THE DEVELOPMENT, ENGINEERING AND APPLICATION OF TAL EFFECTOR NUCLEASES FOR TARGETED GENOME MODIFICATION

A DISSERTATION SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF THE UNIVERSITY OF MINNESOTA

Michelle L. Christian

IN PARTIAL FULFILLIMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Advisor: Daniel F. Voytas

September 2013

ACKNOWLEDGEMENTS

Completing a dissertation is a humbling endeavor, one that would be nearly impossible without the guidance and wisdom of my mentors, peers, family and friends. I would like to express my most sincere gratitude:

To my partner in science and life, who has brought a balance and perspective to both for which I am eternally grateful. I cannot wait to see what discoveries we'll make together!

To my mentor in things critical to success in both science and life, Dr. Daniel Voytas. His unwavering support and advice has allowed me to grow as a scientist and a person. I have yet to meet anyone who knows Dan and has anything but genuine respect and praise for both the work and the man. Our ride from beginning to end has been a fast-paced, exciting and often gut-wrenching one, in the most meaningful ways possible to a growing young scientist – thanks for sticking with me through all of it, Dan! Your intuition and perspective of the science is inspiring, and I am truly grateful to be able to walk away from my graduate experience with a long list of personal and professional attributes I will strive to emulate in my own career.

I would also like to thank Dr. Adam Bogdanove for providing the opportunity to experience a truly fantastic, productive and rewarding collaboration over the past four years. Adam, your perpetual enthusiasm for the science is inspiring, and I can't thank you enough for being so open and willing to share your ideas and drive with us! It has been tremendously rewarding working with you and all the members of your lab.

Thanks to those who served on my committee over the years: To my committee chair, Dr. Nik Somia – it was always a sincere pleasure talking science with you. My sincere thanks to Dr. Scott McIvor and Dr. Reuben Harris for your continued support, helpful critique and insightful feedback. I would also like to thank Dr. Louis Manksy for participating in my qualifying exams and allowing me the chance to rotate in the lab, a time that sparked several long-lasting relationships with some wonderful people (not to mention my fiancé!).

To my lab mates, a thousand resounding "thank-you's"! How thrilling it was to work with you all day in and out with hardly a dull moment. A heartfelt thanks to my in-lab mentors Yiping Qi and Feng Zhang who helped me find my place and set high standards in the lab. The Voytas lab members: Colby Starker, Nick Baltes, Tomáš Čermák, Mick Nyquist, Paul Atkins and Javi Gil, you made the good times *grand*, the minutia bearable, and everything in between worthwhile. A special thanks to former lab members Josh Baller, Zach Demorest, Yong Zhang, Kim Nguyen and Xiahong Li, it was truly a pleasure working alongside you all.

The relationships we forged make the leaving so bittersweet, and I am left with lofty expectations for a truly supportive and enjoyable work environment. Don't forget to bust out those TALE t-shirts every once in a while ☺

I've had the privilege of collaborating with some amazing people on many levels. At the U of M: Dr. Mark Osborne, Dr. Dan Carlson, Spring Tan and several members of the Largaspada lab. To our thoughtful collaborator and genome engineering legend Dr. Dana Carroll, whose friendly face and warm regards were always welcome treats over the

years. Thanks also to our collaborators in Seattle, Dr. Barry Stoddard and Dr. Phillip Bradley, for shedding much needed light on the structural aspects of our work.
I thank the Office of Technology Commercialization at the University of Minnesota and the Doctoral Dissertation Fellowship trustees for providing valuable funding that aided in the completion of this dissertation.

Last but most certainly not least, I would like to thank my family and dear friends for their support throughout the MANY years of my education. Their unwavering confidence in my abilities fueled my efforts and brought me incredible comfort. I feel indescribably fortunate to have parents that made continual efforts to feed my curiosities as a child (enter a toy microscope, rock tumbler and telescope, the best presents EVER), nurture my interests as I grew, and celebrate my pursuits – success or failure. You have provided the most valuable framework a person could ask for, one that made it possible to accomplish goals – large and small – I have deemed important in my life. I would have been lost to the abyss of academic rigor if not for the comfortable solace any time spent with my siblings provided. You will forever be the two weirdest people I know, and of that fact I am immensely proud.

I have received endless support from so many colleagues, friends and family over the years it is truly astounding, I could never thank you enough.

DEDICATION

For my parents, who provided for me an environment rich in self-discovery and careful curiosity that has continued to shape my world and teach me to embrace life's lessons with insight, grace and a healthy dose of sarcasm.

And to my siblings, the only two people on this planet with whom I can converse using only lines from The Princess Bride. Inconceivable!

You are truly amazing.

ABSTRACT

The ability to make precise changes to chromosomal DNA has been a long sought goal for geneticists. Targeted genome modification has a variety of applications – ranging from correcting genetic defects in human cells to creating novel, agronomically important traits in crops. The ability to make such targeted DNA modifications has been enabled by nucleases that bind to specific sequences within a gene and create double-strand breaks. Through the action of cellular repair pathways, these targeted breaks lead to localized mutagenesis via non-homologous end joining and to gene editing or insertion via homologous recombination. A primary focus in the field of genome engineering has been to develop tools and techniques that allow precise manipulation of the DNA of various organisms. Zinc finger nucleases and meganucleases are well established as DNA targeting reagents; however, both have limitations in terms of the ease in which they can be engineered to recognize new target sites and their targeting range.

Several years ago, a novel class of DNA binding domain known as Transcriptional Activator-like (TAL) effectors was described. TAL effectors proteins bind to specific sequences in plant genomes and turn on plant genes that promote bacterial infection. DNA target specificity of TAL effectors is conferred by a central array of typically 14-24 repeats, with each repeat recognizing one DNA nucleotide. The one-to-one correspondence of a TAL repeat to a single DNA base constitutes a simple code that can be used to design a TAL effector to targ*et al*most any sequence in a given genome.

Studies in this dissertation were directed at developing, engineering and applying a novel tool for genetic manipulation, called TAL effector Nucleases, or TALENs. The

first efforts of this work demonstrate that fusions of TAL effector proteins to a non-specific nuclease created a targeted DNA break, the repair of which can resulted in site-specific modification of the target sequence. To advance TALEN technology and extend it to target novel DNA targets, we developed a method for rapid construction of engineered TALENs. Using our Golden Gate system, custom TALENs for target sites in essentially any gene of interest can be constructed in 5 days. Finally, We used this method to engineer TALENs targeting genes in *Nicotiana tabacum* (tobacco) and *Arabidopsis thaliana* and tested their ability to create mutations at endogenous loci in both protoplasts and whole plants. Our results indicate that TALENs indeed cleave their intended targets in plant protoplasts of Arabidopsis, and these mutations are heritable. Studies conducted in this dissertation were the first to develop the TALENs, a tool that promises to facilitate the manipulation of natural genomic loci in organisms to a much greater degree than previous targeting reagents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 5

LIST OF PUBLICATIONS

1. Beumer KJ, Trautman JK, **Christian M**, Dahlem T, Lake C, Hawley RS, Grunwald DJ, Voytas DF and Carroll D. 2013. Comparing ZFNs and TALENs for Gene Targeting in Drosophila. ***G3***. [Accepted].

2. **Christian M**, Qi Y, Zhang Y, Bogdanove, AJ, Voytas DF. 2013. Targeted Mutagenesis of *Arabidopsis thaliana* using Engineered TAL effector Nucleases (TALENs). ***G3***. [Accepted].

3. Wendt T, Bach-Holm P, Voytas DF, Starker CG, **Christian M**, Brinch-Pedersen H, Bæksted-Holme. 2013. TAL effector nucleases induce mutations at a pre-selected location in the genome of primary barley transformants. ***Plant Molecular Biology***. doi:10.1007/s11103-013-0078-4.

4. 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. ***PNAS*** 109(43):17382-7.

5. **Christian ML**, Demorest Z, Osbourne M, Starker C, Nyquist M, Carlson D, Bradley P, Bogdanove AJ, and Voytas DF. 2012. Targeting G with TAL effectors: a comparison of activities of TALENs constructed with NN and NK repeat variable di-residues. ***PLoS One*** 7(9):e45383.

6. 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.

7. **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):757-61.

8. Zhang F, Maeder ML, Unger-Wallace E, Hoshaw JP, Reyon D, **Christian M**, Li X, Pierick CJ, Dobbs D, Peterson T, Joung JK, Voytas DF. 2009. High frequency targeted mutagenesis in Arabidopsis thaliana using zinc finger nucleases. ***PNAS*** 107(26):12028-33.

CHAPTER 1

GENERAL INTRODUCTION

1.2. Toward Achieving Precise Modification of the Genomes of Living Organisms

The ability to manipulate specific regions of the genome in model organisms such as yeast and mice has enabled researchers to investigate the roles of genes in both biology and disease. Pioneer work in the field of genome engineering has focused on the development of methods to enable precise manipulation of genomes, and has provided a platform and level of precision never before achieved in the field. In particular, the use of engineered, sequence-specific endonucleases has ushered in a new era of experimental and therapeutic applications. These technologies allow researchers to hone in on a particular sequence in any genome and create a precise DNA sequence modification. In many cases, creating such desired modifications has relied upon harnessing the cell's natural DNA repair processes, particularly the resolution of DNA double-strand breaks (DSBs).

1.2.1. DNA Double Strand Break Repair

Modern methods of genome editing rely on the repair of a DNA double-strand break (DSB). Cells are constantly subjected to various forms of DNA damage, including exposure to UV light, ionizing radiation and chemical mutagens. These agents cause a range of DNA damage products like single- and double-strand breaks, which can lead to unwanted genome rearrangements or trigger apoptosis if left unrepaired by cell machinery (Cahill *et al.*, 2006). Single- and double-strand DNA breaks can also arise spontaneously from errors occurring during DNA replication (Eggleston, 2007). DSBs activate DNA repair mechanisms to ligate the broken ends along with any modifications such as mutations, insertions and deletions (Rouet *et al.*, 1994). Programmed DSBs are

necessary in most organisms to carry out meiosis, but unintended chromosomal breaks in somatic cells threaten the integrity of genomes (Carroll, 2004). With emerging technologies, however, DSBs can also be harnessed for targeted modification by defined outcomes of the repair pathways. In eukaryotic cells, DNA DSBs are repaired via one of two highly conserved pathways: an error-prone non-homologous end-joining (NHEJ) pathway or error-free homology-directed repair (HDR) [reviewed in (Waterworth *et al.*, 2011) (Figure 1-1)].

NHEJ is the most efficient pathway for the repair of DNA double-strand breaks in eukaryotes, but it is often referred to the 'error-prone' prone pathway in comparison to HDR. Breaks repaired via NHEJ result in either perfect restoration of the original sequence or give rise to mutations. In yeast and mammalian cells, 25-50% of nuclease-mediated DSBs are repaired precisely by NHEJ (Budman & Chu, 2005, Fattah *et al.*, 2010). The remaining breaks are subject to imperfect repair that results in small (or less frequently large) insertions or deletions (indels) between the broken DNA ends. NHEJ requires several proteins to resolve a break: first, the DNA-end binding protein Ku70 is loaded onto broken ends and serves as a dock at which a nuclease, polymerase and ligase can bind and proceed through break repair. Nuclease activity of a DNA-PK_{cs} complex chews back damaged ends, which are then filled in by polymerases. Finally, a ligase complex (XLF:XRCC4 in vertebrates) has the ability to ligate across gaps and fuse incompatible DNA ends, thus restoring integrity to the DNA strands.

Homology-directed repair is characterized by its requirement for sequences homologous to the region surrounding a DSB. When sequence homology is present is the form of a sister chromatid or a provided exogenous donor, DSBs are principally repaired

by homologous recombination (HR). Upon initiation of HR by a DSB, processing proceeds by nucleolytic resection to produce 3'- single-strand DNA ends; strand invasion by the single-strand DNA ends produces intermediate structures (D-loops) and a search for sequence homology takes place. In cases when two strands cross over in a DNA exchange, a Holliday junction (HJ) is formed that is resolved to give crossover (gene conversion) or non-crossover products. If, instead, after the initial steps of DSB resection, DNA strand invasion, and repair DNA synthesis, the invading strand dissociates from the D-loop and recombines with the original chromatid, no HJ is formed and only non-crossover products are made (San Filippo *et al.*, 2008). This repair pathway involves DNA synthesis followed by strand annealing, and is thus termed synthesis-dependent strand annealing (SDSA) (Figure 1-1). Most of the evidence to date suggests SDSA is the major HR pathway in mitotic mammalian cells and in plants (Larocque & Jasin, 2010, Waterworth *et al.*, 2011). A DSB that is flanked by direct repeats is repaired using the repetitive DNA sequences in a process called single-strand annealing (SSA). This repair pathway is often utilized in gene reporter assays to assess induction of DSBs by various biological reagents (see Chapter 2).

Repair pathway choice, although complicated, is largely influenced by cell cycle phase in eukaryotes. DSBs incurred during the G1 phase of the cell cycle are predominantly repaired by the NHEJ pathway. Knowing the timing and key players involved in NHEJ and HR has allowed researchers to manipulate the pathways by knocking down/out function of DNA repair proteins to bias pathway choice and DNA break junction outcomes [see (Qi *et al.*, 2013)].

1.2.2. Genome Editing

The first demonstration of targeted gene modification involved the disruption of a user-specified gene in mice via positive and negative selection, in the absence of an induced DSB (Thomas *et al.*, 1986). This approach was subsequently termed gene targeting (GT). The feature that differentiates gene targeting from other gene editing outcomes (*i.e.* disruption mediated by NHEJ) was the inclusion of regions of sequence homology flanking selectable gene marker that enabled the recovery of rare insertion events via HR. However, there were two major limitations of this approach: first, targeted integration of a donor via HR occurs at a low frequency in mammalian cells (1 in 100,000 cells); second, high rates of random integration are observed (Dray & Gloor, 1997, Thomas *et al.*, 1986). Modern techniques of gene editing were borne out of the discovery that an induced DSB significantly enhances the frequency of HR-mediated repair at a target locus up to 10,000-fold (Choulika *et al.*, 1995, Puchta *et al.*, 1993, Rouet *et al.*, 1994, Smih *et al.*, 1995). Thus, genome manipulation stimulated by DSB-inducing reagents is fundamentally different than that achieved by traditional gene targeting. Gene editing that proceeds via HR employ the terms ‘gene correction’ and ‘gene addition’ to more accurately reflect the nature of the repair event following targeted DSBs (Urnov *et al.*, 2010) (Figure 1-2). Alternatively, targeted gene disruption (TGD) occurs when repair of induced DSBs is allowed to proceed via NHEJ, often resulting in small indels that lead to altered/termination of gene products (Figure 1-2). It has become clear that a number of genome modifications can be achieved by exploiting the DNA repair pathways.

1.2.3. Traditional Methods of Genome Manipulation

Traditional methods of making mutations in genes to study their function have relied on random chance. Natural and chemical mutagens like ionizing radiation and ethyl methanesulfonate (EMS) introduce mutations at random, and large-scale screens are necessary in order to find a single organism with the desired mutation from a large population with other random mutations. In plants, traditional approaches to modify genes include targeted-induced local lesions in genomes (TILLING), insertional mutagenesis using transfer DNA (T-DNA) and transposon mutagenesis (Alonso *et al.*, 2003, Ehrhardt & Frommer, 2012, Henikoff *et al.*, 2004, McCallum *et al.*, 2000). These approaches, however, are labor-intensive and provide little control over where the mutations occur. Moreover, many genes are difficult to target (e.g. small micro RNA genes and duplicated gene arrays). Other reverse genetics approaches, such as RNAi and artificial micro RNAs, provide specificity for intended targets, but it is impossible to achieve a null phenotype by knocking down target gene expression (Schwab *et al.*, 2006; Ossowski *et al.*, 2008; Park *et al.*, 2009; Sablok *et al.*, 2011; Chang *et al.*, 2012; Li *et al.*, 2013). While these more targeted methods allowed researchers to better connect observed phenotypes to gene function, there still lacked a rapid and efficient method to introduce small DNA changes at desired locations within a genome.

1.2.4. Current Methods for Targeted Modification of Genomes

Current techniques for precise genome manipulation often involve the use of site-specific nucleases, and these tools will be the focus of the remainder of this dissertation. First,

however, it is worth mentioning a technique for site-specific modification that does not involve a sequence-specific nuclease.

Oligonucleotide-mediated mutagenesis (OMM) is a technique used to introduce specific nucleotide changes to a defined site in a genome using a chemically synthesized oligonucleotide. The technique has many different names in the literature (*e.g.* targeted nucleotide exchange, oligonucleotide–mediated gene editing, chimeric oligonucleotide-dependent mismatch repair, oligonucleotide–mediated gene repair, triplex–forming oligonucleotides induced recombination); each technique is based on the central premise that gene modification occurs via the introduction of a synthesized oligonucleotide containing 1) homology to a target gene and 2) any desired nucleotide changes to be inserted into the target gene (Breyer *et al.*, 2009). Different forms of oligonucleotides are utilized for OMM, including single-strand DNA, chimeric RNA/DNA hybrids, and triplex-forming oligonucleotides (Igoucheva *et al.*, 2004, Parekh-Olmedo *et al.*, 2005, Storici, 2008). The technique has been applied to crop plants such as tobacco, maize and wheat embryos, along with commercial interest of OMM by Cibus Global. Moreover, plants developed through OMM have been declared as non-genetically modified organisms (non-GMO) by U.S. regulatory bodies USDA and APHIS, which could provide a fast-track for the development of agriculturally important novel crops (Breyer *et al.*, 2009). However, there has been discrepancy in the frequency and reproducibility of gene correction, likely influenced by the efficiency of oligonucleotide delivery into the nucleus and the long-term stability of these molecules.

Meganucleases were the first nucleases used to demonstrate definitively that induced, targeted DSBs enhance gene editing (Puchta *et al.*, 1993). Meganucleases are

engineered based on natural homing endonucleases (HEs), which are encoded as genes within introns. These endonucleases have large DNA sequence recognition sites ranging from 12 to 40 bp that generally occur only once in a given genome, providing great specificity. I-SceI, a yeast mitochondrial enzyme, has been used in proof-of-principle experiments to create targeted DSBs (at pre-integrated I-SceI targets) in many model organisms (Haber, 1995, Jasin, 1996, Johnson & Jasin, 2001). Pioneer work on DSB repair in plants was accomplished using I-SceI to cleave a chromosomal reporter, demonstrating that a DSB increased HR frequencies by two orders of magnitude (Puchta *et al.*, 1996).

The scaffold from which most engineered meganucleases are derived comes from the largest of the HE structural families, the LAGLIDADG homing endonucleases (LAGLIDADG represents the conserved aa structural motif representative of each HE family) (Chevalier & Stoddard, 2001). Both academic groups and biotech companies (*e.g.* Pregenen and Precision Biosciences) have described methods to create variants of LAGLIDADG enzymes that have altered DNA target specificities, for the purpose of extending genome editing capabilities to endogenous genes (reviewed in (Stoddard *et al.*, 2007). While meganucleases are highly specific, their engineering is complicated by the overlap between the cleavage and DNA binding domains of the protein, which has hindered widespread adoption of the technology for genome engineering purposes.

DNA Binding Domain-Effector Protein Fusions

The creation of a DSB at a targeted locus can stimulate desirable genomic modifications. Customizable DNA binding proteins offer the opportunity to target such

modifications to unique sequences in a genome. Over the years, a vast array of tools for genome engineering has emerged by fusing customizable DNA binding domains to different effector domains (Figure 1-3). Of particular interest has been the fusion of a DNA binding domain motif derived from transcription factors to a non-specific endonuclease domain from a restriction enzyme, to create a widely used type of DSB-stimulating agent. Together, the combination has made for a potent genome-engineering tool.

The FokI Nuclease

Features of the FokI endonuclease have enabled the rapid development of customizable nucleases as tools gene editing. FokI is a type IIS restriction endonuclease isolated from the bacterium *Flavobacterium okeanokoites* (Sugisaki & Kanazawa, 1981). There are two separate domains of FokI: an N-terminal DNA-binding domain recognizes the palindromic site 5'-GGATG-3' and cleaves 9 and 13 nucleotides downstream; and a C-terminal catalytic domain that provides non-specific cleavage of both methylated and non-methylated DNA (Li & Chandrasegaran, 1993, Li *et al.*, 1992). FokI exists as a monomer in solution and becomes an active dimer when bound to its DNA target (Vanamee *et al.*, 2001). The catalytic residues of FokI are Asp-450, Asp-467, and Lys-469, and (importantly for later nuclease fusion applications) the mutation D450A renders it catalytically inactive (Bitinaite *et al.*, 1998). Additional residues shown to be important in maintaining the dimer interface have be altered to turn the FokI homodimer into an obligate heterodimer, thus minimizing unintended cleavage (Doyon *et al.*, 2011, Guo *et al.*, 2010, Miller *et al.*, 2007, Mineta *et al.*, 2008, Ramalingam *et al.*, 2011, Szczepek *et al.*, 2007). Chandrasegaran and collegues took advantage of the separable DNA-binding

and non-specific cleavage domains of FokI and created the first zinc finger nuclease (ZFN) by substituting the native FokI binding domain with a well-characterized natural zinc finger array, Zif268 (Kim *et al.*, 1996). Researchers have since created hundreds of novel ZFNs, as well as FokI fusions with other sequence-specific DNA binding domains like TAL effectors.

The Zinc Finger Binding Domain and ZFNs

The structure of a zinc finger protein (ZFP) was first described based on transcription factor TFIIIA purified from *Xenopus* oocyte extracts (Miller *et al.*, 1985). This study revealed a new protein fold for nucleic acid binding and a novel type of DNA recognition that was distinctly different from the *helix-turn-helix* motif found in previously described DNA-binding proteins. Since their discovery, more than 700 genes coding for zinc finger proteins have been found in the human genome (Venter *et al.*, 2001). A zinc finger motif is characterized by a Cys2-Hys2 (C2H2) repeat domain that is coordinated by a zinc ion (Pabo *et al.*, 2001, Wolfe *et al.*, 2000). Notable characteristics of this novel DNA-binding motif suggested that the conserved amino acids of each repeat unit provided the framework for tertiary folding, whereas the variable residues determined the specificity of each finger.

The solved crystal structure of the three-finger ZFP of the mouse transcription factor *Zif268* in complex with a DNA substrate revealed the mechanism of zinc finger array (ZFA) binding (Pavletich & Pabo, 1991). Each finger unit or 'module' interacts with 3-bp on one strand of the DNA via the 'recognition helix', specifically with amino acid residues in positions −1, 3 and 6 of the helix. A cross-strand interaction (hydrogen

bond) is predicted between position 2 on the recognition helix and the second DNA strand. Zinc finger modules in transcription factors are often linked tandemly to recognize DNA or RNA sequences in a linear fashion (Figure 1-4). Because zinc fingers function as independent modules, units with user-defined triplet specificities can be linked to provide recognition sequences of different lengths. This seemingly modular design was a key feature in the development of engineered zinc finger proteins, and in particular ZFNs.

The first ZFP-nuclease fusion used the C-terminal portion of FokI to cleave DNA reporters *in vitro* (Kim *et al.*, 1996). Shortly thereafter, it was discovered that FokI needed to dimerize in order to cleave its substrate (Bitinaite *et al.*, 1998). This observation was the rationale for the current architecture of chimeric nucleases, in which two target sites are arranged in a tail-to-tail fashion and separated by a short sequence (5-7 bp) called a spacer (Bibikova *et al.*, 2001). Thus, two ZFN monomers–or a ZFN pair–are required to orient the FokI nuclease in the spacer region to allow dimerization and cleavage of the target (see Figure 1-4B).

Several types of genomic alterations can be introduced with ZFNs, including deletions, insertions, inversions, duplications and genomic translocations, thus permitting the generation of directed mutations in different types of genes and organisms. Studies of the first ZFNs engineered at endogenous genes came quickly: ZFN activity *in* vivo and at an endogenous gene in a model organism (Bibikova *et al.*, 2001, 2002); homologous recombination stimulated by a custom ZFN (Bibikova *et al.*, 2003); and the first data demonstrating ZFNs can cleave genes in human cells (Porteus & Baltimore, 2003). Since these seminal studies, the use of targeting nucleases has spread beyond standard

organisms (Carroll, 2011, Urnov *et al.*, 2010). Moreover, they have been implemented in therapeutic settings, with ZFNs targeting the HIV co-receptor CCR5 gene in clinical trials for the treatment of HIV/AIDS (Perez *et al.*, 2008).

A major bottleneck of early ZFN technology was developing a method to assemble zinc finger proteins that had both high specificity and affinity for their DNA targets. Most ZFNs were engineered by a method termed 'modular assembly', which involved stitching together multiple individual ZFP units or 'modules', (each recognizing a single DNA triplet) to create three or four-finger ZF arrays (Segal *et al.*, 2003). An interesting observation was made when comparing the activity of ZFNs made by modular assembly: only a small number of the ZFNs made by modular assembly were deemed functional (Ramirez *et al.*, 2008). A group of researchers forming the Zinc Finger Consortium sought to develop an assembly method that took into consideration contextual effects between zinc finger units. Thus, the Oligomerized Pool ENgineering (OPEN) platform was developed (Maeder *et al.*, 2008, 2009). ZFNs created with the OPEN platform are highly active nucleases; however the method is labor intensive and requires a skilled molecular biologist to complete successfully (Sander *et al.*, 2011). The most successful and easily-adoptable method for ZFA assembly is termed CoDA (Context-Dependent Assembly) (Sander *et al.*, 2011). Unlike OPEN, CoDA does not require labor-intensive selection steps, and is instead built upon a large archive of pre-selected, optimized two-finger units that are used to assemble full ZFAs. Recent development of additional ZFA assembly methods and the accessibility of publically-available software to identify possible target sites and design ZFNs have greatly improved the dissemination of ZFN technology (Bhakta *et al.*, 2013, Osborn *et al.*, 2011,

Sander *et al.*, 2007). To date, ZFNs have been used to successfully perform gene disruption and addition in a wide variety of organisms (Urnov *et al.*, 2010).

Engineered sequence-specific nucleases must exhibit certain qualities to be considered useful tools for genome engineering: affinity for DNA/RNA targets within the context of a cell nucleus; specificity for long DNA sequences, such that a unique target is only found once in a particular genome; and a modularity that enables customization to target user-defined sequences. Meganucleases possess superior specificity for their cognate targets, but their widespread use is severely limited by complexities in their engineering. ZFNs can be more easily engineered to recognize and bind unique sequences in a variety of genomes (Carroll, 2011, Urnov *et al.*, 2010). However, ZFN engineering has been hindered by complex and time-consuming assembly methods to generate highly specific and active zinc-finger arrays and nucleases (Carroll, 2011). Moreover, ZFNs have a limited targeting range due to context-dependent limitations, with about one ZFN target site for every 200–500 bp (Maeder *et al.*, 2008, Sander *et al.*, 2011). Therefore, other novel DNA binding protein domains (*e.g.* TAL effectors) offer new opportunities to refine and expand genome-editing technologies.

1.3. Transcription Activator-Like Effector Proteins

Transcription activator-like effectors (TALEs) are a class of proteins with DNA-binding capabilities found in many members of the plant pathogenic bacterial genus *Xanthomonas*. Upon infecting plant tissue, TALEs are secreted into plant cells via the type III secretion pathway. Once inside the nucleus, they bind to specific sequences in

plant the genome to activate genes necessary to promote the bacterial infection (Bogdanove *et al.*, 2010).

1.3.1. TAL effector Biology

Many species and pathovars of *Xanthomonas* affect a wide variety of plants, including rice, citrus, pepper, cotton and soybean, and their TAL effectors are critical to the ability of the bacterium to infect its host (Mansfield *et al.*, 2012). Early studies in pepper plants showed that TALEs entered the nucleus of cells via the type III secretion system (T3SS) and took control of a gene that regulates cell size, causing plant cells to expand and burst (Kay *et al.*, 2007, Römer *et al.*, 2007). Individual strains of *Xanthomonas* may have anywhere from one to several dozen TAL effectors (Bonas *et al.*, 1989, Yang & White, 2004). TALEs have been shown to play roles in both plant disease and resistance. Some TAL effectors interact with importinα (impα) in the host plant cytoplasm and translocate to the nucleus, where they bind to promoters and transcriptionally activate matching host susceptibility (S) genes such as UPA20, Os8N3, OsTFIIAg1, or OsTFX1, resulting in disease (Figure 1-5). The arms race existing between plant and pathogen has enabled many plant species to evolve mechanisms to resist infection. For example, S gene activation can be suppressed by a) mutations in the individual promoter target sites (*e.g.* xa13 mutation) or b) activation traps for resistance in which executor resistance (R) genes are under transcriptional control of TALE target sequences that would typically be present in S gene promoters (Figure 1-5).

The first TAL effector discovered was AvrBs3 from *Xanthomonas campestris pv. Vesicatoria*, a pepper pathogen (Bonas *et al.*, 1989). This TAL effector triggers an

immune response only in plants carrying the disease resistance *Bs3* gene. A striking observation was made regarding the nature of the *AvrBs3* gene: the open reading frame contained 17 direct 102 bp repeats which shared 91% to 100% sequence homology (Bonas *et al.*, 1989). Thus, the hunt began to identify the mechanism by which TAL proteins recognize and bind their promoter targets. Two key studies demonstrating that AvrBs3 can elicit either resistance or susceptibility phenotypes in the same species led to the discovery of a conserved nucleotide sequence that is required for host gene activation by TAL effectors, subsequently named the 'up-regulated by AvrBs3' (UPA) or 'up-regulated by TALE' (UPT) box (Kay *et al.*, 2007, Römer *et al.*, 2007). Together, these studies showed for the first time evidence for sequence specific DNA binding by a TAL effector.

Several features distinguish this class of DNA binding proteins: TAL effectors have and N-terminus required for type III secretion and a C-terminus that contains several nuclear localization signals (NLS), along with an acidic activation domain (AAD) that is a common feature of typical transcription factors (Figure 1-6). Like ZFNs, TAL effectors contain a central region that is responsible for binding DNA. Target specificity of TAL effector proteins is conferred by a central array of near-perfect repeats. Each repeat consists of 34 aa (though 30 to 42 aa repeats exist), and nearly all repeat arrays terminate with a final, truncated repeat containing the first 20 aa of a canonical repeat, referred to as a half repeat (Boch & Bonas, 2010). Additionally, TAL effectors that contain 33 aa repeats contain a deletion of the residue in the 13^{th} position of a repeat. The number of repeats within a TALE array varies from protein to protein, ranging from 1.5–33.5, with an average of 17.5 (Boch & Bonas, 2010). There also exist

polymorphisms across TALE proteins, both in the N- and C-termini and within the repeats. The most distinct variability exists at residues 12 and 13 of a repeat array, referred to as the repeat-variable diresidue (RVD) (Moscou & Bogdanove, 2009). Amongst naturally occurring TAL effectors, over 20 distinct RVDs have been observed; however, HD, NG, NI, and NN account for over 75% of the total (Moscou & Bogdanove, 2009).

1.3.2. TAL Effector Protein-DNA Recognition: Breaking the Code

The central repeat arrays of TALEs constitute a novel DNA-binding domain that shows little resemblance to existing DNA binding motifs. Therefore, it has been historically difficult to elucidate the residues responsible for contacting DNA (Kay *et al.*, 2007). Early data indicated that TAL effector proteins had affinity for DNA, and that the repeat domains controlled the specificity of TAL protein binding (Herbers *et al.*, 1992, Yang *et al.*, 2000). Nearly a decade later, the basis for sequence-specific recognition by TAL effectors was discovered independently by two groups (Boch *et al.*, 2009, Moscou & Bogdanove, 2009). One immediate and striking feature of this novel DNA binding domain is that each repeat of an array binds to one DNA nucleotide, thus providing recognition and binding of TAL effectors to specific DNA sequences within a gene. Both groups were able to derive a DNA binding 'code' for the one-to-one RVD-nucleotide recognition of TAL repeats (Figure 1-6). The association frequencies for the most common or canonical RVDs are as follows: HD binds C most frequently; NG associates with T; NI with A; and NN recognizes both G and A. Additional common RVDs include NS which recognizes A, C and G, while HG mimics NG in binding T. N* (missing

residue 13, for a 33 aa repeat) is the most promiscuous RVD having been found to associate with all four nucleotides at some frequency (Figure 1-6B). Association frequencies for RVDs that occur rarely are less predictive of true RVD-nucleotide interactions (Moscou & Bogdanove, 2009). A key feature of nearly all sequences bound by TAL effectors is a conserved T at the -1 position. In functional studies of targeted gene activation using TAL effectors, permutation of the conserved T just upstream of the TALE repeat binding sequence resulted in loss of gene activation (Boch *et al.*, 2009). Further insight into the significance of the conserved T and -1 position within a repeat array was gained by subsequent functional and structural studies of TAL effector proteins. Taken together, the one-to-one correspondence of a TAL repeat to a single DNA base constitutes a simple code that can be used to predict TAL effector DNA binding sites and to design TAL effector arrays to recognize unique sequences in any genome.

Insight into TALE-DNA interaction from Structural Studies

Recent structural studies of TAL effectors have provided both the structural basis for TAL effector-DNA recognition and valuable insight to further aid the development of TALE engineering technology. The first structure solved was an artificial (exact array composition not yet found in nature) TAL effector dHAX3 with 11 repeats of 34 aa, plus the half repeat, representing three common RVDs (HD, NG and NS) (Deng *et al.*, 2012). It was solved to high resolution in both an unbound state (1.85 Å) and in a complex with DNA (2.4 Å). When unbound, the dHAX3 TAL effector displayed an extended conformation nearly two times than that of its bound conformation. The structure of the

naturally occurring TAL effector PthXo1 from the rice pathogen *Xanthomonas oryzae* was solved in a second study (Mak *et al.*, 2012). Although the resolution was lower (d_{min} = 3 Å), PthXo1 contains 23.5 repeats bound to two full turns of DNA, and provides DNA interactions for six unique RVDs (HD, NN, NG, HG, NI and N*). In addition, the structure of PthXo1 was solved bound to its natural DNA target, which exhibits several 'mismatches' between RVDs and their predicted nucleotide associations. Such interactions provide further insight into the mechanism of DNA recognition exhibited by TAL effectors. Both structures reveal that TAL repeats each form left-handed, two-helix bundles that are connected by a loop comprised of residues 12 and 13 of a repeat. When situated in an array, the self-associating repeats form a right-handed superhelix that wraps around the major groove the length of the DNA target (Figure 1-7).

DNA recognition mechanism

These first structural studies also reveal the nature of sequence-specific contacts between the TAL effector and the DNA major groove. The second residue in the RVD loop (position 13 of each repeat) determines nucleotide specificity by making contacts at each base along a single, contiguous strand of the DNA target (Figure 1-7). The first residue in the loop (position 12) likely plays a stabilizing role in establishing a 'pre-bound' conformation by forming a hydrogen bond between the side chain (of residue 12) and the backbone (of the first helix) in each repeat (Figure 1-7C**)**. The residue in position eight of the same repeat also interacts with residue 12, thereby likely providing additional stability within the individual repeat units. The glutamine (Gln) at position 17 is the only repeat residue that directly contacts the phosphate backbone.

Both structure prediction simulations and evidence from solved crystal structures provide an explanation for observations made in previous functional and molecular studies regarding the TAL effector-DNA binding (Bradley 2012, Mak *et al.*, 2012). Specificity determining interactions can be classified by the type of contact made between RVD loop-residing residues and the DNA base: first, strong hydrogen-bonds provide recognition for RVDs such as HD to cytosine or NN to a purine (A or G); second, in the absence of hydrogen bonds, weaker van der Waals contacts and efficient packing between the backbone of a glycine and a thymidine (T) nucleotide provides the basis of NG and HG contacts; third, relaxed specificity due to the steric exclusion of alternate bases serves to explain the preference of NI for A nucleotides; and finally, N* RVDs that lack a residue at position 13 and thus lack an appropriate side chain are shown to accommodate any base, with little to no impact on overall affinity of the array (Mak *et al.*, 2012).

The impact of TAL repeats on the coordination of RVD-nucleotide binding is abundantly clear from molecular, functional, and now structural studies. The mechanism by which these proteins seek out and identify their cognate DNA targets, however, remains less clear. Additional evidence from structural and other recent studies suggests a likely role for the N- and C-terminal sequences that flank the TALE repeats (Deng *et al.*, 2012, Mak *et al.*, 2012, Miller *et al.*, 2011, Mussolino *et al.*, 2011). Deletion of approximately 150 aa from the N-terminus of the AvrBs3 TAL effector removed features required for secretion, but integrity of the protein remained unaltered (Szurek *et al.*, 2002). However, further truncations result in loss of binding activity, indicating possible structural features are necessary for TAL binding (Miller *et al.*, 2011, Mussolino *et al.*,

2011). The region immediately preceding the repeat array (residues 120-254) in naturally occurring TAL effectors contains 11 conserved lysine and arginine residues that contribute to the overall basic charge of the protein N-terminus. The structure of the PthXo1 TALE bound to its DNA target revealed two 'cryptic' repeats (termed the -1 and 0 repeats) in the N-terminal region just upstream of the canonical repeats (Mak *et al.*, 2012). Residues 221 to 239 and residues 256 to 273 each form a helix and loop similar to the helix and RVD loop of the canonical repeats. A tryptophan residue at position 232 forms a nonpolar contact with the conserved thymidine (T) nucleotide at position 0 of nearly all TAL effector DNA target sites (Mak, *et al.*, 2012). A more recent investigation of the structure of an extended N-terminal region of the dHAX3 effector indicates that as many as four additional cryptic repeats are formed (Gao *et al.*, 2012), which suggests that this region provides the necessary binding energy for high affinity binding between TAL effectors and specific DNA targets. In contrast to contributions of the N-terminal portion of TAL effectors to overall binding affinity, the C-terminal region appears to be less critical. Studies of various C-terminal truncations and the contributions thereof are discussed further in the context of TAL effector engineering in chapter 4 of this dissertation.

1.4. Genome editing using TAL effector technology

The development of ZFN technology set precedence for the general application of DNA binding domains fused to a non-specific FokI DNA cleavage domain (Kim *et al.*, 1996). However, at the time of their creation, TAL effector–nuclease fusion proteins were not necessarily expected to function as ZFNs, based on data strongly suggesting that

TAL effector proteins form dimers in the cytosol of host cells before being transported into the nucleus (Gürlebeck *et al.*, 2005, 2006, Kay *et al.*, 2005, 2009). This notion would have posed problems in engineering a TAL effector nuclease using a similar platform that was successful for ZFNs (see Figure 1-4). The first demonstration of TAL effector arrays fused to a FokI nuclease indicated that the proteins were indeed able to cleave their predicted targets with efficiencies comparable to ZFNs.

1.4.1. TAL effector Nuclease and other Protein Fusions

Parts of this section were reprinted with permission and originally appeared in "TAL effector-Mediated DNA Modification" United States Patent Application 20110145940, Kind Code: A1. Voytas, Daniel F., Bogdanove, Adam, Zhang, Feng, Christian, Michelle, Cermak, Tomas, Schmidt, Clarice Lauer, Doyle, Erin, Wang, Li., 2011.

Work presented in this dissertation describes the first efforts to target DSBs using TAL effector–FokI nuclease fusion proteins, and are described in further detail in Chapter 2. The first proof-of-principle papers that demonstrated the creation of functional TAL effector nucleases used 1) naturally occurring TAL effector repeat arrays with pre-determined DNA targets, and 2) designed arrays to target novel DNA sequences. TALEN-DNA binding and cleavage tests were performed both *in vitro* and *in vivo* in the two studies (Christian *et al.*, 2010, Li *et al.*, 2010).

The TAL effector nucleases (TALENs) function as dimers, with each monomer composed of TALE DNA recognition repeats fused to the non-specific cleavage domain from the FokI endonuclease (Figure 1-8). TAL nuclease monomers bind to one of two DNA half-sites that is separated by a spacer sequence. This spacing allows the FokI

monomers to dimerize and create a double-strand DNA break (DSB) in the spacer sequence between the half-sites (Figure 1-8). To explore the potential of the TAL effector DNA recognition domain, experiments were conducted to determine whether native TAL effectors can function as nucleases when fused with the FokI nuclease domain. A yeast-based assay was carried out by using a TAL nuclease expression construct and a target reporter construct (see Chapter 2) (Christian *et al.*, 2010, Townsend *et al.*, 2009).

In this assay, a target plasmid and TALEN expression plasmids are brought together in the same cell by mating. The target reporter has a disrupted *lacZ* reporter gene with a 125 bp duplication of coding sequence that flanks a target sequence (composed of two half sites and a spacer sequence) recognized by TAL effector DNA binding domains. If the TAL effector nuclease binds and generates DNA double-strand breaks at the target site, such breaks, in yeast, are repaired predominantly by homologous recombination between the duplicated *lacZ* sequences through single strand annealing (SSA) (Haber, 1995). Recombination results in reconstitution of a functional *lacZ* gene and loss of *URA3* (conferring 5-fluoroorotic acid resistance). Relative cleavage activity of TALENs was measured by determining β-gal enzyme activity. In these studies, two native TAL effectors, AvrBs3 and PthXo1, were cloned into a nuclease expression vector, and their cognate target sites (two binding sites arranged in an inverted orientation) with an 18 bp spacer sequence placed the target reporter vector (see Chapter 2, figure 2-1). The results showed that the LacZ activity from yeast cells transformed with both the TALENs and the cognate target reporters was significantly higher (15.8- 30 fold higher) than the control yeast cells that contained only the targets (see Chapter 2 of this dissertation). These data represent the first TAL effector nucleases tested *in vivo* and indicate that

TALENs can function as site-specific nucleases that cleave intended target sequence in yeast. In particular, the ability to create TALENs that recognize novel, arbitrarily selected target sequences suggested a potential for engineering TALENs for targeted genome modification.

A second study demonstrating the initial development of TALE nucleases (termed TALNs in the publication) tested the fusion of the FokI nuclease to both the N- and C-terminus of the full length AvrXa7 TAL effector (Li *et al.*, 2010) (Figure 1-8). Fusions of FokI to the N-terminus of the effector protein were less active than those made to the C-terminus (approximately14-fold change versus a 290-fold change over negative control, respectively.) Therefore, the currently adopted architecture of TALENs has followed a C-terminal fusion of FokI nuclease to a TAL effector protein. Later in the same year, a third study was published that tested TALE–nuclease fusions based on the Hax3 effector from the *Brassicaceae* pathogen, and verified TALEN activity *in vivo* using a transient cleavage assay in tobacco leaves (Mahfouz *et al.*, 2010). Taken together, the first publications on TALEN technology made several important findings: first, TAL effector–nuclease fusions retain the ability to bind DNA; second, TALENs cleave their predicted DNA targets both *in vitro* and *in vivo*, with activities comparable to active ZFNs; third, TALE arrays can be engineered to recognize novel, predicted DNA targets; and lastly–unlike ZFNs–TALENs tested in these studies retain cleavage activity over a range of target spacer lengths, which became a point of improvement in future studies. Importantly, the seemingly modular nature of TALE repeat arrays overcame a significant targeting limitation imposed by previous gene editing technologies, and made TAL effector-based reagents ideal for targeting any given sequence in a genome. A significant

bottleneck in initial studies, however, was the inability to rapidly engineer novel TALE arrays to a desired sequence. The response by researchers to this limitation was astoundingly swift, and an onslaught of TALE engineering methods have since been made available.

1.4.2. Methods to Engineer/Assemble Novel TALEs

The number of both publicly and commercially available platforms from which to engineer novel TAL effectors is a testament to the immense and almost immediate interest in TALE-based technology and applications, and the overall ease with which this technology can be adopted and practiced by nearly any competent bench scientist. To date, approximately 15 different TAL effector engineering methods have been developed by academic labs, and at least 8 commercial outlets now offer TAL engineering and synthesis services. In addition, several core facilities have been set up at various intuitions in the U.S. (see Table 1-2). Each existing TALE assembly method offers its own set of advantages and drawbacks, many of which are summarized in Table 1-2. One of the first and most widely adopted TALE assembly methods was based on a highly efficient cloning system developed independently of TAL effector protein engineering (Engler *et al.*, 2008, 2009). Termed Golden Gate cloning, this method allows for the ordered assembly of multiple DNA fragments in a single reaction, capitalizing on properties of certain Type IIS restriction endonucleases. Type IIS restriction enzymes cleave outside their recognition sequence to create unique 4 bp overhangs; the sequence of these 'sticky ends' can be designed such that after cycles of digestion and ligation, only fragments with corresponding overhangs will come together in the desired order (Figure 1-10). Golden Gate-based methods have the advantage of avoiding the need for

PCR-based assembly that can lead to PCR errors, and no gel purification steps are required, which are tedious and time-consuming. Similar versions of our method have been developed that have comparable output (Table 1-2). Yet other groups have devised methods to scale up the production of engineered TAL effectors and are thus considered to be 'high-throughput' (indicated in Table 1-2). Finally, a number of companies have begun to offer TAL design and engineering services for a fee. These services, though based on proprietary methods, are becoming less costly and offer an alternative for academic researchers who wish to use the technology without implementing 'in-house' assembly methods.

1.4.3. Efforts to Modify Genes using TAL effector Fusions

With the assembly of novel TALE arrays and fusion to numerous effector domains no longer a bottleneck, the pace at which the technology has been applied to genome engineering is astounding. In the two short years following the publication of the first proof-of-principle studies, there have been over 80 published examples of TALE-mediated genome modification, spanning 17 different model organisms (summarized in Table 1-1). Many report the use of single TALE nuclease pairs to mediate the formation of small insertions and deletions (indels) at a gene of interest, thereby altering the gene product. More complex chromosomal modifications can be generated when two TALENs pairs are used to create DSBs simultaneously, leading to the recovery of large deletions and inversions (Ansai *et al.*, 2013, Carlson *et al.*, 2012, Ma *et al.*, 2012). Recent work has also shown that single-strand oligonucleotides can be used to correct a TALEN-induced break via homologous gene repair in zebrafish (Bedell *et al.*, 2012).

In a key study demonstrating efficient modification of endogenous gene targets, TALENs were shown to be active in a mammalian cell environment through targeted modification of human genes *NTF3* and *CCR5* (Miller *et al.*, 2011). An important finding in this study and critical step for improving the TAL effector technology was to define an optimal TALEN architecture sufficient for binding and cleavage, which is addressed in detail in chapters 2 and 4 of this dissertation. Subsequent to this work, TALENs have been utilized in a large number of applications (Table 1-1).

TALENs offer a desirable method with which to introduce targeted alterations into the genomes of organisms that have been traditionally difficult to manipulate (*e.g.* plants, roundworm, zebrafish, frog, mouse and rat). They have also been used to modify endogenous genes in less common organisms such as pig, cow, cricket, mosquito and silkworm, some leading to the development of new animal models of human disease (Aryan *et al.*, 2013, Bedell *et al.*, 2012, Cade *et al.*, 2012, Carlson *et al.*, 2012, Ding *et al.*, 2012, Gupta *et al.*, 2013, Hockemeyer *et al.*, 2011, Huang *et al.*, 2011, Lei *et al.*, 2012, Li *et al.*, 2012, Liu *et al.*, 2012, Ma *et al.*, 2012, Moore *et al.*, 2012, Navis *et al.*, 2013, Sajwan *et al.*, 2012, Sander, Cade, *et al.*, 2011, Sun *et al.*, 2012, Tesson *et al.*, 2011, Tong *et al.*, 2012, Watanabe *et al.*, 2012, Wood *et al.*, 2011). Thus, efforts have begun to move past proof-of-principle studies and extend into creating mutated alleles and chromosomal translocations to model human diseases. Several studies demonstrated the use of TALENs for rapid and efficient gene modification to recapitulate disease models in human cell culture, importantly embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs) (Ding *et al.*, 2012, Hockemeyer *et al.*, 2011, Miller *et al.*, 2011, Osborn *et al.*, 2013, Piganeau *et al.*, 2013, Sakuma *et al.*, 2013, Sun *et al.*, 2012).

For example, one such report demonstrated cellular phenotypes directly linked to human disease (*e.g.* dyslipidemia, insulin resistance, hypoglycemia, lipodystrophy, motor-neuron death, hepatitis C infection) after modification with TALENs, while another study utilized a TALEN-mediated DSB followed by repair with a homologous donor template to correct a mutation in the *COL7A1* gene that causes epidermolysis bullosa (Ding *et al.*, 2012, Osborn *et al.*, 2013). Studies such as these provide a glimpse as to how TALENs can be used to study disease and may lead to critical therapeutic applications in the future.

Nuclease-mediated gene editing is advantageous for modifying the genomes of organisms that have longer life cycles and present more complicated technical hurdles, such as large animals and agriculturally important crops. TALENs have been used to successfully knock out the low-density lipoprotein (LDL) receptor in miniature pigs to create animals that can be used to study cholesterol-related diseases (Carlson *et al.*, 2012). TALENs have been used to edit genes in several plant species as well, with the first demonstration in tobacco and Arabidopsis protoplasts (Cermak *et al.*, 2011, Mahfouz *et al.*, 2010, Zhang *et al.*, 2011). TALENs have been successfully used to modify the genomes of three additional plants, namely barley, rice and Brachypodium (Li *et al.,* 2012, Shan *et al.*, 2013). For example, a TALEN pair was used to disrupt the natural target of TALEs from pathogenic strains of *Xanthomonas* in order to render rice plants resistant to infection (Li *et al.*, 2012). Taken together, these studies demonstrate the great potential that TALENs offer for modifying crop and livestock genomes and developing new animal models of disease.

Demonstration that fusions of a TAL effector DNA binding domain to an endonuclease function as intended has opened the door for additional applications in order to direct activity of other functional domains to a predetermined sequence (Figure 1-3). To date, there have been several reports of successfully altering endogenous gene expression using TALE effector arrays fused to transcriptional activator, repressor or chromatin-modifying domains (see Table 1-1). The acidic activation domain (AAD) present in full-length TALE proteins has been harnessed to activate expression of endogenous plant genes using naturally occurring TAL effectors (Boch *et al.*, 2009, Morbitzer *et al.*, 2010, Römer *et al.*, 2009, Scholze & Boch, 2010). More recently, engineered TAL effector arrays have been fused to well-characterized activation domains (e.g. VP16 and VP64) or repressors (e.g. SID and KRAB) to create artificial transcription factors that can induce changes in gene or protein expression in the range of two-fold to 30-fold (Bultmann *et al.*, 2012, Cermak *et al.*, 2011, Cong *et al.*, 2012, Garg *et al.*, 2012, Geissler *et al.*, 2011, Li *et al.*, 2012, Maeder *et al.*, 2013, Mahfouz *et al.*, 2012, Miller *et al.*, 2011, Perez-Pinera *et al.*, 2013, Politz *et al.*, 2012, Tremblay *et al.*, 2012, Wang *et al.*, 2012, Zhang *et al.*, 2011). Synergistic effects on endogenous gene expression have also been observed upon co-expression of multiple TALE activators targeting a single promoter in mammalian cells (Maeder *et al.*, 2013, Perez-Pinera *et al.*, 2013).

Features of TALE technology have enabled large-scale studies never before seen with nuclease technology (Chen *et al.*, 2013, Kim *et al.*, 2013, Maeder *et al.*, 2013, Schmid-Burgk *et al.*, 2012). One group assembled TALEN plasmids for 18,740 protein–coding genes spanning the human genome, and showed that all TALENs tested in a pilot study (targeting 124 genes) were active and disrupted their target genes at high

frequencies (Kim *et al.*, 2013). In addition to the ability to efficiently scale up studies using the TALE technology, fusions to yet other functional domains to create TALE recombinases, deaminases and methylases have also been considered to extend the power and utility of TAL effector DNA binding arrays (Mercer *et al.*, 2012, Pennisi, 2012).

In addition to swapping out different functional domains at the ends of TALE arrays, there have been efforts to create monomeric nucleases that retain the ability to cleave DNA while employing a single DNA recognition component (Beurdeley *et al.*, 2013, Fonfara *et al.*, 2011, Kleinstiver *et al.*, 2012). One significant advantage of monomeric nuclease architecture is the reduced size of nuclease protein. TALENs are roughly three times larger than other engineered nucleases (~1,800 aa total), which can be a limitation in delivery to a cell, whether as DNA, RNA or protein. Furthermore, the highly repetitive nature of TAL effector arrays can complicate efforts to clone, deliver and integrate a TALEN pair into the genome. Two recent proof-of-principle studies demonstrated the fusion of either a ZFA or TALE protein to the catalytic domain of the I-TevI HE resulted in a functional designer nuclease with efficiencies of mutagenesis similar to those reported for ZFNs and TALENs. (Beurdeley *et al.*, 2013, Kleinstiver *et al.*, 2012).

1.4.4. Strategies to Optimize TALE design and binding specificity

Concurrent with studies demonstrating the use of TALE fusions for genome modification in model systems, much work has focused on optimizing the technology, both to enhance targeting capacity and eliminate undesirable effects. A potential drawback for the use of TAL effector technologies in cells is their size. Compared to the other two nuclease platforms, TAL effectors are the largest proteins (~950 amino acids

for a TALEN monomer compared with ~310 amino acids for a ZFN monomer and ~165 amino acids for a meganuclease monomer). Following in the steps of zinc finger nuclease development, engineering efforts have focused on determining the minimal TALE architecture required to achieve maximum binding and nuclease activity. This included: optimizing both the N- and C-terminal regions surrounding the repeat array, termed the 'backbone'; determining the number of repeat arrays necessary for optimal nuclease activity, in light of a spacer length requirement. As stated previously, fusions of TAL effector proteins to the FokI nuclease were initially constructed using a full-length or near full-length, naturally occurring TAL protein (Figure 1-8B) (Christian *et al.*, 2010, Li *et al.*, 2010). In the TAL effector PthXo1, for example, 288 and 295 amino acids at the N- and C-termini flank the repeat region, respectively. Features present within the N-terminal region include up to four cryptic repeats; within the C-terminus lie two functional nuclear localization signals (NLS), along with an acidic activation domain (present in the full-length TALEN backbone only) (Figure 1-9).

To delimit the DNA binding domain, different research groups have made and tested various N- and C-terminal truncations (Figure 1-8B, Figure 1-9). For example, an N-terminal truncation at residue 152 (retains 136 aa, N152) and C-terminal truncations leaving just 18 amino acids after the repeat domain still allow for effective DNA binding (Miller *et al.*, 2011, Mussolino *et al.*, 2011, Sun *et al.*, 2012). The structure of TAL effectors bound to DNA is consistent with truncations of this extent still being able to bind DNA (Deng *et al.*, 2012, Mak *et al.*, 2012). An earlier study on TAL effector secretion found that the first 152 residues of the N-terminus specifies transport into plant cells, but is dispensable for other functions (Szurek *et al.*, 2002). However, additional

groups have reported that further truncation of this region leads to a loss of TAL effector and TALE nuclease function (Meckler *et al.*, 2013, Mussolino *et al.*, 2011, Sun *et al.*, 2012). For example, both TALE activators and nucleases retaining 50–94 residues of the N-terminus directly upstream of the first repeat display little or no activity, suggesting that specific residues or structural features within this region influence the binding ability and likely the stability of TAL effector proteins (Meckler *et al.*, 2013, Sun *et al.*, 2012). Indeed, the reported structure of the N-terminal residues bound to DNA reveals cryptic repeats that might initiate critical stabilizing interactions, specifically between the tryptophan (W232) in the -1 repeat of PthXo1 effector and the conserved T at the 0 position in the TALE target sites. N-terminal residues 148-288 autonomously bind DNA non-specifically, suggesting that the region is responsible for initiating nucleation of the repeat super–helical structure that wraps around the target (Gao *et al.*, 2012, Mak *et al.*, 2012).

The length of the C-terminus immediately after the last half-repeat of a TAL effector DNA binding domain has a direct impact on the activity of TAL effector nucleases. In contrast to N-terminus, it appears that little of the C-terminus is required for TALEN activity, and truncations as this end appear to dictate the optimal spacer length necessary for efficient cleavage by FokI (Miller *et al.*, 2011, Christian *et al.*, 2012, Zhang *et al.*, 2013). Removal of all but the first 63 aa following the 20th residue of the half-repeat of an array has been shown to be an effective backbone architecture for both TALE activators and nucleases (Miller *et al.*, 2011). TALEN architectures with even further truncation of the C-termini (termed C17, C18 retain 17 or 18 aa) are also active, and require shorter spacer lengths for optimal activity (12-14 bp) compared to the spacer

length requirement for the C63 architecture (15-18 bp) (Christian *et al.*, 2012, Mussolino *et al.*, 2011). These results are described in detail in Chapter 3 of this dissertation. We have further tested these observations in tobacco protoplasts and confirmed mutagenesis activity by high-throughput sequencing. We frequently observed a 25-fold increase in mutagenesis at endogenous targets with the truncated TALENs relative to those with the near full-length backbone architecture (Zhang *et al.*, 2013).

Number of repeats vs. spacer length The length of the TAL effector DNA binding domain also affects binding and efficiency, and thus has been subject to optimization. TAL effectors found in nature vary in the number of repeats within a particular array, with an average of 17.5 repeats (Boch & Bonas, 2010). A systematic study of TALE binding indicates that at least 6.5 repeats are required for activation of a downstream reporter gene, and activity greatly increases with 9.5 repeats and continues to rise as repeats are added thereafter, until a plateau is reached after 14.5 repeats (Boch *et al.*, 2009). The general consensus amongst data from TALEN studies is that between 14 and 18 repeats is ideal to ensure specific binding. The first investigation into the effects of spacer length on TALEN efficiency is presented in Chapter 2 of this dissertation.

Concern over complexities in TALE–DNA binding The notion of complexities in TALE–DNA binding was initially based on observations that single base pair mutations in the AvrBs3 binding site result in both qualitative and quantitative effects (Kay *et al.*, 2009, Römer *et al.*, 2009). Although there is, on average, no influence of immediate neighbors on RVD–nucleotide associations (Moscou & Bogdanove, 2009), individual RVD contribution, RVD composition, neighbor effects and length of an array might affect the stringency of RVD–nucleotide specificities at different positions.

General contributions of individual RVDS, context dependency and location bias Streubel *et al*., (2012) aimed to study properties of individual RVD types using artificial TALE activators. Several interesting observations were made: TALEs containing a string of HD and NN RVDs were able to bind polycytosine (C) and polygaunine (G) targets, respectively, and activate the downstream reporter gene; however, the same NN TALE was unable to activate a polyadenine (A) reporter, despite the known specificity of NN for both G and A nucleotides. Somewhat surprisingly, neither TALEs containing NI nor NG RVDs were able to activate their cognate reporters (string of A or T, respectively), lending to the notion that there are both 'strong' (HD, NN) and 'weak' RVDs (NI, NG, NK), and–contrary to the simple TALE code reported thus far–TALEs with certain RVD compositions are not functional. A report in contradiction to the finding that NG is a weak RVD suggested that the lack of activity was a result the use of uninterrupted runs of six identical repeats in the assays (Meckler *et al.*, 2013).

New RVD specificities (NH, HN) In addition to the four most common or canonical RVDs (HD, NI, NN and NG), several uncommon or rare RVDs have been shown to provide novel specificities (Cong *et al.*, 2012, Miller *et al.*, 2011, Streubel *et al.*, 2012). In two independent studies, the RVD NH had specificity for guanine only, unlike the standard NN RVD that recognizes both G and A nucleotides (although perhaps not with equal affinity (Meckler *et al.*, 2013). The RVD NP also showed a novel specificity for three different bases (A, C and T), and the RVD NS, which was previously designated as nonspecific, showed a clear preference for A and G, although activation levels were low. In vitro SELEX data first suggested that NK RVDs may have higher specificity for G than NN (Miller *et al.*, 2011); however, modification of a locus in zebrafish was more

effective when engineered TALENs were made using repeats with NN RVDs rather than NK (Huang *et al.*, 2011). In vitro binding studies of TALENs containing NN or NK RVDs revealed that binding affinity was greatly reduced with the addition of multiple NK RVDs to an array (Christian *et al.*, 2012). Further efforts described in this dissertation have added to the understanding of the basis for the RVD NK–G nucleotide recognition (Chapter 4).

The new guanine-specific NH RVD was shown to be much more efficient at recognizing G than the NK RVD in TAL effector activators (Cong *et al.*, 2012, Streubel *et al.*, 2012). Interestingly, NH-containing TALEs exhibited significantly higher specificity for a G-based reporter than the TALEs containing the canonical NN RVD in both reporter activation assays and at endogenous loci (Cong *et al.*, 2012). Computational analysis of the relative binding affinities between the TALE and its bound DNA revealed that substitution of a single NH for the NN RVD results in a gain in free energy in the DNA-bound state. Taken together, these data suggest that NH may be a more suitable RVD for targeting G nucleotides. Furthermore, the same study revealed that the NA RVD showed similar specificity for all four bases, and may be useful for targeting degenerate DNA sequences.

Strong repeats (HD or NN) are key for overall efficiency It is now clear that individual RVDs have differing contributions to the overall TAL effector-DNA binding interaction. The concept of designating RVDs as strong or weak was supported by a recovery of TAL effector function when several NI RVDs flanking weak NK RVDs in a nonfunctional artificial TALE array were replaced with HD (Streubel *et al.*, 2012). It is likely that flanking strong RVDs might stabilize the TALE protein–DNA interaction

sufficiently to also allow weak RVD combinations without decreasing overall TALE activity. Quantitative assessment of TAL effector–DNA affinity and recognition has suggested additional complexities in the previously reported simple TAL effector binding code (Christian *et al.*, 2012, Meckler *et al.*, 2013). In the study by Meckler *et al*., (2013), the repeat type was the largest factor in affinity differences between TAL effectors. Designer TALEs (dTALEs) with an even distribution of 10 NG, HD and NN RVDs bound their targets with high affinity compared to those containing NI or NK. Although there was not a clear linear correlation of affinity to activation of a reporter gene, the contribution of particular RVDs to affinity was: NG > HD ~ NN >>NI > NK; while the following was observed for fold-activation: HD ≥ NG ≥ NN >> NI > NK. This work suggested that the average binding contribution for a single repeat is small, 1–4 kJ/mol.

It has been suggested that additional residues within the backbone of TAL effectors might contribute to DNA binding or specificity (Streubel *et al.*, 2012). Outside the designated repeat variable residues at positions 12 and 13, there exists variability amongst natural TAL effectors at positions 4 and 32 (Boch & Bonas, 2010). There is little evidence that these residues affect RVD specificity, but rather play a role in positioning the α–helices (Streubel *et al.*, 2012). The differences in amino acid composition in the regions outside of the repeats, between the different TAL effector backbones (e.g. AvrBs3, PthXo1, Tal1c, Hax3) also appear to have little effect on TALE binding, despite previous conjecture (Bogdanove *et al.*, 2010, Streubel *et al.*, 2012). Overall length of a TALE repeat array also influences efficiency. Early experiments have shown that the efficiency of TALE-controlled transcription drastically increases at a threshold of nine to ten repeats (Boch *et al.*, 2009). This is in agreement with later

molecular data in which a short TALE array with only 10.5 repeats containing either the RVD NN or NH at three fixed positions; only the NN-containing TALE was able to activate the reporter. When additional repeats were added to create a 17.5 repeat array, both the NN and NH-containing TAL effectors were active (Streubel *et al.*, 2012). However, there is little structural evidence to support a claim that TALE efficacy is dependent on repeat number, thus it remains unexplained (Bochtler, 2012).

Polarity effects. The directionality of TAL effector protein binding begs the question of whether affinity and specificity are evenly distributed across arrays. Several studies have suggested that RVD-nucleotide mismatches at the 5'end of both TALE and TALEN targets have greater effects than those made at the 3' end (Blount *et al.*, 2012, Christian *et al.*, 2012, Meckler *et al.*, 2013). Specifically, substitution of the first three or six bp of a target reduced TALE binding 15-fold and 370-fold, respectively; whereas the same number of substitutions at the 3' end had no significant effect (Meckler *et al.*, 2013). Another question surrounding TAL effector binding and specificity is the requirement for the conserved 5' T nucleotide of TALE targets. The same study provided an example in which a dTALE still bound targets containing A, C and G at position 0 both *in vitro* and *in vivo* (albeit with lower affinity), and efficiency was only slightly less than the standard T target (Meckler *et al.*, 2013). It may be possible to alter the requirement for a 5' T nucleotide by replacing the conserved tryptophan (W232) with another residue. There is some evidence to suggest that a proline replacement (W232P) at this position relaxes overall specificity for all four nucleotides (Doyle *et al*., unpublished). Efforts described in Chapter 3 of this dissertation contribute to a better understanding of 1) RVD specificity of NN versus NK to target G nucleotides, 2)

influence of strongly- versus weakly-binding RVDs on overall binding affinity, and 3) TALEN backbone architecture, in context with target spacer lengths.

Binding to methylated DNA. A major drawback for nuclease-mediated genome engineering that is the often overlooked is the sensitivity of DNA binding fusion proteins to cytosine methylation and other epigenetic marks that are considered to be ubiquitous in different cell types and across species. It has been reported that TALE DNA binding domains are affected by the presence of 5-methyl cytosine (5mC) at their endogenous targets (Bultmann *et al.*, 2012). This epigenetic modification is found in approximately 70% of CpG dinucleotides in mammalian and plant somatic and pluripotent cells, as well in CpA, CpT and CpC dinucleotides (Jaenisch & Bird, 2003, Vanyushin & Ashapkin, 2011, Ziller *et al.*, 2011). Moreover, regions of genes most often targeted for gene editing (i.e. promoters and proximal exons) contain extensive 5mC, particularly in CpG islands (Jones, 2012). The standard RVD HD is most frequently used to specify cytosine. However, in cases where cytosine is methylated, the structural data predict a clash between this 5-methyl group and the side chain of aspartic acid (D of HD). Hence, 5-methylation should work against this pairing. Therefore, strategies to overcome this limitation have been developed, and are focused on altering the specificity of TAL effectors to recognize methylated cytosine nucleotides.

Two groups have independently reported that both NG and N* (asterisk indicates a deletion at residue 13 in the repeat unit–see Figure 1-9) RVDs can accommodate 5mC *in vitro* and *in vivo* (Deng *et al.*, 2012, Valton *et al.*, 2012). Based on solved crystal structures containing TALE with NG RVDs bound to DNA targets, it was hypothesized that the lack of a side chain in Gly (of NG RVD) provides sufficient space to

accommodate the 5-methyl group of thymine (T) nucleotides (Deng *et al.*, 2012). This observation suggested that Gly13 of RVDs might also recognize 5mC due to its structural similarity to T, with the only difference being at a position of 5mC not involved in TALE repeat binding. TAL effector arrays were tested for their binding affinity for 5mC using electrophoretic mobility shift assays (EMSA), and only those containing NG RVDs in place of the standard HD to specify C were able to bind their cognate targets (Deng *et al.*, 2012). The RVD N* was also used accommodate 5mC as the first example of TALEN-mediated modification of a methylated locus in human cells (Valton *et al.*, 2012). In the consensus of all repeats, the variable residues (aa 12 and 13) are followed immediately by two conserved glycine ('RVD'GG) residues; thus N* is highly similar to NG except for a shortened RVD loop. The DNA binding capacity and activity of N*–containing TALEs remained unaffected by 5mC, whether present on one or both strands of the DNA (Valton *et al.*, 2012). The RVD HG has similar binding specificity to NG and might also be a suitable alternative to specify 5mC. Although this work aimed to extend the application of TAL effector-mediation genome manipulation in epigenetic modification and other specialized areas of research (*e.g.* cancer research), it has also revealed limitations for the applications of TALEs.

Off-target concerns using TALENs. A longstanding concern amongst users of nuclease technologies is the potential risk of toxicity effects associated with non-specific binding and cleavage, commonly referred to as off-targeting events. Some ZFNs have found to be toxic in living cells (Cornu *et al.*, 2008, Gabriel *et al.*, 2011, Miller *et al.*, 2007, Mussolino & Cathomen, 2011, Porteus & Baltimore, 2003, Szczepek *et al.*, 2007), but it is less well known for TALENs. Several reports have performed *in vivo* analyses

using deep sequencing of genomes treated with TALENs, and none have revealed the level of toxicity as seen with ZFNs (Garg *et al.*, 2012, Kim *et al.*, 2013, Li *et al.*, 2011). These findings are encouraging, but they are not nearly comprehensive enough to provide certainty than TALENs have less off-targeting. Future large-scale and comprehensive sequencing studies will be required to make that determination.

1.5. Dissertation Objectives

The work presented in this dissertation represents three phases of research: the development, engineering, and application of TAL effector nuclease for targeted genetic manipulation. Unlike ZFNs and meganucleases, the one-to-one correspondence of a TAL repeat to a single DNA base constitutes a simple code that can be used to design a TAL effector to recognize *any* sequence in a given genome.

The first aim of this dissertation was the proof-of-principle work demonstrating that fusions of different naturally occurring TAL effector proteins to FokI nuclease were capable of binding and cleaving specific DNA targets. Previous work with well-characterized DNA binding domains showed that fusion to a non-specific nuclease creates functional proteins whose cutting activity can be directed to a specific sequence of DNA. This dissertation shows that fusion of TAL effector DNA binding repeats to the FokI nuclease can direct a double-strand break to a desired DNA sequence, thus making TALENs an exciting new tool for genetic manipulation. Second, we sought to develop a method to efficiently engineer custom TALENs to recognize novel DNA target sequences and test their cleavage activity. To realize the potential of TALENs for gene modification, it was necessary to develop a method to quickly and efficiently construct

the TAL effector DNA biding arrays in any desirable order. The method presented here allows users to easily and efficiently assemble TAL effector repeat domains to create functional TALENs. Furthermore, this system was made publically available through a non-profit plasmid repository (Addgene) to any researcher or institution that desires to use it. Another objective was to elucidate properties of TAL effector - DNA recognition and binding to inform both the biology of protein-DNA interactions and the design of TALEN for applied purposes. The final aim was to implement and assess engineered TAL effector nucleases for site-specific genome manipulation, specifically to achieve targeted gene disruption in plants. The breadth of application of any new technology must be rigorously tested to determine its usefulness to the field. Therefore, we assessed the ability of TALENs to create desired genomic modifications in relevant model systems, in particular important plant models like tobacco and *Arabidopsis thaliana.* These proof-of-concept experiments will likely serve as a framework for enabling the TALEN technology in plants.

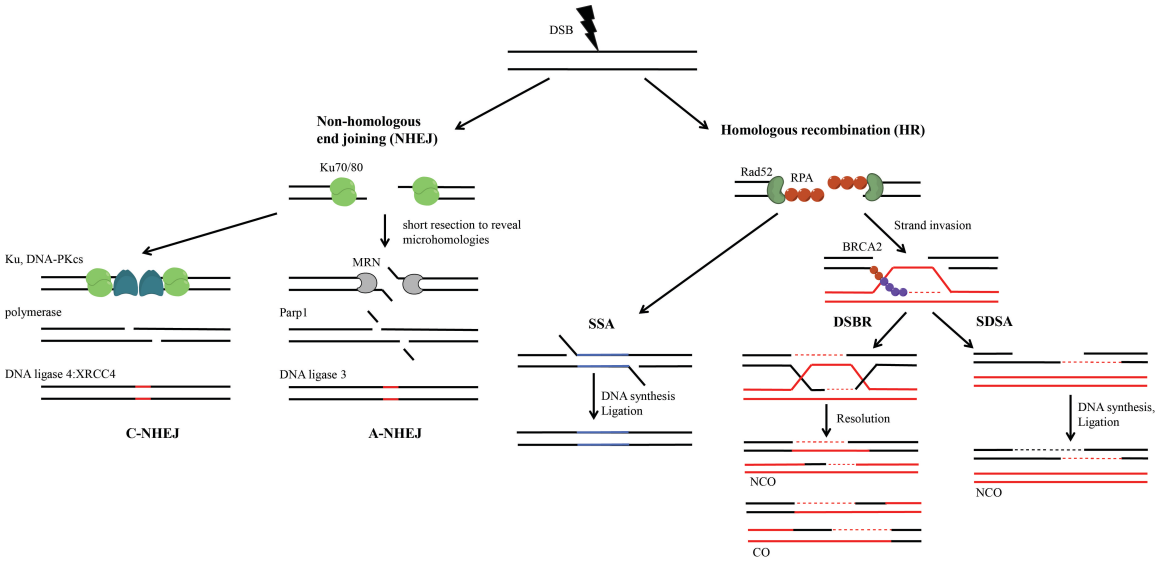


Figure 1-1. Outcomes of common DNA double-strand break repair pathways. A DSB is introduced by a sequence-specific nuclease. Repair through nonhomologous end-joining (NHEJ) can occur through different subpathways. In classic NHEJ (cNHEJ) DNA ends are recognized by the Ku70/Ku80 heterodimer and DNA-dependent protein kinase catalytic subunit (DNA-PKcs). Microhomology is not required, and the break is ligated by DNA ligase4:XRCC4 complex. Alternative NHEJ (aNHEJ) is Ku independent, and free DNA ends are resected by MRN complex to reveal microhomologies that can anneal to repair breaks. Repair by NHEJ can lead to small insertions or deletions at the break site (red lines). Homologous recombination (HR) uses existing sequence homology to repair DSBs. Resected ssDNA ends are coated with RPA proteins and RAD proteins (*e.g. RAD52*) are recruited. RAD51 and BRCA2 are involved in homology search and strand invasion. Single-strand annealing (SSA) involves annealing of homologous single-strand ends (blue region) resulting in deletion of sequence between repeats. Classic DSB repair (DSBR) involves 5' resection to reveal 3' ends responsible for invading homologous regions on sister chromatids to form a D-loop. When both ends invade resolution can

yield non-crossover (NCO) or crossover (CO) products (gene conversion). In synthesis-dependent strand annealing (SDSA), the invading strand is displaced; DNA synthesis and ligation yields NCO products. Abbreviations: MRN complex, Mre11–Rad50–Nbs1 complex; RPA, Replication protein A; RAD51, RADiation sensitive 51 protein; BRCA2, breast cancer type 2 susceptibility protein. Adapted from (San Filippo *et al.*, 2008).

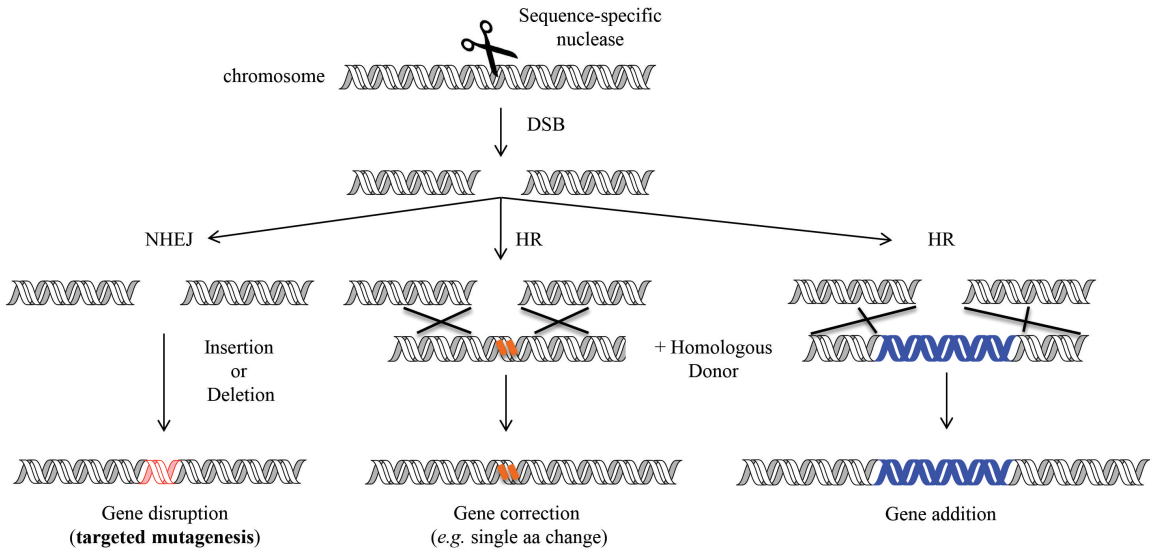


Figure 1-2. Targeted genome modifications using DNA double-strand breaks. Shown Possible genome engineering applications based on the generation of a sequence-specific DNA break by engineered nucleases are shown. Gene disruption or targeted mutagenesis is achieved when the broken DNA are ligated back together imperfectly (via NHEJ), resulting in small insertions or deletions that alter gene function. Alternatively, the HR pathway can be influenced by adding a user-designed donor DNA that contains small point mutations or single aa changes (gene correction) or larger segments of new DNA, *e.g.* novel open reading frame (ORF) (gene addition). Adapted from (Podevin *et al.*, 2012)

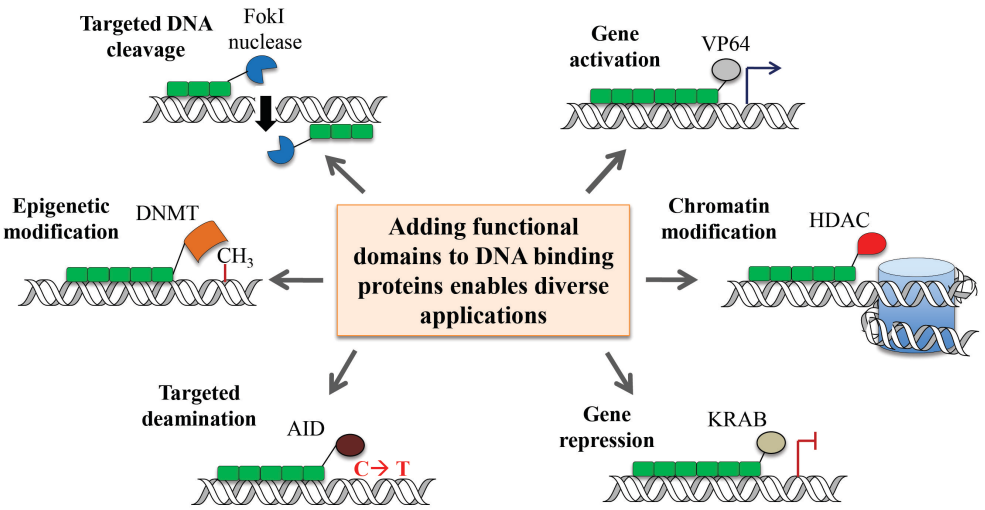


Figure 1-3. Adding functional domains to DNA binding proteins enables diverse applications. Modular DNA binding proteins can be engineered to recognize new sites in the genome. These protein arrays can be fused to various effector domains in order to direct a particular gene modification event to a specific location in the genome (*e.g.* DNA breaks, gene activation, repression, methylation, acetylation, deamination). Abbreviations: DNMT, DNA methyl transferase; AID, activation-induced deamination; KRAB, repressor domain; VP64 activation domain. Adapted from (Jamieson *et al.*, 2003).

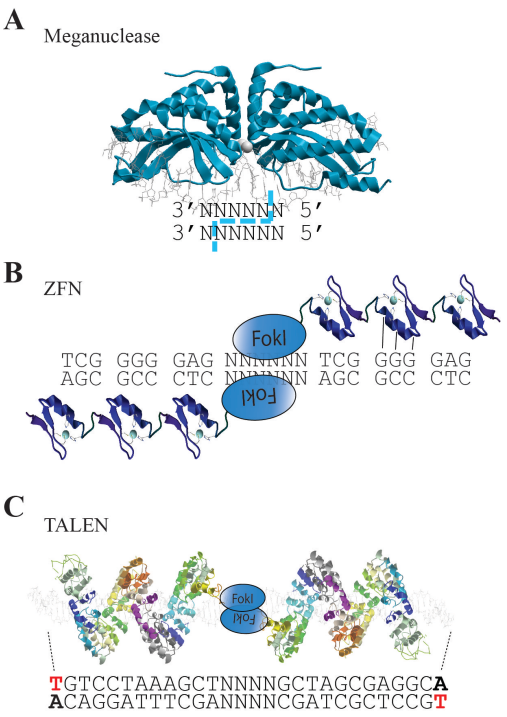


Figure 1-4. Types of custom nucleases. (A) Structural representation of a homing endonuclease (HE) or meganuclease (MN) bound to DNA. The catalytic domain determines specificity and provides nuclease function. Shown below is a schematic of the type of cleavage overhangs created by HEs and MNs. (B) Structural representation of a zinc-finger nuclease (ZFN) pair, composed of two ZFN monomers. A monomer is made of three to six zinc finger arrays (ZFAs), with each unit providing specificity for a DNA triplet. The ZFAs are fused to the catalytic domain of FokI endonuclease (blue). ZFNs are positioned in an inverse orientation on the target to allow for FokI dimerization and cleavage. (C) Structural representation of two transcription activator-like (TAL) effector nucleases (TALENs) bound to DNA. Each array is composed of a series of repeats. Repeat variable di-residues (RVDs) in each repeat make specific DNA contacts. A target sequence is shown below with the conserved 'T' base highlighted in red.

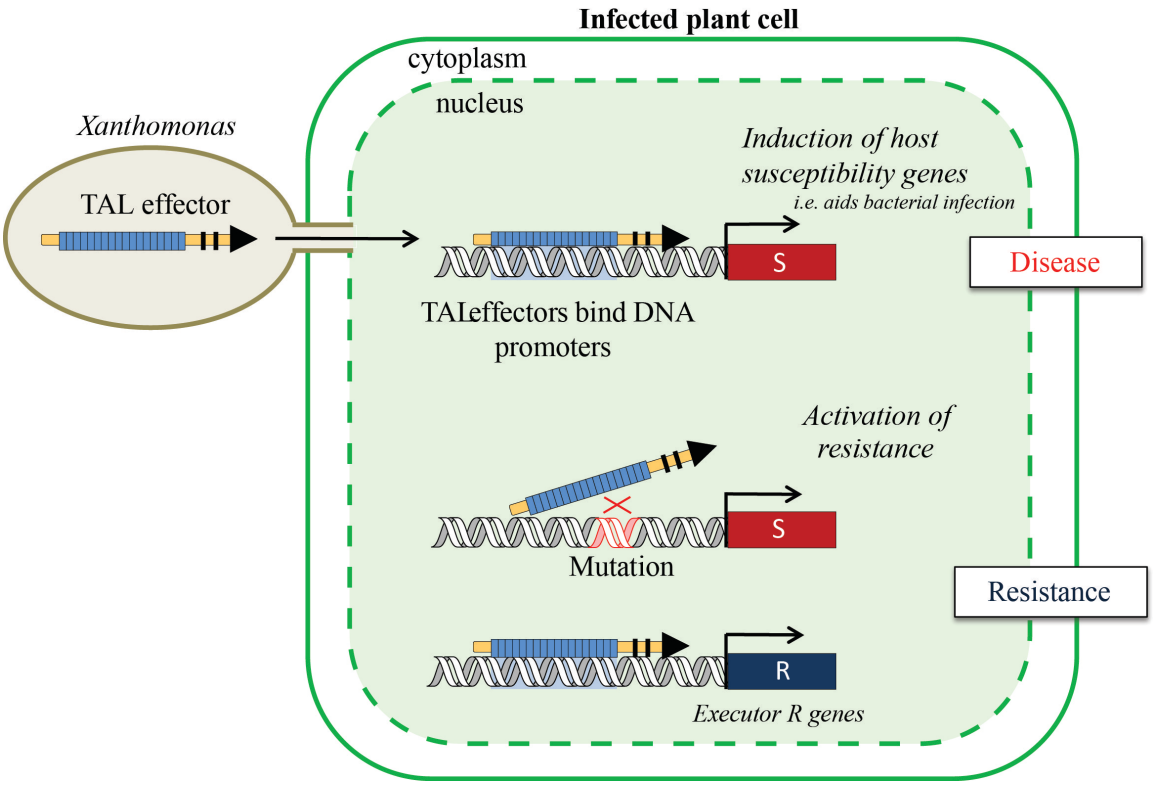


Figure 1-5. Transcription Activator-like effectors (TALEs). TAL effector proteins are derived from a species of plant pathogenic bacteria called *Xanthomonas*. Upon infection, bacteria inject TALEs into host plant cells via the type III secretion system. Once in the nucleus, TALEs bind the promoters of susceptibility (S) genes (red) that help facilitate colonization and promote disease. TALEs can avoid activation of S genes through accumulation of mutations in plant gene promoters normally recognized. Plants have also evolved mechanisms for resistance, such as executor R genes (blue) that are under transcriptional control of TALE target sequences that would normally be present only in S gene promoters. In both cases, plants can become resistant to infection by certain TALEs. Figure adapted from (Bogdanove *et al.*, 2010a).

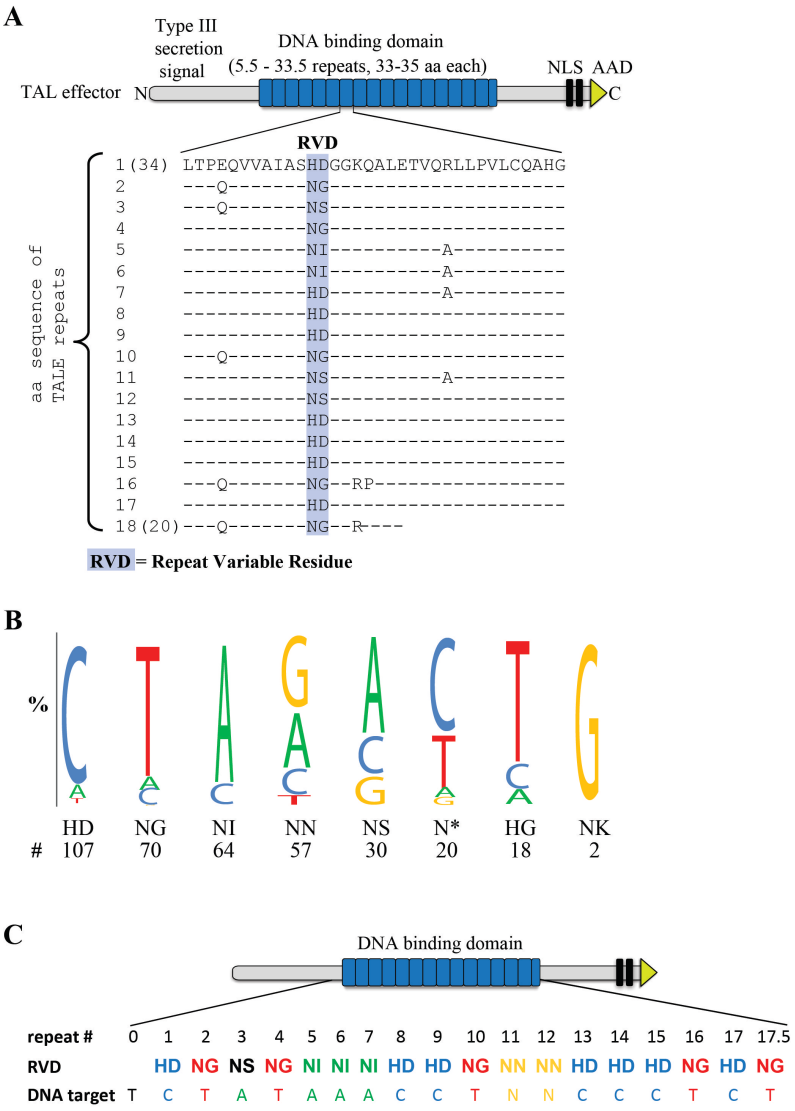


Figure 1-6. Schematic of TAL effector and the RVD – DNA code. (A) Schematic of a TALE effector protein. Central region of repeats (blue squares) specifies DNA binding. The nuclear localization signals (NLS) and acidic activation domains (AAD) are also depicted. The aa sequences for individual repeats of a representative array are shown and repeat variable diresidues (RVDs) are highlighted in blue. Dashes indicate conserved aa amongst repeats in the array. (B) RVD-nucleotide association frequencies. Logo size reflects frequency of RVD association indicated below, based on the number (#) of

associations analyzed in (Moscou & Bogdanove, 2009). (C) Schematic of a TALE repeat array depicts repeat position, RVD identity and specified nucleotide target. A conserved ‘T’ base is depicted at the ‘0’ repeat position.

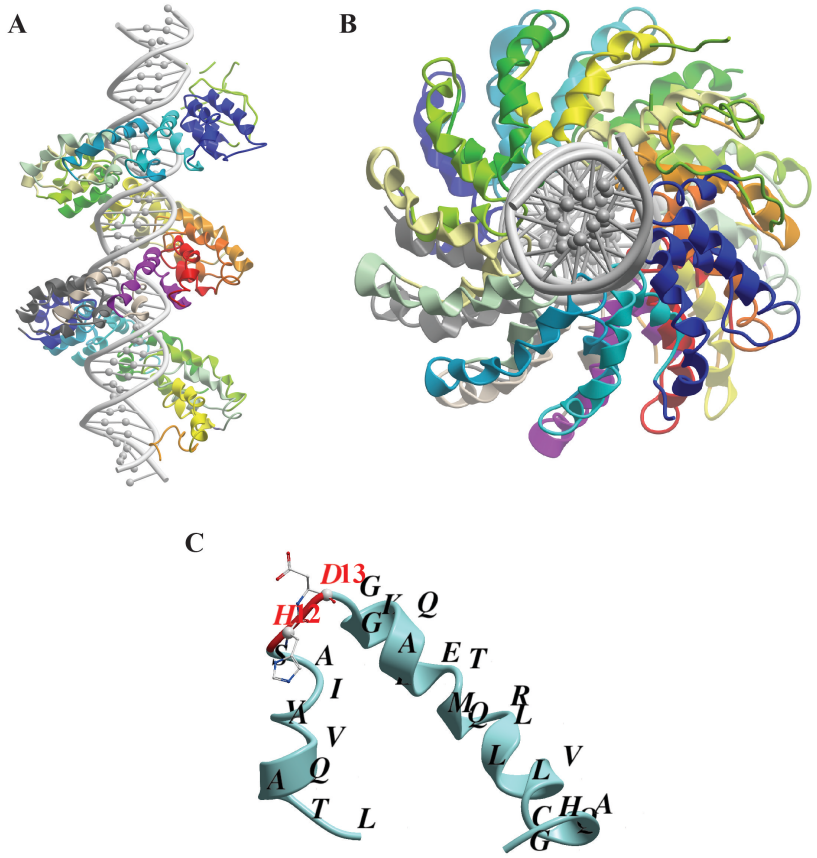


Figure 1-7. Crystal structure of a TAL effector bound to DNA. The solved crystal structure of the PthXo1 TAL effector DNA binding region in complex with its target site. (A) side view (B) top view. PthXo1 contains 23.5 canonical repeats. The coloring reflects individual TAL repeats. (C) The structure of a representative repeat. The 'HD' RVD residues (aa positions 12 and 13) that recognize a cytosine are highlighted in red. All images were rendered using MolSoft ICM browser software, 2011.

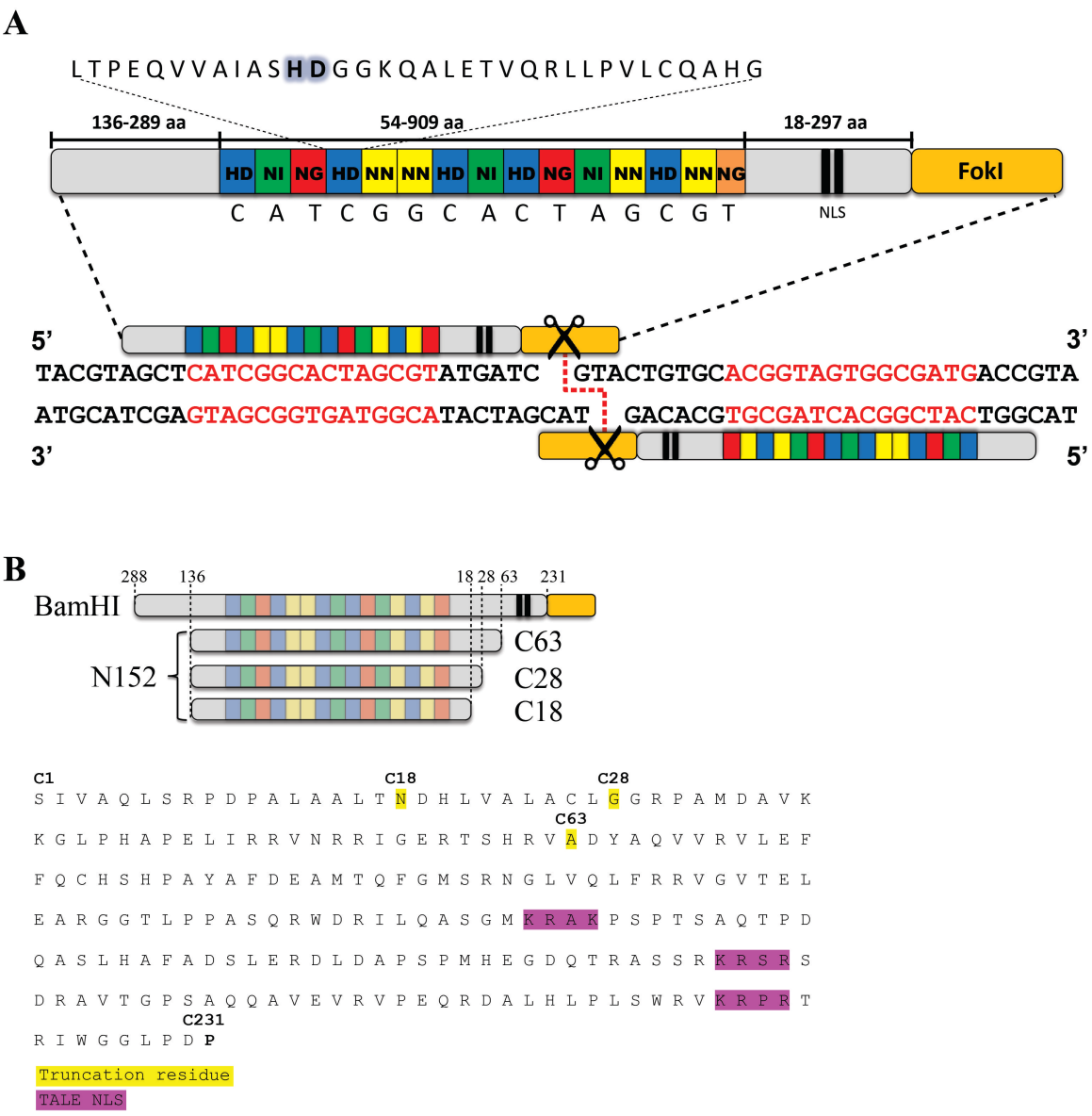


Figure 1-8. Schematic of TAL effector nuclease and architectures. (A) A single TAL effector nuclease is shown above a TAL effector nuclease (TALEN) pair. Length of TAL effector segments are indicated, along with a representative TAL effector array and its cognate nucleotide target. The orientation of a full TALEN pair is shown with the full TALEN target sequence highlighted in red. (B) Truncated TALEN backbone architectures. A series of truncations were made to the N-terminal and C-terminal ends of

the TALE protein. N152 represents removal of aa up to residue 152 in the N-terminal portion, while truncations made to the C-terminus are indicated in the aa sequence provided below (*e.g.* C63 truncation includes the first 63 aa after the last repeat). Repeats corresponding to four RVD types are color-coded: HD, blue; NI, green; NN, yellow; NG, red, and specify C, A, G, and T, respectively.

MDPIRPRRPSPARELLPGPQPDRVQPTADRGVSAPAGSPLDGLPARRTVSRTRLPSPPA
PSPAFSAGSFSDLLRPFDPSLLDTSLLDSMPAVGTPHTAAAPAEWDEAQSALRAADDPP
PTVRVAVTAARPPRAKPAPRRRAAQPSDASPAAQ**VDLRTLGYSQQQQEKIKPKVRSTVA**
QHHEALVGHGFTHAHIVALSQHPAALGTVAVTYQHIITA
-1 **LPEATHEDIVGVGKQWSGARALEALLTDAGELRGPPLQ**
0 **LDTGQLVKIAKRGGVTAMEAVHASRNALTGAPLN**
1 LTP**A**QVVAIASNIGGKQALETVQRLLPVLCQDHG
2 LTPDQVVAIASN*GGKQALETVQRLLPVLCQDHG
3 LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG
4 LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG
5 LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG
6 LTPDQVVAIANNNGGKQALETVQRLLPVLCQDHG
7 LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG
8 LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG
9 LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG
10 LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG
11 LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG
11.5 LTPDQVVAIASHDGGKQALE
SIVAQLSRRDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPEFIRRVNRRIAERTS
HRVADYAHVVRVLEFFQCHSHPAHAFDEAMTQFGMSRHGLVQLFRRVGVTEFEARYGTL
PPASQRWDRILQASGMKRAKPSPTSAQTPDQTSLHAFADSLERDLDAPSPMHEGDQTRA
SSRKRSRSDRAVTGPSAQQAVEVRVPEQRDALHLPLSWRVKRPRTRIWGGLPDPISRS
QLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKH
LGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINP
NEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKA
GTLTLEEVRRKFNNGEINF

N152 aa; C63 aa
W232
cryptic repeats
FokI

Figure 1-9. Amino acid sequence and features of a TALEN. Amino acids that constitute the amino (N) and carboxy (C) terminal regions are separated from the repeats (indented). The position of each repeat within the array is indicated to the left of the repeat sequence. RVDS are underlined. Residues included the truncated form of TALEN

are indicated. The FokI aa sequence is highlighted in yellow. Other features are highlighted.

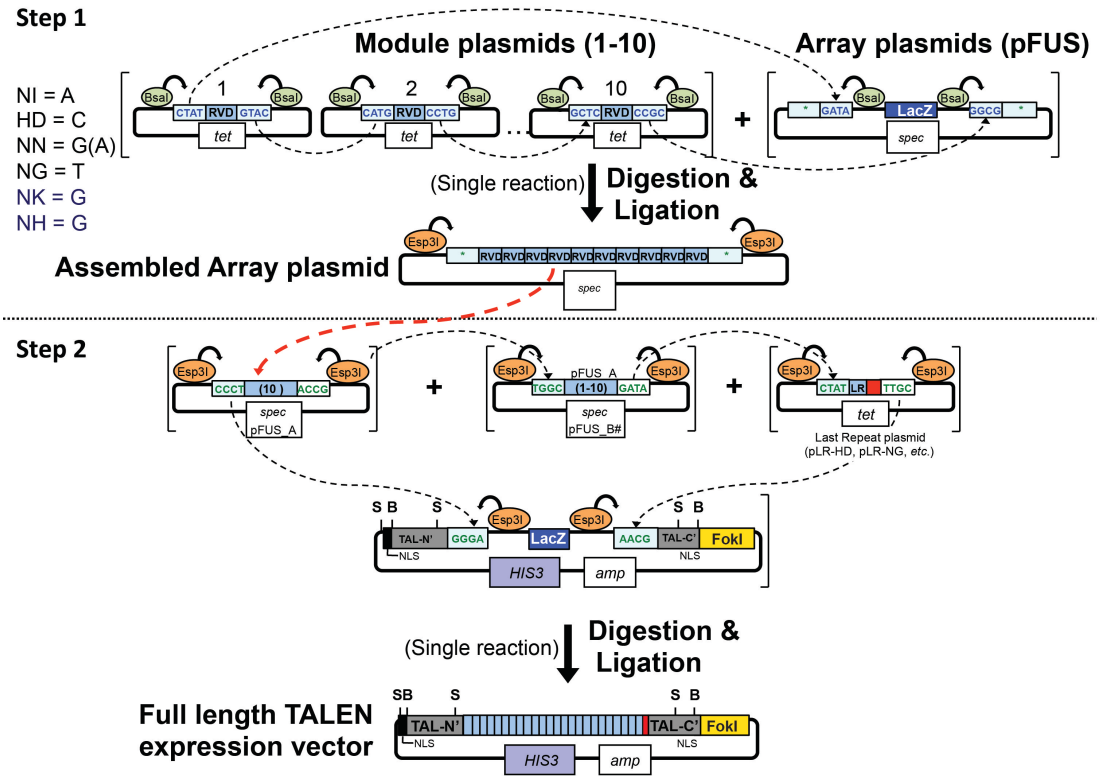


Figure 1-10. Golden gate cloning method for assembly of TAL effector arrays. The Golden Gate method can be used to assemble TAL effector arrays. Type IIS restriction endonucleases (*e.g. Bsa*I and *Esp*3I) allow users to rapidly clone modules containing the desired RVDs (light blue boxes); RVDs can be released with unique cohesive ends for ordered, single-reaction assembly into array plasmids in a first step, and those arrays are subsequently released and assembled in sequential order in a second step into a backbone plasmid to create full length constructs with custom repeat arrays (see Chapter 3 for details). (s); AD, transcriptional activation domain; *tet*, tetracycline resistance; *spec*, spectinomycin resistance; *amp*, ampicillin resistance; *HIS3*, histidine marker gene; NLS, nuclear localization signal; B, *Bam*HI, and S, *Sph*I, useful for subcloning custom repeat arrays. Unique restriction enzyme sites flanking the coding sequences, useful for subcloning the entire constructs into other vectors, are not shown.

Table 1-1. Summary of Published Studies Related to TAL effector Technologies

Functional Domain	**Organism**	**Target or Genes**	**Assembly Method** GG, standard cloning, solid-phase (FLASH, LIC)	**Modification Type**	**Reference**	**Year**
ACTIVATOR						
	pepper leaves	*Bs3* resistance gene	GG-Voytas	activation, measured by HR response (phenotypic)	Cermak	2011
	tobacco; Human	GUS activation; PUMA, IFNa1, IFNb1	GG (Golden TAL tech.)	Activation	Geissler (J.Boch)	2011
	Human	-	Sangamo	Activation	Miller	2011
		-	GG	Activation	Zhang	2011
		-	GG	Activation	Bultman	2012
	e.coli	-	GG	Activation	Politz	2012
		-	GG	Activation	Cong (F.Zhang)	2012
		-	GG	Activation	Li, L.	2012
		-	GG	Activation	Tremblay	2012
	Human (HeLa)	*ROCK1*	GG	Activation	Wang	2012
	N.benthamiana	-	GG	Activation	Streubel	2012
	Human	-	FLASH	Activation	Maeder	2013
		-	GG	Activation	Perez-Pinera	2013
		-	GG	Activation	Li, T. (B.Yang)	2013
REPRESSOR	*S. cerevisiae*	PFY1p; syntehtic promoter	Standard cloning	repression	Blount, BA.	2012

	Arabidopsis	RD29A::LUC transgene; *RD29A*	GG	repression	Mahfouz	2012
	Human	*SOX2, CACNA1C*	GG (PCR-based)	activation, repression	Cong (F.Zhang)	2012
	Human	activation of CFP reporters	standard cloning;also have aglorithm	activation, repression of reporters	Garg	2012
NUCLEASE	**Endogenous genes modified by TALENs**					
Nuclease	Human, Arabidopsis	*HPRT, ADH1*	GG-Voytas	NHEJ	Cermak	2011
Nuclease	*C.elegans*	*ben-1*	Sangamo	NHEJ	Wood	2011
Nuclease	Human, ES, iPSCs	*OCT4, PITX3, PPP1R12C (AAVS1)*	Sangamo	HDR	Hockemeyer	2011
Nuclease	Human	*CCR5, IL2RG*	Golden Gate - Lahaye	NHEJ	Mussolino, C.	2011
Nuclease	Human cells	many	GG	NHEJ/HR	Kim, H. (Kim, J. S.)	2011
Nuclease	Human	*NTF3, CCR5*	Sangamo	NHEJ/HDR (CCR5)	Miller, J.	2011
Nuclease	Rat	*IgM*	Sangamo	NHEJ	Tesson, L.	2011
Nuclease	Yeast	ade2, lys3, ura3	Golden Gate - Yang	NHEJ/HDR (*ura3*)	Li, T.	2011
Nuclease	zebrafish	*gria3a, hey2*	REAL	NHEJ	Sander, J. D.	2011
Nuclease	zebrafish	*tnikb, dip2a*	Standard Cloning - B.Zhang	NHEJ	Huang, P.	2011
Nuclease	Bovine/ Porcine	*GGTA, ACAN, GDF8, PRNP/SRY, RELA (p65), DMD, AMELY, LDLR, RAG2, IL2Rg, GHRHR, GGTA, GDF8*	GG-Voytas, Goldy	NHEJ	Carlson	2012
Nuclease	cricket	*Gb'lac2*	Cellectis	NHEJ	Watanabe	2012

Nuclease	Fruit fly	*yellow, (CG9797)*	Standard Cloning - B.Zhang	NHEJ	Liu, J.	2012
Nuclease	Human	*AXIN2, BRCA1, CDC73, CHD7, CYLD, ERCC2, JAK2, MYCN, NCOR2, NHN, XPC (84 total, see table?)*	FLASH	NHEJ	Reyon, D.	2012
Nuclease	Human	*MAOA*	Solid Phase - Wang	NHEJ	Wang, Z.	2012
Nuclease	Human	PPP1R12C (AAVS1)	Golden Gate - Zhang	NHEJ	Sanjana, N. E.	2012
Nuclease	Human	*HBB*	Golden Gate - Voytas	NHEJ	Sun, N.	2012
Nuclease	Rat ES cells	BMPR2	Golden Gate - Voytas	NHEJ	Tong, C.	2012
Nuclease	Rice	*Os11N3*	Golden Gate - Yang	NHEJ	Li, T. (B.Yang)	2012
Nuclease	silkworm	*BmBlos2*	Gene Synthesis (Zhang)	NHEJ	Ma, S.	2012
Nuclease					Lei, Y.	2012
Nuclease	Zebrafish		Golden Gate - Voytas		Bedell, V. M.	2012
Nuclease	Zebrafish	*elmo1, epas1b, fh, hif1ab, ptmt1, scl6a3*	FLASH	NHEJ	Cade, L.	2012
Nuclease					Moore, F. E.	2012
Nuclease					Dahlem, T. J.	2012
Nuclease					Sajwan, S.	2012
Nuclease	Human, CHO-KI	*XPC1*	Cellectis	NHEJ; TM	Valton, J.	2012
Nuclease					Ding, Q.	2012
Nuclease					Briggs, A. (Church)	2012
Nuclease	CHO cells	*FUT8*		NHEJ, Knock-in	Cristea, S.	2012
Nuclease					Schmid-Burgk, J.	2012

Nuclease					Shan, Q.	2013
Nuclease					Mashimo, T.	2013
Nuclease		*CXADR, CFTR, AAVS1, ERCC5, CDH1, HOXD13, FANCE, KIT, TGFBR2*	GG-Voytas	NHEJ	Lin, Y.	2013
Nuclease			GG-Voytas	NHEJ, germline	Ansai, S.	2013
Nuclease					Sakuma, T.	2013
Nuclease	*Nicotiana tabacum* protoplasts	*SurA, SurB*	GG-Voytas	NHEJ, GT	Zhang, Yong	2013
Nuclease					Kim, Y. (J.S. Kim)	2013
Nuclease					Chen, S	2013
Nuclease		*tyrosine hydroxylase (th); fam46c; smad5*		GT, HR	Zu, Y.	2013
Nuclease	mouse	*chocolate (Rab38)*		GT w/oligos	Wefers, B.	2013
Nuclease			GG-Zhang		Stroud, DA.	2013
Nuclease		*F8, F9, CCR5*		NHEJ	Kim Y. (JS Kim)	2013
Nuclease		*apoea, flt4 (*TALENs*); slc24a5 (golden)*-ZFN *and lepa*-TALEN; LCR for *globin* locus; *linc-birc6 (megamind*) locus	**GG-Voytas;** used N- and C-term backbones from Reyon *et al*. REAL assembly	NHEJ, large deletions/inversions	Gupta, A. (S.Wolfe)	2013
Nuclease	zebrafish	Cftr	GG-Voytas; Miller architecture	NHEJ, KO	Navis, *et al*.	2013
Nuclease	*Aedes aegypti* (mosquito)	*kmo* (eye pigmentation)	Cellectis	NHEJ	Aryan, *et al*.	2013
Nuclease	mouse	*Zic2 (curly tail)*	GG-Voytas	NHEJ	Davies, *et al*.	2013

Nuclease	zebrafish	*s1pr2, s1pr5a*		NHEJ	Hisano, Y.	2013
	Other TALE, TALEN studies					
	TAL code				Moscou, Bogdanove	2009
	TAL code				Boch, J. *et al*.	2009
Nuclease					Christian	2010
Nuclease					Li, T.	2010
Nuclease					(Mafouz)	2010
Nuclease					Morbitzer	2011
Nuclease					Zhang, F.	2011
Structure					Mak	2012
Structure					Deng	2012
Nuclease					Christian, ML.	2012
Nuclease			GG	NHEJ, HR, KI	Maresca, M.	2013
					Neff, KL.	2013
Binding assay					Meckler, J. (Segal)	2013
					Grau at al. (Boch)	2013
					Mendenhall, *et al*. (Joung on paper)	2013
Recombinase					Mercer, A.	2012

Table 1-2. Summary of TAL effector Engineering Methods

	GG	PCR toolbox	REAL and REAL-FAST	Unit Assembly (UA)	FLASH	Iterative capped assembly	LIC	ULtiMATE	EZ-TAL effector kit	Academic Core Facilites	Companies selling TALENs
	(Cermak *et al*.	Sanjana *et al*.	Sander *et al*.,	Huang *et al*.	Reyon *et al*.	Briggs *et al*.	(Schmid-Burgk *et al*.	Yang *et al*.	Uhde-Stone *et al*.,	University of Utah	Cellectis (7,200 TALENs per year)
Lab	Voytas Lab		Joung Lab	Bo Zhang lab	Joung Lab	George Church group				University of Wisconsin - Madison	Excellgen
Fidelity	colony PCR: 50-99%	N/A	standard cloning	standard cloning	colony PCR: N/A	Solid-Phase	colony PCR: 80%	colony PCR: 100%	colony PCR: 40-90%	UCSF ES Cell Targeting Core	GeneCopoeia
	sequencing : 100%	na			sequencing : 8-100%		sequencing : 100%	sequencing: 66%	sequencing: 80-90%		Life Technologies/ GeneArt
Throughput /day	10	low	mid	mid	96 (high)	high	192/1536 high	unknown	N/A		PNA Bio
Turnover time	5 days	3-4 days			3 days	2-5	2-4 days	3 days	3 days		System Biosciences
Includes PCR	no	yes	yes	no	yes	yes	no	yes	yes		Transposagen
Includes ligation	yes	yes	yes	NheI and SpeI	yes	-	no	yes	yes		
Includes gel purification step	no	yes	yes	yes	PCR cleanup	-	Yes, for BKB stocks	yes	yes		
RVDs	NI, HD, NN, NG, NK, NH	NI, HD, NN, NG	NI, HD, NN, NG	NI, HD, NN, NG	NI, HD, NN, NG, NH, NK	NI, HD, NN, NG	NI, HD, NK, NG	NI, HD, NN, NG	NI, HD, NN, NG		
source library	60 RVD modules	4 RVD monomers	-	-	376 TALE multimer plasmids	-	64 x 2-mer plasmids	64 x 3-mer plasmids	72 monomers		

	13 array plasmids	8 BKB plasmids	-	-	4 backbone plasmids	-	4 x 1-mer plasmids	2 x backbone plasmids	4 backbones		
	4 LR plasmids	12 total	28 plasmid (REAL)	376 (REAL-Fast)	380 total	-	7 backbone plasmids	66 total			
	6 backbone plasmids						75 total				
	83 total						(or 3072 x 5-mers)				
architecture	BamHI and d152/+63	d152/+63	d152/+63	single-unit vectors; Simple double-digestion	d152/+63	d152/+63	d152/+63	unknown	only TFs		
Flexibility (RVDs)	11.5 – 30.5	1 - 24.5			8.5 – 20.5		9.5 – 18.5	17.5	12.5 – 18.5		
price of kit	$425	$325	Addgene (REAL); by request (REAL-Fast)	by request	N/A		$350	N/A	$1480 (including primers – for 6-12 assemblies)		
Sequencing primers not included	-	$180	-	3' term of *pthA* from *Xanthomonas axonopodis pv. citri*	$8	-	-	$477	-		

CHAPTER 2

TARGETING DNA DOUBLE-STRAND BREAKS WITH TAL EFFECTOR NUCLEASES

Introduction

Engineered nucleases that cleave specific DNA sequences in vivo are valuable reagents for targeted mutagenesis. Here we report a new class of sequence-specific nucleases created by fusing transcription activator-like effectors (TALEs) to the catalytic domain of the FokI endonuclease. Both native and custom TALE-nuclease fusions direct DNA double-strand breaks to specific, targeted sites.

Results and Discussion

Zinc finger nucleases (ZFNs) and meganucleases cleave specific DNA target sequences in vivo and are powerful tools for genome modification (D Carroll, 2008, Cathomen & Joung, 2008, Galetto *et al.*, 2009). Chromosome breaks created by these engineered nucleases stimulate homologous recombination or gene targeting: in the presence of a template for repairing the double-strand break, specific DNA sequence changes in the template become incorporated into chromosomes at or near the break site. In the absence of a repair template, broken chromosomes are rejoined by non-homologous end joining, which often introduces short DNA insertions or deletions that create targeted gene knockouts. For both ZFNs and meganucleases, a barrier to their widespread adoption has been the challenge in engineering new DNA binding specificities. While significant progress has been made in recent years, generating the necessary reagents to modify new DNA targets is still resource intensive and to some degree empirical (Fajardo-Sanchez *et al.*, 2008, H. J. H. Kim *et al.*, 2009, Maeder *et al.*, 2008, Urnov *et al.*, 2005). A novel DNA binding domain was recently described in a family of proteins known as transcription activator-like effectors (TALEs) (Boch *et al.*,

2009, Moscou & Bogdanove, 2009). TALEs are produced by plant pathogens in the genus *Xanthomonas*, which deliver the proteins to plant cells during infection through the type III secretion pathway (Bogdanove *et al.*, 2010b). Once inside the plant cell, TALEs enter the nucleus, bind effector-specific DNA sequences, and transcriptionally activate gene expression. Typically, activation of target genes increases plant susceptibility to pathogen colonization, but in some cases, it triggers plant defense. TALE binding to DNA is mediated by a central region of these proteins that contains as many as 30 tandem repeats of a 33- to 35-amino-acid-sequence motif (Figure 2-1A). The amino acid sequence of each repeat is largely invariant, with the exception of two adjacent amino acids (the repeat variable diresidue or RVD). Repeats with different RVDs recognize different DNA base pairs, and there is a one-to-one correspondence between the RVDs in the repeat domain and the nucleotides in the target DNA sequence, constituting a straightforward cipher (Boch *et al.*, 2009, Moscou & Bogdanove, 2009). Using this cipher, targets of new TALEs have been correctly predicted, and functional targets of TALEs composed of randomly assembled repeats have been generated (Boch *et al.*, 2009, Moscou & Bogdanove, 2009, Römer *et al.*, 2010). The ability to predict the DNA binding specificity of native or artificial TALEs suggests a variety of applications for these proteins in the targeted modification of genomes. In particular, the DNA recognition domain could direct a fused nuclease to a specific DNA sequence. Further, the apparent modularity of the repeats should enable rapid construction of TALE nucleases (TALENs) with novel specificities to target double-strand breaks at specific locations in the genome. To test this idea, we modified TALEs by adopting the molecular architecture used for ZFNs. ZFNs function as dimers, with each monomer comprising a

DNA binding domain (a zinc finger array) fused to the catalytic domain of the FokI restriction enzyme (D Carroll, 2008, Cathomen & Joung, 2008). Two zinc finger arrays are engineered to recognize target sequences in the genome (each typically 9–12 bp) that are separated by a spacer sequence (typically 5–7 bp). Binding of the zinc finger arrays to the targ*et al*lows FokI I to dimerize and create a double-strand break within the spacer. We reasoned that the zinc finger arrays could be substituted with the DNA recognition domain of TALEs to create TALENs that recognize and cleave DNA targets (Figure 2-1A). To assess TALEN function, we adapted a yeast assay in which LacZ activity serves as an indicator of DNA cleavage (Townsend *et al.*, 2009). In this assay, a target plasmid and a TALEN expression plasmid are brought together in the same cell by mating. The target plasmid has a LacZ reporter gene with a 125-bp duplication of coding sequence. The duplication flanks the TALEN target site. When a double-strand DNA break occurs at the target site, it is repaired through single-strand annealing between the duplicated sequences, creating a functional LacZ gene whose expression can be measured by standard assays. We first used two well-characterized TALEs, namely AvrBs3 from the pepper pathogen *Xanthomonas campestris pv. vesicatoria* and PthXo1 from the rice pathogen *X. oryzae pv. oryzae* (Bonas *et al.*, 1989, Yang *et al.*, 2006). Both TALE-encoding genes have a Bam HI restriction fragment that encompasses the coding sequence for the repeat domain and 287 aa prior and 231 aa after (Figure 2-1A). Absent from the Bam HI fragment is the TALE transcriptional activation domain, which was replaced with a FokI nuclease monomer. Because the FokI monomers must dimerize to cleave, it was unclear what an appropriate spacer length between the two DNA recognition sites might be. For ZFNs, in which the zinc finger array is separated from

FokI by a 4- to 7-amino-acid linker, the typical spacer between the two recognition sites is 5–7 bp. Since 235 amino acids separate the repeat domain from FokI in our TALEN constructs, we selected a 15-bp spacer to separate the two recognition sites, assuming that a larger spacer may be required to enable FokI to dimerize. As a positive control we used a well-characterized zinc finger nuclease with a DNA binding domain derived from the mouse transcription factor Zif268 (Porteus & Baltimore, 2003). As negative controls, the TALE domains were fused to a catalytically inactive FokI variant or tested against non-cognate DNA targets. Robust activity was observed for both the AvrBs3 and the PthXo1 TALENs (Figure 2-1B). The activity of the PthXo1 TALEN approximated that of the ZFN positive control. We next varied the distance between the TALE binding sites (11 length variants between 12 and 30 bp) to identify spacer lengths that enable FokI to dimerize most efficiently (Figure 2-1C). Both enzymes showed two spacer length optima—one at 15 bp and the other at either 21 bp (AvrBs3) or 24 bp (PthXo1). For PthXo1, activity was observed for all tested spacer lengths 13 bp and longer. Some spacer lengths for AvrBs3 showed no activity, however, suggesting that spacer length is critical for certain TALENs. The above experiments tested activity of homodimeric TALENs, which bind two identical recognition sequences placed in opposition on either side of the spacer. Since such palindromic sites are unlikely to occur naturally in genomic targets, we tested whether TALENs could function as heterodimers. AvrBs3 and PthXo1 recognition sites were placed on either side of a 15-bp spacer (Figure 2-1D). The resulting activity of the heterodimeric TALEN approximated an average of the activities observed with the two homodimeric enzymes. To realize the potential of TALENs for genome modification, it will be necessary to engineer them to recognize novel

chromosomal DNA sequences. Randomly assembled repeat domains have been shown to bind DNA targets predicted by the cipher (Boch *et al.*, 2009); however, custom TALEs designed to recognize new target sequences have not been reported. This is an important distinction, because the former experiments validate the cipher, whereas the latter would demonstrate that DNA binding domains can be engineered with the requisite specificity for in vivo manipulations. To test whether repeat domains can be assembled to target TALENs to arbitrary chromosomal sequences, we chose two genes previously targeted for mutagenesis with ZFNs—*ADH1* from Arabidopsis and *gridlock* from zebrafish (Foley *et al.*, 2009, Zhang *et al.*, 2010). We first searched for 12- to 13-bp sequences in the coding regions, preceded by a 5'-T and with a nucleotide composition similar to that of TALE binding sites identified by Moscou and Bogdanove (2009). In *ADH1* and *gridlock*, such sites occurred on average every 7–9 bp. Four 12-bp sites were selected in *ADH1* (at positions 360, 408, 928, and 975 of the chromosomal gene sequence) and one 13-bp site in *gridlock* (at position 2356 of the chromosomal gene sequence) (Figure 2-2A, Table 2-1). We then constructed TALE repeat domains to recognize these targets, using the most abundant RVDs from native TALEs (NI for A, HD for C, NN for G, and NG for T). Sequences encoding the repeat domain of the native TALE tal1c were replaced with the custom repeat domains, and BamHI fragments from these engineered TALEs were then fused to sequences encoding the catalytic domain of FokI (see Figure 2-3). The resulting custom TALENs were tested in the yeast assay as homodimeric TALENs; that is, the identical DNA binding site was duplicated in inverse orientation on either side of a spacer. It is important to note that heterodimeric TALENs would need to be constructed to direct cleavage at naturally occurring DNA targets. Robust nuclease activity was

observed for the ADH1-360-12 and *gridlock*-2356-13r TALEN (Figure 2-2B). The ADH1-928-12 TALEN had modest activity that was nonetheless significantly above the negative controls. For each TALEN that gave positive results, nuclease activity was specific to the cognate target. These results indicate that novel, functional TALENs can be created by assembly of customized repeat domains. The experiments described here suggest that TALENs hold much promise for applications requiring sequence specific nucleases. In particular, the ability to create TALENs that recognize novel, arbitrarily selected target sequences suggests a potential for engineering TALENs for targeted genome modification. That said, the failure of some custom TALENs suggests that yet unknown rules govern the assembly of functional repeat domains. For example, repeat composition may influence protein stability, or interactions among repeat domains may affect DNA binding activity as has been observed for finger–finger interactions in zinc finger arrays (Elrod-Erickson *et al.*, 1996, Isalan *et al.*, 1997). Alternatively, the spacer lengths we used may have prevented dimerization of FokI, as appeared to be the case for some spacers with AvrBs3. Clearly, it will be important to gain a better understanding of the relationship of spacer length to function for TALENs with different repeat domains. Ascertaining the minimal DNA binding domain might help accomplish this; however, we believe the repeats alone are not sufficient for adequate DNA binding, as TALENs constructed with just the repeat domain did not function in the yeast assay (data not shown). In the short term, we will test whether custom TALENs can be created that recognize and cleave endogenous chromosomal targets, and we will evaluate the efficiency with which custom TALENs create genome modifications by non-homologous

end joining and homologous recombination. Such experiments will be key to assessing the full utility of these reagents for eukaryotic genome engineering.

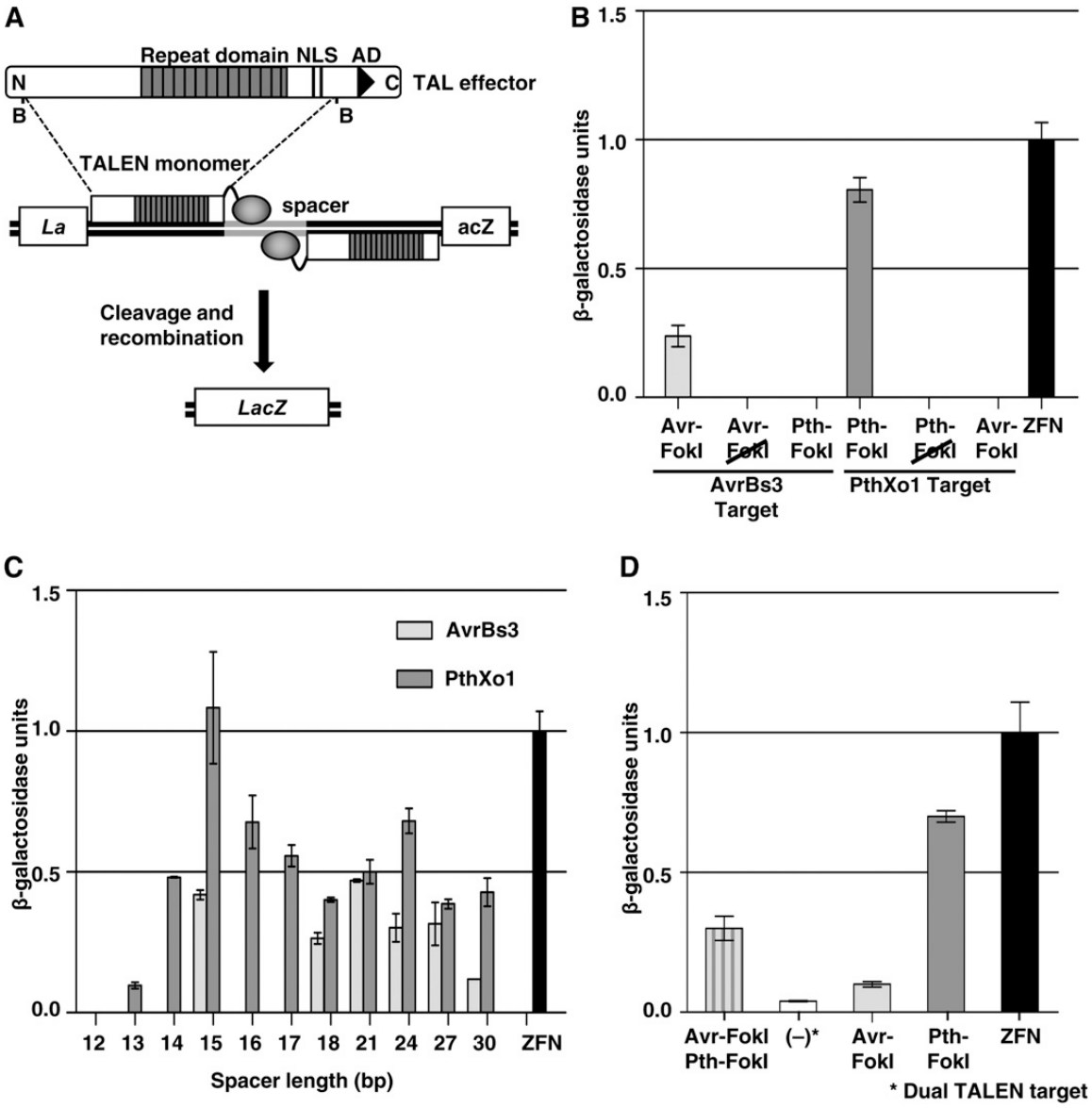


Figure 2-1. **Structure and activity of TALE nucleases (TALENs).** (**A**) Schematic of a transcription activator-like effector (TALE) protein. BamHI fragments were fused to the catalytic domain of the FokI endonuclease to create TALENs. TALEN activity was measured in an in vivo yeast assay that is described in the text (Townsend *et al.*, 2009). N, N terminus; NLS, nuclear localization signal; B, BamHI site; AD, acidic activation domain. (**B**) Activity of TALENs constructed with TALEs AvrBs3 and PthXo1. Haploid cell types containing either TALEN expression or target plasmid in 200 ml of overnight

culture were mated in YPD medium at 30˚. After 4 hr, the YPD medium was replaced with 5 ml of selective medium and incubated overnight at 30˚. Mated cultures were lysed, ONPG substrate added, and absorbance read at 415 nm using a 96-well plate reader (Townsend *et al.*, 2009). β-galactosidase levels were calculated as a function of substrate cleavage velocity. Avr–FokI, AvrBs3 TALEN; Pth–FokI, PthXo1 TALEN; Avr–FokI and Pth–FokI, AvrBs3 and PthXo1 fusions to a catalytically inactive the version of FokI (Bitinaite *et al.*, 1998); ZFN, zinc finger nuclease containing the Zif268 DNA binding domain as a positive control (Porteus & Baltimore, 2003). (**C**) Activity of AvrBs3 and PthXo1 TALENS on targets with different spacer lengths. ZFN, Zif268- derived zinc finger nuclease. (**D**) Function of a heterodimeric TALEN. Activity in yeast containing PthXo1–FokI and AvrBs3–FokI expression vectors and a plasmid with a target consisting of recognition sites for each, in head-to-tail orientation separated by 15 bp is shown (Avr–FokI, Pth–FokI). Also shown for reference is activity of AvrBs3 (Avr–FokI) and PthXo1 (Pth–FokI) TALENS individually and Zif268 (ZFN) on their respective targets. As a negative control, a yeast culture with only the target site plasmid for Avr–FokI Pth–FokI was assayed for LacZ activity [denoted as (-)].

A

Custom TALEN	1	2	3	4	5	6	7	8	9	10	11	12	13
ADH1-360-12	NI A	NG T	HD C	NI A	NI A	NN G	NI A	NG T	NG T	HD C	NG T	HD C	
ADH1-408-12r	HD C	HD C	HD C	NI A	NN G	NI A	NI A	NN G	NG T	NI A	NI A	NI A	
ADH1-928-12	HD C	HD C	NN G	NN G	NI A	NG T	NN G	HD C	NG T	HD C	HD C	NG T	
ADH1-975-12r	NI A	NN G	NI A	HD C	NI A	NI A	NI A	HD C	HD C	NI A	HD C	NI A	
gridlock-2356-13r	NI A	HD C	HD C	HD C	HD C	NG T	HD C	NG T	HD C	HD C	NN G	HD C	NG T

* NI = A HD = C NN = G NG = T

B

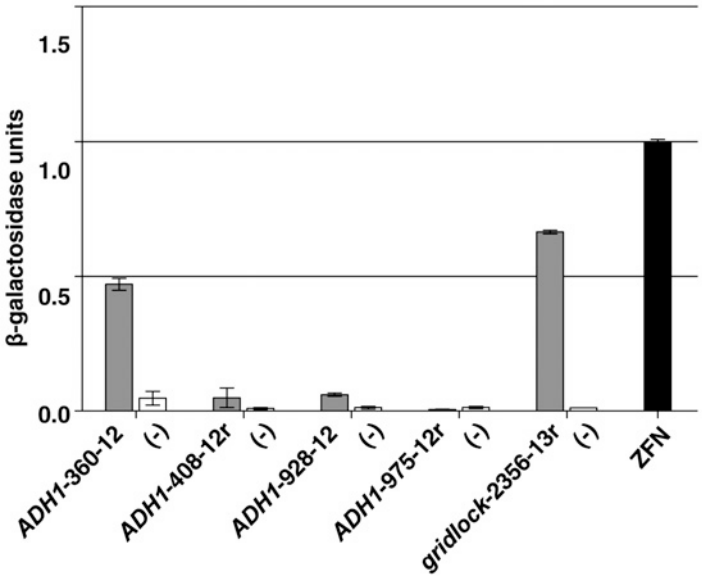


Figure 2-2. Custom TALENs. To construct custom TALENs, repeats with the RVDs NI, HD, NN, and NG, specifying A, C, G, and T, respectively, were synthesized individually and assembled into modules of one, two, or three repeats. These modules were ligated sequentially into a derivative of the *tal1c* gene from which the original repeats had been removed (Moscou & Bogdanove, 2009). Five custom TALENs targeted to *ADH1* from Arabidopsis and the zebrafish *gridlock* gene were created. (**A**) The RVD sequences of individual custom TALENs and their respective DNA recognition sequences are shown. (**B**) The yeast in vivo assay was used to determine the activity of custom TALENs. Plasmids containing two recognition sites for the respective TALEN in

head-to-tail orientation separated by 16- to 18-bp spacers were used as targets. We chose spacer lengths on the basis of the distance closest to 15 bp from the 39 end of the next neighboring (and opposing) candidate site, since we anticipated constructing heterodimeric TALENS that recognize native chromosomal sequences. Sixteen-base pair spacers were used for ADH1-360-12, ADH1-408-12r and 18-bp spacers for ADH1-928-12, ADH1-975-12r, and *gridlock*-2356-13r. The yeast assay was performed as described in Fig. 1B. (-), negative control with target site plasmids only; ZFN, zinc finger nuclease positive control.

Figure 2-3. Supplementary figure of the amino acid sequences of TALENs used in this study.

>avrBs3_TALEN
MASSPPKKKRKVSWKDASGWSRMHADPIRSRTPS
PARELLPGPQPDGVQPTADRGVSPPAGGPLDGLP
ARRTMSRTRLPSPPAPSPAFSAGSFSDLLRQFDP
SLFNTSLFDSLPPFGAHHTEAATGEWDEVQSGLR
AADAPPPTMRVAVTAARPPRAKPAPRRRAAQPSD
ASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHH
EALVGHGFTHAHIVALSQHPAALGTVAVKYQDMI
AALPEATHEAIVGVGKQWSGARALEALLTVAGEL
RGPPLQLDTGQLLKIAKRGGVTAVEAVHAWRNAL
TGAPLNLTPEQVVAIASHDGGKQALETVQRLLPV
LCQAHGLTPQQVVAIASNGGGKQALETVQRLLPV
LCQAHGLTPQQVVAIASNSGGKQALETVQRLLPV
LCQAHGLTPEQVVAIASNGGGKQALETVQRLLPV
LCQAHGLTPEQVVAIASNIGGKQALETVQALLPV
LCQAHGLTPEQVVAIASNIGGKQALETVQALLPV
LCQAHGLTPEQVVAIASNIGGKQALETVQALLPV
LCQAHGLTPEQVVAIASHDGGKQALETVQRLLPV
LCQAHGLTPEQVVAIASHDGGKQALETVQRLLPV
LCQAHGLTPQQVVAIASNGGGKQALETVQRLLPV
LCQAHGLTPEQVVAIASNSGGKQALETVQALLPV
LCQAHGLTPEQVVAIASNSGGKQALETVQRLLPV
LCQAHGLTPEQVVAIASHDGGKQALETVQRLLPV
LCQAHGLTPEQVVAIASHDGGKQALETVQRLLPV
LCQAHGLTPEQVVAIASHDGGKQALETVQRLLPV
LCQAHGLTPQQVVAIASNGGGRPALETVQRLLPV
LCQAHGLTPEQVVAIASHDGGKQALETVQRLLPV
LCQAHGLTPQQVVAIASNGGGRPALESIVAQLSR
PDPALAALTNDHLVALACLGGRPALDAVKKGLPH
APALIKRTNRRIPERTSHRVADHAQVVRVLGFFQ
CHSHPAQAFDDAMTQFGMSRHGLLQLFRRVGVTE
LEARSGTLPPASQRWDRILQASGMKRAKPSPTST
QTPDQASLHAFADSLERDLDAPSPMHEGDQTRAS
SRKRSRSDRAVTGPSAQQSFEVRVPEQRDALHLP
LSWRVKRPRTSIGGGLPDPISRSQLVKSELEEKK
SELRHKLKYVPHEYIELIEIARNSTQDRILEMKV
MEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDY
GVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRN
KHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKA
QLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLT
LEEVRRKFNNGEINF

>pthXo1_TALEN
MASSPPKKKRKVSWKDASGWSRMHADPIRPRRPS
PARELLPGPQPDRVQPTADRGVSAPAGSPLDGLP
ARRTVSRTRLPSPPAPSPAFSAGSFSDLLRPFDP
SLLDTSLLDSMPAVGTPHTAAAPAEWDEAQSALR
AADDPPPTVRVAVTAARPPRAKPAPRRRAAQPSD
ASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHH
EALVGHGFTHAHIVALSQHPAALGTVAVTYQHII
TALPEATHEDIVGVGKQWSGARALEALLTDAGEL
RGPPLQLDTGQLVKIAKRGGVTAMEAVHASRNAL
TGAPLNLTPAQVVAIASNNGGKQALETVQRLLPV
LCQAHGLTPAQVVAIASHDGGKQALETMQRLLPV
LCQAHGLPPDQVVAIASNIGGKQALETVQRLLPV
LCQAHGLTPDQVVAIASHGGGKQALETVQRLLPV
LCQAHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQAHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQAHGLTPDQVVAIASNGGKQALETVQRLLPVL
CQAHGLTPDQVVAIASHDGGKQALETVQRLLPVL
CQTHGLTPAQVVAIASHDGGKQALETVQQLLPVL
CQAHGLTPDQVVAIASNIGGKQALATVQRLLPVL
CQAHGLTPDQVVAIASNGGGKQALETVQRLLPVL
CQAHGLTPDQVVAIASNGGGKQALETVQRLLPVL
CQAHGLTQVQVVAIASNIGGKQALETVQRLLPVL
CQAHGLTPAQVVAIASHDGGKQALETVQRLLPVL
CQAHGLTPDQVVAIASNGGGKQALETVQRLLPVL
CQAHGLTQEQVVAIASNNGGKQALETVQRLLPVL
CQAHGLTPDQVVAIASNGGGKQALETVQRLLPVL
CQAHGLTPAQVVAIASNIGGKQALETVQRLLPVL
CQDHGLTLAQVVAIASNIGGKQALETVQRLLPVL
CQAHGLTQDQVVAIASNIGGKQALETVQRLLPVL
CQDHGLTPDQVVAIASNIGGKQALETVQRLLPVL
CQDHGLTLDQVVAIASNGGKQALETVQRLLPVLC
QDHGLTPDQVVAIASNSGGKQALETVQRLLPVLC
QDHGLTPNQVVAIASNGGKQALESIVAQLSRPDP
ALAALTNDHLVALACLGGRPAMDAVKKGLPHAPE
LIRRVNRRIGERTSHRVADYAQVVRVLEFFQCHS
HPAYAFDEAMTQFGMSRNGLVQLFRRVGVTELEA
RGGTLPPASQRWDRILQASGMKRAKPSPTSAQTP
DQASLHAFADSLERDLDAPSPMHEGDQTRASSRK
RSRSDRAVTGPSAQQAVEVRVPEQRDALHLPLSW
RVKRPRTRIWGGLPDPISRSQLVKSELEEKKSEL
RHKLKYVPHEYIELIEIARNSTQDRILEMKVMEF
FMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVI
VDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHI
NPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLT

RLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEE
VRRKFNNGEINF

>adh1-360-12_TALEN
MASSPPKKKRKVSWKDASGWSRMHADPIRPRRPS
PARELLPGPQPDRVQPTADRGVSAPAGSPLDGLP
ARRTVSRTRLPSPPAPSPAFSAGSFSDLLRPFDP
SLLDTSLLDSMPAVGTPHTAAAPAEWDEAQSALR
AADDPPPTVRVAVTAARPPRAKPAPRRRAAQPSD
ASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHH
EALVGHGFTHAHIVALSQHPAALGTVAVTYQHII
TALPEATHEDIVGVGKQWSGARALEALLTDAGEL
RGPPLQLDTGQLVKIAKRGGVTAMEAVHASRNAL
TGAPLNLTPAQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIANNNGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALESIVAQLSR
RDPALAALTNDHLVALACLGGRPALDAVKKGLPH
APEFIRRVNRRIAERTSHRVADYAHVVRVLEFFQ
CHSHPAHAFDEAMTQFGMSRHGLVQLFRRVGVTE
FEARYGTLPPASQRWDRILQASGMKRAKPSPTSA
QTPDQTSLHAFADSLERDLDAPSPMHEGDQTRAS
SRKRSRSDRAVTGPSAQQAVEVRVPEQRDALHLP
LSWRVKRPRTRIWGGLPDPISRSQLVKSELEEKK
SELRHKLKYVPHEYIELIEIARNSTQDRILEMKV
MEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDY
GVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRN
KHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKA
QLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLT
LEEVRRKFNNGEINF

>adh1-408-12r_TALEN
MASSPPKKKRKVSWKDASGWSRMHADPIRPRRPS
PARELLPGPQPDRVQPTADRGVSAPAGSPLDGLP
ARRTVSRTRLPSPPAPSPAFSAGSFSDLLRPFDP
SLLDTSLLDSMPAVGTPHTAAAPAEWDEAQSALR
AADDPPPTVRVAVTAARPPRAKPAPRRRAAQPSD

ASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHH
EALVGHGFTHAHIVALSQHPAALGTVAVTYQHII
TALPEATHEDIVGVGKQWSGARALEALLTDAGEL
RGPPLQLDTGQLVKIAKRGGVTAMEAVHASRNAL
TGAPLNLTPAQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIANNNGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIANNNGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALESIVAQLSR
RDPALAALTNDHLVALACLGGRPALDAVKKGLPH
APEFIRRVNRRIAERTSHRVADYAHVVRVLEFFQ
CHSHPAHAFDEAMTQFGMSRHGLVQLFRRVGVTE
FEARYGTLPPASQRWDRILQASGMKRAKPSPTSA
QTPDQTSLHAFADSLERDLDAPSPMHEGDQTRAS
SRKRSRSDRAVTGPSAQQAVEVRVPEQRDALHLP
LSWRVKRPRTRIWGGLPDPISRSQLVKSELEEKK
SELRHKLKYVPHEYIELIEIARNSTQDRILEMKV
MEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDY
GVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRN
KHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKA
QLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLT
LEEVRRKFNNGEINF

>adh1-928-12_TALEN
MASSPPKKKRKVSWKDASGWSRMHADPIRPRRPS
PARELLPGPQPDRVQPTADRGVSAPAGSPLDGLP
ARRTVSRTRLPSPPAPSPAFSAGSFSDLLRPFDP
SLLDTSLLDSMPAVGTPHTAAAPAEWDEAQSALR
AADDPPPTVRVAVTAARPPRAKPAPRRRAAQPSD
ASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHH
EALVGHGFTHAHIVALSQHPAALGTVAVTYQHII
TALPEATHEDIVGVGKQWSGARALEALLTDAGEL
RGPPLQLDTGQLVKIAKRGGVTAMEAVHASRNAL
TGAPLNLTPAQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIANNNGGKQALETVQRLLPV
LCQDHGLTPDQVVAIANNNGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV

LCQDHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQDHGLTPDQVVAIANNNGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNGGGKQALESIVAQLSR
RDPALAALTNDHLVALACLGGRPALDAVKKGLPH
APEFIRRVNRRIAERTSHRVADYAHVVRVLEFFQ
CHSHPAHAFDEAMTQFGMSRHGLVQLFRRVGVTE
FEARYGTLPPASQRWDRILQASGMKRAKPSPTSA
QTPDQTSLHAFADSLERDLDAPSPMHEGDQTRAS
SRKRSRSDRAVTGPSAQQAVEVRVPEQRDALHLP
LSWRVKRPRTRIWGGLPDPISRSQLVKSELEEKK
SELRHKLKYVPHEYIELIEIARNSTQDRILEMKV
MEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDY
GVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRN
KHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKA
QLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLT
LEEVRRKFNNGEINF

>adh1-975-12r_TALEN
MASSPPKKKRKVSWKDASGWSRMHADPIRPRRPS
PARELLPGPQPDRVQPTADRGVSAPAGSPLDGLP
ARRTVSRTRLPSPPAPSPAFSAGSFSDLLRPFDP
SLLDTSLLDSMPAVGTPHTAAAPAEWDEAQSALR
AADDPPPTVRVAVTAARPPRAKPAPRRRAAQPSD
ASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHH
EALVGHGFTHAHIVALSQHPAALGTVAVTYQHII
TALPEATHEDIVGVGKQWSGARALEALLTDAGEL
RGPPLQLDTGQLVKIAKRGGVTAMEAVHASRNAL
TGAPLNLTPAQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIANNNGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNIGGKQALESIVAQLSR
RDPALAALTNDHLVALACLGGRPALDAVKKGLPH
APEFIRRVNRRIAERTSHRVADYAHVVRVLEFFQ

CHSHPAHAFDEAMTQFGMSRHGLVQLFRRVGVTE
FEARYGTLPPASQRWDRILQASGMKRAKPSPTSA
QTPDQTSLHAFADSLERDLDAPSPMHEGDQTRAS
SRKRSRSDRAVTGPSAQQAVEVRVPEQRDALHLP
LSWRVKRPRTRIWGGLPDPISRSQLVKSELEEKK
SELRHKLKYVPHEYIELIEIARNSTQDRILEMKV
MEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDY
GVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRN
KHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKA
QLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLT
LEEVRRKFNNGEINF

>gridlock-2356-13r_TALEN
MASSPPKKKRKVSWKDASGWSRMHADPIRPRRPS
PARELLPGPQPDRVQPTADRGVSAPAGSPLDGLP
ARRTVSRTRLPSPPAPSPAFSAGSFSDLLRPFDP
SLLDTSLLDSMPAVGTPHTAAAPAEWDEAQSALR
AADDPPPTVRVAVTAARPPRAKPAPRRRAAQPSD
ASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHH
EALVGHGFTHAHIVALSQHPAALGTVAVTYQHII
TALPEATHEDIVGVGKQWSGARALEALLTDAGEL
RGPPLQLDTGQLVKIAKRGGVTAMEAVHASRNAL
TGAPLNLTPAQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIANNNGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNGGGKQALESIVAQLSR
RDPALAALTNDHLVALACLGGRPALDAVKKGLPH
APEFIRRVNRRIAERTSHRVADYAHVVRVLEFFQ
CHSHPAHAFDEAMTQFGMSRHGLVQLFRRVGVTE
FEARYGTLPPASQRWDRILQASGMKRAKPSPTSA
QTPDQTSLHAFADSLERDLDAPSPMHEGDQTRAS
SRKRSRSDRAVTGPSAQQAVEVRVPEQRDALHLP
LSWRVKRPRTRIWGGLPDPISRSQLVKSELEEKK
SELRHKLKYVPHEYIELIEIARNSTQDRILEMKV
MEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDY
GVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRN

KHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKA
QLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLT
LEEVRRKFNNGEINF

Table 2-1. Table of TALEN reagents used in the study

TALEN name	pMAT	Gene	RVDs	Target name	Description	Spacer length (bp)
PthXo1 TALEN			24	P15	PthXo1 target	15
AvrBs3 TALEN			17	A21	AvrBs3 target	21
AvrBs3 TALEN			17	A24	AvrBs3 target	24
Zif268 ZFN	-	-	-	1715	Zif268 target	6
ADH1-360-12	58	ADH1	12	T58	ADH1-360-12 target	16
ADH1-408-12r	63	ADH1	12	T63	ADH1-408-12r target	16
ADH1-928-12	68	ADH1	12	T68	ADH1-928-12 target	18
ADH1-975-12r	73	ADH1	12	T73	ADH1-975-12r target	18
gridlock-2356-13r	106	gridlock	13	T106	gridlock-2356-13r target	18

CHAPTER 3

EFFICIENT DESIGN AND ASSEMBLY OF CUSTOM TALEN AND OTHER TAL EFFECTOR-BASED CONSTRUCTS FOR DNA TARGETING[1]

[1] Contributors to this work are as follows: Tomas Cermak, Erin L. Doyle, Michelle Christian, Li Wang, Yong Zhang, Clarice Schmidt, Joshua A. Baller, Nikunj V. Somia, Adam J. Bogdanove, and Daniel F. Voytas.

Introduction

Transcription activator-like (TAL) effectors are a newly described class of specific DNA binding protein, so far unique in the simplicity and manipulability of their targeting mechanism. Produced by plant pathogenic bacteria in the genus *Xanthomonas*, the native function of these proteins is to directly modulate host gene expression. Upon delivery into host cells via the bacterial type III secretion system, TAL effectors enter the nucleus, bind to effector-specific sequences in host gene promoters, and activate transcription (Bogdanove *et al.*, 2010c). Their targeting specificity is determined by a central domain of tandem, 33-35 amino acid repeats, followed by a single truncated repeat of 20 amino acids (Figure 3-1A). The majority of naturally occurring TAL effectors examined have between 12 and 27 full repeats (Boch & Bonas, 2010). Members of our group, and another lab independently, discovered that a polymorphic pair of adjacent residues at positions 12 and 13 in each repeat, the "repeat-variable di-residue" (RVD), specifies the target, one RVD to one nucleotide, with the four most common RVDs each preferentially associating with one of the four bases (Figure 3-1A) (Boch *et al.*, 2009, Moscou & Bogdanove, 2009). Also, naturally occurring recognition sites are uniformly preceded by a T that is required for TAL effector activity (Boch *et al.*, 2009, Moscou & Bogdanove, 2009). These straightforward sequence relationships allow the prediction of TAL effector binding sites (Antony *et al.*, 2010, Boch *et al.*, 2009, Moscou & Bogdanove, 2009, Römer *et al.*, 2010) and construction of TAL effector responsive promoter elements (Römer, Recht, *et al.*, 2009), as well as customization of TAL effector repeat domains to bind DNA sequences of interest (Christian *et al.*, 2010, Miller *et al.*, 2011, Morbitzer *et al.*, 2010, Zhang *et al.*, 2011b).

As a result, TAL effectors have attracted great interest as DNA targeting tools. In particular, we and other groups have shown that TAL effectors can be fused to the catalytic domain of the *Fok*I nuclease to create targeted DNA double strand breaks (DSBs) *in vivo* for genome editing (Christian *et al.*, 2010, T. Li *et al.*, 2010, Mahfouz *et al.*, 2010, Miller *et al*., 2011). Since *Fok*I cleaves as a dimer, these TAL effector nucleases (TALENs) function in pairs, binding opposing targets across a spacer over which the *Fok*I domains come together to create the break (Figure 3-1B) (Christian *et al.*, 2010). DSBs are repaired in nearly all cells by one of two highly conserved processes, non-homologous end joining (NHEJ), which often results in small insertions or deletions and can be harnessed for gene disruption, and homologous recombination (HR), which can be used for gene insertion or replacement (Stoddard, 2011, Urnov *et al.*, 2010). Genome modifications based on both of these pathways have been obtained with high frequency in a variety of plant and animal species using zinc finger nucleases (ZFNs) and homing endonucleases. However, for each of these platforms, engineering novel specificities has generally required empirical and selection-based approaches that can be time and resource intensive. And, despite a significant recent advance for ZFNs that takes finger context into account to achieve high success rates (Sander, Dahlborg, *et al.*, 2011), targeting capacity (the diversity of sequences that can be recognized) still suffers limitations (J K Joung *et al.*, 2000, Maeder *et al.*, 2008, Sander *et al.*, 2009). TALENs thus far appear not to be subject to these constraints. And in at least one study, mutagenesis frequency was estimated to be as high as 25% of transfected cells, on par with or better than ZFNs (Miller *et al*., 2011).

The TAL effector repeat domain also has been successfully customized to make targeted transcription factors, both in plants in the native protein context and in human cells with the TAL effector activation domain replaced by VP64 (Morbitzer *et al.*, 2010, Zhang *et al.*, 2011b). Fusions to other protein domains for chromatin modification, gene regulation, or other applications can also be envisioned. Thus, an efficient method for assembling genetic constructs to encode TAL effectors and TAL effector fusions to other proteins, with repeat arrays of user-defined length and RVD sequence, is highly desirable.

In our previous work, we constructed TALENs with customized repeat arrays through sequential cloning of sequence-verified single, double, and triple repeat modules (Christian *et al.*, 2010). We sought a more rapid approach that would not rely on commercial synthesis, which is costly, or PCR-based methods, which can result in mutations or recombined repeats. We opted for Golden Gate cloning, a recently developed method of assembling multiple DNA fragments in an ordered fashion in a single reaction (Engler *et al.*, 2008, 2009). The Golden Gate method uses Type IIS restriction endonucleases, which cleave outside their recognition sites to create unique 4 bp overhangs (sticky ends) (Figure 3-2). Cloning is expedited by digesting and ligating in the same reaction mixture because correct assembly eliminates the enzyme recognition site.

We report here a complete set of plasmids for assembling novel repeat arrays for TALENs, TAL effectors, or TAL effector fusions to other proteins, using the Golden Gate method in two steps. We also describe software for TALEN targeting based on guidelines we developed to reflect naturally occurring TAL effector binding sites and on

our previous TALEN study. We show that TALENs targeted with this software and constructed using the plasmid set are active in a yeast DNA cleavage assay, and effective in gene targeting in human cells and *Arabidopsis thaliana* (hereafter Arabidopsis) protoplasts. Finally, we demonstrate successful construction of a functional analog of the *avrHah1* TAL effector gene of *X. gardneri* (Schornack *et al.*, 2008).

Materials and Methods

Protocol for assembly of custom TALEN, TAL effector, or TAL effector fusion-ready constructs

Assembly of a custom TALEN or TAL effector construct is accomplished in five days (Figure 3-3) and involves two steps: 1) assembly of repeat modules into intermediary arrays of 1-10 repeats, and 2) joining of the intermediary arrays into a backbone to make the final construct. A schematic is shown in Figure 3-2, and the complete set of required plasmids is displayed in Figure 3-4. Construction and features of the plasmids themselves are described in the section following. The assembly protocol differs slightly for arrays of 12-21 modules versus arrays of 22-31 modules. We use as an example here construction of a TALEN monomer with a 16 RVD array, and note differences in the protocol where they occur for making constructs with arrays of 22-31 RVDs.

Day 1. Consider the RVD array NI HD HD NN HD NI NI NG HD NG HD NI NI NG HD NG, targeting the sequence 5'-AGCCCAATCTCACTCT-3'. Note that the 5'-T preceding the RVD-specified sequence is not shown and need not be considered in the

assembly, although based on evidence to date (Boch *et al.*, 2009, Moscou & Bogdanove, 2009), it should be considered during site selection. Select from the module plasmids those that encode RVDs 1-10 in the array using plasmids numbered in that order. For example the plasmid for the first RVD would be pNI1, the second pHD2, the third pHD3, *etc*. Modules from these plasmids will be cloned into array plasmid pFUS_A. Next, select modules for RVDs 11-15 in the 16 RVD array again starting with plasmids numbered from 1. Thus for RVD 11 pHD1 would be used, for RVD 12 pNI2, *etc*. Note that the sixteenth and last RVD is encoded by a different, last repeat plasmid and is added later, in the second step (see Day 3). Modules encoding RVDs 11-15 are cloned into a pFUS_B array plasmid. The pFUS_B plasmids are numbered 1-10 and should be selected according to the number of modules going in. Thus, in our example, pFUS_B5 should be used. If arrays of 22-31 modules are to be assembled, the first 10 modules are cloned into pFUS_A30A, the second 10 modules into pFUS_A30B, and the remaining modules into the appropriate pFUS_B plasmid, again according to the number of modules going in.

The module and array plasmids (150 ng each) are subjected to digestion and ligation in a single 20 ml reaction containing 1 ml *Bsa*I (10 units, New England BioLabs), and 1 ml T4 DNA Ligase (2,000 units, New England BioLabs) in T4 DNA ligase buffer (New England BioLabs). The reaction is incubated in a thermocycler for 10 cycles of 5 min at 37˚C and 10 min at 16˚C, then heated to 50˚C for 5 min and then 80˚C for 5 min. Then, 1 ml 25 mM ATP and 1 ml Plasmid Safe DNase (10 units, Epicentre) are added. The mixture is incubated at 37˚C for 1 hr, then used to transform *E. coli* cells. Cells are plated on LB agar containing 50 mg/ml spectinomycin, with X-gal and IPTG for blue/white screening of recombinants, as described (Sambrook *et al.*, 1989).

Treatment with Plasmid Safe DNase is an important step to prevent linear DNA fragments, including partial arrays, from recombining into and circularizing the linearized array plasmids following transformation, due to the presence of partial repeat sequences at the termini of the array plasmids.

Day 2. Pick up to three white colonies from each transformation and start overnight cultures.

Day 3. Isolate plasmid DNA and identify clones with the correct arrays by restriction enzyme digestion and agarose gel electrophoresis. *Afl*II and *Xba*I will release the repeat arrays, which will be 1048 bp for pFUS_A, pFUS_A30A, or pFUS_A30B plasmids, and of varying sizes for pFUS_B plasmids.

The next step is to join the intermediary arrays, along with a last repeat, into the desired context, using one of four backbone plasmids. A 20 ml digestion and ligation reaction mixture is prepared as in the first step but with 150 ng each of the pFUS_A and pFUS_B plasmids containing the intermediary repeat arrays (or the pFUS_A30A, pFUS_A30B, and pFUS_B plasmids carrying the intermediaries for final arrays of 22-31 RVDs), 150 ng of the backbone plasmid, in this case pTAL3 or pTAL4 for constructing a TALEN monomer, and importantly, 150 ng of the appropriate last repeat plasmid. In our example, pLR-NG, for the 16th and last RVD, would be used. The reaction is treated and used to transform *E. coli* as above, except that Plasmid Safe DNAse treatment is omitted because the backbone plasmid termini have no homology with the array. Also, in this step, ampicillin (100 mg/ml) is used in place of spectinomycin for selection of transformants.

Day 4. Pick up to 3 white colonies from each transformation and start overnight cultures.

Day 5. Isolate plasmid DNA and identify clones containing the final, full-length repeat array. Array length can be verified by digestion with *Bst*API (or *Stu*I) and *Aat*II, which cut just outside the repeats, or with *Sph*I, which cuts farther out. Array integrity can be checked using *Bsp*EI, which cuts only in HD modules 2-10. The array can also be characterized by DNA sequencing.

Construction of module, last repeat, array, and backbone plasmids

Repeat modules with the RVDs HD, NG, NI, NK, and NN, across ten staggered positions and with a *Bsa*I site added to each end, were synthesized. The modules were cloned between the unique *Xba*I and *Xho*I sites of pTC14, replacing the spectinomycin resistance gene in that plasmid, to create a set of 50 module plasmids (pHD1 through pHD10, pNG1 through pNG10, *etc*.). pTC14 is a derivative of the Gateway entry and TOPO cloning vector pCR8 (Invitrogen) in which the Gateway cassette was replaced with a gene for tetracycline resistance using the flanking *Eco*RV and *Hpa*I sites. Aside from the RVD codons, the modules at each position are identical, except for a *Bsp*EI site introduced into HD modules 2-10 for testing full-length array integrity by digestion. The modules are based on the first repeat of *tal1c* of *X. oryzae* pv. oryzicola strain BLS256 (Moscou & Bogdanove, 2009), which matches the consensus repeat and is made up of common codons.

Similarly, one module for each of the five RVDs containing the last, truncated repeat of the TAL effector repeat domain was synthesized and cloned in plasmid pCR8 (carrying the spectinomycin resistance gene) using *Apa*I and *Xba*I and replacing the Gateway cassette, to create five last repeat plasmids (e.g., pLR-HD).

Next, array plasmids pFUS_A, pFUS_A30A, and pFUS_A30B were created by cloning, using *Afl*II and *Xba*I, synthesized fragments into pCR8 that contain two internal *Bsa*I sites oriented to cut outward into flanking sequences such that linearizing the vector with the enzyme leaves the appropriate overhangs to accept an array of 10 repeat modules (i.e., complementary on one side to the 5' end of position 1 modules and on the other to the 3' end of position 10 modules). The series of array plasmids pFUS_B1 through pFUSB10 were made similarly to be complementary on one side to the 5' end of position 1 modules, but complementary on the other to the 3' end of modules in position 1 through 10, respectively, to accept arrays ranging from 1 to 10 modules. A DNA fragment containing the *lacZ* gene for blue/white screening was cloned between the two *Bsa*I sites. For this, the multiple cloning site between the *Hinc*II and *Eco*53kI sites in phagemid pBCSK+ (Stratagene) was deleted and the *lacZ* gene PCR amplified with primers carrying *Kas*I and *Age*I overhangs. These sites were included in the synthesized fragments, allowing the *lacZ* gene to be placed between the *Bsa*I sites, maintaining the overhang sequences for accepting modules. The inserts in the array plasmids all contain terminal *Esp*3I (another type IIS enzyme) sites positioned to cut inward and release the arrays with appropriate overhangs for ordered ligation into a backbone plasmid for complete arrays of 12-21 (a pFUS_A array with a pFUS_B array, plus a last repeat) or 22-31 repeats (a pFUS_A30A with a pFUS_A30B and a pFUS_B array, plus a last repeat). These sites, or flanking *Afl*II and *Xba*I sites in the vector (enzymes that are generally less expensive), can also be used to screen assembled clones for the correct size.

Backbone plasmid pTAL3 was derived from pFZ85, a precursor to the TALEN yeast expression vector we created previously (Christian *et al.*, 2010). Derived from pDW1789 (Townsend *et al.*, 2009), pFZ85 contains the counterselectable *ccdB* gene flanked by *Bam*HI sites downstream of the yeast TEF promoter and a sequence encoding a nuclear localization signal, and upstream of a sequence encoding a linker and the *Fok*I nuclease catalytic domain. For our previous TALEN constructs, we used *tal1c* as a context for custom repeat arrays. First, solely for expediency of later adding the *lacZ* gene, the *Sph*I fragment of *tal1c* was replaced with the *Sph*I fragment of TAL effector gene *pthXo1* (Yang *et al.*, 2006), which has minor polymorphisms flanking the repeat region that create convenient restriction enzyme sites. The spanning *Bam*HI fragment of the resulting gene was then cloned between the *Bam*HI sites of pFZ85. Finally, the repeat region within the *Sph*I fragment was deleted by digestion with *Bst*API and *Aat*II and replaced with a fragment carrying the *lacZ* gene for blue/white screening (cloned into this fragment as described above), flanked by outward cutting *Esp*3I sites and the necessary sequences to create a specific overhang on either end to accept final arrays and reconstitute a complete TAL effector domain. Importantly, the *Sph*I sites, which are highly conserved among TAL effectors and are useful for swapping the entire repeat region into other TAL effector constructs, are preserved. The architecture of the constructs is the same as reported in our earlier work (Christian *et al.*, 2010), encoding 287 and 230 amino-acids of the TAL effector upstream and downstream of the repeats, respectively, with an additional 6 amino acids linking the TAL effector and *Fok*I domains. To create pTAL4, which is identical to pTAL3 except that it carries *LEU2* in place of *HIS3*, first the *LEU2* gene was PCR amplified using primers having 20 bp

extensions with homology to the region at the 5' end of the *Bpu*10I and 3' end of the *Afe*I site in pDW1789. Then, pDW1789 was linearized with *Bpu*10I and *Afe*I (removing the *HIS3* gene) and the PCR amplified *LEU2* gene was inserted by *in vivo* recombination in *E. coli* (Hoffmann *et al.*, 1995). Finally, into this plasmid, the *Xba*I-*Sac*I fragment of pTAL3 containing the TALEN backbone construct was introduced at the corresponding sites.

pTAL1 was created by replacing the *Sph*I fragment of *tal1C* in pCS691 with the corresponding *Sph*I fragment of pTAL3, containing the *lacZ* gene and the *Esp*3I sites and flanking sequences for accepting final arrays. pCS691 is a derivative of Gateway entry vector pENTR-D (Invitrogen) containing between the attL sites the complete *tal1c* gene preceded by both Kozak and Shine-Dalgarno consensus sequences for efficient translation in eukaryotic or bacterial cells respectively. In pCS691, the kanamycin resistance gene of pENTR-D is replaced by the *Bsp*HI fragment of pBlueScript SK(-) (Stratagene) for ampicillin resistance. To create pTAL2, the stop codon of *tal1c* in pTAL1 was deleted using the QuickChange mutagenesis kit (Stratagene) to allow translational fusion to other protein domains following Gateway recombination into a destination vector. A schematic of all module, last repeat, array, and backbone plasmids (Figure 3-4) are included.

Software to identify candidate TALEN target sites

The software used to design TALENs in this study was written in Python 2.6.4. and runs in Linux (Ubuntu 10.10). It is available for use as an online tool (TAL Effector-Nucleotide Targeter, TALE-NT; http://boglabx.plp.iastate.edu/TALENT/). The tool

provides a window to input DNA sequences (Figure 3-5A), which are then scanned for sites based on TALEN design guidelines we established, described in the results section. The software identifies sets of TALEN recognition sites between 15 and 30 bp in length and separated by a spacer. The default spacer lengths are 15 and 18-30 bp, but other lengths can be specified by the user. In addition, buttons allow users to exclude design guidelines individually. The output is tab-delimited text, which can be imported into standard spreadsheet software (Figure 3-5B). It provides coordinates and sequences of identified targets indicating the recognition sites for the left and right TALEN monomers and the spacer sequence. Because naturally occurring TAL effector recognition sites are uniformly preceded by a T, which is required for TAL effector activity (Boch *et al.*, 2009, Moscou & Bogdanove, 2009), only TALEN monomer recognition sites preceded by a T are included. The T itself is not part of the output. Finally, the software provides the RVD sequences needed to construct the corresponding custom TALENs.

Testing TALEN function in yeast

The yeast assay for TALEN function was adapted from one we developed previously for ZFNs (Christian *et al.*, 2010, Townsend *et al.*, 2009) in which cleavage of the target, positioned between partially duplicated fragments of the *lacZ* gene, reconstitutes the gene via subsequent recombination to provide a quantitative readout (Figure 3-6A). For typical heterodimeric target sites (*i.e*. such as would typically occur in a native DNA sequence), paired TALEN constructs, in pTAL3 and pTAL4, are transformed together into yeast strain YPH500 (a mating type) using histidine and leucine prototrophy for selection. Individual TALEN monomers can be tested on homodimeric sites using just one of these

plasmids. The target is made using synthesized complementary oligonucleotides that produce *Bgl*II- and *Spe*I-compatible ends, and cloned between the *lacZ* fragments in the high copy DNA cleavage reporter plasmid pCP5 (Townsend *et al.*, 2009) cut with those enzymes (Figure 3-6B). The target plasmid is transformed into yeast strain YPH499 (a mating type), using tryptophan prototrophy for selection, but also excluding uracil from the growth medium: in addition to the target cloning site, pCP5 carries also the *URA3* gene between the *lacZ* fragments so that selection for *URA3* ensures that the strain has not undergone spontaneous recombination (and loss of *URA3*) prior to the assay.

Three transformants each of YPH500 carrying the TALEN construct(s) and of YPH499 carrying the target plasmid are cultured overnight at 30°C, with rotary shaking at 800 rpm, in synthetic complete medium lacking histidine and/or leucine (TALENs), or tryptophan and uracil (target). TALEN and target transformants are next mated (3 pairs) by combining 200-500 ml of the overnight cultures, adding 1 ml of YPD medium, and incubating for 4-6 hrs at 30°C, shaking at 250-300 rpm. Cells are harvested by centrifguation, washed in 1 ml synthetic complete medium lacking histidine and/or leucine and tryptophan, but now containing uracil, then resuspsended in 5 ml of that medium and incubated overnight again at 30°C, with shaking (800 rpm), to an OD_{600} between 0.1 and 0.9. Cells are harvested by centrifugation, then resuspended and lysed using YeastBuster Protein Extraction Reagent (Novagen) according to the manufacturer's protocol for small cultures. 100 ml of lysate is transferred to a microtiter well plate and b-galactosidase activity measured and normalized as previously described (Townsend *et al.*, 2009). For high throughput, yeast may be cultured and mated (using a gas permeable

seal) as well as lysed in 24 well blocks. We typically express activity relative to a Zif268 ZFN (Townsend *et al.*, 2009).

Expression of custom TALENs in human cells and *Arabidopsis* protoplasts and detection of site-specific mutations

One of the pairs of TALENs targeting the human *HPRT1* gene was subcloned into the mammalian expression vector pCDNA3.1(-) (Invitrogen) using *Xho*I and *Afl*II. These enzymes excise the entire TALEN from pTAL3 or pTAL4 and place the coding sequence under control of the CMV (cytomegalovirus) promoter. The resulting plasmids were introduced into HEK293T cells by transfection using Lipofectamine 2000 (Invitrogen) following the manufacturer's protocol. Cells were collected 72 hours after transfection, and genomic DNA isolated and digested with *Hpy*188I, which cuts in the spacer sequence of the TALEN target site. After digestion, a chromosomal fragment encompassing the target site was amplified by PCR. Upon completion, the reactions were incubated for 20 min at 72°C with 4 µl of Taq DNA polymerase. PCR products then were digested with *Hpy*188I and cloned in a TOPO TA vector (Invitrogen). Independent clones containing the full-length PCR product were sequenced to evaluate mutations at the cleavage site.

The TALENs targeting the Arabidopsis *ADH1* gene were subcloned into the plant expression vector pFZ14 (Zhang *et al.*, 2010) using *Xba*I and *Sac*I. These enzymes excise the entire TALEN from pTAL3 or pTAL4 and place the coding sequence under control of the CaMV (cauliflower mosaic virus) 35S promoter. Recombinant plasmids were transformed into Arabidopsis protoplasts as previously described (Zhang *et al.*, 2010). Forty-eight hours after transformation, DNA was prepared and digested with *Pfl*FI, which

cuts in the spacer sequence of the TALEN target site. After digestion, a chromosomal fragment encompassing the target site was amplified by PCR, and the reaction products were once again digested with *Pfl*FI and run on an agarose gel. The band corresponding in size to undigested product was excised and cloned, and individual clones were sequenced to evaluate mutations at the cleavage site.

Expression of a custom TAL effector in *Xanthomonas* and *in planta* activity assay

An analog of *avrHah1* was assembled into pTAL1 using the Golden Gate method with HD, NI, NG, and NN modules, ordered to match the AvrHah1 binding site in the promoter of the *Bs3* gene (Schornack *et al.*, 2008). A native *avrHah1* construct was made by replacing the *Bam*HI fragment of *tal1c* in pCS495 with that of *avrHah1*. pCS495 is *tal1c* preceded by Shine-Dalgarno and Kozak consensus sequences in pENTR-D (Invitrogen). The analog and native *avrHah1* constructs and *tal1c* were moved into pKEB31 by Gateway cloning (LR reaction). pKEB31 Is a derivative of pDD62 (Mudgett *et al.*, 2000) that contains a Gateway destination vector cassette (Invitrogen) between the *Xba*I and *Bam*HI sites and a tetracycline resistance gene in place of the gene for gentamycin resistance. The resulting plasmids were introduced into *X. campestris* pv. vesicatoria strain 85-10 by electroporation, and transformants were inoculated to 6 wk. old pepper plants by syringe infiltration, as described (Schornack *et al.*, 2008). After 48 hours, infiltrated leaves were cleared in 70% ethanol and 10% glycerol and photographed.

Results

Efficient assembly of custom repeat arrays into TALEN and other TAL effector-based constructs

Our implementation of the Golden Gate method accomplishes custom TAL effector construct assembly in two steps (Figure 3-2 and Figure 3-4). In the first step, it uses five sets of 10 staggered repeat clones, one for each of the four most common RVDs HD, NI, NG, and NN, which associate most frequently with C, A, T, and G, respectively, and one for the less common NK, which at least in some contexts appears to have higher specificity for G than NN does (Miller *et al*., 2011, Morbitzer *et al.*, 2010). Inserts in these "module" plasmids carrying the desired RVDs are released and assembled in order in one or two sets of 10 and one set of 1-10 into "array" plasmids, using a type IIS enzyme. In the second step, the resulting array fragments are joined, along with a final, truncated repeat from a collection of five "last repeat" plasmids (one for each RVD), into any of four different "backbone" plasmids, using a different type IIS enzyme, for a final array of 12 (10 + 1 + the last) to 31 (10 + 10 + 10 + the last) RVDs. Counting the 5' T that precedes the RVD specified sequences in TAL effector binding sites, the corresponding target ranges from 13 to 32 nucleotides.

The backbone plasmids include 1) pTAL1 for assembling a custom TAL effector gene preceded by Shine-Dalgarno and Kozak sequences for efficient translation in bacteria and eukaryotes, respectively; 2) pTAL2, identical to pTAL1, but without a stop codon so that the effector can be fused to other protein domains; 3) pTAL3 for assembling a custom TALEN and expressing it in yeast using the selectable marker *HIS3*;

and 4) pTAL4, identical to pTAL3 but containing the marker *LEU2*, so that two TALEN monomers can be paired in the yeast assay (see below). The TAL effector constructs are flanked by *attL* sites for transfer by Gateway recombination (Invitrogen) into destination vectors of choice. The TALEN constructs, though not Gateway compatible, are flanked by restriction enzyme sites convenient for subcloning into different expression vectors. All constructs retain the internal *Sph*I sites flanking the repeat domain as well as the *Bam*HI sites farther out that are conserved in most TAL effectors and can be used to readily swap a custom array into other TAL effector based constructs.

All of the array and backbone plasmids contain within the cloning site the *lacZ* gene for blue/white screening to identify recombinants (Sambrook *et al.*, 1989). For the work presented here, we successfully assembled more than 30 custom TALENs (Table 3-1) and one custom TAL effector, ranging in array length from 15 to 30 RVDs. We never failed to obtain the correctly assembled array plasmid clone or the correctly assembled, final backbone plasmid clone for any of these by screening only three white colonies per cloning reaction transformed into *E. coli.* We routinely pick just two colonies, and usually both are correct (not shown). Assembly of one or more constructs takes just five days (Figure 3-3 and Materials and Methods).

Guidelines and software for TALEN site selection and repeat array design

To facilitate TALEN design for genome editing, we wrote a computer program that analyzes DNA sequences, identifies suitable, paired and opposing TAL effector target sites across a spacer, and generates corresponding RVD sequences using the four most common RVDs (see Materials and Methods). The software uses guidelines for TAL

effector targeting that reflect naturally occurring TAL effectors and their binding sites, and spacer lengths that we observed to function well in our previous study using TALENs derived from naturally-occurring TAL effectors (Christian *et al.*, 2010). We established the targeting guidelines by examining the 20 TAL effector-target pairs identified by Moscou and Bogdanove (Moscou & Bogdanove, 2009). We looked for positional biases, neighbor effects, and overall trends in nucleotide and RVD composition. To examine position effects for sequences of different lengths, we confined the analysis to the five positions at either end. We compared observed nucleotide and RVD frequencies to expected frequencies, taken as the frequencies in the entire set of sequences (Figure 3-7). The binding sites showed a strong bias against T at position 1 (5' end), a bias against A at position 2, biases against G at the last and next-to-last (3') positions, and a moderate bias for T at the last position. RVD sequences showed corresponding positional biases: NG was disfavored at position 1; NI was disfavored at position 2; and NG was favored and NN disfavored at the last position. The bias for NG at the last position was particularly striking: NG occurs at this position in 85% of the sequences compared to its overall observed frequency of 18%. No neighbor effects were detected in the binding sites or RVD sequences. Average nucleotide composition of the binding sites was 31 +/- 16 % A, 37 +/- 13 % C, 9 +/- 8 % G, and 22 +/- 10% T. To expand on this dataset, we used the weight matrix developed by Moscou and Bogdanove (Moscou & Bogdanove, 2009) to identify the best-scoring binding sites (preceded by a T) for each of 41 *X. oryzae* TAL effectors in each of approximately 57,000 rice promoters. We retained those in genes shown by microarray analysis (www.plexdb.org, experiment OS3) to be up-regulated during infection. This analysis yielded close to 100 putative

additional TAL effector-target pairs. These reflected the same positional biases (not shown). The guidelines are therefore as follows: 1) As noted previously for TAL effector binding sites (Boch *et al.*, 2009, Moscou & Bogdanove, 2009), TALEN monomer binding sites should be preceded by a 5' T; 2) they should not have a T at position 1; 3) they should not have an A at position 2; 4) they should end with a T, so that the corresponding TALENs will reflect the strong bias for NG at this position; and 5) they should have a base composition within two standard deviations of the averages we observed.

We did not systematically test the guidelines, but data from intermediate constructs we obtained while building full-length TALENs with our earlier sequential ligation method provide some support (Table 3-2). Of the four intermediate length TALEN-target pairs showing no detectable activity in the yeast assay for DNA cleavage (Christian *et al.*, 2010), one did not match overall target nucleotide composition, one did not have an RVD sequence ending in NG, and another did not meet either of these guidelines. Two out of seven with activity <25% of the Zif268 ZFN used as a control did not match overall target nucleotide composition. One of four with activity 25-50% of Zif268 did not have an RVD sequence ending in NG. TALENs with 50% or greater activity of Zif268 m*et al*l of the guidelines. The impact of the number of repeats in a TALEN was also considered. In general, longer TALENs that m*et al*l of the guidelines or medium-length TALENs that m*et al*l guidelines and had a high percentage of HDs showed the highest activity. Longer TALENs that failed to meet one or more guidelines showed reduced activity when compared to those of the same length that m*et al*l

guidelines. Thus in addition to providing preliminary support for the guidelines, the results also suggest that array length positively correlates with activity.

Toward validating our method for making custom TAL effector arrays, we used the software to first identify candidate TALEN sites in seven plant (Arabidopsis, tobacco), animal (human, zebrafish, *Drosophila*) and protist (*Plasmodium*) genes as well as in GFP and eGFP. In these genes, the software found unique TALEN sites on average every 35 bp (range = 15-120 bp).

Activity of custom TALENs in a yeast-based DNA cleavage assay

Custom TALEN pairs for fifteen target sites (30 TALENs total; Table 3-1) were made using the Golden Gate method and plasmids described above and tested in the yeast-based DNA cleavage assay we described previously (Christian *et al.*, 2010). All TALEN pairs showed significant activity above the target-only negative controls, and 14 of 15 showed activity greater than or equal to 25% of our positive control, a Zif268 ZFN (Figure 3-8). We have generally found for ZFNs that this level of activity is sufficient for targeted mutagenesis of endogenous plant loci (Townsend *et al.*, 2009, Zhang *et al.*, 2010).

Targeted mutagenesis in human cells and Arabidopsis protoplasts using custom TALENs

To validate the activity of our custom TALENs outside of yeast, we used one of the TALEN pairs for the human *HPRT1* gene (*HPRT1* B in Figure 3-8) and the TALEN pair for the Arabidpsis *ADH1* gene and to carry out targeted mutagenesis in human

embryonic kidney cells and Arabidopsis protoplasts, respectively. In both cases the custom TALENs generated mutations at the recognition site through imprecise repair of the cleaved chromosomes by non-homologous end-joining (Figure 3-9). Our method of detection used an enrichment step, so it was not possible to quantify mutagenesis frequency. However, we obtained for *HPRT1* 17 independent mutations including two single base pair substitutions and deletions ranging from 1-27 bp roughly centered on the spacer, and for *ADH1* six independent mutations consisting of deletions ranging from 4 to 15 bp, also centered on the spacer.

Replication of AvrHah1 TAL effector activity with a Golden Gate assembled clone

To assess our plasmids for construction of custom TAL effectors, we assembled an analog of the *avrHah1* TAL effector gene of *X. gardneri*, which elicits a hypersensitive reaction in pepper by transcriptionally activating the *Bs3* resistance gene (Schornack *et al.*, 2008). We chose AvrHah1 because it is highly divergent relative to other characterized TAL effectors, carrying predominantly 35 amino acid repeats (in contrast to the more common 34 amino acid repeat on which our modules are based) as well as other deviations from the consensus sequences both within and outside the repeat region. Introduced into *X. campestris* pv. vesicatoria strain 85-10, which lacks AvrHah1, that was then inoculated into pepper leaves, the Golden Gate assembled clone triggered a *Bs3* specific hypersensitive reaction indistinguishable from that elicited by the native effector (Figure 3-10). This recreation of AvrHah1 specificity using our modular reagents demonstrates their utility for making custom transcription factors and underscores the sufficiency of the RVD sequences for targeting.

Discussion

The hallmark feature of TAL effectors that makes them such remarkably powerful tools for DNA targeting, their long arrays of 33-35 amino acid repeats that specify nucleotides in the recognition site in a straightforward and modular fashion, also make them challenging to engineer. Commercial synthesis is effective (Miller *et al*., 2011) but costly. PCR based methods (Zhang *et al.*, 2011a) carry the risk of artifact and recombination. Assembly by sequential ligation of sequence-verified modules (Christian *et al.*, 2010) is cheap and assures array integrity but is time consuming. The Golden Gate method using the reagents we describe here provides a cost-effective, robust, and rapid solution. TAL effector constructs with arrays of up to 31 RVDs are assembled in just two cloning steps using a set of sequence-verified modules. Furthermore, the reagents provide great flexibility for cloning arrays in different contexts and expressing them in different organisms, either in our set of backbone plasmids for TALENs, TAL effectors, or TAL effector fusions to additional proteins, or by simple subcloning or Gateway recombination into other vectors.

Zhang and colleagues recently presented a protocol and set of templates for Golden Gate-like assembly that involves PCR amplification of modules, intermediary arrays, and full length arrays to yield TAL effector DNA binding domains with 13 RVDs fused in a backbone vector to VP64 (see also www.taleffectors.com). This marked a significant advance that enabled the authors to rapidly assemble custom arrays and demonstrate the utility of TAL effector-based proteins as custom transcription factors to

activate endogenous genes in human cells. However, the method and plasmids we describe here offer more versatility for broader utility, not only with regard to the available contexts and portability of the arrays, as noted above, but also in array length. The ability with our reagents to construct arrays ranging from 12 to 31 RVDs allows fine tuning for targeting, and will be important for testing the important outstanding question of the relationship of length to affinity and specificity. The broad range in array length also offers greater flexibility to systematically address other important questions including the contributions of individual RVD-nucleotide associations to affinity and specificity, as well as the effect of position on mismatch tolerance (Bogdanove *et al.*, 2010b). This could be accomplished for example by starting with an array of minimal functional length and comparing the effects of adding or interspersing additional RVDs aligned to different nucleotides in the target.

Our method has the technical advantage of involving no PCR. Although the Zhang *et al*. repeat templates for different RVDs are codon engineered to guard against slippage and inter-repeat recombination during PCR amplification, this strategy does not prevent recombination between repeats carrying the same RVD, particularly if they are present in tandem. Also, in part because our method involves no PCR, though it is two days longer it is less labor intensive and time consuming day to day.

Though all of the custom arrays made for this study use just the four most common RVDs, our plasmids set includes modules with NK, which users might opt to substitute for NN to specify G, because NN sometimes associates with A. We note however, based on data presented by Miller *et al*., that NK also associates substantially

with A in some contexts. Modules with yet additional RVDs can be generated readily by mutagenesis of an existing set.

Among the genes we selected for targeting with TALENs, we deliberately chose some for which targeting with ZFNs has proven difficult. For example, one of the most common mutations in patients with cystic fibrosis is a deletion of three nucleotides (DF508) in *CFTR*; however, best efforts to engineer a ZFN for this position only succeeded in targeting a site more than 120 bp away, a distance that would likely compromise gene targeting efficiency (Maeder *et al.*, 2008). For our CFTR TALENs, the DF508 mutation resides within the spacer sequence at the site of TALEN cleavage. Similarly, we previously created herbicide resistant tobacco plants by gene targeting with ZFNs that recognize and cleave the acetolactate synthase gene (Townsend *et al.*, 2009). The nearest ZFN that could be engineered to the desired site of modification was 188 bp away, whereas our TALENs cleave within 10 bp of the desired sequence modification. Finally, AT-rich sequences have been difficult to target with ZFNs; we successfully targeted two sites in the AT-rich (75.5%) Plasmepsin V gene of *Plasmodium falciparum*, which has an overall genome content of 80.6% AT (Gardner *et al.*, 2002). Generally, the high success rate of TALENs designed using our software, which found sites in diverse sequences on average every 35 bp, suggests that targetability of TALENs will prove superior to the public ZFN platforms, which are estimated to be capable of targeting on average every 500 bp (Maeder *et al.*, 2008, Sander, Dahlborg, *et al.*, 2011). Indeed, we anticipate our estimate of targeting range is conservative, as some TALENs that do not follow our design principles still recognize and cleave DNA efficiently (Table 3-2) (Miller *et al*., 2011).

Activity varied among the TALENs we tested in the yeast assay. The reason for this is not clear. It could relate to expression levels or variability in the assay itself, but more likely, the data reflect inherent differences in the DNA binding affinity of the arrays, possibly related to their length and composition. The relationship of array length and composition to overall affinity is still an open question that must be addressed. The important conclusion for this study is that all of the TALENs were active, demonstrating that the targeting approach as well as the Golden Gate methods and plasmids for assembly are robust. And, our results in Arabidopsis protoplasts and human cells, along with recent results from other groups (Mahfouz *et al.*, 2010, Miller *et al*., 2011), indicate that TALENs are likely to be broadly effective for genome engineering.

We have deposited all of our plasmids for constructing and expressing TALENs, as well as TAL effectors with or without a stop codon in the non-profit clone repository AddGene (www.addgene.org). To complement our method and reagents we have also made our software for TALEN site selection and design freely accessible as an online tool, the TAL Effector Nucleotide Targeter at http://boglabx.plp.iastate.edu/TALENT/. Although our success rate was high with TALENs designed using the software, we have not shown that it is *necessary* to follow the guidelines on which the software is based. So, even though the guidelines place only relatively minor constraints on targeting, the online tool allows users to exclude them individually to increase candidate target site frequency. Also, because optimal spacing may differ for different TALEN architectures, the software provides the option to specify desired spacer lengths. In making these resources available, we hope to facilitate further characterization of TAL effector DNA targeting

properties, broad adoption of TALENs and other TAL effector-based tools, and further development of the utility of these unique DNA binding proteins.

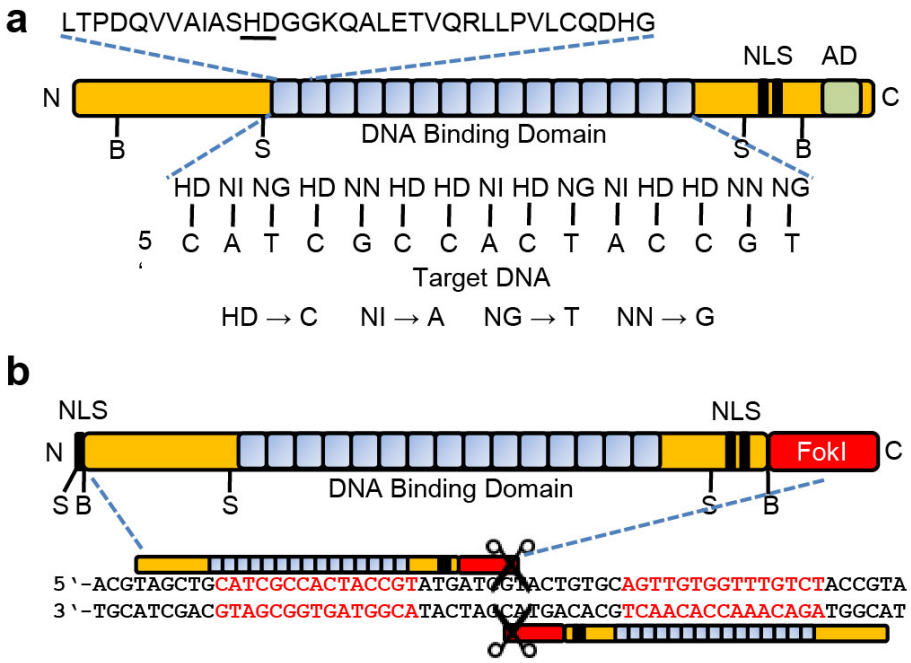


Figure 3-1. **TAL effector and TALEN structure.** (a) Structure of a naturally occurring TAL effector. A consensus repeat sequence is shown with the repeat variable diresidue (RVD) underlined. The sequence of RVDs determines the target nucleotide sequence. The four most common RVDs, on which our designs and plasmids are based, are shown with their most frequently associated nucleotide. Some evidence suggests that the less common RVD NK (not shown) has greater specificity for G than NN does, and for that reason our plasmid s*et al*so includes NK modules. (b) Structure of a TALEN. Two monomeric TALENs are required to bind the target site to enable *Fok*I to dimerize and cleave DNA.

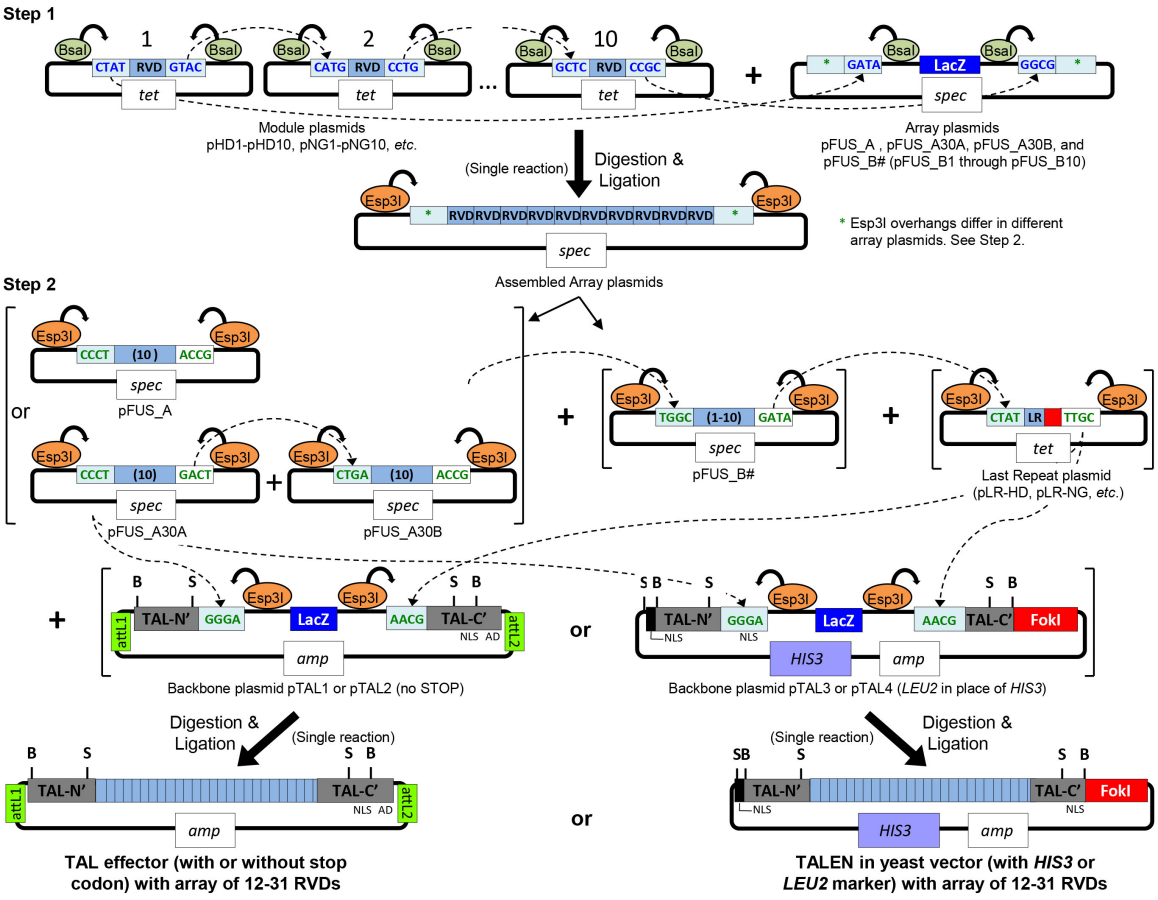


Figure 3-2. **Golden Gate assembly of custom and TAL effector and TALEN constructs using module, array, last repeat, and backbone plasmids.** By using the type IIS restriction endonucleases *Bsa*I and *Esp*3I, modules containing the desired RVDs can be released with unique cohesive ends for ordered, single-reaction assembly into array plasmids in a first step, and those arrays subsequently released and assembled in order in a second step into a backbone plasmid to create full length constructs with custom repeat arrays (see text for details). NLS, nuclear localization signal(s); AD, transcriptional activation domain; *tet*, tetracycline resistance; *spec*, spectinomycin resistance; *amp*, ampicillin resistance; attL1 and attL2, recombination sites for Gateway cloning; B, *Bam*HI, and S, *Sph*I, useful for subcloning custom repeat arrays. Unique restriction enzyme sites flanking the coding seqeunces, useful for subcloning the entire constructs into other vectors, are not shown.

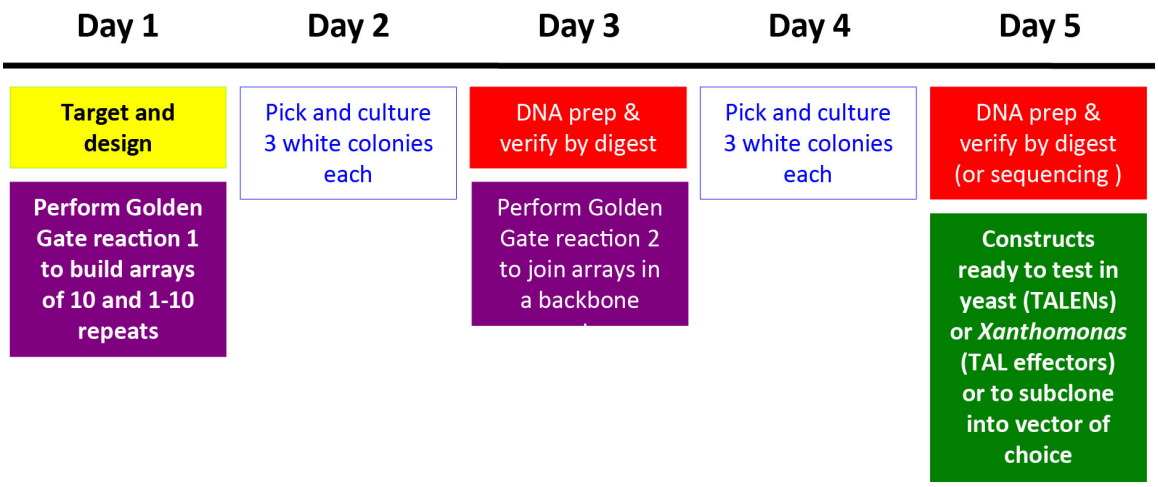


Figure 3-3**. TALEN or TAL effector construct assembly timeline.**

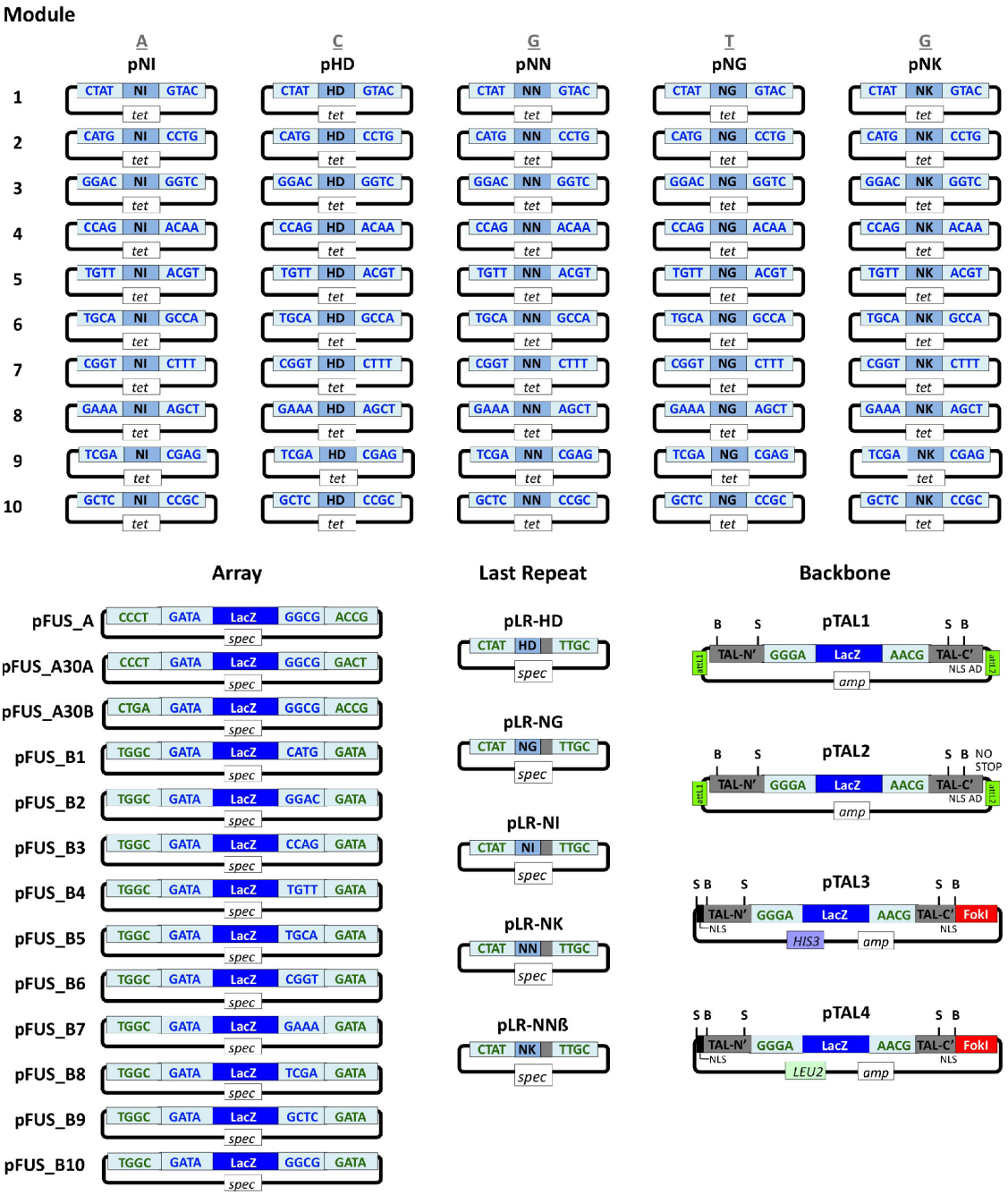
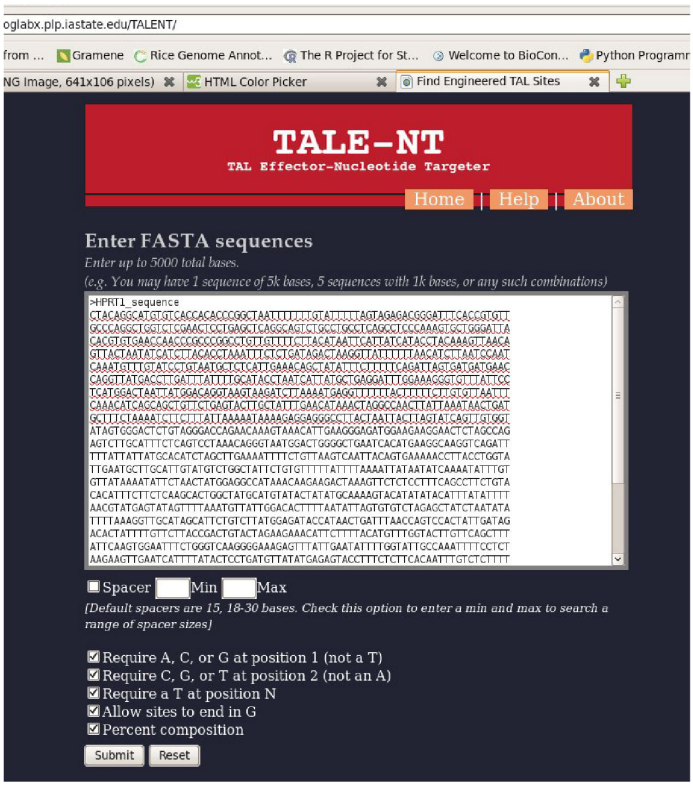


Figure 3-4 **Supplemental figure of module, array, last repeat, and backbone plasmids.** Overhangs left by *Bsa*I digestion are shown in blue font and by *Esp*3I in green. HD modules 2-10 contain a *Bsp*EI site (not shown) not present in the the other modules. NLS, nuclear localization signal(s); AD, transcriptional activation domain; *tet*, tetracycline; *spec*, spectinomycin ; *amp*, ampicillin ; attL1 and attL2, recombination sites

for Gateway cloning; B, *Bam*HI, and S, *Sph*I. Unique restriction enzyme sites flanking the entire coding sequences of the backbone constructs are not shown but can be found in the sequence files.

a



b

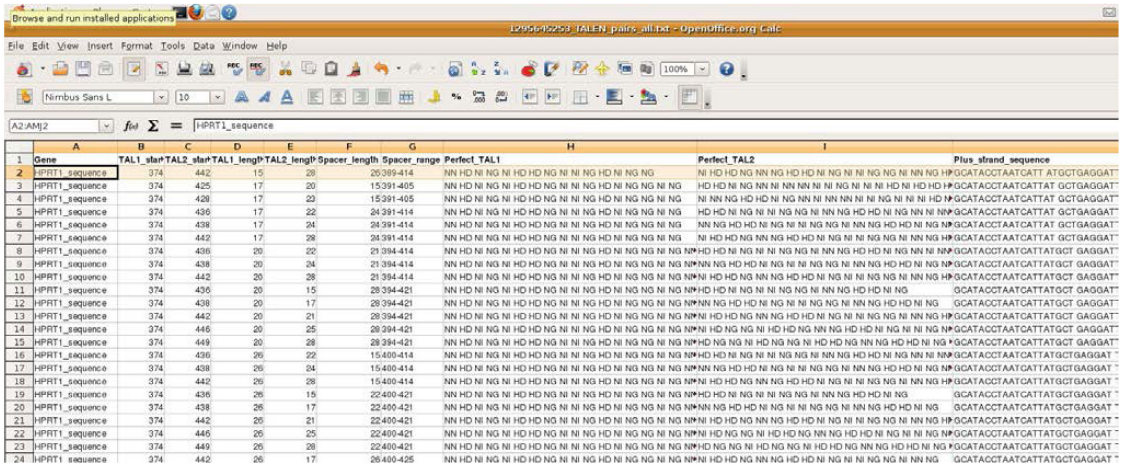


Figure 3-5. Supplemental figure of TAL Effector-Nucleotide Targeter. (a) Screenshot showing the DNA sequence input window and options to include or exclude specific guidelines and to specify spacer lengths. (b) Sample output for the human *HPRT1* gene showing candidate target sequences and corresponding TALEN RVD sequences. Highlighted in the top row is the *HPRT1* A (Fig. 3-8) target site information. The targeter URL is boglabx.plp.iastate.edu/TALENT/.

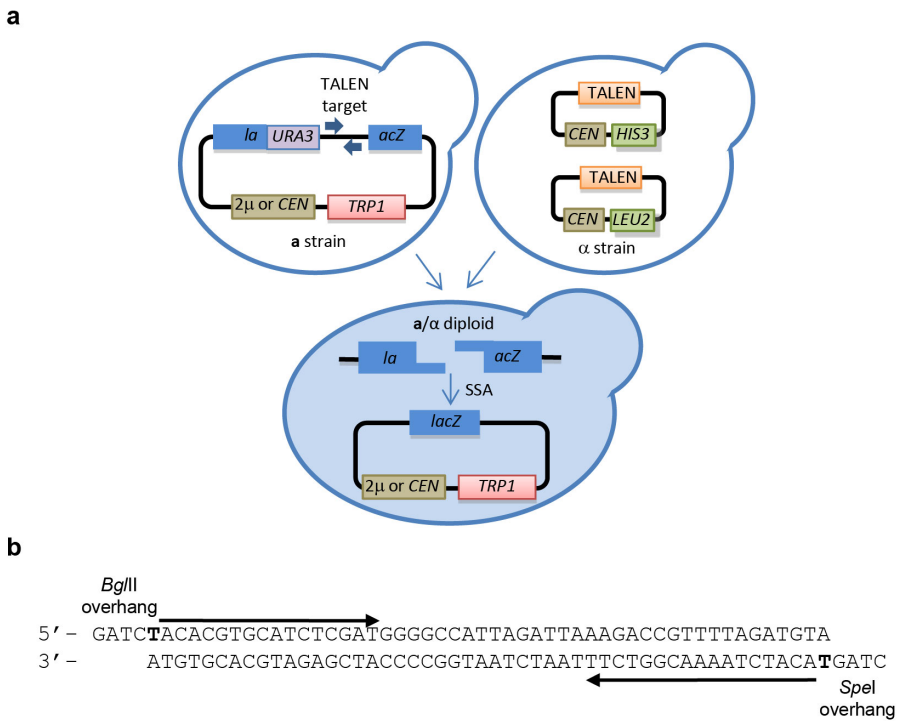


Figure 3-6. Supplemental figure of the yeast assay for testing TALEN function. (a) Schematic of the assay (see text for details). SSA, single-strand annealing. (b) Example of target site oligonucleotides that are synthesized, annealed and cloned into the target plasmid pCP5. Arrows indicate the opposing TALEN monomer binding sites, separated by a spacer The T preceding each site is in bold.

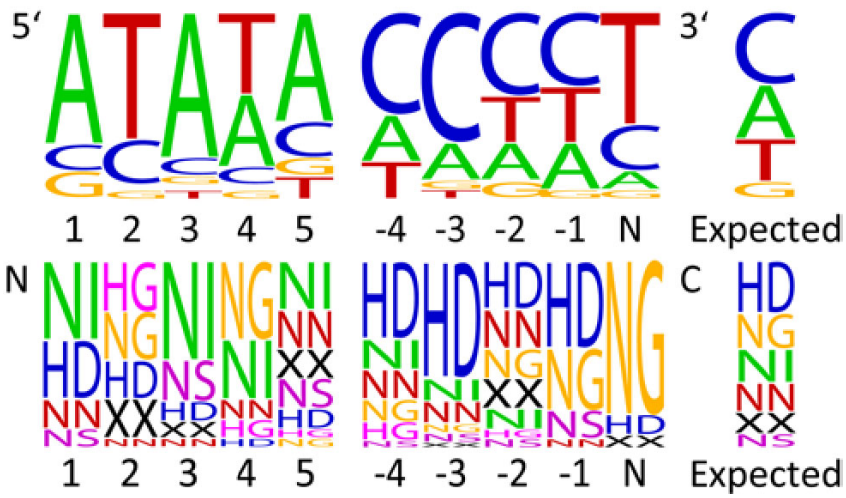


Figure 3-7. Nucleotide and RVD frequencies at the termini of 20 target and TAL effector pairs. RVDs that have a frequency of 20% or greater at one or more of the positions are shown. "XX" represents all other RVDs.

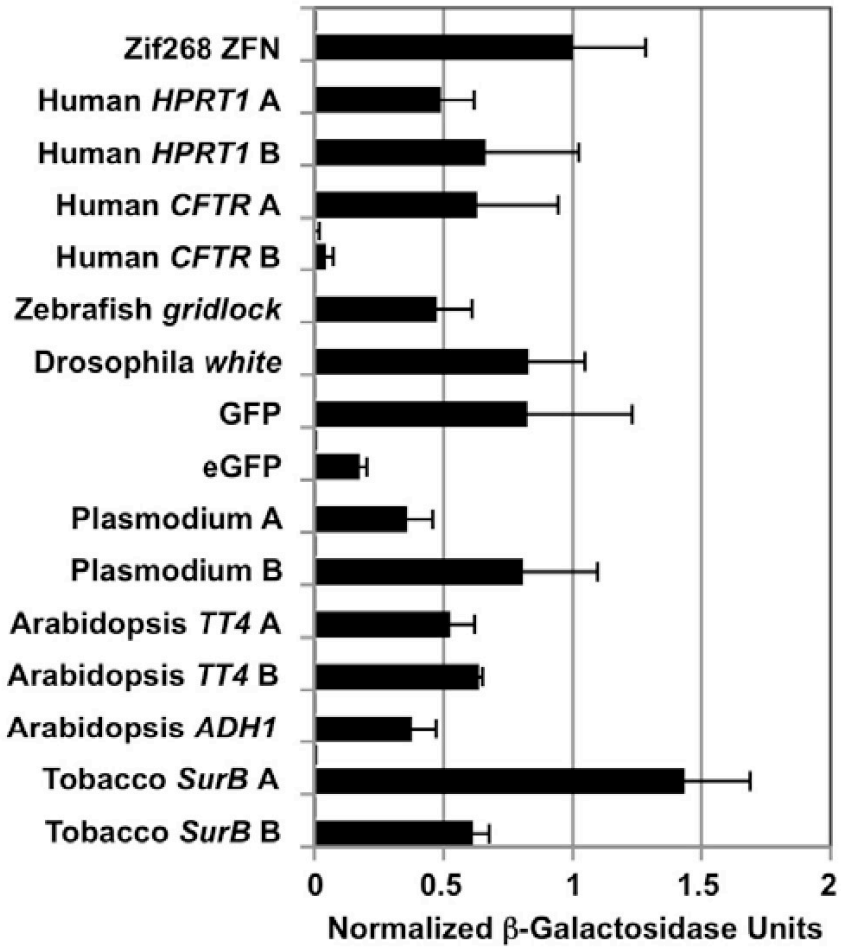


Figure 3-8. Activity of 15 custom TALEN pairs targeting diverse sequences in a reporter based yeast assay. TALENS were targeted to gene sequences from the indicated organisms and to *GFP* and *eGFP* using the software and constructed using the Golden Gate method and plasmids described in the text. Activity was measured in a yeast-based assay in which cleavage and recombination reconstitutes a functional *lacZ* gene (see text for details). Activity was normalized to a Zif268 ZFN positive control. Activity of target-only controls for each is plotted above the target-plus-TALEN values; in each case the activity was undetectable. Error bars denote s.d.; n = 3.

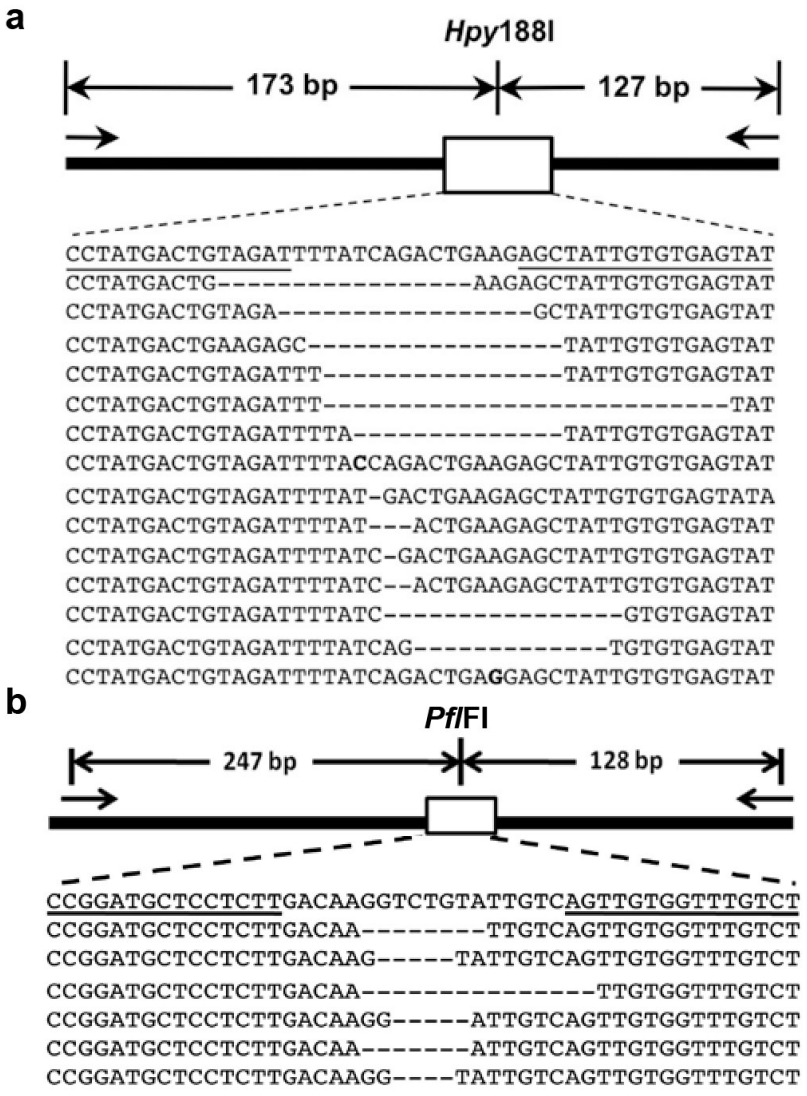


Figure 3-9. Site-directed mutagenesis in human embryonic kidney cells and Arabidopsis protoplasts using custom TALENs. TALENS targeted to the human *HPRT1* gene (pair HPRT1 B in Fig. 3-8) and the Arabidopsis *ADH1* gene (Fig. 3-8) were transiently expressed in human embryonic kidney cells and in Arabidopsis protoplasts, respectively, and the targets subsequently amplified and sequenced (see text for details). Prior to amplification, genomic DNA was digested with a restriction endonuclease having a site present in the TALEN target site to reduce amplification of wild type sequences and enrich the amplicon pool for mutated ones. Results for *HPRT1* are shown in panel a

and *ADH1* in panel b. For each, the schematic at the top shows the chromosomal locus, short arrows designate primers used for PCR amplification following TALEN transient expression, sequence of the wild type gene (top line) and unique mutated alleles obtained are shown below, binding sites for the TALEN monomers are underlined, and the coincident restriction endonuclease site is indicated.

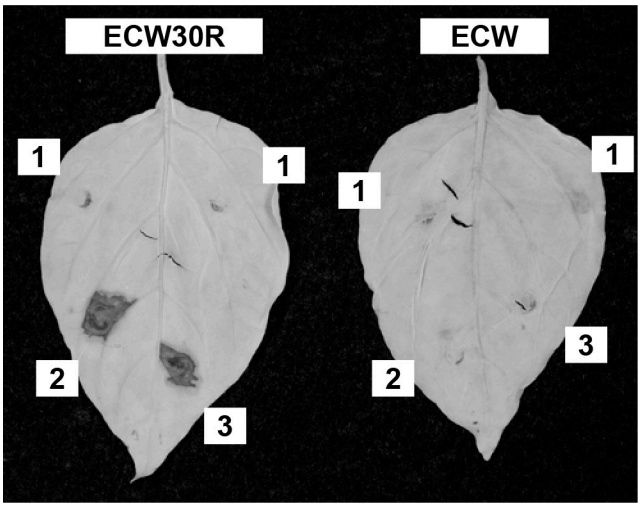


Figure 3-10. Activity of an AvrHah1 analog created using the Golden Gate method and our plasmid set. Shown are leaves of pepper varieties ECW30R, carrying the *Bs3* resistance gene, and ECW, lacking it, 48 hours following spot-infiltration with suspensions of *X. campestris* pv. vesicatoria strain 85-10 transformed to deliver (1) Tal1c (the effector used to make the backbone plasmids in this study), (2) native AvrHah1, or (3) an AvrHah1 analog encoded by a construct made using the Golden Gate method and our plasmid set. Leaves were cleared with ethanol to reveal the accumulation of phenolic compounds, visible as dark stained areas, indicative of the hypersensitive reaction induced by TAL effector driven transcriptional activation of *Bs3*.

Table 3-1 Custom TALENs and target sites.

Gene	Organism	TAL1, TAL2 Plasmids	TAL1, TAL2 Repeat number	Spacer	TAL1, TAL2 RVD Sequences	DNA Target Sequence with Spacer (TAL1, TAL2 underlined)
HPRT1 A	Human	pTAL141N, pTAL142N	17,20	15	TAL1: NN HD NI NG NI HD HD NG NI NI NG HD NI NG NG NI NG TAL2: HD HD NI NG NN NI NN NN NI NI NG NI NI NI HD NI HD HD HD NG	5'- GCATACCTAATCATTAT GCTGAGGATTTGGAA AGGGTGTTTATTCCTCATGG
HPRT1 B	Human	pTAL143N, pTAL144N	20,17	15	TAL1: HD HD NI NG NG HD HD NG NI NG NN NI HD NG NN NG NI NN NI NG TAL2: NI NG NI HD NG HD NI HD NI HD NI NI NG NI NN HD NG	5'- CCATTCCTATGACTGTAGAT TTTATCAGACTGAAG AGCTATTGTGTGAGTAT
CFTR A	Human	pTAL148N, pTAL149N	29,30	28(25)	TAL1: NN NN NI NG NG NI NG NN HD HD NG NN NN HD NI HD HD NI NG NG NI NI NI NN NI NI NI NI NG TAL2: NN NN HD NI NG NN HD NG NG NG NN NI NG NN NI HD NN HD NG NG HD NG NN NG NI NG HD NG NI NG	5'- GGATTATGCCTGGCACCATTAAAGAAAAT ATCATCTTTGGTGTTTCCTATGATGAAT ATAGATACAGAAGC

						GTCATCAAAGCATG CC
CFTR B	Human	pTAL150N, pTAL151N	23,28	25(22)	TAL1: NN NN HD NI HD HD NI NG NG NI NI NI NN NI NI NI NI NG NI NG HD NI NG TAL2: NN NN HD NI NG NN HD NG NG NG NN NI NG NN NI HD NN HD NG NG HD NG NN NG NI NG HD NG	5'-GGCACCATTAAAGA AAATATCAT CTTTGGTGTTTCCTA TGATGAATAT AGATACAGAAGCGT CATCAAAGCATGCC
gridlock	Zebrafish	pTAL112, pTAL117	17,16	18	TAL1: NI NG NG NG HD NN NI HD NN HD NG HD NI NG NG HD NG TAL2: HD HD HD NN NN NI NI NN HD HD NN NI NG NN HD NG	5'-ATTTCGACGCTCATT CT CTGGCCATGGACTTC TTG AGCATCGGCTTCCG GG
white	Drosophila	pTAL152N, pTAL153N	22,24	24	TAL1: HD HD NG NN NI HD HD NI NG NG HD NI NG HD NI NN HD HD NN NG HD NG TAL2: NN HD HD HD NG HD NN NN HD HD NI NG HD NI NN NI NI NN NN NI NG HD NG NG	5'-CCTGACCATTCATCA GCCGTCT TCCGAGCTGTTTGAG CTCTTTGAC AAGATCCTTCTGATG GCCGAGGGC
GFP	Jelly Fish	pTAL156N, pTAL157N	26,25	25	TAL1: HD HD NI NG NN NN HD HD NI NI HD NI HD NG NG NN NG HD NI HD NG NI HD NG NG NG TAL2: NN HD HD NN HD NG NG HD NI NG NI NG NN NI NG HD NG NN NN NN NG NI NG HD NG	5'-CCATGGCCAACACT TGTCACTACTTT CTCTTATGGTGTTCA ATGCTTTTCA AGATACCCAGATCA

						TATGAAGCGGC
eGFP	Jelly Fish	pTAL158N, pTAL159N	24,26	21	TAL1: NI HD NI NI HD NG NI HD NI NI HD NI NN HD HD NI HD NI NI HD NN NG HD NG TAL2: HD NG NG NN NI NI NN NG NG HD NI HD HD NG NG NN NI NG NN HD HD NN NG NG HD NG	5'- ACAACTACAACAGC CACAACGTCT ATATCATGGCCGAC AAGCAGA AGAACGGCATCAAG GTGAACTTCAAG
Plasmepsin A	Plasmodium	pTAL160N, pTAL161N	19,22	15	TAL1: HD NN HD NI NI NI NN NI NI NG NG NG HD NG NG NG NI NI NG TAL2: NI HD NI HD NN NN NN NI NI NI HD NG NG NI NI HD NN NI NI NN NI NG	5'- CGCAAAGAATTTCTT TAAT TCTAGATACAGGTTC ATCTTCGTTAAGTTT CCCGTGT
Plasmepsin B	Plasmodium	pTAL162N, pTAL163N	23,24	15	TAL1: NI NG NG NI NG NG NN NN HD HD NG NG NI NI HD HD NI NG NI HD NG HD NG TAL2: HD HD NI NG NN NG NI NI NI NI HD NI NG NG NG NG NG HD NG NN NI HD NI NG	5'- ATTATTGGCCTTAAC CATACTCT TATCTCTTATTCTTT ATGTCAGAAAAATG TTTTACATGG
TT4 A	Arabidopsis	pTAL164N, pTAL165N	19,22	15	TAL1: HD HD NI NG HD NG HD HD NG HD NI NI NN NN NI NG NN NG NG TAL2: HD NG NI NN NI HD NG HD NG NG HD NI HD NI NI NG NN NG NG HD NG NG	5'- CCATCTCCTCAAGG ATGTT CCCGGCCTCATCTCC AAGAACATTGTGAA GAGTCTAG
TT4 B	Arabidopsis	pTAL166N, pTAL167N	15,16	15	TAL1: NN NG HD NN NG HD NG NG HD NG NN HD NI HD NG	5'- GTCGTCTTCTGCACT

					TAL2: NI NN NG HD NI NN HD NI HD HD NI NN NN HD NI NG	ACCTCCGGCGTCGAC ATGCCTGGTGCTGACT
ADH1	Arabidopsis	pTAL 69N, pTAL 74N	15,15	18	TAL1: HD HD NN NN NI NG NN HD NG HD HD NG HD NG NG TAL2: NI NN NI HD NI NI NI HD HD NI HD NI NI HD NG	5'-CCGGATGCTCCTCTT GACAAGGTCTGTAT TGTC AGTTGTGGTTTGTCT
SuRB A	Tobacco	pZHY096 pZHY106	17,20	15	TAL1: NI HD NG NN NI NI NG NI NI NG HD NI NI HD NI HD NG TAL2: NI NG NI NN NI NI HD HD NN NI NG HD HD NG HD HD HD NI NG NG	5'-ACTGAATAATCAAC ACT TGGGAATGGTGGTT C AATGGGAGGATCGG TTCTAT
SuRB B	Tobacco	pZHY108 pZHY109	19,19	15	TAL1: NI HD HD NG HD NI NG HD NI NN NN NI NI HD NI NG NN NG NG TAL2: HD NG NG NG NN NI NI NI NN HD HD HD HD NN HD HD NI HD NG	5'-ACCTCATCAGGAAC ATGTT CTACCTATGATTCCC AGTGGCGGGGCTTT CAAAG

Table 3-2. Activity, conformity to rules, and length of TALENs tested in the yeast assay.

TALEN	RVDs	Activity	% GATC	Ends in NG	RVD Sequence
ADH1_1R12	12	-	y	n	HD HD HD NI NN NI NI NN NG NI NI NI
ADH1_2R12	12	-	n	n	NI NN NI HD NI NI NI HD HD NI HD NI
tt4_1L15	15	-	y	y	NN NN HD NI HD NG NN HD NG NI NI HD HD HD NG
telo_1R15	15	-	n	y	HD NN NN NG NG NG NN HD NG NI NG HD NN NG NG
ADH1_1R15	15	+	y	y	HD HD HD NI NN NI NI NN NG NI NI NI HD NI NG
ADH1_2L12	12	+	y	y	HD HD NN NN NI NG NN HD NG HD HD NG
grid_1R10	10	+	n	n	NI HD HD HD HD NG HD NG HD HD
telo_1L15	15	+	y	y	HD NG NG NN NG HD HD NN HD NI NG NN NI NG NG
telo_2L16	16	+	y	y	NI NG NG HD HD HD HD NI HD NN NI NN HD NG HD NG
telo_2R17	17	+	n	y	NI NN NI HD NI NN NN NI NI NN NG NN NN NI NN HD NG
ADH1_1L12	12	++	y	n	NI NG HD NI NI NN NI NG NG HD NG HD
ADH1_2R15	15	++	y	y	NI NN NI HD NI NI NI HD HD NI HD NI NI HD NG
grid_1R13	13	++	y	y	NI HD HD HD HD NG HD NG HD HD NN HD NG
ADH1_2L15	15	+++	y	y	HD HD NN NN NI NG NN HD NG HD HD NG HD NG NG
grid_1L16	16	++++	y	y	NI HD HD HD HD NG HD NG HD HD NN HD NG NG HD NG

Activity: – no activity, + activity less than 25% of ZFN Zif268, ++ activity 25-50% of ZFN Zif268, +++ activity 50-75% ZFN Zif268, ++++ activity >75% ZFN Zif268.

CHAPTER 4

TARGETING G WITH TAL EFFECTORS: A COMPARISON OF ACTIVITIES OF TALENS CONSTRUCTED WITH NN AND NK REPEAT VARIABLE DI-RESIDUES

Reprinted from: Christian ML, Demorest Z, Osbourne M, Starker C, Nyquist M, Carlson D, Bradley P, Bogdanove AJ, and Voytas DF. 2012. Targeting G with TAL effectors: a comparison of activities of TALENs constructed with NN and NK repeat variable di-residues. Public Library of Science One 7(9).

Introduction

The ability to target proteins to specific DNA sequences makes it possible to manipulate nucleic acids *in vivo*, for example, by creating artificial transcriptional regulators that alter gene expression or by engineering sequence-specific nucleases that modify genetic loci of interest. Work in the past few years has made it increasingly evident that the protein domain of choice for DNA targeting is the DNA binding domain derived from Transcription Activator-Like (TAL) effectors (Bogdanove & Voytas, 2011). TAL effectors are proteins produced by bacterial plant pathogens of the genus *Xanthomonas*, and they are delivered to plant cells during infection where they activate the expression of select target genes and thereby make the plant more prone to bacterial colonization (Bogdanove *et al.*, 2010c). Targeting of TAL effectors to plant gene promoters is achieved by a simple and elegant mechanism of DNA binding (Deng, Yan, *et al.*, 2012, Mak, Bradley, Cernadas, *et al.*, 2012a). This mechanism enables the pathogen to rapidly evolve TAL effectors with new DNA sequence specificities, and by extension, it enables scientists to engineer DNA binding domains that recognize novel sites for various genome engineering applications.

DNA binding by TAL effectors is mediated by the central domain of the protein, which comprises approximately 13-28 tandem repeats of a 34 amino acid motif (Boch & Bonas, 2010). Amino acid sequences of the repeats are largely invariant with the exception of two residues at positions 12 and 13, called the repeat variable di-residues (RVDs). Each repeat forms two alpha helices that are joined at one end by a short loop that contains the RVD: residue 12 reaches back to interact with one of the alpha helices to stabilize the loop, and residue 13 contacts a specific base within the major groove of

DNA (Deng, Yan, *et al.*, 2012, Mak, Bradley, Cernadas, *et al.*, 2012b). Among the most common RVDs in TAL effectors are HD, NG and NI, which specify the nucleotides C, T and A, respectively; these RVDs are only infrequently associated with other bases (Boch & Bonas, 2010). The mechanistic basis for this sequence specificity has been explained by the recently reported structure of two TAL effectors bound to DNA (Deng, Yan, *et al.*, 2012, Mak, Bradley, Cernadas, *et al.*, 2012b). The most common RVD that specifies G, namely NN, also interacts well with A, and this interaction is achieved by hydrogen-bonding between the second asparagine in the RVD and the N_7 group on these purines (Deng, Yan, *et al.*, 2012, Mahfouz *et al.*, 2010, Mak, Bradley, Cernadas, *et al.*, 2012b).

The lack of specificity for G is one potential drawback in the use of engineered TAL effector proteins for DNA targeting. NK, an RVD only rarely found in nature, also interacts with G (Miller *et al*., 2011). In vitro SELEX data suggested that NK RVDs may have higher specificity for G than NN (Miller *et al*., 2011); however, DNA targeting to a locus in zebrafish was more effective when engineered TAL effector proteins were made using repeats with NN RVDs rather than NK (Huang *et al.*, 2011). Two recent studies also indicated that NK TAL effectors have less activity than NN-containing proteins, and further, the RVD NH was shown to provide G-specificity (Cong *et al.*, 2012, Streubel *et al.*, 2012). Clearly there is considerable interest in both understanding and increasing G-specificity of engineered TAL proteins.

Other efforts to adopt TAL effectors for genome engineering have focused on defining the minimal region of the protein required for DNA binding (Christian *et al.*, 2010, Huang *et al.*, 2011, Miller *et al*., 2011, Mussolino *et al.*, 2011). In the TAL effector PthXo1, for example, the repeat region is flanked by 288 and 295 amino acids at the N-

and C-termini, respectively. To delimit the DNA binding domain, different research groups have made and tested various N- and C-terminal truncations. An N-terminal truncation at residue 152 and C-terminal truncations up to 18 amino acids after the repeat domain still allow for effective DNA binding (Miller *et al*., 2011, Mussolino *et al.*, 2011). The structure of TAL effectors bound to DNA is consistent with truncations of this extent still being able to bind DNA (Deng, Yan, *et al.*, 2012, Mak, Bradley, Cernadas, *et al.*, 2012b).

The length of the C-terminus after the TAL effector DNA binding domain has a direct impact on the activity of TAL effector nucleases (TALENs) – fusions of the TAL repeat arrays to the catalytic domain of the FokI endonuclease (Christian *et al.*, 2010). FokI functions as a dimer, and DNA cleavage is achieved using two TALENs that bind opposing DNA target sites separated by a spacer. The length of the spacer must allow for efficient FokI dimerization so that DNA cleavage is achieved. In all TALEN architectures tested to date, DNA cleavage is occurs across a broad range of spacer lengths (Christian *et al.*, 2010, Huang *et al.*, 2011, Miller *et al*., 2011, Mussolino *et al.*, 2011), and this variability further motivates defining the minimal DNA binding domain so that TALENs can be engineered with maximal activity over a narrow range of spacer lengths.

Toward improving TAL protein engineering, we compared the activity of TALENs that use repeats with either NN or NK RVDs to target G. We conclude that whereas TALENs made with NK RVDs may have more specificity for G-containing targets, this specificity is offset by a considerable reduction in activity that is due to having less affinity for G than their NN-containing counterparts. We also compared the

activity of TALENs with various N- and C-terminal deletions and found a novel architecture that cleaves optimally over a narrower spacer length range than any reported to date.

Results

Activity of TALENs with NN versus NK RVDs

We compared the activities of engineered TALENs that specify the nucleotide G with either NN or NK RVDs. Using the Golden Gate assembly method for TALEN engineering (Cermak *et al.*, 2011), we constructed 15 TALEN pairs that incorporate either NN or separately NK RVDs to specify G. The total number of G's in the corresponding target DNA sequences ranged from 3 to 17, and no bias was imposed as to how the G's were distributed in the target. To test the activity of the TALENs, we used a yeast-based single-strand annealing (SSA) assay in which LacZ activity serves as an indicator of DNA cleavage by TALENs (Christian *et al.*, 2010, Townsend *et al.*, 2009) (Appendix A). Briefly, the assay employs a target plasmid containing a *lacZ* reporter gene with an internal sequence duplication disrupted by a TALEN target sequence. Cleavage of the target by the corresponding TALENs results in reconstitution of a functional *lacZ* gene whose expression is measured by standard LacZ enzymatic assays. TALEN pairs containing NN RVDs showed 14-90% activity of the positive control TALEN pair, SurB, which was slightly more active than the potent zinc-finger nuclease (ZFN), Zif268 (Figure 4-1). When NN was exchanged for the NK RVD, nuclease activity was reduced for 14 of 15 TALEN pairs, and for several NK-containing TALENs, no activity was detected. One exception was TALEN pair 248/249, which contained 6 NK

RVDs in total (2 in TALEN 248 and 4 in 249). The activities of both NN- and NK-versions of this TALEN pair were comparable. Another exception was 143/144 in which the NK version had slightly higher activity than the NN version.

We sought to validate the trends observed in yeast by testing several TALEN pairs made with either NN or NK RVDs in human cells. As in yeast, we used a SSA assay in which a luciferase gene is interrupted by an internal sequence duplication and a TALEN recognition site. When the TALEN cleaves the target, the break is repaired, restoring the luciferase open reading frame and thereby allowing for comparison and quantification of TALEN activity. Three target sites were tested, one of which (143/144) was an exception in the yeast assay and showed higher activity for the NK-containing TALEN pair. For all three targets, the activity of TALENs with NN versus NK RVDs recapitulated in human cells what was observed in yeast, although in some cases the magnitude of the difference was considerably less (Figure 4-2). For example, with TALEN pair 351/352, the NK version was 30-fold less active in yeast, whereas only 1.6-fold less activity was observed for the NK version in human cells.

TAL effector arrays containing NK RVDs have lower affinities for DNA targets

To understand the basis for the lower activity of NK-containing TALENs *in vivo*, we performed a series of electrophoretic mobility shift assays (EMSAs) using purified TAL effectors 166, 167, and 312, which contain 3, 4, or 5 NN or NK RVDs, respectively. Each NN- or NK-containing array was expressed as a GST fusion protein, purified from *E. coli*, and used to perform EMSAs with cognate targets or a scrambled substrate. All NN-containing arrays bound their target significantly better than the scrambled substrate. For TAL effector 166, both the NN- and NK-containing proteins bound target substrates

with similar affinities, as evidenced by the shift of the target probe in the presence of increasing concentration of protein (Figure 4-3A). TAL effector 312NN, which contains five NN/NK repeats, showed the most dramatic difference between binding affinities when compared to TAL effector 312NK. A shift of the target probe can be seen for TAL effector 312NN at 40nM of protein, while there is a lack of visible shift at the highest concentration (250nM) of TAL effector 312NK. Notably, TAL effector 166NN and 312NN also bound the scrambled substrate with greater affinity than the NK versions. TAL effector 167NN bound its target substrate with higher affinity than the NK-containing array.

The data for all six TAL effector proteins were plotted as the fraction of bound substrate at each protein concentration from 0nM to 250nM (Figure 4-3B). These data taken together indicate that substituting NK for NN RVDs does contribute to overall lower binding affinities of the TAL effector proteins. Furthermore, the TAL effector protein binding affinities correlate well with the nuclease activities of the corresponding TALENs *in vivo*.

Effects on TALEN activity of number and position of NK RVDs in repeat arrays

Having observed predominantly reduced nuclease activity for TALEN pairs with NK versus NN RVDs, we next evaluated the effect on activity of number and/or position of NK RVDs within the array. The Golden Gate TALEN assembly method allows rearrangement of RVDs within existing TALEN pairs with relative ease (Cermak *et al.*, 2011), and so we altered the NN/NK composition of the 166 and 167 TALENs to create four variants of each (Figure 4-4A). The 166 and 167 arrays and their variants were then tested as TALENs on homodimeric target sites (i.e. the target site for either the 166 or

167 array was duplicated in inverse orientation and separated by a 15 bp spacer). Two consistent observations were made: 1) as described above, TALENs with NK RVDs had the least activity compared to their NN-containing counterparts, and 2) the most significant impact on activity occurred when the NK RVDs were located within the N-terminal half of the array. This latter observation was particularly pronounced for NK substitutions in the 167 repeat array.

When the 166/167 array variants were tested together as heterodimeric TALENs the impact of the NK RVDs was diminished (Figure 4-4B**,** G-containing targets). For example, when the 166 array variants were tested with a fully NN-containing 167 array, no difference in TALEN activity was observed. Further, the 167-N variant, which showed 5-fold reduction in activity when tested as a homodimer, showed only a modest decrease in activity when tested with the fully NN-containing 166 array. This suggests that pairing with a functional TALEN partner can ameliorate NK substitutions that decrease TALEN activity.

NK substitutions were made in repeat arrays for a second pair of TALENs (272/273) (Figure 4-4B), and the same general observations were made with respect to the impact of the substitutions on TALEN activity. When both the 272/273 TALEN pairs contained all NK RVDs to specify G, little or no activity was observed. Activity was also compromised when a fully NN-containing TALEN 273 was paired with a TALEN 272 variant with three NK substitutions at the N-terminus. In contrast, four NK substitutions at the C-terminus of a 273 variant had no effect on activity when this TALEN was paired with a fully NN-containing TALEN 272. It appears that the number of NK substitutions also impacts activity, as a single NK substitution at the N-terminus of TALEN 273 did

not affect activity when paired with the NN-containing 272 TALEN. We conclude, therefore, that NK substitutions are more likely to compromise TALEN activity when located at the N-terminus of the protein and that the loss of activity is amplified when there are multiple NK substitutions.

Testing TALENs on G- and A-containing targets

A previous report showed a specificity of NK RVDs for G nucleotides over A nucleotides (Miller *et al*., 2011). In contrast, the NN RVD appears to recognize both G and A (Boch *et al.*, 2009, Deng, Yan, *et al.*, 2012, Mak, Bradley, Cernadas, *et al.*, 2012b, Miller *et al*., 2011). We therefore tested several of our TALEN variants against targets in which G's had been replaced with A's (Figure 4-4B, Table 4-1). The fully NN-containing 166/167 and 272/273 TALEN pairs showed comparably high levels of activity on both G- and A-containing targets, whereas little or no activity was observed for the fully NK-containing arrays on both targets. All four 166/167 arrays that varied in the number and distribution of NK RVDs showed higher activity on G-containing targets, and the same was observed for 2 of 3 272/273 TALEN pairs. These data suggest that NK has higher specificity for G over A.

Modeling of NN and NK RVD-DNA interactions

To assess whether our conclusions regarding the affinity and specificity of NN and NK RVDs for G and A have any structural basis, molecular modeling simulations of potential NN:G, NK:G, NN:A, and NK:A interactions were performed using the two solved TAL-DNA co-complexes as structural templates. As determined from the PthXo1 crystal structure, an NN RVD can donate a hydrogen bond to the N_7 atom of G or A and accept a hydrogen bond from the backbone nitrogen of residue 13 in the preceding repeat

(Mak, Bradley, Cernadas, *et al.*, 2012b) (Figure 4-5). When threaded onto the available RVD loop structures, the lysine at position 13 in an NK RVD is positioned to donate a hydrogen bond to the guanine O_6 atom of guanine, but is unable to participate in the two hydrogen bonds formed by the asparagine at that position in an NN RVD. The potential desolvation of the N_7 atom, together with the differential loss of sidechain entropy upon binding, may explain the lower affinity of the NK RVD for G relative to the NN:G interaction. In addition, the prediction that NK:G does form a hydrogen bond with guanine and not with adenine may provide the basis for its apparent G-specificity. It should be noted that the NK interactions are theoretical models based on existing TAL crystal structures; large-scale rearrangement of the RVD loop or interacting residues would likely change the set of potential interactions.

Comparison of TALEN backbone architectures

The first TALEN architecture described contains 288 aa of the N-terminus of Tal1c upstream of the repeat arrays, and 231 aa of the C-terminus downstream of the repeats (Christian *et al.*, 2010). This architecture, termed the pTAL-BamHI backbone, has been shown to be active for many TALEN pairs (Cermak *et al.*, 2011). A second TALEN architecture was described with enhanced nuclease activity, referred to here as the NΔ152/C+63-backbone (Miller *et al*., 2011). The first 152 aa were deleted from the N-terminal portion of the TAL effector protein, upstream of the repeat array (Δ152). A second deletion to the C terminus leaves 63 aa immediately following the last repeat in the repeat array (+63). Other C-terminal truncations of TAL effector proteins have been described, and we constructed an additional TALEN backbone with only 18 aa following

the half-repeat of the TAL effector array, designated NΔ152/C+63 (Mussolino *et al.*, 2011).

To evaluate the relationship between TALEN activity, spacer length and the length of the C-terminus, we cloned a single pair of TAL effector DNA-binding domains in the three architectures described above and tested their activities on targets with spacer lengths varying from 0 to 39 bases. We found that the TALEN pair with the pTAL-BamHI backbone yielded activity that was 25% or greater than the positive control on targets separated by 14-33 bp, with an optimal activity range on spacer lengths of 24-27 bp (Figure 4-6). Similarly, the TALEN pair with the NΔ152/C+63-backbone had activity at least 25% of the positive control across spacer lengths ranging from 14-33 bp. In contrast, the NΔ152/C+18 TALEN displayed more focused activity, showing activity exceeding 25% of the positive control for spacer lengths of 13-16 bp. For spacer lengths of 13-15 bp, activities of the NΔ152/C+18 TALENs were 2-3 fold higher than the NΔ152/C+63 versions. We conclude that shorter C-terminal truncations make it possible to achieve optimal TALEN activity over a narrow range of spacer lengths.

Discussion

Our analysis of the activity of TALENs that target G with either NN or NK RVDs revealed three general conclusions: 1) NN-containing TALENs are more active than their NK-containing counterparts, which can be explained by their overall higher binding affinity for G-containing targets; 2) TALENs made with NK RVDs are more specific for G-containing targets, and 3) NK RVDs have the biggest impact on TALEN activity when they occur in the N-terminal half of the repeat array.

Our study presents, to the best of our knowledge, the first affinity data comparing the effect of RVDs specifying G to the overall affinity of the TAL effector protein. The EMSAs revealed that TAL effectors containing four or five NK RVD substitutions had lower binding affinities than their NN counterparts for the same target. Importantly, the data indicate that the difference in activity of the TALENs can be attributed to the affinity of the TAL effectors for their DNA targets. For example, the NK version of TALEN pair 166 retained over half of the activity of the NN version, which can be explained by the comparable binding affinities of the 166 NN/NK TAL effector arrays *in vitro*. Furthermore, TAL effector 166NN also bound its non-target substrate, indicating it has higher affinity for DNA in general, which could account for the increased nuclease activity observed for TALEN pair 166NN (Figure 4-1). The nuclease activities of TALEN pairs 167 and 312 NN/NK also reflect the observed binding affinities of the TAL effector proteins, in that the NK-containing arrays showed both decreased nuclease activity and lower binding affinity.

Our conclusions are consistent with other studies that evaluated the activity of engineered TAL effector proteins with NN and/or NK RVDs. In one study, the activity of a single TALEN pair was evaluated that contained a mixture of both NK and NN RVDs (Miller *et al*., 2011). These TALENs were active and able to introduce mutations in the intended target site at frequencies exceeding 3%. The TALENs making up this pair contained no NK RVDs within the first five repeats of either array, and of the 36 RVDs in both arrays, 5 were NKs (2 in one array and 3 in the other). The observed activity of this TALEN, therefore, is consistent with our finding that weak RVD/nucleotide associations are tolerated if they occur in the C-terminal half of the array. SELEX assays

were also performed with four engineered TAL effector proteins to profile base specificity, and consistent with our data, NK was found to be more specific for G than NN.

A second study compared the activity of a TALEN pair targeting an endogenous gene in zebrafish (Huang *et al.*, 2011). Both NN and NK versions of the TALEN pair were evaluated, and the NK version was 6-fold less active in mutagenesis than its NN counterpart. Each array of this TALEN pair contained a single NK RVD: one was located in the first repeat of the left array and the other in the seventh repeat of the right. The observed difference in activity of the two TALEN pairs is consistent with our data showing that NK RVDs located in the N-terminal half of the array have the greatest impact on activity.

Two additional studies have appeared with data regarding the activity of TAL effectors targeting the nucleotide G. Both presented evidence consistent with our own observations, namely that TAL effectors with NK RVDs had compromised activity in a transcriptional activation assay compared to the corresponding NN-containing proteins (Cong *et al.*, 2012, Streubel *et al.*, 2012). This led to the conclusion that NK is a 'weak' RVD and does not contribute significantly to the affinity of the protein, but does retain specificity for G. The authors also suggested that NN is a 'strong' RVD and thus contributes more to the overall binding of the TALE. Our EMSA results are consistent with this conclusion, as illustrated by the 166NN TAL effector and TALEN data. Both studies also identified NH as an RVD that targets G with good specificity. It would be interesting to create NH-containing variants of the proteins tested here to directly measure their affinity and compare them to NN- and NK-containing counterparts.

It should also be noted that there is a difference in amino acid composition of the repeats used to make TALENs in this study as compared to the studies mentioned previously. TALENs assembled by our Golden-Gate method have a serine adjacent to the NN RVD (*i.e.* SNN), whereas previous studies use an NN preceded by asparagine (NNN). Although this difference is unlikely to contribute to the overall activity of the TALENs, the impact of this difference was not addressed experimentally in this study.

The recently reported crystal structures of TAL effectors bound to DNA provides some insight into the activity we observe for TALENs made with NN versus NK RVDs (Deng, Yan, *et al.*, 2012, Mak, Bradley, Cernadas, *et al.*, 2012b). The N at position 13 in the NN RVD can form a hydrogen bond with N_7 of either guanine or adenine. For NK, the longer side chain of lysine likely cannot form this bond, and this would negatively impact the affinity of NK-containing repeat arrays for their targets. Additional structures of TAL effectors with NK RVDs should inform this hypothesis and also shed light on differences in how NN and NK RVDs achieve their base specificity.

The crystal structure of the PthXo1TAL effector, which contains 23.5 repeats, showed some disorder at the C-terminus with respect to its interaction with DNA (Mak, Bradley, Cernadas, *et al.*, 2012b). Whether this disorder reflects a diminished role for the C-terminus in DNA binding remains to be determined; however, it is consistent with our finding that suboptimal RVDs at the C-terminus are better tolerated than those at the N-terminus. The activity of other engineered TAL effector proteins with so-called 'mismatches' between RVDs and their intended bases are also consistent with our findings. One study tested the impact of mismatches on transcriptional activation by engineered TAL effector transcriptional activators (Zhang *et al.*, 2011b). Mismatches

near the N-terminus of the engineered TAL effectors typically had a greater negative impact on transcriptional activation than those near the C-terminus, and negative effects were more pronounced if two mismatches occurred in a row or there were more than two total mismatches in the array. We recognize that all mismatches are not equivalent due to differences in affinity and specificity of each RVD for its preferred base; further, TALEN cleavage or transcriptional activation by TAL effectors are indirect readouts of protein/DNA interactions. Direct biochemical measurements of TAL effector DNA binding affinity and specificity as reported here will provide further insight into how to optimize the engineering of these proteins for *in vivo* use.

In addition to issues pertaining to targeting the nucleotide G, TALEN specificity may also be compromised by the fact that a broad range of spacer lengths support FokI dimerization and target cleavage. For therapeutic purposes, it would be desirable to have a TALEN pair that only cleaves when both monomers are bound to target sites separated by a fixed spacer length. Other groups have accessed the effective range of spacers lengths over which TALENs can function, and their conclusions have been similar to ours, namely that truncations to the C-terminus narrows the spacer range over which TALENs are optimally active (Christian *et al.*, 2010, Miller *et al*., 2011, Mussolino *et al.*, 2011, Sun *et al.*, 2012). In this study we completed a side-by-side comparison of three different TAL effector architectures using the same TAL effector DNA binding domains. We observed that TALENs with 63 or 183 amino acids after the repeat array cleave spacer lengths ranging from 12-39 bp at 25% of the activity of our highly active positive control. This contrasts with zinc finger nucleases (ZFNs), which are most active when the range of spacer length is limited to a few base pairs (M Bibikova *et al.*, 2001). Our data

indicate that shortening the C-terminus to 18 residues after the repeat array considerably narrows the range of spacer lengths that support cleavage with the optimal range spanning only 3 bp. Further definition of the TAL effector DNA binding domain may additionally focus the range of spacer lengths that permit cleavage.

We evaluated two aspects of TAL effector protein engineering, namely the suitability of NN and NK RVDs for targeting the nucleotide G and the impact on TALEN activity of both the length of the spacer and the N- and C-termini on either side of the DNA binding domain. We used our yeast-based TALEN activity assay to quickly and quantitatively evaluate these design parameters. In the future, we hope to complement such *in vivo* studies with additional *in vitro* data on the binding activity of TAL effector proteins. We are optimistic that such an integrated approach will provide a deeper understanding of how these proteins interact with DNA and thereby establish best practices for TAL effector protein design.

Material and Methods

TALEN construction. TALEN target sites were identified using the TAL effector-Nucleotide Targeter (TALE-NT) program (Christian *et al.*, 2010). All targets chosen had a T at the -1 position. TALENs recognizing the target sites were constructed using our previously reported assembly method that is based on Golden Gate cloning (Cermak *et al.*, 2011). In brief, Golden Gate cloning uses Type IIS restriction endonucleases (e.g. BsaI, Esp3I) to create unique 4 bp overhangs on DNA fragments so that they can be assembled in a precise, sequential order. Our library of plasmids encodes TAL effector repeats with five different RVDs (NI, HD, NN, NG and NK) that can be released by

digestion with BsaI to create unique 4 bp ends. This enables up to 10 RVD-encoding plasmids to be ligated in the correct order in a single reaction. Sub-arrays of 10 TAL effector repeats can then be joined simultaneously in a second, similar reaction, resulting in a fully-assembled array (11-31 repeats) cloned upstream a FokI nuclease domain in a yeast expression vector.

The pTAL-BamHI expression vector contains 288 aa in the N-terminus of the repeat array and 231 aa in the C-terminus. Truncations of the pTAL-BamHI backbone were constructed by first PCR amplifying fragments encoding the desired length of N or C-termini, along with the sequences necessary to perform subsequent Golden Gate cloning reactions. These amplicons were then cloned into a linearized pTAL backbone lacking the TALE termini to yield the NΔ152/C+63 and NΔ152/C+18 truncated TALEN plasmids. These plasmids are compatible with the Golden Gate assembly platform.

Yeast single-strand annealing assay. The yeast-based assay for testing TALEN function has previously been described (Cermak *et al.*, 2011, Townsend *et al.*, 2009) (Appendix A). Briefly, a yeast strain expressing an engineered TALEN(s) is mated with a yeast strain carrying a corresponding target plasmid. Cleavage of the target and subsequent recombination by single-strand annealing reconstitutes a functional *lacZ* gene and provides a quantitative readout of TALEN activity. β-galactosidase measurements for all data were normalized to the TALEN, SurB, which is highly active in both yeast and at its endogenous target (Y. Zhang *et al.*, 2013). Statistical significance was assessed using a paired t-test. Differences at $p < 0.05$ were considered to be statistically significant. For

some datasets, a ZFN positive control derived from the Zif268 zinc finger array was also included.

Mammalian single-strand annealing assay. Activity of TALENs was measured in HEK293 cells using a single-strand annealing assay that reconstitutes a functional luciferase reporter. The template plasmid for the assay (pSSA-1-3) was kindly provided by Dr. David Segal (UC Davis). PCR-based mutagenesis was used to insert TALEN cleavage sites in pSSA-1-3 between the two halves of the luciferase gene.

Twenty-four hours prior to transfection, HEK293 cells were seeded in quadruplicate in a 24-well dish at a density of 200,000 cells/well in DMEM media containing 10% serum. TALEN monomer plasmids at doses of 100 or 200 μg with 25 ng of the SSA reporter target were transfected into the cells using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. As a control for any basal activity generated by the SSA reporter, the reporter was transfected along with pUC19 plasmid DNA instead of the TALEN. In addition, all reactions included 4 ng of the pRL-TK Renilla luciferase plasmid (Promega) to normalize transfection. At twenty-four hours post transfection, the cells were lysed in 150 μl 1X passive lysis buffer (Promega). Lysate luminescence was measured using the Dual-Luciferase® Reporter Assay System (Promega) according to the manufacturer's instructions. The fold activity of a given TALEN was determined by normalizing the firefly luciferase values to Renilla luciferase, and that value was then divided by the value obtained with the SSA reporter alone.

Expression and purification of recombinant TALE protein. The bacterial expression

vector pGEX6P2-TALE was created by ligating a Golden Gate compatible fragment with the NΔ152/C+63 architecture into pGEX6P2 (GE Healthcare). RVD arrays for specific TALE proteins were then cloned in as described above. Expression constructs encoding TAL effector proteins 166, 167 and 312 were then transformed into Rosetta cells and selected on media containing carbenicillin and chloramphenicol. 200 mL cultures were grown to log phase at 37˚C before induction for 3 hours with 1 mM IPTG. The cells were pelleted by centrifugation and lysed in GST lysis buffer (25 mM HEPES pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 130 µM $CaCl_2$, 0.5% Triton X-100, 10% glycerol, 1 mM PMSF, 1 µg/mL Leupeptin, 100 nM Aprotinin, 1 µg/mL Pepstatin A). The lysates were treated with RNase A (20 µg/mL) and DNase I (10U/mL), clarified by centrifugation (21,000 x g, 10 minutes) and then loaded onto a column containing equilibrated Glutathione Sepharose (GE Healthcare). The columns were washed with GST lysis buffer and subsequently by cleavage buffer (50 mM Tris-HCl pH 8.0, 1 mM EDTA, 1 mM DTT, 10% glycerol). Elution of untagged purified TALE protein was performed by overnight incubation at 4˚C with PreScission protease (GE Healthcare). Purified TALE proteins were separated by electrophoresis and stained with Coomassie to determine the purity of the samples (Figure 4-7).

Electrophoretic mobility shift assay (EMSA). Double stranded DNA substrates were prepared by annealing fluorescently tagged complementary oligos. Sequences for substrates used were 5’-TGGACACGACTTGAGCTGTCGTCTTCTGCACTCGTAGTGCTGTGATGA for 166, 5’- TGGACATGACTTGAGCTAGTCAGCACCAGGCATCGTAGTGCTGTGCTGA for 167,5’-

TGGACACGACTTGAGCTGGCGAAAGAGTCCACCACCATCGTAGTGCTGTGCT GA for 312 and 5'-TGGACACGACTTGAGCTCGACGCTCAGGCAACCGTAGTGCTGTGCTGA for the scrambled target. The purified proteins were diluted into binding buffer (10 mM HEPES pH 7.6, 10% glycerol, 100 mM KCl, 10 mM MgCl2, 100 μM EDTA, 500 μ M DTT) at varying concentrations with a fixed concentration of the labeled DNA substrate (20 nM). The reactions were incubated for 30 minutes at room temperature and then separated by electrophoresis on a 7% TBE-acrylamide gel. Detection of the labeled substrate was then performed on a fluorescent scanner (Storm 860, Molecular Dynamics).

Modeling of NN and NK RVD-DNA interactions. To gain insight into potential determinants of affinity and specificity, NN and NK RVDs were modeled onto all RVD loop backbones in the available crystal structures. Corresponding DNA target site positions were mutated sequentially to adenine and guanine, and the energy of the protein-DNA complex was optimized by Monte Carlo sampling of rotameric sidechain conformations together with gradient-based minimization, allowing small shifts to the protein and DNA backbones. Simulations were performed with the Rosetta software package (Leaver-Fay *et al.*, 2011).

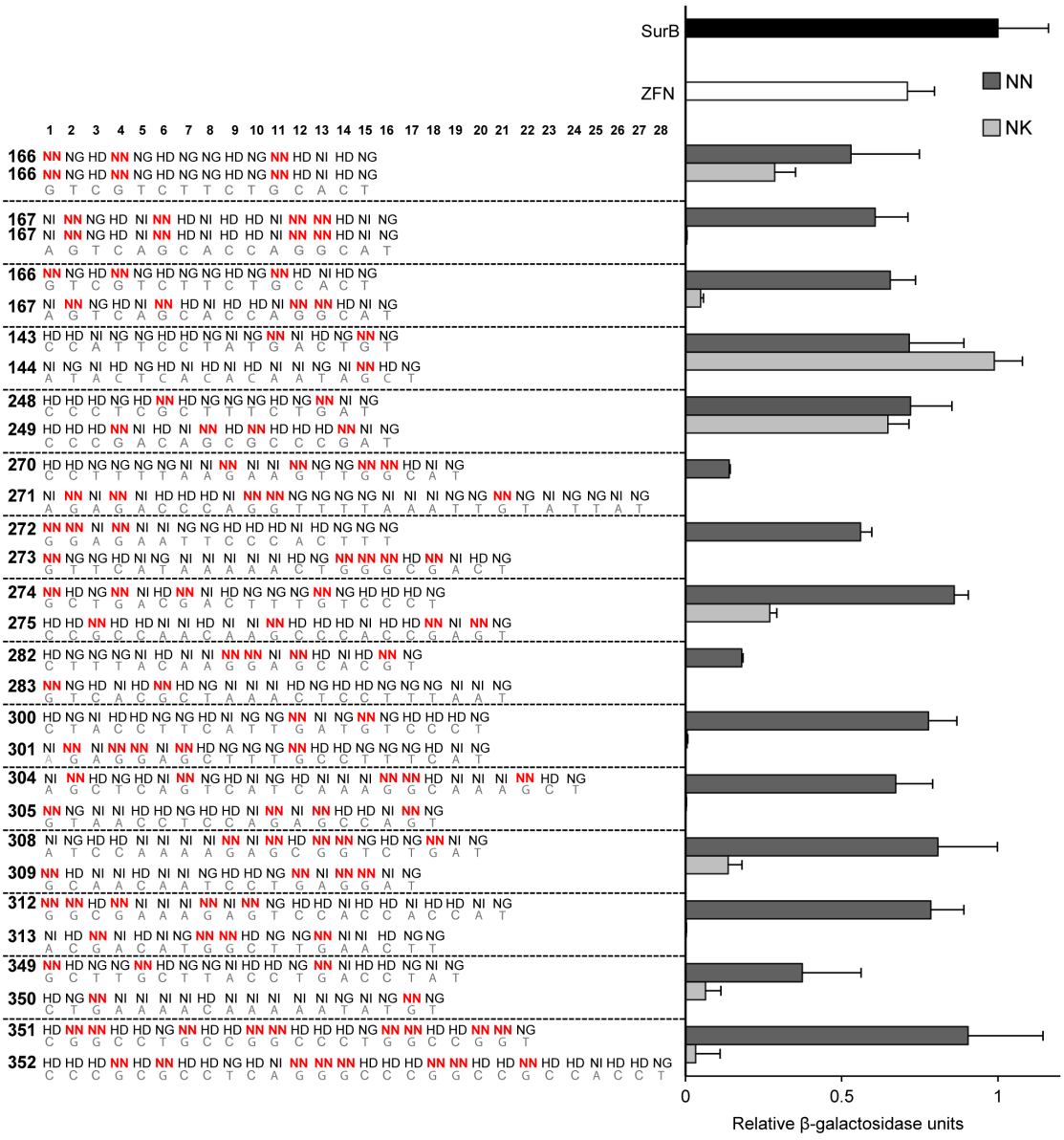


Figure 4-1. Activity of TALENs containing NN or NK RVDs. Plotted on the right are the relative activities of TALENs with either all NN (dark gray) or NK (light gray) RVDs for specifying G nucleotides. Activities are relative β-galactosidase units as determined in our yeast-based TALEN activity assay (see Appendix A). All values are normalized to a TALEN positive control, SurB (black). Also shown for reference is the activity obtained

with the well-characterized ZFN, Zif268 (white). Left and right arrays comprising TALEN pairs are numbered (e.g. 166, 167) followed by RVD sequences of the various TALEN DNA binding domains with repeat 1 being nearest to the N-terminus of the protein. NN RVDs that were replaced with NK are in red. The DNA target sequences are shown below the RVDs. Values are expressed as the mean of triplicates with standard deviation (S.D.).

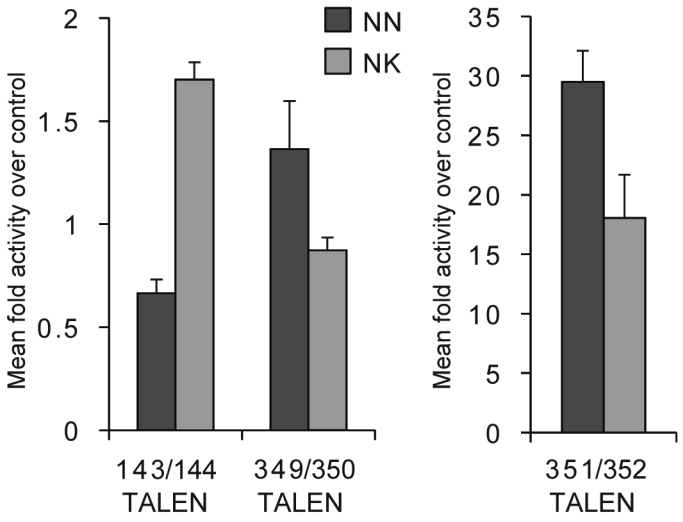


Figure 4-2. Activity of TALENs containing NN or NK RVDs in human cells. Three TALENs with either all NN (dark gray) or NK (light gray) RVDs were tested in a single-strand annealing (SSA) assay in HEK293 cells. Fold activation levels were determined by normalizing the levels of luciferase activity mediated by TALEN conversion of the SSA luciferase target to the basal level of luciferase exhibited by the SSA plasmid alone. Each treatment group contained a transfection/toxicity control consisting of the Renilla luciferase plasmid. Values are expressed as the mean of quadruplicates with S.D.

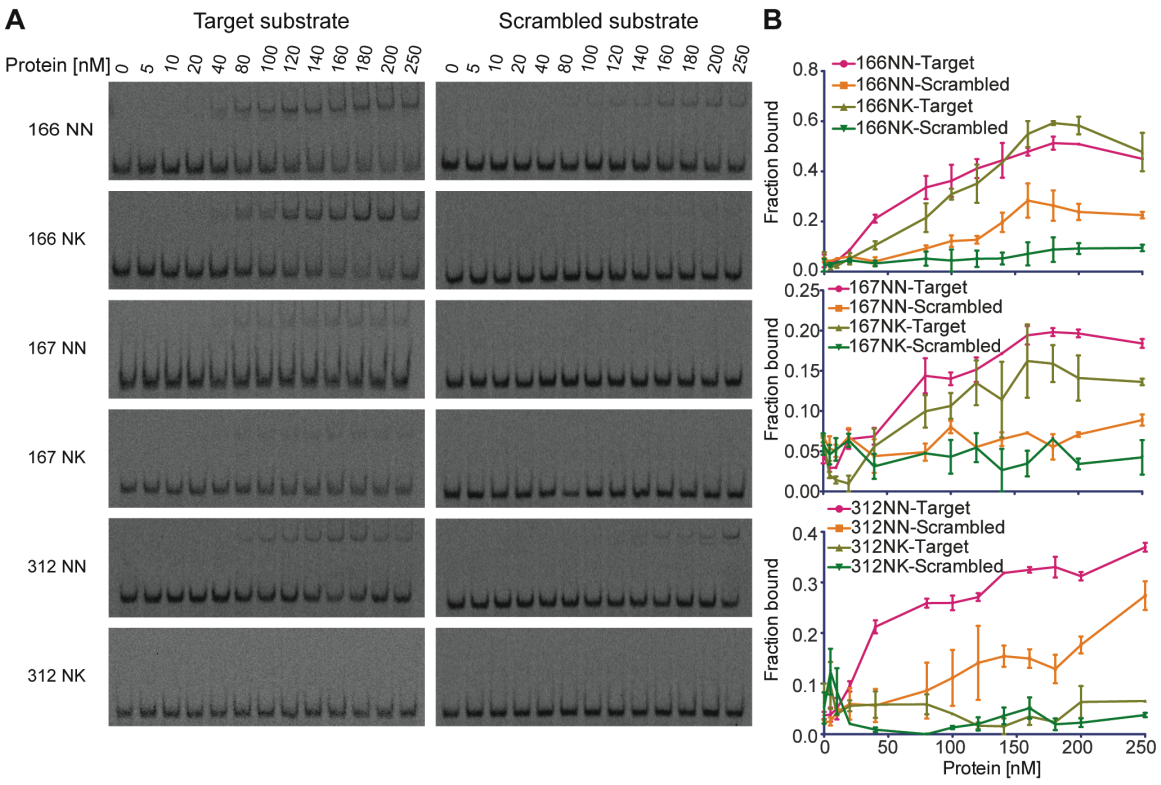


Figure 4-3. Relative binding affinities of purified TAL effector proteins for DNA targets. (**A**) Representative images of electrophoretic mobility shift assays. Fluorescently-labeled double-stranded oligonucleotide substrates corresponding to either TAL effector targets or a scrambled substrate were incubated with purified TAL effector proteins comprised of either NN or NK RVDs. (**B**) Graphs depicting results from densitometry performed on gel images. Shifted labeled substrate was detected using a fluorescent scanner and quantified by densitometry. The fraction of substrate bound by the TAL effector protein was plotted for each protein concentration. Data indicate the mean ± S.D. for 3 independent experiments.

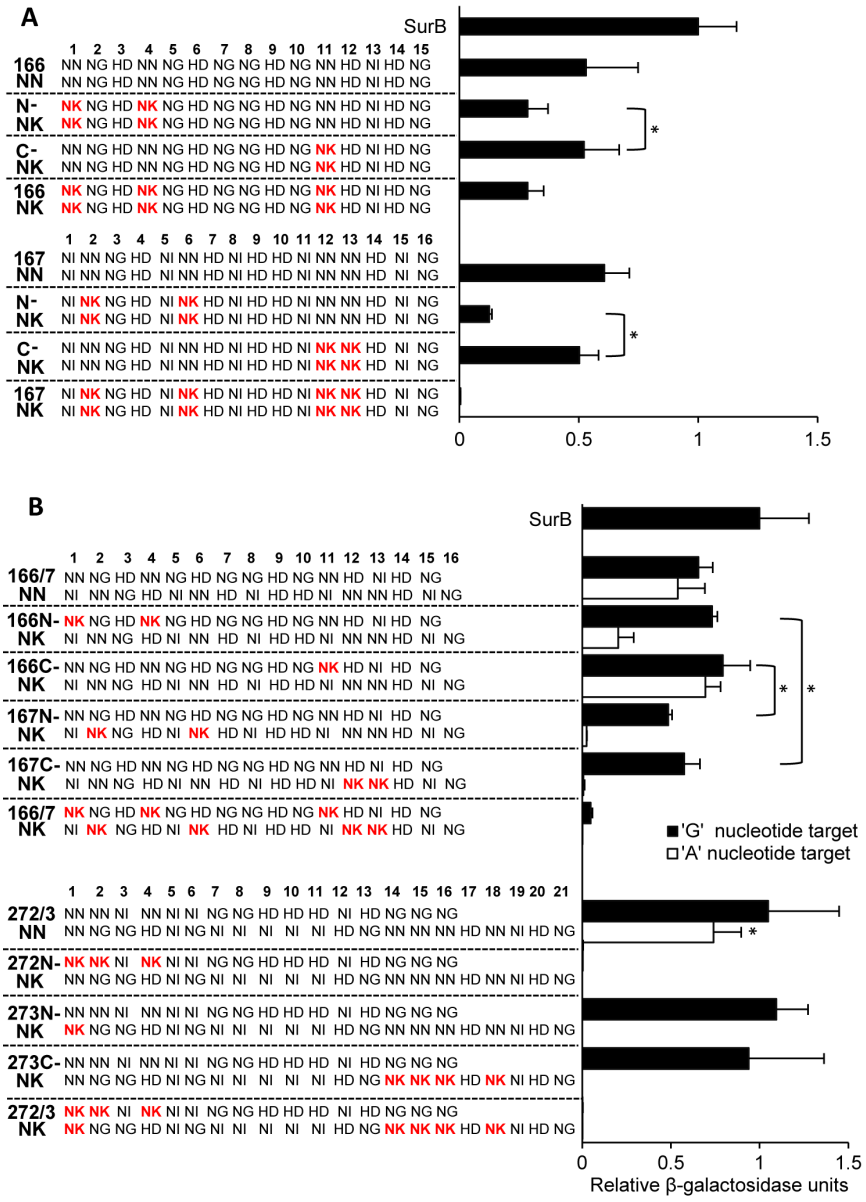


Figure 4-4. Effects on TALEN activity of number and position of NK RVDs in repeat arrays. (**A**) Activity of homodimeric TALENs. TALEN RVD sequences are shown on the left. Note that the left and right TALENs are identical except for the presence of NN versus NK RVDs. NK RVDs are shown in red for ease of visualization. The activity of the TALEN pairs is plotted on the right. (**B**) Activity of heterodimeric TALENs. TALEN arrays with NK substitutions were tested with TALEN array partners that contain only NN RVDs. Black bars denote activity of TALENs at target sequences

with G nucleotides at positions corresponding to NN or NK RVDs. White bars denote TALEN activity with A nucleotides in place of G. All values are normalized to the SurB TALEN positive control. Values are expressed as the mean of quadruplicates with S.D. *, differences in relative β-galactosidase levels between the two indicated samples at $p < 0.05$ were considered to be statistically significant.

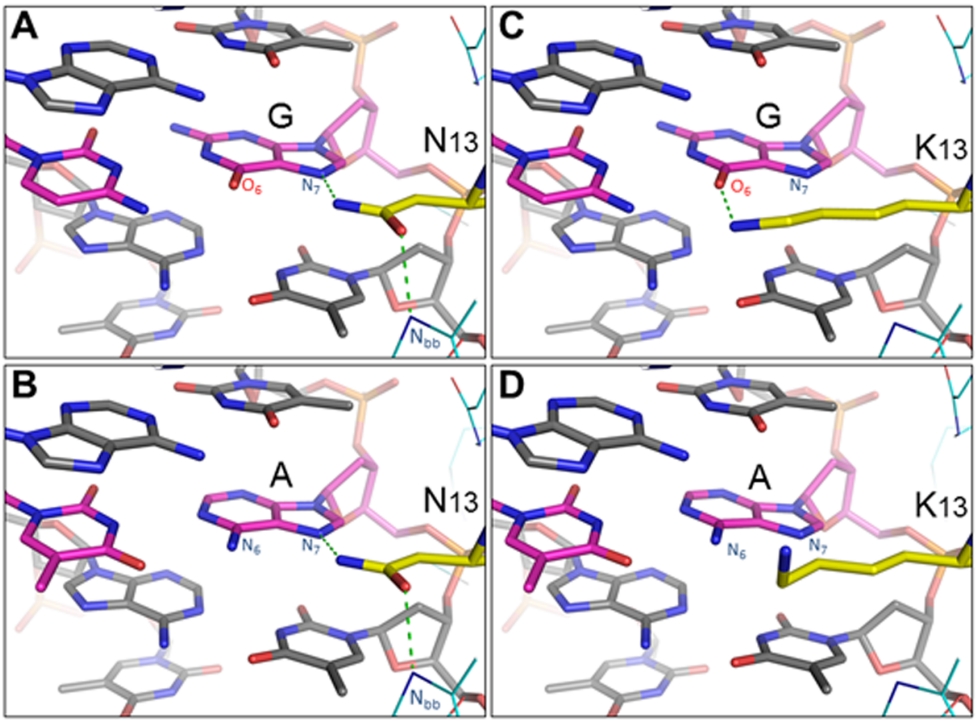


Figure 4-5. Representative low-energy models of NN:G, NK:G, NN:A, and NK:A interactions. (A,B) An NN RVD recognizing either G or A can donate a hydrogen bond to the purine N_7 atom and accept a hydrogen bond from the backbone nitrogen of residue 13 in the preceding repeat. **(C,D)** When threaded onto the available RVD loop structures, the lysine at position 13 (K13) in an NK RVD is positioned to donate a hydrogen bond to the guanine O_6 atom, but is unable to participate in the two hydrogen bonds formed by an asparagine (N13) at that position. The NK RVD is not predicted to form hydrogen bonds with adenine. (Image courtesy of Phil Bradley, University of Washington)

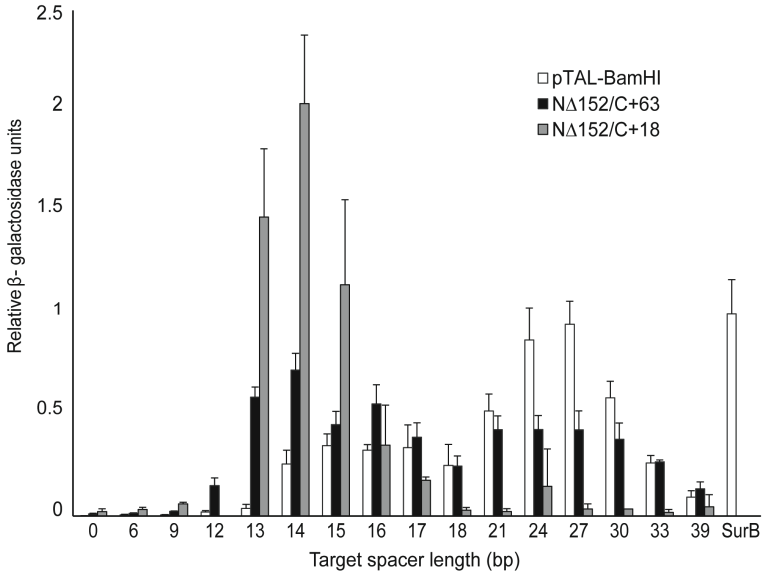


Figure 4-6. TALEN backbone architectures and spacer length optima. A single TALEN repeat array was placed in either the pTAL-BamHI backbone (white), the NΔ152/C+63 backbone (black) or the NΔ152/C+18 backbone (light gray) and tested against its corresponding palindromic target. The left and right TALENs (23.5 repeats) were separated by spacers of lengths indicated by the values shown along the x-axis. All values are normalized to SurB TALEN positive control. Values are expressed as the mean of quadruplicates with S.D.

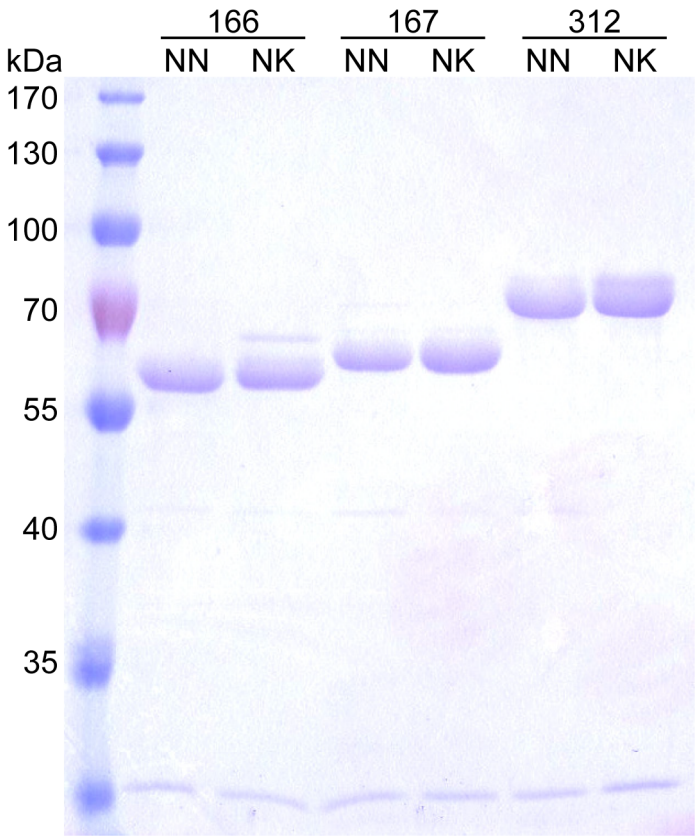


Figure 4-7. Supplementary figure of coomassie stained gel confirming the purity of the purified TALE proteins.

Table 4-1. Supplementary table of TALEN target sequences used in G- vs. A-target yeast SSA assays. Nucleotides positions at which 'T' was replaced with 'A' are shown in red text.

TALEN	Target sequence
166	5'-GTCGTCTTCTGCACT
166A	5'-ATCATCTTCTACACT
167	5'-AGTCAGCACCAGGCAT
167A	5'-AATCAACACCAAACAT
272	5'-GGAGAATTCCCACTTT
272A	5'-AAAAAATTCCCACTTT
273	5'-GTTCATAAAAACTGGG
273A	5'-ATTCATAAAAACTAAA

CHAPTER 5

TARGETED MUTAGENESIS OF *ARABIDOPSIS THALIANA* USING ENGINEERED TAL EFFECTOR NUCLEASES

Introduction

In *Arabidopsis thaliana*, reverse genetic approaches, such as T-DNA and transposon insertional mutagenesis and Targeting Induced Local Lesions in Genomes (TILLING) have enabled the recovery of mutations in a variety of genes (Alonso *et al.*, 2003b, Henikoff *et al.*, 2004, McCallum *et al.*, 2000). These mutagenesis approaches, however, are labor-intensive and provide little control over where the mutations occur. Moreover, many genes are difficult to target (*e.g.* small micro RNA genes and duplicated gene arrays), and there are still thousands of genes for which no mutations have yet been recovered (Alonso *et al.*, 2003b). Other reverse genetics approaches, such as RNAi and artificial micro RNAs, provide specificity for intended targets; however, it is sometimes difficult to achieve a null phenotype by knocking down target gene expression (Waterhouse *et al.*, 1998, Schwab *et al.*, 2006, Felippes *et al.*, 2012, J.-F. Li *et al.*, 2013). The ability to make specific, targeted gene knock-outs would clearly be a valuable addition to the repertoire of genetic tools available for Arabidopsis.

Recent efforts to achieve targeted mutagenesis in plants have focused on the use of sequence-specific nucleases, which include zinc finger nucleases (ZFNs), meganucleases, transcription activator-like effector nucleases (TALENs), and the CRISPR/Cas systems (Carroll, 2011, Stoddard, 2011, Voytas, 2012, Cong *et al.*, 2013, Jinek *et al.*, 2013, Joung & Sander, 2013). All of these nucleases work by introducing a targeted DNA double strand break in the plant genome that is repaired by one of two pathways. In non-homologous end-joining (NHEJ), the break is simply rejoined, sometimes imprecisely, and this can result in insertions or deletions at the break site that

disrupt gene function. Repair by homologous recombination relies on sequence homology provided by a donor DNA molecule, and information from the donor is copied to the broken chromosome (Waterworth *et al.*, 2011). Homologous recombination allows a wide variety of DNA sequence alterations to be made at or near the break site.

Previous work using ZFNs enabled the recovery of plants with mutations in endogenous Arabidopsis genes (Townsend *et al.*, 2009, Zhang *et al.*, 2010, Osakabe *et al.*, 2010, Qi *et al.*, 2013). One challenge in using ZFNs, however, is that they are difficult to engineer to recognize new target sites (J K Joung *et al.*, 2000, Maeder *et al.*, 2008, Sander *et al.*, 2009, Lam *et al.*, 2011). The DNA binding domain of TALENs, on the other hand, is easy to engineer to have new DNA sequence specificities (Boch, 2011). TALENs have been used to transiently edit genes in tobacco and Arabidopsis protoplasts (Cermak *et al.*, 2011, Mahfouz *et al.*, 2010, Zhang *et al.*, 2011a). More recently, TALENs delivered by Agrobacterium have successfully created mutations in rice, barley and Brachypodium (T. Li, Liu, Spalding, Weeks, & Yang, 2012, Shan *et al.*, 2013). In all three cases, somatic mutagenesis was observed in plant tissue expressing the TALENs. One group successfully recovered modified rice plants that were resistant to a plant pathogen due to a TALEN-induced mutation (T. Li, Liu, Spalding, Weeks, & Yang, 2012).

In this study, we used TALENs to create targeted, heritable mutations in Arabidopsis genes. Our goal was to implement TALEN-based mutagenesis using the most commonly practiced transformation technique, namely the robust Agrobacterium-mediated floral dip transformation method. We stably integrated TALEN expression constructs and induced expression during germination with an estrogen-inducible

promoter. Using this so-called *in planta* mutagenesis strategy, we were able to recover TALEN-induced mutations in endogenous genes in 1.5-12% of the progeny.

Materials and Methods

TALEN design, validation and expression in plants

TALEN target sites were identified using the TAL effector-Nucleotide Targeter (TALE-NT) website (Doyle *et al.*, 2012). All target sites retained a T at the -1 position. Corresponding TAL effector arrays were constructed using a Golden Gate-based method of cloning as previously described (Cermak *et al.*, 2011). Information for all of the TAL effector arrays and target sites is listed in Table 5-2. TALENs were assembled in vectors with the truncated N152/C63 backbone architecture (pZHY500 and pZHY501) (Y. Zhang *et al.*, 2013). Fully assembled TAL effector arrays and surrounding N- and C-terminal regions were cloned into the gateway compatible entry clone, pZHY013, using XbaI and BamHI (generating pMC89–95, pZHY013-TALN1–4). pZHY013 contains the EL/KK FokI heterodimers and a viral T2A ribosomal skipping motif that allows translation of both the left and right TALENs from a single transcript (Halpin *et al.*, 1999, J. C. Miller *et al.*, 2007). An estrogen-inducible, Gateway-compatible expression vector, pFZ19, was used to generate transgenic Arabidopsis plants (F.Zhang *et al.*, 2010). A Gateway LR cloning reaction was performed by recombining the entry clones with pFZ19 to generate the vectors pMC105 (*ADH1*), 107 (*TT4*), 108 (*MAPKKK1*), pTALEN5+6 (*DSK2Ba*), pTALEN7+8 (*DSK2Bb*), pTALEN13+14 (*NATA2a*), pTALEN15+16 (*NATA2b*), and pTALEN-*GLL22* (*GLL22*a and *GLL22*b; accession ID *At1g54000* and *At1g54010*). In the final T-DNA expression vectors, the XVE promoter

drives expression of both the left and right TALENs (Zuo *et al.*, 2000) (Figure 5-3). Several TALEN pairs were also cloned into T-DNA expression vectors containing a constitutive 35S promoter. Entry clones pMC89 and pZHY013-TALN1–4 were recombined with pMDC32, a 35S T-DNA expression vector, using LR clonase, generating pMC100–102 (TALENs *ADH1*, *TT4*, *MAPKKK1*) and p35STALEN1–4 (TALENs *DSK2Ba/b* and *NATA2a/b*) (Figure 5-3).

Testing TALEN activity in plants

Stably transformed Arabidopsis lines were generated by the Agrobacterium-mediated floral dip method (Clough & Bent, 1998). T-DNA expression vectors were transformed into *Agrobacterium tumefaciens* strain GV3101 (Koncz *et al.*, 1989, Lee & Gelvin, 2008). Floral dip transformation was conducted using the Columbia ecotype. Seeds from dipped plants were collected and plated onto solid Murashige and Skoog (MS) medium containing 25 mg/L hygromycin to select for transformants containing the transgene (XVE and 35S) and 10-20 μM 17β-estradiol to induce TALEN expression (XVE only). The MS plates containing seeds were placed at 4°C for 3-4 d in the dark for stratification. Plates were then moved to a growth chamber and grown under a regime of 16h light/8h dark at 21°C in a growth chamber

T7 endonuclease and PCR digestion assays to detect somatic mutations

To assess somatic mutagenesis, genomic DNA (gDNA) was extracted from 5-10 pooled T1 seedlings for each TALEN tested. The T7 endonuclease assay was then performed using a modification of a previously described protocol (Guschin *et al.*, 2010). For each reaction, 400 ng of purified PCR product was denatured and reannealed in NEBuffer 2 (New England Biolabs) in a thermocycler using the following regime: 95°C for 5 min,

95-85°C at -2°C/s, 85-25°C at -0.1°C/s; hold at 4°C. The resulting products were digested with 10 U of T7 Endonuclease I at 37°C for 30 min in a final volume of 20 µl. Reactions were then column purified and visualized on a 2% sodium borate gel. For the *NATA2b* TALEN, a 25 µl PCR reaction was performed to amplify target sites using 50-200 ng genomic DNA from pooled seedlings; 10-15 µl of the PCR reaction was subsequently digested with the appropriate restriction enzyme that cleaves wild-type DNA in the spacer of the TALEN target site. Reactions were run on 2% agarose gels and viewed under UV light. Digested bands from both samples were quantified by densitometry using ImageJ software (Rasband, W.S. 1997-2005) and analyzed as previously described (Guschin *et al.*, 2010).

Additional analysis was performed on PCR products from the *ADH1* and *TT4* loci. Undigested PCR products were cloned into TOPO (Invitrogen) vectors, and 84 *ADH1* and 47 *TT4* clones were randomly selected for DNA sequencing. For samples from TALENs targeting *MAPKKK1*, *DSK2Ba/b, NATA2a/b* and *GLL22*, pooled genomic DNA was first pre-digested with the appropriate restriction enzyme, followed by PCR amplification and a second digestion reaction. Resistant bands from these enrichment PCR assays were purified and cloned into TOPO vectors. Several clones were randomly chosen for sequence analysis.

Screening for germline transmission of mutations

To evaluate germinal mutations, hygromycin resistant seedlings (T1) were transferred to soil at 10 days and grown to maturity. Rosette leaves were collected from individual T1 parental lines at the 6-10 leaf stage and genomic DNA was isolated. Limited cycle PCR followed by restriction digestion was performed, and digested bands were visualized on

2% agarose gels. The amount of digestion-resistant product was quantified by densitometry using ImageJ software and expressed as a percentage of the total amplicon (Rasband, W.S. 1997-2005). To screen for *ADH1* mutant progeny, seeds were collected from 11 transgenic plants carrying the *ADH1* TALEN XVE constructs for which transgene expression had been induced. Genomic DNA from 133 sibling plants (originating from four individual T1 lines) was subjected to limited-cycle, high fidelity PCR (F.Zhang *et al.*, 2010). PCR products were digested with PflFI. Mutations in PCR reactions for which a resistant band was detected were confirmed by DNA sequencing. The same protocol was followed to recover inherited mutations at a second locus, *NATA2b*. Seeds from eighteen T1 plants carrying the *NATA2b* 35S TALEN construct were collected. Genomic DNA from 189 sibling plants originating from two distinct parental lines were assayed by PCR and digested with NspI. Four mutant plants were recovered and deletions within the TALEN target site were confirmed by sequencing the PCR product.

Results

Design of TALENs targeting endogenous genes in *Arabidopsis thaliana*

We designed seven different TALEN pairs targeting unique sequences in five Arabidopsis genes (Figure 5-1). Target sites were identified using the publicly available TALE-NT program. Each target had a T nucleotide at the -1 position (Doyle *et al.*, 2012). The number of repeats in each TAL effector array ranged from 15 to 18, and the length of the spacers within the target sites were between 14 and 22 bp (see Table 5-2). One

TALEN pair was designed to recognize the same target sequence in two homologous genes organized in tandem.

We initially engineered four TALENs using the so-called BamHI backbone architecture, which has long N- and C-terminal regions derived from the native TAL effector that flank the DNA binding domain (pTAL-BamHI) (Cermak *et al.*, 2011, M. Christian, Cermak, Doyle, Schmidt, Zhang, Bogdanove, *et al.*, 2010b). Although TALENs made with the BamHI backbone showed consistently high activity in a yeast-based activity assay (Figure 5-2), we found that this activity was not always reflective of mutagenesis efficiencies in somatic tissue of whole plants (Figure 5-2B). Among three TALENs made with the BamHI backbone, only the *ADH1* TALEN showed detectable mutagenesis activity *in planta* (Figure 5-2B) that was confirmed by DNA sequencing (Figure 5-2C). A second TALEN backbone architecture with enhanced nuclease activity was reported that lacks the first 152 aa at the N-terminus of the TALE effector and only has 63 aa after the last repeat in the array (designated N152/C63) (J. C. Miller *et al.*, 2011a). We redesigned our TALEN pairs using this truncated architecture and tested their activity *in planta*.

TALENs induce mutations at endogenous loci in somatic plant tissue

We tested the ability of TALENs to induce mutations at native sites in five genes in Arabidopsis. TALEN pairs were introduced into plants (ecotype Columbia) using the Agrobacterium-mediated floral dip transformation method (Clough & Bent, 1998). TALEN expression was controlled by an estrogen-inducible promoter, which we previously used in our lab for ZFN mutagenesis (Kumar *et al.*, 2005; F.Zhang *et al.*, 2010) (Figure 5-3). Briefly, seeds were collected from plants transformed with

Agrobacterium carrying the TALEN expression construct (T0 plants) and plated on Murashige-Skoog (MS) media with hygromycin and β-estradiol to select for the transgene and to induce TALEN expression, respectively. Primary transgenic seedlings (T1) grown on MS media were pooled and genomic DNA was assessed using two previously described assays to detect NHEJ-induced mutations (H. J. H. Kim *et al.*, 2009, Lloyd *et al.*, 2005). One assay required the presence of a restriction enzyme site in the spacer of the TALEN target. Following TALEN cleavage, restriction sites are often destroyed by NHEJ-induced mutations. Mutated sequences were identified by PCR-amplifying the target site and then quantifying the fraction of amplicons resistant to enzyme digestion. If an appropriate restriction site was not available, the T7 endonuclease I assay was employed (H. J. H. Kim *et al.*, 2009). T7 endonuclease I recognizes and cleaves mismatched heteroduplex DNA. Genomic DNA was isolated from pooled seedlings, and regions surrounding the targets for seven TALENs were amplified by PCR. The amplicons were subjected to rounds of denaturation and annealing to form homo- and heteroduplexes. T7 endonuclease I treatment cleaves the heteroduplexes, and the digested fragments were quantified using densitometry. All TALEN pairs were found to be active at their endogenous targets (Figure 5-4). The somatic mutation frequency of pooled samples ranged from 2% (for the *MAPKKK1* TALENs) to 14% (for the *ADH1* TALENs) (Figure 5-4B).

To confirm the presence of NHEJ-induced mutations at the TALEN target sites, PCR amplicons were cloned and sequenced. Mutations within the spacer sequence of the targets confirmed that TALEN-induced breaks resulted in imprecise repair (Figure 5-4C). Efficiencies of the *ADH1* and *TT4* TALENs were 15% (13 of 87 sequenced clones) and

7% (3 of 43 sequenced clones), respectively (Figure 5-4C). This corroborated the frequencies of somatic mutagenesis observed for these TALENs using the T7 endonuclease I assay. The recovered mutations contained small deletions ranging from 1 to 50 bp in length. The TALEN pairs targeting exons one and two in the *DSK2B* gene showed the lowest levels of activity, namely around 3%. Taken together, these results indicate that TALENs can create targeted mutations in a number of endogenous genes in Arabidopsis.

TALEN-induced mutations are transmitted to the next generation

NHEJ-induced mutations that occur in cells of the L2/L3 layers of the shoot apical meristem should be incorporated into the germline (Furner & Pumfrey, 1992). To determine whether TALEN-induced mutations can be transmitted germinally, we first identified parental transgenic lines with high frequencies of somatic mutagenesis. We collected leaf tissue from individual transgenic T1 plants expressing TALENs and performed the aforementioned PCR and restriction enzyme digestion assay that identifies mutations through loss of a restriction enzyme recognition sequence within the TALEN target site.

We analyzed eleven T1 plant lines transformed with the *ADH1* TALENs and identified four with high levels of somatic mutagenesis (Figure 5-5A). Mutagenesis frequencies in lines 1, 7 and 11 were 12%, 42%, and 33%, respectively. Line 6 had a lower frequency of 5%. Seed was collected from these lines, and individual T2 seedlings were screened for mutations by PCR amplification and restriction enzyme digestion. One germinal event was recovered from line 1, three events from line 7, and one from line 11

(Figure 5-5B). The PCR amplicons were then sequenced to verify the presence of mutations in the TALEN spacer sequence. Indeed, all T2 plants contained small deletions (2, 5, and 9 bp) within the spacer (Figure 5-5C). Thirty-four T2 seedlings from line 6 (low somatic mutagenesis) were also screened, but no germinal events were recovered. Line 7 had the highest number of germinal mutants (12%; 3 mutants/24 seedlings screened) and also had the highest frequency of somatic mutagenesis in the parental line (42%). Seed from one of the ADH1 mutants from line 7 was analyzed in the next generation (T3). Consistent with transmission of a Mendelian factor, we recovered heterozygous individuals (9 of 19 plants), homozygous mutants (3 of 19 plants), and homozygous wildtype plants (7 of 19 plants). The overall germinal transmission frequency of TALEN-induced mutations at the *ADH1* locus was ~4% (Table 5-1). Our data indicate that TALEN-induced mutations can be transmitted to the next generation at reasonable frequencies using parental lines that efficiently express the TALENs.

Constitutive TALEN expression enhances somatic mutagenesis and creates germinally transmitted mutations

Previous work in our lab suggested that there are toxic effects associated with constitutive expression of some ZFNs in Arabidopsis; thus an inducible expression system was adopted (F.Zhang *et al.*, 2010). Because TALENs appear to be less toxic than ZFNs in other eukaryotic systems (Chen *et al.*, 2013), we sought to constitutively express TALENs in Arabidopsis. We placed TALENs under the control of a strong cauliflower mosaic virus (CaMV) 35S promoter and assessed cleavage efficiency amongst individual T1 plants. We focused the analysis on the *NATA2* locus, which had not been previously

targeted using engineered nucleases. Eighteen hygromycin-resistant T1 plants expressing the *NATA2b* TALEN pair were analyzed by the PCR and restriction enzyme digestion assay described above. Evidence for NHEJ-induced mutations was detected in 14 of the 18 lines, with somatic mutagenesis frequencies reaching as high as 73% in line 10 (Figure 5-5D). Predicted insertion or deletion (indel) mutations were verified by sequencing PCR amplicons. Progeny seedlings from line 10 and line 15 (30% mutagenesis frequency) were screened for the presence of inherited mutations at the *NATA2b* target site (Figure 5-5E). We recovered 4 out of 93 T2 plants (4%) from line 10 with mutations in the TALEN spacer sequence (Figure 5-5F; Table 5-1; Table 5-4). No germline mutants were recovered from line 15. For one of the mutant plants derived from line 10, segregation of the mutation was followed in the next generation (T3). We recovered heterozygous individuals (18 of 42 plants), homozygous mutants (9 of 42 plants), and homozygous wild-type plants (15 of 42 plants), consistent with transmission of a Mendelian factor Taken together, we demonstrated that constitutive promoters can be used to express some TALENs without any apparent toxic effects, thereby enabling the recovery of germinal mutants.

TALENs can delete tandemly duplicated genes in somatic cells

In order to expand the utility and versatility of TALENs as tools for Arabidopsis genome modification, we assessed the ability of TALENs to target a duplicated gene array. Duplicated gene arrays are relatively common in many plant species, with nearly 17% of all genes in Arabidopsis arranged as tandem duplicates (The Arabidopsis Initiative, 2000). We designed a TALEN pair to target the same sequence within two

genes of a duplicated gene cluster whose DNA sequence is highly conserved. TALEN pair GLL22 was designed to target and cleave two identical sequences within the tandem gene array *GLL22*, such that simultaneous cleavage of both cut sites might result in a deletion of 4.4 kb of intervening sequence (Figure 5-6A). TALEN expression was induced in T1 plants and cleavage efficiency at each target site was measured using enrichment PCR of pooled DNA samples.

The *GLL22* TALENs cleaved the predicted target sequence in each gene as shown by the fraction of uncut product in the TALEN-induced sample (Figure 5-6B,D). Subsequent cloning and sequencing of the resistant fraction revealed expected indel mutations within the spacer of the TALEN target site in each gene (Figure 5-6C,E). To determine if the TALEN pair could induce a large deletion, primers flanking the upstream TALEN target site GLL22a and the downstream target site GLL22b were used. Presence of an amplicon in the TALEN treated (+) versus untreated (-) sample indicated the expected deletion event of 4.4 kb only in the treated sample. Sequence analysis of the deletion product showed expected deletion junctions between the two TALEN cut sites (Figure 5-6G,F). Both perfect ligations and a range of deletions (3 bp to 50 bp) were recovered from the deletion in somatic cells. These results demonstrate the feasibility of using TALENs to target repetitive gene clusters as well as their potential to delete large spans of DNA between two target sites.

Discussion

Here we demonstrate targeted mutagenesis of endogenous Arabidopsis genes using TALENs. Key to recovery of heritable mutations was identifying plants with high

frequencies of somatic mutagenesis. Leaves from plants that displayed 5% to 73% mutagenesis at the *ADH1* and *NATA2* loci, for example, gave rise to mutants in 12% and 4% of progeny, respectively.

Both optimal nuclease activity and high levels of TALEN expression were important in achieving efficient somatic mutagenesis. We found that the truncated N152/C63 TALEN architecture had higher activity in somatic cells than our original TALEN architecture, which has longer N- and C-terminal extensions flanking the DNA binding domain (the so-called BamHI architecture). When TAL effector DNA binding domains targeting the *ADH1* and *TT4* loci were tested in both architectures, BamHI TALENs showed very little *in vivo* activity (Figure 5-2B). The high activity of N152/C63 TALENs is consistent with our observations in tobacco, where we compared a variety of TALEN N- and C-terminal truncations to identify optimal architectures for activity in plants (Y. Zhang *et al.*, 2013).

Different lines of Arabidopsis with integrated TALEN constructs varied considerably in somatic mutagenesis frequencies, likely because of variable expression of the TALEN transgenes. This could be due to differences in copy number or influences from the genomic site of T-DNA integration (i.e. position effects). Although all eight TALEN pairs tested in this study created mutations at their targets, the efficiencies ranged from 2% to 73% (see Table 5-3). Transmission of TALEN-induced mutations to the next generation was observed only in progeny of T1 plants with the highest frequencies of somatic mutations, a correlation observed recently in zebrafish (Chen *et al.*, 2013). Therefore screening plants for high levels of somatic mutagenesis is recommended to recover heritable mutations.

Nearly all TALENs examined in this study showed comparable or even higher frequencies of somatic mutagenesis than ZFNs previously reported to generate heritable mutations in Arabidopsis (Lloyd *et al.*, 2005; Osakabe *et al.*, 2010; F.Zhang *et al.*, 2010). Osakabe *et al.* (2010) reported ZFNs with relatively low somatic mutagenesis frequencies (0.26 – 3%), and yet heritable mutations were recovered in two of nine lines analyzed. We were not able to recover germinal mutants with comparably active TALENs. The *ADH1* and *TT4* TALENs induced mutations in somatic cell populations at frequencies nearly identical to two ZFNs targeting the same loci (5 – 15% for TALENs and 7 – 16% for ZFNs) (F.Zhang *et al.*, 2010). However, transmission of TALEN-induced mutations to the next generation was considerably lower at the *ADH1* locus (69% for ZFNs vs. 12% for TALENs). For TT4, ZFNs induced mutations in 33% of progeny whereas no mutations were recovered using the *TT4* TALENs. These results were unexpected considering that both the gene targets, method of nuclease expression, and levels of somatic mutagenesis were nearly identical between the two studies.

Is there an inherent difference between ZFNs and TALENs in their ability to generate heritable mutations in Arabidopsis? TAL effector proteins are derived from plant pathogens, and in some plants, TAL effectors trigger defense mechanisms (Boch & Bonas, 2010). Thus, it is possible that a feature of the TALEN scaffold is recognized by the plant and this influences activity in the meristematic cells giving rise to the germline. Alternatively, the small size of ZFNs may allow for greater protein stability or expression in Arabidopsis germ cells. That said, we have also had difficulty in recovering germinal mutants for some ZFNs (Y. Qi, *G3*, in press). To date, the number of endogenous loci in Arabidopsis that have been successfully mutagenized with sequence-specific nucleases is

low, and it remains to be determined if the observed differences in effectiveness of the two nuclease platforms will prove generalizable.

One difference in implementing TALEN-mediated mutagenesis in Arabidopsis versus other models, including Drosophila, mice, rats and human cells, is that the TALEN constructs are stably integrated into the genome. This so-called *in planta* approach presents a number of challenges in obtaining consistent results. As mentioned above, one challenge results from variable nuclease expression in the transgenic lines due to copy number variation and position effects. Additionally, T-DNA transfer can be incomplete, resulting in partial transgene integration, particularly at the right border (Rossi *et al.*, 1996). We observed such incomplete insertions in several lines in which portions of the right TALEN were missing, and this correlates with the proportion of T1 plants with detectable NHEJ levels versus the total number of hygromycin resistant plants (see Table 5-3).

A third challenge facing *in planta* mutagenesis strategies is that some nucleases are cytotoxic. Cytotoxicity with ZFNs led us to use the estrogen-inducible expression system to better control nuclease expression (F.Zhang *et al.*, 2010; Qi *et al.*, 2013). TALENs have been reported to be less cytotoxic than ZFNs (Chen *et al.*, 2013, Mussolino *et al.*, 2011), and so we tested the constitutive CaMV 35S promoter, hoping to enhance mutagenesis frequencies. Primary transgenic plants constitutively expressing TALENs showed somatic mutagenesis efficiencies comparable to or higher than the estrogen-inducible TALENs (see Table 5-3). Further, a higher proportion of plants with constitutively expressed TALENs had somatic efficiencies above 10% relative to those regulated by estrogen. This was true across multiple targets (6 of 8 vs. 4 of 11 for *ADH1*;

9 of 18 versus 2 of 9 for *NATA2b*). Although we did not observe any obvious toxic effects of the *NATA2*b TALEN pair that had somatic mutagenesis frequencies approximating 73%, for some TALEN pairs we obtained few or no primary transgenic plants, which may be attributed to cytotoxicity. This possible relationship between transformation efficiency and cytotoxicity, however, was not tested directly (see Table 5-3). Clearly, there is still considerable room to optimize expression to maximize recovery of heritable mutations while simultaneously minimizing cytotoxic effects and off-target cleavage. A potential disadvantage of the 35S promoter is that it is weakly expressed in germ cells, and this could explain the lower-than-expected transmission frequency for the for the highly active *NATA2b* TALENs (Mascarenhas, 1990). The use of the promoters such as EASE, which confines nuclease expression to the egg cell, may increase germinal recovery of targeted mutations (Even-Faitelson *et al.*, 2011).

TALENs have been successfully used to modify the genomes of three additional plants, namely barley, rice and Brachypodium (T. Li, Liu, Spalding, Weeks, & Yang, 2012, Shan *et al.*, 2013). In all three species, tissue was transformed and plants regenerated that were resistant to the marker gene adjacent to the TALEN. In rice, approximately two-thirds of randomly selected T1 progeny carried mono- or biallelic mutations, suggesting that regeneration of plants directly from transformed cells may be a desirable alternative to *in planta* approaches, at least for plants amenable to cell culture and regeneration.

Our findings in Arabidopsis demonstrate that 1) TALENs can be used to achieve high levels of mutagenesis in somatic cells, 2) TALEN-mediated mutations can be recovered in the next generation from plants with high frequencies of somatic

mutagenesis, and 3) a single TALEN pair can be used to target a tandem gene array and create large deletions. It is expected that the work presented here will provide an easily adoptable framework for TALEN-mediated targeted mutagenesis in Arabidopsis, and future experiments will seek to optimize expression and better understand the relationship between somatic and germinal mutagenesis.

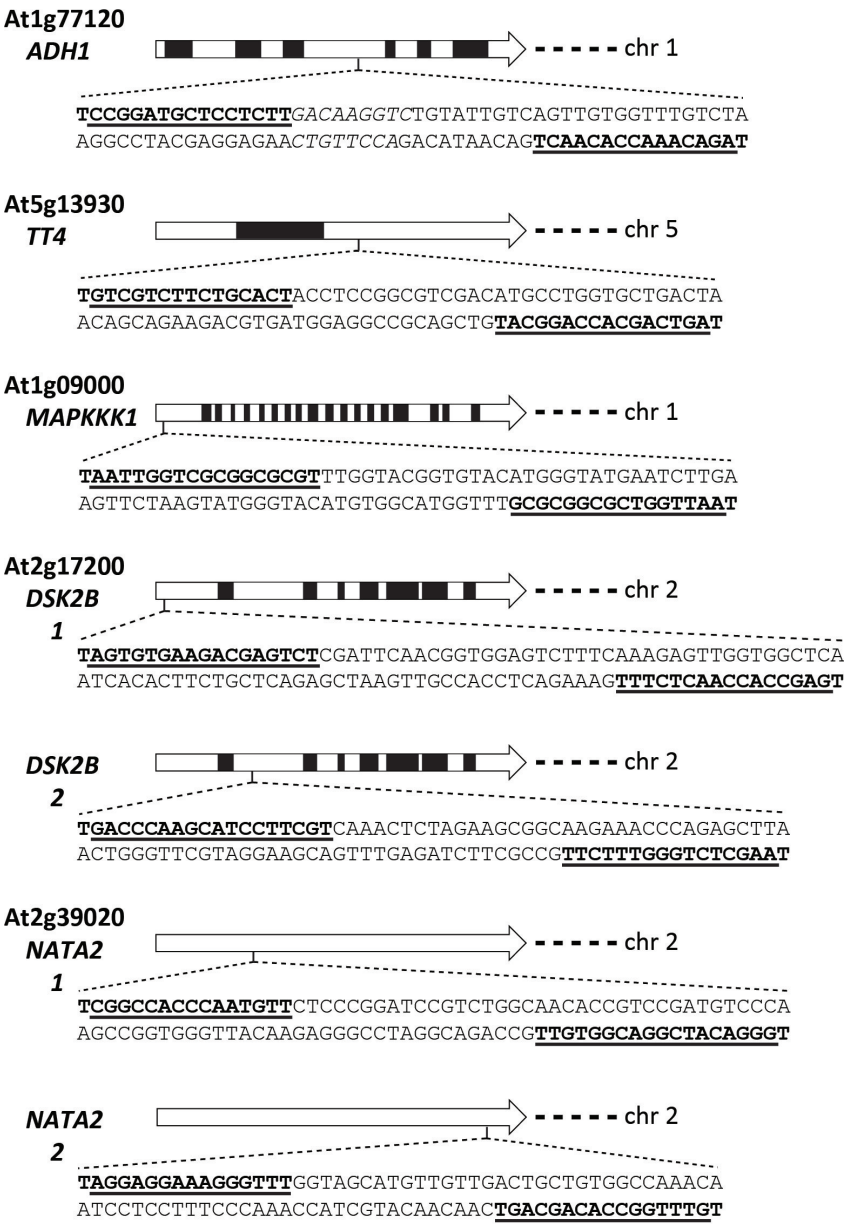


Figure 5-1. TALEN target sites in the coding sequences of Arabidopsis genes. Schematic gene models depicting seven TALEN targets are shown as arrows; white rectangles represent exons. The location of a TALEN target is indicated, and its DNA sequence is provided below the gene model. The sequence recognized by a TALEN pair is underlined and in bold, separated by a spacer sequence. The locus name, gene name and chromosomal location of target genes are also provided.

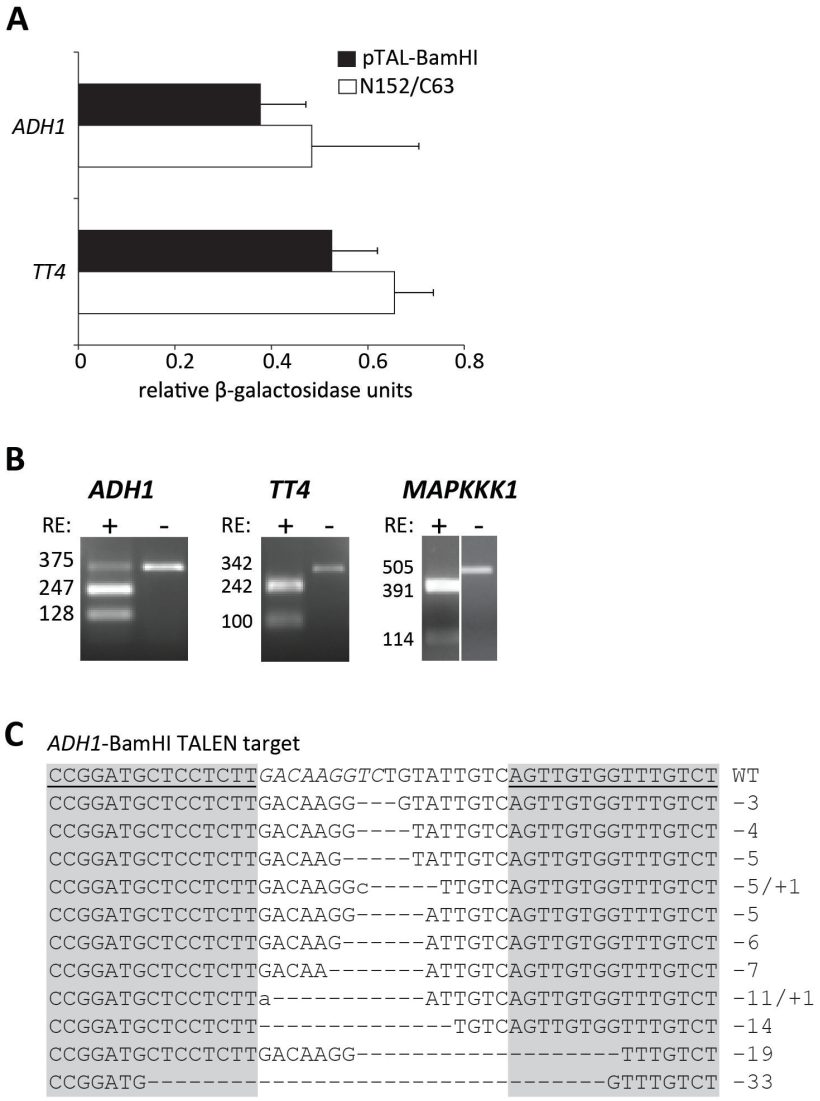


Figure 5-2. Supplementary figure of activity of BamHI-TALENs in yeast and at endogenous targets in Arabidopsis somatic cells. (A) Single-strand annealing (SSA) yeast assay results for representative TALEN pairs. Briefly, *ADH1* and *TT4* TALEN pairs with either the pTAL-BamHI backbone architecture (black bars) or truncated N152/C63 (white bars) were transformed to yeast of one mating type while cognate target plasmids harboring a disrupted *LacZ* gene were transformed to yeast of another mating type. Upon mating, active TALENs cleaved their cognate targets resulting in restoration of *LacZ* that is measured in a b-galactosidase enzyme assay (Christian *et al.*, 2010). Values are normalized to a TALEN positive control and expressed as the mean of duplicates with

standard deviation. (B) Schematic representations of the PCR amplicons assayed by enrichment PCR. TALEN target sites are shown (gray rectangles) and restriction enzyme sites present within spacer sequences of the targets are indicated. (C) Analysis of pTAL-BamHI TALEN activity by enrichment PCR. gDNA from 5-10 pooled seedlings for each BamHI-TALEN was digested with the corresponding restriction enzyme. Mutated target sites preferentially amplified and digested a second time. Presence of a resistant band in the digested (+) TALEN sample versus undigested (-) sample indicates targeted mutations. Fully digested products in the (+) sample indicate absence of TALEN-induced NHEJ mutations. (D) Sequences of ADH1 pTAL-BamHI TALEN TOPO clones. Wild type and mutant genomic *ADH1* targets reveal TALEN-induced mutations. Left and right TALEN target sequences are underlined in the WT sequence; gray rectangles encompass the TALEN target sequence in the clones. Deletions in the sequences alignments are represented by dashes and inserted bases are lowercased. The lengths of insertions (+) or deletions (-) are indicated to the right of the sequences.

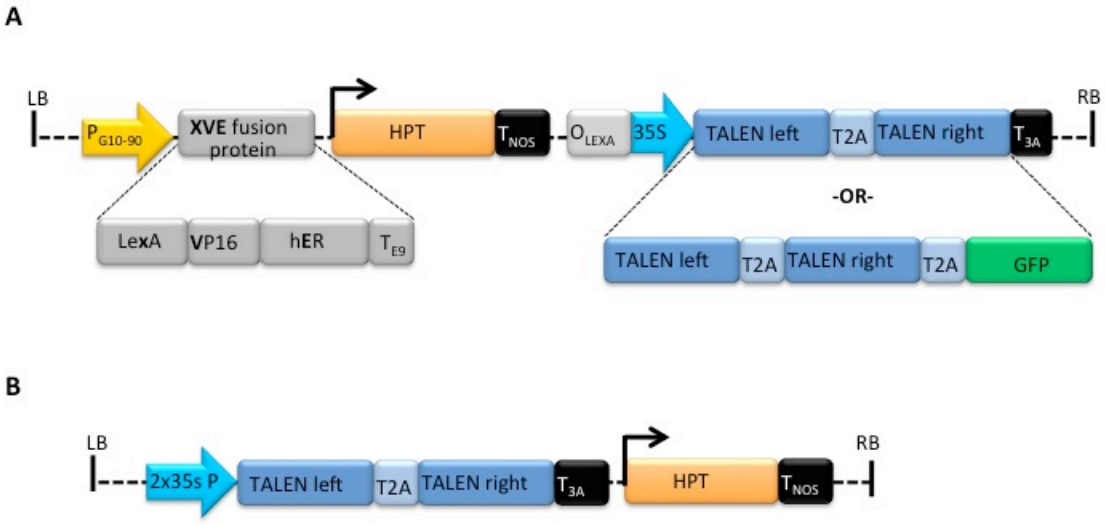


Figure 5-3. Schematic of TALEN T-DNA vectors. (**A**) Estrogen-inducible expression of T2A-linked TALENs. (**B**) Constitutive expression of T2A-linked TALENs

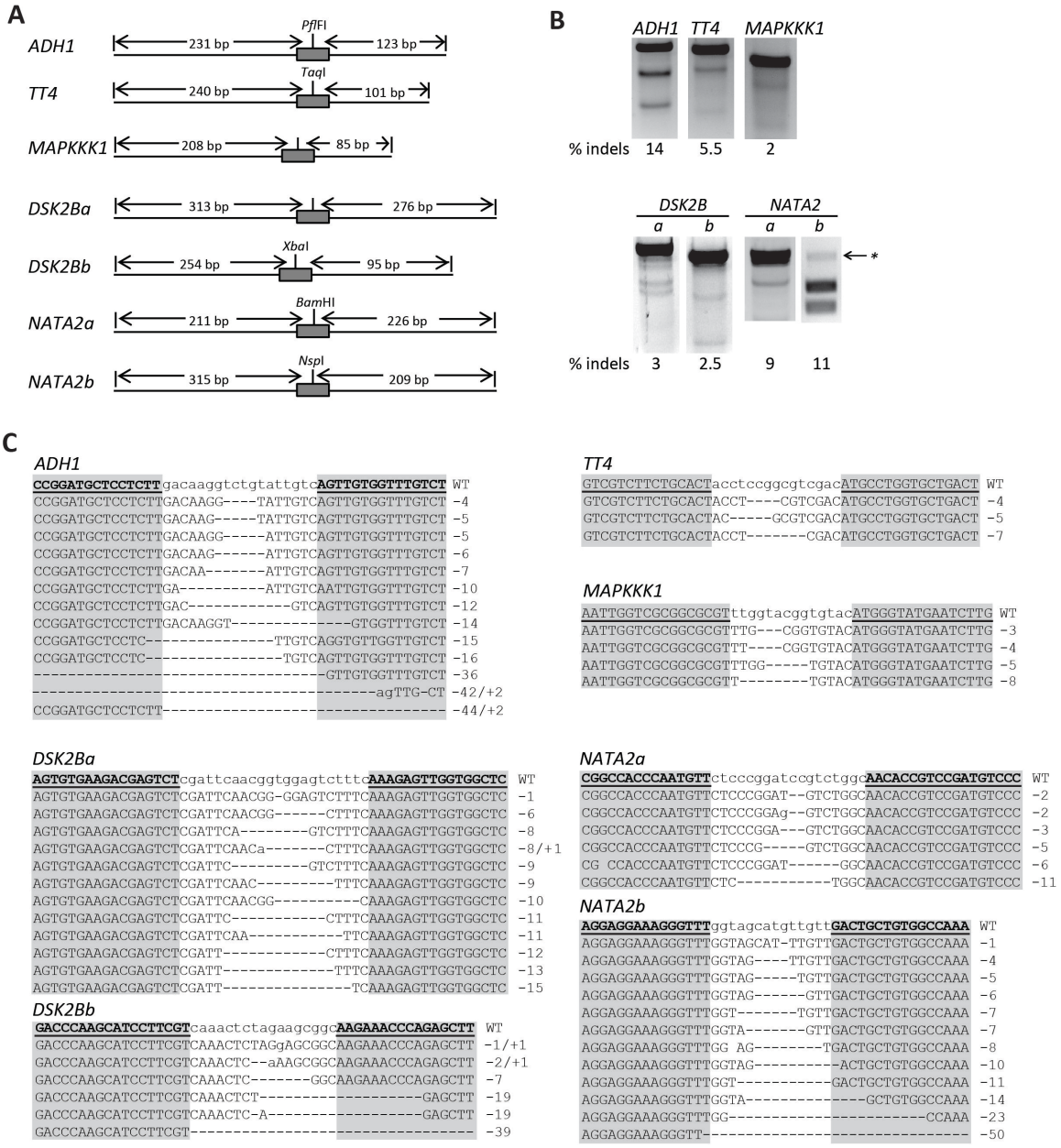


Figure 5-4. Detection of NHEJ mutations in somatic cells of plants with induced TALEN expression. (A) Schematic of the PCR amplicons and features used to detect mutations at TALEN target sites (gray rectangles). Lengths of expected cleavage products after digestion with T7 endonuclease I are indicated between arrows. Restriction enzyme sites present within spacer sequences of the targets are also indicated. Loss of these restriction enzyme sites was used as a measure of TALEN activity (*NATA2* lane b*;*

and in Figure 3). (B) T7 endonuclease I (T7E1) assays to detect TALEN-induced mutations in pooled leaf samples. Genomic DNA was isolated from samples of 5-10 pooled seedlings collected 7-10 days after induction of TALEN expression. Mutations frequencies (% indels) were calculated by measuring band intensities. Treatment with T7E1 yielded unclear results for sample *NATA2b*, so PCR followed by digestion with NspI was used to detect target site mutations. A black arrow with asterisk indicates the position of a DNA band that is resistant to digestion due to TALEN-induced mutations in the enzyme recognition sequence within the TALEN target site. (C) DNA sequencing reveals TALEN-induced mutations. Left and right TALEN target sequences are underlined in bold in the wild type sequence; gray rectangles encompass the TALEN target sequence in the clones. Deletions in the sequences alignments are represented by dashes and inserted bases are in lowercase. The lengths of insertions (+) or deletions (-) are indicated to the right of the sequences.

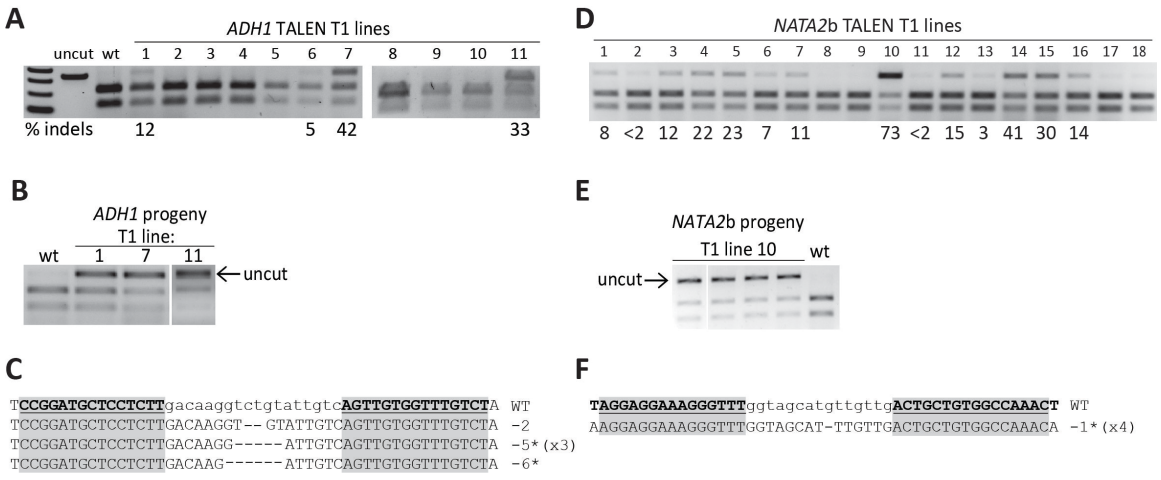


Figure 5-5. Detection of germinally transmitted mutations induced by TALENs. (A, D) Somatic mutation frequencies detected by PCR and digestion assay in individual T1 plant lines expressing the *ADH1* and *NATA2b* TALENs. Genomic DNA from 11 *ADH1* T1 plants and 18 *NATA1-2* *2b* plants was analyzed by amplification of the target locus and subsequent digestion with PflFI (*ADH1*) or NspI (*NATA2b*). Resistant bands in samples 1, 6, 7 and 11 that are equal in size to the uncut product indicate mutations present in the TALEN target sites. The intensity of the resistant band was measured and used to calculate the mutation frequencies (% indels). Samples for which no resistant fraction was detected are unlabeled. (B, E) Germ line mutants recovered from *ADH1* and *NATA2b* TALEN plant lines. Representative T2 seedlings from the next generation of indicated T1 lines were screened by PCR and digestion assay. Uncut bands represent an inherited mutation. (C, F) Sequences of germinal mutations at *ADH1* and *NATA2* loci. Left and right TALEN target sequences are shaded in gray with spacer sequences in lowercase. Deletion lengths are indicated to the right of the aligned sequences. The asterisk indicates a mutation that was identified in the somatic tissue of the T1 plant. Some mutations were identified in more than one mutant progeny seedling as indicated by the number in parentheses.

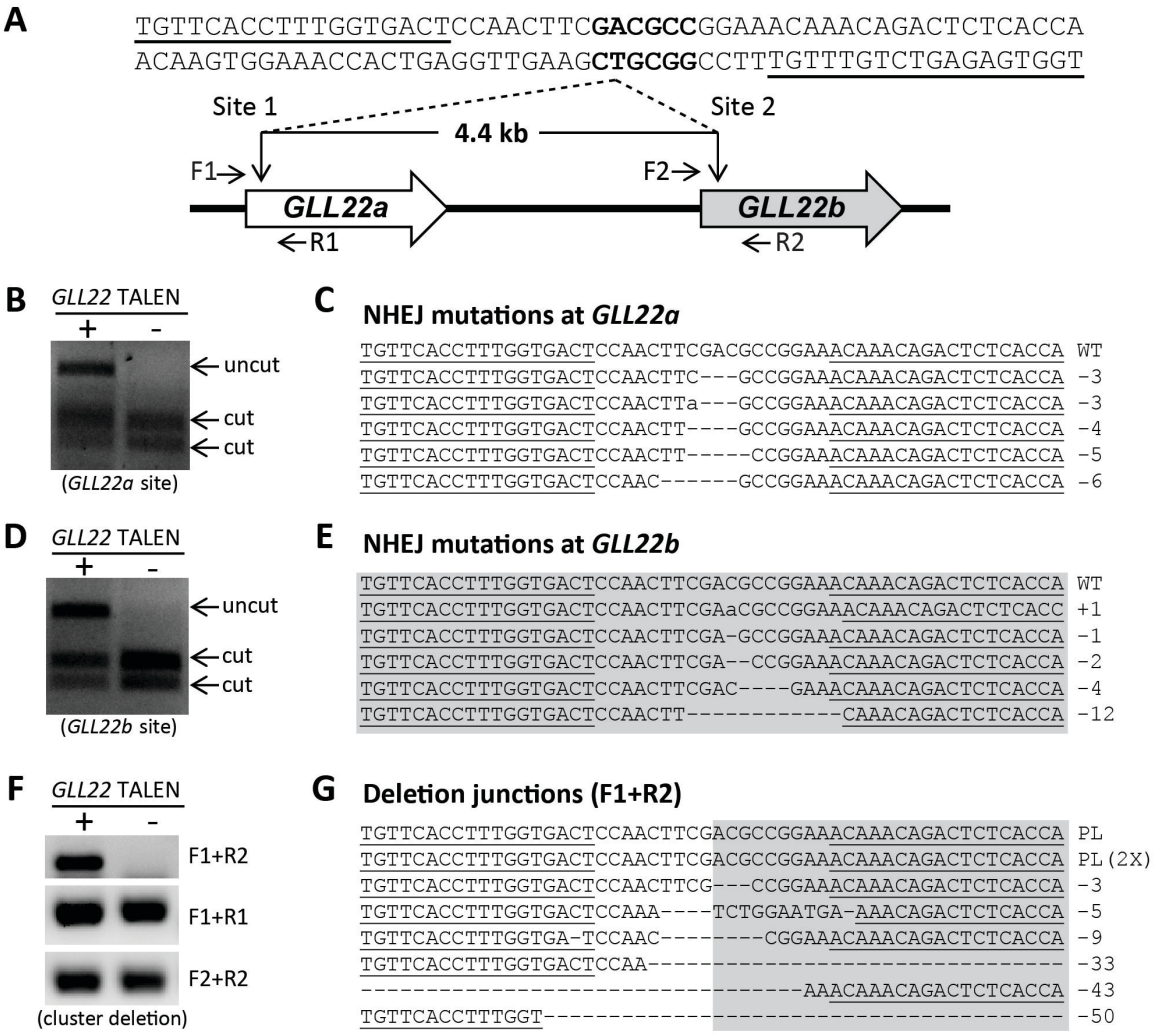


Figure 5-6. Targeted deletion of a *GLL22* gene cluster by TALENs in somatic cells. (A) Schematic representation of the deletion of a gene cluster generated by the *GLL22* TALENs. TALEN binding sites in both gene copies are underlined, and the BsaHI restriction site in the spacer is in bold. (B) NHEJ mutagenesis activity of the *GLL22* TALEN pair at the *GLL22a* site. TALEN activity was assessed by the enrichment PCR assay. Either untreated (-) or TALEN treated (+) genomic DNA was pre-digested with BsaHI, and the TALEN target site was PCR amplified using primers unique to the region in *GLL22a*. Amplicons were digested again with BsaHI and separated on an agarose gel. (C) The uncut band in panel B representing mutated sequences was cloned and randomly selected clones were sequenced to reveal the mutations shown. (D) NHEJ mutagenesis activity of the *GLL22* TALENs at the *GLL22b* site. The activity of *GLL22* TALEN at the

GLL22b site was also assessed by enrichment PCR. (E) The resistant band generated by TALENs at the *GLL22b* site was cloned and randomly selected clones were sequenced as in panel C. (F) Gene cluster deletion was confirmed by PCR with specific flanking primers (F1 and R2, panel A). PCR reactions amplifying both TALEN target sites were used as controls. (G) The deletion-specific PCR product was cloned and sequences compared to the sequence of a perfect ligation (PL). One-week old T1 transgenic (+) and Columbia WT (-) seedlings germinated on MS media with estradiol were used for this experiment.

Table 5-1. Germline transmission frequencies of TALEN-induced mutations in *Arabidopsis thaliana*

TALEN target	T1 line no.	Somatic NHEJ levels (%)	Number of T2 plants screened	Mutant progeny recovered	Transmission frequency[a] (%)
ADH1	1	12	65	1	1.5
	6	5	14	0	0
	7	42	24	3	12.0
	11	33	30	1	3.3
NATA2b	10	73	93	4[b]	4.3
	15	31	96	0	0

[a] The values shown are transmission frequencies as determined by the number of mutant/total number progeny screened

[b] The same mutation was found in the individual mutant plants recovered

Table 5-2. Engineered TALEN information

TALEN ID	Gene name	Number of repeats	Spacer length	RVDs	Target sequence
105-left	*ADH1*	15	18	HD HD NN NN NI NG NN HD NG HD HD NG HD NG NG	CCGGATGCTCC TCTT
105-right		15		NI NN NI HD NI NI NI HD HD NI HD NI NI HD NG	GACAAGGTCTG TATTGTC AGTTGTGGTTT GTCT
107-left	*TT4*	15	15	NN NG HD NN NG HD NG NG HD NG NN HD NI HD NG	GTCGTCTTCTG CACT
107-right		16		NI NN NG HD NI NN HD NI HD HD NI NN NN HD NI NG	ACCTCCGGCGT CGAC ATGCCTGGTGC TGACT
108-left	*MAPKKK1*	17	14	NI NI NG NG NN NN NG HD NN HD NN NN HD NN HD NN NG	AATTGGTCGCG GCGCGT
108-right		16		HD NI NI NN NI NG NG HD NI NG NI HD HD HD NI NG	TTGGTACGGTG TAC ATGGGTATGAA TCTTG
Dsk2-1-left	*DSK2* (exon 1)	17	22	NI NN NG NN NG NN NI NI NN NI HD NN NI NN NG HD NG	AGTGTGAAGAC GAGTCT
Dsk2-1-right		16		NN NI NN HD HD NI HD HD NI NI HD NG HD NG NG NG	CGATTCAACGG TGGAGTCTTTC

					AAAGAGTTGGT GGCTC
Dsk2-2-left	*DSK2* (exon 2)	18	17	NN NI HD HD HD NI NI NN HD NI NG HD HD NG NG HD NN NG	GACCCAAGCAT CCTTCGTCAAA
Dsk2-2-right		16		NI NI NN HD NG HD NG NN NN NN NG NG NG HD NG NG	CTCTAGAAGCG GCAAGAAACC CAGAGCTT
Nata1-1-left	*NATA1*	15	18	HD NN NN HD HD NI HD HD HD NI NI NG NN NG NG	CGGCCACCCAA TGTT
Nata1-1-right	(exon 1)	18		NN NN NN NI HD NI NG HD NN NN NI HD NN NN NG NN NG NG	CTCCCGGATCC GTCTGGCAACA CCGTCCGATGT CCC
Nata1-2-left	*NATA1*	15	18	NI NN NN NI NN NN NI NI NI NN NN NN NG NG NG	AGGAGGAAAG GGTTT
Nata1-2-right	(exon 1)	16		NN NG NG NG NN NN HD HD NI HD NI NN HD NI NN NG	GGTAGCATGTT GTTG ACTGCTGTGGC CAAAC
At1g54-left	*At1g54000* *At1g54010*	18	18	NG NN NG NG HD NI HD HD NG NG NG NN NN NG NN NI HD NG	TGTTCACCTTT GGTGACTCCAA
At1g54-right	*At1g54000* *At1g54010*	18		NG NN NN NG NN NI NN NI NN NG HD NG NN NG NG NG NN NG	CTTCGACGCCG GAAACAAACA GACTCTCACCA

Table 5-3. Summary of TALEN activity in *Arabidopsis thaliana*

Gene target	Expression system	Number of parental T1 plants with somatic indels (total transgenic)	Somatic NHEJ efficiency	Number of progeny plants screened	Number of mutant progeny recovered
ADH1	XVE	6 (13)	5 – 42%	107	4
ADH1	35S	6 (8)	10 – 60%	nt	_
TT4	XVE	1 (3)[a]	6 – 7%	134	0
MAPKKK 1	XVE	1 (4)[b]	5%	108	0
***DSK2*-1**	XVE	5 (10)	3 – 9%	386	0
***DSK2*-2**	XVE	4 (6)	2.5 – 7%	248	0
***NATA1*-1**	XVE	10 (13)	2.5 – 28%	99	0
***NATA1*-1**	35S	2 (6)	nt	60	0
***NATA1*-2**	XVE	6 (9)	4 – 27%	172	0
***NATA1*-2**	35S	14 (18)	2 – 73%	189	4

nt = not tested

[a,b] Consistently low recovery of transgenic plants over three separate transformation experiments

Table 5-4. Mendelian segregation of the *NATA2*b mutation to the next generation.

LINE	No. WT	No. Hets	No. Homo	TOTAL
B	2	4	5	10
C	4	6	2	12
E	5	3	0	8
F	5	5	2	12
TOTALS	16	18	9	42
AVG	**0.38**	**0.43**	**0.21**	

CHAPTER 6

DISSERTATION SUMMARY

Dissertation Summary and Contributions to the Field

One of the primary goals in the field of genome engineering has been to develop tools that allow precise manipulation of the genomes of various organisms. The ability to make targeted DNA sequence modifications has been enabled through the development of proteins that create double-strand breaks at a specific site in the genome. This method of targeted modification or disruption relies upon a 'toolbox' of sequence-specific nucleases, in which proteins that bind DNA in a predictable way are fused to a nuclease protein that then cleaves the DNA at the predicted target site (Baker, 2011). ZFNs have been the primary tool used in the past (Urnov *et al.*, 2010), but they suffer from a number of design and implementation limitations that have been a significant barrier for many researchers trying to adopt the technology.

The work presented in this dissertation provides a framework for genome modification using TAL effector nucleases. Chapter 2 provided the first published demonstration of TALENs, which have proven to be an exciting new addition to the field of genome engineering (Boch, 2011). To date, TALENs have been used to modify the genomes of dozens of organisms with high efficiency and throughput (Table 1-1). We tested fusions of 5 different naturally occurring TAL effector proteins to the FokI nuclease domain. Several variants of each TALE-FokI fusion will be tested in order to determine the minimum TALE sequence required to create a functional nuclease. A yeast-based recombination assay will be used to gauge the relative cleavage activities of the TALENs.

The next aim of this work was to devise a method to rapidly and efficiently engineer custom TALENs to recognize novel DNA target sequences (Chapter 3). To

realize the potential of TALENs for gene modification, it was necessary to have a reliable method to quickly and efficiently construct the TAL effector DNA biding arrays in any desirable order. Our Golden Gate cloning system has been deposited to a plasmid repository, where is has now been distributed to over 1,000 researchers around the world (Addgene). The demand still continues to surprise and impress, and it underlies the desire to use such a tool for a wide variety of purposes, many of which we hope to see come to fruition in the coming years. Another contribution of my research as set out in Chapter 4 was to help elucidate the intricacies of TAL effector-DNA recognition and binding. We were able show how different RVD-DNA binging preferences influenced the affinity of TAL effector proteins for their DNA substrates. This work added to the understanding of the nature of TAL effector-DNA interactions and also helped inform TALEN design and engineering practices.

The final chapter of this dissertation examined the use of TALENs to achieve targeted mutagenesis in Arabidopsis plants. These proof-of-concept experiments applying TALENs in Arabidopsis using standard transformation techniques demonstrate the feasibility of using TALENs to modify plant genomes. The technology does not come without limitations. We were able to make close comparisons of the efficiency of TALENs versus ZFNs at two loci in Arabidopsis. While somatic mutagenesis frequency was high with TALENs, it did not correlate with transmission of the mutation to the next generation, and transmission rates were significantly lower than with ZFNs. It is clear that there is still work to be done, as addressed in the discussion of chapter 5. For example, in the near future we will test different promoters (*e.g.* embryo specific, ubiquitin promoters) in an effort to increase nuclease expression in the cells that give rise

to the seed and therefore the next generation of plants. Achieving effective levels of expression, timing and delivery methods (*e.g.* viral-based delivery using Gemini viruses) of TALENs will be a necessary barrier to overcome if the technology is to become generally applied in plants.

TALENs have become an important and integral tool in the repertoire of genome engineering technologies. Several years ago, the ability to rapidly engineer reliable and efficient reagents to create targeted genetic modifications–and specifically DNA double-strand breaks–was the limiting step for most researchers in the field of genome engineering. No longer does it take a skilled team of molecular biologists over six months to engineer a good ZFN. Now, a single researcher can make dozens of TALENs in under a week; hundreds of TALENs can be produced using open-access, high-throughput methods in just a few days. It is truly astonishing the pace at which TALE technology has progressed.

These advances, however, beg an important question: Now that genome engineers can quickly and reliably make a DSB, what do you do with it? The tools to perform complicated genomic gymnastics are now readily available, so it stands to reason that future work in the field should focus on the next, downstream steps. One possible direction is to manipulate the DNA repair pathways in order to shift the outcome of a DSB in a user-specified manner. Preliminary data and details of these types of experiments are explored in Appendix A of this dissertation. Maximizing the potential of capturing desired mutagenic events and at *high* frequencies will be especially important to genome editing endeavors in plants, especially if recovery of these events in the next generation continues to prove difficult in agriculturally important crops (Chapter 5;

Seichii Toki – personal communication). It is also entirely possible that a genome-editing tool new to the scene – CRISPR/Cas9 – will help overcome existing barriers and provide another exciting tool with which to precisely modify genomes (Dana Carroll, 2012, Cho *et al.*, 2013, Cong *et al.*, 2013, Damian & Porteus, 2013, Ding *et al.*, 2013, Gaj *et al.*, 2013, Gasiunas *et al.*, 2012, Hwang *et al.*, 2013, Jiang *et al.*, 2013, Jinek *et al.*, 2012, Mussolino & Cathomen, 2013, H. Wang *et al.*, 2013, Wei *et al.*, 2013). In all likelihood, these tools – each with its own set of advantages and disadvantages – will be useful depending on the ultimate goal of the researcher.

The work presented in this dissertation has helped lay the groundwork for TALE-based technologies, and has played a pivotal role in enabling researchers to use the TALEN technology. Future endeavors will further facilitate precise genome manipulation in plants. Importantly, these and continued efforts will provide the proof-of-concept experiments that demonstrate the use of TALENs to facilitate generation of non-transgenic, non-genetically modified plants with desired, agriculturally important traits.

BIBLIOGRAPHY

Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., Stevenson, D. K., *et al.* (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science (New York, N.Y.)* **301**(5633), 653–7.

Ansai, S., Sakuma, T., Yamamoto, T., Ariga, H., Uemura, N., Takahashi, R. & Kinoshita, M. (2013) Efficient Targeted Mutagenesis in Medaka Using Custom-Designed TALENs. *Genetics* 1–38.

Antony, G., Zhou, J., Huang, S., Li, T., Liu, B., White, F. & Yang, B. (2010) Rice xa13 recessive resistance to bacterial blight is defeated by induction of the disease susceptibility gene Os-11N3. *The Plant cell* **22**(11), 3864–76.

Aryan, A., Anderson, M. a. E., Myles, K. M. & Adelman, Z. N. (2013) TALEN-Based Gene Disruption in the Dengue Vector Aedes aegypti. (I. A. Hansen, Ed.) *PLoS ONE* **8**(3), e60082.

Baker, M. (2011) Gene-editing nucleases. *Nature Methods* **9**(1), 23–26. *Nature*. Nature Publishing Group. doi:10.1038/nmeth.1807.

Bedell, V. M., Wang, Y., Campbell, J. M., Poshusta, T. L., Starker, C. G., Ii, R. G. K., Tan, W., *et al.* (2012) In vivo genome editing using a high-efficiency TALEN system. *Nature*. Nature Publishing Group. doi:10.1038/nature11537.

Bennardo, N., Gunn, A., Cheng, A., Hasty, P. & Stark, J. M. (2009) Limiting the persistence of a chromosome break diminishes its mutagenic potential. *PLoS genetics* **5**(10), e1000683.

Beurdeley, M., Bietz, F., Li, J., Thomas, S., Stoddard, T., Juillerat, A., Zhang, F., *et al.* (2013) Compact designer TALENs for efficient genome engineering. *Nature Communications* **4**, 1762.

Bhakta, M. S., Henry, I. M., Ousterout, D. G., Das, K. T., Lockwood, S. H., Meckler, J. F., Wallen, M. C., *et al.* (2013) Highly active zinc-finger nucleases by extended modular assembly. *Genome research* **23**(3), 530–8.

Bibikova, M, Carroll, D., Segal, D. J., Trautman, J. K., Smith, J., Kim, Y. & Chandrasegaran, S. (2001) Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Mol Cell Biol* **21**, 289–297.

Bibikova, M, Golic, M., Golic, K. G. & Carroll, D. (2002) Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. *Genetics* **161**(3), 1169–1175.

Bibikova, Marina, Beumer, K., Trautman, J. K. & Carroll, D. (2003) Enhancing gene targeting with designed zinc finger nucleases. *Science (New York, N.Y.)* **300**(5620), 764.

Bitinaite, J., Wah, D. a, Aggarwal, a K. & Schildkraut, I. (1998) FokI dimerization is required for DNA cleavage. *Proceedings of the National Academy of Sciences of the United States of America* **95**(18), 10570–5.

Blount, B. a., Weenink, T., Vasylechko, S. & Ellis, T. (2012) Rational Diversification of a Promoter Providing Fine-Tuned Expression and Orthogonal Regulation for Synthetic Biology. (M. F. Tuite, Ed.)*PLoS ONE* **7**(3), e33279.

Boch, J. (2011) TALEs of genome targeting. *Nature biotechnology* **29**(2), 135–6.
Boch, J. & Bonas, U. (2010) Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annual review of phytopathology* **48**(April), 419–36.
Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., *et al.* (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science (New York, N.Y.)* **326**(5959), 1509–12.
Bochtler, M. (2012) Structural basis of the TAL effector-DNA interaction. *Biological chemistry* **393**(10), 1055–66.
Bogdanove, A. J., Schornack, S. & Lahaye, T. (2010) TAL effectors: finding plant genes for disease and defense. *Current opinion in plant biology* **13**(4), 394–401.
Bogdanove, A. J. & Voytas, D. F. (2011) TAL effectors: customizable proteins for DNA targeting. *Science (New York, N.Y.)* **333**(6051), 1843–6.
Bonas, U., Stall, R. E. & Staskawicz, B. (1989) Genetic and structural characterization of the avirulence gene avrBs3 from Xanthomonas campestris pv. vesicatoria. *Molecular & general genetics : MGG* **218**(1), 127–36.
Bradley, P. (2012) Structural modeling of TAL effector-DNA interactions_DRAFT. *Protein science : a publication of the Protein Society* **000**, 1–4.
Breyer, D., Herman, P., Brandenburger, A., Gheysen, G., Remaut, E., Soumillion, P., Doorsselaere, J. Van, *et al.* (2009) Genetic modification through oligonucleotide-mediated mutagenesis. A GMO regulatory challenge? *Environmental biosafety research* **8**(2), 57–64.
Budman, J. & Chu, G. (2005) Processing of DNA for nonhomologous end-joining by cell-free extract. *The EMBO journal* **24**(4), 849–60.
Bultmann, S., Morbitzer, R., Schmidt, C. S., Thanisch, K., Spada, F., Elsaesser, J., Lahaye, T., *et al.* (2012) Targeted transcriptional activation of silent oct4 pluripotency gene by combining designer TALEs and inhibition of epigenetic modifiers. *Nucleic Acids Research* 1–10.
Cade, L., Reyon, D., Hwang, W. Y., Tsai, S. Q., Patel, S., Khayter, C., Joung, J. K., *et al.* (2012) Highly efficient generation of heritable zebrafish gene mutations using homo- and heterodimeric TALENs. *Nucleic acids research* **40**(16), 8001–10.
Cahill, D., Connor, B. & Carney, J. P. (2006) Mechanisms of eukaryotic DNA double strand break repair. *Frontiers in Bioscience*11:1958-76.
Carlson, D. F., Tan, W., Lillico, S. G., Stverakova, D., Proudfoot, C., Christian, M., Voytas, D. F., *et al.* (2012) Efficient TALEN-mediated gene knockout in livestock. *Proceedings of the National Academy of Sciences of the United States of America* **109**(43), 17382–7.
Carroll, D. (2008) Progress and prospects: zinc-finger nucleases as gene therapy agents. *Gene therapy* **15**(22), 1463–8.
Carroll, D. (2004) Using nucleases to stimulate homologous recombination. *Methods in molecular biology (Clifton, N.J.)* **262**, 195–207. doi:10.1385/1-59259-761-0:195
Carroll, D. (2011) Genome engineering with zinc-finger nucleases. *Genetics* **188**(4), 773–82.
Carroll, D. (2012) A CRISPR approach to gene targeting. *Molecular therapy : the journal of the American Society of Gene Therapy* **20**(9), 1658–60.

Cathomen, T. & Joung, J. K. (2008) Zinc-finger nucleases: the next generation emerges. *Molecular therapy : the journal of the American Society of Gene Therapy* **16**(7), 1200–7.

Cermak, T., Doyle, E. L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J. A., *et al.* (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res* **39**(12), e82.

Certo, M. T., Gwiazda, K. S., Kuhar, R., Sather, B., Curinga, G., Mandt, T., Brault, M., *et al.* (2012) SI_Coupling endonucleases with DNA end-processing enzymes to drive gene disruption. *Nature methods* **9**(10), 973–5.

Chang, C.-H., Wang, H.-I., Lu, H.-C., Chen, C.-E., Chen, H.-H., Yeh, H.-H. & Tang, C. Y. (2012) An efficient RNA interference screening strategy for gene functional analysis. *BMC genomics* **13**(1), 491.

Chen, S., Oikonomou, G., Chiu, C. N., Niles, B. J., Liu, J., Lee, D. a, Antoshechkin, I., *et al.* (2013) A large-scale in vivo analysis reveals that TALENs are significantly more mutagenic than ZFNs generated using context-dependent assembly. *Nucleic acids research* 1–10.

Chevalier, B. S. & Stoddard, B. L. (2001) Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic acids research* **29**(18), 3757–74.

Cho, S. W., Kim, S., Kim, J. M. & Kim, J.-S. (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature Biotechnology* (November 2012), 3–5.

Choulika, A., Perrin, A., Dujon, B. & Nicolas, J. (1995) Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. *Mol. Cell. Biol.* **15**(4), 1968–1973.

Christian, M., Cermak, T., Doyle, E. L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A. J., *et al.* (2010) Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**(2), 757–61.

Christian, M. L., Demorest, Z. L., Starker, C. G., Osborn, M. J., Nyquist, M. D., Zhang, Y., Carlson, D. F., *et al.* (2012) Targeting G with TAL effectors: a comparison of activities of TALENs constructed with NN and NK repeat variable di-residues. *PloS one* **7**(9), e45383

Clough, S. J. & Bent, A. F. (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *The Plant journal : for cell and molecular biology* **16**(6), 735–43.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)* **339**(6121), 819–23.

Cong, L., Zhou, R., Kuo, Y.-C., Cunniff, M. & Zhang, F. (2012) Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature communications* **3**, 968.

Cornu, T. I., Thibodeau-Beganny, S., Guhl, E., Alwin, S., Eichtinger, M., Joung, J. K., Joung, J. K., *et al.* (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), 352–8.

Damian, M. & Porteus, M. H. (2013) A Crisper Look at Genome Editing: RNA-guided Genome Modification. *Molecular therapy : the journal of the American Society of Gene Therapy* **21**(4), 720–2.

Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J.-K., Shi, Y., *et al.* (2012) Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science (New York, N.Y.)* **335**(6069), 720–3.

Deng, D., Yin, P., Yan, C., Pan, X., Gong, X., Qi, S., Xie, T., *et al.* (2012) Recognition of methylated DNA by TAL effectors. *Cell research* **22**(10), 1502–4.

Ding, Q., Lee, Y.-K., Schaefer, E. A. K., Peters, D. T., Veres, A., Kim, K., Kuperwasser, N., *et al.* (2012) A TALEN Genome-Editing System for Generating Human Stem Cell-Based Disease Models. *Cell stem cell*.

Ding, Q., Regan, S. N., Xia, Y., Oostrom, L. A., Cowan, C. A. & Musunuru, K. (2013) Enhanced Efficiency of Human Pluripotent Stem Cell Genome Editing through Replacing TALENs with CRISPRs. *Cell Stem Cell* **12**(4), 393–394.

Doyle, E. L., Booher, N. J., Standage, D. S., Voytas, D. F., Brendel, V. P., Vandyk, J. K. & Bogdanove, A. J. (2012) TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic acids research* **40**(Web Server issue), W117–22.

Doyon, Y., Vo, T. D., Mendel, M. C., Greenberg, S. G., Wang, J., Xia, D. F., Miller, J. C., *et al.* (2011) Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *Nature methods* **8**(1), 74–9.

Dray, T. & Gloor, G. B. (1997) Homology requirements for targeting heterologous sequences during P-induced gap repair in Drosophila melanogaster. *Genetics* **147**(2), 689–99.

Eggleston, A. K. (2007) DNA replication and repair. *Nature* **447**(7147), 923–923.

Ehrhardt, D. W. & Frommer, W. B. (2012) New technologies for 21st century plant science. *The Plant cell* **24**(2), 374–94.

Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. (1996) Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure* **4**(10), 1171–80.

Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. (2009) Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PloS one* **4**(5), e5553.

Engler, C., Kandzia, R. & Marillonnet, S. (2008) A one pot, one step, precision cloning method with high throughput capability. *PloS one* **3**(11), e3647.

Even-Faitelson, L., Samach, A., Melamed-Bessudo, C., Avivi-Ragolsky, N. & Levy, A. a. (2011) Localized egg-cell expression of effector proteins for targeted modification of the Arabidopsis genome. *The Plant journal : for cell and molecular biology* **68**, 929–37.

Fajardo-Sanchez, E., Stricher, F., Pâques, F., Isalan, M. & Serrano, L. (2008) Computer design of obligate heterodimer meganucleases allows efficient cutting of custom DNA sequences. *Nucleic acids research* **36**(7), 2163–73.

Fattah, F., Lee, E. H., Weisensel, N., Wang, Y., Lichter, N. & Hendrickson, E. A. (2010) Ku regulates the non-homologous end joining pathway choice of DNA double-strand break repair in human somatic cells. *PLoS genetics* **6**(2), e1000855.

Felippes, F. F. de, Wang, J. & Weigel, D. (2012) MIGS: miRNA-induced gene silencing. *The Plant journal : for cell and molecular biology* **70**(3), 541–7.
Foley, J. E., Yeh, J.-R. J. R., Maeder, M. L., Reyon, D., Sander, J. D., Peterson, R. T. & Joung, J. K. (2009) Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by Oligomerized Pool ENgineering (OPEN). *PLoS ONE* **4**(2),
Fonfara, I., Curth, U., Pingoud, A. & Wende, W. (2011) Creating highly specific nucleases by fusion of active restriction endonucleases and catalytically inactive homing endonucleases. *Nucleic acids research* 1–14.
Furner, I. J. & Pumfrey, J. E. (1992) Cell fate in the shoot apical meristem of Arabidopsis thaliana. *Development* **115**, 755–764.
Gabriel, R., Lombardo, A., Arens, A., Miller, J. C., Genovese, P., Kaeppel, C., Nowrouzi, A., *et al.* (2011) An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nature biotechnology* **29**(9), 816–23.
Gaj, T., Gersbach, C. a & Barbas, C. F. (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in biotechnology* 1–9.
Galetto, R., Duchateau, P. & Pâques, F. (2009) Targeted approaches for gene therapy and the emergence of engineered meganucleases. *Expert Opinion on Biological Therapy* **9**(10), 1289–1303.
Gao, H., Wu, X., Chai, J. & Han, Z. (2012) Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell research* **22**(12), 1716–20.
Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., *et al.* (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**(6906), 498–511.
Garg, A., Lohmueller, J. J., Silver, P. a & Armel, T. Z. (2012) Engineering synthetic TAL effectors with orthogonal target sites. *Nucleic acids research* **40**(15), 7584–95.
Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **109**(39), E2579–86.
Geissler, R., Scholze, H., Hahn, S., Streubel, J., Bonas, U., Behrens, S.-E. & Boch, J. (2011) Transcriptional activators of human genes with programmable DNA-specificity. *PloS one* **6**(5), e19509.
Guo, J., Gaj, T. & Barbas, C. F. I. (2010) Directed evolution of an enhanced and highly efficient FokI cleavage domain for Zinc Finger Nucleases. *Journal of Molecular Biology*.
Gupta, A., Hall, V. L., Kok, F. O., Shin, M., McNulty, J. C., Lawson, N. D. & Wolfe, S. a. (2013) Targeted Chromosomal Deletions and Inversions in Zebrafish. *Genome research*.
Gürlebeck, D., Szurek, B. & Bonas, U. (2005) Dimerization of the bacterial effector protein AvrBs3 in the plant cell cytoplasm prior to nuclear import. *The Plant journal : for cell and molecular biology* **42**(2), 175–87.
Gürlebeck, D., Thieme, F. & Bonas, U. (2006) Type III effector proteins from the plant pathogen Xanthomonas and their role in the interaction with the host plant. *Journal of plant physiology* **163**(3), 233–55.

Guschin, D. Y., Waite, A. J., Katibah, G. E., Miller, J. C., Holmes, M. C. & Rebar, E. J. (2010) A Rapid and General Assay for Monitoring Endogenous Gene Modification. (J. P. Mackay & D. J. Segal, Eds.)*Methods in molecular biology (Clifton, N.J.)* **649**, 247–256. Totowa, NJ: Humana Press.

Haber, J. E. (1995) In vivo biochemistry: physical monitoring of recombination induced by site-specific endonucleases. *BioEssays : news and reviews in molecular, cellular and developmental biology* **17**(7), 609–20.

Halpin, C., Cooke, S. E., Barakate, A., Amrani, A. El & Ryan, M. D. (1999) Self-processing 2A-polyproteins--a system for co-ordinate expression of multiple proteins in transgenic plants. *The Plant journal : for cell and molecular biology* **17**(4), 453–9.

Henikoff, S., Till, B. J. & Comai, L. (2004) TILLING. Traditional mutagenesis meets functional genomics. *Plant physiology* **135**(2), 630–6.

Herbers, K., Conrads-Strauch, J. & Bonas, U. (1992) Race-specificity of plant resistance to bacterial spot disease determined by repetitive motifs in a bacterial avirulence protein. *Nature* **356**(6365), 172–174.

Hockemeyer, D., Wang, H., Kiani, S., Lai, C. S., Gao, Q., Cassady, J. P., Cost, G. J., *et al.* (2011) Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology* **29**(8), 731–4.

Hoffmann, U., Hecht, R., Boldyreff, B. & Issinger, O. G. (1995) Characterization of the cDNA and three processed pseudogenes from the murine protein kinase CK2 alpha subunit. *Biochimica et biophysica acta* **1260**(3), 337–40.

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, W. Y., Fu, Y., Reyon, D., Maeder, M. L., Tsai, S. Q., Sander, J. D., Peterson, R. T., *et al.* (2013) Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature Biotechnology* (January), 1–3.

Igoucheva, O., Alexeev, V. & Yoon, K. (2004) Oligonucleotide-directed mutagenesis and targeted gene correction: a mechanistic point of view. *Current molecular medicine* **4**(5), 445–63.

Isalan, M., Choo, Y. & Klug, A. (1997) Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proceedings of the National Academy of Sciences of the United States of America* **94**(11), 5617–21.

Jaenisch, R. & Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics* **33 Suppl**, 245–54.

Jamieson, A. C., Miller, J. C. & Pabo, C. O. (2003) Drug discovery with engineered zinc-finger proteins. *Nature reviews. Drug discovery* **2**(5), 361–8.

Jasin, M. (1996) Genetic manipulation of genomes with rare-cutting endonucleases. *Trends in genetics* **12**(6), 224–8.

Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. a. (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature Biotechnology* (November 2012), 1–9.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. & Charpentier, E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)* **337**(6096), 816–21.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E. & Doudna, J. (2013) RNA-programmed genome editing in human cells. *eLife* **2**, e00471.

Johnson, R. D. & Jasin, M. (2001) Double-strand-break-induced homologous recombination in mammalian cells. *Biochemical Society transactions* **29**(Pt 2), 196–201.

Jones, P. A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics* **13**(7), 484–92.

Joung, J K, Ramm, E. I. & Pabo, C. O. (2000) A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **97**(13), 7382–7.

Joung, J Keith & Sander, J. D. (2013) TALENs: a widely applicable technology for targeted genome editing. *Nature reviews. Molecular cell biology* **14**(1), 49–55.

Kay, S., Boch, J. & Bonas, U. (2005) Characterization of AvrBs3-like effectors from a Brassicaceae pathogen reveals virulence and avirulence activities and a protein with a novel repeat architecture. *Molecular plant-microbe interactions : MPMI* **18**(8), 838–48.

Kay, S., Hahn, S., Marois, E., Hause, G. & Bonas, U. (2007) A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science (New York, N.Y.)* **318**(5850), 648–51.

Kay, S., Hahn, S., Marois, E., Wieduwild, R. & Bonas, U. (2009) Detailed analysis of the DNA recognition motifs of the Xanthomonas type III effectors AvrBs3 and AvrBs3Drep16 859–871.

Kim, H. J. H., Lee, H. J., Cho, S. W. & Kim, J.-S. (2009) Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. *Genome research* **19**(7), 1279–88.

Kim, Y., Kweon, J., Kim, A., Chon, J. K., Yoo, J. Y., Kim, H. J., Kim, S., *et al.* (2013) A library of TAL effector nucleases spanning the human genome. *Nature Biotechnology* **31**, 251–8.

Kim, Y.-G., 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), 1156–60.

Kleinstiver, B. P., Wolfs, J. M., Kolaczyk, T., Roberts, A. K., Hu, S. X. & Edgell, D. R. (2012) Monomeric site-specific nucleases for genome editing. *Proceedings of the National Academy of Sciences of the United States of America* **109**(21), 8061–6.

Koncz, C., Martini, N., Mayerhofer, R., Koncz-Kalman, Z., Körber, H., Redei, G. P. & Schell, J. (1989) High-frequency T-DNA-mediated gene tagging in plants. *Proceedings of the National Academy of Sciences of the United States of America* **86**(21), 8467–71.

Kumar, S., Franco, M. & Allen, G. C. (2005) Gene Targeting : Development of Novel Systems for Genome Engineering in Plants. *Plant Biotechnology*.

Lam, K. N., Bakel, H. van, Cote, A. G., Ven, A. van der & Hughes, T. R. (2011) Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. *Nucleic acids research* **39**(11), 4680–90.

Larocque, J. R. & Jasin, M. (2010) Mechanisms of recombination between diverged sequences in wild-type and BLM-deficient mouse and human cells. *Molecular and cellular biology* **30**(8), 1887–97.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., *et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology* **487**, 545–74.

Lee, L.-Y. & Gelvin, S. B. (2008) T-DNA binary vectors and systems. *Plant physiology* **146**(2), 325–32.

Lei, Y., Guo, X., Liu, Y., Cao, Y., Deng, Y., Chen, X., Cheng, C. H. K., *et al.* (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), 17484–9.

Li, J.-F., Chung, H. S., Niu, Y., Bush, J., McCormack, M. & Sheen, J. (2013) Comprehensive Protein-Based Artificial MicroRNA Screens for Effective Gene Silencing in Plants. *The Plant cell*.

Li, Lin & Chandrasegaran, S. (1993) Alteration of the cleavage distance of Fok I restriction endonuclease by insertion mutagenesis **90**(April), 2764–2768.

Li, Lin, Wu, L. & Chandrasegaran, S. (1992) Functional domains in Fok I restriction endonuclease. *Proceedings of the National Academy of Sciences of the United States of America* **89**(10), 4275–9.

Li, Lixin, Piatek, M. J., Atef, A., Piatek, A., Wibowo, A., Fang, X., Sabir, J. S. M., *et al.* (2012) Rapid and highly efficient construction of TALE-based transcriptional regulators and nucleases for genome modification. *Plant molecular biology* **78**(4-5), 407–16.

Li, T., Huang, S., Jiang, W. Z., Wright, D., Spalding, M. H., Weeks, D. P. & Yang, B. (2010) TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic acids research* **39**(1), 359–72.

Li, T., Huang, S., Zhao, X., Wright, D. a, Carpenter, S., Spalding, M. H., Weeks, D. P., *et al.* (2011) Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic acids research* **39**(14), 6315–25.

Li, T., Liu, B., Spalding, M. H., Weeks, D. P. & Yang, B. (2012) High-efficiency TALEN-based gene editing produces disease-resistant rice. *Nature biotechnology* **30**(5), 390–2. doi:10.1038/nbt.2199

Lieber, M. R. (2010) The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. *Annual review of biochemistry* (October 2009), 181–211.

Liu, J., Li, C., Yu, Z., Huang, P., Wu, H., Wei, C., Zhu, N., *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), 209–15.

Lloyd, A., Plaisier, C. L., Carroll, D. & Drews, G. N. (2005) Targeted mutagenesis using zinc-finger nucleases in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **102**(6), 2232–7.

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

Maeder, M. L., Linder, S. J., Reyon, D., Angstman, J. F., Fu, Y., Sander, J. D. & Joung, J. K. (2013) Robust, synergistic regulation of human gene expression using TALE activators. *Nature Methods*. doi:10.1038/nmeth.2366

Maeder, M. L., Thibodeau-Beganny, S., Osiak, A., Wright, D. A., Anthony, R. M., Eichtinger, M., Jiang, T., *et al.* (2008) Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Molecular cell* **31**(2), 294–301. doi:10.1016/j.molcel.2008.06.016

Maeder, M. L., Thibodeau-Beganny, S., Sander, J. D., Voytas, D. F. & Joung, J. K. (2009) Oligomerized pool engineering (OPEN): an "open-source" protocol for making customized zinc-finger arrays. *Nature protocols* **4**(10), 1471–501.

Mahfouz, M. M., Li, L., Piatek, M., Fang, X., Mansour, H., Bangarusamy, D. K. & Zhu, J.-K. (2012) Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein. *Plant molecular biology* **78**(3), 311–21.

Mahfouz, M. M., Li, L., Shamimuzzaman, M., Wibowo, A., Fang, X. & Zhu, J.-K. (2010) De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. *Proceedings of the National Academy of Sciences of the United States of America* **108**(6), 2623–8.

Mak, A. N.-S., Bogdanove, A. J. & Stoddard, B. L. (2012) The cocrystal structure of TAL effector PthMak, A. N.-S., Bogdanove, A. J., & Stoddard, B. L. (n.d.). The cocrystal structure of TAL effector PthXo1 bound to its DNA target, 4031.Xo1 bound to its DNA target. *Science* **4031**.

Mak, A. N.-S., Bradley, P., Bogdanove, A. J. & Stoddard, B. L. (2012) TAL effectors: function, structure, engineering and applications. *Current opinion in structural biology*. doi:10.1016/j.sbi.2012.11.001

Mansfield, J., Genin, S., Magori, S., Citovsky, V., Sriariyanum, M., Ronald, P., Dow, M., *et al.* (2012) Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular plant pathology* **13**(6), 614–29.

Mascarenhas, J. P. (1990) Gene Activity During Pollen Development. *Annual Review of Plant Physiology and Plant Molecular Biology* **41**(1), 317–338.

Mazur, D. J. & Perrino, F. W. (1999) Identification and Expression of the TREX1 and TREX2 cDNA Sequences Encoding Mammalian 3'-5' Exonucleases. *Journal of Biological Chemistry* **274**(28), 19655–19660.

McCallum, C. M., Comai, L., Greene, E. A. & Henikoff, S. (2000) Targeted screening for induced mutations. *Nature biotechnology* **18**(4), 455–7.

Meckler, J. F., Bhakta, M. S., Kim, M.-S., Ovadia, R., Habrian, C. H., Zykovich, A., Yu, A., *et al.* (2013) Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Research* 1–11.

Mercer, A. C., Gaj, T., Fuller, R. P. & Barbas, C. F. (2012) Chimeric TALE recombinases with programmable DNA sequence specificity. *Nucleic acids research* **40**(21), 11163–72.

Miller, J. C., Holmes, M. C., Wang, J., Guschin, D. Y., Lee, Y.-L., Rupniewski, I., Beausejour, C. M., *et al.* (2007) An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature biotechnology* **25**(7), 778–85.

Miller, J. C., Tan, S., Qiao, G., Barlow, K. A., Wang, J., Xia, D. F., Meng, X., *et al.* (2011) A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **29**(December), 143–8.

Miller, J., McLachlan, A. D. & Klug, A. (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *The EMBO journal* **4**(6), 1609–14.

Mineta, Y., Okamoto, T., Takenaka, K., Doi, N., Aoyama, Y. & Sera, T. (2008) Enhanced cleavage of double-stranded DNA by artificial zinc-finger nuclease sandwiched between two zinc-finger proteins. *Biochemistry* **47**(47), 12257–9.

Moore, F. E., Reyon, D., Sander, J. D., Martinez, S. a, Blackburn, J. S., Khayter, C., Ramirez, C. L., *et al.* (2012) Improved somatic mutagenesis in zebrafish using transcription activator-like effector nucleases (TALENs). *PloS one* **7**(5), e37877.

Morbitzer, R., Römer, P., Boch, J. & Lahaye, T. (2010) Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **107**(50), 21617–22.

Moscou, M. J. & Bogdanove, A. J. (2009) A Simple Cipher Governs DNA Recognition by TAL Effectors. *Science* (December), 500-11.

Mudgett, M. B., Chesnokova, O., Dahlbeck, D., Clark, E. T., Rossier, O., Bonas, U. & Staskawicz, B. J. (2000) Molecular signals required for type III secretion and translocation of the Xanthomonas campestris AvrBs2 protein to pepper plants. *Proceedings of the National Academy of Sciences of the United States of America* **97**(24), 13324–9.

Mussolino, C. & Cathomen, T. (2011) On target? Tracing zinc-finger-nuclease specificity. *Nature methods* **8**(9), 725–6.

Mussolino, C. & Cathomen, T. (2013) RNA guides genome engineering. *Nature Biotechnology* **31**(3), 208–209.

Mussolino, C., Morbitzer, R., Lütge, 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* 1–11.

Navis, A., Marjoram, L. & Bagnat, M. (2013) Cftr controls lumen expansion and function of Kupffer's vesicle in zebrafish. *Development*. doi:10.1242/dev.091819.

Orel, N., Kyryk, A. & Puchta, H. (2003) Different pathways of homologous recombination are used for the repair of double-strand breaks within tandemly arranged sequences in the plant genome. *The Plant Journal* **35**(5), 604–612.

Osakabe, K., Osakabe, Y. & Toki, S. (2010) Site-directed mutagenesis in Arabidopsis using custom-designed zinc finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America* **107**(26), 12034–9.

Osborn, M. J., DeFeo, A. P., Blazar, B. R. & Tolar, J. (2011) Synthetic Zinc Finger Nuclease Design and Rapid Assembly. *Human Gene Therapy* 1–43.
Osborn, M. J., Starker, C. G., McElroy, A. N., Webber, B. R., Riddle, M. J., Xia, L., DeFeo, A. P., *et al.* (2013) TALEN-based Gene Correction for Epidermolysis Bullosa. *Molecular Therapy* 1–9.
Ossowski, S., Schwab, R. & Weigel, D. (2008) Gene silencing in plants using artificial microRNAs and other small RNAs. *The Plant journal : for cell and molecular biology* **53**(4), 674–90.
Pabo, C. O., Peisach, E. & Grant, R. A. (2001) Design and selection of novel Cys2His2 zinc finger proteins. *Annual review of biochemistry* **70**, 313–40.
Parekh-Olmedo, H., Ferrara, L., Brachman, E. & Kmiec, E. B. (2005) Gene therapy progress and prospects: targeted gene repair. *Gene therapy* **12**(8), 639–46.
Park, W., Zhai, J. & Lee, J.-Y. (2009) Highly efficient gene silencing using perfect complementary artificial miRNA targeting AP1 or heteromeric artificial miRNA targeting AP1 and CAL genes. *Plant cell reports* **28**(3), 469–80.
Pavletich, N. P. & Pabo, C. O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science (New York, N.Y.)* **252**(5007), 809–17.
Pennisi, E. (2012) The tale of the TALEs. *Science (New York, N.Y.)* **338**(6113), 1408–11.
Perez, E. E., Wang, J., Miller, J. C., Jouvenot, Y., Kim, K. a, Liu, O., Wang, N., *et al.* (2008) Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature biotechnology* **26**(7), 808–16.
Perez-Pinera, P., Ousterout, D. G., Brunger, J. M., Farin, A. M., Glass, K. A., Guilak, F., Crawford, G. E., *et al.* (2013) Synergistic and tunable human gene activation by combinations of synthetic transcription factors. *Nature Methods*. doi:10.1038/nmeth.2361.
Perrino, F. W., Harvey, S., McMillin, S. & Hollis, T. (2005) The human TREX2 3' -> 5'-exonuclease structure suggests a mechanism for efficient nonprocessive DNA catalysis. *The Journal of biological chemistry* **280**(15), 15212–8.
Piganeau, M., Ghezraoui, H., Cian, A. De, Guittat, L., Tomishima, M., Perrouault, L., Rene, O., *et al.* (2013) Cancer translocations in human cells induced by zinc finger and TALE nucleases. *Genome research*.
Plchova, H., Hartung, F. & Puchta, H. (2003) Biochemical characterization of an exonuclease from Arabidopsis thaliana reveals similarities to the DNA exonuclease of the human Werner syndrome protein. *The Journal of biological chemistry* **278**(45), 44128–38.
Podevin, N., Davies, H. V, Hartung, F., Casacuberta, J. M. & Nogue, F. (2012) Site-directed nucleases : a paradigm shift in predictable , knowledge-based plant breeding. *Trends in biotechnology* 1–9. doi:10.1016/j.tibtech.2013.03.004
Politz, M. C., Copeland, M. F. & Pfleger, B. F. (2012) Artificial repressors for controlling gene expression in bacteria. *Chemical communications (Cambridge, England)* 2–4.
Porteus, M. H. & Baltimore, D. (2003) Chimeric nucleases stimulate gene targeting in human cells. *Science (New York, N.Y.)* **300**(5620), 763.
Puchta, H., Dujon, B. & Hohn, B. (1993) Homologous recombination in plant cells is enhanced by in vivo induction of double strand breaks into DNA by a site-specific endonuclease. *Nucleic acids research* **21**(22), 5034–40.

Puchta, H., Dujon, B. & Hohn, B. (1996) Two different but related mechanisms are used in plants for the repair of genomic double-strand breaks by homologous recombination. *Proceedings of the National Academy of Sciences of the United States of America* **93**(10), 5055–60.

Qi, Y., Zhang, Y., Zhang, F., Baller, J. a, Cleland, S. C., Ryu, Y., Starker, C. G., *et al.* (2013) Increasing frequencies of site-specific mutagenesis and gene targeting in Arabidopsis by manipulating DNA repair pathways. *Genome research* **23**, 547–54.

Ramalingam, S., Kandavelou, K., Rajenderan, R. & Chandrasegaran, S. (2011) Creating designed zinc-finger nucleases with minimal cytotoxicity. *Journal of molecular biology* **405**(3), 630–41. Elsevier B.V.

Ramirez, C. L., Foley, J. E., Wright, D. a, Müller-Lerch, F., Rahman, S. H., Cornu, T. I., Winfrey, R. J., *et al.* (2008) Unexpected failure rates for modular assembly of engineered zinc fingers. *Nature methods* **5**(5), 374–5.

Römer, P., Hahn, S., Jordan, T., Strauss, T., Bonas, U. & Lahaye, T. (2007) Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene. *Science (New York, N.Y.)* **318**(5850), 645–8.

Römer, P., Recht, S. & Lahaye, T. (2009) A single plant resistance gene promoter engineered to recognize multiple TAL effectors from disparate pathogens. *Proceedings of the National Academy of Sciences of the United States of America* **106**(48), 20526–31.

Römer, P., Recht, S., Strauss, T., Elsaesser, J., Schornack, S., Boch, J., Wang, S., *et al.* (2010) Promoter elements of rice susceptibility genes are bound and activated by specific TAL effectors from the bacterial blight pathogen, Xanthomonas oryzae pv. oryzae. *The New phytologist* **187**(4), 1048–57.

Römer, P., Strauss, T., Hahn, S., Scholze, H., Morbitzer, R., Grau, J., Bonas, U., *et al.* (2009) Recognition of AvrBs3-like proteins is mediated by specific binding to promoters of matching pepper Bs3 alleles. *Plant physiology* **150**(4), 1697–712.

Rossi, L., Hohn, B. & Tinland, B. (1996) Integration of complete transferred DNA units is dependent on the activity of virulence E2 protein of Agrobacterium tumefaciens. *Proceedings of the National Academy of Sciences of the United States of America* **93**(1), 126–30.

Rouet, P., Smih, F. & Jasin, M. (1994) Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **91**(13), 6064–8.

Rouet, P., Smih, F. & Jasin, M. (1994) Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol. Cell. Biol.* **14**(12), 8096–8106.

Sablok, G., Pérez-Quintero, A. L., Hassan, M., Tatarinova, T. V & López, C. (2011) Artificial microRNAs (amiRNAs) engineering - On how microRNA-based silencing methods have affected current plant silencing research. *Biochemical and biophysical research communications* **406**(3), 315–9.

Sajwan, S., Takasu, Y., Tamura, T., Uchino, K., Sezutsu, H. & Zurovec, M. (2012) Efficient disruption of endogenous Bombyx gene by TAL effector nucleases. *Insect biochemistry and molecular biology* **43**(1), 17–23.

Sakuma, T., Hosoi, S., Woltjen, K., Suzuki, K.-I., Kashiwagi, K., Wada, H., Ochiai, H., *et al.* (2013) Efficient TALEN construction and evaluation methods for human cell and animal applications. *Genes to cells : devoted to molecular & cellular mechanisms*. doi:10.1111/gtc.12037.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor, NY: Cold Spring Harbor Lab Press.

San Filippo, J., Sung, P. & Klein, H. (2008) Mechanism of eukaryotic homologous recombination. *Annu Rev Biochem* **77**, 229–57.

Sander, J. D., Cade, L., Khayter, C., Reyon, D., Peterson, R. T., Joung, J. K. & Yeh, J. J. (2011) Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nature biotechnology* **29**(8), 697–698.

Sander, J. D., Dahlborg, E. J., Goodwin, M. J., Cade, L., Zhang, F., Cifuentes, D., Curtin, S. J., *et al.* (2011) Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nature methods* **8**(1), 67–9.

Sander, J. D., Zaback, P., Joung, J. K., Voytas, D. F. & Dobbs, D. (2007) Zinc Finger Targeter (ZiFiT): an engineered zinc finger/target site design tool. *Nucleic acids research* **35**(Web Server issue), W599–605.

Sander, J. D., Zaback, P., Joung, J. K., Voytas, D. F. & Dobbs, D. (2009) An affinity-based scoring scheme for predicting DNA-binding activities of modularly assembled zinc-finger proteins. *Nucleic acids research* **37**(2), 506–15.

Schmid-Burgk, J. L., Schmidt, T., Kaiser, V., Höning, K. & Hornung, V. (2012) A ligation-independent cloning technique for high-throughput assembly of transcription activator-like effector genes. *Nature biotechnology* **31**(1), 76–81.

Scholze, H. & Boch, J. (2010) TAL effector-DNA specificity. *Science* **1**(5), 428–432.

Schornack, S., Minsavage, G. V, Stall, R. E., Jones, J. B. & Lahaye, T. (2008) Characterization of AvrHah1, a novel AvrBs3-like effector from Xanthomonas gardneri with virulence and avirulence activity. *The New phytologist* **179**(2), 546–56.

Schwab, R., Ossowski, S., Riester, M., Warthmann, N. & Weigel, D. (2006) Highly specific gene silencing by artificial microRNAs in Arabidopsis. *The Plant cell* **18**(5), 1121–33.

Segal, D. J., Beerli, R. R., Blancafort, P., Dreier, B., Effertz, K., Huber, A., Koksch, B., *et al.* (2003) Evaluation of a modular strategy for the construction of novel polydactyl zinc finger DNA-binding proteins. *Biochemistry* **42**(7), 2137–48.

Shan, Q., Wang, Y., Chen, K., Liang, Z., Li, J., Zhang, Y., Zhang, K., *et al.* (2013) Rapid and efficient gene modification in rice and Brachypodium using TALENs. *Molecular plant* 1–11.

Smih, F., Rouet, P., Romanienko, P. J. & Jasin, M. (1995) Double-strand breaks at the target locus stimulate gene targeting in embryonic stem cells. *Nucleic acids research* **23**(24), 5012–9.

Stark, J. M., Pierce, A. J., Oh, J., Pastink, A. & Jasin, M. (2004) Genetic Steps of Mammalian Homologous Repair with Distinct Mutagenic Consequences Genetic Steps of Mammalian Homologous Repair with Distinct Mutagenic Consequences.

Stoddard, B. L. (2011) Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. *Structure* **19**(1), 7–15.

Stoddard, B. L., Scharenberg, A. M. & Monnat, R. J. J. (2007) Advances in Engineering Homing Endonucleases for Gene Targeting : Ten Years After Structures. *Progress in Gene Therapy: Autologous and cancer stem cell gene therapy* 135–167.

Storici, F. (2008) RNA-mediated DNA modifications and RNA-templated DNA repair. *Current opinion in molecular therapeutics* **10**(3), 224–30.

Streubel, J., Blücher, C., Landgraf, A. & Boch, J. (2012) TAL effector RVD specificities and efficiencies. *Nature biotechnology* **30**(7), 593–5.

Sugisaki, H. & Kanazawa, S. (1981) New restriction endonucleases from Flavobacterium okeanokoites (FokI) and Micrococcus luteus (MluI). *Gene* **16**(1-3), 73–8.

Sun, N., Liang, J., Abil, Z. & Zhao, H. (2012) Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. *Molecular bioSystems* **8**(4), 1255–63.

Swoboda, P., Gal, S., Hohn, B. & Puchta, H. (1994) Intrachromosomal homologous recombination in whole plants. *The EMBO journal* **13**(2), 484–9.

Szczepek, M., Brondani, V., Büchel, J., Serrano, L., Segal, D. J. & Cathomen, T. (2007) Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nature biotechnology* **25**(7), 786–93.

Szurek, B., Rossier, O., Hause, G. & Bonas, U. (2002) Type III-dependent translocation of the Xanthomonas AvrBs3 protein into the plant cell. *Molecular microbiology* **46**(1), 13–23.

Tesson, L., Usal, C., Ménoret, S., Leung, E., Niles, B. J., Remy, S., Santiago, Y., *et al.* (2011) Knockout rats generated by embryo microinjection of TALENs. *Nature biotechnology* **29**(8), 695–6.

The Arabidopsis Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**(6814), 796–815.

Thomas, K. R., Folger, K. R. & Capecchi, M. R. (1986) High frequency targeting of genes to specific sites in the mammalian genome. *Cell* **44**(3), 419–28.

Tinland, B., Hohn, B. & Puchta, H. (1994) Agrobacterium tumefaciens transfers single-stranded transferred DNA (T-DNA) into the plant cell nucleus. *Proceedings of the National Academy of Sciences of the United States of America* **91**(17), 8000–4.

Tong, C., Huang, G., Ashton, C., Wu, H., Yan, H. & Ying, Q.-L. (2012) Rapid and cost-effective gene targeting in rat embryonic stem cells by TALENs. *Journal of genetics and genomics = Yi chuan xue bao* **39**(6), 275–80.

Townsend, J. a, Wright, D. a, Winfrey, R. J., Fu, F., Maeder, M. L., Joung, J. K. & Voytas, D. F. (2009) High-frequency modification of plant genes using engineered zinc-finger nucleases. *Nature* **459**(7245), 442–5.

Tremblay, J. P., Chapdelaine, P., Coulombe, Z. & Rousseau, J. (2012) Transcription activator-like effector proteins induce the expression of the frataxin gene. *Human gene therapy* **23**(8), 883–90.

Tzfira, T., Frankman, L. R., Vaidya, M. & Citovsky, V. (2003) Site-specific integration of Agrobacterium tumefaciens T-DNA via double-stranded intermediates. *Plant physiology* **133**(3), 1011–23.

Urnov, F. D., Miller, J. C., Lee, Y.-L., Beausejour, C. M., Rock, J. M., Augustus, S., Jamieson, A. C., *et al.* (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**(7042), 646–51.

Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. (2010) Genome editing with engineered zinc finger nucleases. *Nature reviews. Genetics* **11**(9), 636–46. Nature Publishing Group.

Valton, J., Dupuy, A., Daboussi, F., Thomas, S., Maréchal, A., Macmaster, R., Melliand, K., *et al.* (2012) Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. *The Journal of biological chemistry* **287**(46), 38427–32.

Vanamee, E. S., Santagata, S. & Aggarwal, A. K. (2001) FokI requires two specific DNA sites for cleavage. *Journal of molecular biology* **309**(1), 69–78.

Vanyushin, B. F. & Ashapkin, V. V. (2011) DNA methylation in higher plants: past, present and future. *Biochimica et biophysica acta* **1809**(8), 360–8.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., *et al.* (2001) The sequence of the human genome. *Science (New York, N.Y.)* **291**(5507), 1304–51.

Voytas, D. F. (2012) Plant Genome Engineering with Sequence-Specific Nucleases. *Annual Review of Plant Biology* **64**(1), 327–50.

Wang, H., Yang, H., Shivalila, C. S., Dawlaty, M. M., Cheng, A. W., Zhang, F. & Jaenisch, R. (2013) One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. *Cell* 1–9.

Wang, Z., Li, J., Huang, H., Wang, G., Jiang, M., Yin, S., Sun, C., *et al.* (2012) An integrated chip for the high-throughput synthesis of transcription activator-like effectors. *Angewandte Chemie (International ed. in English)* **51**(34), 8505–8.

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

Waterhouse, P. M., Graham, M. W. & Wang, M.-B. (1998) Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. *Proceedings of the National Academy of Sciences* **95**(23), 13959–13964.

Waterworth, W. M., Drury, G. E., Bray, C. M. & West, C. E. (2011) Repairing breaks in the plant genome: the importance of keeping it together. *The New phytologist* **192**(4), 805–22.

Wei, C., Liu, J., Yu, Z., Zhang, B., Gao, G. & Jiao, R. (2013) TALEN or Cas9 -- rapid, efficient and specific choices for genomic modifications. *Journal of Genetics and Genomics*. Elsevier Ltd. doi:10.1016/j.jgg.2013.03.013

Wolfe, S. A., Nekludova, L. & Pabo, C. O. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Annual review of biophysics and biomolecular structure* **29**, 183–212. doi:10.1146/annurev.biophys.29.1.183

Wood, A. J., Lo, T.-W., Zeitler, B., Pickle, C. S., Ralston, E. J., Lee, A. H., Amora, R., *et al.* (2011) Targeted genome editing across species using ZFNs and TALENs. *Science (New York, N.Y.)* **333**(6040), 307.

Yang, B., Sugio, A. & White, F. F. (2006) Os8N3 is a host disease-susceptibility gene for bacterial blight of rice. *Proc. Natl Acad. Sci. USA* **103**, 10503–10508.

Yang, B. & White, F. (2004) Diverse members of the AvrBs3/PthA family of type III effectors are major virulence determinants in bacterial blight disease of rice. *Mol. Plant-Microbe Interact*, 1192–1200.

Yang, B., Zhu, W., Johnson, L. B. & White, F. F. (2000) The virulence factor AvrXa7 of Xanthomonas oryzae pv. oryzae is a type III secretion pathway-dependent, nuclear-localized, double-stranded DNA binding protein. *Proc. Natl. Acad. Sci. U.SA* **97**, 9807–9812.

Yoo, S.-D., Cho, Y.-H. & Sheen, J. (2007) Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nature protocols* **2**(7), 1565–72.

Zhang, F., Cong, L., Lodato, S., Kosuri, S., Church, G. M. & Arlotta, P. (2011) Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **29**, 149–53.

Zhang, F., Cong, L., Lodato, S., Kosuri, S., Church, G. M. & Arlotta, P. (2011) Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **29**(2), 149–53.

Zhang, F., Maeder, M. L., Unger-Wallace, E., Hoshaw, J. P., Reyon, D., Christian, M., Li, X., *et al.* (2010) High frequency targeted mutagenesis in Arabidopsis thaliana using zinc finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America* **107**(26), 12028–33.

Zhang, Y., Zhang, F., Li, X., Baller, J. A., Qi, Y., Starker, C. G., Bogdanove, A. J., *et al.* (2013) Transcription activator-like effector nucleases enable efficient plant genome engineering. *Plant physiology* **161**(1), 20–7.

Ziller, M. J., Müller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., Boyle, P., *et al.* (2011) Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS genetics* **7**(12), e1002389.

Zuo, J., Niu, Q. W. & Chua, N. H. (2000) Technical advance: An estrogen receptor-based transactivator XVE mediates highly inducible gene expression in transgenic plants. *The Plant journal : for cell and molecular biology* **24**(2), 265–73.

APPENDIX A

EXPLORING THE USE OF MEGANUCLEASES COUPLED WITH THREE-PRIME REPAIR EXONUCLEASE TO ENHANCE TARGETED MUTAGENESIS IN ARABIDOPSIS

Summary of Preliminary Data

The work of this dissertation surrounds the use of engineered nucleases to achieve site-specific genome modification, with particular focus on the application of such tools to modify genes in plants (Chapter 5). This section explores the use of DNA end-processing enzymes to influence the outcome of DSB repair in order to recover more mutagenic events at desired sites in the genome. Preliminary data for this study is presented here in Appendix A.

Engineered nucleases are powerful tools that enable specific modification of user-defined targets. Even with numerous advances in developing the technology, efforts are still ongoing to enhance the efficiency of nuclease – mediated genome editing after the induction of a DSB. The resolution of a DSB by the non-homologous end joining (NHEJ) pathway can result in deleterious events at the break site, which can be harnessed to achieve targeted gene knockouts (Lieber, 2010). This approach has been demonstrated in plants using both ZFNs and TALENs at varying levels of success (Voytas, 2012).

The NHEJ repair pathway is known to proceed via distinct subpathways, namely classic NHEJ (c-NHEJ) or alternative NHEJ (a-NHEJ). Each subpathway can be characterized by mutagenic outcomes directed by the DNA repair proteins that are involved. Classic NHEJ requires KU proteins, DNA-PKcs/Lig4/XRCC4 complex, that often ligate broken ends precisely with minimal processing (Lieber, 2010). Failure to repair a break by cNHEJ leads to increased end processing and typically involves regions of microhomology, making a-NHEJ a more mutagenic process (Lieber, 2010).

A key feature of DSBs induced by current engineered nuclease technologies (*e.g.* ZFNs, TALENs, meganucleases) is the type of DNA end resulting from the break: all

three nucleases create compatible DNA overhangs, which are excellent substrates for repair by cNHEJ and are thus often repair perfectly. A key requirement of targeted mutagenesis following NHEJ repair is to recover altered DNA ends (*i.e.* indels) that impair gene function. Therefore, it is likely that many of the DSBs created by engineered nucleases are repaired accurately and are lost to genome modification purposes.

It has been previously shown that expression of DNA end-processing enzymes makes it possible to influence DNA repair outcomes at a targeted DBS break. Three-prime Repair Exonucleas (Trex) is the major mammalian 3'–5' exonuclease (Perrino *et al.*, 2005). Identification and expression of both Trex1 and Trex2 mammalian cDNAs revealed both were capable of degrading 3'- DNA ends in vitro (Mazur & Perrino, 1999). It has been demonstrated that cotransfection of the mouse Three-prime Repair Exonuclease 2 (Trex2) with the homing endonuclease I–SceI reduced the number of perfectly repaired end-joining events in mammalian cells (Bennardo *et al.*, 2009).

The observation that coexpression of I-SceI and Trex2 limits persistent DSBs has prompted recent studies using Trex2 and other DNA end-processing enzymes to drive gene disruption in mammalian cells (Beurdeley *et al.*, 2013, Certo *et al.*, 2012). One study demonstrated that coupled expression of Trex2 and a panel of meganucleases significantly enhanced mutagenesis frequencies up to 25-fold at targeted DSBs (Certo *et al.*, 2012). A significant enhancement of gene disruption was seen upon expression of Trex2 for every meganuclease tested. Interestingly, expression of Trex2 was able to 'rescue' the activity of inefficient meganucleases and in one case by more than 20-fold enhancement.

Cleavage by meganucleases leaves 2 bp 3'overhangs, which are then presumably chewed back by the 3' exonuclease activity of Trex2 and repaired imperfectly. FokI–derived nucleases create 5' overhangs, so it was not surprising that significant enhancement was not observed when Trex2 was co-expressed with a ZFN or TALEN (Beurdeley *et al.*, 2013, Certo *et al.*, 2012). These results support the notion that Trex2 is most effective at degrading 3' DNA substrates.

The majority of DSBs repaired via NHEJ are repaired precisely; however, the presence of Trex2 results in significantly increased gene disruption. Therefore, we were interested in determining whether similar stimulating effects could be achieved in plants. We have tested whether coexpression of Trex2 would enhance mutagenesis frequencies in plant cells using a well–characterized homing endonuclease, I–SceI (Stark *et al.*, 2004). Next, we hope to determine if these results can be recapitulated in Arabidopsis plants at both an integrated and endogenous, user-defined loci.

Materials and Methods

Protoplast assay using I-SceI and Trex2 expression plasmids[2]

Experiments were performed to test whether co-expression of a meganuclease with Trex2 increases frequencies of targeted mutagenesis in plant cells. Tobacco (*Nicotiana tabacum*) protoplasts were isolated from young leaves of approximately 4-week-old plants grown at 22°C with 16 h of daylight. Protoplast isolation and transformation protocols were adapted from previous work and performed by CPS (Yoo *et al.*, 2007). I-SceI and Trex2 were cloned into separate plasmids downstream of the 35S promoter that

[2] I-SceI, Trex2 and meganuclease constructs were provided by Feng Zhang, Anita Lowy and Cellectis Plant Sciences (CPS). Zhang carried out protoplast assays and 454 sequencing at CPS.

provides high levels of constitutive expression in plant cells. Plasmids encoding I-SceI and Trex2 were introduced into tobacco protoplasts by PEG-mediated transformation (Yoo *et al.*, 2007). The protoplasts were derived from a tobacco line with an integrated I-SceI recognition site (Tzfira *et al.*, 2003).

Four batches of protoplasts, each with 1 X 10^6 cells, were transformed with different combinations of plasmids: 1) 15 mg each of plasmids encoding I-SceI and YFP; 2) 15 mg each of plasmids encoding I-SceI and Trex2; 3) 15 mg each of plasmids encoding Trex2 and YFP; 4) 15 mg of a plasmid encoding YFP. Twenty-four hours after transformation, genomic DNA was isolated from the protoplasts. An approximately 350 bp region encompassing the I-SceI recognition site was amplified by PCR using the primer SSA gFW 5'–TCTCCAATTTGCATACCTCATACC and SSA gRV 5'–TCAGCTTTGTCATTGTGGGATAA. The PCR product was subjected to 454 pyro-sequencing (Roche). Sequencing reads with insertion/deletion (indel) mutations at the cleavage site were considered as having been derived from imprecise repair of a cleaved target site by non-homologous end joining.

Meganuclease and Trex2 T-DNA expression vectors used in this study.

Meganucleases (MN47, MN62) were designed and engineered by Cellectis and Cellectis Plant Sciences (CPS). A summary of the T-DNA engineered MN expression constructs transformed to *Arabidopsis* plants in of the study can be found in Table 1.

T-DNA vectors containing MN47 and MN62 (no Trex2) were stably integrated into the Arabidopsis genome by the floral dip method using *agrobacterium* (Clough & Bent, 1998). The transgenic plants were selected on either: hygromycin plus β-estradiol plates

(XVE) at 22C for 22 hour and 37C for 2 ours, 16 h light/8 dark for 2 weeks; or in soil after spray treatment with BASTA (Orel *et al.*, 2003) to select for the presence of a phosphinotricin-resistance gene (*BAR*). Seeds from both lines were harvested and used in subsequent experiments. Constructs containing MN–T2A–Trex2 (MN47+Trex2, MN62+Trex2) or fusions of MN–Trex2 (MN47+F-Trex2, MN62+F-Trex2) were transformed to *agrobacterium* and introduced into plants by floral dip (Clough & Bent, 1998). Transformants were selected for on hygromycin +β-estradiol plates or in soil using spray treatment with BASTA (Orel *et al.*, 2003).

T-DNA constructs encoding the I-SceI homing endonuclease with Trex2 (35S::MN335 and XVE::MN336) were transformed to *agrobacterium* and introduced into plants by floral dip (Clough & Bent, 1998).

Plant lines and SSA GUS assay.

An *Arabidopsis* line with an integrated pDGU.US reporter was used to generate results presented in figure 2. This vector contains the I-SceI target sequence flanked by two halves of a disrupted β-glucuronidase (GUS) reporter gene with 557 bp of overlapping sequence (Orel *et al.*, 2003, Tinland *et al.*, 1994), and was integrated into Arabidopsis using the floral dip method (Clough & Bent, 1998). Upon induction of a DSB by expression of I-SceI, repair by SSA can occur. The β-glucuronidase gene will then be expressed and can be visualized by histochemical staining as blue colored product after the addition of the substrate X-Gluc. T-DNA constructs encoding the I-SceI homing endonuclease with Trex2 (335/336) were transformed to plants homozygous for the GU.US SSA transgene, and seed was collected and plated on MS media under the following conditions: Arabidopsis plants carrying both the SSA target constructs and

335/336 transgene were identified using the herbicide BASTA (Orel *et al.*, 2003) to select for the presence of a phosphinotricin-resistance gene (*BAR*) (GU.US SSA transgene) as well as Hygromycin (HYG) to select for the presence of the MN – Trex2 transgene.

GUS Assay

The X-Gluc solution was prepared as previously described, but NaN_3 was omitted (Swoboda *et al.*, 1994). Transgenic seedlings (*BAR*+/HYG+) were collected after 7–10 days and subjected to staining with X-Gluc. Seedlings were placed in X-Gluc solution for 24h. After staining, chlorophyll was removed by placing seedlings in an 80% ethanol solution overnight.

Preliminary results and discussion

Coexpression of Trex2 enhances the frequency of mutagenesis in tobacco protoplasts.

Experiments were performed to test whether co-expression of a well-characterized meganuclease with the 3' exonuclease Trex2 increases frequencies of targeted mutagenesis in plant cells. I-SceI and Trex2 were each cloned downstream of a promoter that provides high levels of constitutive expression in plant cells. Plasmids encoding I-SceI and Trex2 were then transformed into tobacco protoplasts that were derived from a tobacco line with an integrated I-SceI recognition site, providing the target for a DSB (Tzfira *et al.*, 2003). Genomic DNA was isolated from batches of protoplasts transformed with I-SceI only, I-SceI plus Trex2, Trex2 only or YFP control. 454

pyrosequencing was performed on the samples (Figure 1A). The sequencing results provided clear evidence that Trex2, in combination with a meganuclease, greatly increases frequencies of targeted mutagenesis in plants. Indels were recovered from cells expressing both I-SceI and Trex2 at a frequency 16-fold greater than with I-SceI alone (Figure 1B). The fold enhancement was not affected when the native nuclear localization signal (NLS) was replaced with SV40 NLS (Figure 1B). In addition to increasing the mutation frequency, the expression of Trex2 also altered the mutation spectrum. For example, the frequency of insertions in cells expressing Trex2 (7 out of 1238; 0.57%) was significantly lower than in cells expressing I-SceI alone (9 out of 46; 19.5%) (Figure 1C). Furthermore, the size of the deletions created by co-expression of I-SceI and Trex2 was smaller than what was observed with I-SceI alone. In the former case, 96.6% of the deletions were less than 10 bp, whereas in the latter, the number of deletions less than 10 bp was 81.1% (Figure 1C). Taken together, we conclude that co-expression of Trex2 with an active meganuclease enhances the frequency of mutagenesis and promotes the formation of small deletions at the cleavage site in plant cells.

Both constitutive and induced expression of I-SceI plus Trex2 mediates SSA in Arabidopsis plants.

Next, we wanted to test if the results in tobacco protoplasts could be recapitulated *in planta*. One approach was to target endogenous loci. For this, we chose two engineered meganucleases that were previously tested for activity in Arabidopsis (Cellectis Plant Sciences). The progress of these experiments is discussed in the next section.

At the same time, we also sought to express the well-characterized I-SceI homing endonuclease in whole plants. Because the native I-SceI target sequence does not exist in the Arabidopsis genome, we utilized a transgenic line of plants homozygous for an integrated reporter containing the I-SceI target (Orel *et al.*, 2003). The transgene is a single-strand annealing (SSA) reporter with overlapping halves of a non-functional β-glucuronidase (GUS) gene separated by the I-SceI cut site (Figure 2; GU.US). Upon cleavage and repair, a functional GUS gene is restored and the gene product can be detected as blue color after staining with X-Gluc (Figure 2A). Transgenic GU.US plants were transformed with a constitutively expressed I-SceI plus Trex2 (MN335) or an inducible I-SceI plus Trex2 (MN336). Seeds from these plants were collected and transgenic seedlings were screened for the presence of MN335 (Bar) or MN336 (hygromycin). Resistant transgenic seedlings were then stained with X-Gluc to assess cleavage efficiencies resulting from meganuclease and Trex2 coexpression.

High levels of GUS activity was observed in both cotyledons (Figure 2B) and five-leaf stage seedlings (Figure 2C) constitutively expressing I-SceI plus Trex2 (MN335). Levels of GUS activity were lower when I-SceI + Trex2 expression was induced (Figure 2D, E). The GUS only control seedlings showed little detectable GUS activity, as has been previously reported (Figure 2F, Table S1) (Orel *et al.*, 2003). The high level of GUS activity observed when I-SceI and Trex2 were co-expressed is certainly striking, and suggests that perhaps coexpression of Trex2 is influencing repair of the GUS gene. For example, nearly all of the tissue from 12 individual seedlings stained fully blue (Figure 2C, 6 shown). The lower levels observed amongst the MN336 seedlings likely reflect poor induction, as we have observed in other similar experiments

(Voytas lab). It is important to note that a comparison between expression of I-SceI only and I-SceI plus Trex2 could not be made in this experiment, and ongoing experiments will address this caveat. Previous studies in have been conducted using I-SceI and the same GU.US Arabidopsis plant lines, and can offer insight into the expected levels of repair (as a proxy for cleavage efficiency), albeit anecdotally. In addition, others have observed that expression of Trex2 in the presence of a DSB decreases SSA (Bennardo *et al.*, 2009). Although these details remain to be fleshed out with further experiments, our observations suggest that 1) our approach to integrate and co-express a meganuclease and Trex2 is likely to succeed without prohibiting nuclease activity, and 2) There appears to be no global toxicity effects or phenotypic consequences of constitutively expressing both a site-specific endonuclease and a non-selective exonuclease.

Ongoing and Future Experiments

Testing engineered meganucleases with Trex2 at endogenous loci in Arabidopsis plants.

Coexpression of a meganuclease and Trex2 appears to increase gene disruption in plant protoplasts (Figure 1). Furthermore, the combination of meganuclease and Trex2 can be integrated into the Arabidopsis genome and constitutively expressed without obvious toxic effects, and coexpression of Trex2 appears to increase the number of recombination events at an integrated reporter locus when compared to previous reports (Orel *et al.*, 2003, Plchova *et al.*, 2003) (Figure 2).

Next, we would like to determine if coexpression of meganucleases and Trex2 could enhance gene disruption at endogenous loci in Arabidopsis. A series of T-DNA constructs were made containing a combination of 1) two different engineered meganucleases, MN47 and MN62; 2) T2A or fusion of Trex2; 3) constitutive or inducible promoters (Table 1, Figure 3A). To date, Arabidopsis plants have been transformed these constructs in two separate experiments; and transgenic plants have been selected (Figure 3). The transformation efficiencies have been low, and we failed to recover any transgenic plants (Hyg or Bar resistant) expressing the meganucleases plus Trex2 in the first experiment (data not shown). Genomic DNA has been isolated from transgenic seedlings recovered in the second experiment (Figure 3B). Molecular characterization of mutation efficiencies have been attempted for samples MN47 and MN62 with and without Trex2, both co-translated (T2A) and as a fusion protein (see Table 1), but results remain inconclusive due to complications with PCR and T7E1 assays (data not shown). Efforts are underway to troubleshoot these experiments.

Proposed future studies for this project include:

1. 454 sequencing of candidate samples to determine more precise frequency of mutagenesis with and without expression of Trex2 (performed in-house at Cellectis).
2. Testing MN47 and MN67 minus and plus Trex2 using a rapid reporter system in tobacco protoplasts.
3. Test TALEN (5' overhang)/CRISPR (blunt end) plus Trex2 in tobacco protoplasts.

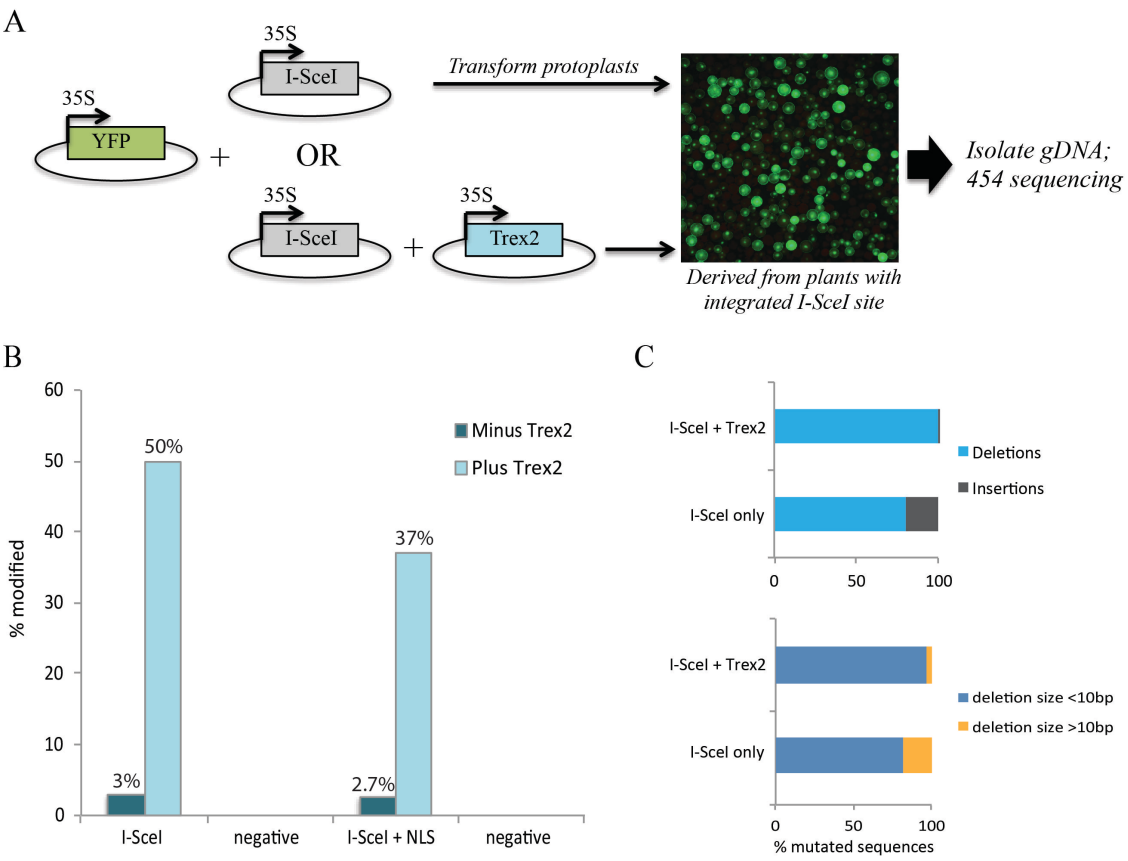


Figure 1. Coupling of I-SceI meganuclease with Trex2 enhances frequency of mutagenesis in protoplasts. (A) Schematic of protoplast transformation and sequencing workflow. Plasmids expressing I-SceI only (gray) or I-SceI plus Trex2 (blue) were introduced into tobacco protoplasts harboring an integrated I-SceI target site. A plasmid expressing YFP transformation control (green) was also delivered with each batch of transformations. Genomic DNA was isolated 24 hr after transformation and 454 sequencing was performed (see methods). (B). Plot of modified sequencing reads for each batch of transformations. Minus Trex2 (dark blue) and plus Trex2 (light blue) conditions are plotted as percent (%) modified sequences. Negative control represents sequencing on samples with no I-SceI or Trex2. (C) Types and characteristics of recovered indels. Top panel represents the proportion of mutated sequences that are deletions (light blue) or insertions (dark gray). Bottom panel shows the proportion of

recovered deletions by size; < 10 bp (blue) and > 10 bp (orange). Total indels for I-SceI: 46; for I-SceI plus Trex2: 1238.

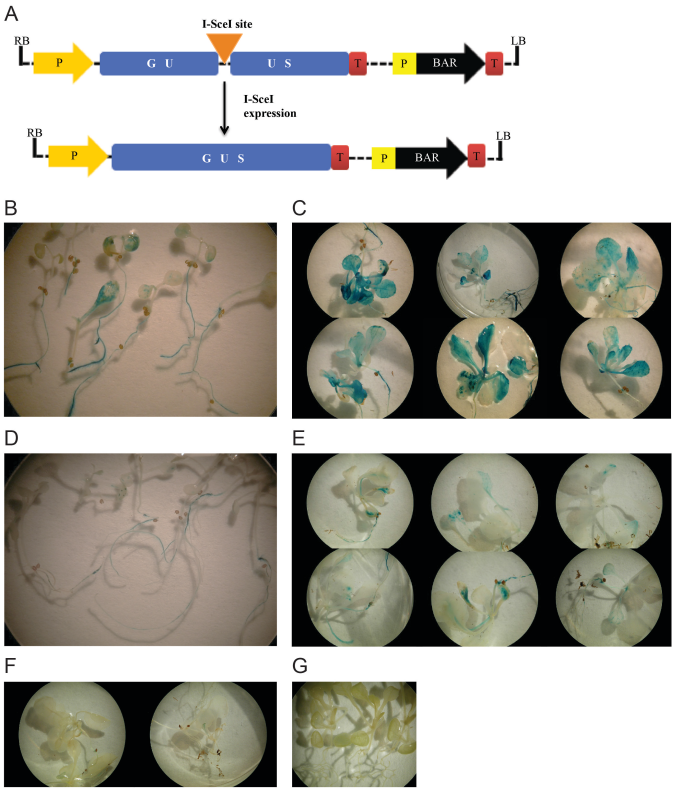


Figure 2. Activity of I-SceI plus Trex2 in SSA assay in Arabidopsis plants. (A) Schematic of the SSA assay system. The GU.US T-DNA construct contains an I-SceI target site (orange triangle) between overlapping halves of a GUS gene (blue). Expression is driven by a strong plant promoter, 35S (P; yellow arrow). A BAR gene is present as a transformation marker. RB:right border; LB:left border. After cleavage and repair, a functional GUS gene is restored and results in blue staining. (B–F) GUS assays with transgenic seedlings (see methods). GUS assays were performed on (B) cotyledons, line MN335 (constitutive I-SceI expression); (C) 5-leaf stage seedlings, line MN335 (constitutive I-SceI expression) (D) cotyledons, line MN336 (XVE: induced I-SceI expression); (E) 5-leaf stage seedlings, line MN336 (XVE: induced I-SceI expression). (F) pDGU.US control with no I-SceI (G) X-Gluc staining control seedlings, no GUS transgene.

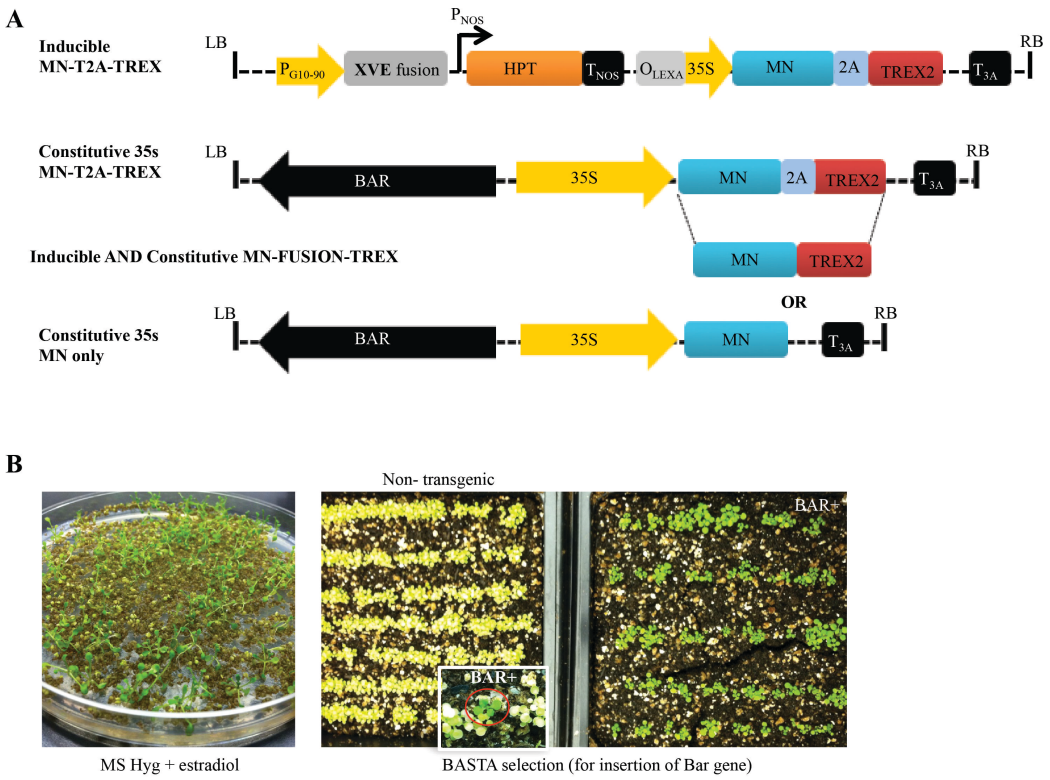


Figure 3. Expressing Meganuclease–Trex2 in Arabidopsis. (A) Schematic of Meganuclease–Trex2 plant T-DNA expression cassettes. An inducible XVE promoter (gray boxes) and constitutive 35S (yellow arrow) drive expression of the transgene. Engineered meganucleases (blue box) were co-expressed with Trex2 (red box) using a T2A translational skip element (light blue box) or as a fusion protein. A NOS promoter drives expression of a hygromycin selection marker gene. LB: left T-DNA border; RB: right T-DNA border. (B) Selection of transgenic seedlings co-expressing MN + Trex2. Image on left depicts hygromycin resistant seeds plated on MS media containing estradiol to induce expression of the MN plus Trex2. Images in the middle and right depict selection of transgenic seedlings constitutively expressing MN plus Trex2 using the herbicide BASTA to select for the presence of the Bar gene. Seedlings that do not contain an integrated transgene will turn yellow and die from exposure to BASTA (middle, *non-transgenic*), while plants with the transgene (and Bar marker) confer resistance to BASTA and thus remain healthy (middle inset, BAR+). Bar+ positive BASTA control seedlings treated with BASTA are shown on the right.

Table 1. MN–Trex2 Constructs transformed into Arabidopsis thaliana plants.

CONSTRUCT NAME	PROMOTER	SELECTION	DESCRIPTION
MN335	35s	HYG	T2A; I-SceI+Trex2
MN336	XVE	HYG	T2A; I-SceI+Trex2
MN47	XVE	HYG	No Trex2
MN62	XVE	HYG	No Trex2
MN47	35s	BAR	No Trex2
MN62	35s	BAR	No Trex2
XVE::MN47+Trex2	XVE	HYG	T2A+Trex2
XVE::MN62+Trex2	XVE	HYG	T2A+Trex2
GW6::MN47+Trex2	35s	BAR	T2A+Trex2
GW6::MN62+Trex2	35s	BAR	T2A+Trex2
XVE:: MN47+F-Trex	XVE	HYG	Fusion-Trex2
XVE:: MN622+F-Trex	XVE	HYG	Fusion-Trex2
GW6:: MN47+F-Trex	35s	BAR	Fusion-Trex2
GW6::3 MN62+F-Trex	35s	BAR	Fusion-Trex2

*GW6 = BAR
*GW2 = HYG

Table S1.* Recombination frequencies in SSA lines with and without induction of DSB by I-SceI.

Line	No. of seedlings	Sectors	Sectors per seedling	Enhancement
SSA	31	47	1.5	
SSA + I-SceI**	22	2380	108.2	72.13

*from (Orel *et al.*, 2003)

**DMC1 promoter

This table provides information from the literature to use as comparison to SSA results in this study for the SSA I-SceI only control.

References

Bennardo, N., Gunn, A., Cheng, A., Hasty, P. & Stark, J. M. (2009) Limiting the persistence of a chromosome break diminishes its mutagenic potential. PLoS genetics 5(10), e1000683. doi:10.1371/journal.pgen.1000683

Beurdeley, M., Bietz, F., Li, J., Thomas, S., Stoddard, T., Juillerat, A., Zhang, F., *et al*. (2013) Compact designer TALENs for efficient genome engineering. Nature Communications 4, 1762. Nature Publishing Group. doi:10.1038/ncomms2782

Certo, M. T., Gwiazda, K. S., Kuhar, R., Sather, B., Curinga, G., Mandt, T., Brault, M., *et al*. (2012) SI_Coupling endonucleases with DNA end-processing enzymes to drive gene disruption. Nature methods 9(10), 973–5. doi:10.1038/nmeth.2177

Clough, S. J. & Bent, a F. (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. The Plant journal : for cell and molecular biology 16(6), 735–43.

Lieber, M. R. (2010) The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annual review of biochemistry (October 2009), 181–211. doi:10.1146/annurev.biochem.052308.093131

Orel, N., Kyryk, A. & Puchta, H. (2003) Different pathways of homologous recombination are used for the repair of double-strand breaks within tandemly arranged sequences in the plant genome. The Plant Journal 35(5), 604–612. doi:10.1046/j.1365-313X.2003.01832.x

Perrino, F. W., Harvey, S., McMillin, S. & Hollis, T. (2005) The human TREX2 3' -> 5'-exonuclease structure suggests a mechanism for efficient nonprocessive DNA catalysis. The Journal of biological chemistry 280(15), 15212–8. doi:10.1074/jbc.M500108200

Plchova, H., Hartung, F. & Puchta, H. (2003) Biochemical characterization of an exonuclease from Arabidopsis thaliana reveals similarities to the DNA exonuclease of the human Werner syndrome protein. The Journal of biological chemistry 278(45), 44128–38. doi:10.1074/jbc.M303891200

Stark, J. M., Pierce, A. J., Oh, J., Pastink, A. & Jasin, M. (2004) Genetic Steps of Mammalian Homologous Repair with Distinct Mutagenic Consequences Genetic Steps of Mammalian Homologous Repair with Distinct Mutagenic Consequences. doi:10.1128/MCB.24.21.9305

Swoboda, P., Gal, S., Hohn, B. & Puchta, H. (1994) Intrachromosomal homologous recombination in whole plants. The EMBO journal 13(2), 484–9.

Tinland, B., Hohn, B. & Puchta, H. (1994) Agrobacterium tumefaciens transfers single-stranded transferred DNA (T-DNA) into the plant cell nucleus. Proceedings of the National Academy of Sciences of the United States of America 91(17), 8000–4.

Tzfira, T., Frankman, L. R., Vaidya, M. & Citovsky, V. (2003) Site-specific integration of Agrobacterium tumefaciens T-DNA via double-stranded intermediates. Plant physiology 133(3), 1011–23. doi:10.1104/pp.103.032128

Voytas, D. F. (2012) Plant Genome Engineering with Sequence-Specific Nucleases. Annual Review of Plant Biology 64(1), 130301143929006. doi:10.1146/annurev-arplant-042811-105552

Yoo, S.-D., Cho, Y.-H. & Sheen, J. (2007) Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. Nature protocols 2(7), 1565–72. doi:10.1038/nprot.2007.199

APPENDIX B

COPYRIGHT PERMISSIONS

This is a License Agreement between Michelle Christian ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3172120763549
License date	Jun 18, 2013
Licensed content publisher	Oxford University Press
Licensed content publication	Nucleic Acids Research
Licensed content title	Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting:
Licensed content author	Tomas Cermak, Erin L. Doyle, Michelle Christian, Li Wang, Yong Zhang, Clarice Schmidt, Joshua A. Baller, Nikunj V. Somia, Adam J. Bogdanove, Daniel F. Voytas
Licensed content date	09/01/2011
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	The Development, engineering and application of TAL effector nucleases and other modular DNA binding proteins for targeted genome modification
Publisher of your work	n/a
Expected publication date	Jun 2013
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD