.2
CRTAC1 APPL2 RBM41 PPP1R9A NXF3 MRAP TMPRSS4 ZFAT
+ - + - + - + - + - + - + - + -
Indels (%) 15 37 3.4 26 13 25 27 5.0
ACE2 SMOC1 TNMD TRIM45 SLC19A3 PPP1R1B ABCC11 TMEM67
+ - + - + - + - + - + - + - + -
Indels (%) 6.9 2.2 9.6 14 39 40 42 17
DRP2 EDNRB ACAT1 AGXT ALOX15B C8orf38 BEND2 PASD1
+ - + - + - + - + - + - + - + -
Indels (%) 28 21 9.6 0.8 8.2 10 15 6.7
CHAT CHI3L1 BMP1 C4BPB CDH11 ACHE ANK2 ARVCF
+ - + - + - + - + - + - + - + -
Indels (%) 14 31 14 6.6 12 8.4 3.9 4.2
BMP2 KRIT1 CD40 CNGA2 CHN2 COL9A1 CEL COCH
+ - + - + - + - + - + - + - + -
Indels (%) 0.9 1.2 3.0 4.8 5.7 14 1.9 13
DSC2 EGF GRPR ERCC2 FUT2 DNM1 EYA1 GIF
+ - + - + - + - + - + - + - + -
Indels (%) 2.3 27 2.3 5.8 18 2.9 4.3 11

(continued)

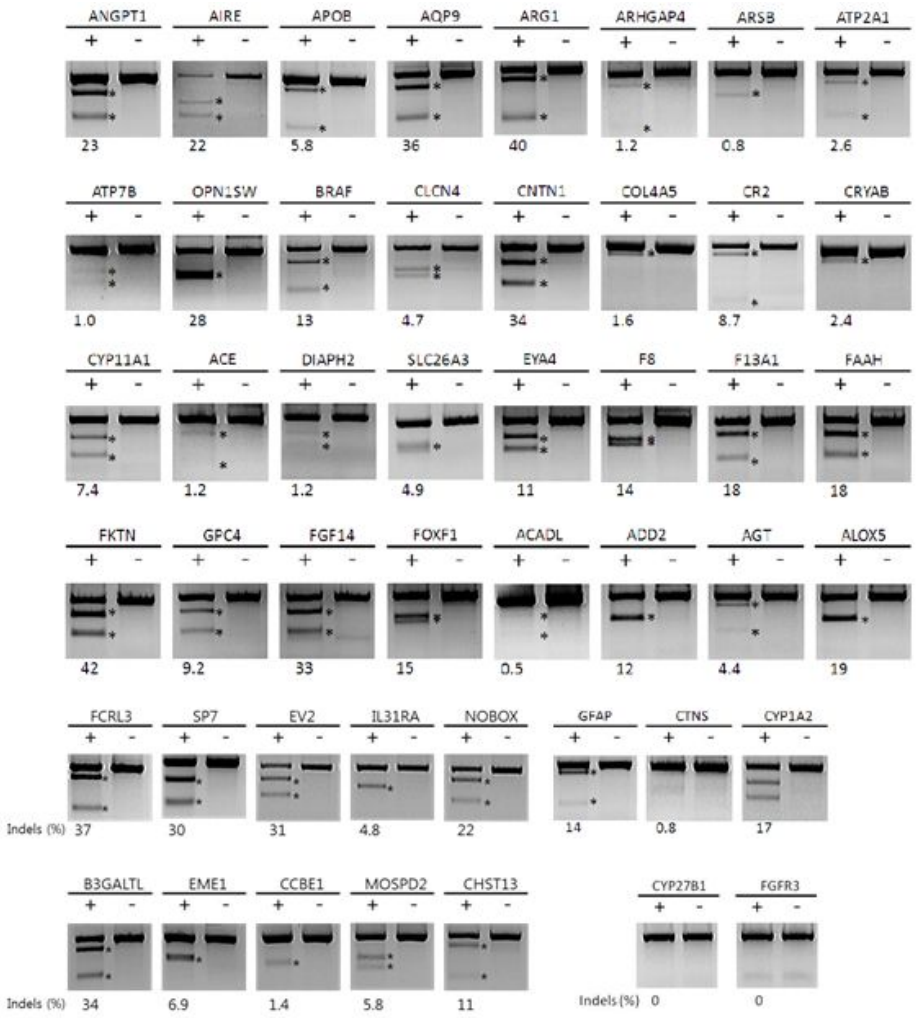


Figure 9. 103 of TALENs test using the T7E1 assay. The genome-editing activities of 103 TALENs were measured by the T7E1 assay. Asterisks indicate the expected positions of DNA bands cleaved by T7E1. The numbers at the bottom of the gels indicate mutation frequencies measured by the intensities of cleaved bands. + or – indicates the presence or absence of TALENs in transfected cells.

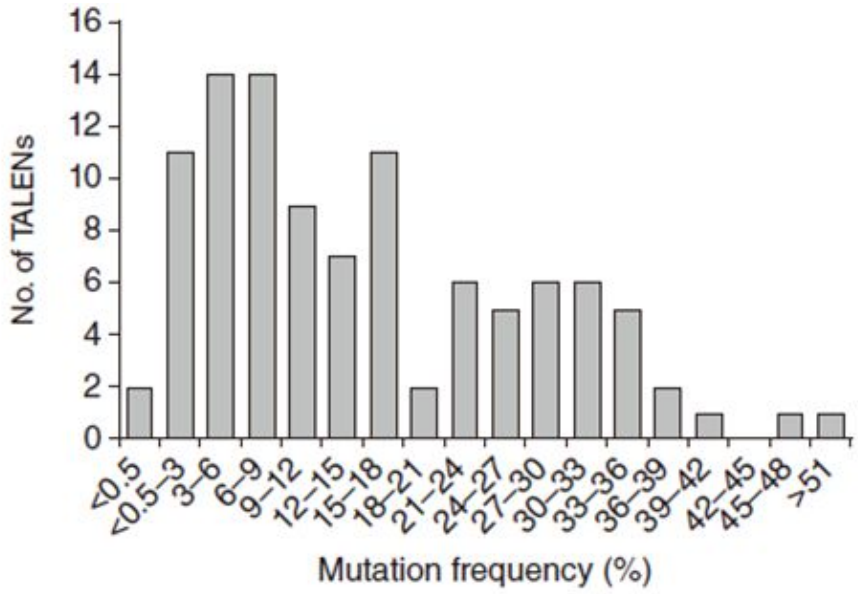


Figure 10. Distribution of TALEN activities. Detected mutaion frequencies by the T7E1 assay were analyzed. These TALENs were highly active: 76% (78/103) of TALENs were associated with mutation frequencies of >5% (or indel %), and 55% (57/103) of TALENs showed frequencies of >10%

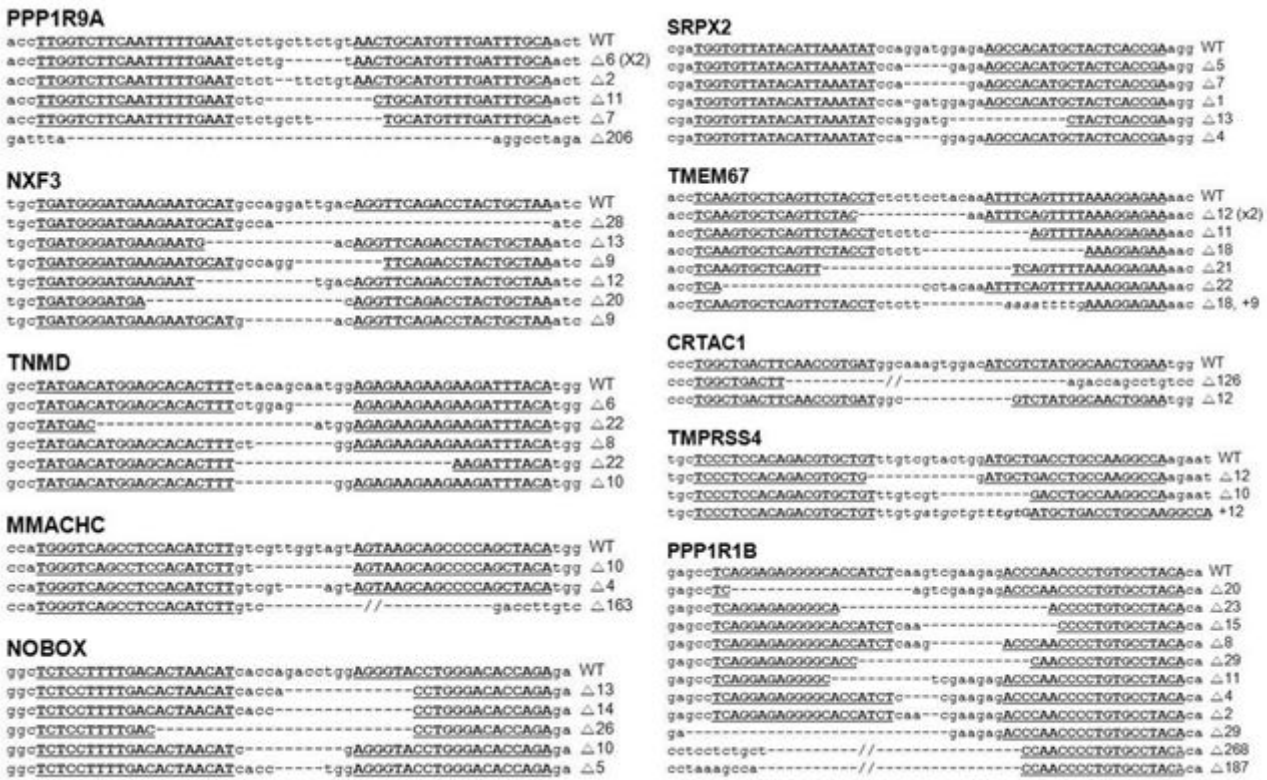
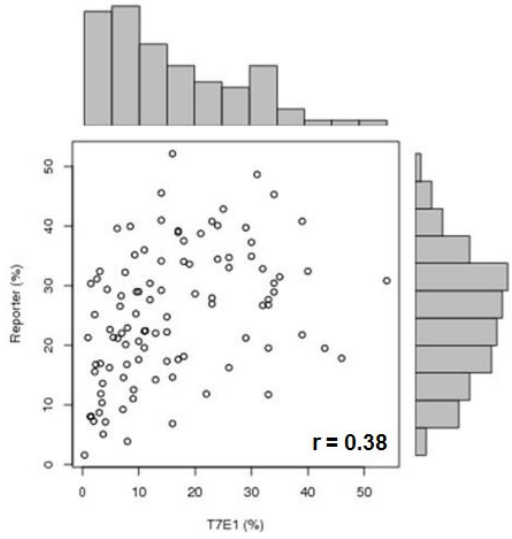

PPP1R9A
WT
△6 (X2)
△2
△11
△7
△206
NXF3
WT
△28
△13
△9
△12
△20
△9
TNMD
WT
△6
△22
△8
△22
△10
MMACHC
WT
△10
△4
△163
NOBOX
WT
△13
△14
△26
△10
△5
SRPX2
WT
△5
△7
△1
△13
△4
TMEM67
WT
△12 (x2)
△11
△18
△21
△22
△18, +9
CRTAC1
WT
△126
△12
TMPRSS4
WT
△12
△10
+12
PPP1R1B
WT
△20
△23
△15
△8
△29
△11
△4
△2
△29
△268
△187

Figure 11. DNA sequences of indels induced by TALENs. Endogenous mutations induced by 10 different TALENs were confirmed by dideoxy DNA sequencing. The numbers of inserted or deleted bases are shown on the right side of each mutant sequence. WT, wild-type sequence.

Furthermore, to know what is the main factors that influenced on genome editing frequencies of nucleases, I investigated correlation between nuclease mediated episomal reporter activity and endogenous mutation activity of 103 TALEN pairs. The estimated values of linear (Pearson's r) and non-linear (Spearman, ρ) correlation coefficients between endogenous activities and episomal reporter activities were 0.38 and 0.44, respectively (Figure 12a, b). This result showed that nuclease's genome editing activity is little correlate with episomal reporter based activity. In other words, some endogenous factors such as chromatin accessibility or DNA methylation influenced on TALEN's gene disruption frequency.

a



b

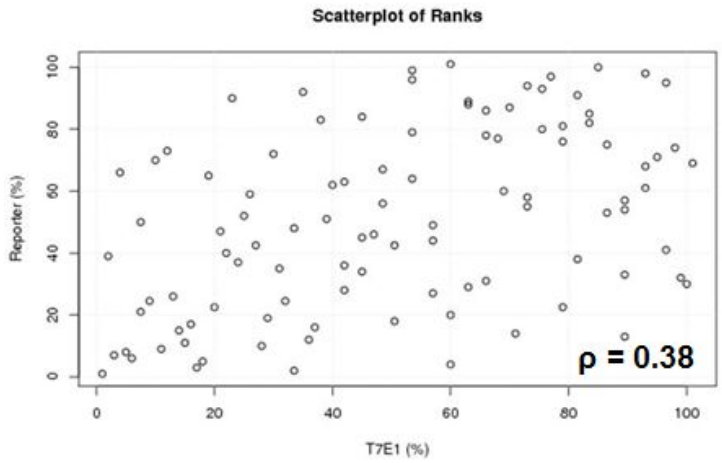


Figure 12. Comparison between episomal reporter assay and the T7E1 assay. The estimated values of linear (Pearson's r) and non-linear (Spearman, ρ) correlation coefficients between endogenous activities and episomal reporter activities were 0.38 and 0.44, respectively.

4. Rescue of inactive TALENs

I investigated why two of the 118 TALENs (including 15 TALENs used in the pilot study), which were designed to target *CYP27B1* and *FGFR3*, failed to show any genome-editing activity in the T7E1 assay in contrast with highly active in reporter assay (Figure 13 and Appendix 1). First, I sequenced the two target sites in HEK293 cells and found no poly-morphisms or mutations at these sites. Thus, the target sites do not contain sequence variations that prevent the TALENs from binding. To resolve this unexpected discrepancy in cleavage of episomal and chromosomal DNA, I hypothesized that these TALENs could not access the endogenous sites owing to the local chromatin structure or DNA methylation. To test this idea, I treated cells with either trichostatin A, an inhibitor of histone deacetylase (HDAC), or 5-aza-2-deoxycytidine (5-aza-dC), an inhibitor of DNA methyltransferase. The two TALENs both induced mutations when cells were pretreated with the inhibitor of DNA methylation but not with the HDAC inhibitor, as shown by the T7E1 assay and by DNA sequencing (Figure 14a, c). In addition, bisulfite DNA sequencing analysis revealed that all the six CpG dinucleotides in each of the two target sites were, indeed, methylated (Figure 14b). These results showed that TALENs cannot recognize heavily methylated DNA and that 5-aza-dC rescued the TALENs that did not cleave chromosomal DNA initially, consistent with recent biochemical and structural studies (Bultmann et al. 2012; Deng et al. 2012; Valton et al. 2012).

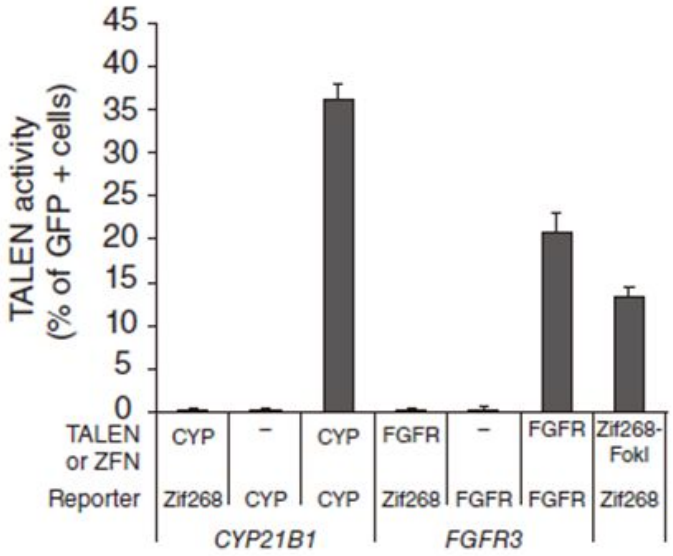
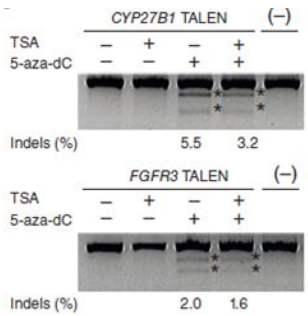
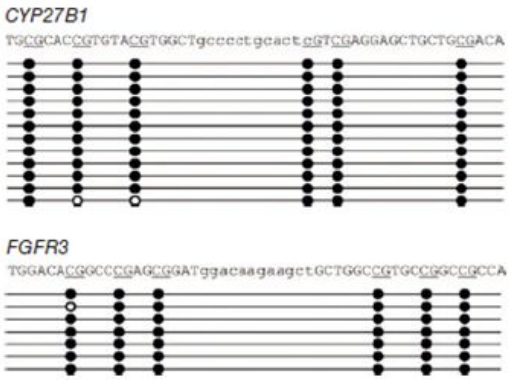


Figure 13. Episomal reporter assays of two genome-inactive TALENs. Dual fluorescent reporter assays of two TALENs that could not detected mutations by the T7E1 assay. Zif268-FokI (ZFN) was used as a positive control. Error bars, means ± s.e.m.

a



b



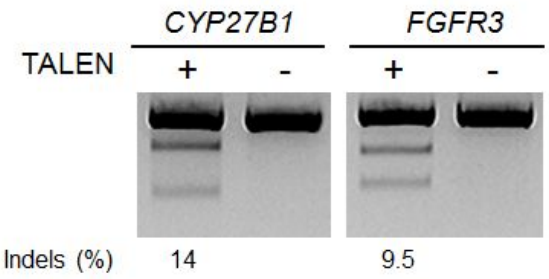
c

Figure 14. TALEN-driven mutations in drug-treated cells. (a) I treated cells with either trichostatin A, an inhibitor of histone deacetylase(HDAC), or 5-aza-2-deoxycytidine (5-aza-dC), an inhibitor of DNA methyltransferase. The two TALENs both induced mutations when cells were pretreated with the inhibitor of DNA methylation but not

with the HDAC inhibitor, as shown by the T7E1 assay. (b) Cytosine methylation at two initially unmodified sites revealed by bisulfite sequencing. The DNA sequences of two half-sites and spacers are shown in upper case and lower case, respectively. CpG dinucleotides are underlined. Closed and open circles indicate methylated and unmethylated cytosines, respectively. (c) Sequencing validation of TALEN induced mutations in 5-aza-dC treated cells.

To avoid treating cell lines with 5-aza-dC, a mutagen, First, I synthesized two new TALENs that HD RVD modules replaced with NG RVD modules in genome-inactive TALENs for binding methylated cytosine as previous study (Deng et al. 2012). The two methylated cytosine binding TALENs (mC-TALENs) were rescued mutagenesis activity (Figure 15a). Next, I tested whether the two genome-inactive TALENs could be replaced with other TALENs that target the same gene. I synthesized new TALENs specific to the *CYP27B1* and *FGFR3* genes and found that these TALENs were able to induce targeted genome modifications at the two new sites (Figure 15b).

a



b

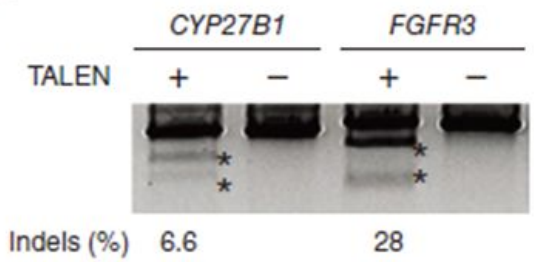


Figure 15. Targeted mutagenesis using alternative sets of TALENs. (a) HD RVDs modules in inactive TALENs were replaced with NG RVDs and transfected to HEK293 cells. The mutation bands were detected in the T7E1 assay. (b) Targeted mutagenesis using a second set of TALENs in TALEN library.

5. Undetectable off-target mutations

Both ZFNs and TALENs can induce mutations at sites other than their intended target sites (Gabriel et al. 2011; Hockemeyer et al. 2011; Mussolino et al. 2011; Pattanayak et al. 2011). As noted above, I chose all the TALEN sites carefully to minimize off-target effects. I investigated whether my designed TALENs could still induce unintended mutations at sites whose sequences are homologous to those of on-target sites. I chose ten TALENs that were highly active, showing mutation frequencies of 30% or greater (36% on average) at their on-target sites, searched for the most likely off-target sites based on the sequence homology in the genome, and tested the genome-editing activities of these TALENs at these potential off-target sites using the T7E1 assay (Table 5). None of these TALENs elicited any measurable mutations at their most likely off-target sites (assay sensitivity, ~0.5%), demonstrating exquisite specificities of these TALENs (Figure 16).

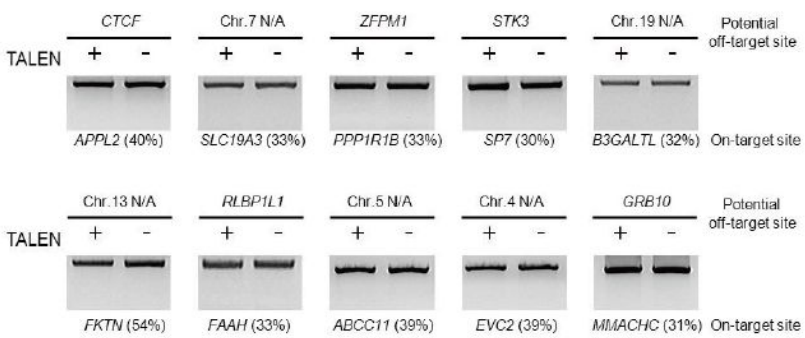
The specificities of RVDs are not absolute. Among the four RVDs, NN appears most promiscuous, as it cannot distinguish G from A efficiently (Boch et al. 2009; Moscou and Bogdanove 2009; Miller et al. 2011). To address this issue, I searched for additional sites of the ten TALENs that would be homologous to the on-target TALEN sites if the promiscuity of NN is considered. Nine TALEN sites were not associated with any potential off-target sites that carry six or fewer mismatches and thus still belong to group A. Only one TALEN site, which contained an unusually high number (17) of guanines in the

40-bp sequence, was associated with additional off-target sites; in this case, there were three potential off-target sites, each of which carries 6-base mismatches (Table 5). The T7E1 assay showed that this TALEN did not induce off-target mutations at these sites (Fugure 16). Among the 17,120 group A sites, the vast majority (86%) still satisfy the group A criterion when I take the promiscuity of NN into account. As expected, the other 14% of the sites contain an unusually high number of guanines in their target sequences. These results suggest that we would still choose most of the same sites even when I consider the promiscuity of NN. Because I designed up to ten sites in each gene, potential users who concerned about NN promiscuity may avoid sites that contain too many guanines.

Table 5. Potential off-target sites of highly active TALENs in the human genome.

	Chromosome number	Gene name	Left half-site (5' to 3')	No. of mismatches	Right half-site (5' to 3')	No. of mismatches	Spacer (bp)
ON		*APPL2*	TTGGCAGACACAATGGTTCT		TGAGATCCTTTTCTCGGAAT		12
OFF	16	*CTCF*	TgGGgAcACACAgTGGTcCa	6	TGAGcTtCTTTTCagtGAAT	5	14
ON		*SLC19A3*	TATGTCCGCTACAAGCCAGT		TAATGATGAAACTGATACCT		12
OFF	7	N/A	TtTtTCCcCaACAAGCCAGT	4	TAtTtATaAAACTaATAgCa	6	13
ON		*PPP1R1B*	TCAGGAGAGGGGCACCATCT		TGTAGGCACAGGGGTTGGGT		12
OFF	16	*ZFPM1*	cCAGGgacGtGGCACCATCT	5	gGTgGGCAaAGGGGTTGGGa	4	13
ON		*SP7*	TTTGGTGGCTCTAGCCCTCT		TGCCTGCTTTGCCCAGAGTT		12
OFF	8	*STK3*	TTTGGaGtCcaaAGCCCTCT	5	TGaCTGCaTaGCaCAGAGaT	5	13
ON		*B3GALTL*	TCATCCAGAGTCAAAGTAAT		TTTAACTGCTCTGCTCTCTT		12
OFF	19	N/A	TCATCCAGtGTCAAAGggtT	4	TgTcACTGCTgTtCTCaCTT	5	13
ON		*FKTN*	TGGAGCTGGTTTGTCAAAAT		TGTGCTATCAAATCCAATTC		12
OFF	13	N/A	TGGtGtTGcTTTGTtAtAAT	5	TGTtCTATaAAAaCCAATTa	4	13
ON		*FAAH*	TCAGAGGCGCTGCTAGCCCT		TGTGTAACTTCTGCACCAGC		12
OFF	8	*RLBP1L1*	TCAaAGGaGCaGCTgGCCCg	5	TGTGTggCTcCcaCACCAGa	6	14
ON		*ABCC11*	TACTCTCCAAGATGGCCCCT		TGGAGCCTCAGGATTTCTCT		12
OFF	5	N/A	TgCTCTCCcAacTGGCCgCT	5	caGAaCCTCAGGAaTaCTCT	5	13
ON		*EVC2*	TGCTCTTGCTGGACAGCATT		TTTCCCACAGAGTCCCAAAT		12
OFF	4	N/A	TGCTCTTcCaGGAttaCtTT	6	TTTaCtACAGAGTCCCgAAT	3	12
ON		*MMACHC*	TGTAGCTGGGGCTGCTTACT		TGGGTCAGCCTCCACATCTT		12
OFF	7	*GRB10*	TGagGCTGGtGCTGCTTACa	4	TGccTCAGttTCCtCATCTg	6	13
ON		*PPP1R1B*	TCAGGAGAGGGGCACCATCT		TGTAGGCACAGGGGTTGGGT		12
OFF	6	N/A	aCAAGAAcAAAACtCCATCT	(3)	TGTAAGCACAAcAAcTGAcT	(3)	13
OFF	3	*MFN1*	TCAAGAAAAAGACACCAaag	(3)	TGTAcACACcAGGATTAGAc	(3)	13
OFF	2	*ERBB4*	TCAAAAGcAAAACtaCATCT	(3)	TGTAAGCACAAcAAgaAAcT	(3)	12

a



b

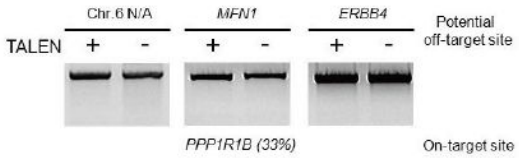


Figure 16. Undetectable off-target mutations with TALENs. (a) The genome was searched *in silico* for potential off-target sites (listed in Table 6) of 10 highly active TALENs, and the genome-editing activities of TALENs at these sites were tested by the T7E1 assay. Mutation frequencies at on-target sites are indicated at the bottom of the gels. (b) To account for the promiscuity of the NN RVD, additional sites were searched *in silico* that would be homologous to the on-target sites of the 10 highly active TALENs if all adenines were changed to guanines in the human genome. Only the *PPP1R1B*-specific TALEN was associated with any potential off-target sites (3 sites, listed in Table 6). No mutations were detected at these sites.

6. TALEN induced genome rearrangement

In many cases, homologous genes are often clustered in a chromosome. Many of these homologous genes are redundant in their functions. A single gene knockout often fails to cause any phenotypic changes owing to these redundancies. I propose that our TALENs could be used to induce large chromosomal deletions in a targeted manner to remove a cluster of homologous and functionally redundant genes. The TALEN library reported here consists of 18,740 TALEN pairs whose target sites are distributed all across the human genome. Thus, two neighboring TALEN target sites are separated by 170 kbp, on average, in a chromosome.

As a proof of principle experiment, I investigated whether and how efficiently two TALEN pairs can induce large chromosomal deletions in human cell lines. Two combinations of TALEN pairs were tested: (i) a *BRAF*-specific TALEN and a *NOBOX*-specific TALEN and (ii) a *EDNRB*-specific TALEN and a *FGF14*-specific TALEN (Figure 17a). The two target sites of the first pair are separated by 3.6 Mbp on chromosome 7, and those of the second pair are separated by 24 Mbp on chromosome 13. Targeted deletions of these large chromosomal segments were detected by PCR and confirmed by DNA sequencing (Figure 17b, d). The deletion frequencies measured by limiting dilution and PCR detection were 0.6% (*BRAF* and *NOBOX*) and 0.4% (*EDNRB* and *FGF14*) (Figure 17c), which are in line with those obtained with ZFNs (Lee et al. 2010). Recently, two TALEN pairs were used to

induce a 7-kbp chromosomal deletion within a single gene in porcine fibroblasts at a frequency of 10% (Carlson et al. 2012). Apparently, large deletions of Mbp by TALENs are at least tenfold less efficient than are those of kbp.

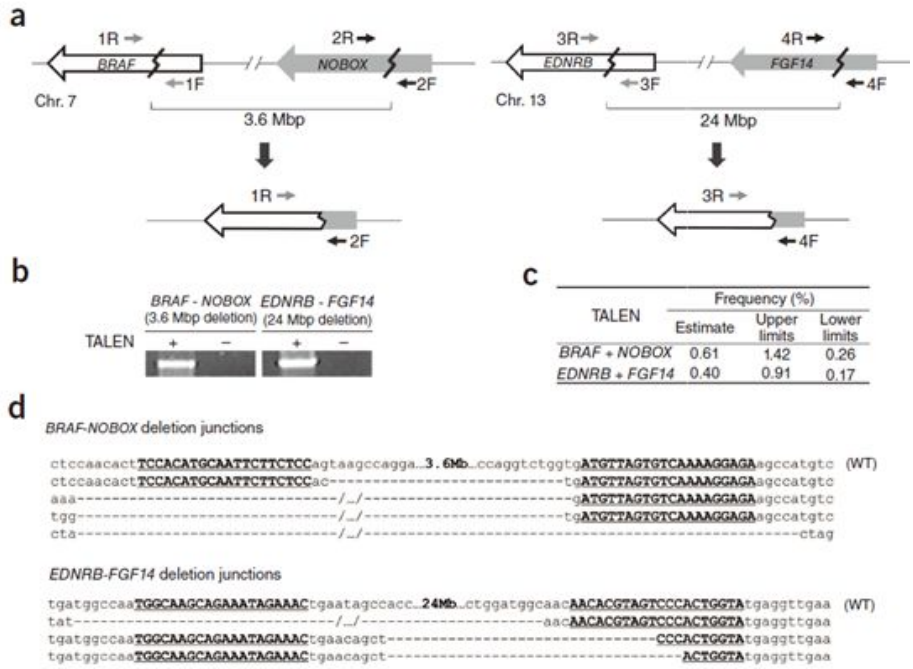


Figure 17. TALEN-mediated targeted genomic deletions. (a) Scheme of targeted chromosomal deletions induced by TALENs. Zigzag lines indicate TALEN target sites. Small arrows are PCR primers. (b) PCR products corresponding to large chromosomal deletions. (c) Deletion frequencies measured by dilution PCR. (d) DNA sequences of deletion junctions. TALEN recognition sequences are underlined and shown in boldface.

In addition, the large collection of TALENs reported here could also be used to induce other types of genome rearrangements, which include translocations, inversions, and duplications (Figure 18-20). Indeed, I was able to induce these structural variations in a targeted manner using various combinations of TALENs at frequencies ranging from 0.01% to 0.7. Structural variations in individual human genomes contribute to genetic diversity and often are associated with genetic diseases and cancer. Although thousands of structural variations have been identified, the biological consequences of these variations are largely unexplored. Previous studies have reported that the repair of two concurrent DSBs in the genome induced by ZFNs gives rise to these variations in a targeted manner in cultured human cells (Lee et al. 2012). To study their functions, one could also use TALENs to induce specific variations in cell lines. The high density of our TALENs all across the human genome could facilitate the creation of cell lines whose genomes are custom-designed.

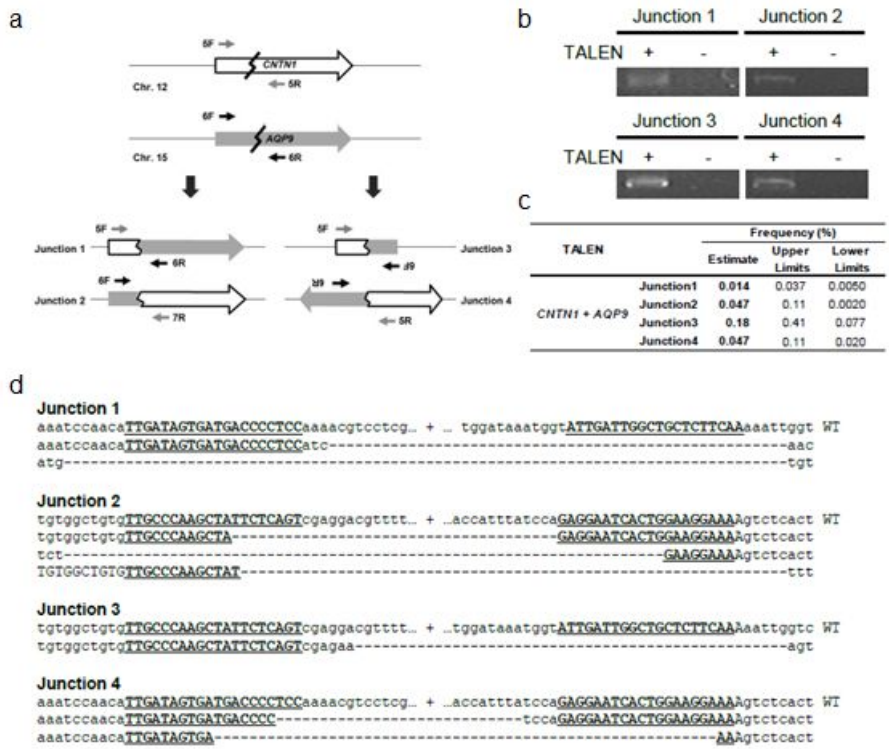
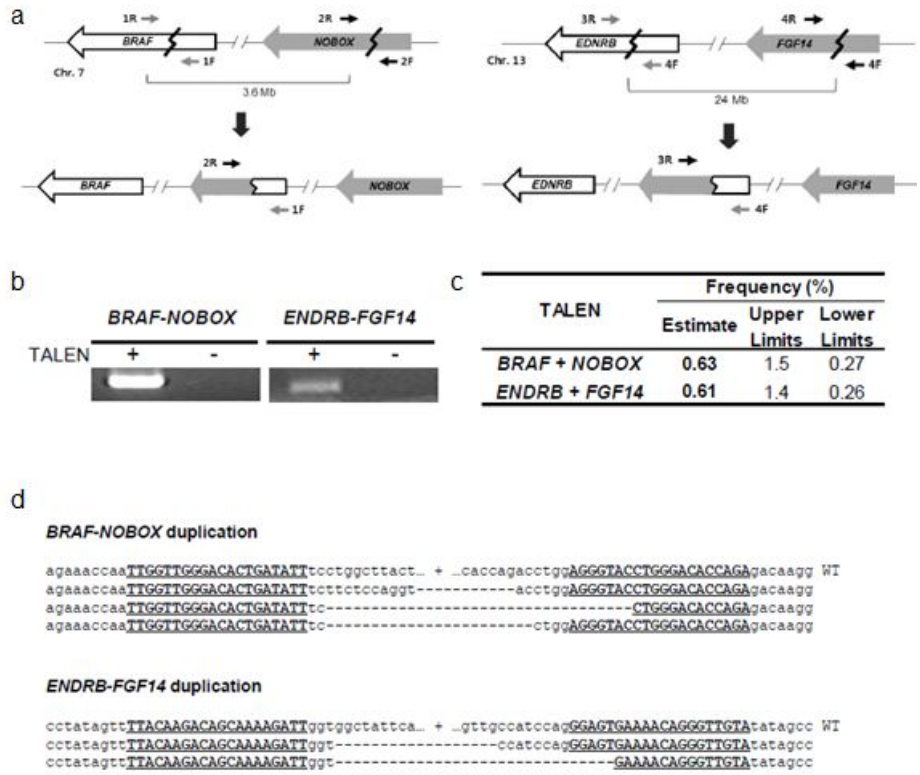


Figure 18. TALEN-mediated targeted genomic inversions. (a) Scheme of targeted chromosomal inversions induced by TALENs. Zigzag lines indicate TALEN target sites. Small arrows are PCR primers. (b) PCR products corresponding to chromosomal inversions. (c) Inversion frequencies measured by dilution PCR. (d) DNA sequences of inversion junctions. TALEN recognition sequences are underlined and shown in boldface.



TALEN	Frequency (%)		
	Estimate	Upper Limits	Lower Limits
BRAF + NOBOX	0.63	1.5	0.27
ENDRB + FGF14	0.61	1.4	0.26

Figure 19. TALEN-mediated targeted genomic duplications. (a) Scheme of targeted chromosomal duplications induced by TALENs. Zigzag lines indicate TALEN target sites. Small arrows are PCR primers. (b) PCR products corresponding to chromosomal duplications. (c) Duplication frequencies measured by dilution PCR. (d) DNA sequences of duplication junctions. TALEN recognition sequences are underlined and shown in boldface.

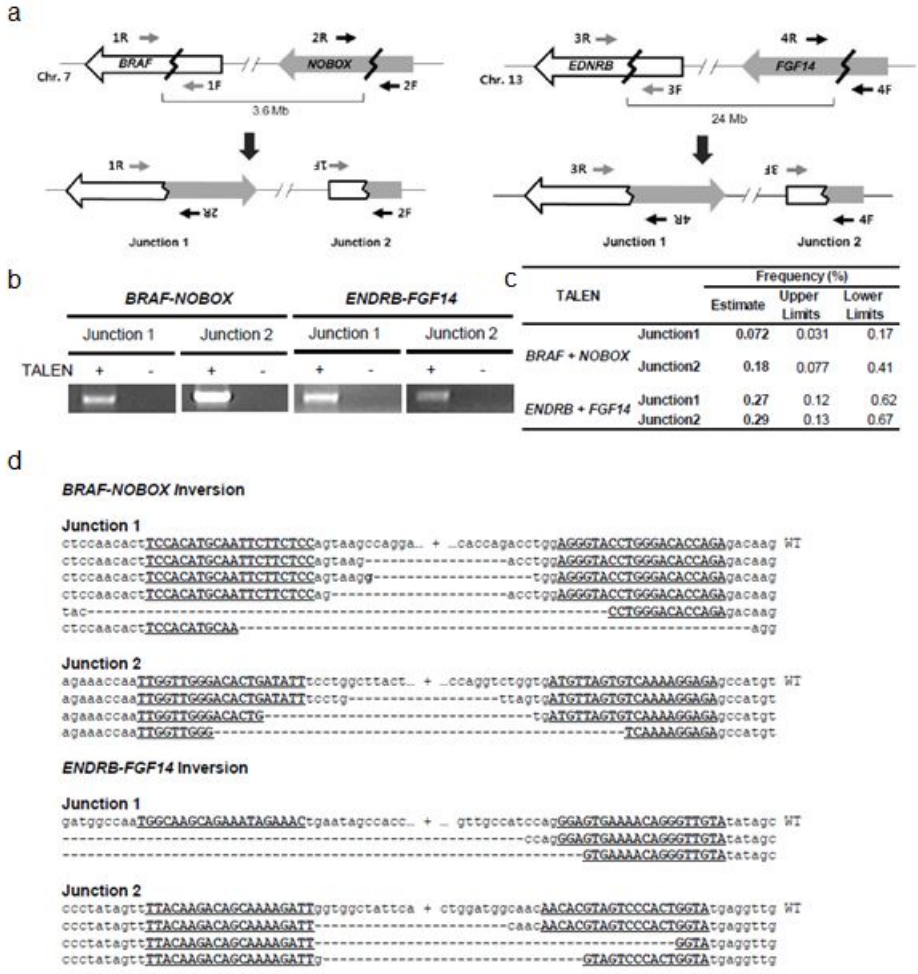


Figure 20. TALEN-mediated targeted genomic tranlocations. (a) Scheme of targeted chromosomal translocations induced by TALENs. Zigzag lines indicate TALEN target sites. Small arrows are PCR primers. (b) PCR products corresponding to chromosomal translocations. (c) Translocation frequencies measured by dilution PCR. (d) DNA sequences of translocation junctions. TALEN recognition sequences are underlined and shown in boldface.

C. TALEN-mediated knockout cell lines

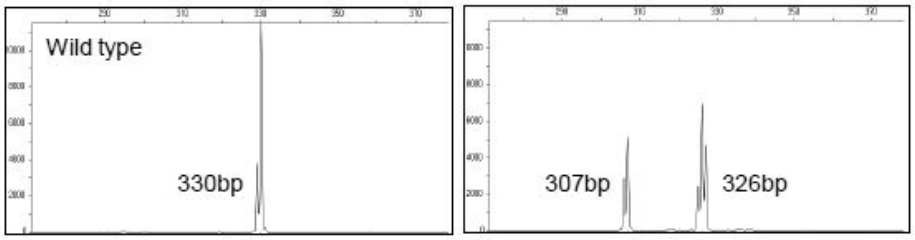
1. Establishment of knockout cell lines for NF-κB pathway study

To investigate functional gene knockout study in human cells, the cell lines, a perpetuating strain of cells in laboratory culture, have commonly used. But most model cell lines such as HEK293 cells and HeLa cells contain more than three copies of most chromosomes (Macville et al. 1999; Bylund et al. 2004). Thus, the disruption of one or two alleles does not yield gene-knockout cell lines. In addition, some engineered nucleases are cytotoxic (Cornu et al. 2008). Thus, cells that contain nuclease-induced mutations often die out, making it difficult to isolate rare clonal populations of gene-knockout cells (Kim et al. 2009).

As a proof-of-principle experiment for creating gene-knockout cells, I used TALENs to disrupt a series of genes associated with NF-κ B, a transcription factor linked to inflam-mation, septic shock, apoptosis, oncogenesis and DNA repair (Perkins 2007; Volcic et al. 2012). I transfected HEK293 cells or HeLa cells with both TALEN plasmids and surrogate reporters, which enabled enrichment of cells in which a gene of interest is completely disrupted (Appendix 2) (Kim et al. 2011; Kim et al. 2013a). After mutant clones were isolated using T7E1 assay, I inspected clonal populations of cells using fluorescent PCR for confirming disruption of all alleles that include gene of

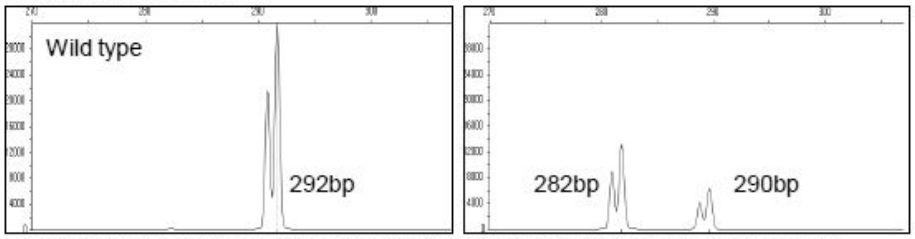
interests. Gene-knockout cells produced amplicon peaks corresponding to indels but not the wild-type peak. Note that all three alleles were mutated in a clonal population of cells. DNA sequencing confirmed the presence of frameshift mutations at target loci in these cells (Figure 21).

HeLa *TNFR1* KO clone



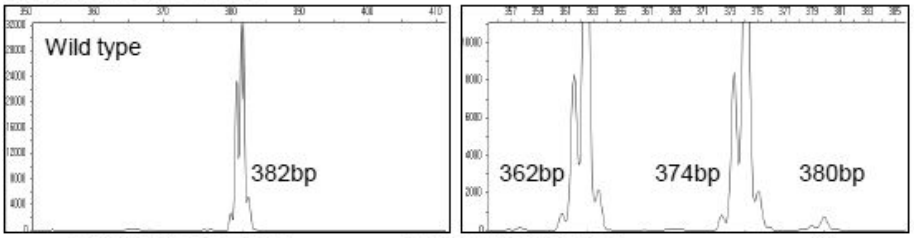
cttctccctg**TCCCCTAGGTGAGGGACCAG**tccaataaccccc**TGAGGGGTATATTCCCACCA**acagctccaggagcacct WT
cttctccctg**TCCCCTAGGTGAGGG**----------**GTATATTCCCACCA**acagctccaggagcacct △23
cttctccctg**TCCCCTAGGTGAGGGACCAG**----ataaccccc**TGAGGGGTATATTCCCACCA**acagctccaggagcacct △4

HeLa *IL1R1* KO clone



attgatg**TTCGTCCCTGTCCTCTTAAC**ccaaatgaacac**AAAGGCACTATAACTTGGTA**taaagatgacagcaaga WT
attgatg**TTCGTCCCTGTCCTCTTAAC**----------ac**AAAGGCACTATAACTTGGTA**taaagatgacagcaaga △10
attgatg**TTCGTCCCTGTCCTCTTAAC**ccaa--tgacac**AAAGGCACTATAACTTGGTA**taaagatgacagcaaga △4,+2

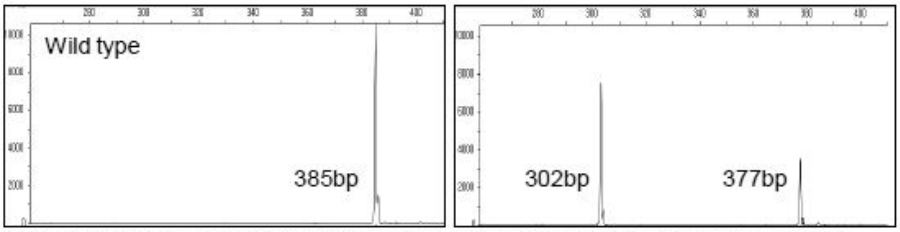
HeLa *IKKa* KO clone



acttga**TCTCAAAATAGCAATTAAGT**cttgtcgcctag**AGCTAAGTACCAAAAACAGA**gaacgatggtgccatga WT
acttga**TCTCAAAATAGCAATTAAGT**cttgt--cctag**AGCTAAGTACCAAAAACAGA**gaacgatggtgccatga △2
acttga**TCTCAAAATAGCAATTAAGT**c--------tag**AGCTAAGTACCAAAAACAGA**gaacgatggtgccatga △8
acttga**TCTCAAAATAGCAAT**----------**TAAGTACCAAAAACAGA**gaacgatggtgccatga △20

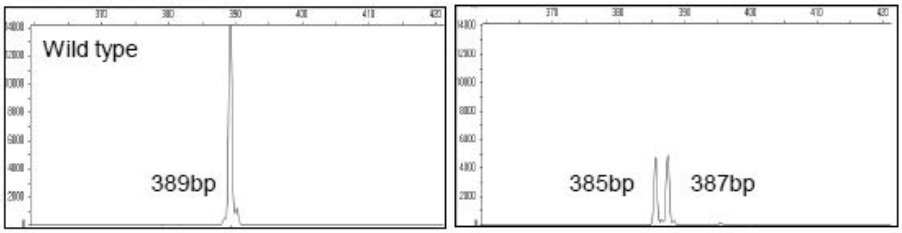
(Continued)

HeLa *TNFR2* KO clone



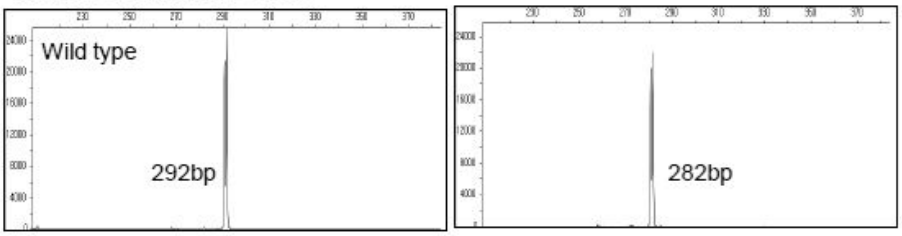
ttcccttcttccttccaggTGGCATTTACACCCTACGCCccggagcccgggAGCACATGCCGGCTCAGAGAatacta WT
aggggc----------//----------TCAGAGAatacta △83
ttcccttcttccttccaggTGGCATTTACACCCTACGCCccg--------gAGCACATGCCGGCTCAGAGAatacta △8

HeLa *p50* KO clone



atccttctttgacTCATACAATATTTAATCCAGaagtatttcaacCACAGATGGCACTGCCAACAggtaagaaaactcat WT
atccttctttgacTCATACAATATTTAATCCAGaagt--ttcaacCACAGATGGCACTGCCAACAggtaagaaaactcat △2
atccttctttgacTCATACAATATTTAATCCAGaagt----caacCACAGATGGCACTGCCAACAggtaagaaaactcat △4

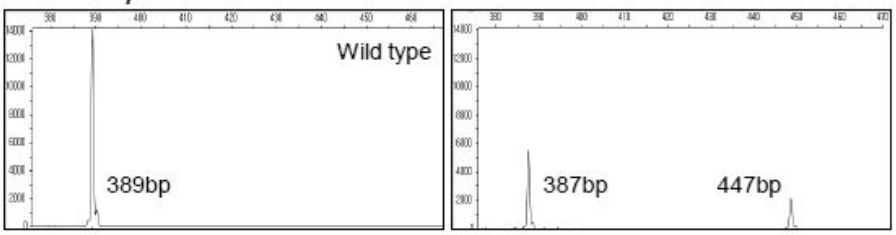
HEK 293 *IL1R1* KO clone



ttaTACCAAGTTATAGTGCCTTTgtgttcatttggGTTAAGAGGACAGGGACGAAcatcaatttcatttgcaga WT
ttaTACCAAGTTATAGTGCCTTTg----------gGTTAAGAGGACAGGGACGAAcatcaatttcatttgcaga △10

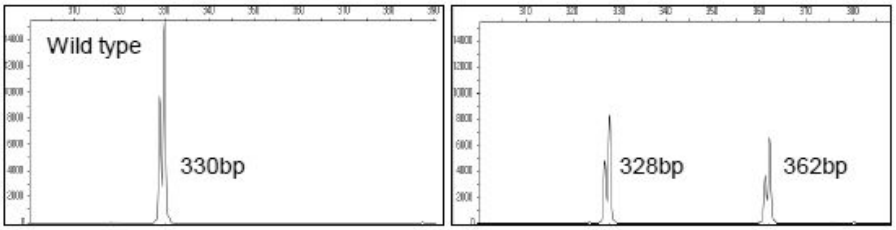
(Continued)

HEK 293 *p50* KO clone



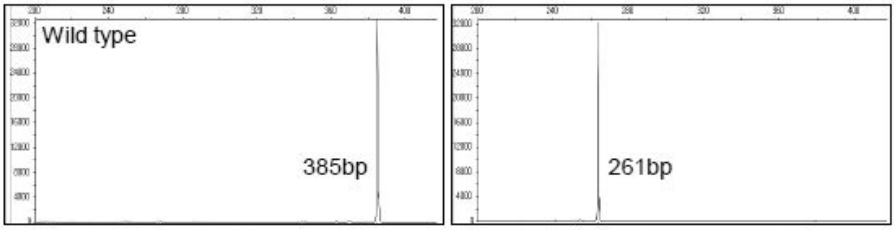
ac**TCATACAATATTTAATCCAG**aagtatttcaac**CACAGATGGCACTGCCAACA**ggt WT
ac**TCATACAATATTTAATCCAG**aagtat--caac**CACAGATGGCACTGCCAACA**ggt △2
TTAATCCAGaagtattt*gctatgcaaacataaaacgattattgagatacatttaaagttatagggaaaactggaaa*caac**CACAGATGGCAC** +58

HEK 293 *TNFR1* KO clone



agctgt**TGGTGGGAATATACCCCTCA**ggggttattgga**CTGGTCCCTCACCTAGGGGA**caggg WT
agctgt**TGGTGGGAATATACCCCTCA**ggggtt--tgga**CTGGTCCCTCACCTAGGGGA**caggg △2
TGGTGGGAATATACCCCTCAggg*tatacctgacatacccctgacatacccctaagtgtgtgtctctg***GGTCCCTCACCTAGGGGA** △11,+43

HEK 293 *TNFR2* KO clone



cttccttccagg**TGGCATTTACACCCTACGCC**ccggagcccggg**AGCACATGCCGGCTCAGAGA**atactatgac WT
gcaggggga----------//----------ggg**AGCACATGCCGGCTCAGAGA**atactatgac △124

(Continued)

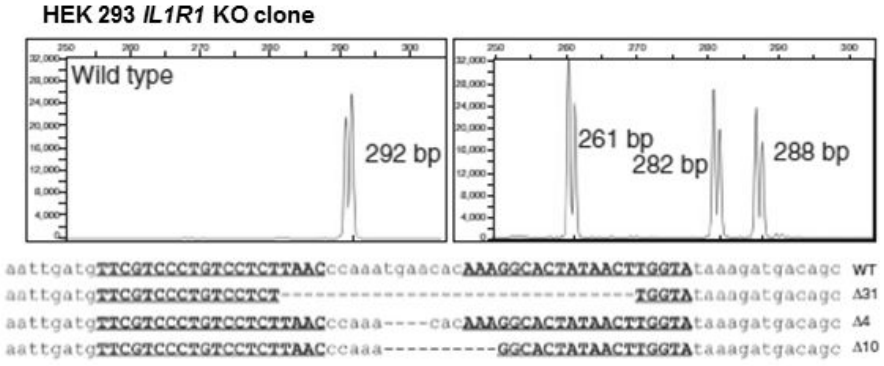


Figure 21. Validation of knockout cell lines. Genotypes of single clones were validated by F-PCR and sequencing analyses. TALEN recognition sequences are underlined and shown in boldface. The numbers of deletions and insertions are indicated. WT, wild-type.

I searched potential off-target sites that are highly homologous to the target site, and no mutations were detected using T7E1 assay (Table 6 and Figure 22). And then, to validate that target protein were not expressed in gene-knockout cells, I performed western blot analysis and RT-PCR (Figure 23a, b). This result showed that these knockout clones are functionally gene disruptions without no mutations at other potential off-target sites.

Table 6. Potential off-target sites of TALENs used for knockout cell lines

	Chromosome number	Gene name	Left half-site (5' to 3')	No. of mismatches	Right half-site (5' to 3')	No. of mismatches	Spacer (bp)
ON		*TNFR1*	TGGTGGGAATATACCCCTCA		TCCCCTAGGTGAGGGACCAG		12
OFF	5	*FAM169A*	TGcTGGcAATggACCtCaCA	6	TCCCCTAGGaaAGGctCCAG	4	14
OFF	16	N/A	cGGTaGaAAcAgACCCCTCt	6	TggCCatGGTGAGGGcCCAG	5	12
OFF	3	N/A	TGGTatGtATtTACCCCTtA	5	atCCCagGGaGAGGGACtAG	6	13
OFF	5	N/A	TGGTGaGgAcATgCtCCTCA	5	TCCaCcAtGTGAGGacaCAG	6	13
OFF	6	N/A	TtGTGGGAtcAgAgCCCcCA	6	TCCaCTAGGTGgtGccCCAG	5	13
ON		*IL1R1*	TTCGTCCCTGTCCTCTTAAC		TACCAAGTTATAGTGCCTTT		12
OFF	6	N/A	aTCtTCCCTGgCCcCTTgAC	5	atCCTAGTTATAAtGgCTTT	5	13
ON	3	*VPRBP*	TgCGaaCCTGcCCTgcTAAC	6	aAaCAAGTTATAcctCCTTT	5	12
OFF	2	N/A	TTgGTgCtTcTCCTCaTAAC	5	TttCAAaTTATAaTGaCgTT	6	13
ON	7	N/A	TTgcTCtCTGTCtTCTgAAC	5	TtCCAAcaTAaAGTGCCTTg	5	14
OFF	4	N/A	TTtcTCtCTcTCCTCTcAAC	5	TACCcAtTTcTAGTtCtTTg	6	14

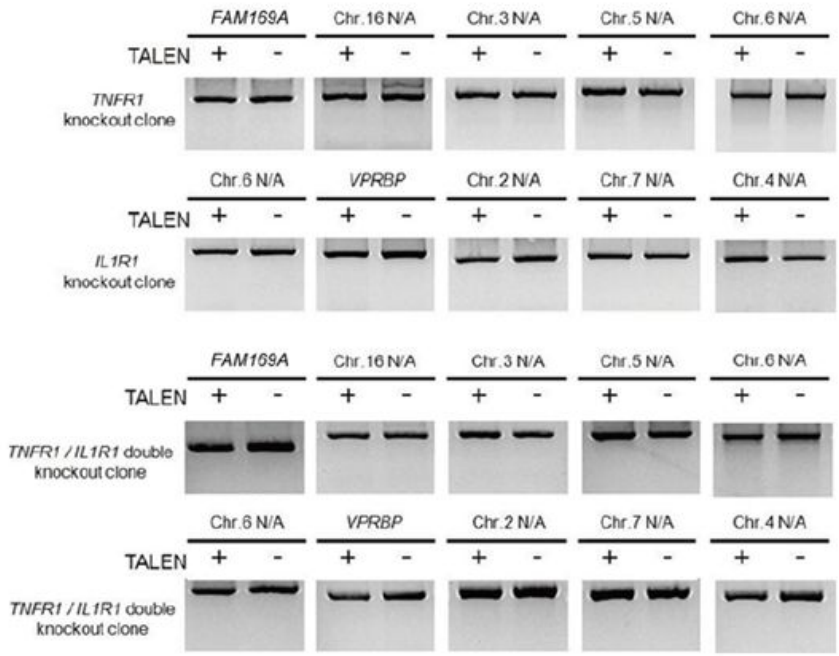
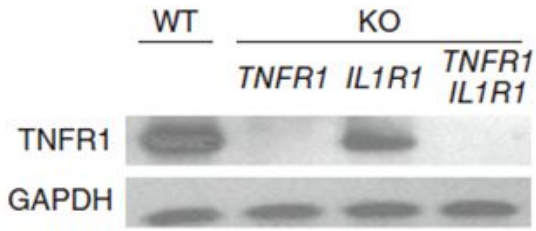


Figure 22. Undetectable off-target mutations in TALEN-mediated knockout cell lines. The top 5 potential off-target sites of *TNFR1* and *IL1R1* TALENs were searched for *in silico*. The T7E1 assay was used to investigate whether these sites were mutated in clonal populations of single- and double-gene knockout cells. No mutations were detected at these sites.

a



b

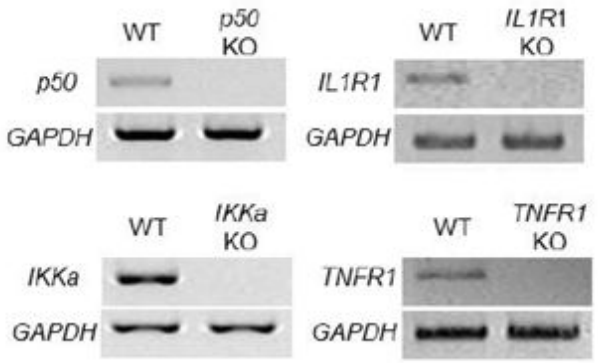


Figure 23. Undetectable gene expression level in gene knockout cells.

(a) No protein expressions in single- and double-gene-knockouk cells by Western blotting. (b) Total RNA was extracted from each knockout cell line, and cDNA was synthesized using reverse transcriptase. Target mRNA levels were analyzed by PCR using appropriate primers (Kim et al. 2013b). *GAPDH* RNA levels were used as an internal control.

2. Episomal reporter assay

For the study of NF-κB pathway study with these knockout cell lines, I constructed luciferase reporter plasmid that inserted multi-copy of p65 binding site in upstream of luciferase gene (Figure 24). Luciferase was regulated by induction of NF-κB signaling pathway such as treatment of tumor necrosis factor alpha (TNFα) or interleukin-1 beta (IL-1β), two well-known cytokines that activate NF-κ B signaling. First, we compared TALEN-mediated gene knockout with siRNA-mediated gene knockdown (Figure 25a). Two validated siRNAs, each specific to one of the two receptor genes, *IL1R1* or *TNFR1*, only partially suppressed the NF-κB signaling by IL-1β or TNFα, respectively. Thus, the two cytokines still activated the NF-κ B-dependent reporter activities in siRNA-transfected wild-type cells, compared to cells not treated with cytokines ($P < 0.01$, Student's *t*-test). In contrast, both *IL1R1* and *TNFR1* knockout cells showed complete suppression of IL-1β- and TNFα-mediated NF-κB signal transduction, respectively. In addition, the two cytokines did not activate the luciferase reporter in double-knockout cells in which both *IL1R1* and *TNFR1* were disrupted. Transfection of the *IL1R1* cDNA into *IL1R1* knockout cells restored IL-1β-mediated signal transduction, demonstrating that the gene-knockout cells were not impaired in NF-κB signaling (Figure 25b). Unexpetedly, over-expression of *TNFR1* in *TFNR1* KO cells activated NF-κB-dependent reporters even in the absence of TNFα. Similar results have been reported by previous study

(Gaeta et al. 2000).

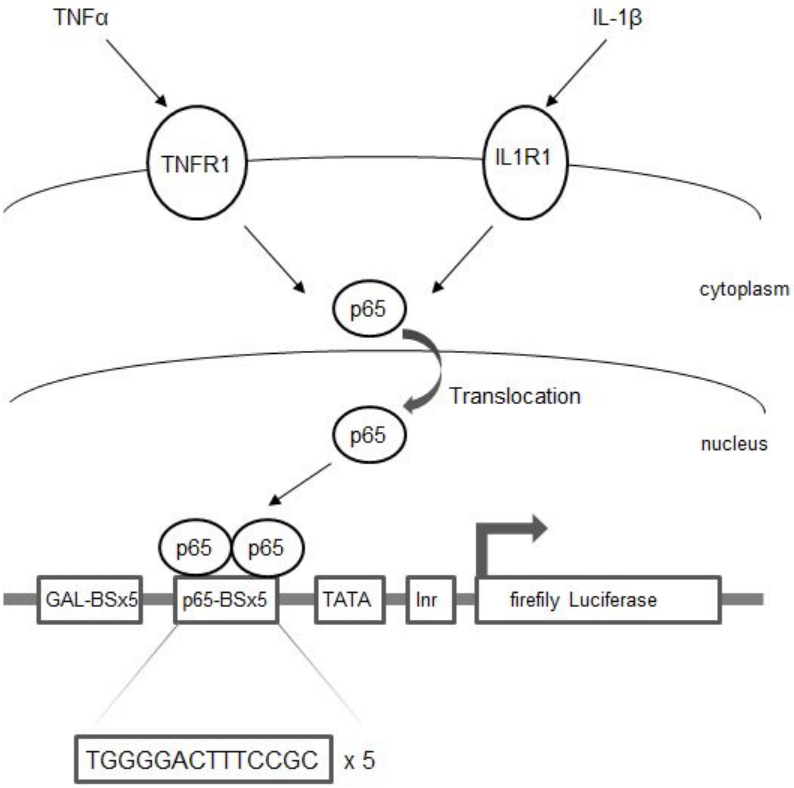
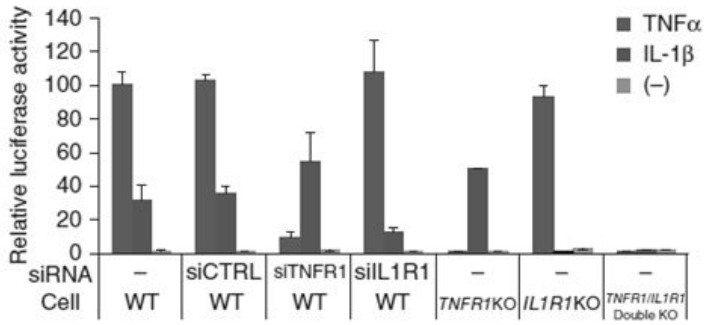


Figure 24. Schematic of reporter assay related NF-κB signaling. Gene-knockout or wild-type cells were co-transfected with the luciferase reporter plasmid and the Renilla luciferase plasmid. After 24 h of incubation, cells were treated with TNFα or IL-1β and incubated for 15 h. Cells were lysed and detected luciferase activities.

a



b

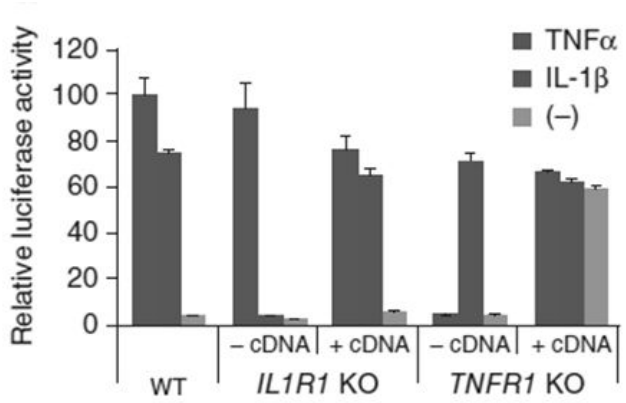


Figure 25. Functional assay in knockout cells using reporter.

(a) Comparison of TALEN-mediated gene knockout (KO) with siRNA-mediated gene knockdown. (b) Transfection of cDNA into gene-knockout cells restores cytokine-dependent NF-κB activation. The *IL1R1* or *TNFR1* cDNA was transfected into *IL1R1* or *TNFR1* knockout cells, respectively. Note that the overexpression of the *TNFR1* cDNA in *TNFR1* knockout cells induced NF-κB activation even in the absence of TNFα.

Next, mitoxantrone (Novantrone), an anti-cancer drug that induces nonspecific DSBs in cells, strongly activated the reporter gene in the single- and double-gene-knockout cells, demonstrating that NF-κB activation triggered by DNA damage is independent of signaling through the two cytokines (Figure 26). These gene-knockout cells can be used for dissecting NF-κB pathways and screening for reagents or factors that selectively modulate NF-κB activation independent of signaling via IL-1β or TNFα or both.

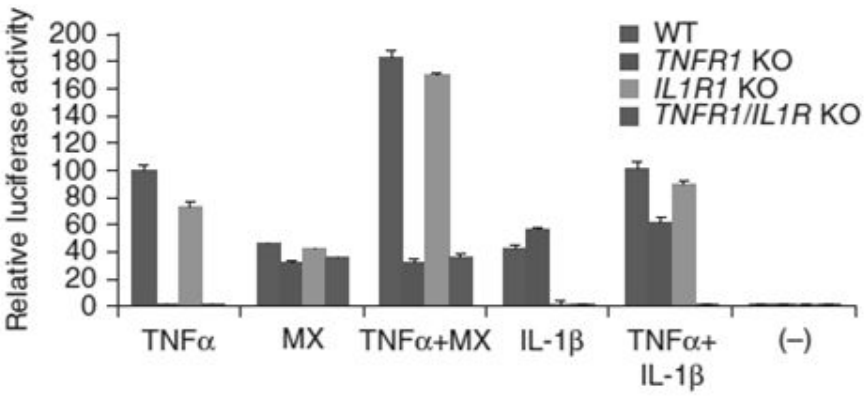
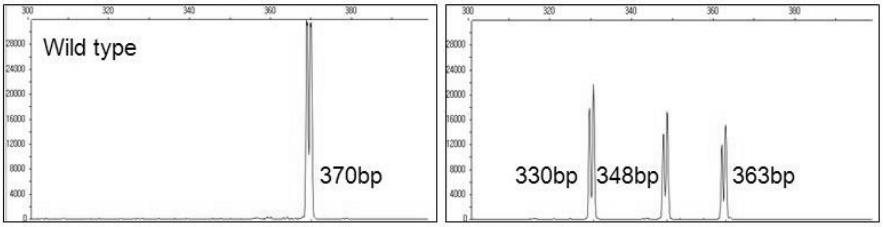


Figure 26. Cytokine-independent NF-κB activation by mitoxantrone. Gene-knockout or wild-type cells were co-transfected with the luciferase reporter plasmid and the Renilla luciferase plasmid. After 24 h of incubation, cells were treated with TNFα/IL-1β/MX and incubated for 15 h. Cells were lysed and detected luciferase activities.

Furthermore, to demonstrate the usefulness of TALEN-mediated gene knockout further, we disrupted the *NR1H4* and *SMEK1* gene in HEK293 cells (Figure 27). These genes had been identified as a potential mediator of NF-κB activation in a recent genome-wide siRNA screen (Gewurz et al. 2012). I prepared knockout cell lines with TALENs and surrogate reporters and confirmed by fluorescent PCR and Sanger sequencing. Both IL-1β and TNFα strongly activated the luciferase reporter gene in three or two independent knockout clones, respectively (Figure 28,29). To validate expression level of each genes, I performed RT-PCR and It appears that *NR1H4* is not expressed even in wild-type HEK 293 cells. In case with *SMEK1* gene, gene expression level was detected in wild-type cells but not detected in two independent knockout cell lines. These results showed that these genes were a false positive in siRNA screen.

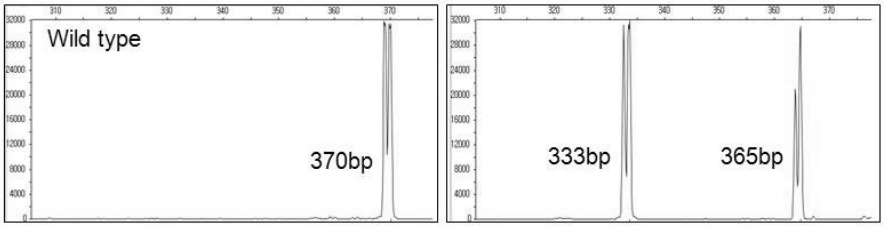
a

HEK 293 *NR1H4* clone #1



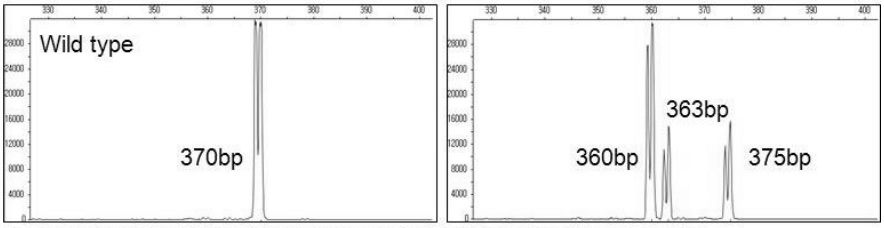

```
ttatTCCAACCTGGGTTTCTACCCCCAGCAGCCTGAAGAGTGGTACTCTCCTGGAAtatatgaac WT
ttatTCCAACCTGGGTTTCTACCCC-------CTGAAGAGTGGTACTCTCCTGGAAtatatgaac △7
ttatTCCAAC----------------------CTGAAGAGTGGTACTCTCCTGGAAtatatgaac △22
ttatTCCAACCTGG----------------------------------------AAtatatgaac △40
```

HEK 293 *NR1H4* clone #2




```
ttatTCCAACCTGGGTTTCTACCCCCAGCAGCCTGAAGAGTGGTACTCTCCTGGAAtatatgaac WT
ttatTCCAACCTGGGTTTC----------------------------------tatatgaac △37
ttatTCCAACCTGGGTTTCTACCCCCAGggg-----AGAGTGGTACTCTCCTGGAAtatatgaac △8,+3
```

HEK 293 *NR1H4* clone #3




```
ttatTCCAACCTGGGTTTCTACCCCCAGCAGCCTGAAGAGTGGTACTCTCCTGGAAtatatgaac WT
ttatTCCAACCTGGGTTTC----------AGCCTGAAGAGTGGTACTCTCCTGGAAtatatgaac △10
ttatTCCAACCTGGGTTTCTAC-------AGCCTGAAGAGTGGTACTCTCCTGGAAtatatgaac △7
ttatTCCAACCTGGGTTTCTACCCCCAGgtagaagtaCTGAAGAGTGGTACTCTCCTGGAAtata △4,+9
```

(Continued)

b

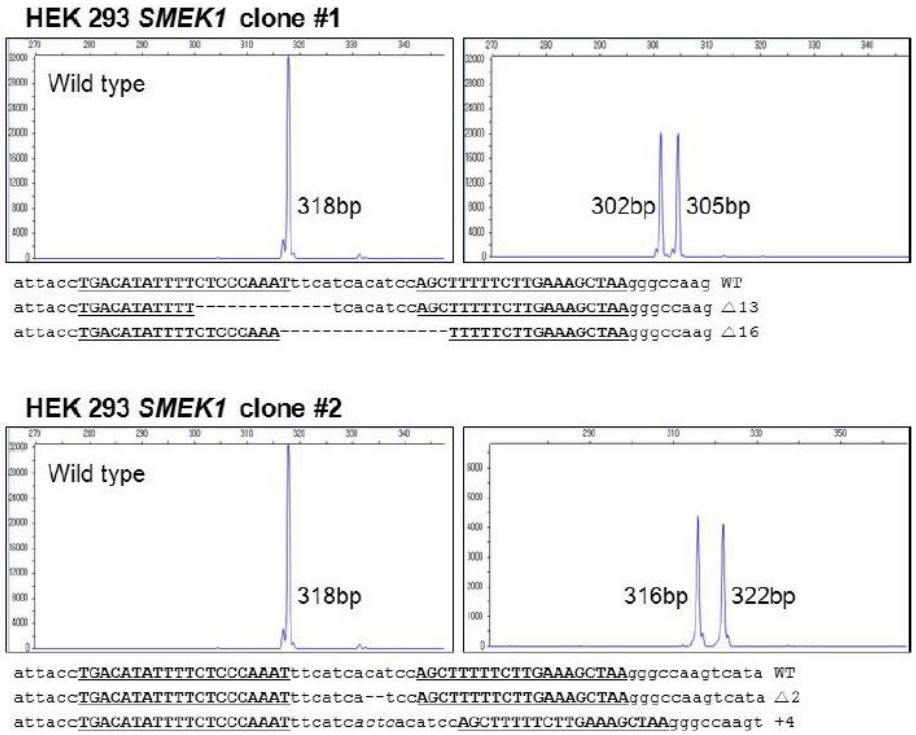
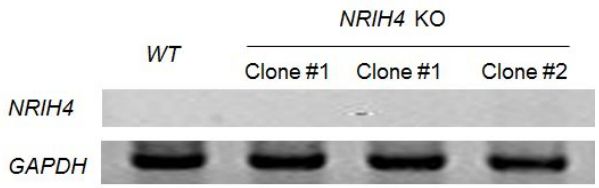


Figure 27. Validation of knockout cells related siRNA screening. Genotypes of single clones were validated by F-PCR and sequencing analyses. (a) NR1H4 gene, (b) SMEK1 gene.

a



b

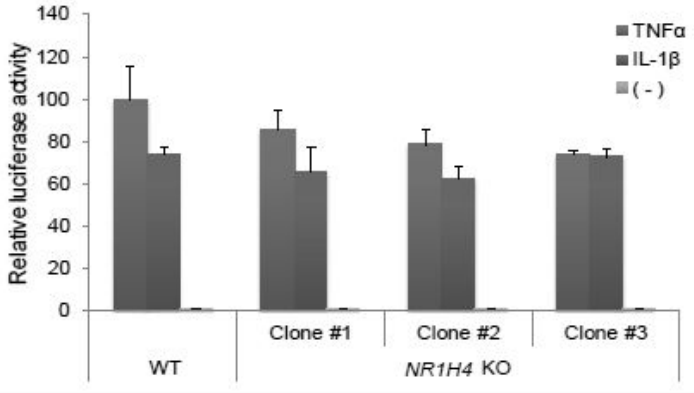
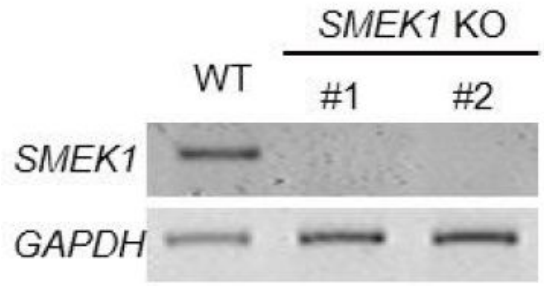


Figure 28. Invalidation of a false-positive gene identified in a genome-wide siRNA screen: *NR1H4*. (a) *NR1H4* gene was not expressed in HEK 293 cells. Total RNA was extracted from each knockout cell line, and cDNA was synthesized using reverse transcriptase. Target mRNA levels were analyzed by PCR using appropriate primers. (b) Three independent clones of *NR1H4* knockout cells were treated with cytokines, and the luciferase reporter activities were measured. Error bars indicate S.E.M. from at least three independent experiments

a



b

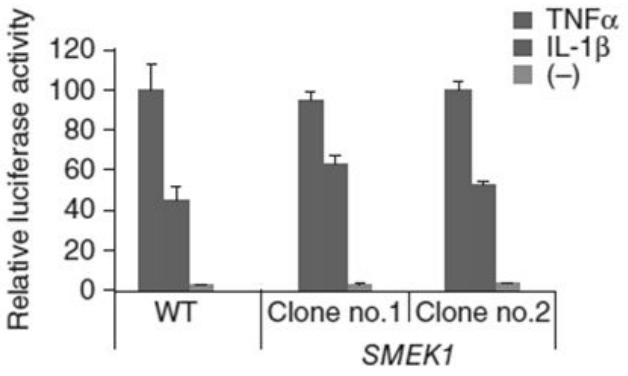


Figure 29. Invalidation of a false-positive gene identified in a genome-wide siRNA screen: *SMEK1*. (a) No expression of *SMEK1* gene in two independent *SMEK1* gene-knockout clones. (b) Two independent clones of *SMEK1* knockout cells were treated with cytokines, and the luciferase reporter activities were measured. Error bars indicate S.E.M. from at least three independent experiments.

D. Expansion of TALEN library

1. Design strategy of TALEN library

In described above, because human genome-wide TALEN library was designed for gene disruption, they do not used other functional studies such as gene correction, targeted transcript disruption, specific cancer related oncogene mutations. For the expanded use of TALEN library, I designed TALEN pairs for every exon for human protein coding genes. I obtained the DNA sequences of defined human protein coding genes from the Ensembl project in April 2013 (Birney and Ensembl 2003). Then I developed improved computational strategy to identify TALEN target sites in each exon according to the following criteria.

1) From the exon sequence database, I identified 40-bp target sequences with 12- or 13-bp spacers and start with the base T and end with A in each exon.
2) As described above, I searched for unique TALEN target sites with the minimum number of potential off-target sites using Bowtie preferred without potential methylated sequences.
3) Then, I comfirmed that the on-target sites indeed exist in human genome sequence by searching sequences using Bowtie.
4) To identify additional target sites, I loosened the criteria about off-target that mentioned 2). Therefore, I selected TALEN target sites with minimum potential off-target sites preferred on-target

site alone.

As a result, I designed TALEN target sites for 99.4 % (22,567 / 22,693) of genes and 87.8% (541,985 / 617,233) of exons (included overlapped exons). I identified a total of 3,242,203 target sites (144 sites per gene, 6.0 sites per exon on average (Table 7)

To validate the TALEN mediated mutagenesis activity I selected *NRAS* gene, commonly mutated in myeloid leukemia (Brose et al. 2002), and synthesized TALENs targeted every exons in *NRAS*. *NRAS* gene consists of 7 exons, and I synthesized 6 TALEN pairs (exon 6 is too small to design target sites) (Table 8). These TALENs were transfected to HeLa cells and I performed T7E1 assay. All TALENs could induce targeted mutations with high efficiencies (Figure 30).

Table 7. Summary of TALEN target sites for exons in human protein-coding genes.

Human								
Criteria		Genes			Exons			Target sites
Off-target	CpG sites	number of genes	%	Avg.	number of exons	%	Avg.	
>6 (3+3)	0	21360	94.1	101.3	340,940	55.2	6.3	2,162,853
	<3	21644	95.4	129.7	458,616	74.3	6.1	2,807,219
	<5	21674	95.5	134.6	483,157	78.3	6.0	2,916,278
	all	21681	95.5	136.2	491,834	79.7	6.0	2,953,457
all		**22,567**	**99.4**	**143.7**	**541,985**	**87.8**	**6.0**	**3,242,203**

Table 8. List of TALEN target sites for targeting *NRAS*

Exons	Target sequence	spacer (bp)
Exon1	TTTTCCCGGCTGTGGTCCTAAATCTGTCCAAAGCAGAGGCAGTGGAGCTTGA	12
Exon2	TGGTTGGAGCAGGTGGTGTTGGGAAAAGCGCACTGACAATCCAGCTAATCCA	12
Exon3	TGGACATACTGGATACAGCTGGACAAGAAGAGTACAGTGCCATGAGAGACCAA	13
Exon4	TGATGTACCTATGGTGCTAGTGGGAAACAAGTGTGATTTGCCAACAAGGACA	12
Exon5	TTACACACTGGTAAGAGAAATACGCCAGTACCGAATGAAAAAACTCAACAGCA	13
Exon7	TTCAGTCTCACAGAGAAGCTCCTGCTACTTCCCCAGCTCTCAGTAGTTTAGTA	13

a

b

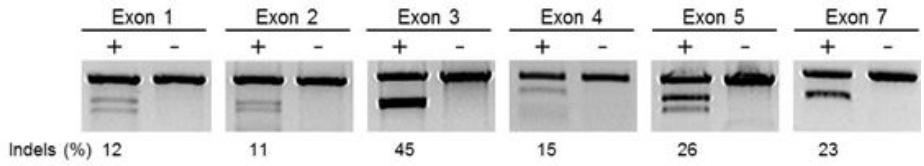
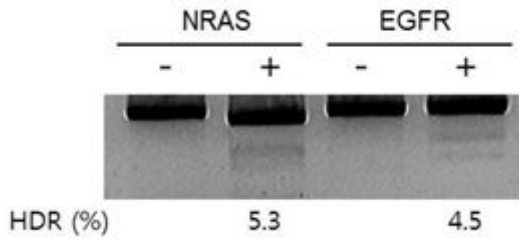


Figure 30. TALEN-mediated mutations in each exons of *NRAS*. (a) schematic of *NRAS* gene. *NRAS* is consist of 7 exons. (b) Targeted mutation in each exons of *NRAS* were detected by the T7E1 assay. all of TALENs were highly active.

As a proof-of-principle experiment for targeted gene modification, I performed TALEN-mediated oncogenic mutations at two genes, *NRAS* and *EGFR*. Q61K mutation at *NRAS* gene and L858R mutation at *EGFR* gene are well known for hyperactive oncogenic mutations in primary cancers (Brose et al. 2002;Paez et al. 2004). With designed TALEN pairs targeting *NRAS* and *EGFR* with appropriate ssODN contained PstI and HindIII respectively, I transfected human K562 cells. PCR amplicons of these chromosomal regions were partially digested by appropriate restriction enzymes, demonstrating ~5% genome-editing efficiency (Figure 31a). I determined the DNA sequences of PCR products representing the targeted genomic regions and found that substitutions were induced in targeted regions (Figure 31b).

a



b

Figure 31. TALEN-mediated homology depenent rapair (HDR). (a) HDR mediated gene-editing in *NRAS* and *EGFR*. TALENs and ssODN including each PstI and HindII restriction enzymes, respectively were transfected to K562 cells and after 3 d of incubation, genomic DNA was isolated, and the target locus was amplified with appropriate primers. TALEN with ssODN-mediated mutations were detected appropriate restriction enzyme treatment. (b) sequence analysis of targeted modifications.

2. TALEN library for several organisms

TALENs are also applicable for several organisms such as mouse, rat, zebrafish and C. elegans. In the previous studies, TALENs have been successfully used for making gene-knockout animals and plants. Therefore, as human TALEN library, I designed TALENs targeting protein coding genes for mouse, rat, zebrafish and C. elegans (Table 9).

To validate these designed TALENs, I synthesized TALEN pairs targeting several genes in zebrasifh (Table 10). Then TALENs transcripts were injected to zebrafish's embryo and performed T7E1 assay to detect mutagenesis activities. The TALEN pairs were highly active from 10 to 50 % of populations as results of TALENs for human genes (Figure 32). I verified the TALEN induced mutation sequences by dedeoxy sequencing (Figure 33). These result showed that TALEN libraries for several organisms are useful tools for gene editing. The TALEN library reported here provides a foundation for functional genomic studies.

Table 9. Summary of TALEN library for several organisms

Mouse								
Criteria		Genes			Exons			Target sites
Off-target	CpG sites	number of genes	%	Avg.	number of exons	%	Avg.	
>6 (3+3)	0	20,552	90.4	60.4	214,446	58.6	5.8	1,241,791
	<3	20,790	91.5	81.9	288,750	78.9	5.9	1,703,261
	<5	20,842	91.7	84.1	299,762	82.0	5.8	1,753,450
	all	20,850	91.8	84.5	303,208	82.9	5.8	1,762,478
all		**22,509**	**99.1**	**84.8**	**326,654**	**89.3**	**5.8**	**1,908,519**

Zebrafish								
Criteria		Genes			Exons			Target sites
Off-target	CpG sites	number of genes	%	Avg.	number of exons	%	Avg.	
>6 (3+3)	0	22,986	87.6	39.7	170,174	54.4	5.4	911,721
	<3	23,517	89.6	59.7	239,671	76.6	5.9	1,403,493
	<5	23,547	89.8	61.0	244,915	78.3	5.9	1,436,532
	all	23,551	89.8	61.1	245,305	78.4	5.9	1,438,631
all		**26,139**	**99.6**	**62.8**	**274,157**	**87.7**	**6.0**	**1,640,393**

Rat								
Criteria		Genes			Exons			Target sites
Off-target	CpG sites	number of genes	%	Avg.	number of exons	%	Avg.	
>6 (3+3)	0	19,712	85.9	31.1	115,404	54.9	5.3	613,438
	<3	20,444	89.1	43.6	158,799	75.6	5.6	891,864
	<5	20,580	89.7	44.5	163,505	77.8	5.6	915,305
	all	20,599	89.8	44.6	164,436	78.2	5.6	917,917
all		**22,397**	**97.6**	**45.2**	**178,838**	**85.1**	**5.7**	**1,011,929**

Drosophila								
Criteria		Genes			Exons			Target sites
Off-target	CpG sites	number of genes	%	Avg.	number of exons	%	Avg.	
>6 (3+3)	0	13,397	96.1	21.0	45,445	58.8	6.2	281,360
	<3	13,674	98.1	34.6	69,176	89.5	6.8	473,747
	<5	13,680	98.2	36.7	72,788	94.2	6.9	501,596
	all	13,680	98.2	36.9	73,138	94.6	6.9	504,289
all		**13,931**	**100.0**	**36.7**	**74,054**	**95.8**	**6.9**	**511,880**

Table 10. List of TALEN target sites for zebrafish

Genes	Target sites	spacer (bp)
BRAF	TTAAACAGATGATTAAGTTAACTCAAGAGCACCTAGAAGCCCTTTTAGATAA	12
Rad51	TGGGCAGTGATGTTCTGGATAACGTGGCCTACGCCAGAGCCTTCAACACTGA	12
ATM	TCACCAGACACTGTGGAAGAACTTGATCGTACATCAGGAAGCAAAGGCTCCA	12
PTENB	TTTATCCTAACATAATAGCTATGGGTTTCCCTGCTGAAAGACTGGAGGGTGTA	13
CASP7	TTCTCTGTGTAGTACTTACAACGTTTATTTGGAATAGCAGAATACAAGACTGA	13
Rock2a	TGAACCATATAAGGGAGCTCCAGATGAGGCCAGAGGATTTTGACAGAGTGAA	12
Fzd2	TGGTGGGACACTACAACCAAGATGATGCCGGGCTTGAGGTCCATCAGTTTTA	12
Stat3	TCAGTTGCAGCAGTTGGAGACGCGGTATCTGGAGCAGCTGTATCACCTGTACA	13

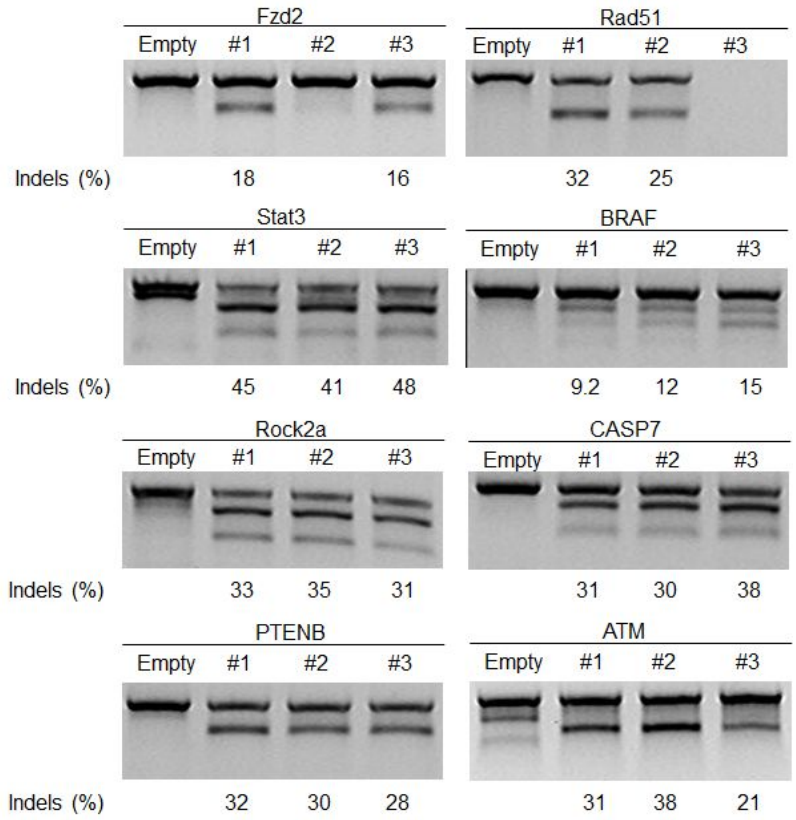


Figure 32. TALEN-mediated gene disruption in zebrafish. TALEN-encoding mRNAs were injected zebtrafish embryos and after 1 d incubation, genomic DNA was extracted from pools of 8 embryos. TALEN induced mutations were detected by the T7E1 assay. each numbers of genes indicated independent pools of embryos.

Figure 33. Sequencing validation of TALEN-induced indels in zebrafish. Endogenous mutations induced by 10 different TALENs were confirmed by dideoxy DNA sequencing. The numbers of inserted or deleted bases are shown on the right side of each mutant sequence. WT, wild-type sequence.

E. Comparison of TALENs and ZFNs

TALENs and ZFNs share the same FokI nuclease domain and induce site-specific DNA cleavages, whose repair via error-prone NHEJ gives rise to indels at the target sites. I investigated difference of TALENs with ZFNs in patterns of mutations (Kim et al. 2013c).

First, I compared mutation signatures of ZFNs and TALENs that have been reported in the literature. I calculated frequencies of insertions, deletions, and complex patterns that accompany both insertions and deletions at each target site in the following categories: mammalian cells, mammalian organisms, non-mammalian animals, and plants. my analysis included a total of 1,456 mutant sequences at 122 target sites reported in 43 independent studies (Figure 34). ZFN-induced mutations are much more evenly distributed between deletions and insertions in any organism or cell line than are TALEN-induced mutations. In particular, insertions are rarely obtained with TALENs. For example, ZFNs induce insertions in mammalian cells at a frequency of 43%, comparable to that of deletions (52%). In sharp contrast, TALENs induce insertions at a frequency of 1.6%, much lower than that of deletions (89%). In all other systems combined, the ZFN-induced insertion frequency is much higher than the TALEN-induced insertion frequency (24% vs. 4.1%) (Student's *t*-test, $P < 0.05$). I also found that ZFNs are associated with lower deletion frequencies (59%) as compared to TALENs (81%) ($P < 0.05$). Complex patterns are obtained at comparable frequencies (17% with ZFNs and

15% with TALENs).

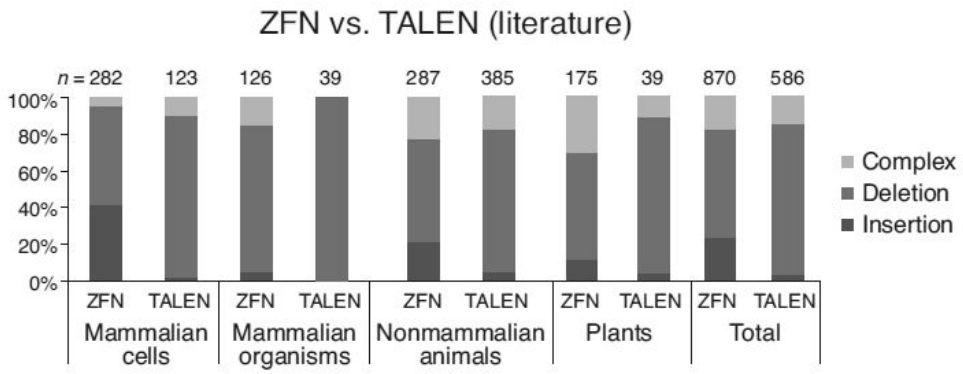


Figure 34. ZFN and TALEN mutation patterns reporterd in the literature. Nuclease-induce mutant sequences were classified as three different type: insertions, deletions or comlex (that is, having both insertions and deletions) patterns. The number of mutant events is shown above each other.

Next, I investigated that pattern of TALENs induced mutations different with that of ZFNs by several TALENs and ZFNs whose target sites overlap with each other (Figure 35). I transfected HEK 293 cells with them, and compared mutation patterns using Sanger sequencing methods. In line with my analysis described above, TALENs, unlike ZFNs, rarely produced insertions. Thus, five TALENs induced insertions at an overall frequency of 2.6%. In sharp contrast, ZFNs induced insertions much more frequently (39%) ($P < 0.05$) (Figure 36). This results showed that the differential mutation signatures induced by TALENs and ZFNs do not arise from the differences in target loci.

Figure 35. Overlapping target sites of ZFNs and TALENs. Both the left and right half-sites are underlined. Spacers are shown in small latters.

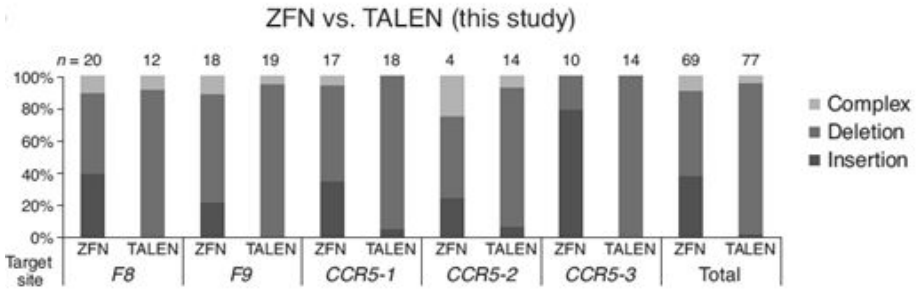


Figure 36. Comparison of ZFNs and TALENs that target overlapping sites. Nuclease-induce mutant sequences were classified as three different type: insertions, deletions or comlex (that is, having both insertions and deletions) patterns. The number of mutant events is shown above each other.

Furthermore, I investigated the molecular basis of the difference in mutation patterns. I hypothesized that the larger spacers (12 to 21 bp) in the TALEN sites as compared to the ZFN sites (mostly 5 to 6 bp) might cause the difference. To test this idea, I expressed two homodimeric ZFNs in cells whose genome was modified to contain ZFN sites with 15-bp spacers (Figure 37a). Unlike ZFNs that target sites with 5- to 6-bp spacers, these ZFNs produced insertions much less frequently than deletions (1.8% vs. 86%) (Figure 37b). I speculated that the larger spacers either in ZFN sites or in TALEN sites may give rise to heterogeneous overhangs when cleaved by nucleases. In contrast, ZFNs that target sites with 5- to 6-bp spacers produce defined 4- or 5-bp overhangs, which might be efficiently filled in before ligation to yield insertions. Defined overhangs may facilitate targeted insertions of plasmid DNA at genomic sites (Maresca et al. 2013).

a

ZFN 1

Deletion

```
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaagcttgaagtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA WT
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaag-------tatAAACTGCAAAAGGCTGAAGAGCATGACTGACA (3)
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTac----ttgaagtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA (2)
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaagcttgaa------ACTGCAAAAGGCTGAAGAGCATGACTGACA (4)
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTT------------------------GGTTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaa-------------ACTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaagcttga-gtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTG---------------aagtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaagcttgaa-----AACTGCAAAAGGCTGAAGAGCATGACTGACA (2)
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTaca--cttgaagtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTac---cttgaagtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTT-----cttgaagtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaagcttg------------CAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaagcttg----atAAACTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaagc------------------AAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTT--------gaagtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaagcttgaa-------CTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTT-------ttgaagtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA
```

Complex

```
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaagcttgaagtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA WT
ATCCTCATCCGGAATTCTAGAAACTTTTGCAG-------acttgaagtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTTacaagcttga----cAAACTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCC--------------------------------------atAAACTGCAAAAGGCTGAAGAGCATGACTGACA
ATCCTCATCCGGAATTCTAGAAACTTTTGCAGTTT----cttgaaagtatAAACTGCAAAAGGCTGAAGAGCATGACTGACA
```

ZFN 2

Insertion

```
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCccaagcttgaagttgGATGAGGATGACTTCTAGATCTACGTAAAC WT
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCccaagcaagcttgaagttgGATGAGGATGACTTCTAGATCTACGT
```

Deletion

```
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCccaagcttgaagttgGATGAGGATGACTTCTAGATCTACGTAAAC WT
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCcca--cttgaagttgGATGAGGATGACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCccaagcttga-------TGAGGATGACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCccaagc----agttgGATGAGGATGACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCc----cttgaagttgGATGAGGATGACTTCTAGATCTACGTAAAC (3)
TTTGGTTTTGTGGGCAACATGCTG--------------------------gGATGAGGATGACTTCTAGATCTACGTAAAC
AAAACCTA-----------------//----------------------gGATGAGGATGACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGC-------------------------------------------ATGACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCcc---cttgaagttgGATGAGGATGACTTCTAGATCTACGTAAAC (2)
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCcc-------aagttgGATGAGGATGACTTCTAGATCTACGTAAAC (3)
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCA---------tgaagttgGATGAGGATGACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCAT------cttgaagttgGATGAGGATGACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCc-----ttgaagttgGATGAGGATGACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCAT-------------------------GACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCcca----------------------CTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATG----------------------aagttgGATGAGGATGACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATGCTGGT----------------------tgGATGAGGATGACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCc-----------------------------------------TAAAC
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCccaa---------tgGATGAGGATGACTTCTAGATCTACGTAAAC
TTTGGTTTTGTGGGCAACATGCT--------------------tgaagttgGATGAGGATGACTTCTAGATCTACGTAAAC
```

Complex

```
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCccaagcttgaagttgGATGAGGATGACTTCTAGATCTACGTAAAC WT
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCctga-------------------------------------TAAAC (2)
TTTGGTTTTGTGGGCAACATGCTGGTCATCCTCATCcc-----aaaagttgGATGAGGATGACTTCTAGATCTACGTAAAC
```

b

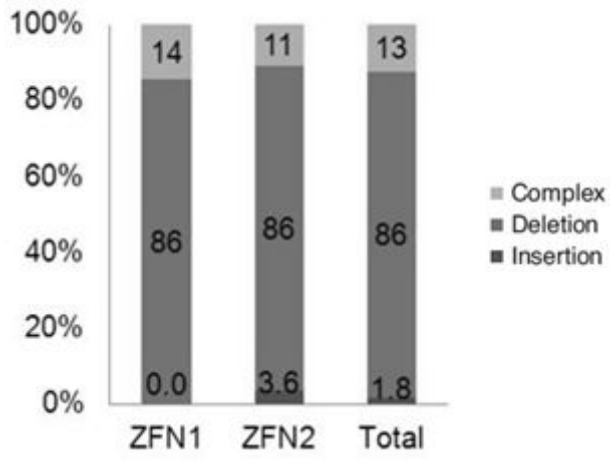


Figure 37. Mutations induced by ZFNs with 15-bp spacer. (a) I modified the *CCR5* target site of ZFN-224 by cotransfecting K562 cells with ZFN-224 plasmids and oligonucleotides as described previous study (Chen et al. 2010). The two monomers that constitute ZFN-224 were used as homodimers to cleave the modified sites with 15-bp spacers. Sequences recognized by ZFNs are underlined. Spacers are shown in small letters. (b) Mutation signatures of ZFNs that target sites with 15-bp spacers.

IV. Discussion

The TALEN technology has recently developed, which can be rationally designed *de novo* due to its highly predictable modularity. Unlike ZFNs, which have sequence-bias toward guanine-rich sites (Maeder et al. 2008), TALENs have no sequence-bias toward any base and, thus, can be designed to target any sequence.

Over the last two years, TALENs have been reported for genome editing in yeast (Li et al. 2011), roundworm (Wood et al. 2011), fruitfly (Liu et al. 2012), cricket (Watanabe et al. 2012), zebrafish (Huang et al. 2011;Sander et al. 2011;Bedell et al. 2012), frog (Lei et al. 2012), mouse (Sung et al. 2013), rat (Tesson et al. 2011), pig (Carlson et al. 2012), cow (Carlson et al. 2012), thale cress (Cermak et al. 2011), rice (Li et al. 2012), silkworm (Ma et al. 2012) and human cells (Hockemeyer et al. 2011;Miller et al. 2011;Reyon et al. 2012;Kim et al. 2013b). Although previous TALEN architectures are highly active with high success rates, for genome-scale study, they are limited some aspects. 1) The success rates of TALEN-mediated gene disruption range from 64% to 88% in mammalian cells (Cade et al. 2012;Carlson et al. 2012;Reyon et al. 2012). These success rates are certainly impressive compared to those achieved with ZFNs. On a genomic scale about 20,000 genes, however, failure rates of 12–36% correspond to 2,000–7,000 genes, if one TALEN is used per gene. 2) The previous TALEN architectures have poorly defined fusion junctions between TAL effectors and FokI cleavage domain (Christian et al.

2010;Li et al. 2011;Miller et al. 2011). Different with ZFNs that have a narrow range of spacers (5- to 6-bp) (Kim et al. 2009), these TALEN constructs induce mutagenesis in broad spacers (12- to 22-bp) that may induce indels at unwanted loci in genome.

To this end, I developed TALEN architectures with 12- to 14-bp spacers that might reduce unwanted mutations in genome. Furthermore, using obligatory heterodimeric FokI nuclease domains, the TALEN constructs were improved target specificities through elimination of TALEN monomer's homodimeric activity. And my thesis suggested that a nearly 100% success rate could be achieved with TALENs by avoiding heavily methylated sites, making it feasible to initiate genome-scale gene-knockout studies. This high success rates were important for TALEN mediated genome scale study.

I developed a scalable Golden-Gate assembly system that consists of a total of 432 plasmids (424 TAL effector array plasmids and 8 FokI plasmids) and used this system to construct a TALEN library, a collection of 18,740 TALEN pairs, to disrupt or modify every protein-coding gene in the human genome. Recently, two different solid surface–based assembly methods for the synthesis of TALEN plasmids were reported: iterative capped assembly and fast ligation-based automatable solid-phase high-throughput (FLASH) systems (Briggs et al. 2012;Reyon et al. 2012). Although these methods can be automated at least partially, multiple cycles of ligation and washing steps in addition to PCR are involved before subcloning into an expression vector. Furthermore, FLASH uses up to microgram quantities of TAL effector

array plasmids in each reaction, making it cost-inefficient and cumbersome to scale up. Compared to these methods, My Golden-Gate cloning system allowed PCR-free TALEN assembly with much less hands-on time and greatly reduced plasmid quantities. In addition, cloning efficiency was higher in general with Golden-Gate systems than with solid surface–based methods.

The one-step Golden-Gate cloning systems can use several other applications of TAL effectors. The DNA sequence based binding protein is valuable resource for targeted gene regulation by fusing VP64 activation domain or KRAB repression domain (Maeder et al. 2013b; Perez-Pinera et al. 2013). TAL effectors fused histone demethlytransferase are also useful options for targeted gene regulations (Maeder et al. 2013a). These gene regulators enable the gain of function studies. In addition, because they are not induced permanent gene modifications like engineered nucleases, they can be used for transient or inducible gene regulations. TAL effector's random or target library pools also can used phenotype screening as zinc finger protein library.

TAL effectors also can be used targeted chromosomal visualizers to combine with fluorescent proteins (Miyanari et al. 2013). Unlike other chromosomal visualizers such as FISH that need fixed cells by chemicals and only processed in death cells, TALE fluorescent visualizers enable to detect target sequences in living cells. So, they can use to detect the shape and loci of targeted chromosomes during the cell cycles.

Then, I established computational strategy for TALENs targeting almost protein coding gens in several organisms. As described results above, I identified 40-bp target sequences with 12- or 13-bp spacers in each gene, which were recognized by TALEN pairs that consist of two 18.5 RVDs with the minimum number of potential off-target sites. The ten of TALENs that were highly active at their on-target sites, did not induce any measurable mutations at their most likely off-target sites.

In the short term, TALEN-mediated gene-knockout experiments should be useful for validating or invalidating the functions of genes identified by genome-scale siRNA library screens, which have revealed many potential drug targets associated with various diseases (Sigoillot and King 2011). Unfortunately, many of the newly identified genes have turned out to be false positives that resulted from off-target effects (Lin et al. 2005;Adamson et al. 2012). To distinguish true positives from false positives, one could use TALENs to knock out candidate genes as demonstrated here, rather than using another siRNA to downregulate the genes. In the long term, TALENs could enable the creation of a series of cell lines in which a family of druggable target genes encoding proteases, kinases, phosphodiesterases and G protein–coupled receptors are disrupted. In principle, genetic screens can be performed using these cell lines instead of siRNA libraries. The TALEN library reported here provides a foundation for functional genomic studies based on gene knockout rather than knockdown in mammalian cells.

Furthermore, the TALEN pairs in advanced libraries for every exon in protein coding genes of several organisms can be used to generate not only gene disruption, but also precise genetic modifications by homology directed repair. For example, as described above, single nucleotide changes can be made in a gene of interest to study phenotypic differences associated with single nucleotide polymorphisms in an otherwise isogenic background. A reporter gene can be fused to a gene of interest to monitor the expression of the gene or to track the protein.

Although the primary use of my TALENs in drug discovery will be to identify and validate druggable target genes, I envision that some TALENs or their inproved versions would be useful therapeutically themselves. An example for a targeted nuclease that is in clinical trials in Sangamo's CCR5-specific ZFN for the treatment of AIDS, whish is now under clinical investigation in the United States (Holt et al, 2010). TALENs can also be used to disrupt disease-associated genes or correct genetic defects in stem and somatic cells. I propose that the one-step Golden-Gate cloning system and TALEN library presented in this thesis will be widely used for basic and biomedical research.

V Appendix

1. List of TALEN activity in reporter assay

Gene ID	sequence	Avg.	S.E.M
1544	TTGTCCCAGTCTGTTCCCTTCTCGGCCACAGAGCTTCTCCTGGCCTCTGCCA	28.6	1.3
5354	TCCATGCCTTCCAGTATGTCATCTATGGAACTGCCTCTTTCTTCTTCCTTTA	3.4	0.1
7007	TTAAGTTGGCCATCCCAGTTTTCTTCTTTGGCGTTCCTTACCGCACTGTCTA	14.2	1.0
2875	TGGTGGCCGGCGAGGGCCACACACGCACGGGTGTGCTCATCCCCATCCCCCA	32.9	2.0
11023	TCACCGCGGAGCAGCTCTATCGGCTGGAGATGGAGTTCCAGCGCTGCCAGTA	3.3	0.3
866	TCTCTGGAGTCTATGACCTCGGAGATGTGCTGGAGGAAATGGGCATTGCAGA	20.7	3.3
146059	TCACCCCCAATCAGCTGTGTCCCCAGTTCCCAACCCTCAGCCCTGGACACTA	5.3	0.9
4893	TACAAAACAAGCCCACGAACTGGCCAAGAGTTACGGGATTCCATTCATTGAA	4.0	0.7
119	TGGAGAAGGGCAGCAGCTGCTTCCCAGTGGACACCACAGGCTTCTGTCTGCA	11.8	0.7
1401	TTCACTGTGTGCCTCCACTTCTACACGGAACTGTCCTCGACCCGTGGGTACA	12.5	1.2
257	TCCAAAGCTGCCAGCTTCCCCCAGCTGCCCTTGGACTGCCGAGGGGGCCCCA	25.2	0.8
1675	TTCTCCTGGGCGCGCACTCCCTGTCGCAGCCGGAGCCCTCCAAGCGCCTGTA	6.4	0.5
8908	TCTTCAGGAACTGGTCGACCACAGACATCCACAAGCACCTGCCGTTCATCTA	22.1	1.7
1588	TCATCTCCCACGGCAGATTCCTGTGGATGGGGATCGGCAGTGCCTGCAACTA	22.0	1.5
958	TCTGCGAGCCCTGCCCAGTCGGCTTCTTCTCCAATGTGTCATCTGCTTTCGA	16.2	2.2
1759	TGGCATTGTCACCCGACGTCCCCTGGTCTTGCAGCTGGTCAATGCAACCACA	21.3	2.5
9228	TGCCCTCGACCTCCCGGGATGTTTCCGAACAAGGAGTCACAGCTACCTTCGA	9.3	1.2
491459	TCTGGAGAACATGGAATTGTGGTTTTCTCTTTGGGATCAATGGTCTCAGAAA	38.4	1.5
5309	TCGCTGTCAGACGCTGGCACTCCGCACCCCCAGCTCCCAGAGCACGGCTGCA	38.9	1.0
11277	TGCAGACCCTCATCTTTTCGACATGGAGGCCACTGGCTTGCCCTTCTCCCA	14.1	1.3
2902	TTAGCCATCCACCTACCCCCAACGACCACTTCACTCCCACCCCTGTCTCCTA	14.2	2.4
2321	TACTTAGAGGCCATACTCTTGTCCTCAATTGTACTGCTACCACTCCCTTGAA	17.0	1.0
2693	TGGCCGACCTGGACTGGGATGCTTCCCCCGGCAACGACTCGCTGGGCGACGA	17.1	2.2
54716	TCCCCTACATCATCATGCTTATCGTGGAGGGAATGCCGCTCTTGTACCTGGA	4.9	0.5
722	TTTCTACGCATACGGCTTTTCTGTCACCTACAGCTGTGACCCCCGCTTCTCA	12.5	1.2
5798	TACAGGACATCCCCACTGGCTCCGCCCCTGCTGCCCAGCATCGGCTTCCACA	40.6	2.3
8745	TATGGTCTTCCAATGGCAGTGGCACAAGTATTATCGCAGAGCCTGGCTCAAA	24.0	2.7
2671	TACTACCCCGACCTGCCCACCCCAGAACAGCAGCAAGACATGGCCCAGTTCA	17.9	2.3
4747	TGGCTACTCCCAGAGCTCCCAGGTCTTTGGCCGATCTGCCTACGGCGGTTTA	6.6	0.8
59344	TCTTCCCCTCCTCCTGGATCACAGGACACGGGAGCTCCGGGCCCGACAAGAA	34.3	4.2
7433	TTTATTTGCATCATCCGAATCCTGCTTCAGAAACTGCGGCCCCCAGATATCA	12.3	1.0
4920	TTGCCTGTGCACGCTTCATTGGCAACCGGACCATTTATGTGGACTCGCTTCA	13.4	1.1
3816	TGGGAGTGTGAGCAGCATTCCCAGCCCTGGCAGGCGGCTCTGTACCATTTCA	26.3	2.2
3697	TCCCAGAACTCAGCAACCATGCCTCAATACTCATCATGTTGACAGATGGCGA	17.5	1.5
7441	TGTCCTCGGCCCTTGGAACCACAATCCGCCTCACCTGCACCCTGAGGAACGA	2.0	0.4
50943	TCGGCCCCTTCCTTGGCCCTTGGCCCATCCCCAGGAGCCTCGCCCAGCTGGA	14.6	1.2
5345	TGCTCAGAACCACACGTTGCAGAGGCTGCAACAGGTGCTGCACGCAGGCTCA	23.9	2.5
20	TGCAATCGCTGTGCCCGGACGGCCAGCGAGACGAGTTCGGCTTCCTGCAGTA	6.7	0.9
81031	TCCTGCCTTTGTGTGCCTCTGTGTCTTTGCTGGGTGGCCTGACCTTTGGTTA	2.4	0.3
4602	TTCCCAAGTCTGGAAAGCGTCACTTGGGGAAAACAAGGTGGACCCGGGAAGA	23.4	2.7
4867	TCTTCCAGGAATGAGCATACAGGTTCTCAGCAGACATGTACGCCTCTGTCTA	8.3	0.5
6622	TGGAGGGAGCAGGGAGCATTGCAGCAGCCACTGGCTTTGTCAAAAAGGACCA	20.9	0.7
268	TGGACCGCCCTGCGGGGGCCTGGCGCGGCTCCGGGCTGGCCTTGACCCTGCA	26.3	3.0
2100	TTAGTGGGAACCGTTGCGCCAGCCCTGTTACTGGTCCAGGTTCAAAGAGGGA	26.6	1.6
6331	TCCTCAGTCCCTTCCACCCCATCGGGAGAGCGGCTGTGAAGATTCTGGTTCA	20.9	2.7
6338	TACCCAGGCATTGACAGAGTGGTACATCCTGCAGGCCACCAACATCTTTGCA	36.2	2.7
4669	TGTGGTGGGACTGGGCCCGCTGGGAGCGAGAGATAGACTGGATGGCGCTGAA	40.3	1.9
7976	TGTTCTCGGGATTTCCGGCCTTTTCTTTGTGCACTCTACGCTCCTATTTGTA	24.4	2.0
6531	TCCTGGTCCCCTACCTGCTCTTCATGGTCATTGCTGGGATGCCACTTTTCTA	16.4	3.3

Gene ID	sequence	Avg.	S.E.M
1311	TCCACTGCGCGCCCGGCTTCTGCTTCCCCGGCGTGGCCTGCATCCAGACGGA	18.6	1.3
26254	TCCTCCATCGATAATGATGCCTTCCGCCTGCTACATGCCCTCCAGGACCTCA	14.2	1.8
1183	TTGGGAATAGCCGTCTCGTTCTCTTTTATGTGGAATACCACACGCCCTGGTA	22.0	1.4
2239	TGGTCAGCGAACAGTGCAATCATTTGCAAGCTGTCTTTGCTTCACGTTACAA	22.0	2.5
197	TGCACCTCCGTCCCCTCCACTTGGCGCACCTGGACTCCCTCCAGCTGGCTCA	26.5	2.5
5837	TGCCTGGCTATCGCAACAATGTTGTCAACACCATGCGCCTCTGGTCTGCCAA	13.1	1.9
8224	TGGAAAACCAGCTTGACTTCCAGGACATCACCAGCGTGGTCGCCATGGCCAA	25.7	3.2
5365	TCCGCACCTAATACCACTCTCAACCACTTGGCACTGGCACCTGGCCGAGGCA	22.4	2.1
785	TCTGACTCCGATGTCTCTTTGGAAGAGGACCGGGAAGCAATTCGACAGGAGA	31.5	0.7
57211	TTTGGAGGATGGAACACGTCAGGATGTGTTGCACACAGAGATTCAGATGCAA	1.6	0.2
2036	TCCCAAACCACCGGTCAGCCAAGAGACTGTGGAAGGTCTGCATCGAGCATCA	25.5	1.8
6017	TCGGGGGAGGAGCTGGCGGTGGCCGTGGCGGAGAGGGTGCAAGAGAAGGACA	25.3	4.6
4486	TGGGTGACAGTGGGCAGCCCGTGCAGCGGGATGTCAGTCGTCTTGGGGACCA	20.1	2.5
6556	TGCTGCAGGCGGTGGGCATTGTTGGCGCCATCATCATGCCCCACAACATCTA	11.6	1.6
1369	TGTGCGAGGAGTTCCGGAACAGGAACCAGCGCATCGTCCAGCTCATCCAGGA	22.3	1.3
3786	TCTGGGCTGCTGGATGTTGCTGCCGATACAAAGGCTGGCGGGGCCGACTGAA	39.0	1.2
26047	TACAGCATCCGATTATTGCCCGCTATGTGCGCATAGTGCCTCTGGATTGGAA	19.0	3.9
7137	TCAGACGCCGCTCCTCCAACTACCGCGCTTATGCCACGGAGCCGCACGCCAA	27.2	2.5
2645	TGGGGCTGCGACCCTCGACCACCGACTGCGACATCGTGCGCCGCGCCTGCGA	22.3	2.5
5459	TGGCGGCGGTGGATATCGTCTCCCACGGCAAGAACCATCGTTCAAGCCCGA	15.5	3.7
6091	TGCAGTAATGGAATGCCAACCTCCACGAGGCCATCCTGAGCCCACCATTTCA	14.3	2.5
11107	TGCTACCAACCCACGGCACTCCAACTGGCTTCGCTTCGTTCATGAGGCACCA	33.3	3.1
3046	TTCACCCCTGAAGTGCAGGCTGCCTGGCAGAAGCTGGTGTCTGCTGTCGCCA	11.1	1.1
63976	TCCCCGTGGGGCCACCGTCCCCCTTCCCCACCAGCGAGGACTTCACCCCCAA	20.0	1.5
90	TTCGCAGTATGTCTTTTAGCCTGCCTGCTGGGAGTTGCTCTCCGAAAATTTA	5.2	0.4
81693	TGGAGGAGGCACGAGGCGGCGGCCCCGGCTGGAGCGCCCCTCGGCTTCCGCA	29.3	1.6
7054	TGCGCCGAGGGGACCTGGCCGCCCTGCTCAGTGGTGTGCGCCAGGTGTCAGA	16.6	1.0
5346	TGGCCTGCCGAGGCTTGGACCACCTGGAGGAAAAGATCCCCGCCCTCCAGTA	2.3	0.4
10743	TGGTTTCCATGGCAAACAACAGAACTACCAGCAGACCTCGCAGGAAACATCA	28.5	3.4
489	TGCGGCCAGTTCGACGGGCTGGTGGAGCTGGCGACCATCTGCGCCCTGTGCA	27.1	2.5
650	TTGGAAGAACTACCAGAAACGAGTGGGAAAACAACCCGGAGATTCTTCTTTA	8.0	1.3
3640	TGGTGGCCGACAGTAATCTCACGCTGGGACCTGGCCTGCAGCCCCTGCCCCA	32.7	2.4
7224	TGTACTACAAAAAGGTCAACTACTCACCGTACAGAGACCGCATCCCCCTGCA	3.0	0.1
50939	TACCAAAAAGAAACAGCCTCTGGACCGCAGAGAAACTGAAAGACAGTGGTTA	1.9	0.3
5071	TGGGGGGCGTGTTATGCCCCCGCCCTGGCTGTGGAGCGGGGCTGCTGCCGGA	15.2	0.8
2588	TTTTCCTCTACTGGGCTGTCGACGCCACGCACGCACCCGTCTATGCCTCCAA	25.7	1.3
51206	TTTTGCCAAACCCTCGCTCTCAGCCCAGCCCGGCCCGGCGGTGTCGTCAGGA	24.1	0.8
2516	TGCAGAACAACAAGCACTACACGTGCACCGAGAGCCAGAGCTGCAAGATCGA	12.9	1.0
1789	TCGATCCTCGTCAACGGGGCCTGCAGCGACCAGTCCTCCGACTCGCCCCCAA	23.2	1.7
50508	TTTCTTCAACCTGGAACGCTACCACTGGAGCCAGTCCGAGGAGGCCCAGGGA	0.8	0.1
63970	TCTTCCTCTGAGGCGAGCTTCAGATCTGCTCAAGCTTCCTGCAGTGGGGCCA	37.3	1.1
9333	TCCAGAGCTCCAGAAATAATGTGCGGCACCACACGGAGGAGATCACTGTGGA	20.5	1.7
1583	TGGTGCAAGTGGCCATCTATGCTCTGGGCCGAGAGCCCACCTTCTTCTTCGA	22.4	1.0
2253	TCATCGTGGAGACGGACACCTTTGGAAGCAGAGTTCGAGTCCGAGGAGCCGA	27.1	1.5
6492	TGTCCCCCTTCGACGGCTGCTACCAAAACGTGGGCCTGGTGGCCGTGGGCCA	36.0	3.4
6491	TCAAGCAGGATGGTACCTTTCCACTTTCCTCCATCAAAATGTGCACTTTGGA	19.0	0.6
2914	TCCGCCGCCTCCTGGAGACTTCGAACGCCAGGGCAGTCATCATCTTTGCCAA	13.6	0.4
11173	TCGAGTCGACGCGGGGGGCTCCTTCCTGTCCTACGAGCTGTGGCCCCGCGCA	12.0	0.5
51196	TACAGTTGGATCTCTACTCCATTTCCTCACCAAGCTCCCAGCCTCCGAGACA	23.1	1.0

Gene ID	sequence	Avg.	S.E.M
10195	TCTTCACCTCCAACTTCATTGGCATCTGCTTCAGCCGCTCCCTCCACTACCA	18.1	1.3
6010	TCAAGCCGGAGGTCAACAACGAGTCTTTTGTCATCTACATGTTCGTGGTCCA	24.8	1.7
29760	TCAAGTCACAGACAAGAAGCTGTGCAGTCACCAGTGTTTCCTCCTGCCCAGA	27.1	1.5
3363	TGGCTTCTCTCCGCCTCCATCACCTTACCTCCACTCTTTGGATGGGCTCAGA	37.1	1.5
56606	TCTCCCGGACAGCCCACGCTACCTGCTCTTGGAGAAGCACAACGAGGCAAGA	13.6	1.1
6530	TCGCCCTGTACGTTGGCTTCTACTACAACGTCATCATCGCCTGGTCACTCTA	7.1	0.4
6517	TTCTCTGCGGTGCTTGGCTCCCTGCAGTTTGGGTACAACATTGGGGTCATCA	28.2	2.2
2532	TCTCCCCCTCAACTGAGAACTCAAGTCAGCTGGACTTCGAAGATGTATGGAA	32.1	1.3
8022	TCCTCAAGGCTCTGGACCGCCACTGGCACAGCAAGTGTCTCAAGTGCAGCGA	24.3	1.6
415	TCCTCCTCTTTGTTTCCTTTCTACACGTTCACATCCCTCTTATCACTATGGA	20.2	1.2
3730	TTCAAAGACGACGACCCACTCGCCCGCTGGAAGTCGGAGCTCCCTTCTATCA	13.8	1.2
1748	TCCGAGCGGCGCCCTCAGGCCCCCGCCAAAAAGCTCCGCAAGCCGAGGACCA	29.6	1.0
3918	TGTCTGCCAGGCTTCCACATGCTCACGGATGCGGGGTGCACCCAAGACCAGA	22.6	0.8
367	TGCCCATTGACTATTACTTTCCACCCCAGAAGACCTGCCTGATCTGTGGAGA	30.2	1.1
81030	TCTACAGGTTTCTCAAAGACAATGGTCCCCAGAGGGCCCTGGTCATCGCCCA	30.0	1.7
4330	TCGAGCCTCCGTGCCCACCTGGCGGTCCCTGCATTCTGACATCTCCAACAGA	27.9	2.7
4693	TGGTGCTCCTGGCCAGGTGCGAGGGGCACTGCAGCCAGGCGTCACGCTCCGA	34.3	1.3
84667	TCCGGTTTCCGCGAGTGCCTGCTTCGCTTGGCGGCCTTCGCGCACGACGCCA	24.3	2.3
6528	TGTTCTACACTGACTGCGACCCTCTCCTCCTGGGGCGCATCTCTGCCCCAGA	16.0	0.6
5317	TGGCTGACAATTACAACTATGGGACCACCAGCAGGAGCAGCTACTACTCCAA	17.3	1.6
8038	TCCAAGGGACCACCATCGGCATGGCCCCAATCATGAGCATGTGCACGGCAGA	6.6	0.5
6524	TCGGCGGCCGCCGCATCCGCCTCTACCTGTCTGTGCTCTCCCTTTTCCTGTA	13.3	2.5
1592	TCTACAAGACGCATCTGTTCGGGCGGCCCACCGTACGGGTGATGGGCGCGGA	10.3	0.4
2122	TTCGCTCTCAGCATGTCGGTGCCCGGGCCCATGCATGCCCGGAGTGTGGCAA	31.4	2.8
6497	TCCCAGTCCCCCGCGCCTTCCGAAAAGGACAAGCCGTCCAGCTGGCTGCGGA	31.0	2.1
3977	TCCAGGAAGGGTGACAGCGTTGGTGGGCCCACGTGCTACAAGCTACACTTTA	18.8	0.9
1180	TCCTACGCGCAGATGCAGCCCAGCCTTCCTCTGCAGTTCCTGGTCTGGGTCA	28.9	0.5
8862	TCCGCCACCTGGTGCAGCCCAGAGGGTCAAGGAATGGGCCAGGGCCCTGGCA	33.9	2.7
215	TACAGCGCTCCTACCAGGACCTGGCCTCGCAGATCAACCTCATCCTTCTGGA	18.4	2.0
132884	TGCTCTTGCTGGACAGCATTGCTGGTCTCACCATTTGGGACTCTGTGGGAAA	21.7	2.5
2153	TGGCCAGCCGCCCCTATAGCATTTACCCTCATGGAGTGACCTTCTCGCCTTA	12.2	0.8
4916	TGGAGCACTGCATCGAGTTTGTGGTGCGTGGCAACCCCCCACCAACGCTGCA	31.6	1.7
2903	TCTGAAGTGGTGGACTTCTCTGTGCCCTTTGTGGAAACGGGAATCAGTGTCA	36.2	1.4
360	TGGCCCAGGTTGTGCTCAGCCGGGGCACCCACGGTGGTTTCCTCACCATCAA	20.0	2.1
9907	TCTTCACGGTGCTCTCCAGCGGCCACCGCTTCACAGACGACCAGTGGCTGAA	31.9	1.0
50814	TGGCTCTGGATTCCTGGGGCAGCACATGGTGGAGCAGTTGCTGGCAAGAGGA	7.1	0.6
55315	TGTCTCAGAGGACGACTTTCAGCACAGTTCAAACTCCACCTACAGAACCACA	30.1	1.8
3795	TCTCCCTGCTCGGAGCCCCCTGTGCCTTCATGGGCTCAATGGCTCCTGGCCA	34.3	2.5
6716	TCATTCTCAGAGGCACTGCCTTCTGCACTGGAAATGGAGTCCTTCAAGGCTA	2.9	0.4
43	TTTCTGGTTTACGGGGCCCCAGGCTTCAGCAAAGACAACGAGTCTCTCATCA	17.6	1.1
9568	TGGTGGGAGCAGGTGCACACGGAAGCCAACTCATCCCGCTGCCTCCGGAAGA	13.3	1.6
5972	TCAAAGTCGTCTTTGACACTGGTTCGTCCAATGTTTGGGTGCCCTCCTCCAA	7.4	0.2
51168	TGGAATCTATGGGCCGGAGCAGGTGCAGCAGTACAACGGACGGGCCCTGGGA	10.5	1.5
9248	TGGGGGCTGGGATCTGAGCCAGTTACAGTGCCAGATGGTCGGGTTCATCACA	16.9	2.1
3049	TCCCACCTGGACCTGAGCCCGGCTCCTCACAAGTCAGAGCCCACGGCCAGA	23.6	2.4
3954	TTGGCTGCTGCACTCCCATCCACCCTGTGTACACATCCTCCAGAGGCGATCA	31.0	2.7
9759	TGCCTCTGCAAGTGGCCCCCTCGGCAGTGCCCATGGACCTGCGCCTGGACCA	21.3	1.4
3667	TGGCGAGCCCTCCGGAGAGCGATGGCTTCTCGGACGTGCGCAAGGTGGGCTA	18.1	1.6
1280	TCCTGGCCCCCCTGGTCCTCCCGGCCCCCCTGGTCCCCCTGGTCTTGGTGGA	30.4	2.9

Gene ID	sequence	Avg.	S.E.M
83990	TTGCAGATGATTATAAAATTGCGATTCAACAGACTTACTCCTGGACAAATCA	13.2	0.9
5653	TGGAGACCACCTCCGAGGCCTTGTGTCATGGGGTAACATCCCCTGTGGATCA	36.0	2.5
1244	TGTGGATTCCCTTGGGCTTCCTATGGCTCCTGGCCCCCTGGCAGCTTCTCCA	29.2	2.5
6519	TGCACGACATTGTCCGCAGCTTCCGGCAGACCATGGACCAATACAGCACGGA	3.8	0.3
10157	TACATAGGACTGGACTGGGCAGGCTGGGGGTCAGCCTCTCCAAGGGTCTTCA	33.3	2.3
3748	TGCCCAGCCATTGACCAGCCTGCCATGTCCCCGGAAGACAAGAGCCCCATCA	26.6	2.1
85358	TGGACTTCCGCACTCGCGATGGGCTCACTGCCGTGCACTGTGCCACACGCCA	20.1	1.5
11005	TGCACCAGGGAGCATAATCCTGTCCGTGGCCCAGATGGCAAAATGCATGGAA	16.0	1.3
3850	TGGTCAGCAGCAGCACGACTTCCGCCTCCGCAGGTGGCTATGGAGGAGGTTA	9.6	0.3
7306	TTTGTCCGGGCCCTGGATATGGCAAAGCGCACAACTCACCCTTTATTTGTCA	42.0	2.6
2563	TCACCTCCACTGTGGCCTGCGACATGGACCTGGCCAAATACCCCATGGACGA	26.9	2.4
6862	TGGACCCCAACGCCATGTACTCCTTCCTGCTGGACTTCGTGGCGGCGGACAA	21.0	1.9
1586	TGGCGATGGCCACCTTTGCCCTGTTCAAGGATGGCGATCAGAAGCTGGAGAA	5.9	0.6
6927	TCGTCACGGAGGTGCGTGTCTACAACTGGTTTGCCAACCGGCGCAAAGAAGA	20.3	3.1
11081	TTATGGACCTCATCTTCGCTACCTCCGTCTGGATGGAAATGAAATCAAACCA	12.8	1.2
6493	TGTCCCTGTACGACTCCTGCTACCAGATTGTGGGGCTGGTGGCCGTGGGCCA	16.4	1.7
79400	TGCTCATCCATGGCAGCCCCATGGACAAACTCAAATTCCTCTTCCAGGTGTA	7.7	0.7
2538	TCATTACCTTCTTCCTGTTCAGCTTCGCCATCGGATTTTATCTGCTGCTCAA	10.9	1.0
54453	TCCGCCTGCCCTGTGAATTTGGGGCCCCACTCAAGGAATTTGCCATAAAGGA	10.8	0.9
4018	TGCCGAAATCCAGATCCTGTGGCAGCCCCTTGGTGTTATACAACAGATCCCA	43.4	1.3
23316	TGGCCCTTAGTAAGAGAAGTCAGGAGGCGGAGGCTGCTTTTCTGAGTGTTTA	6.1	0.7
53630	TCTCCTACTTGTCTCACACCATCCCCGATTTCACCGACAACTGCCTGATCAA	11.9	2.2
4914	TCAACGCTCTGGAGTCTCTCTCCTGGAAAACTGTGCAGGGCCTCTCCTTACA	18.4	0.7
1103	TGCCCAAACTGCCCGTGCCCCCGCTGCAGCAGACCCTGGCCACGTACCTGCA	14.6	0.3
3762	TGCTCTGGGGCCACCGATTCACACCAGTCCTCACCTTGGAAAAGGGCTTCTA	14.3	0.9
54221	TGGCTGGACACCTTGTCCGTGCCTCTGTCCATGGCTCGCATCTCAAGGTACA	26.0	0.7
2675	TCAGCGCCAAGAGCAACCATTGCCTGGATGCTGCCAAGGCCTGCAACCTGAA	12.3	0.5
2161	TGTGAACATGCCAAGCGGCCCCCACTGTCTCTGTCCACAACACCTCACTGGA	36.5	3.7
6506	TCAATGTTGTGGGTGACTCTTTTGGGGCTGGGATAGTCTATCACCTCTCCAA	0.1	0.0
89	TTGCGGGCGGCGGCGGGGGCGGCGAGTACATGGAACAGGAGGAGGACTGGGA	26.1	0.6
2904	TCTCACCCCCTTTCCGCTTTGGGACCGTGCCCAACGGCAGCACAGAGAGAAA	23.0	1.2
81607	TTCTGCCTTCTGGTGGTGGTGGTGGTGCTCATGTCCCGATACCATCGGCGCA	20.3	0.8
64218	TCAAGATTCCTACCTGTTGCCCATCTCGGAGGACAAGGTCATGGAGGGAAAA	21.0	1.7
9732	TCCGAGGAACCGTTCCATATGGCCTGTCATTGGAAATTGGAGATACAGTTCA	5.6	0.5
26960	TCTCGAGATGCCACCCTGCTGCTCTGGTACTGGAGTGGGCGGCACCATATCA	23.0	1.7
8817	TCCGGCTGGTACGTGGGCTTCACCAAGAAGGGGCGGCCGCGGAAGGGCCCCA	44.6	2.1
51156	TTTGCTGACCTTAGTGAACTCTCAGCTACTGGAAGAAATCTCCAAGTATCCA	28.6	2.1
27324	TCAGACCCTGCCCTAGGCATGCCGGATGTACTGCTACCCTTTCAAGCCCTCA	30.7	3.5
4284	TCCAGTTCGTGCTCTGCATCTTTGCCACATACGACGAGAGGCGGAATGGCCA	39.0	1.1
2318	TGTCCCAGTTCCCCAAGGCCAAGCTCAAACCTGGTGCCCCTGTTCGATCCAA	24.0	1.7
6330	TGGCTGTCGTGGGCGGGGTCATCGGGCTCCTCATCCTCATCCTGCTGATCAA	26.4	1.5
6869	TACATGGCCATCATACATCCCCTCCAGCCCCGGCTGTCAGCCACAGCCACCA	36.4	1.0
5284	TGTCCATCACGTGCTACTACCCACCCACCTCTGTCAACCGGCACACCCGGAA	28.3	2.5
6295	TACTTCTCCCGGGTCCAGGTGTATCCTCCTGTGGGGGCCGCGAGCACCCCCA	25.1	1.0
6337	TGTCGAGGCTGCCAGAGACTCTGCCATCCCTGGAGGAGGACACGCTGGGCAA	23.9	1.2
27031	TCCCCCAACACACTGATTCTTTCCCATTTTGTGGGAAGGCCCATGTCAACCA	34.5	1.6
2138	TATGGGCAAACACAGTTTACCACAGGAATGCAACAAGCTACAGCCTATGCCA	22.6	2.3
640	TGGTGGCTGGCCAGGTCACTCGTCACAGGAAGAGAAGGCTATGTGCCCAGTA	11.4	1.5
92359	TCCCTCTTGGCTGCCTTGCTCCTGGCTGTGGGGCTGGCACTGTTGGTGCGGA	26.1	1.4

Gene ID	sequence	Avg.	S.E.M
55615	TACCTGCGCCTGGAGACGCTGGTCCAGAAGGTGGTGTCGCCATACCTGGGCA	34.3	1.2
2524	TTCCCATGGCCCACTTCATCCTCTTTGTCTTTACGGTTTCCACTATATTTCA	34.4	1.1
5913	TGGTCCAGATCGACACGGCCCGGGAGCTGGAGGATGCCGACTTCCTCCTGGA	23.6	0.8
60529	TACCCCCTGACTCTGACACTGTGGGGATGGACAGCAGCTACCTGAGTGTCAA	38.3	1.6
1535	TCCTCATCACCGGGGGCATCGTGGCCACAGCTGGGCGCTTCACCCAGTGGTA	23.1	2.3
10840	TGGGGGCCGAGCTCAACGTCCTGCCCTTCTGCAGCCAATTCATCCCCATGGA	42.9	3.0
1300	TATAGCAGTAAGAGGAGAGCAAGGTACTCCTGGTCCACCAGGCCCTGCTGGA	23.9	1.7
9723	TGGCTCGAGACCCTTACTGTGCCTGGGATGGCATATCCTGCTCCCGGTATTA	35.2	1.7
1811	TGCTCAGTGATATTGTTTCTGGTATCAGCACAGGGATTGTGGCCGTACTACA	11.0	0.5
1356	TTCGGTCGTGTGGTACTTATTCAGCGCCGGAAATGAGGCCGATGTACATGGA	32.9	0.7
525	TTGCCCGCGGCCAGAAGATCCCCATCTTCTCAGCAGCCGGGCTCCCCCACAA	41.6	0.9
3663	TCAGCAACCCCCATGGCTGCCGGCTCTTCTACAGCCAGCTGGAGGCCACCCA	35.5	1.1
8483	TGCCTGCATGTGCCAGGACTTCATGCTTCATGGGGCTGTCTCCCTTCCCGGA	35.2	5.3
4621	TCCAGTACTTTGCAACAATTGCAGCTACTGGGGACCTGGCCAAGAAGAAGGA	2.0	0.5
79755	TGGGCCCCCTCAACCTCTCCAAGAAATCAGAGATAAACCTGGCAGCCACCCA	37.9	2.4
3982	TCTGTGCCTGGGTGGGGACCATCCTCCTGGTGGTGGCCATGGCAACAGACCA	21.1	2.5
6608	TGCCCAGCCGTACCCTCTGCCAGGCCACCCGAGGCCCCTGTGCCATCGTGGA	39.9	1.5
2917	TCAAACAGCTCCTGGACACCCCCAACTCCAGGGCCGTCGTGATTTTTGCCAA	9.6	0.4
3866	TCGACAACTCCCGGGTCATCCTGGAGATCGACAATGCCAGGCTGGCTGCGGA	33.1	1.0
4143	TTGATGTTCGGCTATGCTACCGACGAGACAGAGGAGTGCATGCCCCTCACCA	29.9	2.5
860	TGCTTCAGAACTGGGCCCTTTTTCAGACCCCAGGCAGTTCCCAAGCATTTCA	14.6	3.0
4016	TGGTCCCAGACCCCAACTATGTGCAAGCATCCACTTATGTGCAGAGAGCCCA	34.8	3.9
7021	TGGTGTCCCGAGCCACAGCTCGCGGCTCTCCCAGCTGGGCTCGGTGTCCCAA	30.6	2.8
1591	TCGGTGCACCTGGGCTCGCCATGCCTGCTGGAAGCGCTGTACCGCACCGAGA	21.3	1.8
3561	TCCCTCAGTGTTTCCACTCTGCCCCTCCCAGAGGTTCAGTGTTTTGTGTTCA	5.3	0.8
1950	TTTGCTGTTTGCCAATTCTCAAGATATTCGACACATGCATTTTGATGGAACA	39.7	1.5
6564	TTGTGTTTGTCCTTGGCAGTGGGATGTACAAGAAGTTCAAGCCACAGGGCAA	28.1	0.7
2623	TCACCATATGCCGGCTGGGCCTACGGCAAGACGGGGCTCTACCCTGCCTCAA	30.3	0.8
8013	TCTGAAAGGGAGGAGAGGTCGTCTGCCTTCCAAACCAAAGAGCCCATTACAA	35.8	1.7
2048	TTTCCAGTGTCAATGAGACCTCCCTCATGCTGGAGTGGACCCCTCCCCGCGA	0.2	0.0
3955	TACCCGCACACGCGGGACGTCTACGTCGGCAAGCCCAGCCTGGACAGGCCCA	17.1	1.1
79133	TTATGTCGGCGACCTTGGGCGGCGAGGGTCCCAGCGGAGAATCTTGGCCGTA	21.0	2.3
55613	TTGTCCCCTGACCCTCCGCTGCAAGAATTTCCGGGTGGCCCACTTTGTTTTA	21.8	1.0
6261	TCAAGCTCTGCCTGGCCGCCGAGGGCTTCGGCAACCGCCTGTGCTTCCTGGA	39.2	1.0
2277	TCCCAAAGATCTAATCCAGCACCCCAAAAACTGCAGTTGCTTTGAGTGCAAA	1.9	0.3
3082	TCTGGTTCCCCTTCAATAGCATGTCAAGTGGAGTGAAAAAAGAATTTGGCCA	11.5	1.9
10560	TACCTCCGTTATAAACCTGTTGTTCTACTGCAGGGGCTCAGCCTTATTGTTA	7.2	0.6
8323	TGTTCACCAAACATTGAAACTTTCCTCTGCAAAGCATTTGTACCAACCTGCA	14.6	1.6
23426	TCTCCATCGATGGAACCAGCATGGAGTACTGTACACTTGCAGAAGCAACCCA	37.4	1.0
8838	TTGTCACTGGCCCTGCAAATGCCCTCAGCAGAAGCCCCGTTGCCCTCCTGGA	18.0	1.0
64241	TCTCCTCTGAAAGTGACAACAGCCTGTACTTCACCTACAGTGGCCAGCCCAA	17.1	1.2
79444	TTCTTCTGCTATGGGGGCCTGCAGAGCTGGAAGCGCGGGGACGACCCCTGGA	24.9	1.6
1000	TCGTGTCTCAGGCTCCAAGCACCCCTTCACCCAACATGTTTACAATCAACAA	51.1	2.3
2735	TCTGCCCGGGGTGGGATGATCCCACATCCTCAGTCCCGGGGACCCTTCCCAA	39.8	0.8
862	TCCAACAATGCCACCTCCCCCAACTACTCAAGGAGCTCCAAGAACCAGTTCA	1.0	0.2
79152	TGGTGGACTGGCGAAAGCCTCTCCTGTGGCAGGTGGGCCACTTGGGAGAGAA	19.1	1.5
9719	TCTCCAGCAAACCGTGTGACCTGCACTGTACCACCGTGGACGGCCAGCGGCA	0.4	0.1
56000	TTAGCAGTAGGTCTGAACCTGTCAATCCTGGCATGCATTCTTCATCCCATCA	34.0	0.9
84623	TTCATCTCCCCTGGTGACGTGGAGAATGGCCAGAGCATCGTGTGTCGTGCCA	47.3	3.4

Gene ID	sequence	Avg.	S.E.M
7223	TGGGACATGTGGCATCCCACTCTGGTGGCAGAGGCTTTATTTGCTATTGCAA	31.5	0.5
3984	TCATGCGATGCCTGGAACACCCCAACGTGCTCAAGTTCATCGGGGTGCTCTA	19.3	1.2
1813	TGTCCTGGTATGATGATGATCTGGAGAGGCAGAACTGGAGCCGGCCCTTCAA	39.1	0.7
26278	TGTGTACAACAACGCGGTTTTCACCCCAGAGGACTGGCACGGCATTCAAGAA	18.7	0.7
3680	TCTCCATCTTCCTCCCGGGCTCCATCAACATCACAGCGCCTCAGTGTCACGA	6.4	0.4
92745	TCCTGCCCAGTCGTGGTCCTGCTCCTGGGAGCAAGCCGGTCCAGTTCATGGA	12.8	1.0
1420	TCAGCGAGATCCGTTCCCTCCACGTGCTGGAGGGCTGCTGGGTCCTCTACGA	24.3	1.1
288	TGTTATGGAAATTTGGCCCCACTTACCAAAGGAGGACAGCAACTTGTTTTTA	1.9	0.2
4855	TCTGCCTGTGTCAGCCTGGCTATTCGGGGCCCACCTGCCACCAGGACCTGGA	34.8	0.5
11283	TCTGTAACATCAACATCTTCGCAATCCATCACAACCCCTCAGTCTGGCCAGA	44.4	1.6
5979	TGGTGGCCACGCTGCGTGTCTTCGATGCAGACGTGGTACCTGCATCAGGGGA	24.3	0.7
1757	TGGAGGTGGAGCTTCTGGCCCACACTCGGCGGGTGGTGAGCCGGGAGCTGGA	27.2	1.4
5393	TGCTGCAATCGTGGCCTTATGTCATTTCCGAAGACCTGATGTCTCTGTCCAA	19.9	2.6
6948	TCCATGACAGCGTGGTGGACAAACTTCTGTATGCTGTGGAACCTTTCCACCA	26.7	0.7
3779	TGTGTGCCGTCATCACCTACTACATCCTGGTCACGACTGTGCTGCCCCTCTA	24.5	0.9
7068	TGTCATAGACAAAGTCACGCGAAATCAGTGCCAGGAATGTCGCTTTAAGAAA	16.3	0.8
8029	TTTCACCCAACAGCCCTGCCTGCACGCTGGACAGAGACGAGTGCAGCTTCCA	29.4	0.9
2147	TCCAGCGGGTCCGGCGAGCCAACACCTTCTTGGAGGAGGTGCGCAAGGGCAA	45.4	1.8
2564	TGTCTATGGCCCCCAGCCCCAGCCTCTGGAAAATCAGCTCCTCTCTGAGGAA	3.2	0.5
7075	TGCGCCTGGCGCGCAACGGTTCGCACCAGGTCACGCTTCGCGGCTTCTCCAA	20.0	1.0
2925	TTGTGGGCCTGTTCGCCTTCTGCTGGCTCCCCAATCATGTCATCTACCTGTA	8.7	0.3
56159	TACTCCATACTGACATGCTTTTACAAAGGAAGGAAGAACTTCTTGCCAAGGA	15.0	2.2
3662	TCCCATGACGTTTGGACCCCGCGGCCACCACTGGCAAGGCCCAGCTTGTGAA	27.8	2.2
50937	TCAGCAGCAGCCCTCCTCCTGTGGTCCCTGTGGTAGCACCTTATCCTCAGGA	26.6	0.3
1181	TGCCAAAGAGGAAGCTGCTCGGATTCGCCTGGGAGGGCCTGAACCCTGGAAA	19.9	0.9
28	TCTGGGAGGGCACATTCAACATCGACATCCTCAACGAGCAGTTCAGGCTCCA	16.3	0.2
1594	TGCGCACCGTGTACGTGGCTGCCCCTGCACTCGTCGAGGAGCTGCTGCGACA	24.6	1.0
2591	TGGTCCACGTTGCTTAGAACTGTCCACAGTGTGCTCTATTCTTCACCTGCAA	20.8	1.2
4481	TCGGGGAGCAATTGGCTTTCCTGGAAGTCGAGGACTCCCAGGATATGCCGGA	24.4	1.1
7010	TATCTGATGCTGAAACATCTCTCACCTGCATTGCCTCTGGGTGGCGCCCCCA	15.0	0.3
1815	TGCGACGCCCTCATGGCCATGGACGTCATGCTGTGCACCGCCTCCATCTTCA	28.5	0.6
10660	TCTATGAATTGGAAAAGCGCTTTCTATACCAGAAGTACCTGTCCCCCGCCGA	8.3	2.8
55145	TCCCCCACCTCCTTTACCGCCTCCTGTTTCCCAGGTTGATGCTGCTATTGGA	6.0	0.3
5799	TGCCACACTCGGGGGCCTCCTGCAGGACCACGGGTCTCGACTCTTACCTGGA	24.1	1.5
3141	TCCCAGCTCCAACATAGTGCAAACTCCAGAAGATTTTAACTTGCTCAAGTCA	21.1	2.1
3195	TCCTTCACACGCCTGCAGATCTGCGAGCTGGAGAAGCGCTTCCACCGCCAGA	15.9	0.2
631	TTCATTGAAACTCCCATTCCCCTGTTCACCCAGAGCCATGGAGTCTCTCTCA	30.7	2.5
114327	TCGGACCTACATCATTCATTACTATCTTATGGATGATACGGTGGAAATTCGA	19.0	0.7
6557	TTTTTTTCCCAGCAGCTACTGGGATTCTTGCTGGTGCCAATATCTCAGGAGA	24.3	1.8
2294	TCAACGCCATGGCGTCCTCTTCCATGCACTCGGCCGGCGGGGGCTCCTACTA	18.1	0.2
6005	TCGTTGCTGCTTTGGGCCTCCAGTGGGGCACTATTGTACAGGGAATCCTGCA	35.8	1.3
326	TCTACAAGCACCTGCCGGCTCCGCCTTCTGCAGCCCCGCTGCCAGGGCTGGA	19.5	1.0
3886	TCAAGAGCCCAGGAGGGGCCAGTGGGGGCTCCACTCGTGTGTCCGCAATGTA	28.0	1.5
10243	TTAAAAGAGGGGAATGTGTTTTGGCCAAAGGAACCCACATGGGCCCCTCAGA	30.2	1.4
2000	TACCCTGATTTACTGCATCTGTACTCGGGACTGGAGTTGGACGACGTTCACA	18.5	1.5
3815	TGTGCCTGTTGTGTCTGTGTCCAAAGCAAGCTATCTTCTTAGGGAAGGGGAA	21.7	2.1
5890	TCTCACCAGCATTCTTATCTACTACCCTTTCTGCTTTGGACGAAGCCCTGCA	22.4	1.7
4547	TTGATTCAAAGTACACGGCCATTCCCATTGTGGGGCAGGTCTTCCAGAGCCA	31.1	0.5
1326	TTGGTTCTGATTTTATTCCTCGGGGCGCCTTTGGAAAGGTATACTTGGCACA	4.8	0.6

Gene ID	sequence	Avg.	S.E.M
4318	TCTCTGGGGCCTGCGCTGCTGCTTCTCCAGAAGCAACTGTCCCTGCCCGAGA	5.4	0.4
2571	TCTTCGACCCCATCTTCGTCGGCAACCTCCTCGAACGCGGGAGCGGACCCCA	21.2	1.7
611	TGGGGCCGTGGGATGGGCCTCAGTACCACATTGCCCCTGTCTGGGCCTTCTA	18.8	2.4
5626	TGTCTCCTGCCGCCTTTTCCAGCTTCTTGCCAGAGTCCACTGCTTGCCCCTA	3.8	0.8
12	TGCCCTTTGACCCCCAAGATACTCATCAGTCAAGGTTCTACTTGAGCAAGAA	16.5	1.4
3062	TGGAACATCATCTGTAGTTCAGAGAAAATGGAAGCCCCTGCAGCCTGTTTCA	48.1	1.2
4353	TGTCCAAGTCAAGCGGCTGCGCCTACCAGGACGTGGGGGTGACTTGCCCGGA	3.6	0.3
6584	TCCCACAACATTCTGGATCTGCTTCGAACCTGGAATATCCGGATGGTCACCA	8.2	0.7
7512	TGTCACTCACAGCCCTCCGCCAGCAGATGCAGACCCAGAATCTCTCAGCCTA	15.1	0.8
6855	TGGTGGCTGGGGGTCAGTTCCGGGTGGTCAAGGAGCCCCTCGGCTTTGTGAA	23.9	1.0
9118	TACCTGCTCCCACCTAGAATCCTCAGTGCTACAACCTCCAAAGTCTCATCCA	31.7	1.7
3575	TTATCCAGCACAAAGCTGACACTCCTGCAGAGAAAGCTCCAACCGGCAGCAA	5.2	0.9
2157	TTACTGCTTCATCCTACTTTACCAATATGTTTGCCACCTGGTCTCCTTCAAA	33.0	1.7
23414	TACACTGCTGATTCCGTGATCAACTTTCACCAACACCTGTTCTCCCATCTCA	27.8	1.1
4010	TGCTTCCCTCGCGGGCAGACGGACTGCGCCAAGATGTTGGACGGCATCAAGA	18.5	1.4
2103	TCATCAAGACTGAGCCGTCCAGCCCGTCCTCGGGCATCGATGCCCTCAGCCA	9.9	0.4
5105	TTGTCCAGGGAAGCCTGGACAGCCTACCCCAGGCAGTGAGGGAGTTTCTCGA	14.8	1.1
4317	TGTATCCCAACTATGCTTTCAGGGAAACCAGCAACTACTCACTCCCTCAAGA	17.1	2.2
8854	TGCAAGCTTTTTATGTGGATTTGCAGGGCGTCATCAAAACCTTTCGATATTA	10.6	0.5
23093	TTTGTGTGCCAAGATCCTGCCCAGCGGGCATCAACTCGGCCAATTTATCCCA	25.4	1.8
383	TTTGGCAATTGGAAGCATCTCTGGCCATGCCAGGGTCCACCCTGATCTTGGA	31.4	2.0
242	TGCACAAAGAGCGGTACGCCTTCTTCCCCAAGGACCCTTGGTACTGCAACTA	29.4	1.2
6329	TCTACACCTTTGAGTCCCTCATCAAGATACTGGCCCGAGGCTTCTGTGTCGA	1.8	0.2
3425	TGCCACACAGCCAGGCTGACCAGTACGTCCTCAGCTGGGACCAGCAGCTCAA	12.0	1.6
2587	TGGAATCTCCTGGCTGCCGCACCACATCATCCATCTCTGGGCTGAGTTTGGA	21.5	2.1
6332	TTCAGATTCAATGCGGCTTCCATCTTGTGTACATTGTCTCCTTTCAATTGTA	4.4	0.2
5967	TGCCCCAGGCCCGGATCAGCTGCCCAGAAGGCACCAATGCCTATCGCTCCTA	14.1	0.2
84152	TCAGGAGAGGGGCACCATCTCAAGTCGAAGAGACCCAACCCCTGTGCCTACA	26.7	1.4
5924	TGCGCTACAACGAGGGGCACGCCCTGTACCTGGCCTTTCTGGCGCGCAAGGA	17.9	0.6
887	TGCTGCTTCTGCTCTTGTTCTTCATCCCGGGTGTGGTTATGGCCGTGGCCTA	25.4	0.5
9152	TTCTACCTGTTTGCCTCCTTTGTGTCTGTACTACCCTGGGGCTCCTGCAACA	21.3	1.2
9499	TTACCCACCACGTTTCATTCAAGTGCCAGAGAACATGTCGATTGATGAAGGA	13.0	0.7
115352	TCCAGGAACAGAACAGGCCTTACCGCTGCGGGAATCACGGGGCTGGTGCTCA	26.6	1.6
1993	TCCTATGCTCGCCCAAGTTCAGCTTCTATCAGAGATGCAAATTTATATGTCA	22.0	1.6
1124	TGGATTGAACGTACACAAACAGTGTTCCAAGCACGTTCCCAATGACTGCCAA	22.3	0.3
6315	TCTTGCAGAACCAGCCCTTTTTCGTGTGGTCCAGGAAAGTCCATGCCTGGCA	21.9	0.7
23533	TCACACGGTTCTTCAAACTTCAGTTCTTCTACGTGCCTGTGAAGCGAAGTCA	29.2	0.8
2824	TTCTCCGGGGTGGCCTTATTCTGCGGCTGTGGGCATGTGGCTCTCGCAGGCA	17.2	1.8
6514	TCACTGGGACCCTGGTTTTCACTGTCATCACTGCTGTGCTGGGTTCCTTCCA	15.8	0.7
50615	TCCACAGGTCGGCCCACAATGCCACGCATGCCACCTACACCTGCCACATGGA	21.4	2.4
2166	TCAGAGGCGCTGCTAGCCCTGCCCCTGCCTCAGCTGGTGCAGAAGTTACACA	27.6	1.3
4868	TGGAGCTGCGTTGTGGGGTCAGCACCCCTGGCAGTGCGGTGCAATGGGCCAA	4.8	0.8
1308	TACAGGAGGGCTCACTCACCTGCCTCCACTCTGCCCAACTCCCCAGGCTCAA	40.6	0.8
6928	TGCAGCAACACAACATCCCCCAGAGGGAGGTGGTCGATGTCACCGGCCTGAA	44.9	1.0
2334	TTGGCCACTCCCCAGCCCCCACCTGCAGTGCAAGCCAGCGGGGGTTCTGGCA	16.8	0.1
1013	TCCTGCAGCAGGGCAGCCCCGAGCTCTTCAGCATCGACGAGCTCACAGGAGA	31.3	2.9
5002	TGGGAGCTGTCCTCAGCTTCACCTGCATCCCCGCCAGCACCAAAGGGGCCAA	25.4	0.7
695	TGGGCAGTATCTCTGCTGCTCTCAGACAGCCAAAAATGCTATGGGCTGCCAA	19.6	0.7
6097	TGTCAAGTTCGGCCGCATGTCCAAGAAGCAGAGGGACAGCCTGCATGCAGAA	25.3	2.2

Gene ID	sequence	Avg.	S.E.M
356	TGCCTCTGGAATGGGAAGACACCTATGGAATTGTCCTGCTTTCTGGAGTGAA	8.8	0.1
7345	TGGCACAATCGGACTTATTCACGCAGTGGCCAATAATCAAGACAAACTGGGA	10.1	1.0
11281	TCAACCAGCCAATCCTCATTCCCTTCAACATGGCGGGACAGCTAGGAGGCCA	4.4	0.2
4359	TTTACACCGACAGGGAGGTCCATGGTGCTGTGGGCTCCCGGGTGACCCTGCA	33.4	1.6
5551	TTCAACGCCTCCACCCAGCCCGCCTACCTCAGGCTTATCTCCAACTACGGCA	0.3	0.1
6850	TGCTGCACTATCGCATCGACAAAGACAAGACAGGGAAGCTCTCCATCCCCGA	20.0	0.6
2042	TCAGTGGTGTGGATGAACATTACACACCCATCAGGACTTACCAGGTGTGCAA	21.4	1.9
154881	TCGGAGGGGCTCACTTCACTACACGCCTGTCCACACTGCGGTGCTACGAAGA	15.2	3.4
2898	TCTTGGGGTGGCTGCCATCTTCGGGCCTTCACACAGCTCATCAGCAAACGCA	32.8	1.1
1723	TCAGTGGTGGAACACAGGTTACGGGCCAGACAGCAGAAGCAGGCCAAGCTCA	18.4	1.1
4036	TCTGTGCTCTGCCTTGAACTGCCAGTACCAGTGCCATGAGACGCCGTATGGA	25.0	0.7
10462	TTAGCAACTTCACCTCAAACACTGTGGCGGAGATCCAGGCACTGACTTCCCA	8.9	1.9
80144	TTGCAATCATGGGGAAGTCCGATGTACCCCCCAACCATGCCCACCGCTGTCA	2.5	0.2
2357	TCATCACTTATCTGGTATTTGCAGTCACCTTTGTCCTCGGGGTCCTGGGCAA	21.2	0.9
88	TCTGCCAAAGAAGGTCTGCTGCTTTGGTGTCAGAGGAAAACTGCTCCTTATA	9.5	0.3
5143	TCCTGGCCAGTCTGCGGACCGTTCGGAGCAACGTGGCGGCCCTTGCCCGCCA	16.2	2.2
5083	TCAGCATCCAGCATGGCTCCTTACCCTACCCCAGCCCAAGTGTCGCCTTACA	27.4	2.0
1258	TGGGTCCAGAGGGTGCTGCCTCAGCCCCCAGGGACCCCTCGGAAGACCAAGA	39.5	1.4
4057	TACTCTGCCCAGACAACACTCGGAAGCCAGTGGACAAGTTCAAAGACTGCCA	27.6	1.3
7187	TCTATTGTCGGAATGAAAGCAGAGGTTGTGCAGAGCAGTTAATGCTGGGACA	42.9	1.3
23211	TCAGCCCCAGTGAGAAAGGTCACCGCAAGTACAGAGAGTACAGCCCCCCATA	12.3	0.8
5191	TCGCTGCAGCTCTGGGACACTGCCAAAGCTGCAGGGCCACTGCAAGTCTATA	22.4	1.1
27	TGCAACCGAGAAGAGGTGACTGCAGTTGTGCTGCTCTACATGGCCACTCAGA	37.0	1.2
1773	TTGCCATTGTTCCCCTGCATGCGGCCCCGGGGGACGCAGTAGCCGAGATCGA	23.2	1.5
5452	TGCTTGTGCCAGGCCACCACCTCCAGCCACCTGCTCAGTTCCTGCTACCGCA	12.1	0.3
7392	TTGCCTATTTCCCAGCGTCCAGTGTGGGAGATACTACGGCTGTGTCCGTACA	25.3	1.9
54457	TGCACTGCTGATGGTGATATCCACCTTTCTCCAGAAGAACCAGCTGCCTCTA	25.0	0.9
9060	TTGCCCTGGAGGAGTACCTTGTCTCCCATGCCATCCCTTGTTACTCCCTGGA	30.5	3.2
51151	TTGGAGGTGCCCTGGGTTACCTTTTGGGTGCTATAGACTGGGCCCATCTGGA	19.6	0.7
4038	TCCAGTGTGCCTATGGACGCTGCATCCTCGACATCTACCACTGCGATGGCGA	26.6	1.4
6583	TTCTTCGTAGGCGTGCTCCTCGGCTCCTTCGTGTCCGGGCAGCTGTCAGACA	21.7	0.7
5608	TCAAGCCTTCTAATGTACTCATCAATGCTCTCGGTCAAGTGAAGATGTGCGA	19.9	1.5
1814	TCTGGAGCTGAAGCGTTACTACAGCATCTGCCAGGACACTGCCTTGGGTGGA	23.8	1.0
54806	TTTCAGAGTTTTACCTCATCCTTCTTTTGTTTACACGGCTAAATTCCATCCA	1.9	0.5
229	TCTCTGTGGACAGTTCCATCAACCAGAGCATCGGGGGTGTGATCCTTTTCCA	23.7	3.8
64131	TCCCCAAAGACTTTGAGAATGTCGACAACAGCAACTTCGCACCCAGGACTCA	18.4	0.8
2780	TGCTGCAGTCCATCCTGGCTATCATCCGGGCCATGACCACACTGGGCATCGA	25.2	2.1
6683	TACTCAGGAAGTGACCTAACAGCTTTGGCAAAAGATGCAGCACTGGGTCCTA	9.1	0.7
3670	TTGGTTGCGGCAATCAGATTCACGATCAGTATATTCTGAGGGTTTCTCCGGA	7.7	1.2
338	TGCTCCACTCACTTTACCGTCAAGACGAGGAAGGGCAATGTGGCAACAGAAA	22.9	1.2
5091	TCCGTGTGTTCCGGGCCTGCACGGAGCTGGGCATCCGCACCGTAGCCATCTA	16.0	0.5
3785	TCCTCCTGGTTTTCTCCTGCCTCGTGCTGTCTGTGTTTTCCACCATCAAGGA	22.2	1.1
4056	TCTACCTGTTCGCGCGCCTCCGCTACTTCCAGGGCTACGCGCGCTCCGCGCA	23.1	1.0
5630	TCCGCGCGCAGTACGAGAGCATCGCCGCGAAGAACCTGCAGGAGGCGGAGGA	13.5	1.0
84000	TGGCAGGGCCACACAGGGATCAGGTACAAGGAGCAGAGGGAGAGCTGTCCCA	31.1	1.2
5654	TGGCTAGTGGGTCTGGGTTTATTGTGTCGGAAGATGGACTGATCGTGACAAA	16.8	1.0
5618	TGTCCAGGTTCGCTGCAAACCAGACCATGGATACTGGAGTGCATGGAGTCCA	30.3	0.8
3587	TTGCCTTTGTCCTGCTGCTCTCCGGAGCCCTCGCCTACTGCCTGGCCCTCCA	29.1	1.1
5308	TTCAATGGGCTCATGCAGCCCTACGACGACATGTACCCAGGCTATTCCTACA	22.7	1.2

Gene ID	sequence	Avg.	S.E.M
4211	TACTTGTACCCCCCGCGAGCCGGGGGTGGCGGGCGGGGACGTCTGCTCGTCA	29.5	1.3
1823	TCCAAAGCATTTCTCCATACACCCAGATACCGGTGTCATCACCACAACTACA	0.6	0.1
5139	TTTCTTCCACTTGGACCACCACCACCTCGGCCACAGGTCTACCCACCTTGGA	5.4	0.5
4624	TGCTGCTGGTCACCAACAATCCCTACGACTACGCCTTCGTGTCTCAGGGAGA	4.2	0.3
579	TGTGCAGCGGGGCCGGCGGCGGGGGCGGCAGCGGGCCGGCAGGCGTCGCGGA	3.1	0.1
84839	TGGCGCCGCCAGGAGCGGCTGGAGTCAGGCTCGGGTGCCGTGGCAGCTCCGA	19.0	3.8
25833	TCCACAGCAACAAAGCGGTCTCCTCCTCCCACAGACTGGGCCGGGACTGGCA	5.8	0.6
22806	TCAATAACGCCATCAGCTATCTTGGCGCCGAAGCCCTGCGCCCCTTGGTCCA	27.0	1.2
38	TGCCTTGAAGCAAGGAGAATACGGTCTTGCCAGTATTTGCAATGGAGGAGGA	25.3	0.5
189	TCTCTGGCTCGGGACACTGTGCCCTGGAGGCCGCCCTGGTCAATGTGCTGGA	7.2	0.1
247	TGCCTGGATGAAAAGACAGTGGAAGACTTGGAGCTCAATATCAAATACTCCA	34.1	0.6
284	TGGGATAAAGTGGCACTACTTCAAAGGGCCCAGTTACTCCTTACGTTCCACA	40.1	0.3
366	TTGCCCAAGCTATTCTCAGTCGAGGACGTTTTGGAGGGGTCATCACTATCAA	20.2	0.3
393	TTGCACTGCTGGGCGGTGCTGCTGCAGCACACGCGGCAGCAGAGCCGGGAGA	30.3	1.4
411	TCACACGATGTGCTCTTGATTTTCGAGATGGCGAAGAAGTTGCAACAGGATA	8.0	0.4
487	TGGGCATCGTGGCCACCACTGGTGTGGGCACCGAGATTGGGAAGATCCGAGA	29.3	1.5
540	TGGTGCAAACTACAGATGGTACACCTACATCTGTGCAGGAAGTGGCTCCCCA	25.1	0.5
673	TTGGTTGGGACACTGATATTTCCTGGCTTACTGGAGAAGAATTGCATGTGGA	6.8	0.1
1272	TTGAAGAGCAGCCAATCAATACCATTTATCCAGAGGAATCACTGGAAGGAAA	45.3	0.8
1287	TCCTCCTGGCCCTCCTGGATTTCCTGGAGAAAGGGGTCAGAAAGGTGATGAA	10.3	0.7
1380	TGTTGGTACCGTGATAAGGTACAGTTGTTCAGGTACCTTCCGCCTCATTGGA	39.9	1.2
1410	TGGAGAAGGACAGGTTCTCTGTCAACCTGGATGTGAAGCACTTCTCCCCAGA	32.4	1.0
1497	TCCTGGCCTGCCATCGGCTTCCTGGTGCTCGCGTGGCTCTTCGCATTTGTCA	21.3	1.0
1636	TATGACCGGACATCCCAGGTGGTGTGGAACGAGTATGCCGAGGCCAACTGGA	15.5	1.8
1730	TTACAAGTTTTAGAAAATCTACTGTCAAAAAAGAAAAACCTCTTATTCAACA	13.6	0.1
1821	TTGTCAGCTCAGCTGCCCCTACAGGGGGATGTGGCCCTGGTGCAACAGGAGA	52.1	1.8
1910	TTACAAGACAGCAAAAGATTGGTGGCTATTCAGTTTCTATTTCTGCTTGCCA	29.2	0.5
2070	TTCTTTCTACACCAGCAGCTCAAACAATGTCTGCCTATGCAGGCCAGACTCA	34.7	1.3
2162	TATGAAAACAACAAGCTGATTGTCCGCAGAGGGCAGTCTTTCTATGTGCAGA	17.6	0.7
2218	TGGAGCTGGTTTGTCAAAATCCAAAGGAAGCCGAATTGGATTTGATAGCACA	30.8	0.6
2259	TACCAGTGGGACTACGTGTTGTTGCCATCCAGGGAGTGAAAACAGGGTTGTA	30.4	0.9
2261	TGGACACGGCCCGAGCGGATGGACAAGAAGCTGCTGGCCGTGCCGGCCGCCA	36.4	0.8
25974	TGTAGCTGGGGCTGCTTACTACTACCAACGACAAGATGTGGAGGCTGACCCA	48.6	1.5
25984	TGTTATCCGAGACCCAGTCTCGGTACTCCTGCAAGCTCCAGGACATGCAAGA	22.2	0.9
27130	TGCATGCTGCTGCTCTTTCTGGCCATGTCAGCACCGTGAAGTTATTACTGGA	5.1	0.4
27241	TATTTTGAAAAACAGGGAGTCAAAGATTTTGCATGTTCTTTTTCGGGATCTA	14.2	2.0
27286	TGGTGTTATACATTAAATATCCAGGATGGAGAAGCCACATGCTACTCACCGA	33.6	0.7
51314	TTTATTGAGAGACAATGGCTTGCAATACTGGAAACAATTACTGGGACCAAGA	40.9	1.3
51733	TGGAATACAGCCGTGGTGATCTCCAATTCCGGAGCAGTCCTGGGAAAGACCA	28.9	0.5
55084	TTCTGCAGCGAGAAGTGCTTTGCGGCCTGCCGACGAGCCTACTTCAAGAGAA	20.1	0.2
55118	TGGCTGACTTCAACCGTGATGGCAAAGTGGACATCGTCTATGGCAACTGGAA	37.5	2.5
55198	TTGGCAGACACAATGGTTCTACCTATCATACAATTCCGAGAAAAGGATCTCA	32.4	0.7
55285	TGTCTAAGAAAGAGAGCTTTGCTCCTGGTACTATGTACAAGCCCTTTGGGAA	9.2	1.1
55607	TGCAAATCAAACATGCAGTTACAGAAGCAGAGATTCAAAAATTGAAGACCAA	21.2	0.7
56246	TGGACTATCTGGACCTCATTCCCGTGGACGAGAAGAAGCTGAAAGCCCACAA	27.9	0.5
56649	TGGCCTTGGCAGGTCAGCATCCAGTACGACAAACAGCACGTCTGTGGAGGGA	0.3	0.1
57623	TCGGAGAGGCTACCAGGAATACGCCATTCAGCAGACACCTTATGAGCAACCA	32.2	1.9
59272	TGCCTTTTTAAAGGAACAGTCCACACTTGCCCAAATGTATCCACTACAAGAA	20.6	0.5
64093	TGGGAAGCCCATCAGTGGCTCTTCTGTGCAGAATAAAACTCCTGTATGTTCA	7.1	0.3

Gene ID	sequence	Avg.	S.E.M
64102	TATGACATGGAGCACACTTTCTACAGCAATGGAGAGAAGAAGAAGATTTACA	24.7	0.7
80263	TCACCCTGCTTTGTAAGGATGCCGCAGGAGAAATCATGGGCAGGGGAGGAGA	17.3	0.9
80704	TATGTCCGCTACAAGCCAGTCATCATCTTGCAAGGTATCAGTTTCATCATTA	11.7	0.7
85320	TACTCTCCAAGATGGCCCCTGGAGTCAGCAAGAGAGAAATCCTGAGGCTCCA	40.8	1.4
91147	TCAAGTGCTCAGTTCTACCTCTCTTCCTACAAATTTCAGTTTTAAAGGAGAA	16.2	0.6
121340	TTTGGTGGCTCTAGCCCTCTGCGGGACTCAACAACTCTGGGCAAAGCAGGCA	37.2	1.0
123263	TGGAGGTGGTCACAATGCCTTCCCCATCACCAAAAGGACTGCCAGTGAAGCA	6.6	0.2
133396	TGGCTACAACATATGGTACTATCCAGAAAGCAACACTAACCTCACAGAAACA	35.1	1.2
135935	TCTCCTTTTGACACTAACATCACCAGACCTGGAGGGTACCTGGGACACCAGA	40.7	1.3
137682	TTGATATATTCCACCCATCTTTACAGCAGAAGAATACATTACTTCCATTATA	3.9	0.4
139105	TGTCGGCTAGATACCTTATTCAGAAACTCTTCACAAAAGATGTCCTGGTCCA	38.7	1.6
139135	TTGCAGAGGTTGAGCAGTATGGACCACAAGAAAACGTTCACATGTTTGTAGA	28.3	1.9
145173	TCATCCAGAGTCAAAGTAATTCTTTTCATGCAAAGAGAGCAGAGCAGTTAAA	32.8	1.4
146956	TACACACAGAAGCCCAGGCTCAAATTGTGCAGAGCTGGAAAGAGCTGGCCGA	30.4	1.8
147372	TGGCCGAGTGCTGTGTACTTGTTATCCGGGATACCGATATGACCGGGAGAGA	31.1	0.5
158747	TACGTTAAAAGACAATGCTTTCAATATGTCAGATAAAACCAGTGAAGATATA	16.8	0.4
166012	TGGGCTCCAGCTGGCTTGGTGGGGAGAAGAGAAGCCCCCTGCAGAAGCTCTA	19.6	0.3
33	TCTTTTCTCCAGAGCATGACATTTTCCGGAAAAGTGTAAGGAAGTTTTTCCA	1.6	0.1
183	TCGCCCAGGCTGAGCTGCCCGCCATTCTGCACACCGAGCTGAACCTGCAAAA	26.5	0.6
240	TGGAGAACCTGTTCATCAACCGCTTCATGCACATGTTCCAGTCTTCTTGGAA	19.5	0.5
287	TTTACTCCTCTAGCTGTGGCACTCCAGCAAGGACACAACCAGGCGGTGGCCA	14.6	0.3
421	TGGAGGACTGCAATGTGCACTCGGCCGCCAGCATCCTGGCCTCGGTGAAGGA	21.1	0.6
492	TGGCGGACAACACCAATGACCTGGAGAAGCGCAGGCAGATCTACGGGCAGAA	25.6	1.1
649	TTCCAGGAAACACTTCTACCCCCAGCTGCCAGAGCACCAACGGGCAGCCTCA	42.8	0.3
725	TCACGTTTATGTGCAATGACCACTACATCCTCAAGGGCAGCAATCGGAGCCA	12.5	0.6
889	TGTACTAATGAAAAAATTTCCTCTGGATGGAGAGAAGATGGGCAGAGAAGCA	16.8	0.5
1009	TTGATGACAAATCAGGGAACATTCATGCCACCAAGACGTTGGATCGAGAAGA	39.2	1.4
1056	TCTTCAAGGGCATCCCCTTCGCAGCTCCCACCAAGGCCCTGGAAAATCCTCA	16.9	0.4
1260	TTCGCCTGGTGGGGATCATCAGAGAATGGGCCAACAAGAATTTCCGAGAGGA	39.6	1.0
1297	TGTCCCAATGCCTGTCCACCAGGTCGCTCAGGATATCCAGGCCTACCAGGCA	44.8	0.5
1690	TTGCTATCACATGTTTTACCAGAGGCTTGGACATCAGGAAAGAGAAAGCAGA	45.5	1.0
1824	TTTTGGAGGATGGTTCAGTCTATACAACAAATACTATTCTATTGTCCTCGGA	11.8	0.4
2068	TTTATAGCTACCACTACCTCCTGGACCCCAAGATTGCAGACCTGGTGTCCAA	29.0	0.6
2670	TGGAGGAAGAGGGGCAGAGCCTCAAGGACGAGATGGCCCGCCACTTGCAGGA	36.0	0.8
2694	TTCCCTCAGCACAGGAGCCCTTGGTCAATGGAATACAAGTACTCATGGAGAA	38.9	0.7
2892	TTGCCGTGCAGTTATACAACACCAACCAGAACACCACCGAGAAGCCCTTCCA	26.4	0.7

2. List of TALEN activity in the T7E1 assay

	Gene name	TALEN binding site	Number of CpG sites			Mutation frequency (%)
			Left site	Right site	Total	
Genome-scale study	*AGXT*	TCTCTGGCTCGGGACACTGTgccctggaggccGCCCTGGTCAATGTGCTGGA	1	1	2	2
	ALOX15B	TGCCTGGATGAAAAGACAGTggaagacttggaGCTCAATATCAAATACTCCA	0	0	0	14
	APOB	TGCTCCACTCACTTTACCGTcaagacgaggaaGGGCAATGTGGCAACAGAAA	1	0	1	8
	AQP9	TTGCCCAAGCTATTCTCAGTcgaggacgttttGGAGGGGTCATCACTATCAA	0	0	0	34
	ARG1	TTTGGCAATTGGAAGCATCTctggccatgccaGGGTCCACCCTGATCTTGGA	0	0	0	35
	ARHGAP4	TTGCACTGCTGGGCGGTGCTgctgcagcacaGCCGCAGCAGAGCCGGGAGA	1	3	4	1.5
	ARSB	TCACACGATGTGCTCTTGATtttcgagatggccGAAGAAGTTGCAACAGGATA	1	1	2	1.6
	ATP2A1	TGGGCATCGTGGCCACCACTggtgtgggcaccGAGATTGGGAAGATCCGAGA	1	2	3	4.4
	ATP7B	TGGTGCAAACTACAGATGGTacacctacatctGTGCAGGAAGTGGCTCCCCA	0	0	0	2.2
	OPN1SW	TGGGGCCGTGGGATGGGCCTcagtaccacattGCCCCTGTCTGGGCCTTCTA	1	0	1	46
	BRAF	TTGGTTGGGACACTGATATTtcctggcttactGGAGAAGAATTGCATGTGGA	0	0	0	16
	CLCN4	TTGGGAAATAGCCGTCTCGTTctctttttatgtgGAATACCACACGCCCTGGTA	2	1	3	7
	CNTN1	TTGAAGAGCAGCCAATCAATaccatttatccaGAGGAATCACTGGAAGGAAA	0	0	0	34
	COL4A5	TCCTCCTGGCCCTCCTGGATttcctggagaaaGGGGTCAGAAAGGTGATGAA	0	0	0	3.5
	CR2	TGTTGGTACCGTGATAAGGTacagttgttcagGTACCTTCCGGCTCATTGGA	1	1	2	8.5
	CRYAB	TGGAGAAGGACAGGTTCTCTgtcaacctggatGTGAAGCACTTCTCCCCAGA	0	0	0	3.1
	CTNS	TCCTGGCCTGCCATCGGCTTcctggtgctcgcGTGGCTCTTCGCATTTGTCA	1	2	3	1
	CYP1A2	TTGTCCCAGTCTGTTCCCTTctcggccacagaGCTTCTCCTGGCCTCTGCCA	0	0	0	20
	CYP27B1	TGCGCACCGTGTACGTGGCTgcccctgcactcGTCGAGGAGCTGCTGCGACA	3	3	6	<0.5
	ACE	TATGACCGGACATCCCAGGTggtgtggaacgaGTATGCCGAGGCCAACTGGA	1	1	2	2.2
	DIAPH2	TTACAAGTTTTAGAAAATCTactgtcaaaaaaGAAAAACCTCTTATTCAACA	0	0	0	3.6
	SLC26A3	TGCTCAGTGATATTGTTTCTggtatcagcacaGGGATTGTGGCCGTACTACA	0	1	1	9
	DRP2	TTGTCAGCTCAGCTGCCCCTacaggggatgtGGCCCTGGTGCAACAGGAGA	0	0	0	16
	EDNRB	TTACAAGACAGCAAAAGATTggtggctattcaGTTTCTATTTCTGCTTGCCA	0	0	0	14
	EYA4	TTCTTTCTACACCAGCAGCTcaaacaatgtctGCCTATGCAGGCCAGACTCA	0	0	0	26
	F13A1	TATGAAAACAACAAGCTGATtgtccgcagaggGCAGTCTTTCTATGTGCAGA	0	0	0	17
	FAAH	TCAGAGGCGCTGCTAGCCCTgccccctgcctcaGCTGGTGCAGAAGTTACACA	1	0	1	33
	FKTN	TGGAGCTGGTTTGTCAAAATccaaaggaagccGAATTGGATTTGATAGCACA	0	1	1	54
	GPC4	TGGTCAGCGAACAGTGCAATcatttgcaagctGTCTTTGCTTCACGTTACAA	1	1	2	13
	FGF14	TACCAGTGGGACTACGTGTTgttgccatccagGGAGTGAAAACAGGGTTGTA	1	0	1	34
	FGFR3	TGGACACGCCCCGAGCGGATggacaagaagctGCTGGCCGTGCGGCCGCCA	3	3	6	<0.5

	Gene name	TALEN binding site	Number of CpG sites			Mutation frequency (%)
			Left site	Right site	Total	
Genome-scale study	*MMACHC*	TGTAGCTGGGGCTGCTTACTactaccaacgacAAGATGTGGAGGCTGACCCA	0	0	0	31
	INVS	TGCATGCTGCTGCTCTTTCTggccatgtcagcACCGTGAAGTTATTACTGGA	0	1	1	3.7
	SRPX2	TGGTGTTATACATTAAATATccaggatggagaAGCCACATGCTACTCACCGA	0	1	1	19
	TXNDC3	TTTATTGAGAGACAATGGCTtgcaatactggaAACAATTACTGGGACCAAGA	0	0	0	14
	UPB1	TGGAATACAGCCGTGGTGATctccaattccggAGCAGTCCTGGGAAAGACCA	1	0	1	4
	SOBP	TTCTGCAGCGAGAAGTGCTTtgcggcctgccgACGAGCCTACTTCAAGAGAA	1	1	2	7.7
	CRTAC1	TGGCTGACTTCAACCGTGATggcaaagtggacATCGTCTATGGCAACTGGAA	1	1	2	18
	APPL2	TTGGCAGACACAATGGTTCTacctatcatacaATTCCGAGAAAAGGATCTCA	0	1	1	40
	RBM41	TGTCTAAGAAAGAGAGCTTTgctcctggtactATGTACAAGCCCTTTGGGAA	0	0	0	7.2
	PPP1R9A	TGCAAATCAAACATGCAGTTacagaagcagagATTCAAAAATTGAAGACCAA	0	0	0	29
	NXF3	TTAGCAGTAGGTCTGAACCTgtcaatcctggcATGCATTCTTCATCCCATCA	0	0	0	18
	MRAP	TGGACTATCTGGACCTCATTcccgtggacgagAAGAAGCTGAAAGCCCACAA	0	0	0	23
	TMPRSS4	TGGCCTTGGCAGGTCAGCATccagtacgacaaACAGCACGTCTGTGGAGGGA	0	1	1	22
	ZFAT	TCGGAGAGGCTACCAGGAATacgccattcagcAGACACCTTATGAGCAACCA	1	0	1	7.6
	ACE2	TGCCTTTTTAAAGGAACAGTccacacttgcccAAATGTATCCACTACAAGAA	0	0	0	10
	SMOC1	TGGGAAGCCCATCAGTGGCTcttctgtgcagaATAAAACTCCTGTATGTTCA	0	0	0	4.1
	TNMD	TATGACATGGAGCACACTTTctacagcaatggAGAGAAGAAGAAGATTTACA	0	0	0	15
	SLC19A3	TATGTCCGCTACAAGCCAGTcatcatcttgcaAGGTATCAGTTTCATCATTA	1	0	1	33
	PPP1R1B	TCAGGAGAGGGGCACCATCTcaagtcgaagagACCCAACCCCTGTGCCTACA	0	0	0	33
	ABCC11	TACTCTCCAAGATGGCCCCTggagtcagcaagAGAGAAATCCTGAGGCTCCA	0	0	0	39
	TMEM67	TCAAGTGCTCAGTTCTACCTctcttcctacaaATTTCAGTTTTAAAGGAGAA	0	0	0	26
	SP7	TTTGGTGGCTCTAGCCCTCTgcgggactcaacAACTCTGGGCAAAGCAGGCA	0	0	0	30
	EVC2	TGCTCTTGCTGGACAGCATTgctggtctcaccATTTGGGACTCTGTGGGAAA	0	0	0	39
	IL31RA	TGGCTACAACATATGGTACTatccagaaagcaACACTAACCTCACAGAAACA	0	0	0	9.3
	NOBOX	TCTCCTTTTGACACTAACATcaccagacctggAGGGTACCTGGGACACCAGA	0	0	0	23
	PASD1	TTGCAGAGGTTGAGCAGTATggaccacaagaaAACGTTCACATGTTTGTAGA	0	1	1	6.9
	B3GALTL	TCATCCAGAGTCAAAGTAATtcttttcatgcaAAGAGAGCAGAGCAGTTAAA	0	0	0	32
	CCBE1	TGGCCGAGTGCTGTGTACTTgttatccgggatACCGATATGACCGGGAGAGA	1	2	3	2.6
	CHST13	TGGGCTCCAGCTGGCTTGGTggggagaagagaAGCCCCCTGCAGAAGCTCTA	0	0	0	11
	ACADL	TCTTTTCTCCAGAGCATGACatttttccggaaaAGTGTAAGGAAGTTTTTCCA	0	0	0	0.6
	ACHE	TTTCTGGTTTACGGGGCCCCaggcttcagcaaAGACAACGAGTCTCTCATCA	1	1	2	10
	ADD2	TGGAGAAGGGCAGCAGCTGCttcccagtggacACCACAGGCTTCTGTCTGCA	0	0	0	22
	ALOX5	TGGAGAACCTGTTCATCAACcgcttcatgcacATGTTCCAGTCTTCTTGGAA	0	0	0	33
	ANK2	TTTACTCCTCTAGCTGTGGCactccagcaaggACACAACCAGGCGGTGGCCA	0	1	1	7.3
	ARVCF	TGGAGGACTGCAATGTGCACtcggccgccagcATCCTGGCCTCGGTGAAGGA	0	1	1	6.3
	BMP1	TTCCAGGAAACACTTCTACCccagctgccagAGCACCAACGGGCAGCCTCA	0	1	1	25
	BMP2	TTGGAAGAACTACCAGAAACgagtgggaaaacAACCCGGAGATTCTTCTTTA	1	1	2	1.4
	C4BPB	TCACGTTTATGTGCAATGACcactacatcctcAAGGGCAGCAATCGGAGCCA	1	1	2	9.1
	KRIT1	TGTACTAATGAAAAAATTTCctctggatggagAGAAGATGGGCAGAGAAGCA	0	0	0	2.4
	CD40	TCTGCGAGCCCTGCCCAGTCggcttcttctccAATGTGTCATCTGCTTTCGA	2	1	3	4.8
	CDH11	TTGATGACAAATCAGGGAACattcatgccaccAAGACGTTGGATCGAGAAGA	0	2	2	17
	CEL	TCTTCAAGGGCATCCCCTTCgcagctcccaccAAGGCCCTGGAAAATCCTCA	1	0	1	3.2
	CHAT	TGCCCAAACTGCCCGTGCCCccgctgcagcagACCCTGGCCACGTACCTGCA	1	1	2	16
	CHI3L1	TTGACCGCTTCCTCTGTACCcacatcatctacAGCTTTGCCAATATAAGCAA	1	0	1	30
	CHN2	TGGATTGAACGTACACAAACagtgttccaagcACGTTCCCAATGACTGCCAA	1	1	2	4.4
	CNGA2	TTCGCCTGGTGGGGATCATCagagaatgggccAACAAGAATTTCCGAGAGGA	1	1	2	6.2
	COL9A1	TGTCCCAATGCCTGTCCACCaggtcgctcaggATATCCAGGCCTACCAGGCA	0	0	0	12
	COCH	TTGCTATCACATGTTTTACCagaggcttggacATCAGGAAAGAGAAAGCAGA	0	0	0	14
	DNM1	TGGCATTGTCACCCGACGTCccctggtcttgcAGCTGGTCAATGCAACCACA	2	0	2	5.5

	Gene name	TALEN binding site	Number of CpG sites			Mutation frequency (%)
			Left site	Right site	Total	
Genome-scale study	*DSC2*	TTTTGGAGGATGGTTCAGTCtatacaacaaatACTATTCTATTGTCCTCGGA	0	1	1	3.3
	EGF	TTTGCTGTTTGCCAATTCTCaagatattcgacACATGCATTTTGATGGAACA	0	0	0	29
	ERCC2	TTTATAGCTACCACTACCTCctggaccccaagATTGCAGACCTGGTGTCCAA	0	0	0	10
	EYA1	TATGGGCAAACACAGTTTACcacaggaatgcaACAAGCTACAGCCTATGCCA	0	0	0	4.9
	FUT2	TTCCCATGGCCCACTTCATCctctttgtctttACGGTTTCCACTATATTTCA	0	1	1	17
	GIF	TTCCCTCAGCACAGGAGCCCttggtcaatggaATACAAGTACTCATGGAGAA	0	0	0	17
	GRPR	TTGTGGGCCTGTTCGCCTTCtgctggctccccAATCATGTCATCTACCTGTA	1	0	1	3
	TRPC6	TCCATTTGGTATGAGAATCTttctggtttacgACAGCAGACAATGGCGGTCA	0	1	1	5.7
	TYRP1	TTTGTCCGGGCCCTGGATATggcaaagcgcacAACTCACCCTTTATTTGTCA	1	0	1	27.7
	USH2A	TCTGGAGATCTTCTCAGATTgcatgcccaatcACATTGCCGTTGCCCTGGCA	0	1	1	8.3
	XRCC2	TTATCACCTAACAGCACGATgtatacttcccaAATCAGAAGGTGGCCTGGAA	1	0	1	10.6
	FOXN1	TGCCGGCTTCAGCTGCTCGTcatttgtgtccgACGGCCCTCCAGAGAGGACA	2	1	3	15.2
	PRSS12	TTTGAAGGCACAGTGGAAGTatatgcaagtggAGTTTGGGGCACTGTCTGTA	0	0	0	21.6
	TNFSF11	TTAATCAGGATGGCTTTTATacctgtatgccAACATTTGCTTTCGACATCA	0	1	1	27.2
	CHRD	TGGGCAGGCAGGCCAGGGCTgcgcatcagtggACACATTGCTGCCAGGAAGA	0	0	0	35.9
	IL18RAP	TTTACCAGAGCCACAGAAATcacatttctgccACAGAAATCGACTCTCACCA	0	1	1	25.6
	WISP3	TTGTCACTGGCCCTGCAAATgccctcagcagaAGCCCCGTTGCCCTCCTGGA	0	1	1	21.3
	ALDH1A2	TGCAAGCTTTTTATGTGGATttgcagggcgtcATCAAAACCTTTCGATATTA	0	1	1	10.2
	SGCE	TGTCCTCTTTGTTCATGTTTtggaaagagaatATTTTAAGGGGGAATTTCCA	0	0	0	14
	PAPSS2	TTGCCCTGGAGGAGTACCTTgtctcccatgccATCCCTTGTTACTCCCTGGA	0	0	0	16
	SLC6A5	TTCTACCTGTTTGCCTCCTTtgtgtctgtactACCCTGGGGCTCCTGCAACA	0	0	0	7.8
	REPS2	TACATTGCCCTGAAATTAATtgctgcagcacaATCTGGCCTCCCGGTACGGA	0	2	2	25.6
	KL	TCAAGCTGGATGGGGTGGATgtcatcgggtatACCGCATGGTCCCTCATGGA	0	1	1	8.8
Backup TALENs	*CYP27B1*	TGTGCCTGGGCCCTGGGGCCgcctctgccgagACTGGGACCAGATGTTTGCA	1	0	1	6.6
	FGFR3	TGGCCGTGGCCATCGTGGCCggcgcctcctcgGAGTCCTTGGGGACGGAGCA	3	1	4	28
NF-kB	*SMEK1*	TTAGCTTTCAAGAAAAAGCTggatgtgatgaaATTTGGGAGAAAATATGTCA	0	0	0	20
	TNFR1	TGGTGGGAATATACCCCTCAggggttattggaCTGGTCCCTCACCTAGGGGA	0	0	0	22
	IL1R1	TTCGTCCCTGTCCTCTTAACccaaatgaacacAAAGGCACTATAACTTGGTA	1	0	1	18
	TNFR2	TGGCATTTACACCCTACGCCccggagcccgggAGCACATGCCGGCTCAGAGA	1	1	2	28
	p50	TCATACAATATTTAATCCAGaagtatttcaacCACAGATGGCACTGCCAACA	0	0	0	26
	IKKA	TCTCAAAATAGCAATTAAGTcttgtcgcctagAGCTAAGTACCAAAAACAGA	0	0	0	25

VI. References

Adamson B, Smogorzewska A, Sigoillot FD, King RW, Elledge SJ. 2012. A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. *Nature cell biology* **14**(3): 318-328.

Bedell VM, Wang Y, Campbell JM, Poshusta TL, Starker CG, Krug RG, 2nd, Tan W, Penheiter SG, Ma AC, Leung AY et al. 2012. In vivo genome editing using a high-efficiency TALEN system. *Nature* **491**(7422): 114-118.

Bibikova M, Beumer K, Trautman JK, Carroll D. 2003. Enhancing gene targeting with designed zinc finger nucleases. *Science* **300**(5620): 764.

Birmingham A, Anderson EM, Reynolds A, Ilsley-Tyree D, Leake D, Fedorov Y, Baskerville S, Maksimova E, Robinson K, Karpilow J et al. 2006. 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature methods* **3**(3): 199-204.

Birney E, Ensembl T. 2003. Ensembl: a genome infrastructure. *Cold Spring Harbor symposia on quantitative biology* **68**: 213-215.

Boch J, Scholze H, Schornack S, Landgraf A, Hahn S, Kay S, Lahaye T, Nickstadt A, Bonas U. 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**(5959): 1509-1512.

Briggs AW, Rios X, Chari R, Yang L, Zhang F, Mali P, Church GM. 2012. Iterative capped assembly: rapid and scalable synthesis of

repeat-module DNA such as TAL effectors from individual monomers. *Nucleic acids research* **40**(15): e117.

Brose MS, Volpe P, Feldman M, Kumar M, Rishi I, Gerrero R, Einhorn E, Herlyn M, Minna J, Nicholson A et al. 2002. BRAF and RAS mutations in human lung cancer and melanoma. *Cancer research* **62**(23): 6997-7000.

Bultmann S, Morbitzer R, Schmidt CS, Thanisch K, Spada F, Elsaesser J, Lahaye T, Leonhardt H. 2012. Targeted transcriptional activation of silent oct4 pluripotency gene by combining designer TALEs and inhibition of epigenetic modifiers. *Nucleic acids research* **40**(12): 5368-5377.

Bylund L, Kytola S, Lui WO, Larsson C, Weber G. 2004. Analysis of the cytogenetic stability of the human embryonal kidney cell line 293 by cytogenetic and STR profiling approaches. *Cytogenetic and genome research* **106**(1): 28-32.

Cade L, Reyon D, Hwang WY, Tsai SQ, Patel S, Khayter C, Joung JK, Sander JD, Peterson RT, Yeh JR. 2012. Highly efficient generation of heritable zebrafish gene mutations using homo- and heterodimeric TALENs. *Nucleic acids research* **40**(16): 8001-8010.

Carlson DF, Tan W, Lillico SG, Stverakova D, Proudfoot C, Christian M, Voytas DF, Long CR, Whitelaw CB, Fahrenkrug SC. 2012. Efficient TALEN-mediated gene knockout in livestock. *Proceedings of the National Academy of Sciences of the United States of America* **109**(43): 17382-17387.

Cermak T, Doyle EL, Christian M, Wang L, Zhang Y, Schmidt C, Baller JA, Somia NV, Bogdanove AJ, Voytas DF. 2011. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic acids research* **39**(12): e82.

Cho SW, Kim S, Kim JM, Kim JS. 2013. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature biotechnology* **31**(3): 230-232.

Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ, Voytas DF. 2010. Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**(2): 757-761.

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA et al. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**(6121): 819-823.

Cornu TI, Thibodeau-Beganny S, Guhl E, Alwin S, Eichtinger M, Joung JK, Cathomen T. 2008. DNA-binding specificity is a major determinant of the activity and toxicity of zinc-finger nucleases. *Molecular therapy : the journal of the American Society of Gene Therapy* **16**(2): 352-358.

Deng C, Capecchi MR. 1992. Reexamination of gene targeting frequency as a function of the extent of homology between the targeting vector and the target locus. *Molecular and cellular biology* **12**(8): 3365-3371.

Deng D, Yin P, Yan C, Pan X, Gong X, Qi S, Xie T, Mahfouz M,

Zhu JK, Yan N et al. 2012. Recognition of methylated DNA by TAL effectors. *Cell research* **22**(10): 1502-1504.

Doyon Y, McCammon JM, Miller JC, Faraji F, Ngo C, Katibah GE, Amora R, Hocking TD, Zhang L, Rebar EJ et al. 2008. Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nature biotechnology* **26**(6): 702-708.

Duan H, Heckman CA, Boxer LM. 2005. Histone deacetylase inhibitors down-regulate bcl-2 expression and induce apoptosis in t(14;18) lymphomas. *Molecular and cellular biology* **25**(5): 1608-1619.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**(6669): 806-811.

Gabriel R, Lombardo A, Arens A, Miller JC, Genovese P, Kaeppel C, Nowrouzi A, Bartholomae CC, Wang J, Friedman G et al. 2011. An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nature biotechnology* **29**(9): 816-823.

Gaeta ML, Johnson DR, Kluger MS, Pober JS. 2000. The death domain of tumor necrosis factor receptor 1 is necessary but not sufficient for Golgi retention of the receptor and mediates receptor desensitization. Laboratory investigation; a journal of technical methods and pathology 80(8): 1185-1194.

Gewurz BE, Towfic F, Mar JC, Shinners NP, Takasaki K, Zhao B, Cahir-McFarland ED, Quackenbush J, Xavier RJ, Kieff E. 2012. Genome-wide siRNA screen for mediators of NF-kappaB activation. *Proceedings of the National Academy of Sciences of*

the United States of America **109**(7): 2467-2472.

Guo J, Gaj T, Barbas CF, 3rd. 2010. Directed evolution of an enhanced and highly efficient FokI cleavage domain for zinc finger nucleases. *Journal of molecular biology* **400**(1): 96-107.

Hockemeyer D, Wang H, Kiani S, Lai CS, Gao Q, Cassady JP, Cost GJ, Zhang L, Santiago Y, Miller JC et al. 2011. Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology* **29**(8): 731-734.

Holt N, Wang J, Kim K, Friedman G, Wang X, Taupin V, Crooks GM, Kohn DB, Gregory PD, Holmes MC, Cannon PM. 2010. Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo. *Nature biotechnology* **28**(8): 839-847.

Huang P, Xiao A, Zhou M, Zhu Z, Lin S, Zhang B. 2011. Heritable gene targeting in zebrafish using customized TALENs. *Nature biotechnology* **29**(8): 699-700.

Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh JR, Joung JK. 2013. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature biotechnology* **31**(3): 227-229.

Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, Li B, Cavet G, Linsley PS. 2003. Expression profiling reveals off-target gene regulation by RNAi. *Nature biotechnology* **21**(6): 635-637.

Khan AA, Betel D, Miller ML, Sander C, Leslie CS, Marks DS. 2009.

Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. *Nature biotechnology* **27**(6): 549-555.

Kim H, Kim MS, Wee G, Lee CI, Kim H, Kim JS. 2013a. Magnetic separation and antibiotics selection enable enrichment of cells with ZFN/TALEN-induced mutations. *PloS one* **8**(2): e56476.

Kim H, Um E, Cho SR, Jung C, Kim H, Kim JS. 2011. Surrogate reporters for enrichment of cells with nuclease-induced mutations. *Nature methods* **8**(11): 941-943.

Kim, H.J. 2012. Targeted mutagenesis of the human *chemokine (C-C motif) receptor 5* gene using zinc-finger nucleases and TAL effector nucleases. *Ph.D. Dissertation*, Seoul National University, Seoul, Korea.

Kim HJ, Lee HJ, Kim H, Cho SW, Kim JS. 2009. Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. *Genome research* **19**(7): 1279-1288.

Kim, S., Lee, M.J., Kim, H., Kang, M. and Kim, J.S. 2011. Preassembled zinc-finger arrays for rapid construction of ZFNs. *Nat Methods,* 8: 7.

Kim Y, Kweon J, Kim A, Chon JK, Yoo JY, Kim HJ, Kim S, Lee C, Jeong E, Chung E et al. 2013b. A library of TAL effector nucleases spanning the human genome. *Nature biotechnology* **31**(3): 251-258.

Kim Y, Kweon J, Kim JS. 2013c. TALENs and ZFNs are associated with different mutation signatures. *Nature methods* **10**(3): 185.

Kim YG, Cha J, Chandrasegaran S. 1996. Hybrid restriction enzymes:

zinc finger fusions to Fok I cleavage domain. *Proceedings of the National Academy of Sciences of the United States of America* **93**(3): 1156-1160.

Krueger U, Bergauer T, Kaufmann B, Wolter I, Pilk S, Heider-Fabian M, Kirch S, Artz-Oppitz C, Isselhorst M, Konrad J. 2007. Insights into effective RNAi gained from large-scale siRNA validation screening. *Oligonucleotides* **17**(2): 237-250.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**(3): R25.

Lee HJ, Kim E, Kim JS. 2010. Targeted chromosomal deletions in human cells using zinc finger nucleases. *Genome research* **20**(1): 81-89.

Lee HJ, Kweon J, Kim E, Kim S, Kim JS. 2012. Targeted chromosomal duplications and inversions in the human genome using zinc finger nucleases. *Genome research* **22**(3): 539-548.

Lei Y, Guo X, Liu Y, Cao Y, Deng Y, Chen X, Cheng CH, Dawid IB, Chen Y, Zhao H. 2012. Efficient targeted gene disruption in Xenopus embryos using engineered transcription activator-like effector nucleases (TALENs). *Proceedings of the National Academy of Sciences of the United States of America* **109**(43):

17484-17489.

Li T, Huang S, Zhao X, Wright DA, Carpenter S, Spalding MH, Weeks DP, Yang B. 2011. Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic acids research* **39**(14): 6315-6325.

Li T, Liu B, Spalding MH, Weeks DP, Yang B. 2012. High-efficiency TALEN-based gene editing produces disease-resistant rice. *Nature biotechnology* **30**(5): 390-392.

Lin X, Ruan X, Anderson MG, McDowell JA, Kroeger PE, Fesik SW, Shen Y. 2005. siRNA-mediated off-target gene silencing triggered by a 7 nt complementation. *Nucleic acids research* **33**(14): 4527-4535.

Liu J, Li C, Yu Z, Huang P, Wu H, Wei C, Zhu N, Shen Y, Chen Y, Zhang B et al. 2012. Efficient and specific modifications of the Drosophila genome by means of an easy TALEN strategy. *Journal of genetics and genomics = Yi chuan xue bao* **39**(5): 209-215.

Ma S, Zhang S, Wang F, Liu Y, Liu Y, Xu H, Liu C, Lin Y, Zhao P, Xia Q. 2012. Highly efficient and specific genome editing in silkworm using custom TALENs. *PloS one* **7**(9): e45035.

Macville M, Schrock E, Padilla-Nash H, Keck C, Ghadimi BM, Zimonjic D, Popescu N, Ried T. 1999. Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. *Cancer research* **59**(1): 141-150.

Maeder ML, Angstman JF, Richardson ME, Linder SJ, Cascio VM, Tsai SQ, Ho QH, Sander JD, Reyon D, Bernstein BE et al. 2013a. Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nature biotechnology*.

Maeder ML, Linder SJ, Reyon D, Angstman JF, Fu Y, Sander JD, Joung JK. 2013b. Robust, synergistic regulation of human gene expression using TALE activators. *Nature methods* **10**(3): 243-245.

Maeder ML, Thibodeau-Beganny S, Osiak A, Wright DA, Anthony RM, Eichtinger M, Jiang T, Foley JE, Winfrey RJ, Townsend JA et al. 2008. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Molecular cell* **31**(2): 294-301.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9. *Science* **339**(6121): 823-826.

Maresca M, Lin VG, Guo N, Yang Y. 2013. Obligate ligation-gated recombination (ObLiGaRe): custom-designed nuclease-mediated targeted integration through nonhomologous end joining. *Genome research* **23**(3): 539-546.

Mendenhall EM, Williamson KE, Reyon D, Zou JY, Ram O, Joung JK, Bernstein BE. 2013. Locus-specific editing of histone modifications at endogenous enhancers. *Nature biotechnology*.

Mercer AC, Gaj T, Fuller RP, Barbas CF, 3rd. 2012. Chimeric TALE

recombinases with programmable DNA sequence specificity. *Nucleic acids research* **40**(21): 11163-11172.

Miller JC, Tan S, Qiao G, Barlow KA, Wang J, Xia DF, Meng X, Paschon DE, Leung E, Hinkley SJ et al. 2011. A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **29**(2): 143-148.

Miyanari Y, Ziegler-Birling C, Torres-Padilla ME. 2013. Live visualization of chromatin dynamics with fluorescent TALEs. *Nature structural & molecular biology* **20**(11): 1321-1324.

Morbitzer R, Elsaesser J, Hausner J, Lahaye T. 2011. Assembly of custom TALE-type DNA binding domains by modular cloning. *Nucleic acids research* **39**(13): 5790-5799.

Moscou MJ, Bogdanove AJ. 2009. A simple cipher governs DNA recognition by TAL effectors. *Science* **326**(5959): 1501.

Mussolino C, Morbitzer R, Lutge F, Dannemann N, Lahaye T, Cathomen T. 2011. A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic acids research* **39**(21): 9283-9293.

Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ et al. 2004. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**(5676): 1497-1500.

Pattanayak V, Ramirez CL, Joung JK, Liu DR. 2011. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature methods* **8**(9): 765-770.

Perez-Pinera P, Ousterout DG, Brunger JM, Farin AM, Glass KA, Guilak F, Crawford GE, Hartemink AJ, Gersbach CA. 2013. Synergistic and tunable human gene activation by combinations of synthetic transcription factors. *Nature methods* **10**(3): 239-242.

Perkins ND. 2007. Integrating cell-signalling pathways with NF-kappaB and IKK function. *Nature reviews Molecular cell biology* **8**(1): 49-62.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic acids research* **40**(Database issue): D130-135.

Reyon D, Tsai SQ, Khayter C, Foden JA, Sander JD, Joung JK. 2012. FLASH assembly of TALENs for high-throughput genome editing. *Nature biotechnology* **30**(5): 460-465.

Sander JD, Cade L, Khayter C, Reyon D, Peterson RT, Joung JK, Yeh JR. 2011. Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nature biotechnology* **29**(8): 697-698.

Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. 2011. genenames.org: the HGNC resources in 2011. *Nucleic acids research* **39**(Database issue): D514-519.

Sigoillot FD, King RW. 2011. Vigilance and validation: Keys to success in RNAi screening. *ACS chemical biology* **6**(1): 47-60.

Sledz CA, Holko M, de Veer MJ, Silverman RH, Williams BR. 2003. Activation of the interferon system by short-interfering RNAs. *Nature cell biology* **5**(9): 834-839.

Smithies O, Gregg RG, Boggs SS, Koralewski MA, Kucherlapati RS. 1985. Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. *Nature* **317**(6034): 230-234.

Sung YH, Baek IJ, Kim DH, Jeon J, Lee J, Lee K, Jeong D, Kim JS, Lee HW. 2013. Knockout mice created by TALEN-mediated gene targeting. *Nature biotechnology* **31**(1): 23-24.

Tesson L, Usal C, Menoret S, Leung E, Niles BJ, Remy S, Santiago Y, Vincent AI, Meng X, Zhang L et al. 2011. Knockout rats generated by embryo microinjection of TALENs. *Nature biotechnology* **29**(8): 695-696.

Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC. 2005. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**(7042): 646-651.

Valton J, Dupuy A, Daboussi F, Thomas S, Marechal A, Macmaster R, Melliand K, Juillerat A, Duchateau P. 2012. Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. *The Journal of biological chemistry* **287**(46): 38427-38432.

Venter JC Adams MD Myers EW Li PW Mural RJ Sutton GG Smith HO Yandell M Evans CA Holt RA et al. 2001. The sequence of the human genome. *Science* **291**(5507): 1304-1351.

Volcic M, Karl S, Baumann B, Salles D, Daniel P, Fulda S, Wiesmuller L. 2012. NF-kappaB regulates DNA double-strand

break repair in conjunction with BRCA1-CtIP complexes. *Nucleic acids research* **40**(1): 181-195.

Watanabe T, Ochiai H, Sakuma T, Horch HW, Hamaguchi N, Nakamura T, Bando T, Ohuchi H, Yamamoto T, Noji S et al. 2012. Non-transgenic genome modifications in a hemimetabolous insect using zinc-finger and TAL effector nucleases. *Nature communications* **3**: 1017.

Weber E, Gruetzner R, Werner S, Engler C, Marillonnet S. 2011. Assembly of designer TAL effectors by Golden Gate cloning. *PloS one* **6**(5): e19722.

Wood AJ, Lo TW, Zeitler B, Pickle CS, Ralston EJ, Lee AH, Amora R, Miller JC, Leung E, Meng X et al. 2011. Targeted genome editing across species using ZFNs and TALENs. *Science* **333**(6040): 307.

Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P. 2011. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **29**(2): 149-153.

국문 초록

TALEN은 최근에 새롭게 개발된 유전자 가위로 DNA 의 염기를 하나씩 특이적으로 인식하는 TALE을 이용한다. 특히 TALEN은 기존의 유전자 가위보다 쉽게 제작할 수 있고 유전자 변이 효율도 상당히 개선되어 다양한 동·식물에 응용되고 있다. 이번 연구에서는 기존의 TALEN의 구조를 개선해 게놈 상에 원치 않는 곳에 변이를 일으킬 가능성을 줄였다. 그리고 컴퓨터 프로그래밍을 통하여 인간 전체의 유전자에 대해, 게놈 상에 유사한 염기서열이 거의 없는 TALEN 적중 염기서열을 선별했다. 선별된 TALEN 적중 서열에 대해 TALEN을 쉽게 만들 수 있는 고속대량 합성법을 개발하였다. 이를 통해 100여개 이상의 유전자를 적중하는 TALEN을 만들었고 실제 변이를 유도한 결과 거의 모든 TALEN이 높은 효율로 유전자 변이를 일으켰다. 변이를 유도하지 못한 두 TALEN의 적중 서열을 분석한 결과, 시토신이 메틸화되어 있었으며 이로 인해 TALEN이 작동하지 못했음을 알 수 있었다. 또한 이 TALEN을 이용하여 유전자 변이가 유도된 세포주를 건립하였으며 실제 유전자의 기능이 없어졌음을 확인했다. 이러한 유전자 녹아웃은 기존의 siRNA를 이용하는 방법보다 매우 효율적으로 완벽히 유전자의 기능을 상실시키는 확인했다. 나아가 인간뿐만 아니라, 다양한 실험동물들에 대해 TALEN 적중 서열을 미리 선별해 놓음으로써 신약개발 등의 다양한 연구에 널리 활용될 수 있을 것으로 생각한다.

학 번: 2008-22719