2013

Computational and experimental analysis of TAL effector-DNA binding

Erin Lynn Doyle
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Agricultural Science Commons, Agriculture Commons, Bioinformatics Commons, Biology Commons, and the Plant Pathology Commons

Recommended Citation
Doyle, Erin Lynn, 'Computational and experimental analysis of TAL effector-DNA binding' (2013). Graduate Theses and Dissertations. 13132.
https://lib.dr.iastate.edu/etd/13132

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Computational and experimental analysis of TAL effector-DNA binding

by

Erin Lynn Doyle

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Adam J. Bogdanove, Co-Major Professor
Daniel Nettleton, Co-Major Professor
Susan Lamont
Leonor Leandro
Steve Whitham

Iowa State University
Ames, Iowa
2013

Copyright © Erin Lynn Doyle, 2013. All rights reserved.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER 1. GENERAL INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Rationale</td>
<td>2</td>
</tr>
<tr>
<td>Dissertation Organization</td>
<td>3</td>
</tr>
<tr>
<td>References</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER 2. THE EFFECTS OF TAL EFFECTORS ON CELL BIOLOGY</td>
<td>7</td>
</tr>
<tr>
<td>Abstract</td>
<td>7</td>
</tr>
<tr>
<td>Introduction</td>
<td>7</td>
</tr>
<tr>
<td>TAL effector structure and function</td>
<td>8</td>
</tr>
<tr>
<td>The natural role of TAL effectors</td>
<td>11</td>
</tr>
<tr>
<td>TAL effectors and improving plant disease resistance</td>
<td>13</td>
</tr>
<tr>
<td>TAL effectors as customizable DNA targeting proteins</td>
<td>14</td>
</tr>
<tr>
<td>Fusions for targeted gene activation and repression</td>
<td>14</td>
</tr>
<tr>
<td>TAL effector fusions for genome editing</td>
<td>15</td>
</tr>
<tr>
<td>Design and construction</td>
<td>17</td>
</tr>
<tr>
<td>Future challenges</td>
<td>17</td>
</tr>
<tr>
<td>References</td>
<td>19</td>
</tr>
<tr>
<td>CHAPTER 3. EFFICIENT DESIGN AND ASSEMBLY OF CUSTOM TALEN AND OTHER TAL EFFECTOR-BASED CONSTRUCTS FOR DNA TARGETING</td>
<td>25</td>
</tr>
<tr>
<td>Abstract</td>
<td>25</td>
</tr>
<tr>
<td>Introduction</td>
<td>26</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>29</td>
</tr>
<tr>
<td>Protocol for assembly of custom TALEN, TAL effector or TAL effector fusion-ready constructs</td>
<td>29</td>
</tr>
<tr>
<td>Construction of module, last repeat, array, and backbone plasmids</td>
<td>31</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Software to identify candidate TALEN target sites</td>
<td>33</td>
</tr>
<tr>
<td>Testing TALEN function in yeast</td>
<td>34</td>
</tr>
<tr>
<td>Expression of custom TALENs in human cells and Arabidopsis protoplasts and detection of site-specific mutations</td>
<td>35</td>
</tr>
<tr>
<td>Expression of a custom TAL effector in Xanthomonas and in planta activity assay</td>
<td>36</td>
</tr>
<tr>
<td>Results</td>
<td>36</td>
</tr>
<tr>
<td>Efficient assembly of custom repeat arrays into TALEN and other TAL effector-based constructs</td>
<td>36</td>
</tr>
<tr>
<td>Guidelines and software for TALEN site selection and repeat array design</td>
<td>37</td>
</tr>
<tr>
<td>Activity of custom TALENs in a yeast-based DNA cleavage assay</td>
<td>39</td>
</tr>
<tr>
<td>Targeted mutagenesis in human cells and Arabidopsis protoplasts using custom TALENs</td>
<td>39</td>
</tr>
<tr>
<td>Replication of AvrHah1 TAL effector activity with a Golden Gate assembled clone</td>
<td>40</td>
</tr>
<tr>
<td>Discussion</td>
<td>41</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>44</td>
</tr>
<tr>
<td>Funding</td>
<td>44</td>
</tr>
<tr>
<td>References</td>
<td>45</td>
</tr>
<tr>
<td>Supplementary Material</td>
<td>47</td>
</tr>
<tr>
<td>CHAPTER 4. TAL EFFECTOR-NUCLEOTIDE TARGETER (TALE-NT) 2.0: TOOLS FOR TAL EFFECTOR DESIGN AND TARGET PREDICTION</td>
<td>53</td>
</tr>
<tr>
<td>Abstract</td>
<td>53</td>
</tr>
<tr>
<td>Introduction</td>
<td>53</td>
</tr>
<tr>
<td>Software and Algorithms</td>
<td>56</td>
</tr>
<tr>
<td>Programming</td>
<td>56</td>
</tr>
<tr>
<td>Tools for designing custom TALENs and TAL effectors</td>
<td>56</td>
</tr>
<tr>
<td>A tool for identifying candidate TAL effector targets</td>
<td>56</td>
</tr>
<tr>
<td>Web Interface</td>
<td>57</td>
</tr>
<tr>
<td>Design and general features</td>
<td>57</td>
</tr>
<tr>
<td>TALEN Targeter</td>
<td>58</td>
</tr>
<tr>
<td>TAL Effector Targeter</td>
<td>59</td>
</tr>
<tr>
<td>Target Finder</td>
<td>59</td>
</tr>
<tr>
<td>Conclusion</td>
<td>61</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>62</td>
</tr>
</tbody>
</table>
CHAPTER 5. ADDITION OF TRANSCRIPTION ACTIVATOR-LIKE EFFECTOR BINDING SITES TO A PATHOGEN STRAIN-SPECIFIC RICE BACTERIAL BLIGHT RESISTANCE GENE MAKES IT EFFECTIVE AGAINST ADDITIONAL STRAINS AND AGAINST BACTERIAL LEAF STREAK

Summary

Introduction

Materials and Methods

Modification of Xa27
Plant transformation and growth
TAL effector clones
Isolation and transformation of Xag EB08 and cultivation of Xanthomonas strains
Plant inoculations, virulence assays, and qPCR
5' RACE
Analysis of EBE sequence representation in rice promoters
Results

Stable integration of EBE-amended Xa27 constructs into rice
EBE-amended Xa27 is activated by Xoo and Xoc independent of AvrXa27
EBE-amended Xa27 is specifically activated by each TAL effector
All six EBEs drive transcription of Xa27 primarily from a shared start site
EBE-amended Xa27 confers resistance to a broader spectrum of Xoo strains and to all tested Xoc strains
The UXO EBEs contain sequences apparently under selection in rice promoters
Discussion

Acknowledgements

References

Supporting Information

CHAPTER 6. THE ROLE OF TRYPTOPHAN 232 IN TAL EFFECTOR DNA BINDING

Abstract
X. oryzae pv. oryzicola and X. oryzae pv. oryzae induce very different gene expression changes in rice leaves.

Many up-regulated genes are predicted to be targets of TAL effectors.

Experimental validation yields 19 confirmed TAL effector targets for X. oryzae pv. oryzicola TAL effectors.

Comparison of validated and un-validated predicted TAL effector binding sites reveal characteristics that may be useful for future TAL effector target prediction.

The sulfate transporter gene targeted by Tal2g is a major S gene for bacterial leaf streak.

The sulfate transporter gene targeted by Tal2g is a major S gene for bacterial leaf streak.

X. oryzae pv. oryzicola and X. oryzae pv. oryzae induce very different gene expression changes in rice leaves.

Many up-regulated genes are predicted to be targets of TAL effectors.

Experimental validation yields 19 confirmed TAL effector targets for X. oryzae pv. oryzicola TAL effectors.

Comparison of validated and un-validated predicted TAL effector binding sites reveal characteristics that may be useful for future TAL effector target prediction.

The sulfate transporter gene targeted by Tal2g is a major S gene for bacterial leaf streak.

Discussion
LIST OF FIGURES

CHAPTER 2

Figure 1. Structure of the TAL effector-DNA association and the basis of specificity. .......... 9
Figure 2. TAL effectors in plant disease, natural disease resistance, and engineered resistance. ................................................................. 12
Figure 3. TALEN-mediated genome editing. ................................................................. 15
Figure 4. Applications of custom TAL effectors. ............................................................. 19

CHAPTER 3

Figure 1. TAL effector and TALEN structure. ................................................................. 27
Figure 2. Golden Gate assembly of custom TAL effector and TALEN constructs using module, array, last repeat and backbone plasmids. ......................................................... 28
Figure 3. TALEN or TAL effector construct assembly timeline. ...................................... 29
Figure 4. Nucleotide and RVD frequencies at the termini of 20 target and TAL effector pairs. ........................................................................................................................................ 38
Figure 5. Activity of 15 custom TALEN pairs targeting diverse sequences in a reporter-based yeast assay ............................................................................................................. 40
Figure 6. Site-directed mutagenesis in human embryonic kidney cells and Arabidopsis protoplasts using custom TALENs. ................................................................. 41
Figure 7. Activity of an AvrHah1 analog created using the Golden Gate method and our plasmid set. ............................................................................................................. 42
Supplementary Figure S1. Module, array, repeat, and backbone plasmids......................... 50
Supplementary Figure S2. The TAL Effector-Nucleotide Targeter. ..................................... 51
Supplementary Figure S3. The yeast assay for testing TALEN function. ........................... 52

CHAPTER 5

Figure 1. Engineered Xa27 constructs used in this study.................................................. 70
Figure 2. Type III secretion system (T3SS)-dependent and AvrXa27-independent activation of effector binding element (EBE)-amended Xa27 by Xanthomonas oryzae pv. oryzae (Xoo) PXO99A and Xanthomonas oryzae pv. oryzicola (Xoc) BLS256. ........ 74
Figure 3. Specific activation of Xa27 in the UXO construct by each of the corresponding transcription activator-like (TAL) effectors individually. ........................................... 75
Figure 4. Xa27 transcriptional start sites observed in the UXO-3 plants in response to individual transcription activator-like (TAL) effectors, Xanthomonas oryzae pv. oryzae (Xoo) PXO99A, BLS256, or a mock inoculation ........................................................................ 77
**Figure 5.** Bacterial blight and bacterial leaf streak resistance in the effector binding element (EBE)-amended Xa27 transgenic plants. .................................................................78

**Figure 6.** Effector binding element (EBE) sequence representation in rice promoters........80

**Figure S2.** Specific activation of Xa27 in UXO-1 plants by each of the corresponding TAL effectors individually. ...........................................................................................................89

### CHAPTER 6

**Figure 1.** GUS activity of TAL effectors with single amino acid substitutions for W232 on targets preceded by A, C, G, or T at the 0th position. .................................................................103

**Figure 2.** GUS activity of chimeric TAL effector RScPthXo1 with single amino acid substitutions for R232 on targets preceded by A, C, G, or T at the 0th position. ............105

**Figure 3.** Activity of TALENs with the TAL868 repeat sequence and W232 substitutions on targets preceded by A, C, G, or T at the 0th position. .........................................................106

**Figure 4.** Relative binding affinities of TAL868 W232 substitution variants for targets preceded by A, C, G, or T at the 0th position.................................................................107

**Figure 5.** Structural models for representative W232 substitutions. ......................................109

**Supplementary Figure S1.** GUS activity for PthXo1 variants with all 19 possible single amino acid substitutions for W232.................................................................121

**Supplementary Figure S2.** ClustalW multiple sequence alignment of N terminal sequences of Ralstonia TAL-like effectors (RTLs) and PthXo1 .......................................................122

**Supplementary Figure S3.** Secondary structure predictions for PthXo1 and Rsc1815 N terminal regions. ..................................................................................................................123

### APPENDIX

**Appendix Figure 1.** Rice transcriptional responses to Xanthomonas oryzae pv. oryzicola BLS256 (Xoc) or X. oryzae pv. oryzae PXO99A (Xoo). .................................................................135

**Appendix Figure 2.** Virulence contribution of X. oryzae pv. oryzicola BLS256 TAL effector Tal2g. ......................................................................................................................137

**Appendix Figure 3.** Determination of Os01g52130 as the relevant target of Tal2g using designer TAL effectors. .....................................................................................................139
LIST OF TABLES

CHAPTER 3

Supplementary Table S1. Custom TALENs and target sites. ..................................................47
Supplementary Table S2. Activity, conformity to rules, and length of TALENs tested in
the yeast assay. ..................................................................................................................49

CHAPTER 4

Supplementary Table S1. Target Finder RVD-nucleotide association probabilities........65

CHAPTER 5

Table S1. Alignments of TAL effector RVDs and EBE sequences used in this study. ........90
Table S2. Relative abundance of Xa27 mRNA in mock-inoculated transgenic lines
compared to IRBB7. ............................................................................................................91

CHAPTER 6

Supplementary Table S1. Primers used for pAH236 tryptophan 232 substitutions.........116
Supplementary Table S2. Results of paired t-tests for significant effects of W232
substitutions on GUS activity. ............................................................................................117

APPENDIX

Appendix Table 1. Predicted X. oryzae pv. oryzicola BL256 TAL effector targets in rice
and target verification. .......................................................................................................146
Appendix Table 2. Predicted X. oryzae pv. oryzae PXO00A TAL effector targets in
rice ......................................................................................................................................148
ABSTRACT

TAL effectors, from the plant-pathogenic bacterial genus Xanthomonas, are DNA binding proteins that can be engineered to bind to almost any sequence of interest. The DNA target of the TAL effector is encoded by a modular central repeat region, with each repeat specifying a single binding site nucleotide. TAL effectors can be targeted to novel DNA sequences by assembling the corresponding repeat sequence. Therefore, custom TAL effectors have become important tools for manipulating gene expression and creating site-specific DNA modifications. This dissertation explores TAL effector-DNA binding through computational and experimental analyses.

I identified positional and composition biases in known TAL effector-target pairs and proposed guidelines for designing custom TAL effectors and TAL effector nucleases (TALENs). Using these guidelines, I created a software tool for TAL effector design. We expanded this tool into a suite of tools for TAL effector/TALEN design and target site prediction. Target site predictions can be used to estimate potential off-target binding of custom TAL effector constructs or to identify unknown targets of natural TAL effectors.

Next, I present a case study in engineering disease resist rice plants. Inserting multiple TAL effector binding elements (EBEs) into the promoter of a rice resistance gene conferred resistance to diverse strains of Xanthomonas oryzae. Analysis of the EBE sequences revealed that TAL effectors have evolved to target specific host regulatory sequences, and caution is warranted when introducing such sequences into the promoter of an executor resistance gene.

Finally, I examine the role of the TAL effector N terminus in DNA binding. Most natural TAL effector binding sites are preceded by a T at the 5’ end (T0). Structural data suggests T0 is encoded by tryptophan 232 (W232) in the cryptic -1st repeat. We show that substitutions for W232 alter TAL effector activity and specificity for T0. However, we find that the TAL effector-T0 interaction is complex and may depend on other residues in the -1st repeat, the 0th cryptic repeat, or repeat sequence context. Better understanding of TAL effector-DNA binding will improve TAL effector design and target prediction and enhance understanding of the role of TAL effectors in plant disease.
CHAPTER 1. GENERAL INTRODUCTION

Abbreviations

CRR: central repeat region
dTALE: designer TAL effector
EBE: effector binding element
R gene: resistance gene
RTL: Ralstonia TAL-like effector
RVD: repeat variable diresidue
S gene: susceptibility gene
TAL effector: transcription activator-like effector
TALEN: TAL effector nuclease
T3SS: type three secretion system

Introduction

Transcription activator-like (TAL) effectors are a class of transcriptional activators found primarily in the plant pathogenic bacterial genus Xanthomonas. During infection, they are secreted by the bacterium directly into the host plant cell and subsequently localized to the nucleus where they bind to effector-specific DNA targets and activate transcription of host genes. Many TAL effectors activate host genes known as susceptibility (S) genes, whose expression is necessary for bacterial spread and multiplication and/or for the development of disease symptoms. A few TAL effectors bind to and activate genes that trigger disease resistance (known as resistance or R genes) (1,2).

In 2009, the mechanism by which TAL effectors bind to their specific DNA targets was discovered independently by two different labs. The TAL effector’s central repeat region (CRR) consists of a variable number of tandem, nearly identical repeats of 33-34 amino acids in length. Repeat-to-repeat variation occurs primarily at amino acids 12 and 13, which are together known as the repeat variable diresidue, or RVD. Computational and experimental analysis revealed a one-to-one correspondence between binding site nucleotides and repeats, such that the sequence of RVDs “encodes” the binding site (3,4). Solved structures of two different TAL effectors bound to their DNA targets later confirmed that each RVD interacted with and specified a single binding site nucleotide (5,6)
Discovery of this RVD-nucleotide binding code has led to increasing interest in TAL effectors, both for their role in plant disease and their potential as research tools. Knowledge of RVD-nucleotide specificity makes it possible to predict previously unknown TAL effector binding sites, potentially leading to the discovery of novel S or R genes and opening up novel paths for engineering disease resistant plants (7-9). Additionally, RVDs specifying each of the four nucleotides have been identified, making it possible to design custom TAL effectors to target a given DNA sequence by assembling a TAL effector CRR with the appropriate sequence of repeats/RVDs. These custom TAL effectors and TAL effector-based constructs have been used in a variety of applications, including targeted gene activation (using the native TAL effector activation domain or replacing it with another activation domain) targeted gene repression (typically by replacing the TAL effector activation domain with a repressor domain), and targeted genome modifications (by using TAL effector-endonuclease fusion proteins known as TAL effector nucleases or TALENs to create targeted double strand breaks in the DNA) (10-17). TAL effector-based constructs have been demonstrated to work effectively in a wide variety of tissue and cell types. TALENs in particular have become a widely used research tool, with potential uses ranging from creating gene knockouts in model and non-model organisms to disease therapies in humans (18).

Rationale

This dissertation focuses on computational and experimental analysis of TAL effector-DNA binding. I begin by analyzing known TAL effector-target pairs from nature and applying the results to create computational tools to design custom TAL effectors and TALENs. The design of maximally efficient TAL effector-based constructs is an important consideration for researchers. Because I reasoned that TAL effectors in nature had likely evolved to have maximum binding affinity for their targets, this set of computational tools focused on selecting target sites and corresponding RVD sequences that were similar in sequence composition to natural TAL effectors and known TAL effector targets. Later, I altered these initial design guidelines to reflect accumulating data on the activity of custom TAL effector-based constructs as well as information from the structures of two TAL effectors bound to their targets.

An additional concern when designing custom TAL effector constructs is their specificity for their intended target sequence. Because the RVD-nucleotide binding code is not absolute, a custom TAL effector can bind to sites other than the intended target but with sufficiently similar sequence. Such off-target binding is especially problematic in applications using TALENs, where it may result in the unintended cleavage and subsequent disruption of non-target genes.
I extended my TAL effector and TALEN design software to include tools to predict potential off-target binding sites based on a mathematical scoring function. In addition to aiding researchers in designing highly specific custom TAL effectors, target prediction tools can also be used to identify previously unknown targets of natural TAL effectors, leading to the identification of novel plant disease S or R genes (a case study using TAL effector target prediction tools to discover a previously unknown S gene is included in the Appendix). I also show that TAL effectors may have evolved to target specific sequences under selection, which has important implications for the use of TAL effectors in engineered plant disease resistance.

Finally, I examine the role of a specific region of the TAL effector’s N terminus in DNA binding. Although the work described above is based primarily on the repeat/RVD sequence of the TAL effector, this N-terminal region located outside of the CRR may also play a role in defining the binding site and in nucleating TAL effector-DNA binding. Understanding the role of this region may lead to the design of more efficient TAL effectors for biotechnology applications and may improve DNA targeting by custom TAL effectors.

Dissertation Organization

This dissertation is organized into five parts. Chapter 2 presents a review of TAL effector literature. Chapters 3 and 4 describe the creation and development of a suite of software tools for designing and predicting targets for TAL effectors and TALENs. Chapter 5 is a case study in engineering plant disease resistance based on naturally occurring TAL effectors, including computational analysis showing that TAL effectors have evolved to target specific gene sequences under selection. Finally, Chapter 6 details experiments to study the role of the TAL effector’s cryptic –1st repeat region in TAL effector-DNA binding. The Appendix presents an additional case study which verifies the functionality and utility of the target prediction tools by using them to identify novel plant disease S genes. My contributions to the individual chapters were as follows:

- For the literature review in Chapter 2, I was responsible for conducting the literature search, outlining the article, writing the majority of the text, and designing the figures. Co-author Barry Stoddard contributed to the section “TAL effector structure and function” and also created Figure 1, co-author Daniel Voytas contributed to the section “TAL effector fusions for genome editing”, and co-author Adam Bogdanove edited the paper.
Chapter 3 describes a novel method for the rapid construction of custom TAL effectors and TALENs. Additionally, it presents guidelines for designing TAL effectors/TALENs and a software tool to assist users in designing constructs that conform to these guidelines. My role in this chapter was analysis of known TAL effector-target pairs to develop the design guidelines, creation of the TAL effector design software and website, design of TALENs targeting genes from a variety of organisms (activity of these TALEN pairs are reported in the manuscript), and writing all of the related sections of the manuscript. Tomas Cermak and I were designated co-first authors of this manuscript.

Chapter 4 discusses a complete suite of software tools that I created for designing custom TAL effector constructs and predicting TAL effector binding sites. I was responsible for the original design and coding of all software tools, overseeing development of the web interface, drafting the manuscript, and preparing the supplementary materials.

Chapter 5 presents a case study in which multiple TAL effector binding sites were inserted into the promoter of rice disease resistance gene Xa27, broadening the plants' resistance to include multiple strains of Xanthomonas oryzae. For this chapter, I designed and conducted the analysis of TAL effector binding site representation in the rice genome, wrote the related sections of the manuscript, and prepared Figure 6.

Chapter 6 details a set of experiments to characterize the role of tryptophan 232 and the cryptic -1st repeat in TAL effector-DNA binding. I was responsible for the conception and initial design of the study, as well as data interpretation, preparing figures (except Figure 5), and writing and assembling the manuscript. I also conducted all of the GUS experiments for TAL effectors PthXo1 and TAL868, and the RTL secondary structure analysis.

The study presented in the Appendix represents part of a larger study currently in preparation for publication. I created the computational tools for target prediction, generated the initial TAL effector target predictions, cross-referenced the predicted target lists with microarray data to identify candidate target genes, and analyzed the confirmed target genes to identify characteristics of TAL effector targets. I also wrote all of the related sections of the manuscript describing the larger study, and modified and compiled relevant sections from the longer manuscript into the Appendix.
References


CHAPTER 2. THE EFFECTS OF TAL EFFECTORS ON CELL BIOLOGY

A paper submitted to Trends in Cell Biology

Erin L. Doyle¹, Barry L. Stoddard², Daniel F. Voytas³, and Adam J. Bogdanove¹ ⁴ ⁵

Abstract

Transcription activator-like (TAL) effectors are specific, customizable DNA binding proteins. DNA recognition occurs via a polymorphic repeat region that encodes the target sequence through a one-to-one correspondence between repeats and nucleotides. TAL effectors are transkingdom transcription factors, injected into plant cells by plant pathogenic bacteria in the genus Xanthomonas. They can act as virulence factors by activating host “susceptibility” genes important for disease, or as avirulence factors by turning on a gene that confers resistance. Understanding TAL effector DNA recognition has facilitated identification of targets and opened new avenues for engineering disease resistance. The customizability of TAL effectors has made them transformative tools for applications such as targeted control of gene expression and genome editing in a variety of organisms. This review discusses the structural basis for TAL effector-DNA specificity, the role of TAL effectors in plant pathology, engineered plant disease resistance, TAL effector-based DNA targeting applications, and future challenges.

Introduction

TAL (transcription activator-like) effectors are transcriptional activators injected into plant cells by plant pathogenic members of the bacterial genus Xanthomonas. In the plant nucleus, TAL effectors act as virulence factors by turning on genes necessary for disease susceptibility (S genes), or in a few cases, as avirulence factors by activating a gene that confers resistance (an R gene) [1, 2]. TAL effectors recognize their targets via a polymorphic repeat region that

¹ Department of Plant Pathology and Microbiology, Iowa State University, 351 Bessey Hall, Ames, IA 50011
² Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N. A3-025, Seattle WA 98109
³ Department of Genetics, Cell Biology & Development and Center for Genome Engineering, 321 Church Street SE, University of Minnesota, Minneapolis, MN 55455
⁴ Department of Plant Pathology and Plant-Microbe Biology, Cornell University, 334 Plant Science, Ithaca NY 14853
⁵ Corresponding author
encodes a specific nucleotide sequence through a one-to-one correspondence between repeats and consecutive bases on one strand of the DNA [3-6]. The modularity of this mechanism facilitates target identification and allows customization to generate DNA binding proteins with specificities of choice. Target identification is improving understanding of the role of TAL effectors in plant diseases and is informing efforts to develop more effective means to control those diseases. The customizability of TAL effectors has led to widely enabling TAL effector-based tools for systems biology and genome engineering, advancing fundamental research, crop and livestock improvement, and medicine [1, 7].

The TAL effector-DNA recognition mechanism hinges on a central protein domain that is composed of a variable number of tandem, 33-34 amino acid repeats. The repeats are nearly identical, with variation occurring primarily at amino acids 12 and 13. These two positions are known as the repeat variable diresidue (RVD). The sequence of RVDs has been shown both computationally and experimentally, and more recently structurally, to “encode” the TAL effector binding site on the DNA, with each RVD specifying a single binding site nucleotide through direct interaction [3-6].

This means of protein-DNA recognition and the functions it enables are so far unique. From the remarkable targeted effects these proteins have on plant cell biology to the transformative effects they are having on the field of cell biology itself, TAL effectors are a fascinating and useful invention of nature. TAL effector research is moving fast. Since publication in 2009 of the discovery of the RVD-nucleotide code that governs TAL effector-DNA recognition, the number of papers related to TAL effector function and application has risen sharply and steadily, from fewer than ten in 2009 to over 110 in 2012 [8]. In this review, we discuss highlights concerning the structural basis for TAL effector-DNA binding and specificity, the roles of TAL effectors in plant pathology and engineered plant disease resistance, TAL effector-based applications, and future challenges for TAL effector research.

**TAL effector structure and function**

The four most common RVDs, HD, NG, NN, and NI, respectively specify C, T, G or A, and A. Other RVDs of interest for DNA targeting include NS and N* (the asterisk represents a gap at position 13, resulting in a 33 amino acid repeat), which have lax specificity, and the less common NH, which has stringent specificity for G [9]. Two recent crystallization studies provided a clear view of the structural basis and context for all but one of these interactions. One illustrated the free and DNA-bound forms of an artificially engineered TAL effector termed 'dHAX3', that contains 11 canonical TAL repeats and three of the most common RVDs (HD, NG
and NS) [6]. The other presented the DNA-bound structure of the naturally occurring TAL effector PthXo1 from X. oryzae [5] (Figure 1). That structure contained over 20 repeats bound to two full turns of DNA and six RVDs (HD, NG, NN, NI, N*, and the rare RVD HG), as well as a portion of a positively charged N-terminal region that interacts nonspecifically with the 5' DNA region preceding the target site.

Each TAL repeat forms a two-helix bundle, in which the RVD forms a well-ordered loop. The individual repeats self-associate to wrap around along the entire length of the unbent DNA target site. Unbound dHAX3 displays a slightly unwound, but still helical structure [6]. Thus, a significantly more extended effector conformation appears to be required for DNA target acquisition (a hypothesis that agrees with published small angle X-ray scattering data on the

![Figure 1. Structure of the TAL effector-DNA association and the basis of specificity.](image)

At top, the structure of PthXo1 bound to its DNA target site [5] is shown (left) from the side of the DNA duplex and (right) looking down the axis of the DNA. The effector contains 22.5 repeat modules each colored separately. In the side view, the N-terminal end of the protein, containing two cryptic repeats that engage the DNA backbone via a series of basic residues, and that also capture the strongly conserved thymine (5' T) at position 'zero' of the binding site, is leftmost. The labeled repeats (14, 15, and 16) are shown in detail at bottom. Bottom left illustrates the contacts made by the 'HD' RVD (residues 12 and 13) in repeat number 14. The histidine at position 12 in the repeat forms a hydrogen bond to the backbone carbonyl oxygen of residue 8 in the first alpha-helix, while the aspartate at position 13 forms a hydrogen bond to the extracyclic amino nitrogen of the cytosine base. Bottom right shows repeats 14, 15, and 16 interacting with the DNA, illustrating that consecutive RVDs (HD, NG and NN, respectively in these repeats) contact consecutive bases (cytosine, thymine, and guanine in this case) on the same DNA strand.
full-length PthA TAL effector [10]).

The first residue in each RVD (at position 12) forms a structural interaction with the protein backbone carbonyl in the preceding helix of the same repeat, while the second residue makes sequence-specific contacts in the major groove to a single nucleotide base; all such contacts are made to a single DNA strand. The majority of observed contacts between the RVDs and DNA bases correspond to (i) directional hydrogen-bonds (observed for ‘HD’ RVDs in contact with cytosine, or for ‘NN’ RVDs in contact with purines), (ii) highly complementary steric packing in the absence of hydrogen bonds (such as between NG or HG RVDs and thymine), or (iii) interactions that appear to achieve reduced (but not completely negligible) specificity through steric exclusion of alternate bases (in particular, the ‘NI’ RVD). The 'N*' RVD, in which the RVD loop is truncated by one residue, appears to accommodate any base with little or no contribution to overall affinity. These observations agree with a recent study that examined artificial TAL effectors with blocks of homopolymeric repeats to determine the specificity and contribution to activity, as a proxy for binding affinity, of the most common RVDs and several rare RVDs. This study demonstrated that 'HD' and 'NN' (or 'HN') repeats provide a significant contribution to overall TAL effector activity (and therefore, presumably binding affinity), whereas strings of 'NI' or 'N*' repeats result in considerably lower activity and reduced specificity [9].

The pattern of contacts and specificity described above can be extended to the recognition of modified bases: the presence of an NG or HG repeat (which are specific for thymine in an unmodified target) can accommodate similar interactions with a 5-methyl-cytosine, thus making it possible to identify or design potential TAL effectors that can discriminate between target sites that contain methylated CpG sequences and those that are unmodified [11]. A more recent study indicates that TAL effectors can also recognize RNA-DNA hybrids by recognizing the sequence of the deoxynucleotide-containing strand [12].

The structure of PthXo1 [5] and the relative performance of artificial TAL constructs containing a variety of N- and C-terminal truncations [13, 14] appear to indicate that an N-terminal region immediately preceding the beginning of the canonical TAL repeats forms a highly basic region of the protein that is critical for high affinity DNA binding, and also engages a thymine base found at position zero of nearly all TAL binding sites via contacts with a single tryptophan residue that is found in all Xanthomonas TAL effectors. A detailed analysis of an extended N-terminal region of dHax3 indicates that as many as four additional cryptic repeats are formed immediately upstream of the central repeat region, and that this region provides the bulk of binding energy required for high affinity target binding and sequence-specific recognition [15].
Finally, comparisons of the structures and individual repeats of dHAX3 and PthXo1 indicate that these proteins display extraordinarily slight structural differences from one another (r.m.s.d. ~0.8 angstroms across all repeats) despite comprising distinct strings of RVDs that target completely different DNA target sites. This observation provides structural confirmation of the modular nature of DNA recognition by TAL effectors, which enables the prediction of targets in plant disease, the manipulation of targets for disease resistance, and the customization of TAL effectors for DNA targeting applications.

**The natural role of TAL effectors**

Different *Xanthomonas* species collectively infect a wide range of economically important crops and ornamentals. Sequenced *Xanthomonas* genomes typically encode zero to six TAL effectors, with some encoding upwards of thirty [16]. These have from 1.5-33.5 repeats in their DNA recognition domains, with an average of 17.5 [2]. The final repeat is the “half repeat”, but, truncated at 20 amino acids and containing an RVD, it is functional. Despite the natural variation in numbers of repeats in naturally occurring TAL effectors, experimental evidence suggests that a minimum of 10 repeats is necessary for TAL effector activity [3]. TAL effector genes encoding fewer repeats likely exist as remnants of, or raw material for, adaptations in TAL effector gene content through recombination.

Delivery of TAL effectors into plant cells takes place through the bacterial type III secretion system within hours of inoculation [17]. The type III secretion signal resides N terminal to the repeat region. C-terminal, the proteins display a highly conserved trio of nuclear localization signals, and the acidic activation domain that confers the namesake function of these proteins. Many TAL effectors activate susceptibility (S) genes, whose expression facilitates bacterial multiplication and spread, or the development of symptoms, or both [1, 2] (Figure 2). TAL effector AvrBs3, from *Xanthomonas campestris* pv. vesicatoria (a pepper pathogen) targets the cell size regulator *UPA20*. The cell hypertrophy triggered by *UPA20* activation is hypothesized to contribute to bacterial exudation to the leaf surface for dissemination [18, 19]. Several *Xanthomonas oryzae* TAL effectors target a family of genes encoding sugar exporters known as SWEET proteins [20-23]. Inducing expression of these genes would ostensibly shunt energy-rich photosynthate into the extracellular space for consumption by the bacteria [24], though this explanation has not yet been tested experimentally. At least one TAL effector induces a host transcription factor, indicating that individual TAL effector contributions to virulence may involve broad and indirect alterations to host transcription [25].

Plants have evolved resistance mechanisms to thwart TAL effector-wielding pathogens
Figure 2. TAL effectors in plant disease, natural disease resistance, and engineered resistance. Following delivery into the host plant cytoplasm via the bacterial type III secretion system (T3SS) and translocation to the nucleus, TAL effectors activate host susceptibility (S) genes that contribute to disease. Resistant plants may harbor a mutation in a TAL effector binding site in a major S gene that blocks activation, or they may harbor an S gene mimic called an executor resistance (R) gene that triggers a localized cell death and limits the infection when activated by a TAL effector. Strategies for engineered resistance to pathogens that depend on TAL effectors as virulence factors include (bottom left) site-directed mutagenesis of TAL effector binding sites sites in major S genes, using engineered nucleases such as TALENs (see also Figure 3), (bottom right) R genes with multiple TAL effector binding sites added to their promoters to provide greater durability in the face of possible TAL effector loss by the pathogen, and to broaden specificity, and (bottom middle) an S gene RNA silencing construct driven by one or more TAL effectors. In this last example, the context for the TAL effector binding site in the silencing construct should be different from that of the S gene to avoid silencing any endogenous expression of the S gene in the absence of the TAL effector.

(Figure 2). These mechanisms include adaptations that prevent key S genes from being activated. The rice resistance gene *xa13* is an allele of the major S gene *SWEET13* that harbors a short deletion in its promoter that destroys the binding site for the corresponding TAL effector PthXo1 [20]. The rice resistance gene *xa5* encodes the gamma subunit of the general transcription factor TFIIA but harbors a mutation that results in a single amino acid substitution. This substitution is thought to disrupt the presumed association of TAL effectors with the host transcriptional machinery, and prevent activation of S genes [26]. Resistance mechanisms also include S gene mimics, in which a TAL effector target site resides in the promoter of a so-called “executor resistance (R) gene” [1]. Activation of such genes (including *Bs3* and *Bs4c* in pepper and *Xa27* in rice) by corresponding TAL effectors results in a localized host cell death and
blocks infection [27-29]. In yet another form of resistance, the *Xanthomonas campestris* TAL effector AvrBs4 is recognized prior to nuclear import by the cytoplasmic host protein Bs4, and this recognition triggers a plant immune response [30].

Though the distribution of TAL effectors in nature appears largely confined to *Xanthomonas*, TAL effector-like proteins are found in another plant pathogenic bacterium, the soil borne species *Ralstonia solanacearum*. These *Ralstonia* TAL like-effectors (RTLs) harbor repeats of 35 amino acids and contain RVDs rarely or not observed in *Xanthomonas*. The RTL repeats were recently shown to mediate specific DNA binding [31], with binding sites determined by their RVD sequence. The specificity of RTL RVDs that differ from a *Xanthomonas* RVD only in the 12th position are consistent with what one would predict from the *Xanthomonas* structures, strongly suggesting a conserved overall protein topology. No RTL activity has yet been demonstrated or RTL-targeted genes identified, so the role of these proteins in plant disease remains unknown.

**TAL effectors and improving plant disease resistance**

Discovery of the TAL-DNA binding code is advancing not only our understanding of TAL effector-dependent mechanisms in disease and defense, but also our ability to identify new sources of resistance. For example, knowledge of S genes allows breeders to search for alternate alleles that lack the TAL effector binding site. New S gene candidates can be identified by scanning the promoters of all genes in a plant genome for matches to TAL effectors that are important for virulence [32] and selecting from among them ones that are upregulated during infection. These candidates are then examined individually for dependence on the corresponding effector for expression, and for a role in disease. Genome wide predictions combined with TAL effector dependent expression data can also expedite the molecular identification of new executor R genes when the corresponding TAL effector is known [33].

The TAL-DNA code also can be used to engineer altogether new types of resistance (Figure 2). Identification of TAL effector binding sites in S gene promoters has enabled targeted mutagenesis of these sites (e.g., by using engineered nucleases; see next section) to generate plants that resist infection [34]. However, such resistance may not prove to be durable; it could be overcome by the evolution of an alternative TAL effector that binds to the new promoter or to an alternate S gene. Likewise, executor R genes might be broken by pathogen loss or modification of the activating TAL effector. For more durable resistance, binding sites for multiple TAL effectors might be added to an executor R gene promoter, selecting sites for TAL effectors that are important for virulence. This site stacking strategy can also broaden the
specificity of an R gene. For example, it was used to generate rice plants resistant to multiple strains of the two pathogenic variants of Xanthomonas oryzae [35]. However, that same study showed that TAL effector binding sites likely overlap endogenous cis regulatory elements; adding such elements to the promoter of a cell death-triggering executor R gene could cause unintended activation under unforeseeable conditions, leading to aborted development or death of the plant [35]. Therefore, alternative strategies should also be explored. One possibility is a TAL effector-activated S gene silencing construct. Using a promoter identical to that of the S gene to drive expression of the silencing construct might seem a desirable failsafe against evolution of alternate TAL effectors that activate the S gene, but it could be problematic since it would lead to silencing of any endogenous expression of the S gene. Using the promoter of an alternate target with a different context for the TAL effector binding site might solve the problem, provided the expression pattern of the target in the absence of the TAL effector has no overlap with that of the S gene. Or, as with the stacked R gene promoter, a construct that contains multiple TAL effector binding sites could be used.

**TAL effectors as customizable DNA targeting proteins**

**Fusions for targeted gene activation and repression**

TAL effectors can be easily targeted to desired DNA sequences by assembling the corresponding sequence of repeats. Custom TAL effectors have been shown to be effective for targeted gene activation in a variety of cell types, often increasing gene expression by more than 20-fold. In plant cells, high levels of gene activation have been achieved using the native TAL effector activation domain [36]. In human and other mammalian cells, activation of target genes was highest when the native activation domain was replaced by the VP16 activation domain from the herpes simplex virus or its tetrameric derivative VP64 [37, 38]. Similarly, targeted gene repression in plant and animal cells has been achieved using custom TAL effectors in which the activation domain is replaced by a repressor domain, or, in yeast, when it is simply removed [39-41]. Custom TAL effectors (as activators) can fail in the face of epigenetic silencing of the target locus [42], though as described in the structure section above, substituting NG for HD should accommodate cytosine methylation. Despite this limitation, TAL effector-based gene regulation has potential in synthetic biology, in which the ability to tightly control components of novel genetic circuits is critical [39, 43]. TAL effector-based gene activators and repressors may also be useful for treating diseases related to gene expression defects. In a proof of concept study, a custom TAL effector was used to increase transcription of
the human frataxin gene. Low expression of frataxin causes Friedreich ataxia, a disease characterized by progressive nervous system damage [44].

**TAL effector fusions for genome editing**

Custom TAL effector domains were rapidly adopted to create sequence specific nucleases – so-called TAL effector nucleases or TALENs [45, 46]. A TALEN comprises a TAL effector repeat array that recognizes a specific target sequence, fused to the catalytic domain of the endonuclease *FokI*. *FokI* functions as a dimer, so two TAL effector arrays are engineered to bind sites on opposing strands of DNA that are separated by a short spacer (typically 15–20 bp). The binding of the TAL effector domains brings the *FokI* monomers into proximity, resulting in DNA cleavage within the spacer sequence. The double strand break (DSB) activates cellular DNA repair pathways, which can be harnessed to achieve desired DNA sequence modifications at or near the break site (Figure 3).

Non-homologous end-joining (NHEJ), which involves the rejoining of the broken ends of the chromosome, is one of the primary means by which DSBs are repaired in eukaryotes [47-49]. NHEJ is sometimes imprecise, and occasionally small deletions or insertions are introduced at the break site. If imprecise repair of a TALEN-induced DSB occurs in a coding sequence, mutations are introduced that can knock out gene function. Alternatively, if two or more breaks are introduced simultaneously in the genome, then a variety of chromosomal rearrangements can result, including large-scale deletions, inversions or translocations. Targeted mutagenesis

![Figure 3. TALEN-mediated genome editing](image-url)

**Figure 3. TALEN-mediated genome editing.** Double strand breaks introduced by TALENs are repaired by non-homologous end joining (NHEJ), leading to short insertions or deletions, or by homologous recombination (HR), which can be used to replace or insert new DNA. TALENs are shown as TAL effector fusions to the catalytic domain of the type IIS restriction endonuclease *FokI*, which cuts as a dimer.
achieved through NHEJ facilitates the study of gene function or enables creation of organisms or cell lines of novel genetic composition.

Another means by which DSBs can be repaired is through homologous recombination (HR), also referred to as gene targeting [47-49]. HR is achieved by introducing into cells a DNA fragment encoding a variant of a sequence of interest. Once in the cell, HR between the incoming DNA and the native locus at the cut site results in an exchange of genetic information, thereby incorporating the sequence alterations at the locus. Sequence alterations can range from single base pair substitutions to large DNA insertions. The diversity of DNA sequence modifications enabled by HR, as well as the ability to target DSBs to any chromosomal locus, provides unprecedented control over the genetic material in a variety of organisms.

The field of targeted genome modification using TALENs has advanced rapidly since the first TALENs were reported three years ago [45, 46]. Whereas the TAL effector repeat array mediates DNA recognition, sequences flanking the array (the so-called TALEN backbone) also influence TALEN activity in vivo. Although targeted chromosome breaks can be achieved using TALEN backbones that include most or all of the N- and C-terminal regions flanking the DNA binding domain [45, 46], several groups reported enhanced activity for TALENs with truncations of the N- and C-termini [13, 14, 50]. For example, truncated backbone architectures in plants result in more than a 25-fold increase in mutagenesis of endogenous targets [51]. We speculate that removal of N- and C-terminal sequences stabilizes the TALEN protein or facilitates folding.

In the short time since the first TALENs were reported, they have proven powerful reagents for reverse genetics in multiple experimental systems. Although not comprehensive, the pantheon of experimental models whose genome has been manipulated by TALENs already includes C. elegans [52], Drosophila [53], zebrafish [54], Xenopus [55], mice [56], rats [57], and various plant [46, 58, 59] and livestock species [60]. TALEN-mediated genome modification has been accomplished in human differentiated cell lines [13] as well as embryonic stem and induced pluripotent stem cells [61], and they are rapidly being deployed to ameliorate genetic diseases through gene therapy and to help solve challenges in agriculture [34, 50, 60, 62-64].

Although TALENs have quickly become the genome editing tool of choice, another potentially powerful approach is the use of TAL effector targeted recombinases. Recombinases (e.g. Cre) mediate recombination between specific DNA recognition sites (e.g. LoxP) and enable a diverse array of genome modifications including deletions, insertions and inversions. TAL effector recombinases (“TALER”s) have recently been engineered that recognize novel sites in genomes and thereby obviate the need to integrate recognition sites such as LoxP [65].
Despite the growing popularity of TAL effector based tools for genome editing, particularly TALENs, a newly developed platform based on the bacterial CRISPR/Cas system for defense against phage may have significant relative advantages. This system relies on a guide RNA to direct the Cas9 nuclease to cleave a target DNA sequence [66]. Where TALENs function as dimers, requiring construction and delivery of two large proteins, Cas9 functions as a monomer and requires only a single RNA of about 75 bases to direct cleavage. Multiple targets can be encoded into longer RNAs [67]. The CRISPR/Cas9 system has been used to mediate HR and NHEJ in human and mouse cells, with activity comparable to that of TALENs [67, 68]. Targeting specificity of the CRISP/Cas system, which depends on RNA-DNA hybridization, has not yet been examined in depth however.

**Design and construction**

Several online tools that aid users in designing custom TAL effectors and TALENs are freely available, including ZiFiT (http://zifit.partners.org/ZiFiT/Introduction.aspx), Mojo Hand (http://www.talendesign.org/mojohand_main.php), and TALE-NT 2.0 (https://tale-nt.cac.cornell.edu) [32, 69, 70]. For construction, modular assembly methods that rely on Golden Gate cloning have been developed that enable individual researchers to make dozens of constructs in a few days [31, 38, 59, 71, 72]. Reagents for many of these methods are available as kits from the non-profit plasmid repository Addgene (www.addgene.org). Solid state synthesis methods that allow high throughput, automatable assembly of custom TAL effector constructs have also been developed [73-75]. Custom TAL effectors and TALENs are available commercially as well.

**Future challenges**

Although custom TAL effectors and TALENs have proven to be powerful tools for a variety of applications, challenges and questions remain. Custom TAL effector-based constructs function with varying degrees of efficiency. Despite the modularity of the repeats, as noted above, RVDs differ in their apparent affinities for their preferred nucleotides. Thus, variation may simply relate to RVD (and target site) composition. Effects of binding site context and chromatin status are still largely unexplored however, and may also play a significant role. Better understanding of these effects combined with the identification of any high affinity RVDs or repeat types that could be used to replace those with lower affinities will be necessary to maximize and standardize the efficiency of TAL effectors and TALENs.

Specificity is also important, and sometimes critical. In nature, all known TAL effector and target alignments contain positions where an RVD associates with a non-preferred nucleotide,
and most contain one or a few RVDs that lack specificity. This degeneracy may be the product of selection for a binding mechanism that can accommodate minor genetic changes in the host before adaptation through effector recombination restores optimal binding. Whether this degeneracy is based in evolution or simple biochemistry, it presents a challenge for designing custom TAL effector-based DNA targeting proteins with perfect specificity, even when using the most specific RVDs. There is some evidence that TAL effector tolerance for mismatches varies depending on the type, position, and context of the mismatch, but this tolerance has not been studied systematically, making it difficult to predict off-targeting. Activity at off-target sites has been documented for both custom TAL effectors and TALENs [14, 71, 76]. Overall, TAL effectors appear to target with very good specificity [reviewed in 7]. But for applications such as human disease therapy, absolute specificity is paramount [63]. Better understanding of mismatch tolerance, and continued identification of any additional RVDs that can be used to replace less stringent ones, such as NH in place of NN for guanine, will aid in designing highly specific constructs [9, 41].

Because of their large size and repeat structure, delivery and integrity of TAL effector proteins or expression constructs might present a challenge, depending on cell type or method. For viral delivery of TAL expression constructs to mammalian cells, for example, size may be limiting. The repeats might also render constructs unstable. Advances in protein and nucleic acid delivery are an important technical frontier for further improving the utility of TAL effector-based DNA targeting tools. For TALENs, the use of a monomeric nuclease domain would decrease complexity [77]. Where possible, incorporating a nuclease domain with inherent specific DNA binding properties might also reduce the number of repeats needed, allowing smaller constructs.

Another future challenge is the continued development of new TAL effector-based tools (Figure 4). Closest on the horizon are probably TAL effector targeted DNA methylases (or demethylases) and histone modifiers for chromatin remodeling, and targeted cytosine deaminases or other mutagens for single base pair editing without recombination. One could also imagine using TAL effectors for DNA labeling or diagnostics, or even as organic linker molecules for DNA based materials [78]. Questions about native TAL effectors also remain. Determining their evolutionary origins and the selective forces acting on them may shed light on their uneven distribution across Xanthomonas species, and their limited distribution outside Xanthomonas. Probing the diversity and functions of TAL effector activated S and R genes is essential to expand our knowledge of disease and defense mechanisms, and to exploit that
knowledge for effective disease control. Advances in both fields of inquiry into natural TAL effectors will undoubtedly also further improve our understanding of TAL effector DNA interaction and our ability to continue to use that understanding in creative and robust ways to investigate and manipulate cell biology.

References


5 Mak, A.N.-S., et al. (2012) The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* 335, 716-719


8 Web of Knowledge, February 3, 2013. Thomson Reuters


18 Wichmann, G. and Bergelson, J. (2004) Effector genes of *Xanthomonas axonopodis* pv. vesicatoria promote transmission and enhance other fitness traits in the field. *Genetics* 166, 693-706


50 Sun, N., et al. (2012) Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. Molecular BioSystems 8, 1255-1263


70 Neff, K.L., et al. (2013) Mojo hand, a TALEN design tool for genome editing applications. *BMC Bioinformatics* 14, 1


75 Briggs, A.W., et al. (2012) Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. *Nucleic Acids Res.* 40, e117


CHAPTER 3. EFFICIENT DESIGN AND ASSEMBLY OF CUSTOM TALEN AND OTHER TAL EFFECTOR-BASED CONSTRUCTS FOR DNA TARGETING

A paper published in *Nucleic Acids Research*¹

Tomas Cermak²,³, Erin L. Doyle²,⁴, Michelle Christian², Li Wang³, Yong Zhang²,⁵, Clarice Schmidt³, Joshua A. Baller²,⁶, Nikunj V. Somia¹, Adam J. Bogdanove²,⁷, and Daniel F. Voytas¹,⁷

Abstract

TALENs are important new tools for genome engineering. Fusions of transcription activator-like (TAL) effectors of plant pathogenic *Xanthomonas* spp. to the FokI nuclease, TALENs bind and cleave DNA in pairs. Binding specificity is determined by customizable arrays of polymorphic amino acid repeats in the TAL effectors. We present a method and reagents for efficiently assembling TALEN constructs with custom repeat arrays. We also describe design guidelines based on naturally occurring TAL effectors and their binding sites. Using software that applies these guidelines, in nine genes from plants, animals, and protists, we found candidate cleavage sites on average every 35 bp. Each of 15 sites selected from this set was cleaved in a yeast-based assay with TALEN pairs constructed with our reagents. We used two of the TALEN pairs to mutate *HPRT1* in human cells and *ADH1* in *Arabidopsis thaliana* protoplasts. Our reagents include a plasmid construct for making custom TAL effectors and one for TAL effector fusions to additional proteins of interest. Using the former, we constructed *de novo* a functional analog of AvrHah1 of *X. gardneri*. The complete plasmid set is available through the non-profit repository AddGene, and a web-based version of our software is freely accessible online.

---

¹ Reprinted with permission from *Nucleic Acids Res*, 2011, 39(12), e82. Published by Oxford University Press. Copyright The Authors.
² The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
³ Department of Genetics, Cell Biology & Development and Center for Genome Engineering, 321 Church Street SE, University of Minnesota, Minneapolis, MN 55455, USA
⁴ Department of Plant Pathology, 351 Bessey Hall, Iowa State University, Ames, IA 50011, USA
⁵ Department of Biotechnology, School of Life Sciences and Technology, University of Electronic Science and Technology of China, Chendu 610054, China
⁶ Biomedical Informatics and Computational Biology Program, University of Minnesota Rochester, 111 South Broadway, Rochester, MN 55904, USA
⁷ To whom correspondence should be addressed.
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 (1). 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 1a). The majority of naturally occurring TAL effectors examined have between 12 and 27 full repeats (2). 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 1a) (3,4). Also, naturally occurring recognition sites are uniformly preceded by a T that is required for TAL effector activity (3,4). These straightforward sequence relationships allow the prediction of TAL effector binding sites (3-6) and construction of TAL effector responsive promoter elements (7), as well as customization of TAL effector repeat domains to bind DNA sequences of interest (8-11).

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 *FokI* nuclease to create targeted DNA double strand breaks (DSBs) *in vivo* for genome editing (8,10,12,13). Since *FokI* cleaves as a dimer, these TAL effector nucleases (TALENs; 8) function in pairs, binding opposing targets across a spacer over which the *FokI* domains come together to create the break (Figure 1b). 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 (14,15). 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
Figure 1. TAL effector and TALEN structure. (a) Structure of a naturally occurring TAL effector. A consensus repeat sequence is shown with the repeat-variable di-residue (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 displayed) has greater specificity for G than NN does and for that reason our plasmid set also includes NK modules. (b) Structure of a TALEN. Two monomeric TALENs are required to bind the target site to enable $FokI$ to dimerize and cleave DNA. NLS, nuclear localization signal(s); AD, transcriptional activation domain; B, BamHI; S, SphI.

intensive. And, despite a significant recent advance for ZFNs that takes finger context into account to achieve high success rates (16), targeting capacity (the diversity of sequences that can be recognized) still suffers limitations (17-19). TALENs thus far appear not to be subject to these constraints. 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 (10)

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 (9,11). 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 (8). 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 (20,21). The Golden Gate method uses Type IIS restriction endonucleases, which cleave outside their recognition sites to create unique 4 bp overhangs (sticky ends) (Figure 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,

Figure 2. Golden Gate assembly of custom TAL effector and TALEN constructs using module, array, last repeat and backbone plasmids. By using the type IIS restriction endonucleases BsaI and Esp3I, 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, BamHI, and S, SphI, 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 but can be found in the sequence files (Supplementary Data).

1 The marker in the Last Repeat plasmid has been changed to spec, according to the correction published in Nucleic Acids Res., 2011. 39(17), 7879.
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 Xanthomonas gardneri (22).

**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) and involves two steps: (i) assembly of repeat modules into intermediary arrays of 1-10 repeats, and (ii) joining of the intermediary arrays into a backbone to make the final construct. A schematic representation is shown in Figure 2 and the complete set of required plasmids is displayed in Supplementary Figure S1. Construction and features of the plasmids themselves are described in the following section. 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.

<table>
<thead>
<tr>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target and design</td>
<td>Pick and culture 3 white colonies each</td>
<td>DNA prep &amp; verify by digest 3 white colonies each</td>
<td>Pick and culture 3 white colonies each</td>
<td>DNA prep &amp; verify by digest (or sequencing ) Constructs ready to test in yeast (TALENs) or Xanthomonas (TAL effectors) or to subclone into vector of choice</td>
</tr>
<tr>
<td>Perform Golden Gate reaction 1 to build arrays of 10 and 1-10 repeats</td>
<td>Perform Golden Gate reaction 2 to join arrays in a backbone</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 3. TALEN or TAL effector construct assembly timeline.*
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'-ACCGCAATCTCAATCT-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 (3,4), 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 BsaI (10U, New England BioLabs), and 1 ml T4 DNA Ligase (2000U, 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, 1ml 25 mM ATP and 1 ml Plasmid Safe DNase (10U, Epicentre) are added. The mixture is incubated at 37°C for 1 hr, then used to transform Escherichia 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 (23). Treatment with Plasmid Safe DNase is an important step to prevent linear DNA fragments, including partial repeat sequences at the termini of the 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

1 This sentence has been corrected to reflect the correction published in Nucleic Acids Res., 2011. 39(17), 7879.
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, 1052 bp for pFUS_A30A, 1040 bp for 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 μl 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 μg/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 *Stul*) 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 (3), 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 Apal and XbaI 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 AflII and XbaI, synthesized fragments into pCR8 that contain two internal Bsal 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 (23) was cloned between the two Bsal sites. For this, the multiple cloning site between the HincII and Eco53kl sites in phagemid pBCSK+ (Stratagene) was deleted and the lacZ gene PCR amplified with primers carrying KasI and AgeI overhangs. These sites were included in the synthesized fragments, allowing the lacZ gene to be placed between the Bsal sites, maintaining the overhang sequences for accepting modules. The inserts in the array plasmids all contain terminal Esp3I (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 AflII and XbaI 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 (8). Derived from pDW1789 (24), pFZ85 contains the counterselectable ccdB gene flanked by BamHI 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 FokI 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 SphI fragment of tal1c was replaced with the SphI fragment of TAL effector gene pthXo1 (25), 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*AI 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 (8), 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 *Afel* site in pDW1789. Then, pDW1789 was linearized with *Bpu*10I and *Afel* (removing the *HIS3* gene) and the PCR amplified *LEU2* gene was inserted by *in vivo* recombination in *E. coli* (26). 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 modules, last repeat, array, and backbone plasmids (Supplementary Figure S1) and a folder containing complete sequences (Supplementary Data) are included in Supplementary Data.

**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 (Supplementary Figure S2a), 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 (8), 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 (Supplementary Figure S2b). 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 (3,4), 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 (8,24) 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 (Supplementary Figure S3a). 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 (α 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 BglII- and SpeI-compatible ends, and cloned between the lacZ fragments in the high copy DNA cleavage reporter plasmid pCP5 (24) cut with those enzymes (Supplementary Figure S3b). The target plasmid is transformed into yeast strain YPH499 (α 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 µl 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 centrifugation, washed in 1 ml synthetic complete medium lacking histidine and/or leucine and tryptophan, but now containing uracil, then resuspended 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 µl of lysate is transferred to a microtiter well plate and β-galactosidase activity measured and normalized as previously described (24). 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 (24).

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 XhoI and AflII. 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 Hpy188I, 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 Hpy188I 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 (27) using XbaI and SacI. 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 (27). Forty-eight hours after transformation, DNA was prepared and digested with PfI, 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 PfI 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 \textit{Xanthomonas} and \textit{in planta} activity assay

An analog of \textit{avrHah1} was assembled into \textit{pTAL1} using the Golden Gate method with HD, NI, NG, and NN modules, ordered to match the \textit{AvrHah1} binding site in the promoter of the Bs3 gene (22). A native \textit{avrHah1} construct was made by replacing the \textit{BamH}I fragment of \textit{tal1c} in \textit{pCS495} with that of \textit{avrHah1}. \textit{pCS495} is \textit{tal1c} preceded by Shine-Dalgarno and Kozak consensus sequences in \textit{pENTR-D} (Invitrogen). The analog and native \textit{avrHah1} constructs and \textit{tal1c} were moved into \textit{pKEB31} by Gateway cloning (LR reaction). \textit{pKEB31} is a derivative of \textit{pDD62} (28) that contains a Gateway destination vector cassette (Invitrogen) between the \textit{Xba}I and \textit{BamH}I sites and a tetracycline resistance gene in place of the gene for gentamycin resistance. The resulting plasmids were introduced into \textit{X. campestris pv. vesicatoria} strain 85-10 by electroporation and transformants were inoculated to 6-week-old pepper plants by syringe infiltration, as described (22). 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 2 and Supplementary Figure S1). 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 (9,10). 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 (i) pTAL1 for assembling a custom TAL effector gene preceded by Shine-Dalgarno and Kozak sequences for efficient translation in bacteria and eukaryotes, respectively, (ii) pTAL2, identical to pTAL1, but without a stop codon so that the effector can be fused to other protein domains, (iii) pTAL3 for assembling a custom TALEN and expressing it in yeast using the selectable marker HIS and (iv) 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 SphI sites flanking the repeat domain as well as the BamHI 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 (23). For the work presented here, we successfully assembled > 30 custom TALENs (Supplementary Table S1) 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 5 days (Figure 3 and refer “Materials and Methods” section).

**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” section). 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 (8). We established the targeting guidelines by examining the 20 TAL effector-target pairs identified by Moscou and Bogdanove (3). 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 4). 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 (3') and next-to-last 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 (3) 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 (data not shown). The guidelines are therefore as follows: (i) As noted previously for TAL effector binding sites (3,4), TALEN monomer binding sites should be preceded by a 5' T, (ii) they should not have a T at position 1, (iii) they should not have an A at position 2, (iv) they should end with a T, so that the corresponding TALENs will reflect the strong bias for NG at this position and (v) 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 (Supplementary Table S2). Of the four intermediate length TALEN-target pairs showing no detectable activity in the yeast assay for DNA cleavage (8), 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 met all of the guidelines. The impact of the number of repeats in a TALEN was also considered. In general, longer TALENs that met all of the guidelines or medium-length TALENs that met all 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 met all 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 15 target sites (30 TALENs total; Supplementary Table S1) were made using the Golden Gate method and plasmids described above and tested in the yeast based DNA cleavage assay we described previously (8). All TALEN pairs showed significant activity above the target-only negative controls, and 14 of 15 showed activity ≥25% of our positive control, a Zif268 ZFN (Figure 5). We have generally found for ZFNs that this level of activity is sufficient for targeted mutagenesis of endogenous plant loci (24,27).

**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 5) and the TALEN pair for the *Arabidopsis ADH1* gene 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 6). 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, 6 independent mutations consisting of deletions ranging from 4 to 15 bp, also centered on the spacer.
Figure 5. 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.

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 (22). 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 7). 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.
Figure 6. 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 Figure 5) and the Arabidopsis ADH1 gene (Figure 5) 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 (a) and ADH1 in (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.

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 (10) but costly. PCR based methods (11) carry the risk of artifact and recombination. Assembly by sequential ligation of sequence-verified modules (8) is inexpensive and assures array integrity but is time consuming. The Golden Gate
Figure 7. 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 h 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*.

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 *et al.* (11) 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 (1). This could be accomplished, e.g. 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. (11) 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 2 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. (Figure 2e in ref. 10), 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 >120 bp away, a distance that would likely compromise gene targeting efficiency (18). 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 (24). 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 (29). 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 (16,18). 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 (10; Supplementary Table S2).
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 (10,13), 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.

Acknowledgements

The authors thank Marit-Nilsen Hamilton and Lee Bendickson for assistance with mammalian cell culture, Divya Mistry for assistance with development of the TALE-NT website, and Jeff Jones for providing a native avrHah1 clone.

Funding

The National Science Foundation (DBI 0923827 and MCB 0209818 to D.V., DBI 0820831 to A.B.); the University of Minnesota; the China Scholarship Council (2009104157 to Y.Z.); the National Natural Science Foundation of China (30900779 to Y.Z.). Funding for open access charge: The National Science Foundation (DBI 0923827 to D.V.).

Conflict of interest statement. None declared.
References


### Supplementary Material

**Supplementary Table S1. Custom TALENs and target sites.**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Organism</th>
<th>TAL1, TAL2 Plasmids</th>
<th>TAL1, TAL2 Repeat number</th>
<th>Spacer</th>
<th>TAL1, TAL2 RVD Sequences with Spacer (TAL1, TAL2 underlined)</th>
</tr>
</thead>
</table>
| HPRT1 A | Human | pTAL141N, pTAL142N | 17,20 | 15 | TAL1:  NN HD NI NG NI HD HD NG NI NG NI NG NI NG NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI NG NI
<table>
<thead>
<tr>
<th>Gene</th>
<th>Organism</th>
<th>TAL1,TAL2 Plasmids</th>
<th>TAL1,TAL2 Repeat number</th>
<th>Spacer</th>
<th>TAL1, TAL2 RVD Sequences with Spacer (TAL1, TAL2 underlined)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gridlock</td>
<td>Zebrafish</td>
<td>pTAL112, pTAL117</td>
<td>17,16</td>
<td>18</td>
<td>5'-ATTTCGAGCGCTATTCC-TC GCATCGATGACCTCCTTGG-5'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3'-GTGCCCATGACCTTGGAGC-3'</td>
</tr>
<tr>
<td>white</td>
<td>Drosophila</td>
<td>pTAL152N, pTAL153N</td>
<td>22,24</td>
<td>24</td>
<td>5'-CCGTCGTCTGACGCTTGGAGC-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3'-CTCCTGACATTTCAG-3'</td>
</tr>
<tr>
<td>GFP</td>
<td>Jelly Fish</td>
<td>pTAL156N, pTAL157N</td>
<td>26,25</td>
<td>25</td>
<td>5'-CCATGACCCAACACTTG-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3'-TACACTACCTTGTCTCAG-3'</td>
</tr>
<tr>
<td>eGFP</td>
<td>Jelly Fish</td>
<td>pTAL158N, pTAL159N</td>
<td>24,26</td>
<td>21</td>
<td>5'-AACACTAACAGCCCA-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3'-CAGAGAGCATACATA-3'</td>
</tr>
<tr>
<td>Plasmepsin A</td>
<td>Plasmodium</td>
<td>pTAL160N, pTAL161N</td>
<td>19,22</td>
<td>15</td>
<td>5'-GGGAGGCAGC-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3'-GAAAAGAGGAGGACAAG-3'</td>
</tr>
<tr>
<td>Plasmepsin B</td>
<td>Plasmodium</td>
<td>pTAL162N, pTAL163N</td>
<td>23,24</td>
<td>15</td>
<td>5'-CCGCGCCCTCATCTCC-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3'-AGTGACCCATCATCGGAC-3'</td>
</tr>
<tr>
<td>TT4 A</td>
<td>Arabidopsis</td>
<td>pTAL164N, pTAL165N</td>
<td>19,22</td>
<td>15</td>
<td>5'-CCGCGGCTCATCTCC-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3'-AGTGACCCATCATCGGAC-3'</td>
</tr>
<tr>
<td>TT4 B</td>
<td>Arabidopsis</td>
<td>pTAL166N, pTAL167N</td>
<td>15,16</td>
<td>15</td>
<td>5'-GGGAGGCAGC-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3'-GAAAAGAGGAGGACAAG-3'</td>
</tr>
<tr>
<td>ADH1</td>
<td>Arabidopsis</td>
<td>pTAL69N, pTAL74N</td>
<td>15,15</td>
<td>18</td>
<td>5'-CCGCGGCTCATCTCC-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3'-AGTGACCCATCATCGGAC-3'</td>
</tr>
<tr>
<td>SuRB A</td>
<td>Tobacco</td>
<td>pZHY096, pZHY106</td>
<td>17,20</td>
<td>15</td>
<td>5'-ACTCTACTTCTCAAAAGATGTT-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3'-CCGCGGCTCATCTCC-3'</td>
</tr>
<tr>
<td>SuRB B</td>
<td>Tobacco</td>
<td>pZHY108, pZHY109</td>
<td>19,19</td>
<td>15</td>
<td>5'-ACTCTACTTCTCAAAAGATGTT-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3'-CCGCGGCTCATCTCC-3'</td>
</tr>
</tbody>
</table>
### Supplementary Table S2. Activity, conformity to rules, and length of TALENs tested in the yeast assay.

<table>
<thead>
<tr>
<th>TALEN</th>
<th>RVDS</th>
<th>Activity</th>
<th>% GATC</th>
<th>Ends in NG</th>
<th>RVD sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADH1_1R12</td>
<td>12</td>
<td>-</td>
<td>y</td>
<td>n</td>
<td>HD HD HD NI NN NI NN NG NI NI NI</td>
</tr>
<tr>
<td>ADH1_2R12</td>
<td>12</td>
<td>-</td>
<td>n</td>
<td>n</td>
<td>NI NN NI HD NI NI HD HD NI HD NI</td>
</tr>
<tr>
<td>telo_1L15</td>
<td>15</td>
<td>-</td>
<td>y</td>
<td>y</td>
<td>NN NN HD NI HD NG NN HD NG NI NI HD HD NG</td>
</tr>
<tr>
<td>ADH1_R15</td>
<td>15</td>
<td>+</td>
<td>y</td>
<td>y</td>
<td>HD HD HD NI NN NI NI NN NG NI NI HD NI NG</td>
</tr>
<tr>
<td>ADH1_2L12</td>
<td>12</td>
<td>+</td>
<td>y</td>
<td>n</td>
<td>HD HD NN NI NN NG NN HD HD NG HD NG</td>
</tr>
<tr>
<td>grid_1R10</td>
<td>10</td>
<td>+</td>
<td>n</td>
<td>y</td>
<td>NI HD HD HD HD NG NG NG HD HD</td>
</tr>
<tr>
<td>telo_1L15</td>
<td>15</td>
<td>+</td>
<td>y</td>
<td>y</td>
<td>HD NG NG NN NG HD HD NN HD NI NG NN NI NG NG</td>
</tr>
<tr>
<td>telo_2L16</td>
<td>16</td>
<td>+</td>
<td>y</td>
<td>y</td>
<td>NI NG NG HD HD HD HD NI HD NN NI NN HD NG HD NG</td>
</tr>
<tr>
<td>telo_2R17</td>
<td>17</td>
<td>+</td>
<td>n</td>
<td>y</td>
<td>NI NN NI HD NI NN NI NI NI NN NG NN NI NN NI NN HD NG</td>
</tr>
<tr>
<td>ADH1_1L12</td>
<td>12</td>
<td>++</td>
<td>y</td>
<td>n</td>
<td>NI NG HD NI NI NN NI NG NG HD NG HD</td>
</tr>
<tr>
<td>ADH1_2R15</td>
<td>15</td>
<td>++</td>
<td>y</td>
<td>y</td>
<td>NI NN NI HD NI NI HD HD NI NI HD NI HD NG</td>
</tr>
<tr>
<td>grid_1R13</td>
<td>13</td>
<td>++</td>
<td>y</td>
<td>y</td>
<td>NI HD HD HD HD NG HD NG HD HD NN HD NG</td>
</tr>
<tr>
<td>ADH1_2L15</td>
<td>15</td>
<td>+++</td>
<td>y</td>
<td>y</td>
<td>HD HD NN NI NN NG NN HD NG HD NG HD NG</td>
</tr>
<tr>
<td>grid_1L16</td>
<td>16</td>
<td>++++</td>
<td>y</td>
<td>y</td>
<td>NI HD HD HD NG HD NG HD NG HD NG</td>
</tr>
</tbody>
</table>

1. – no activity, + activity less than 25% of ZFN Zif268, ++ activity 25-50% of ZFN Zif268, +++ activity 50-75% ZFN Zif268, ++++ activity >75% ZFN Zif268.
Supplementary Figure S1. Module, array, repeat, and backbone plasmids. Overhangs left by BsaI digestion are shown in blue font and by Esp3I in green. HD modules 2-10 contain a BspEI site (not shown) not present in 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, BamHI, and S, SphI. Unique restriction enzyme sites flanking the entire coding sequences of the backbone constructs are not shown but can be found in the sequence files.
Supplementary Figure S2. The 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 (Figure 5) target site information. The targeter URL is boglabx.plp.iastate.edu/TALENT/.
Supplementary Figure S3. 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.
CHAPTER 4. TAL EFFECTOR-NUCLEOTIDE TARGETER (TALE-NT) 2.0: TOOLS FOR TAL EFFECTOR DESIGN AND TARGET PREDICTION

A paper published in *Nucleic Acids Research*¹

Erin L. Doyle², Nicholas J. Booher³, Daniel S. Standage⁴, Daniel F. Voytas⁵, Volker P. Brendel⁶,⁷, John K. VanDyk³ and Adam J. Bogdanove²,⁸

Abstract

Transcription activator-like (TAL) effectors are repeat-containing proteins used by plant pathogenic bacteria to manipulate host gene expression. Repeats are polymorphic and individually specify single nucleotides in the DNA target, with some degeneracy. A TAL effector-nucleotide binding code that links repeat type to specified nucleotide enables prediction of genomic binding sites for TAL effectors and customization of TAL effectors for use in DNA targeting, in particular as custom transcription factors for engineered gene regulation and as site-specific nucleases for genome editing. We have developed a suite of web-based tools called TAL Effector-Nucleotide Targeter 2.0 (TALE-NT 2.0; https://boglab.plp.iastate.edu/) that enables design of custom TAL effector repeat arrays for desired targets and prediction of TAL effector binding sites, ranked by likelihood, in a genome, promoterome or other sequence of interest. Search parameters can be set by the user to work with any TAL effector or TAL effector nuclease architecture. Applications range from designing highly specific DNA targeting tools and identifying potential off-target sites to predicting effector targets important in plant disease.

Introduction

Transcription activator-like (TAL) effectors from the plant pathogenic bacterial genus *Xanthomonas* represent a class of DNA binding proteins that can be readily engineered to

---

¹ Reprinted with permission from *Nucleic Acids Res*, 2012, 40(W1), W117-W122. Published by Oxford University Press. Copyright The Authors.
² Department of Plant Pathology and Microbiology, Iowa State University, Ames, IA 50011, USA
³ Department of Entomology, Iowa State University, Ames, IA 50011, USA
⁴ Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA
⁵ Department of Genetics, Cell Biology and Development and Center for Genome Engineering, University of Minnesota, Minneapolis, MN 55455, USA
⁶ Department of Biology, Indiana University, Bloomington, IN 47405, USA
⁷ School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA
⁸ To whom correspondence should be addressed.
target novel DNA sequences. During infection, TAL effectors are deployed by the bacteria to modulate host gene expression, with each effector directly binding an effector-specific DNA target (1,2). A central region in the protein, composed of a variable number of tandem, near-identical, 33-35 amino acid repeats, determines the target(s) of each TAL effector. Repeat-to-repeat variation occurs primarily at residues 12 and 13 (termed the repeat variable diresidue or RVD). The RVD sequence has been shown both computationally and experimentally to correspond directly to the DNA target site sequence; repeats with different RVDs recognize different nucleotides, in a code-like fashion, with some degeneracy (3,4). Custom TAL effectors can be targeted to novel DNA sequences by assembling an array of repeats that corresponds to the intended target site (5). Designing custom TAL effectors for DNA targeting has proved to be a much simpler and less labor-intensive process than the design of other customizable DNA binding proteins such as zinc fingers (6), and a variety of rapid construction methods for custom TAL effectors and TAL effector fusion proteins have recently been developed (7-12).

Increasingly therefore, TAL effectors and TAL effector-based fusion proteins have been adopted as tools for DNA targeting applications (6).

Site-specific DNA modification has been achieved using TAL effector-endonuclease fusion proteins (TAL effector nucleases or TALENs), which create targeted double-stranded breaks (DSBs) in DNA. TALEN architectures to date combine the central repeat region and some portion of the flanking parts of the TAL effector with the catalytic domain of FokI, which functions as a dimer. Thus, TALENs work in heteromeric pairs, with the monomers binding to opposing target sites oriented 5’ to 3’ on opposite strands across a spacer, allowing the C-terminal FokI domains to dimerize and create a DSB in that spacer (5,13-15). In eukaryotes, such breaks are repaired by either non-homologous end joining (NHEJ), which is prone to errors and useful for creating gene knockouts, or homologous recombination (HR), which replaces the sequences surrounding the break with a supplied template. TALEN-mediated NHEJ and HR have been demonstrated in a variety of cell types and organisms (7,8,13,14,16-20). TAL effectors have also been used as custom transcriptional activators, most effectively in plants in their native form and in mammalian cells with the native acidic activation domain replaced by the VP16 (or tetrameric derivative VP64) activation domain from Herpes Simplex Virus (10,21,22). The TAL effector targeting domain has also been fused to a repressor protein and used successfully for specific gene repression in plants (23).

Custom TAL effector and TALEN architectures using different proportions of the flanking regions on either side of the repeat region have been reported. In the case of TALENs, the optimum spacer length between the TALEN monomers depends on the architecture used...
Additionally, different methods assemble different numbers of repeats in an array, with some allowing a wide range (7-12). Therefore, tools for designing custom TAL effectors and TALENs should allow for a range or a prescribed number of repeats and, for TALENs, various spacer lengths. Most TAL effector targets in nature are preceded by a T at the 5’-end (3), but at least one example of a TAL effector target preceded by a C has been identified (25), and some custom TAL effectors have been reported to be active on sites preceded by a C (13). The preference for T at this position is structurally specified by the region of the protein immediately N-terminal to the central repeat region (26,27). Accounting for the “-1 position” nucleotide is important for target prediction and custom TAL effector design as well.

Because of the degeneracy in the TAL effector–DNA recognition code, off-targeting is a concern in the use of TAL effectors and TALENs. RVDs associate preferentially, but not exclusively, with specific nucleotides. TAL effector-target pairs in nature are observed to have up to 50% mismatches (positions at which the RVD is not aligned with its most frequently associated nucleotide) (3), and TAL effector-based custom transcription factors have been shown to activate transcription of off-target genes whose promoters contain sequences similar but not identical to the intended targets (21). If custom TAL effectors and TALENs are to be a widely adopted biotechnology, accurate prediction of potential off-target activity and design to minimize such activity will be crucial. Prediction of potential TAL effector binding sites that takes the degeneracy into account is also important for the identification of disease targets in plant hosts of _Xanthomonas_ spp.

To aid researchers in adopting TALENs as gene editing tools, we converted a computational script we had created for TALEN design, called “TALEN Targeter”, to a web-based tool and posted it on a site we called TAL Effector-Nucleotide Targeter (TALE-NT 1.0). The tool designed TALEN pairs according to five design guidelines based on positional and composition biases observed in known TAL effector-target pairs (7). Herein, we describe a new version of the web site, TALE-NT 2.0, that offers a suite of tools for TALEN and TAL effector design as well as target prediction. Since the publication of TALE-NT 1.0, it was shown that the design guidelines have little effect on TALEN efficiency (12). Therefore, for TALE-NT 2.0, we removed the guidelines from TALEN Targeter and instead allow users to search for TALENs targeting a specific base. We also updated TALEN Targeter to provide users with additional options for entering sequences, customizing their queries to accommodate different TALEN architectures and other preferences, searching for sites preceded by T or C and viewing output. We added two new tools: “TAL Effector Targeter”, for designing custom TAL effectors for single sites, and “Target Finder”, for predicting candidate TAL effector targets, ranked by likelihood, in the
genomes or gene promoteromes of several model organisms or in a sequence supplied by the user. The new web site (https://boglab.plp.iastate.edu/) makes this suite of tools freely available to the research community. Applications range from designing highly specific DNA targeting reagents and identifying potential off-target sites to predicting effector targets important in plant disease.

**Software and Algorithms**

**Programming**

TALENT 2.0 tools other than Target Finder are written in the Python programming language and use Biopython libraries for parsing input DNA sequences and other sequence operations (28). Target Finder is written in C and uses Kseq.h (http://lh3lh3.users.sourceforge.net/parsefastq.shtml) for sequence parsing.

**Tools for designing custom TALENs and TAL effectors**

TALEN Targeter and TAL Effector Targeter design paired and single custom TAL effector repeat arrays, respectively, for targeting DNA sequences of interest. Both tools require one or more FASTA-formatted DNA sequences as input. TALEN Targeter identifies paired monomer (typically heteromeric) binding sites oriented 5' to 3' on opposite strands of the DNA and separated by a spacer. TAL Effector Targeter identifies single TAL effector binding sites; users have the option of searching the reverse complement in addition to searching the DNA sequence as entered. In either tool the number of TAL effector repeats can vary across a user-specified range. TALEN Targeter allows users also to specify a range for the spacer length; all possible combinations of repeat numbers and spacer lengths are considered. Identified binding sites are converted into RVD sequences using the four most common RVD-nucleotide pairs (NI-A, HD-C, NN-G and NG-T). Target sites, RVD sequences, and other information are returned to the user (the output for all tools is described more fully in the “Web Interface” section). Arrays designed with TAL Effector Targeter are by default designed to meet the five guidelines developed based on TAL effectors observed in nature, as the effect of these guidelines on TAL effector transcription factors has not been determined. Users may choose not to enforce one or more guidelines.

**A tool for identifying candidate TAL effector targets**

Target Finder allows users to enter an RVD sequence and search for candidate targets in the genome or gene promoterome of any of several model organisms or in one or more user-provided, FASTA-formatted DNA sequences. The tool may be used to identify and rank
candidate plant genomic targets of TAL effectors important for disease, or potential off-target binding sites of custom TAL effector proteins.

To predict and rank sites, the tool uses a simple scoring function developed by Moscou and Bogdanove (3) based on RVD-nucleotide association frequencies found in a set of known TAL effector-target pairs. We have used this scoring function to predict 21 previously unknown and subsequently experimentally verified plant targets for 14 TAL effectors (A. Cernadas, E. Doyle and A. Bogdanove, unpublished results). Briefly, for scoring, for each RVD in the set of TAL effector-target pairs, the frequency with which it pairs with each nucleotide was calculated and then converted to a weighted RVD-nucleotide association probability. For RVDs that are not observed in the set of pairs, each nucleotide association was given an equal probability. The score for a DNA/RVD sequence alignment is found by summing the negative logs of the appropriate RVD-nucleotide association probabilities, such that better alignments have lower scores. A detailed description of the scoring function, including weighted RVD-nucleotide association probabilities, is provided in the Supplementary Material.

Target Finder returns a list of the lowest scoring (best) sites in the queried DNA sequence below a cutoff (discussed in the Web Interface section). Candidate binding sites are not required to conform to any of the design guidelines except that they must be preceded by a T, or C if that option is selected. The tool by default searches for binding sites on both strands of the DNA sequence(s), but users may opt to search only the forward strand.

**Web Interface**

**Design and general features**

The TALE-NT 2.0 web site is powered by Drupal 7 on Red Hat Enterprise Linux 6. Job queuing is handled by Celery using Redis as a message broker. All features of the web site have been tested on common web browsers.

TALE-NT 2.0 makes the three tools for design of TAL effectors and prediction of TAL effector targets freely available to all users, with no log-in requirement. Upon submitting a job, users are taken to a bookmarkable page updating them on the status of their query or supplying a link that will take them to their results when the query has finished processing. Processing times are not excessive: a search of the entire rice genome for possible targets of an average length TAL effector (18 RVDs) took less than three minutes. Nevertheless, users have the option to enter an address to receive email notification when their job finishes. Results for each tool are displayed in a sortable table by default. Users may also or instead download results as
a tab-delimited text file. For Target Finder users can download either or both of two formats, standard or GFF3, described further below.

All tools allow users to design or search for TAL effector binding sites preceded by a 5’ T only, C only, or T or C. In our hands however, with either TALENs or TAL effectors, sites preceded by a C are significantly less active than those preceded by a T, so we suggest using T, which is the default selection.

**TALEN Targeter**

The TALEN Targeter web interface allows users to design custom TALENs to target one or more DNA sequences of interest. Users enter their sequence(s) in a text box or upload a file containing the sequence(s). Allowable file size is up to 2 MB. Users may select from four common TALEN architectures (7,13,15,24) with preselected ranges for spacer size and numbers of repeats or enter their own ranges for these parameters. Users may choose to allow binding sites to be preceded by a T only, C only, or by T or C. In addition, users have the option to output just TALEN pairs targeting a specific base, a filtered list including up to one TALEN pair for each base, or the complete list of all TALEN pairs targeting anywhere in the sequence. The targeted base is defined as the base is in the center (or immediately to the left of the center) of the spacer. If users choose to return up to one TALEN pair per base, the pair with the smallest average number of RVDs and shortest spacer targeting a given base will be returned.

Each line of the output describes a pair of monomers that will function together as a TALEN. Information provided includes the name of the input sequence (in case more than one sequence was entered), the starting position and number of repeats for each monomer, the target sequence including the -1 nucleotide and RVD sequences corresponding to each monomer.

The output is ordered by sequence name, with TALENs for each sequence grouped by the start position of the first monomer. The order of TALENs in the output does not relate to how well they might function. No scoring or other prioritization method is used. Users should select TALENs closest to the site of desired cleavage that have the fewest or poorest predicted off-target sites determined using Target Finder and further analysis. This analysis should take into account not only the heterodimer but both of the possible homodimers as well. Users may also wish to choose TALENs that straddle a spacer with a diagnostic restriction endonuclease site to facilitate detection of NHEJ-mediated mutation of the spacer sequence (such restriction endonuclease sites are indicated in the output table) or in the case of HR, a TALEN that uniquely binds the sequence to be replaced, so that following replacement it does not cut again. Users might opt to introduce silent mismatches in the HR template to facilitate this.
TAL Effector Targeter

TAL Effector Targeter allows users to design a single custom TAL effector repeat array to target a DNA sequence of interest. The array may be used for the design of a TAL effector or any TAL effector-based fusion protein that functions as a monomer. As with TALEN Targeter, users enter sequence(s) in a text box or upload a FASTA-formatted file. Additional text boxes allow users to specify a range for the number of repeats in each TAL effector (default is 15-30). Checkbox options allow users to turn off individual design guidelines. By default, the tool searches only the sequence(s) as entered; a checkbox option includes also the reverse complement sequence(s). Users may also choose to search only for sites preceded by T at the 5’ end, or to allow T and/or C.

Each line of the output describes a single custom TAL effector repeat array. Information returned includes the name of the input sequence (in case multiple sequences were entered), the strand and coordinates of the target, the target sequence including the -1 nucleotide, and corresponding RVD sequences.

The output is ordered first by sequence name, with target sites for each sequence sorted by their position in the sequence. As with TALEN Targeter, arrays are not ranked or prioritized. Users should choose according to their own targeting criteria. If there are no target sites in a sequence or no sites in the desired region, users may be able to increase the number of sites by relaxing some of the design rules and/or changing the range for number of repeats. We recommend that users leave the percent composition rule enforced to better assure good overall affinity. Although data regarding relative affinities of different RVDs for their partner nucleotide(s) is lacking, based on the published structures (26,27) we predict that the highest affinity interactions are those that occur most frequently in nature. So, if a user chooses not to enforce the percent composition guideline, at a minimum, arrays should be selected that have overall higher numbers of HDs and NGs than NIs and NNs. Also, in most contexts tested, NN pairs with G or A, so we anticipate that minimizing NN content will maximize specificity.

Target Finder

Target Finder uses the scoring function of Moscou and Bogdanove (3) to identify the best-scoring sites in a DNA sequence for a user-specified string of RVDs. Users input a TAL effector RVD sequence containing from 12 to 35 RVDs, each separated by a space, using single-letter amino acid abbreviations. All possible RVDs constructed using the standard 20 amino acids and ‘*’ to indicate a missing 13th amino acid are allowed. Users may select a genome or gene promoterome from a drop-down list, or provide their own sequence of interest in a text box or as
Promoterome is defined as the collection of sequences 1000 bp upstream of annotated translational start sites in a genome. Available genomes and promoteromes include those of rice \((\text{Oryza sativa})\), Arabidopsis thaliana, human \((\text{Homo sapiens})\), fruit fly \((\text{Drosophila melanogaster})\), mouse \((\text{Mus musculus})\), nematode \((\text{Caenorhabditis elegans})\), and zebra fish \((\text{Danio rario})\). Genome sequences were obtained from Ensembl (www.ensembl.org), except for the rice genome, which was downloaded from the MSU Rice Genome Annotation Project (http://rice.plantbiology.msu.edu, Version 6.1). Promoteromes were downloaded from the UCSC Genome Bioinformatics Site (http://genome.ucsc.edu/index.html), except for the rice promoterome, which was downloaded from the Rice Genome Annotation Project, and the Arabidopsis promoterome, which was downloaded from The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org/). By default, the tool searches both strands, but users may choose to search only the forward strand. Users may choose to search for sites preceded by a 5' T only, C only or T and C.

Target Finder returns all targets scoring under a threshold cutoff defined as a ratio of observed score to best possible score of 3.0 or less. The best possible score is the score for the array aligned to its code-specified, perfect match DNA sequence. Users may relax the cutoff to 4.0. The default threshold of 3.0 was selected because the naturally occurring TAL effector-target site pairs we analyzed to decipher the TAL effector DNA binding code and develop the scoring matrix (3) typically had scores <3.0 times the best possible score for the TAL effector. In addition, for binding sites of naturally occurring TAL effectors predicted using the scoring function, experimentally verified sites also had scores <3.0 times the best possible score (A. Cernadas, E.Doyle and A.Bogdanove, unpublished results\(^1\)). Users wishing to identify more sites should choose the less stringent cutoff. The best possible score for an array is included at the top of the output table or file. Information provided for each target in the output includes the coordinates, the DNA strand being targeted, the score for the target, and the target sequence. Targets in each DNA sequence are ordered from best (lowest scoring) to worst (highest scoring). For searches of genomes or promoteromes from the drop-down list, the output in the default displayed table includes a link for each target that configures a custom track to show the target location in a corresponding genome browser when available. The number of targets displayed in the table is set by default to a maximum of 10, though users can specify a different number. In the downloadable output, all targets below the threshold are included. As noted above, the downloadable output is available in two formats. The standard file includes the same

---

\(^1\) These results are described in the Appendix.
information for each target as is displayed in the table. The GFF3 file contains gene feature coordinates and can be used in many genome browser and related applications.

It is important to point out that the contributions of individual RVD-nucleotide associations to overall binding affinity are not yet worked out. Scoring based on RVD-nucleotide association frequencies provides only an estimate of relative affinities. Thus, not all sites returned may be efficiently bound by the TAL effector, irrespective of their score, and some biologically relevant targets may be missed if their scores rise above the arbitrary cutoff. Target Finder output should be considered a best estimate of the most probable binding sites for a given TAL effector based on available information. It represents a good starting point for further study to identify true targets or off-targets through experimentation, and a tool for initial assessment of probable relative specificities to choose among multiple arrays that might be available to target a sequence of interest.

A final note about the web interface for Target Finder: TALE-NT 2.0 provides a convenient way to use the tool, but working through the web interface restricts users to the model organism sequences available on the site, or relatively short, user-provided sequences. To study large datasets not included on the site, users can download from TALE-NT 2.0 a C-coded version of Target Finder and run it locally under an open source license.

Conclusion

TALEN Targeter and TAL Effector Targeter are versatile tools that allow design of custom RVD arrays for gene editing, engineered gene regulation and other applications. Although other TALEN design tools exist (TALEN Hit, http://talen-hit.cellectis-bioresearch.com/search; ZiFit Targeter 4.0 (16), http://zifit.partners.org/ZiFiTBeta/Introduction.aspx and idTALE (29), http://idtale.kaust.edu.sa/index.html), TALEN Targeter is the only tool that works with any architecture by allowing users to specify ranges for both spacer size and number of repeats. TAL Effector Targeter is the only tool available that targets single custom TAL effector arrays. Although idTALE allows users to search a genome for paired TALEN sites, its search function identifies exact matches only. Target Finder uses a scoring function that allows a biologically relevant number of mismatches in the TAL effector-target alignment, useful for identifying candidate targets of naturally occurring TAL effectors, as well as potential off-targets for custom TAL effector-based proteins.

In addition to these design and targeting tools, TALE-NT 2.0 also includes help pages and tutorials, with guides to interpretation of results. A “Protocols and Reagents” page provides useful links to other TAL effector and TALEN resources. With its wide range of capabilities and
content, TALE-NT 2.0 should be a valuable resource for anyone studying TAL effector function or using custom TAL effectors, TALENs, or other TAL effector-based proteins for DNA targeting applications. A new tool, “Paired Target Finder”, that automates the identification of potential off-target sites for TALENs, as heterodimer or either homodimer, was recently added to the web site.

**Acknowledgements**

The authors thank D. Mistry, W. Kokulapalan and the Iowa State University Bioinformatics and Computational Biology Laboratory for assistance creating the web site, and members of the Bogdanove and Voytas laboratories for beta-testing and suggestions.

**Funding**

The National Science Foundation [IOS\#0820831] (to A.B. in part and ISO\#1221984 to V.B. in part) and the National Institutes of Health [R01GM098861] (to A.B. and D.V.). Funding for the open access charge: NIH grant [R01GM098861].

*Conflict of interest statement.* None declared.

**References**


Supplementary Material

The Target Finder scoring function

To predict and rank sites, the tool uses a scoring function developed based on the RVD nucleotide association frequencies in known TAL effector-target pairs used by Moscou and Bogdanove to establish the RVD-nucleotide binding code (1). Precisely, if $F_{r,n}$ is taken as the observed frequency (percent) with which RVD $r$ associates with nucleotide $n$ in the set of identified pairs, then $P_{r,n} = w F_{r,n} + (1-w) 0.25$ estimates the probability of association of RVD $r$ with nucleotide $n$ in general. Here $w$ is set to 0.9 and corrects for unobserved data (i.e. the
possibility of an RVD that occurs in the observed set participating in an unobserved RVD-nucleotide pair). In this correction, any nucleotide is taken to occur with equal probability. For any RVD $v$ not included in the set of known TAL effector-target pairs, $F_{vn}$ is set so that $P_{vn}$ is 0.25 for all $n$. The score for an RVD sequence/DNA alignment with $l$ RVDs aligned with $l$ consecutive bases is then calculated as

$$S = -\sum_{1}^{l} \log(P_{rn}(l))$$

where $P_{rn}(l)$ is the association probability of the RVD-nucleotide pair at position $l$ in the alignment.

Supplementary Table S1. Target Finder RVD-nucleotide association probabilities.

<table>
<thead>
<tr>
<th>RVD</th>
<th>Total Observations</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>HD</td>
<td>107</td>
<td>0.084</td>
<td>0.858</td>
<td>0.025</td>
<td>0.033</td>
</tr>
<tr>
<td>NI</td>
<td>64</td>
<td>0.841</td>
<td>0.109</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>NG</td>
<td>70</td>
<td>0.102</td>
<td>0.102</td>
<td>0.038</td>
<td>0.758</td>
</tr>
<tr>
<td>NN</td>
<td>57</td>
<td>0.357</td>
<td>0.151</td>
<td>0.436</td>
<td>0.057</td>
</tr>
<tr>
<td>NS</td>
<td>30</td>
<td>0.625</td>
<td>0.205</td>
<td>0.145</td>
<td>0.025</td>
</tr>
<tr>
<td>N*</td>
<td>20</td>
<td>0.070</td>
<td>0.520</td>
<td>0.070</td>
<td>0.340</td>
</tr>
<tr>
<td>HG</td>
<td>18</td>
<td>0.075</td>
<td>0.125</td>
<td>0.025</td>
<td>0.775</td>
</tr>
<tr>
<td>HA</td>
<td>6</td>
<td>0.175</td>
<td>0.625</td>
<td>0.175</td>
<td>0.025</td>
</tr>
<tr>
<td>ND</td>
<td>4</td>
<td>0.025</td>
<td>0.925</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>NK</td>
<td>2</td>
<td>0.025</td>
<td>0.025</td>
<td>0.925</td>
<td>0.025</td>
</tr>
<tr>
<td>HI</td>
<td>1</td>
<td>0.025</td>
<td>0.925</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>HN</td>
<td>1</td>
<td>0.025</td>
<td>0.025</td>
<td>0.925</td>
<td>0.025</td>
</tr>
<tr>
<td>NA</td>
<td>1</td>
<td>0.025</td>
<td>0.025</td>
<td>0.925</td>
<td>0.025</td>
</tr>
<tr>
<td>IG</td>
<td>1</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.925</td>
</tr>
<tr>
<td>H*</td>
<td>1</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.925</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0.250</td>
<td>0.250</td>
<td>0.250</td>
<td>0.250</td>
</tr>
</tbody>
</table>

References
CHAPTER 5. ADDITION OF TRANSCRIPTION ACTIVATOR-LIKE EFFECTOR BINDING SITES TO A PATHOGEN STRAIN-SPECIFIC RICE BACTERIAL BLIGHT RESISTANCE GENE MAKES IT EFFECTIVE AGAINST ADDITIONAL STRAINS AND AGAINST BACTERIAL LEAF STREAK

A paper published in *New Phytologist*¹

Aaron W. Hummel², Erin L. Doyle², and Adam J. Bogdanove²,³

**Summary**

• *Xanthomonas* transcription activator-like (TAL) effectors promote disease in plants by binding to and activating host susceptibility genes. Plants counter with TAL effector-activated executor resistance genes, which cause host cell death and block disease progression. We asked whether the functional specificity of an executor gene could be broadened by adding different TAL effector binding elements (EBEs) to it.

• We added six EBEs to the rice *Xa27* gene, which confers resistance to strains of the bacterial blight pathogen *Xanthomonas oryzae* pv. *oryzae* (*Xoo*) that deliver the TAL effector *AvrXa27*. The EBEs correspond to three other effectors from *Xoo* strain PXO99A and three from strain BLS256 of the bacterial leaf streak pathogen *Xanthomonas oryzae* pv. *oryzicola* (*Xoc*).

• Stable integration into rice produced healthy lines exhibiting gene activation by each TAL effector, and resistance to PXO99A, a PXO99A derivative lacking *AvrXa27*, and BLS256, as well as two other *Xoo* and 10 *Xoc* strains virulent toward wildtype *Xa27* plants. Transcripts initiated primarily at a common site. Sequences in the EBEs were found to occur nonrandomly in rice promoters, suggesting an overlap with endogenous regulatory sequences.

• Thus, executor gene specificity can be broadened by adding EBEs, but caution is warranted because of the possible coincident introduction of endogenous regulatory elements.

**Introduction**

Many crops, including rice, wheat, cotton, citrus, tomato, cassava, banana, soybean, sugarcane, and others, suffer losses as a result of infection by pathogenic members of the bacterial genus *Xanthomonas*. Rice (*Oryza sativa*), a staple for more than half the world’s

¹ Reprinted with permission from *New Phytologist*, 2012, 195(4), 883-893. Copyright The Authors.
² Department of Plant Pathology and Microbiology, Iowa State University, Ames, IA 50011, USA
³ Author for correspondence.
As the rice activated by the Xoo TAL effector AvrXa27 (Gu et al., 2006), which causes bacterial blight, and the leaf mesophyll pathogen, Xanthomonas oryzae pv. oryzicola (Xoc), which causes bacterial leaf streak. These diseases reduce yields by up to 50 and 30%, respectively (Niño-Liu et al., 2006). Like several Xanthomonas species, Xoo relies on transcription activator-like (TAL) effectors to render the host susceptible to colonization (Bai et al., 2000; Yang & White, 2004; Yang et al., 2006; Sugio et al., 2007; Antony et al., 2010). Xoc also deploys TAL effectors, which, though less well studied, are presumed to contribute similarly to virulence (Makino et al., 2006; Bogdanove et al., 2011).

Named for features shared with eukaryotic transcription factors (Yang et al., 2006), TAL effectors are secreted into host cells via the bacterial type III secretion system (T3SS), then directed to the host cell nucleus by C-terminal nuclear localization signals (Yang & Gabriel, 1995; Van den Ackerveken et al., 1996; Szurek et al., 2001, 2002). There they bind to cognate effector binding elements (EBEs) in specific host gene promoters to activate transcription of those genes using a C-terminal, acidic activation domain (Zhu et al., 1998, 1999; Kay et al., 2007; Römer et al., 2007; Boch et al., 2009; Moscou & Bogdanove, 2009). Binding specificity is dictated by a variable number of central, 33–35 amino acid repeats. In each repeat, a pair of variable residues at positions 12 and 13 (together called the repeat-variable diresidue (RVD)) preferentially associates with a different nucleotide to define the length and sequence of the EBE (Boch et al., 2009; Moscou & Bogdanove, 2009). With this modular protein-DNA recognition mechanism, the pathogen can activate multiple susceptibility (S) genes in the host by deploying different TAL effectors. Several bacterial blight S genes in rice that correspond to Xoo TAL effectors important for virulence have been identified (Yang et al., 2006; Sugio et al., 2007; Antony et al., 2010), and several more S gene candidates for bacterial blight and bacterial leaf streak have been predicted computationally based on RVD sequences of uncharacterized Xoo and Xoc TAL effectors (Moscou & Bogdanove, 2009).

Plants counter TAL effector-wielding pathogens with S gene mimics that cause host cell death and block disease progression when transcriptionally activated. Such normally silent, EBE- controlled ‘executor’ resistance genes (Bogdanove et al., 2010) include the pepper (Capsicum annuum) Bs3 gene, which provides resistance against strains of the bacterial spot pathogen Xanthomonas campestris pv. vesicatoria that express the TAL effector AvrBs3 (Bonas et al., 1989; Römer et al., 2007), and the rice bacterial blight resistance gene Xa27, which is activated by the Xoo TAL effector AvrXa27 (Gu et al., 2005). In addition to executor genes, alleles of major S genes that are immune to activation by the corresponding TAL effector, such as the rice xa13 and xa25 bacterial blight resistance genes, provide another form of defense.
A third type of resistance directed at TAL effectors, again exemplified by a rice bacterial blight resistance gene, *xa5*, is a polymorphism in the gamma subunit of general transcription factor TFIIA. A single amino acid substitution in the protein is thought to impair the ability of TAL effectors to recruit the transcriptional machinery to activate target genes (Iyer & McCouch, 2004; Sugio et al., 2007; Gu *et al*., 2009). Each of these types of resistance, however, is subject to defeat by adaptation of the pathogen TAL effector inventory. For example, executor genes can be defeated by mutation or loss of the corresponding TAL effector, provided the TAL effector is dispensable for virulence, as appears to be the case for AvrXa27 (Tian & Yin, 2009). S gene promoter polymorphisms that confer resistance can be overcome by TAL effectors that target the new sequence or an alternative S gene (Antony *et al*., 2010). Also, the *xa5* gene is rendered ineffective by a strain with two apparent adaptations, a TAL effector that activates the corresponding S gene particularly strongly such that the reduction in activity caused by *xa5* might be inconsequential, and a TAL effector that induces expression of a TFIIA gamma paralog that may substitute for the allele found in susceptible plants (Yang *et al*., 2006; Sugio *et al*., 2007; B. Yang, unpublished). Furthermore, the latter two types of resistance are genetically recessive, rendering them less easily bred into elite hybrids. A genetically dominant, broad-spectrum, and durable form of resistance effective against pathogens that deploy TAL effectors would be beneficial, but none has been identified.

In an *Agrobacterium*-mediated transient expression assay in *Nicotiana benthamiana*, Römer *et al*., (2009a) showed that amending the *Bs3* gene promoter with the AvrXa27 EBE (which they call a ‘UPT box’, which stands for up-regulated by TAL effector) and an EBE matching an AvrBs3 variant called AvrBs3Δrep16 rendered the promoter responsive to all three TAL effectors. This pioneering study suggested that broad-spectrum and potentially durable resistance might be achieved by stable integration of an executor gene engineered in this way to respond to TAL effectors from multiple pathogen strains or even different pathogens.

In previous studies, stable integration of Xa27 into rice under conditions in which it was expressed constitutively resulted in reduced tillering, delayed flowering, and stiff, early-senesceding leaves; nonetheless, the expression of the gene indeed conferred resistance to several *Xoo* strains normally virulent on wildtype Xa27 lines, and partial resistance to a strain of *Xoc* (Gu *et al*., 2005; Tian & Yin, 2009). We therefore chose Xa27 to test the notion suggested by Römer *et al*., (2009a) that functional specificity of an executor gene could be broadened, without deleterious effects associated with constitutive expression, by making its promoter responsive to several distinct TAL effectors.
We added to the Xa27 promoter EBEs corresponding to three additional TAL effectors each from the Xoo strain PXO99A, which harbors AvrXa27, and the Xoc strain BLS256, which does not. Stable integration of this construct into rice produced healthy lines exhibiting gene activation by each of the TAL effectors, and resistance to PXO99A, a PXO99A derivative lacking AvrXa27, and BLS256, as well as two other Xoo and 10 Xoc strains from a diverse collection virulent toward wildtype Xa27 plants. Our results establish the efficacy of executor gene promoter engineering for broader specificity. They also demonstrate that a rice gene for bacterial blight resistance can be readily modified to provide complete protection from bacterial leaf streak as well, a disease for which no major gene resistance in rice has been identified.

**Materials and Methods**

**Modification of Xa27**

A 2.4 kb sequence including the Xa27 gene (AY986492.1) was PCR-amplified from rice cultivar IRBB27 (Gu et al., 2005) with primers P250 (CACCTGCAGCTGAACCAACAGTTTACGC) and P251 (GGGCCCATTTACTTTATTATTATTGTGCTGAC) using the Phire Hotstart polymerase (Thermo Fisher Scientific, Waltham, MA, USA) and cloned into the Gateway entry vector pENTR/D-TOPO (Life Technologies, Grand Island, NY, USA). To modify the promoter, custom oligonucleotides (Integrated DNA Technologies, Coralville, IA, USA) were synthesized to generate double-stranded DNA fragments with appropriate overhangs that were then cloned into the unique SpeI site 100 bp upstream of the Xa27 coding sequence and 14 bp upstream of the AvrXa27 EBE. Oligonucleotides were designed such that EBEs are flanked by the weak consensus context of TAT at the 5’ end and CCC at the 3’ end (Moscou & Bogdanove, 2009) and such that EBEs for TAL effectors from Xoo alternate with those for TAL effectors from Xoc. The EBE and control constructs (Fig. 1) were transferred to the binary vector pTF101.1gw1 (Plant Transformation Facility, Iowa State University, http://www.agron.iastate.edu/ptf/), which carries the bar gene for resistance to the herbicide phosphinothricin, using the Gateway LR II Clonase enzyme kit (Life Technologies).

**Plant transformation and growth**

Rice (Oryza sativa L. cv Kitaake) was transformed by the Iowa State University Plant Transformation Facility using Agrobacterium tumefaciens gene transfer in callus tissue, as
Figure 1. Engineered Xa27 constructs used in this study. A 2.4 kb region from rice variety IRBB27, including the Xa27 gene, its promoter and 3’ untranslated region (UTR) sequences (drawn to scale), was modified by inserting new sequences at an SpeI restriction enzyme site at position -100 relative to the coding sequence (CDS), 14 nucleotides upstream of the native AvrXa27 effector binding element (EBE). In the UXO (up-regulated by Xanthomonas oryzae) construct, EBEs corresponding to transcription activator-like (TAL) effectors Tal4c, Tal2g, and Tal4a from Xanthomonas oryzae pv. oryzicola (Xoc) strain BLS256 alternate with EBEs corresponding to TAL effectors PthXo1, PthXo6, and Tal9a from Xanthomonas oryzae pv. oryzae (Xoo) strain PXO99A. In the XOC construct, the EBEs for the three Xoo TAL effectors are substituted with internally randomized sequences. The inverse is true of the XOO construct. In the RAN construct, all six EBE sequences are substituted with the internally randomized ones. The native AvrXa27 EBE (shown at bottom) is maintained in all constructs except XOC, in which the corresponding, nonfunctional sequence from the xa27 allele was substituted (differences are indicated above the EBE; ‘-’ indicates a deletion). The putative TATA box upstream of the AvrXa27 EBE is italicized. In all constructs, the nopaline synthase terminator (NOS-T) was included upstream of the Xa27 promoter to prevent read-through transcription by any neighboring promoter at the point of insertion into the genome.

previously described (http://www.agron.iastate.edu/ptf/protocol/Rice.PDF). Herbicide-resistant lines were advanced to the T2 generation for characterization. Presence of the transgene was assessed using the Kapa3G Plant PCR Kit (Kapa Biosystems, Woburn, MA, USA) with primers P664 (GGCATTCTTCCTTTCTTCAGC) and P352 (GGAGGCAGCTTCTTGGGTGTCTCAG). Transgenic and IRBB27 rice plants were grown in a growth chamber under cycles of 12 h, 28°C : 12 h, 25°C; light : dark. Fertilizer for acid-loving plants (30-10-10; Earl May, Shenandoah, IA, USA) and iron chelate micronutrient (Becker Underwood, Ames, IA, USA) were applied by watering twice a week at rates of 0.185 and 0.595 g l⁻¹, respectively.

TAL effector clones

Clones encoding PthXo1, PthXo6, Tal9a, AvrXa27, and AvrXa10 were provided by B. Yang (Iowa State University), and clones encoding Tal2g, Tal4a, Tal4c, and Tal1c were from C. Schmidt (our laboratory). The SphI fragment of each TAL effector gene, encoding its central repeat region, was transferred into a functionally equivalent tal1c gene backbone (missing its own SphI fragment) in plasmid pCS466, a derivative of the pCR8/GW/TOPO TA entry vector (Life Technologies, Grand Island, NY, USA). Each effector construct was then moved via the
Gateway LR II Clonase enzyme kit (Life Technologies, Grand Island, NY, USA) into pKEB31, a derivative of the broad host range plasmid pDD62 (Mudgett et al., 2000) in which the nptII gene is replaced by the tetR gene of pBR322. pKEB31 places the effector gene under the control of the lac promoter, which is constitutive in Xanthomonas.

Isolation and transformation of Xag EB08 and cultivation of Xanthomonas strains

*Xanthomonas axonopodis* pv. glycines (Xag) strain EB08 was isolated from diseased soybean (*Glycine max* L. Merr.) leaves that had been stored at -20°C after being collected in Iowa in 2008. Leaf tissue was ground in sterile water; an aliquot of supernatant was cultured on glucose yeast extract (GYE) agar amended with 150 µg ml⁻¹ cyclohexamide, and a yellow, mucoid colony was isolated. The isolate was confirmed as the causal agent of bacterial pustule of soybean by reinoculation to healthy soybean plants. Xag EB08 was transformed by electroporation as previously described (Tsuge et al., 2001) and cultured on GYE agar amended with 10 µg ml⁻¹ tetracycline for plasmid selection. *X. oryzae* strains were cultured on GYE.

Plant inoculations, virulence assays, and qPCR

Fresh (24–48 h) bacterial cultures were scraped from GYE plates and suspended in 10 mM MgCl₂ to an optical density at 600 nm (OD600) = 0.5 as inoculum for virulence assays, and OD600 = 0.9 as inoculum for quantitative real-time reverse-transcriptase PCR (qPCR) assays. For *Xoc* virulence assays, 6-wk old plants were inoculated by infiltration of leaves with inoculum using a needleless syringe (Schaad et al., 1996), and the length of water-soaked lesions was measured at 10 d. For *Xoo*, leaf tips of 6-wk-old plants were clipped with scissors dipped in bacterial suspension (Kauffman et al., 1973), and the length of leaf curl was measured at 14 d. For qPCR, bacteria were inoculated with a needleless syringe into the youngest, fully expanded leaf on 4-wk-old plants. Inoculated tissue was flash-frozen after 48 h, and total RNA was extracted with the RNeasy Mini Plant Kit (Qiagen, Hilden, Germany). qPCR was performed on an iCycler Thermo Cycler (Bio-Rad, Hercules, CA, USA) with 100 ng total RNA as template for cDNA synthesis and PCR amplification using the iScript One-Step RT-PCR kit with SYBR Green (Bio-Rad). Gene-specific primers P772 (CCGTCATCCTCAT GCACATGCTCACCAC) and P771 (CACGGAGGAGAACTAGAGAGACCAGAGAC) were used for amplification of Xa27 cDNA, and P787 (CCGGTGGATCTTCATGCTCACCAC) and P788 (CGACGAGTCTCTGCGCAACTGC) were used for amplification of Actin-6 (EU215044) cDNA for normalization. A minimum of three independent biological replicates, each with three qPCR technical replicates, were tested for each treatment. The 2⁻∆∆Ct method (Livak &
Schmittgen, 2001) was used to quantify expression of Xa27 transcript for each treatment relative to mock-inoculated tissue.

5’ RACE
cDNA for 5’-RACE was produced from 1 µg total RNA with the SMARTer RACE cDNA amplification kit (Takara Bio, Inc., Otsu, Shiga, Japan) using the manufacturer’s protocol, from the same samples prepared for qPCR. Xa27 mRNA was amplified with the Universal forward primer (Takara Bio, Inc., Otsu, Shiga, Japan) and P771. For sequencing, PCR products were cloned using the pCR8/GW/TOPO TA entry vector system (Life Technologies, Grand Island, NY, USA).

Analysis of EBE sequence representation in rice promoters
Rice promoters (defined as the sequences from 400 to 50 bases upstream of all translational start sites) were extracted from the MSU Rice Genome Annotation Project database (http://rice.plantbiology.msu.edu/), version 6.1, and concatenated into a single ‘promoterome’. The promoterome was then randomized 5000 times. Next, all six-nucleotide sequences of each EBE were extracted using a sliding window. For each six-nucleotide sequence, the numbers of exact matches in the promoterome and in the 5000 randomized promoteromes were determined. To test if each six-nucleotide sequence occurred at higher- or lower-than-expected frequencies in the actual promoterome, P-values were estimated by dividing by 5000 the number of randomized promoteromes that contained more (or fewer) exact matches. To determine whether the observed numbers of six-nucleotide sequences that occurred at lower- or higher-than-expected frequencies were significant, the EBEs were randomized 100 times and the analysis was repeated for each randomization. P-values were estimated by dividing by 100 the number of EBE randomizations with more six-nucleotide sequences showing higher- or lower-than-expected frequencies than those from the real EBEs.

Results
Stable integration of EBE-amended Xa27 constructs into rice
We amplified Xa27 from rice variety IRBB27 (Gu et al., 2005) and amended its promoter with EBEs from target genes of three TAL effectors important for virulence of the Xoo strain PXO99A and EBEs found in rice for three TAL effectors of the Xoc strain BLS256, whose roles in virulence are as yet uncharacterized (Table S1). These include EBEs from rice genes 8N3/Xa13 (Os08g42350), TFX1 (Os09g29820), and Hen1 (Os07g06970) that correspond to
virulence factors PthXo1, PthXo6, and Tal9a, respectively, from PXO99A (Yang et al., 2006; Sugio et al., 2007; Salzberg et al., 2008; Römer et al., 2010; B. Yang, unpublished), and EBEs from rice genes Os03g37840 (a putative potassium transporter), Os06g37080 (a putative L-ascorbate oxidase precursor) and Os05g34600 (a putative NAC (which stands for NAM (no apical meristem), ATAF, CUC (cup-shaped cotyledon)) transcription factor), predicted to be bound by TAL effectors Tal4a, Tal4c and Tal2g, respectively, of BLS256 (Moscou & Bogdanove, 2009; Bogdanove et al., 2011; E. L. Doyle & A. J. Bogdanove, unpublished). The TAL effectors are reciprocally unique to each strain. The EBEs were inserted upstream of the native AvrXa27 binding site. To maximize space for simultaneous binding of multiple effectors from the same pathogen while keeping the length of the overall modification short, sites for Xoo were alternated with those for Xoc and further separated by 6 bp (Fig. 1 and Supporting Information, Fig. S1). The nopaline synthase terminator was placed upstream of the Xa27 promoter to prevent read-through expression of the transgene should it insert downstream of an active endogenous promoter. This construct, containing all six EBEs, was designated as UXO (up-regulated by X. oryzae). Three other constructs were also made (Figs 1, S1). In the first, termed RAN (for randomized EBEs), sequences of all six EBEs were internally randomized, thereby maintaining the nucleotide composition and length of the UXO construct, but destroying the binding sites. In the second, termed XOO, the Xoo EBEs were kept intact but the Xoc EBEs were randomized as in the RAN construct. In the third, called XOC, the inverse was done to retain only the Xoc EBEs. The native AvrXa27 EBE was unchanged in all except the XOC construct, where, in order to generate strictly Xoc-specific resistance, it was converted to the corresponding sequence of the nonfunctional xa27 allele, which is not activated by AvrXa27 (Gu et al., 2005). None of the constructs contains any sequence that would be predicted to bind a PXO99A or BLS256 TAL effector other than those for which the constructs were designed.

All four constructs were stably integrated into the readily transformed, short-season japonica rice cv. Kitaake. The xa27 allele in this variety is nonfunctional and identical in sequence to that (Os06g39810) in the reference japonica genome, Nipponbare. All recovered events were advanced for characterization. Of three lines from independent transformation events with the UXO construct that were advanced to the T2 generation, lines 1 and 3 were found by PCR amplification to contain the UXO construct, and were analyzed further. Similarly, two independent lines for each of the RAN and XOC constructs were confirmed by PCR and characterized. Only a single line with the XOO construct survived to the T2 generation, and this was also characterized. Finally, two events resulting from transformation with the wild type Xa27 gene were advanced to the T2 generation. One, termed HRB, was found to contain the
herbicide resistance cassette but not the modified Xa27 gene. In the second, called AZY, neither the herbicide resistance cassette nor the Xa27 gene could be detected at T2. These two lines were retained as controls. Plants from all lines advanced to T2 were morphologically and developmentally indistinguishable from untransformed plants with the exception of the XOC lines, which were slightly reduced in stature. In leaves of the UXO and XOO lines, basal levels of expression of Xa27 were similar to that in IRBB27. In the RAN lines and the one tested XOC line, basal expression was moderately elevated relative to that in IRBB27 (Table S2).

**EBE-amended Xa27 is activated by Xoo and Xoc independent of AvrXa27**

To examine responsiveness of the UXO construct to the corresponding TAL effectors, we first analyzed the relative fold-change in Xa27 mRNA 48 h after syringe inoculation of T2 leaves with Xoo PXO99^A, the avrXa27-deficient mutant derivative ME1 (Gu et al., 2005), Xoc BLS256, and T3SS-deficient mutants of PXO99^A and BLS256. qPCR revealed significant (P < 0.05) up-regulation of the transgene in the UXO-1 and UXO-3 lines when inoculated with PXO99^A, ME1, or BLS256 relative to mock-inoculated tissue, but not when inoculated with either of the T3SS mutants (Fig. 2). By contrast, the gene in the RAN-1 and RAN-2 lines was not significantly up-regulated by any of the bacterial strains, unexpectedly including PXO99^A, in spite of the AvrXa27 EBE present in the RAN construct. Inoculation of PXO99^A to the

![Figure 2. Type III secretion system (T3SS)-dependent and AvrXa27-independent activation of effector binding element (EBE)-amended Xa27 by Xanthomonas oryzae pv. oryzae (Xoo) PXO99^A and Xanthomonas oryzae pv. oryzicola (Xoc) BLS256. Vertical bars represent fold-change in abundance of Xa27 mRNA in the UXO and RAN lines and in IRBB27 at 48 h after inoculation with the indicated bacterial strains, relative to mock-inoculated plants, measured by quantitative real-time reverse-transcriptase PCR (qPCR). Each is the average of three technical replicates of at least three biological replicates. Capped vertical lines show standard error. Asterisks indicate values significantly > 1.0 (dashed line) as determined by two-tailed, heteroscedastic t-tests. Relative fold-change was calculated by the 2^−ΔΔCt method. ME1 is an avrXa27-deficient PXO99^A derivative.](image)
nontransgenic variety IRBB27, which carries the wildtype Xa27 gene, resulted in the expected activation, indicating that the failure of the RAN lines to respond was not the result of a defect in the PXO99A isolate.

**EBE-amended Xa27 is specifically activated by each TAL effector**

To assess the function of each EBE in the UXO construct, we measured the fold-change in Xa27 mRNA in leaves of T2 plants of the UXO-3 line 48 h after syringe infiltration with transformants of strain EBO8 of the soybean pathogen Xag expressing each of the corresponding effectors individually. This strain on its own produces no visible response in rice. Xa27 in the UXO-3 plants was significantly (P < 0.05) up-regulated by all seven TAL effectors, that is (in the order of their EBEs) Tal4c, PthXo1, Tal2g, PthXo6, Tal4a, Tal9a, and AvrXa27, when compared with mock-inoculated tissue (Fig. 3). The Xoo TAL effector AvrXa10 (Hopkins et al., 1992), and the Xoc TAL effector Tal1c (Bogdanove et al., 2011), for which UXO-3 contains no EBEs, as well as Tal1c lacking its central repeat region, caused no significant up-

![Figure 3](image-url)

**Figure 3. Specific activation of Xa27 in the UXO construct by each of the corresponding transcription activator-like (TAL) effectors individually.** Vertical bars represent relative abundance of Xa27 mRNA, measured by quantitative real-time reverse-transcriptase PCR (qPCR), in the UXO-3 line and in IRBB27 48 h after inoculation with *Xanthomonas axonopodis* pv. glycines EBO8 expressing one of the seven corresponding *Xanthomonas oryzae* pv. oryzae (Xoo) or *Xanthomonas oryzae* pv. oryzicola (Xoc) TAL effectors, or as controls for specificity, the Xoo TAL effector AvrXa10, the Xoc TAL effector Tal1c, or Tal1c lacking its central repeat region (∆CRR), as indicated, relative to mock-inoculated plants. Each is the average of three technical replicates of at least three biological replicates. Capped vertical lines show the standard error. Asterisks indicate values significantly > 1.0 (dashed line) as determined by two-tailed, heteroscedastic t-tests. Relative fold-change was calculated by the $2^{\Delta\Delta Ct}$ method. Values for the seven corresponding effectors are shown in the order of their effector binding elements (EBEs) in UXO (left to right, 5’ to 3’).
regulation (Fig. 3). Fold induction by the six targeted TAL effectors (Fig. 3) positively correlated with proximity of the corresponding EBE to an apparent TATA box immediately upstream of the native AvrXa27 EBE (Fig. 1; Spearman’s rank correlation coefficient = 0.8857, P < 0.01). The same pattern of induction of Xa27 by the individual TAL effectors was observed in the UXO-1 line (Fig. S2).

**All six EBEs drive transcription of Xa27 primarily from a shared start site**

Previous studies reported start sites for TAL effector-induced transcription at positions distinct from the start site for basal level expression; these were generally located 42–54 nucleotides downstream of the 3’ end of the EBE (Kay et al., 2007, 2009; Römer et al., 2009a; Römer et al., 2009b; Antony et al., 2010). To determine whether the position of the EBE for each TAL effector activating the UXO construct similarly defines the transcriptional start site (TSS), 5’ rapid amplification of cDNA ends (5’-RACE) was conducted on RNA collected for the UXO-3 gene expression assays described earlier (Fig. 4). Based on the 12–29 transcript sequences obtained per sample, despite the distinct positions of their corresponding EBEs, PthXo1, PthXo6, Tal9a, Tal4a, Tal4c, and Tal2g each initiated transcription primarily at a shared site, identical to that of the majority of transcript sequences obtained from basal expression in the mock and AvrXa10-treated negative controls. The site, 60 bp upstream of the start codon of the Xa27 coding sequence, resides 27 bp downstream of the native, putative TATA box that immediately precedes the AvrXa27 EBE, which positions it only 10 bp downstream of the 3’ end of that EBE. Consistent with the observations to date that TAL effector-driven transcription initiates a minimum of 42 bp from the 3’ end of an EBE, none of the AvrXa27-generated transcripts initiated at that common site. Rather, they initiated further downstream, nearly all of them at a distance of 46 bp from the 3’ end of the AvrXa27 EBE, 24 bp upstream of the Xa27 start codon. Xoo PXO99A- and Xoc BLS256-inoculated plants each exhibited a greater diversity of TSSs than was generated by any of the individually delivered effectors, including some TSSs not observed following individual TAL effector delivery. Conversely, however, some sites that were observed upon delivery of the TAL effectors individually were not detected following inoculation with the pathogen strains. The most common TSS in PXO99A-inoculated plants was the shared site at 60 bp upstream of the Xa27 start codon. This was the second most common for BLS256-inoculated plants, the first being a unique site upstream, at -149 relative to the Xa27 coding sequence. Other apparent TSS locations for PXO99A- and BLS256-inoculated plants ranged as far upstream as -165 relative to the Xa27 translational start, with no discernible correlation to EBE positions.
EBE-amended *Xa27* confers resistance to a broader spectrum of *Xoo* strains and to all tested *Xoc* strains

To assess the effectiveness of adding EBEs to *Xa27* to broaden its functional specificity, we first carried out quantitative virulence assays of ME1 and BLS256 in T2 plants of the UXO,
XOO, XOC, and RAN lines, as well as the HRB and AZY controls and IRBB27 (Fig. 5a,b). The extent of bacterial blight symptoms (length of leaf curling) following leaf clip inoculation (Kauffman et al., 1973) with ME1 was significantly (P < 0.01) and markedly reduced specifically in the UXO lines and the XOO line relative to the RAN, XOC, and HRB lines, and IRBB27. The single AZY line, for unknown reasons, showed an intermediate mean leaf curl length. Bacterial leaf streak lesions in the UXO and XOC lines following syringe inoculation (Schaad et al., 1996) with BLS256 were significantly (P < 0.01) and markedly reduced relative to all other lines. The patterns of resistance and susceptibility to ME1 and BLS256 in the EBE-amended and control lines correlate strictly with the patterns of Xa27 activation described earlier (Fig. 2), and demonstrate that the amendment resulted in a specific broadening of the resistance spectrum of Xa27 to include both bacterial blight and bacterial leaf streak.

Figure 5. Bacterial blight and bacterial leaf streak resistance in the effector binding element (EBE)-amended Xa27 transgenic plants. (a) Length of leaf curl 14 d after leaf clip inoculation of T2 plants of the indicated lines with Xanthomonas oryzae pv. oryzae (Xoo) ME1, which lacks avrXa27. Different letters indicate a statistically significant difference between values (P < 0.01; UXO-3 and AZY values are different at P < 0.05). (b) Lesion lengths 10 d after syringe inoculation with Xanthomonas oryzae pv. oryzicola (Xoc) BLS256. Statistically significant differences (P < 0.01) are indicated as in (a). (c) Extent of leaf curl on T2 plants of UXO-1 or RAN-1 14 d following inoculation with Xoo strains isolated from several locations, as indicated (Phil., Philippines; Thai., Thailand). An asterisk indicates significantly reduced length of leaf curling (P < 0.05) relative to that observed on the RAN line. (d) Lesion length 10 d after inoculation with Xoc strains isolated from several locations, as indicated (Malay., Malaysia; Chi., China; Ind., India). Each strain produced significantly shorter lesions (P < 0.01) on UXO-1 than on RAN-1. Two-tailed, heteroscedastic t-tests were used to determine significance. Experiments were repeated at least twice with similar results. Error bars represent SD.
The distribution of the TAL effectors corresponding to the six EBEs in the \textit{UXO} construct and of AvrXa27 in field populations of \textit{Xoo} and \textit{Xoc} is not yet known. Nonetheless, we sought to determine the effectiveness of the \textit{UXO} construct against geographically diverse strains of both pathovars. Seven \textit{Xoo} strains reported to be virulent (1947 from Africa, C4 and ZHE173 from China, K202 from Korea, and 2 from Thailand) or weakly virulent (JW89011 from Korea and PXO71 from the Philippines) on IRBB27 (Gu et al., 2004) were inoculated to \textit{UXO}-1 and, for comparison, to \textit{RAN}-1 by clip inoculation (Fig. 5c). C4 and ZHE173 caused significantly (P < 0.05) and markedly reduced leaf curl lengths on \textit{UXO}-1 relative to \textit{RAN}-1. JW89011 caused significantly (P < 0.05) but less dramatically reduced leaf curl lengths on \textit{UXO}-1. The remaining \textit{Xoo} strains were not significantly less virulent on \textit{UXO}-1 than on \textit{RAN}-1. Ten geographically diverse strains of \textit{Xoc}, each virulent toward wildtype \textit{Xa27} plants (data not shown), were inoculated to \textit{UXO}-1 and \textit{RAN}-1 by syringe infiltration. The strains differed in their virulence on \textit{RAN}-1, but each showed significantly (P < 0.01) and markedly reduced relative virulence on \textit{UXO}-1 (Fig. 5d). Thus, EBE amendment to the \textit{Xa27} promoter broadened its functional specificity to include a diversity of \textit{Xoo} and \textit{Xoc} strains.

\textbf{The \textit{UXO} EBEs contain sequences apparently under selection in rice promoters}

To assess whether the EBEs added to the \textit{Xa27} promoter in the \textit{UXO} construct might contain \textit{cis} regulatory elements, a condition that could prevent TAL effector-mediated activation or cause TAL effector-independent activation under certain environmental conditions or during development, we examined representation of the EBEs in the rice ‘promoterome’, the collection of sequences from 400 to 50 bases upstream of all annotated translational start sites. We chose this sequence space based on the observation that the EBEs in the \textit{UXO} construct in their natural contexts occur from positions -362 (Tal4a) to -86 (AvrXa27) relative to the translational start site. We reasoned that any portions of EBE sequences corresponding to \textit{cis} elements in the promoterome would be under selection and therefore occur at nonrandom frequencies. For every possible six-nucleotide fragment of each EBE in the \textit{UXO} construct, we determined the number of instances of that sequence in the promoterome and in each of 5000 randomized promoteromes. Of the 97 six-nucleotide sequences contained in the EBEs, 40 occurred in the promoterome at lower frequencies than expected based on observed frequencies in the randomized promoteromes, and 54 occurred at higher-than-expected frequencies (P < 0.05, Fig. 6). To distinguish whether the large number of six-nucleotide sequences with nonrandom frequencies was a function of the EBE sequences or instead an artifact of the EBE nucleotide composition, we conducted the same test with six-nucleotide sequences from each of 100
Figure 6. Effector binding element (EBE) sequence representation in rice promoters. (a) Representation of the 97 sequences obtained from moving a six-nucleotide (nt) window through each EBE in the UXO construct. The number of occurrences of each sequence across all annotated rice gene promoters (from 400 to 50 bp upstream of the coding sequences) was determined. The numbers of sequences occurring at lower-than-expected (P < 0.05), expected, or higher-than-expected (P < 0.05) frequency relative to the distribution of frequencies found in 5000 sequence-randomized iterations of the promoter set are shown. (b–d) Examples of six-nt EBE sequences that are represented at a lower-than-expected frequency (b), the expected frequency (c), or a higher-than-expected frequency (dashed lines) (d) relative to the distribution of their frequencies in the randomized set (gray bars).

Discussion

A resistance gene that recognizes an effector that otherwise contributes strongly to virulence can be relatively durable due to the fitness cost to the pathogen of losing or modifying that effector to evade detection (Leach et al., 2001). A striking example is bacterial blight resistance mediated by the rice Xa7 gene, which is triggered by AvrXa7, a major virulence factor in Xoo (Hopkins et al., 1992; Bai et al., 2000; Vera Cruz et al., 2000; Ponciano et al.,
2003). However, a mutation that uncouples the virulence and avirulence (resistance-triggering) properties of an effector could result in defeat of the resistance gene. Also, because most resistance genes exhibit pathogen race specificity linked to recognition of a single effector, evolution or introduction of pathogen strains that use alternative effectors for virulence can render a resistance gene ineffective. Against pathogens that rely on TAL effectors for virulence, an executor gene engineered to respond to multiple TAL effectors important for virulence and conserved across a pathogen population could provide durable and broad-spectrum protection.

Here, we established the feasibility of such an approach by adding EBEs to the promoter of the Xa27 gene for resistance to rice bacterial blight and demonstrating consequent broadening of its functional specificity to include Xoo strains virulent toward wild type Xa27-containing plants as well as strains of the bacterial leaf streak pathogen Xoc, for which no simply inherited resistance genes had been identified in rice. Without altering the coding sequence and by adding less than 200 bp of recombinant DNA to the gene promoter, EBE amendment resulted in a single gene with an effective recognition spectrum similar to that achieved by the arduous and time-consuming process of pyramiding multiple resistance genes. Our selection of EBEs was based on available data from one strain each of Xoo and Xoc. Systematic studies to catalog the diversity of TAL effectors in Xanthomonas field populations should enable the rational modification of executor genes to provide broad and durable resistance on a geographically specific basis.

Importantly, each of the targeted TAL effectors individually activated the transgene in the UXO construct. Also, the observed patterns of resistance and susceptibility to Xoo ME1 and Xoc BLS256 across all transgenic lines were as expected based on the corresponding EBE and TAL effector content in each interaction, and they correlated directly with the patterns of Xa27 activation. We used ME1 in addition to PXO99A for the gene expression assays and in place of PXO99A for the virulence assays to distinguish resistance due to the EBEs added to Xa27 from resistance mediated by the native AvrXa27 EBE. It should be noted that in addition to its deficiency in avrXa27, ME1 is also disrupted in pthXo6, which neighbors avrXa27; though AvrXa27 makes no measurable contribution to virulence, pthXo6 does (Sugio et al., 2007). The virulence reduction due to the disruption of pthXo6 in ME1 is slight, however, and did not mask the resistance conferred by the EBE-amended Xa27 constructs. The disruption of pthXo6 in ME1 did, though, limit the effectors in that strain expected to interact with EBEs in the UXO promoter to PthXo1 and Tal9a. Together, the results strongly suggest that activation of the transgene by the wildtype Xoo and Xoc strains is mediated by at least one and possibly each of the corresponding TAL effectors in those strains.
Each of the TAL effectors we chose from PXO99A plays a role in virulence. None of the ones we chose from BLS256 has been characterized with respect to virulence, but each has a well-matched EBE in a putative target gene promoter in rice. The set of EBEs matching these effectors expanded the resistance spectrum of Xa27 to include not only the PXO99A mutant lacking AvrXa27, and BLS256, but two of seven additional, geographically diverse Xoo strains and each of ten additional, diverse Xoc strains tested. The wide resistance spectrum conferred by these EBEs against the Xoc strains and the narrower spectrum against Xoo strains suggests conservation of one or more of the BLS256 TAL effectors in the Xoc strains, and poor conservation of the PXO99A effectors in that group. This may reflect greater overall TAL effector diversity among Xoo strains, potentially due to diversifying selection exerted by the nearly 30 known bacterial blight resistance genes in rice (Niño-Liu et al., 2006). As noted, no simply inherited genes for bacterial leaf streak resistance have been identified in rice [the only identified source of complete resistance is the Rxo1 gene from maize (Zhao et al., 2005)], and this might explain the apparently lesser TAL effector diversity across Xoc strains. To conclude with confidence whether a difference in TAL effector diversity truly exists between Xoo and Xoc, however, a more comprehensive and direct inventory across a broad collection of strains would be necessary.

In light of the absence of major, native genes for resistance to bacterial leaf streak in rice, it is particularly promising that Xa27 was fully effective against Xoc. The only executor gene in rice cloned to date, Xa27 encodes a 113 amino acid product that has no similarity to functionally characterized proteins, but contains an amino-terminal signal-anchor-like sequence that mediates export to the apoplast and is required for bacterial blight resistance (Wu et al., 2008). Xoo colonizes rice xylem vessels, interacting with cells in the surrounding xylem parenchyma to cause bacterial blight. Xoc multiplies in the mesophyll parenchyma. The mechanism by which Xa27 confers resistance to bacterial blight is unknown. Its activation causes cell death, but it is yet unclear whether the death is programmed or due to toxicity of the Xa27 protein. Its effectiveness against bacterial leaf streak, demonstrated here, indicates that the mechanism is general and not restricted to cells in the xylem parenchyma.

As noted earlier, weak constitutive expression of Xa27, though deleterious to the plant, conferred partial resistance to Xoc strain L8 (Tian & Yin, 2009). The complete resistance we observed is likely due to stronger, localized expression in response to one or more of the targeted TAL effectors delivered by Xoc. Fold induction of Xa27 in the UXO construct by each of the targeted Xoo and Xoc TAL effectors individually correlated positively with proximity of the EBE to a putative TATA box downstream and to the common start site of transcription. Although
differences in expression, delivery, or affinity of the TAL effectors may have contributed to this pattern, the strength and significance of the correlation of EBE position to fold induction suggest that there may be a limit to the number of EBEs that can be effectively added to a promoter, as those farthest upstream might drive expression only weakly or not at all. We also observed, unexpectedly, that PXO99^ failed to activate Xa27 in the RAN lines, despite its ability to do so in IRBB27. The RAN construct maintains the native AvrXa27 EBE, differing from the UXO construct only in its internal randomization of the added EBEs. Thus, the randomized sequences in some way prevent activation by AvrXa27, or less likely, in both lines activation is blocked due to the positions at which the construct integrated. Though the TAL effector-DNA binding code enables prediction of TAL effector binding sites with relative confidence, the above observations highlight the fact that we still know little of the contextual requirements for functional EBEs, both with respect to promoter sequences as well as chromosomal location and chromatin status. Until more is known, successful placement of EBEs in a promoter may require some trial and error.

Another complexity in EBE-amendment is the possibility of interaction among different TAL effectors binding to the same promoter. We sought to minimize this by spacing EBEs six bp apart and alternating those for Xoo with those for Xoc. Nonetheless, our 5' RACE results comparing TSSs detected following individual TAL effector delivery to those detected following inoculation with the corresponding Xoo or Xoc strain provided evidence for such interaction. Some TSSs detected following inoculation with the pathogen strain were not detected on delivery of any of the corresponding individual effectors and, to a lesser extent, vice-versa. Thus, although the collection of transcript sequences may not have been saturating, the data suggest that both cooperative interaction to generate novel TSSs as well as interference that blocks initiation from some sites take place.

Previous analyses showed TAL effector-initiated transcription to occur between 42 and 54 nucleotides downstream of the 3' end of the EBE, even when the EBE was moved out of context (Kay et al., 2007; Römer et al., 2007; Kay et al., 2009; Römer et al., 2009; Römer et al., 2009; Antony et al., 2010). Our results differ. Most transcripts driven by each of the six non-native TAL effector-EBE interactions on the Xa27 promoter in the UXO construct initiated at a site shared by all interactions, rather than at sites corresponding to the relative positions of the EBEs. Because this was the same site of transcript initiation observed in mock-inoculated leaves and leaves treated with the non-targeted negative control AvrXa10, it appears that in some configurations, and perhaps depending on the EBE, TAL effector induced transcript initiation can default to the primary site used for basal expression. The location of the common
TSS at 27 bp downstream of a putative TATA box suggests that presence of such an element might influence the location of TAL effector-driven transcript initiation. Some of the TAL effectors reported to dictate the TSS display a TATA box-like sequence within their EBE, which might explain that ability, but not all of them do. Indeed, the AvrXa27 EBE contains no TATA-like sequence, yet drives transcription from a site downstream of the common TSS, perhaps because it is too close to that TSS. In all, it appears that the site of TAL effector driven transcript initiation is influenced by position of the EBE and by other promoter features in a yet poorly understood way.

We discovered that the EBEs we chose for the UXO construct contain sequences apparently under selection in rice promoters, suggesting coincidence with endogenous regulatory elements. These might include elements for gene activation in response to a particular environmental or developmental cue. Introduction of an EBE with such an element into the promoter of an executor gene could lead to TAL effector-independent cell death on that cue, with potentially disastrous consequences. We examined sequences from only six EBEs, so our findings may not be generalizable. However, TAL effectors might be expected typically to target endogenous regulatory elements in the host genome, by chance due to their localization in promoters, but also because such regulatory elements would likely be relatively immutable, and thereby confer selective advantage on any corresponding TAL effector.

Though the lines included in this study were healthy and fertile under controlled growth conditions, we encountered difficulty retrieving viable, stably transformed lines for some constructs we made. In addition to the UXO construct, in which the EBEs are exactly those found in the rice genome and contain some nucleotides that do not match the RVD at that position in the corresponding TAL effector, we also made a construct containing analogous, perfectly matched EBEs. Plants with this construct were unhealthy, and no lines survived to T2. Also, we were unable to obtain plants transformed with wildtype Xa27. The latter observation, consistent with previous findings (Wu et al., 2008), suggests that the unmodified Xa27 promoter can drive expression of the executor gene and consequent cell death when outside of its native genomic context. Similarly, even with the terminator upstream, transformation events using the XOO construct only yielded one healthy line. Thus, in addition to the possibility that EBEs added to a promoter may contain regulatory elements that could drive TAL effector independent-independent activation, synthetic sequences, sequences out of context, and position effects might also be problematic.

In sum, though the results presented here establish the effectiveness of EBE amendment to expand the spectrum of pathogen genotypes against which an executor gene functions, multiple
alternative constructs may be necessary to obtain effective, stable transformants, and all lines should undergo testing under a variety of growth conditions and field environments prior to distribution or commercialization. Analysis of the representation of EBE sequences in the host promoterome might be a useful preliminary to selection of EBEs, in order to exclude those that contain sequences apparently under selection. Although beyond the scope of the work we have presented here, it might also be useful to determine whether any sequences that are over- or under-represented are found in promoters of genes of a particular functional class, in order to better predict conditions or processes in which the sequence might act as a regulatory element. Taking advantage of the degeneracy in the TAL effector-DNA binding code to generate non-native EBE sequences may enable the rational design of a promoter recognized by a desired suite of TAL effectors without the inclusion of endogenous cis elements. Finally, making the promoter modifications in situ using engineered nucleases such as TALENs (Bogdanove & Voytas, 2011) may help guard against position effects on expression.

Acknowledgements

The authors thank T. Li, B. Yang, J. Baller, and S. Kamoun for helpful discussion, E. Braun for infected soybean leaves, B. Yang and C. Schmidt for TAL effector clones, Z. Yin for IRBB27, F. White for ME1, B. Yang for several Xoo strains, C. Song, R. Sonti, C. Vera-Cruz, and V. Verdier for Xoc strains, and M. Moscou,, K. Thilges, and M. Peterson for technical assistance. This work was supported by the NSF (award 0820831).

References


Gu KY, Tian DS, Qiu CX, Yin ZC. 2009. Transcription activator-like type III effector AvrXa27 depends on OsTFIIA gamma 5 for the activation of Xa27 transcription in rice that triggers disease resistance to Xanthomonas oryzae pv. oryzae. Molecular Plant Pathology 10(6): 829-835.


**Supporting Information**

Supplementary Fig. S1., which contains the complete sequences of the UXO, XOC, XOO, and RAN constructs, is available in the online version of the article.

![Figure S2](image-url)  
**Figure S2.** Specific activation of Xa27 in UXO-1 plants by each of the corresponding TAL effectors individually. Vertical bars represent relative abundance of Xa27 mRNA, measured by qPCR, 48 hours after inoculation with Xag EB08 expressing one of the seven corresponding Xoo or Xoc TAL effectors, or as controls for specificity, the Xoo TAL effector AvrXa10 or the Xoc TAL effector Tal1c, as indicated, relative to mock inoculated plants. Each is the average of three technical replicates of at least three biological replicates. Capped vertical lines show standard error. Asterisks indicate values significantly greater than 1.0 as determined by two-tailed, heteroscedastic t-tests. Relative fold-change was calculated by the $2^{-\Delta\Delta Ct}$ method. Values for the seven corresponding effectors are shown in the order of their EBEs in UXO (left to right, 5' to 3').
<table>
<thead>
<tr>
<th>Pathogen</th>
<th>Effector</th>
<th>Target Gene</th>
<th>Target Locus ID/Accession</th>
<th>RVD/EBE alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>PXO99</td>
<td>PthXo1</td>
<td>Xa13</td>
<td>LOC_Os08g42350</td>
<td>NN HD NI HG HD NG N* HD HD HD NG NI NG NG NI NI NG Ni Ni Ni Ni Ni N* NS NS N*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T G C A T C T C C C C C T A C T G T A C A C C C A C</td>
</tr>
<tr>
<td>PXO99</td>
<td>PthXo6</td>
<td>TFX1</td>
<td>LOC_Os09g29620</td>
<td>Ni HG Ni NN NN NN NN NN NN NN HD HD HG HD Ni N* NS Ni Ni HG HD NS NS NG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T A T A A A A G G C C C C T C A C A A A C C C A T</td>
</tr>
<tr>
<td>PXO99</td>
<td>Tal9a</td>
<td>Hen1</td>
<td>LOC_Os07g06970</td>
<td>HD HD HD NG N* NN HD HD N* Ni Ni NN HD HI ND HD Ni HD NG NG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T C C C C T T C C C C T A A A A C C C C A C T T</td>
</tr>
<tr>
<td>PXO99</td>
<td>AvrXa27</td>
<td>Xa27</td>
<td>AY986492.1</td>
<td>Ni NN N* NG NS NN NN NN NN Ni Ni N* HD HD Ni NG NG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T A G A A G A A G A G A C C C A T A</td>
</tr>
<tr>
<td>BLS256</td>
<td>Tal4a</td>
<td>N/A</td>
<td>LOC_Os03g33780</td>
<td>Ni NG Ni NG Ni NG NN NN NN NN NG HD NN NN NG HD HD NN NS NN NN HD HD HG HG HG HG HG HG HG HG HG HG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T A T A T A T C T C G G G T C A G G C C A G G C C</td>
</tr>
<tr>
<td>BLS256</td>
<td>Tal4c</td>
<td>N/A</td>
<td>LOC_Os06g37080</td>
<td>NS HG Ni NG NS NN ND N* NG N* HN NN HD NS Ni Ni NN HD HD HG HG HG HG HG HG HG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T A T A A A A C C T G G A C A A A G C C T C T C T T</td>
</tr>
<tr>
<td>BLS256</td>
<td>Tal2g</td>
<td>N/A</td>
<td>LOC_Os05g34600</td>
<td>NN NN ND N* NS NS YG Ni SN HD HD HG HD NS NS HD NG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T A A C C A A A A A G C C T C C C A C T</td>
</tr>
</tbody>
</table>

1 Effector sequences taken from PXO99 and BLS256 genome sequences (GenBank accessions NC_010717 and NC_017267, respectively).
2 N* indicates a 33 amino acid repeat in which the 13th residue is missing relative to a 34 amino acid repeat.
3 Locus ID from the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/)
4 Genbank accession.
**Table S2. Relative abundance of Xa27 mRNA in mock-inoculated transgenic lines compared to IRBB7.**

<table>
<thead>
<tr>
<th>Line</th>
<th>Fold difference</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>UXO-1</td>
<td>1.14</td>
<td>0.43</td>
</tr>
<tr>
<td>UXO-3</td>
<td>1.28</td>
<td>0.00</td>
</tr>
<tr>
<td>XOO</td>
<td>1.11</td>
<td>0.19</td>
</tr>
<tr>
<td>RAN-1</td>
<td>6.01</td>
<td>0.01</td>
</tr>
<tr>
<td>RAN-2</td>
<td>3.51</td>
<td>0.03</td>
</tr>
<tr>
<td>XOC-2</td>
<td>7.79</td>
<td>0.04</td>
</tr>
</tbody>
</table>

1 RNA was sampled at 48 hours after mock inoculation and Xa27 mRNA levels assessed by qPCR. Each value is the average of three technical replicates of at least three biological replicates and was calculated by the $2^{-\Delta\Delta Ct}$ method.

2 P values were determined by two-tailed, heteroscedastic t-tests.
CHAPTER 6. THE ROLE OF TRYPTOPHAN 232 IN TAL EFFECTOR DNA BINDING

A paper to be submitted to PLoS One

Erin L. Doyle\textsuperscript{1}, Colby G. Starker\textsuperscript{2}, Zachary L. Demorest\textsuperscript{2}, Aaron Hummel\textsuperscript{1}, Dan Nettleton\textsuperscript{3}, Phillip Bradley\textsuperscript{4}, Barry L. Stoddard\textsuperscript{5}, Daniel F. Voytas\textsuperscript{2}, and Adam J. Bogdanove\textsuperscript{1,6}

Abstract

TAL effectors are customizable DNA binding proteins that can easily be targeted to bind to almost any sequence of interest. Consequently, custom TAL effector-based constructs have rapidly become popular tools for targeted gene regulation and genome engineering. Naturally occurring TAL effector binding sites are nearly all directly preceded by a thymine (T) at the 5’ end of the binding site, and custom TAL effector constructs have typically been designed with this constraint. Structural information suggests that the requirement for T is encoded by a tryptophan residue (W232) in the cryptic -1st repeat of the protein that exhibits Van der Waal’s contacts with the T. We generated TAL effectors with all possible single amino acid substitutions for W232 and showed that many substitutions alter or relax the specificity for T at this position but also reduce the TAL effector’s activity and binding affinity. We were unable to identify any substitutions with consistently better activity that W232 across different TAL effectors and in different types of assays, suggesting previously unknown complexity and context dependence of the interaction between this region of the protein and the DNA preceding the target site. TAL effectors with tryptophan or arginine at position 232 were predicted to have the most favorable DNA binding energies. Notably, TAL effector-like proteins from Ralstonia solanacearum have and arginine at this position, suggesting that TAL effectors and Ralstonia TAL-like effectors have naturally selected the optimal residues at this position for binding.

\textsuperscript{1} Department of Plant Pathology and Microbiology, Iowa State University, Ames, IA, 50011, USA
\textsuperscript{2} Department of Genetics, Cell Biology & Development and Center for Genome Engineering, 321 Church Street SE, University of Minnesota, Minneapolis, MN 55455, USA
\textsuperscript{3} Department of Statistics, Iowa State University, Ames, IA, 50011, USA
\textsuperscript{4} Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N. M1-B514 Seattle, WA 98109, USA
\textsuperscript{5} Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N. A3-025 Seattle, WA 98019, USA
\textsuperscript{6} Current address: Plant Pathology and Plant-Microbe Biology, 334 Plant Science, Cornell University, Ithaca, NY 14850 USA
Therefore, engineering enhanced TAL effectors with high activity levels that bind to sites not preceded by T may require altering portions of the TAL effector other than W232.

Introduction

Transcription activator-like (TAL) effectors from the plant-pathogenic bacterial genus *Xanthomonas* have generated considerable interest as tools for targeted control of gene expression and site-specific genome engineering. In nature, TAL effectors function as transkingdom transcriptional activators that are deployed by *Xanthomonas* spp. to manipulate host plant gene expression. During infection, the TAL effectors are secreted from the pathogen into the host plant cell via the type three secretion system (T3SS) and localized to the nucleus. Once there, they bind to effector-specific DNA sequences to activate transcription of host plant genes necessary for bacterial multiplication and survival (1,2). Each TAL effector’s specific binding site is determined by its central repeat region (CRR), which is composed of a variable number of tandem repeats 34-35 amino acids in length. The repeats are nearly identical with repeat-to-repeat variation occurring primarily at amino acids 12 and 13 (termed the repeat-variable diresidue or RVD). The sequence of RVDs corresponds directly to the DNA binding site, with each repeat/RVD forming base-specific contacts with single binding site nucleotide (3-6). Discovery of this TAL effector-DNA binding code has enabled prediction of binding sites and targets for individual TAL effectors (7). It has also made it possible to engineer custom TAL effectors to bind to novel DNA sequences for a variety of applications.

Custom TAL effector-based fusion proteins can be targeted to bind to novel sequences by assembling a CRR with the appropriate RVD sequence that corresponds to the intended target site. Such custom TAL effectors have been used to activate transcription of specific genes, using either the native TAL effector activation domain or replacing it with the VP16 activation domain from herpes simplex virus or its tetrameric derivative VP64 (8-11). Targeted gene repression using custom TAL effectors has also been demonstrated, either by simply removing the activation domain (yeast) or replacing it with a repressor domain (12-14). The most widely used application of TAL effectors are TAL effector nucleases, or TALENs, which use TAL effector CRRs fused to the catalytic domain of the FokI endonuclease to create precisely targeted double strand breaks (DSBs) in the DNA (8,15-17). In eukaryotic cells, the DSBs are subsequently repaired by one of two pathways. Non-homologous end joining (NHEJ) frequently results in insertions or deletions, and can be used to disrupt a gene and created targeted gene knockouts. Homologous recombination (HR) inserts a repair template of similar sequence at the break site. By supplying a template encoding alterations to the targeted sequence, HR can
be used to create site-specific sequence modifications ranging from changing a few bases to inserting an entire gene (18-20). TALEN-mediated NHEJ and HR have been demonstrated in a wide variety of cell types and organisms. Additionally, custom TALENs have proven to be easier to design than other sequence-specific nucleases, and a variety of kits are readily available for their construction (reviewed in 21,22). Therefore, custom TAL effectors and TALENs have become key tools for genetic engineering and other biotechnology applications.

Because the TAL effector repeats appear to behave modularly with no neighbor or context effects, and can therefore be assembled in any order, custom TAL effectors can be readily targeted to almost any sequence in the genome (5,6). However, the design and targeting of custom TAL effectors and TALENs is constrained by a requirement for the TAL effector binding site to be directly preceded by a thymine (T) at the 5’ end (known as the 0th position). The 0th position T is found in nearly all known TAL effector binding sites in nature (3,4). The single exception is the target of TalC from Xanthomonas oryzae pv. oryzae (Xoo) strain BAI3, which is preceded by a 5’ cytosine at position 0 (23). Replacing the 0th position T with another base has been shown to dramatically reduce or eliminate TAL effector-driven activation of reporter genes as well as DNA binding (4,24,25). Functional TALENs targeting sites preceded by other nucleotides have been reported, but activity was not compared to equivalent targets with a T (8,26,27), and most custom TAL effectors have typically been designed to target binding sites preceded by a T. Thus, the presence of a T preceding the TAL effector binding site appears to be an important requirement for optimal TAL effector activity and represents a constraint on the design of custom TAL effectors and TAL effector based constructs.

A structural explanation was revealed by the structure of Xoo TAL effector PthXo1 bound to its DNA target. This structure included two previously unknown cryptic repeats (designated the 0th and -1st repeats) located N-terminal to the CRR which were similar in structure but dissimilar in sequence to the canonical repeats encoding the binding site. Residue 232 of the protein, a tryptophan located in the -1st repeat (tryptophan 232 or W232), was shown to be in close proximity to the 0th position T and to form energetically favorable, base-specific Van der Waal’s contacts with the methyl group of the T (5). Additionally, W232 and its surrounding residues are highly conserved in Xanthomonas TAL effector sequences, consistent with W232 playing a significant role by encoding the T at the 0th position of the binding site. TAL effector truncation studies have found that the presence of the -1st repeat is necessary for TAL effector binding and function (9,28). Therefore, it is not possible to engineer TAL effectors that have no requirement for a T at the 0th position of the binding site by removing this portion of the protein.
Here we report a series of experiments testing if the requirement for the 0th position T is encoded by W232. We also sought to enhance TAL effector-DNA targeting capacity by identifying substitutions for W232 that would relax or alter specificity for the 0th position T while retaining activity levels similar to wild-type TAL effectors. We constructed a series of TAL effectors with single amino acid substitutions for W232 and tested their activity and binding affinities on targets preceded by A, C, G, or T. None of the substitutions improved on the native W232 by consistently relaxing specificity for the 0th position T and having a high level of activity. We discovered an arginine in the analogous position of Ralstonia TAL-like effectors (RTLs) that, in the context of an RTL-TAL effector chimera, is similarly critical for DNA binding. Computational analysis predicted that TAL effectors with tryptophan or arginine at position 232 had more favorable binding energies than other substitution variants. This suggests that TAL effectors and RTLs have evolved at this position for optimal DNA binding. Although we found substitutions for W232 altered TAL effector activity and specificity for the 0th position T, these effects varied depending on the TAL effector repeat sequence, the N terminal portion of the TAL effector, and across assays. These context–based effects suggest that the cryptic 0th repeat, other portions of the N terminus, and other regions of the protein may be involved in specifying the 0th position T. Therefore, development of TAL effectors with broader DNA targeting capacities will likely require re-engineering multiple positions in the cryptic repeats and the N terminus.

Materials and Methods

Construction of TAL effector -1 repeat variants

To generate the PthXo1 TAL effector variants for GUS assays, the SphI fragment containing the repeat region of the gene encoding TAL effector PthXo1 (clone obtained from Bing Yang, Iowa State University) was ligated into the SphI site of plasmid pCS466 to create plasmid pAH103, which encodes a full length TAL effector with the PthXo1 CRR. pCS466 is a Gateway entry plasmid containing a truncated form of the Xanthomonas oryzae pv. oryzicola (Xoc) tal1c gene, from which the SphI fragment containing the DNA binding domain has been removed (29). To make single amino acid substitutions for W232, the region of pAH103 surrounding W232 was amplified via PCR using a forward primer p885 (5'-CGTCTGGCAACAGGCTCGCCGCACGC-3') crossing the NotI site, and reverse primers crossing the Stul site and introducing both a substitution for W232 and a silent XhoI site to facilitate screening (including arginine: 5'-CGTGAGGAAGCCTCCAGGGCTCGAGCGCCGGATCGCTGTTTGCGC-3', asparagine: 5'-CGTGAGGAAGCCTCCAGGGCTCGAGCGCCGGATCGCTGTTTGCGC-3', asparagine: 5'-CGTGAGGAAGCCTCCAGGGCTCGAGCGCCGGATCGCTGTTTGCGC-3', asparagine: 5'-CG})
CGTGAGCAAGGCTCCAGGCTCGACGCGCCGAATTCTGTTTGGCCG-3', glutamine: 5'-
CGTGAGCAAGGCTCCAGGCTCGACGCGCCGAATTCTGTTTGGCCG-3', proline: 5'-
CGTGAGCAAGGCTCCAGGCTCGACGCGCCGAATTCTGTTTGGCCG-3', and threonine: 5'-
CGTGAGCAAGGCTCCAGGCTCGACGCGCCGAATTCTGTTTGGCCG-3'). The same
forward primer was used for all substitutions. Primers for all substitutions are listed in
Supplementary Table 1. PCR products carrying the W232 substitutions were digested with NotI
and Stul (both New England Biolabs, Inc.) and ligated back into the pAH103 backbone between
the NotI and Stul sites using standard techniques to create genes for full length TAL effectors
with the PthXo1 repeat sequence and W232 substitutions.

To generate pTAL868 variants, a TALEN containing the pTAL868 repeat sequence (Figure
1) was constructed using our Golden Gate method (30), and digested with restriction enzymes
Stul and AatII (both New England Biolabs, Inc.). The 1.7kb fragment containing the repeat
region was purified from an agarose gel. pAH103 variants with W232 substitutions described
above were digested with Stul and AatII and the 4.3kb backbone fragments containing the TAL
effect N and C termini were recovered. TAL868 repeats were ligated into these pAH103-
based backbones containing the W232 substitutions. All TAL effectors (PthXo1 and TAL868
variants) were recombined into the Gateway binary vector (31) pGWB5 using Gateway LR
Clonase II (Life Technologies) according to the manufacturer's instructions.

Construction of chimeric TAL effector RScPthXo1

The Ralstonia TAL-like effector N-terminal sequence was PCR-amplified with Phusion DNA
polymerase (Thermo Fisher Scientific) from GMI1000 genomic DNA (gift of D. Gross, University
of Georgia) with forward primer p1178 (5’-
TTGCATGTAATAGGAGTGCCACCAGAGAATAGGCAATCAAG-3’; which added sequence
upstream of the start codon to match that of the Xanthomonas TAL effector expression
constructs) and p1179 (5’-GAGACTCGTCTCGGCCACGCTGAGCTTCC-3’; which added a
downstream Esp3I restriction site for cloning). This amplicon was tailed with 3’ adenine residues
using Taq Polymerase (Thermo Fisher Scientific) and cloned into the pCR8/GW/TOPO TA
vector (Life Technologies). The resulting plasmid was digested with Esp3I and EcoRV
restriction enzymes (both New England Biolabs, Inc.) and ligated to the BanII/EcoRV restriction
enzyme fragment of pAH103 to produce the full-length, Ralstonia/Xanthomonas chimeric TAL
effector in the Gateway entry vector, designated pAH410. Each -1st repeat amino acid
substitution variant of the chimeric effector was generated by PCR-amplifying a portion of the
Ralstonia N-terminus using forward primer P1178 and each of five reverse primers that
generated five different amino acid substitutions at the -1 repeat (asparagine: 5' - CGGGTAGCAGCGCTGAGCGCCAGGTCAACCCGATTTGCTGC-3', glutamine: 5' - CGGGTAGCAGCGCTGAGCGCCAGGTCAACCCGATTTGCTGC-3', proline: 5' - CGGGTAGCAGCGCTGAGCGCCAGGTCAACCCGATTTGCTGC-3', threonine: 5' - CGGGTAGCAGCGCTGAGCGCCAGGTCAACCCGATTTGCTGC-3', and tryptophan: 5' - CGGGTAGCAGCGCTGAGCGCCAGGTCAACCCGATTTGCTGC-3'). PCR amplifications were done using Phusion DNA polymerase (New England Biolabs, Inc.) Each amplicon was digested with Afel and Bsai restriction enzymes (both New England Biolabs, Inc.) and cloned into the Afel/Bsai restriction enzyme digested pAH410. The chimeric effector and amino acid substitution variants were transferred to the binary expression vector pGWB5 (31) via the Gateway LR II Clonase enzyme kit (Life Technologies) to create the final chimeric TAL effector construct and RScPthXo1 and -1 repeat variants for use in GUS assays.

Construction of GUS reporter plasmids

For GUS reporter constructs, a 343bp region of the Bs3 promoter was PCR amplified as previously described (24). The PCR product was ligated into the Gateway vector pCR8/GW/TOPO TA vector (Life Technologies) and site-directed mutagenesis was used to introduce an AscI site upstream of the naturally occurring binding site for TAL effector AvrBs3. This modified promoter was recombined into the Gateway GUS reporter vector pGWB3 (31) upstream of the GUS gene using LR Clonase II (Life Technologies). Single-stranded DNA oligos containing the naturally occurring binding site for PthXo1 (UptPthXo1) or the binding site corresponding to TAL868 (Upt868) directly preceded by A, G, C, or T at the 0th position and with AscI-compatible ends were annealed and ligated into the AscI site. Oligos used were 5'-CGCGTGCATCTCCCTCTACTGTACACCAC-3' and 5'-CGCGGTGGTGTAGCTAGGGGGAGATGCA-3' for UptPthXo1 (T); 5'-CGCGAGCATCTCCCTCTACTGTACACCAC-3' and 5'-CGCGTGGTGTAGCTAGGGGGAGATGCA-3' for UptPthXo1 (A); 5'-CGCGGGCATCTCCCTCTACTGTACACCAC-3' and 5'-CGCGGTCATCTCCCTCTACTGTACACCAC-3' for UptPthXo1 (C); 5'-CGCGGGCATCTCCCTCTACTGTACACCAC-3' and 5'-CGCGGTCATCTCCCTCTACTGTACACCAC-3' for UptPthXo1 (G); 5'-CGCGGTCATCTCCCTCTACTGTACACCAC-3' and 5'-CGCGGTCATCTCCCTCTACTGTACACCAC-3' for UptPthXo1 (T); 5'-CGCGGTCATCTCCCTCTACTGTACACCAC-3' and 5'-CGCGGTCATCTCCCTCTACTGTACACCAC-3' for UptPthXo1 (A); 5'-CGCGGTCATCTCCCTCTACTGTACACCAC-3' and 5'-CGCGGTCATCTCCCTCTACTGTACACCAC-3'.
CGCGCAGTTAATGGAAGCT-3’ and 5’-CGCGAGCTTCCATTAACGT-3’ for Upt868 (C); and 5’-CGCGGACGTTAATGGAAGCT-3’ and 5’-CGCGAGCTTCCATTAACGT-3’ for Upt868 (G).

Quantification of GUS activity

To measure TAL effector activity, TAL effector and reporter constructs were transformed into Agrobacterium tumefaciens GV3101. Cells were grown overnight and diluted to OD_{600} = 0.8 in infiltration buffer (10mM MgCl₂, 5mM MES pH 5.3, 200µM acetosyringone). Cells carrying TAL effector constructs and cells carrying GUS reporter constructs were mixed 1:1 and infiltrated into the leaves of 6-8 week old Nicotiana benthamiana plants using a needleless syringe, and infiltrated areas were marked. After 48 hours GUS measurements were carried out as described (32): Infiltrated leaf disks were ground in 300µL extraction buffer (50mM sodium phosphate buffer pH 7.0, 10mM EDTA, 10mM β-mercaptoethanol, 0.1% Triton X-100, 0.1% SDS) and centrifuged at 3500rpm for 10 minutes. 10µL supernatant was mixed with 90µL assay buffer (extraction buffer supplemented with 10mM 4-methyl-umbelliferyl-β-D-glucuronide) and incubated at 37°C for 1 hour. Following incubation, 10µL of the reaction was stopped by adding 90µL 0.2M sodium carbonate. Fluorescence was read at 360nm (excitation) and 460nm (emission). Protein amounts were quantified by Bradford assay (Biorad).

Paired T-tests

GUS assays were conducted in multiple independent experiments over several months. Because results varied across experiments, we used paired t-tests to identify significant differences in activity due to W232 substitutions and to estimate effect sizes. Each TAL effector/target combination was considered to be a treatment. GUS measurements for two treatments taken from the same plant on the same day were considered to be paired samples. To test for significant differences between two treatments, the log-transformed paired measurements were subtracted. The set of paired differences for the two treatments were tested using the R function t.test to test the null hypothesis that the average difference between the two treatments was equal to zero. Estimated effect sizes from the t.test function were converted to fold changes.

Construction and testing of TALENs

We used a yeast-based single-strand annealing assay for testing TALEN function that has been previously described (30,33) Briefly, a mating-type cells containing TALEN expressing plasmid were mated to α mating-type cells containing target plasmids. Cleavage of the target plasmid results in reconstitution of a functional lacZ gene, providing a quantitative measurement
of TALEN activity. All data are normalized to a wild type N-terminus TALEN on the target with the 5’ thymine and are the result of two independent experiments with a minimum of five replicates.

Yeast expression vectors were made by amplifying a fragment of the N-terminus of the TAL effector present in pZHY500 with forward primer 5'-ggacgcaagtggtgctagatggtgg-3' for all reactions and a reverse primer incorporating the desired changes (asparagine: 5’-cctccagggcgctgcggcgggaATTctgtttgccgacgcc-3’; proline: 5’-cctccagggcgctgcggcgggaTGGctgtttgccgacgcc-3’; glutamine: 5’-cctccagggcgctgcggcgggaTGTctgtttgccgacgcc-3’; arginine: 5’-cctccagggcgctgcggcgggaTCGctgtttgccgacgcc-3’; and threonine: 5’-cctccagggcgctgcggcgggaTGTctgtttgccgacgcc-3’). PCR products were digested with XbaI, StuI and cloned into the XbaI, StuI sites of pZHY500 or pZHY501 (34). DNA binding repeats were cloned into the resulting vectors as described (35).

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). Variants with substitutions at W232 were created by site-directed mutagenesis (Quickchange II, Agilent) using the same primers described above for the TALEN yeast assays. An array of 15 repeats with the RVD sequence [NI HD NN NG NI NI NG NN NN NI NI NN HD NG], designated as TAL868, was then cloned into each of these vectors using the Golden Gate method as previously described (30). Expression constructs encoding the TAL effector proteins were then transformed into Rosetta (DE3) pLysS cells (EMD Millipore) and selected on media containing carbenicillin (50 µg/ml) and chloramphenicol (30 µg/ml). 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 MgCl₂, 130 µM CaCl₂, 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 (10 U/mL), clarified by centrifugation (21,000× 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 blue to determine the purity of the samples.

**Electrophoretic mobility shift assay (EMSA)**

Double stranded DNA substrates were prepared by annealing fluorescently tagged complementary oligos. Sequences for substrates used were 5′- AACGTTAATGGAAGCT-3′ for Upt868A, 5′- CACGTTAATGGAAGCT-3′ for Upt868C, 5′- GACGTTAATGGAAGCT-3′ for Upt868G, 5′- TACGTTAATGGAAGCT-3′ for Upt868T, and 5′- TCGACGCTCAGGCAAC-3′ 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 μM DTT, 15 ng/μL salmon sperm DNA, 30 ng/μL BSA) at varying concentrations with a fixed concentration of the labelled 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 labelled substrate was then performed on a fluorescent scanner (Storm 860, Molecular Dynamics).

**Structural predictions**

DNA binding specificity calculations were performed for all 20 canonical amino acid substitutions at position W232 using the molecular modelling package Rosetta (36). The structure of the TAL effector PthXo1 bound to a natural target site (PDB ID: 3UGM) was used as a modelling template. We explored two levels of conformational sampling, either keeping the protein backbone completely fixed or allowing limited flexibility in the neighborhood of the mutation. Each mutant was simulated in complex with 16 different DNA target sites that sampled all possible DNA sequences at positions 0 (the canonical T position) and -1 (also contacted by W232 in the crystal structure). At the end of each simulation, we computed a binding energy for the final structure by taking the difference between the energy of the complex and the energies of the unbound partners, allowing limited conformational relaxation in the unbound state prior to computing the unbound energies.

**Results**

**Single amino acid substitutions for W232 alter target specificity and activity of the TAL effector PthXo1**

To test whether W232 encodes TAL effectors’ specificity for T at the 0th position of the binding site, we generated full length TAL effector constructs containing the repeat region of *Xanthomonas oryzae* TAL effector PthXo1 (see Figure 1b) with the wild type W232 as well as all 19 possible single amino acid substitutions at this position. We then tested the ability of the
PthXo1 W232 substitution variants to activate transcription of a GUS reporter gene containing a PthXo1 binding site (UptPthXo1, for up-regulated by TAL effector PthXo1) preceded by adenine, cytosine, guanine, or thymine in their promoters (UptPthXo1 (A), UptPthXo1 (C), UptPthXo1 (G), and UptPthXo1 (T), respectively). PthXo1 variants and reporter constructs were co-delivered into *Nicotiana benthamiana* leaves using *Agrobacterium* and GUS activity was measured after 48 hours.

For initial screening, each PthXo1 W232 substitution variant was tested in at least one experiment. In each experiment, one to three TAL effector constructs were tested against each of the four corresponding UptPthXo1 target reporter constructs. In addition, each experiment included UptTALPthXo1 (T) with no TAL effector as a negative control, and UptPthXo1 (T) co-delivered with the PthXo1 W232 construct as a positive control. During a single experiment, each sample (a control or a TAL effector and target) was inoculated onto five different *N. benthamiana* plants, for a total of five different measurements for the sample. To facilitate comparisons across experiments, within each experiment data were normalized to the positive control. Data for each experiment were reported as the average of three leaf discs (high and low measurements were thrown out).

All 19 of the substitutions for W232 appeared to alter specificity for the 0\(^{th}\) position T (indicated by having the highest activity on a target preceded by A, C, or G, or having nearly equal activity on all targets) (Supplementary Figure 1). These alterations in specificity provide the first experimental evidence that W232 is involved in encoding the requirement for T preceding the binding site. However, many of the substitutions, including aspartate (D), cysteine (C), glutamate (E), isoleucine (I), leucine (L), lysine (K), methionine (M), and valine (V), reduced activity to less than half of the activity level of PthXo1 W232 on the target UptPthXo1 (T).

Because we were interested in identifying W232 substitutions that both altered or relaxed specificity for T at the 0\(^{th}\) position and maintained relatively high levels of activity, we selected the arginine (R), asparagine (N), glutamine (Q), proline (P), and threonine (T) substitutions to characterize in greater detail other assays and contexts.

**The effects of W232 substitutions on TAL effector activity and specificity for T at the 0\(^{th}\) position depend on the RVD sequence**

In our initial screening, we found that W232 substitutions altered the specificity for T preceding the binding site and the activity level of TAL effector PthXo1. Next, we sought to determine whether the effects on specificity and activity depend on the TAL effector’s RVD sequence. We constructed variants of a second TAL effector, TAL868 (see Figure 1), with W at
position 232 and five W232 substitutions and tested their ability to activate a GUS reporter gene driven by the corresponding target preceded by A, C, G, or T (Upt868A, Upt868C, Upt868G, and Upt868T, respectively). As with our initial screening of PthXo1 mutants, one to three TAL effector constructs were tested against each of the four corresponding Upt868 target reporter constructs, and each experiment included both Upt868 (T) with no TAL effector and Upt868 (T) co-delivered with the TAL868 W232 construct as controls. In a single experiment, each was inoculated onto five different N. benthamiana plants, for a total of five measurements for the sample. Each TAL868 W232 mutation construct was tested in at least three experiments. GUS assays were also repeated for the five W232 substitution variants of PthXo1.

We used paired t-tests to identify differences in GUS activity due to W232 substitutions (see Materials and Methods). For each TAL effector, we tested for significant differences between the wild type TAL effector (W232) on the target preceded by T and all other TAL effector-target combinations. We converted the estimated effect sizes to fold changes relative to either PthXo1 W232 or TAL868 W232 on the corresponding target preceded by T (Figure 1). We also tested for significant differences between each of the W232 TAL effector variants on the target preceded by T and each of the other three targets to determine if specificity had been relaxed or altered. Results of all t-tests are reported in Supplementary Table 2.

In general, W232 substitutions relaxed or altered specificity for targets with a T at the 0th position. However, TAL effector context affected base preference for many of these (Figure 1 and Supplementary Table 2). For example, in the context of PthXo1, the W232R substitution shifted specificity for the position 0 base to A or C (p<0.05). In contrast, the W232R substitution in the context of TAL868 appeared to relax specificity entirely (p<0.01). W232P was the only substitution to relax specificity in both TAL effectors (activity on targets preceded by A, C, and G were not significantly different than activity on the target preceded by T, p < 0.05). Therefore, our findings suggest that the effect of W232 substitutions on TAL effector binding site specificity and activity may depend on the TAL effector’s RVD sequence. None of the W232 substitution variants showed significantly higher activity on any of the targets than the wild type TAL effector on the target preceded by T.

**A chimeric RTL-TAL effector construct shows that effects of W232 substitutions depend on N terminus context**

To further examine the effects of context on the W232 substitutions, we tested the effects of single amino acid substitutions on the analogous position in Ralstonia TAL effector-like effectors (RTLs), which are found in the soil-borne bacterium Ralstonia solanacearum. RTLs possess N
and C termini with type III secretion signals and acidic activation domains similar to those of TAL effectors, although with some amino acid sequences differences. RTLs also possess a repetitive CRR composed of repeats 35 amino acids in length and differing in sequence from *Xanthomonas* TAL effector repeats (37,38). Although the RTL repeats possess RVDs typically not found in *Xanthomonas*, the RVD sequence has been shown to encode the RTL's DNA.
binding site (39). Currently, no RTL-targeted genes have been identified, and their role in plant disease is unknown.

Because of the many similarities between *Xanthomonas* TAL effectors and RTLs, we reasoned that the RTLs might possess cryptic repeats similar in structure and function to those of TAL effectors. We identified three RTL sequences in Genbank with complete coding sequences that included both the CRR and the entire N-terminal region upstream of the repeats (Supplementary Figure 2). Using ClustalW (40,41), we aligned the N terminal regions of the RTLs with the N terminal region of *Xoo* TAL effector PthXo1. The multiple sequence alignment showed that the RTL N termini have high sequence similarity to the N terminal region of PthXo1. Tryptophan 232 is replaced by an arginine in the RTL sequences, but the residues immediately on either side are conserved (Supplementary Figure 2). To further assess the similarity of the cryptic repeats, we generated secondary structure predictions for the N termini of the RTLs and PthXo1. For all of the proteins, the N termini form six helices, with four helices corresponding approximately to the -1<sup>st</sup> and 0<sup>th</sup> repeats of PthXo1. W232 (PthXo1) and R232 (RTLs) are located on a short loop between the two helices that make up the -1<sup>st</sup> repeat in each protein context (Supplementary Figure 3).

Because the sequence alignments and secondary structure predictions indicated that RTLs possess regions structurally similar to the -1<sup>st</sup> and 0<sup>th</sup> repeats of TAL effectors and present an arginine at the equivalent position to W232, we tested the effects of single amino acid substitutions for this arginine, which we refer to as R232 because it is located at the RTL position that aligns with W232, on the RTL’s specificity at the 0th position of the binding site. We constructed a chimeric TAL effector, RScPthXo1, by replacing the N terminal region immediately upstream of the CRR of our PthXo1 construct with the N terminal region of RTL RSc1815 (Figure 2). The chimeric TAL effector was co-delivered with each of the four UptPthXo1 GUS reporter constructs into *Nicotiana benthamiana*, and GUS activity was measured after 48 hours.

We found that the chimeric TAL effector with the wild-type N terminal R232 had the highest activity on the target preceded by G, and overall activity was significantly lower than the activity of PthXo1 with W232 on the target UptPthXo1 (T) (Figure 2). The specificity of RScPthXo1 R232 for targets preceded by G is different than the specificities observed for PthXo1 W232R or TAL868 W232R, suggesting that other N terminus amino acid sequence differences may affect the specificity for the 0<sup>th</sup> position T.

Next, we tested the activity of single amino acid substitution variants of RScPthXo1 on targets preceded by A, C, G, or T. All substitutions for R232, including R232W, eliminated
activity almost entirely (Figure 2). Therefore, the RTL amino acid at the position analogous to W232 of Xanthomonas TAL effectors also plays a critical role in determining the TAL effector’s binding site, and substitutions at this position dramatically decrease RTL activity. However, the differences in specificities in the context of TAL effectors or RTLs suggests that sequence differences in other parts of the proteins’ N termini, also play a role in specifying a preferred 0th position nucleotide preceding the binding site.

**W232 substitutions alter TALEN activity and specificity for the 0th position T**

Next, we asked if the effects of W232 substitutions on TAL effector activity and specificity for the 0th position T would be the same in the context of TALENs. We generated TALEN pairs with the same RVD sequence as TAL868 in each monomer (see Figure 1c), and created W232 substitutions variants. We tested their activity on targets preceded by A, C, G, or T in a yeast single-strand annealing assay (15). The assay is based on a target plasmid containing a variant of the LacZ gene with an internal sequence duplication that is disrupted by the TALEN target
sequence. The yeast strain carrying the target plasmid is mated with a yeast strain carrying the TALEN, and cleavage of the target by the TALEN restores LacZ function through single-strand annealing. Measured LacZ activity is used as a proxy for TALEN activity.

As expected, the TAL868 W232 TALEN had highest activity on the target Upt868 (T). Two of the W232 substitutions, W232P and W232R, each relaxed the preference for targets preceded by T, but also reduced TALEN activity to less than 60% of the W232 TALEN on Upt868 (T) (Figure 3). The W232T and W232N substitutions altered specificity to prefer targets preceded by G and C/T, respectively. Additionally, none of the W232 substitutions showed higher activity on any target than the wild type (W232) TALEN on Upt868 (T), although the W232Q TALEN had activity >85% (relative to W232 on Upt868 (T)) on the targets preceded by C, G, or T. Overall, the effects of W232 substitutions on activity and specificity observed in TALENs were not consistent with the changes observed in the context of the GUS assay.

Substitutions for W232 alter the TAL effector’s DNA binding affinity and specificity for targets preceded by A, C, G, or T

We next asked if the effects of W232 substitutions on TAL effector and TALEN activity and binding site specificity are directly related to changes in TAL effector-DNA binding affinity. We expressed and purified a TAL868 construct and five W232 substitution variants and tested their affinities for DNA targets preceded by A, C, G, or T in electrophoretic mobility shift assays (EMSAs) (Figure 4). The protein with the wild type W232 showed the strongest affinity for

Figure 3. Activity of TALENs with the TAL868 repeat sequence and W232 substitutions on targets preceded by A, C, G, or T at the 0th position. Upt868 (A), Upt868 (C), Upt868 (G), and Upt868 (T) indicate activity on the Upt868 target preceded by A, C, G, or T, respectively. Data are normalized so that the TAL868 W232 + Upt868 (T) is equal to 1.0. Error bars report the mean of all replicates ± 1 s.d. N=8 for all TAL868 W232 variants on all targets except W232 (N=15) and W232P (N=7).
Figure 4. Relative binding affinities of TAL868 W232 substitution variants for targets preceded by A, C, G, or T at the 0th position. A. Relative affinity of TAL868 (WT) and W232 substitution variants for the TAL868 target preceded by T, A, C, or G, or a scrambled target, demonstrated by electrophoretic mobility shift assay. For each TAL effector target combination, protein concentration increases from left to right. Bands across the bottom represent unbound DNA. The next bands up represent DNA bound by the TAL effector. The uppermost bands represent higher order complexes. DNA bound at lower protein concentrations indicates higher affinity. B. Relative affinity of TAL868 (WT) and W232 substitution variants for the TAL868 target preceded by T, A, C, or G, reported as the fraction of DNA bound by the protein. Fraction bound was estimated by band densitometric analysis of the images in Panel A. T, A, C, and G indicate the nucleotide at the 0th position of the target site. Scr, scrambled target.
Upt868 preceded by a T. The W232 substitution variants each showed a similar pattern, with the highest affinity for the Upt868 (T) target. However, their binding affinity for Upt868 (T) was reduced relative wild type TAL effector, and the preference for T was less pronounced. Our results suggest that in the context of TAL868, substitutions for W232 lower the TAL effector’s DNA binding affinity and reduce the TAL effector’s specificity for targets preceded by T. However, the changes in binding affinity and specificity did not match the changes in activity and specificity observed in the GUS and TALEN assays. We were unable to obtain sufficient soluble PthXo1 protein to assay effects of W232 substitutions on affinity in that context.

**Structural models of W232 substitutions predict W232 and W232R to have the most favorable binding energies**

We built structural models of all 20 amino acid substitutions at position W232, and calculated binding profiles to DNA target sites with all 16 nucleotide combinations at positions 0 and -1 (representative models are depicted in Figure 5). As described in Materials and Methods, we explored two levels of conformational flexibility, either keeping the protein backbone completely fixed or allowing limited flexibility in the neighborhood of the mutation. In the fixed-backbone simulations, the wild-type Tryptophan in complex with a T at position 0 yielded the highest predicted binding affinity overall (Figure 5a), followed by non-polar substitutions (isoleucine and valine) that maintained packing against the T0 methyl group. We did not see a strong dependence of binding energy on position -1, although some substitutions appeared to disfavor T at -1, likely due to steric clashes. We saw a different profile of binding energies in the flexible backbone simulations, with polar amino acids (arginine, followed by lysine, and glutamine) yielding the highest binding affinities, again generally preferring T at position 0. Examination of the structural models revealed that contacts to the DNA phosphate backbone predominated in these models, with protein backbone shifts allowing the mutated side chains to form hydrogen bonds to phosphate oxygens (Figure 5, panels B and D). In addition, several of the models show that the side chains of the amino acids at position 232 are in close proximity to, and may interact with, the base preceding the 0th position T (designated the -1 position, see Figure 5).

**Discussion**

We tested the effects of single amino acid substitutions for W232 on TAL effector activity, preference for a 0th position binding site nucleotide, and binding affinity. Located in the TAL effector’s cryptic -1st repeat, W232 packs tightly against the thymine residue that directly precedes most TAL effector binding sites in nature (at the so-called “0th position”) and forms
energetically favorable, base-specific Van der Waal’s contacts with this base (5). Therefore, W232 has been predicted to specify the 0\textsuperscript{th} position T. We found that single amino acid substitutions for W232 altered the TAL effector’s activity in a GUS reporter assay, and also altered the specificity of the TAL effector for targets preceded by T. This general reduction in activity and changes to target specificity were also observed in the context of TALENs and were supported by reductions in binding affinity as measured in EMSAs. Taken together, these results provide experimental evidence that W232 plays a role in encoding the requirement for T preceding the binding site.

Although our experimental results generally agreed that substitutions for W232 lowered TAL effector activity or binding affinity, specificity results from different assays did not always agree. In GUS assays of TAL868 W232T we observed low activity (relative to TAL868 W232 on the
target Upt868 (T)) and a very slight preference for targets preceded by A and T, while in TALEN yeast assays TAL868 W232T had a preference for Upt868 (C) and high activity on that target. GUS assays also showed relaxed specificity and relatively high activity for TAL868 W232R. However, in the context of TALENs, the W232R substitution showed relaxed specificity but very low activity. Additionally, EMSAs failed to detect any overall changes in target preference based on W232 substitutions, although the preference for targets preceded by T was less pronounced for all substitutions tested, suggesting some relaxation of target specificity.

There are several factors that may explain the discrepancies between assays. Despite careful standardization of conditions and leaf age, we found the GUS assay to be quite variable, both within experiments (across replicates) and across experiments performed on different days. Although we attempted to account for this inherent variability by using paired T-tests to test for effects on activity levels, the variability may have been too subtle to accurately assess changes in target preference. Another factor that may explain the differences in results across assays is that the W232 substitutions may in fact have differential effects on the different assays themselves. While the GUS assay used a full-length TAL effector protein, the TALEN assay used a TAL effector-nuclease fusion with truncations at the N and C termini, and the EMSAs used a similarly truncated TAL effector. The substitutions may have varying effects on the rates at which these slightly different proteins bind to and release their DNA targets. Similarly, the substitutions may cause structural changes that affect the assay readouts differently. For example, a W232 substitution might alter the TAL effector’s ability to interact with the host cell’s transcriptional machinery, affecting the outcome of the GUS assay but not the other assays.

In addition to disagreements between assays, we found that the effects of the W232 substitutions in the GUS assay varied depending on the TAL effector. These differences may be due to differences in the overall repeat sequence length and RVD sequence composition. Alternatively, the differences observed between array contexts may be due to specific differences in the repeat sequence near the cryptic -1st and 0th repeats. These differences may also be due to the fact that we tested PthXo1 against its naturally occurring target sequence, which contains several mismatches (positions where the RVD is not aligned with its preferred nucleotide) while TAL868 was tested on its perfect targets (no mismatches). One recent study found that the activity of TAL effector AvrBs3 (which contains several "non-standard" RVDs and has several mismatch positions) on its natural target was more sensitive to substitutions for the 0th position T than an engineered TAL effector designed to match the target (42). Therefore, the activity of TAL effector PthXo1, when paired with its natural target, may be more sensitive to
changes to W232. Mismatches, especially those near the N terminus, may interfere with the
ability of that portion of the protein to interact with and bind to the DNA.

We also identified and studied the -1st cryptic repeat of RTLs from Ralstonia solanacearum. Although RTLs are highly similar to TAL effectors in structure, numerous amino acid sequence differences are present throughout the proteins. These differences include an arginine at the analogous position to W232 in the RTL N terminus. In our experiments, a chimeric TAL effector composed of the native RTL N terminus (with R232) and the PthXo1 CRR preferred targets preceded by G. This differs from the Xanthomonas TAL effector W232R substitution variants, which had a preference for targets preceded by A or C (PthXo1) or no preference (TAL868) in GUS assays. This suggests that the sequence differences between the TAL effector and RTL N terminal regions surrounding W232 or R232 affect the specificity for the 0th position nucleotide, and that efforts to engineer TAL effectors with relaxed or altered 0th position specificities may require modifications throughout the entire TAL effector N terminus.

We also observed that the chimeric TAL effector (R232) had significantly lower activity than PthXo1, and that substitutions for R232 in the RTL N terminus essentially eliminated activity. This may simply be due to sequence differences between the RTL and TAL effector N termini, as discussed above, which result in lower activity. It is also possible that sequence and structure differences between the RTL and TAL effector N termini and CRRs caused the two pieces of the chimeric TAL effector to interact differently than an N terminus and CRR that are both derived from the same type of protein, potentially affecting the protein's binding affinity or activity levels.

Our goal for this study was to enhance TAL effector targeting capacity by identifying W232 substitutions that would alter specificity to prefer sites preceded by a base (or bases) other than T while retaining overall affinity and activity. Although a few substitutions showed high activity levels and relaxed specificities in a single assay, none of the substitutions tested behaved consistently in all assays. We also failed to identify any substitutions that had consistently higher activity levels than the native W232. Initially, we screened all 19 possible substitutions in a GUS assay, and then selected five substitutions to characterize more extensively, so substitutions we did not test might have met our criteria for high activity and broadened specificity. However, our conclusion that TAL effectors with W232 have the highest activity level is supported by binding energy predictions in our computational simulations, which found that TAL effectors with tryptophan or arginine (depending on the conditions of the simulation) at position 232 have the highest binding affinity. In nature, TAL effectors from Xanthomonas and Ralstonia TAL-like effectors have either a W (Xanthomonas) or R (RTLs) at this position,
suggesting that these two families of proteins have evolved to have the best possible affinities for their targets. Therefore, it may not be feasible to engineer TAL effectors to have broader DNA binding capacities by altering W232 or other positions in the -1\textsuperscript{st} repeat.

Binding site specificity and activity may also depend on other regions of the -1\textsuperscript{st} repeat, or on other cryptic repeats, whose role in DNA binding is still relatively unknown. Examination of the crystal structure of TAL effector PthXo1 bound to its DNA target shows that the 0\textsuperscript{th} and -1\textsuperscript{st} repeats pack tightly together and the loop of the cryptic 0\textsuperscript{th} repeat also appears to enter the major groove near the 0\textsuperscript{th} position T (5). Therefore, the 0\textsuperscript{th} repeat may also be involved in stabilizing binding or specifying a preference for the 0\textsuperscript{th} position T. This specificity may depend on the 0\textsuperscript{th} repeat directly, or on effects of the cryptic 0\textsuperscript{th} repeat on the -1\textsuperscript{st} repeat or on other regions of the N terminus. The structure also shows that W232 makes specific contacts with the base preceding the T at binding site position 0 (5). Therefore, it is possible that the preference for a 0\textsuperscript{th} position T depends on specific interactions between the TAL effector and other nucleotides near the T.

Thus, engineering TAL effectors to have robust, altered specificities may require changes at multiple positions to re-engineer the entire -1\textsuperscript{st} repeat, changes to the cryptic 0\textsuperscript{th} repeat and other regions of the TAL effector N terminus, or both. The availability of more TAL effector structures will help to shed light on the increasingly complex picture of the role of the N terminus and the cryptic repeats on TAL effector-DNA binding. Additionally, the development of more structures will improve computational modeling and structural predictions of TAL effector-DNA interactions, and will help to guide future efforts to engineer TAL effectors with broadened DNA binding capacities.

References


Supplementary Data

Supplementary Table S1. Primers used for pAH236 tryptophan 232 substitutions.

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>All²</td>
<td>p885</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232A</td>
<td>p886</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232F</td>
<td>p919</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232L</td>
<td>p920</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232I</td>
<td>p921</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232M</td>
<td>p922</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232S</td>
<td>p923</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232P</td>
<td>p924</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232T</td>
<td>p925</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232H</td>
<td>p926</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232G</td>
<td>p927</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232Q</td>
<td>p928</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232R</td>
<td>p929</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232D</td>
<td>p930</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232E</td>
<td>p931</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232C</td>
<td>p932</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232R</td>
<td>p933</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
<tr>
<td>W232G</td>
<td>p934</td>
<td>5'-CGTGGAGCAGCGTCTTGAGCCGACACGACG-3'</td>
</tr>
</tbody>
</table>

¹ Codons introducing W232 substitutions are highlighted in grey. The silent XhoI site is underlined.
² Used in all reactions along with one of the primers listed below in the table to produce the desired mutation.
Supplementary Table S2. Results of paired t-tests for significant effects of W232 substitutions on GUS activity.

<table>
<thead>
<tr>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>P-Value $^1$</th>
<th>T Statistic</th>
<th>DF $^2$</th>
<th>Estimate $^3$</th>
<th>95% CI $^4$</th>
<th>FC $^5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upt868 (A) + TAL868</td>
<td>Upt868 (T) + TAL868</td>
<td>0.000</td>
<td>-5.527</td>
<td>14</td>
<td>-0.519</td>
<td>-0.720 to 0.317</td>
<td>0.595</td>
</tr>
<tr>
<td>Upt868 (A) + TAL868 W232N</td>
<td>Upt868 (T) + TAL868</td>
<td>0.005</td>
<td>-3.736</td>
<td>9</td>
<td>-0.355</td>
<td>-0.570 to -0.140</td>
<td>0.701</td>
</tr>
<tr>
<td>Upt868 (A) + TAL868 W232P</td>
<td>Upt868 (T) + TAL868</td>
<td>0.001</td>
<td>-4.318</td>
<td>14</td>
<td>-0.383</td>
<td>-0.574 to -0.193</td>
<td>0.682</td>
</tr>
<tr>
<td>Upt868 (A) + TAL868 W232Q</td>
<td>TAL868</td>
<td>0.000</td>
<td>-5.481</td>
<td>14</td>
<td>-0.392</td>
<td>-0.545 to -0.238</td>
<td>0.676</td>
</tr>
<tr>
<td>Upt868 (A) + TAL868 W232R</td>
<td>TAL868</td>
<td>0.690</td>
<td>-0.405</td>
<td>19</td>
<td>-0.033</td>
<td>-0.206 to 0.139</td>
<td>0.967</td>
</tr>
<tr>
<td>Upt868 (C) + TAL868</td>
<td>Upt868 (T) + TAL868</td>
<td>0.000</td>
<td>-15.196</td>
<td>14</td>
<td>-0.978</td>
<td>-0.840 to -1.116</td>
<td>0.376</td>
</tr>
<tr>
<td>Upt868 (C) + TAL868 W232N</td>
<td>Upt868 (T) + TAL868</td>
<td>0.000</td>
<td>-6.663</td>
<td>9</td>
<td>-0.584</td>
<td>-0.386 to -0.643</td>
<td>0.558</td>
</tr>
<tr>
<td>Upt868 (C) + TAL868 W232P</td>
<td>Upt868 (T) + TAL868</td>
<td>0.001</td>
<td>-4.112</td>
<td>14</td>
<td>-0.422</td>
<td>-0.202 to -0.703</td>
<td>0.656</td>
</tr>
<tr>
<td>Upt868 (C) + TAL868 W232Q</td>
<td>TAL868</td>
<td>0.000</td>
<td>-8.369</td>
<td>14</td>
<td>-0.560</td>
<td>-0.416 to -0.560</td>
<td>0.571</td>
</tr>
<tr>
<td>Upt868 (C) + TAL868 W232R</td>
<td>Upt868 (T) + TAL868</td>
<td>0.140</td>
<td>-1.542</td>
<td>19</td>
<td>-0.091</td>
<td>-0.214 to 0.032</td>
<td>0.913</td>
</tr>
<tr>
<td>Upt868 (C) + TAL868 W232T</td>
<td>Upt868 (T) + TAL868</td>
<td>0.000</td>
<td>-7.184</td>
<td>14</td>
<td>-0.752</td>
<td>-0.527 to -1.087</td>
<td>0.471</td>
</tr>
<tr>
<td>Upt868 (G) + TAL868</td>
<td>Upt868 (T) + TAL868</td>
<td>0.000</td>
<td>-16.429</td>
<td>12</td>
<td>-0.960</td>
<td>-0.832 to -0.579</td>
<td>0.383</td>
</tr>
<tr>
<td>Upt868 (G) + TAL868 W232N</td>
<td>TAL868</td>
<td>0.158</td>
<td>-1.542</td>
<td>9</td>
<td>-0.235</td>
<td>0.110 to -0.516</td>
<td>0.791</td>
</tr>
<tr>
<td>Upt868 (G) + TAL868 W232P</td>
<td>Upt868 (T) + TAL868</td>
<td>0.014</td>
<td>-2.844</td>
<td>13</td>
<td>-0.293</td>
<td>-0.070 to -0.431</td>
<td>0.746</td>
</tr>
<tr>
<td>Upt868 (G) + TAL868 W232Q</td>
<td>Upt868 (T) + TAL868</td>
<td>0.007</td>
<td>-3.171</td>
<td>14</td>
<td>-0.257</td>
<td>-0.083 to -0.166</td>
<td>0.773</td>
</tr>
<tr>
<td>Upt868 (G) + TAL868 W232R</td>
<td>Upt868 (T) + TAL868</td>
<td>0.905</td>
<td>0.121</td>
<td>19</td>
<td>0.010</td>
<td>0.186 to 1.010</td>
<td></td>
</tr>
<tr>
<td>Upt868 (G) + TAL868 W232T</td>
<td>TAL868</td>
<td>0.000</td>
<td>-5.804</td>
<td>14</td>
<td>-0.755</td>
<td>-0.476 to -1.035</td>
<td>0.470</td>
</tr>
<tr>
<td>Upt868 (G) + TAL868 W232N</td>
<td>Upt868 (T) + TAL868</td>
<td>0.120</td>
<td>-1.715</td>
<td>9</td>
<td>-0.152</td>
<td>0.049 to -0.655</td>
<td>0.859</td>
</tr>
<tr>
<td>Upt868 (T) + TAL868</td>
<td>Upt868 (T) + TAL868</td>
<td>0.000</td>
<td>-5.227</td>
<td>14</td>
<td>-0.464</td>
<td>-0.274 to -0.628</td>
<td></td>
</tr>
</tbody>
</table>

---

1 P value for the t-tests testing the null hypothesis that the expected value of ln(Treatment 1) - ln(Treatment 2) = 0
2 Degrees of freedom
3 Estimate of the average difference ln(Treatment 1)-ln(Treatment 2). Estimate < 0 indicates that the GUS activity of Treatment 1 is lower than the GUS activity of Treatment 2. Estimate > 0 indicates that the GUS activity of Treatment 1 is higher than the GUS activity of Treatment 2.
4 95% Confidence Interval for the Estimate.
5 Fold Change of Treatment 1/Treatment 2. FC = e^Estimate. FC = 1 indicates the two treatments have equal GUS activity. FC > 1 indicates that Treatment 1 has higher GUS activity that Treatment 2. FC < 1 indicates that Treatment 2 has higher GUS activity that Treatment 1.
### Supplementary Table S2. Continued.

<table>
<thead>
<tr>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>P-Value</th>
<th>T Statistic</th>
<th>DF</th>
<th>Estimate</th>
<th>95% CI</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upt868 (T) + TAL868 W232Q</td>
<td>Upt868 (T) + TAL868</td>
<td>0.020</td>
<td>-2.615</td>
<td>14</td>
<td>-0.170</td>
<td>-0.310 to -0.031</td>
<td>0.843</td>
</tr>
<tr>
<td>Upt868 (T) + TAL868 W232R</td>
<td>Upt868 (T) + TAL868</td>
<td>0.549</td>
<td>0.610</td>
<td>19</td>
<td>0.050</td>
<td>0.220 to 0.592</td>
<td>1.051</td>
</tr>
<tr>
<td>Upt868 (T) + TAL868 W232T</td>
<td>Upt868 (T) + TAL868</td>
<td>0.000</td>
<td>-5.140</td>
<td>14</td>
<td>-0.418</td>
<td>-0.244 to -0.484</td>
<td>0.658</td>
</tr>
<tr>
<td>Upt868 (A) + TAL868 W232N</td>
<td>Upt868 (T) + TAL868</td>
<td>0.054</td>
<td>-2.209</td>
<td>9</td>
<td>-0.203</td>
<td>-0.410 to 0.005</td>
<td>0.817</td>
</tr>
<tr>
<td>Upt868 (C) + TAL868 W232N</td>
<td>Upt868 (T) + TAL868</td>
<td>0.002</td>
<td>-4.309</td>
<td>9</td>
<td>-0.431</td>
<td>-0.658 to 0.205</td>
<td>0.650</td>
</tr>
<tr>
<td>Upt868 (G) + TAL868 W232N</td>
<td>Upt868 (T) + TAL868</td>
<td>0.555</td>
<td>-0.613</td>
<td>9</td>
<td>-0.082</td>
<td>-0.386 to 0.221</td>
<td>0.921</td>
</tr>
<tr>
<td>Upt868 (A) + TAL868 W232P</td>
<td>Upt868 (T) + TAL868</td>
<td>0.353</td>
<td>0.962</td>
<td>14</td>
<td>0.081</td>
<td>-0.100 to 0.262</td>
<td>1.084</td>
</tr>
<tr>
<td>Upt868 (C) + TAL868 W232P</td>
<td>Upt868 (T) + TAL868</td>
<td>0.709</td>
<td>0.380</td>
<td>14</td>
<td>0.042</td>
<td>-0.196 to 0.280</td>
<td>1.043</td>
</tr>
<tr>
<td>Upt868 (G) + TAL868 W232P</td>
<td>Upt868 (T) + TAL868</td>
<td>0.115</td>
<td>1.688</td>
<td>13</td>
<td>0.195</td>
<td>-0.055 to 0.445</td>
<td>1.216</td>
</tr>
<tr>
<td>Upt868 (A) + TAL868 W232Q</td>
<td>Upt868 (T) + TAL868</td>
<td>0.000</td>
<td>-4.869</td>
<td>14</td>
<td>-0.221</td>
<td>-0.319 to -0.124</td>
<td>0.801</td>
</tr>
<tr>
<td>Upt868 (C) + TAL868 W232Q</td>
<td>Upt868 (T) + TAL868</td>
<td>0.000</td>
<td>-11.768</td>
<td>14</td>
<td>-0.390</td>
<td>-0.461 to -0.319</td>
<td>0.677</td>
</tr>
<tr>
<td>Upt868 (G) + TAL868 W232Q</td>
<td>Upt868 (T) + TAL868</td>
<td>0.151</td>
<td>1.519</td>
<td>14</td>
<td>-0.087</td>
<td>-0.209 to 0.036</td>
<td>0.917</td>
</tr>
<tr>
<td>Upt868 (A) + TAL868 W232R</td>
<td>Upt868 (T) + TAL868</td>
<td>0.163</td>
<td>-1.452</td>
<td>19</td>
<td>-0.083</td>
<td>-0.203 to 0.037</td>
<td>0.920</td>
</tr>
<tr>
<td>Upt868 (C) + TAL868 W232R</td>
<td>Upt868 (T) + TAL868</td>
<td>0.038</td>
<td>-2.227</td>
<td>19</td>
<td>-0.141</td>
<td>-0.273 to -0.008</td>
<td>0.869</td>
</tr>
</tbody>
</table>

1. P value for the t-tests testing the null hypothesis that the expected value of ln(Treatment 1) - ln(Treatment 2) = 0
2. Degrees of freedom
3. Estimate of the average difference ln(Treatment 1)-ln(Treatment 2). Estimate < 0 indicates that the GUS activity of Treatment 1 is lower than the GUS activity of Treatment 2. Estimate > 0 indicates that the GUS activity of Treatment 1 is higher than the GUS activity of Treatment 2.
4. 95% Confidence Interval for the Estimate.
5. Fold Change of Treatment 1/Treatment 2. FC = e^Estimate. FC = 1 indicates the two treatments have equal GUS activity. FC > 1 indicates that Treatment 1 has higher GUS activity that Treatment 2. FC < 1 indicates that Treatment 2 has higher GUS activity that Treatment 1.
<table>
<thead>
<tr>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>P-Value¹</th>
<th>T Statistic</th>
<th>DF²</th>
<th>Estimate³</th>
<th>95% CI⁴</th>
<th>FC⁵</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upt868 (G) + Upt868 (T) + W232R</td>
<td>Upt868 (T) + TAL868</td>
<td>0.485</td>
<td>-0.712</td>
<td>19</td>
<td>-0.040</td>
<td>0.077</td>
<td>0.961</td>
</tr>
<tr>
<td>Upt868 (A) + Upt868 (T) + TAL868 W232T</td>
<td>Upt868 (T) + TAL868 W232T</td>
<td>0.861</td>
<td>-0.179</td>
<td>14</td>
<td>-0.011</td>
<td>0.126</td>
<td>0.989</td>
</tr>
<tr>
<td>Upt868 (C) + Upt868 (T) + TAL868 W232T</td>
<td>Upt868 (T) + TAL868 W232T</td>
<td>0.000</td>
<td>-5.278</td>
<td>14</td>
<td>-0.334</td>
<td>-0.198</td>
<td>0.716</td>
</tr>
<tr>
<td>Upt868 (G) + Upt868 (T) + W232T</td>
<td>Upt868 (T) + TAL868 W232T</td>
<td>0.008</td>
<td>-3.066</td>
<td>14</td>
<td>-0.338</td>
<td>-0.101</td>
<td>0.714</td>
</tr>
<tr>
<td>UptPthXo1 (A) + UptPthXo1 (T) + PthXo1</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.757</td>
<td>-0.312</td>
<td>29</td>
<td>-0.031</td>
<td>0.173</td>
<td>0.969</td>
</tr>
<tr>
<td>UptPthXo1 (A) + UptPthXo1 (T) + PthXo1 W232N</td>
<td>UptPthXo1 (T) + PthXo1 W232N</td>
<td>0.006</td>
<td>-3.158</td>
<td>17</td>
<td>-0.617</td>
<td>-0.205</td>
<td>0.540</td>
</tr>
<tr>
<td>UptPthXo1 (A) + UptPthXo1 (T) + PthXo1 W232P</td>
<td>UptPthXo1 (T) + PthXo1 W232P</td>
<td>0.313</td>
<td>-1.027</td>
<td>30</td>
<td>-0.067</td>
<td>0.066</td>
<td>0.935</td>
</tr>
<tr>
<td>UptPthXo1 (A) + UptPthXo1 (T) + PthXo1 W232Q</td>
<td>UptPthXo1 (T) + PthXo1 W232Q</td>
<td>0.089</td>
<td>-1.793</td>
<td>19</td>
<td>-0.416</td>
<td>-0.020</td>
<td>1.088</td>
</tr>
<tr>
<td>UptPthXo1 (A) + UptPthXo1 (T) + PthXo1 W232T</td>
<td>UptPthXo1 (T) + PthXo1 W232T</td>
<td>0.000</td>
<td>-4.350</td>
<td>30</td>
<td>-0.576</td>
<td>-0.036</td>
<td>0.562</td>
</tr>
<tr>
<td>UptPthXo1 (C) + UptPthXo1 (T) + PthXo1</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.033</td>
<td>-2.305</td>
<td>18</td>
<td>-0.555</td>
<td>-0.049</td>
<td>0.574</td>
</tr>
<tr>
<td>UptPthXo1 (C) + UptPthXo1 (T) + PthXo1 W232N</td>
<td>UptPthXo1 (T) + PthXo1 W232N</td>
<td>0.013</td>
<td>-2.657</td>
<td>29</td>
<td>-0.280</td>
<td>-0.064</td>
<td>0.755</td>
</tr>
<tr>
<td>UptPthXo1 (C) + UptPthXo1 (T) + PthXo1 W232P</td>
<td>UptPthXo1 (T) + PthXo1 W232P</td>
<td>0.114</td>
<td>-1.629</td>
<td>30</td>
<td>-0.113</td>
<td>-0.029</td>
<td>0.893</td>
</tr>
<tr>
<td>UptPthXo1 (C) + UptPthXo1 (T) + PthXo1 W232Q</td>
<td>UptPthXo1 (T) + PthXo1 W232Q</td>
<td>0.098</td>
<td>-1.739</td>
<td>19</td>
<td>-0.323</td>
<td>-0.071</td>
<td>0.724</td>
</tr>
<tr>
<td>UptPthXo1 (C) + UptPthXo1 (T) + PthXo1 W232R</td>
<td>UptPthXo1 (T) + PthXo1 W232R</td>
<td>0.970</td>
<td>0.038</td>
<td>14</td>
<td>0.005</td>
<td>0.286</td>
<td>1.005</td>
</tr>
<tr>
<td>UptPthXo1 (C) + UptPthXo1 (T) + PthXo1 W232T</td>
<td>UptPthXo1 (T) + PthXo1 W232T</td>
<td>0.000</td>
<td>-4.589</td>
<td>30</td>
<td>-0.673</td>
<td>-0.374</td>
<td>0.510</td>
</tr>
<tr>
<td>UptPthXo1 (G) + UptPthXo1 (T) + PthXo1</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.053</td>
<td>-2.020</td>
<td>29</td>
<td>-0.227</td>
<td>0.003</td>
<td>0.797</td>
</tr>
<tr>
<td>UptPthXo1 (G) + UptPthXo1 (T) + PthXo1 W232N</td>
<td>UptPthXo1 (T) + PthXo1 W232N</td>
<td>0.902</td>
<td>-0.125</td>
<td>18</td>
<td>-0.020</td>
<td>0.322</td>
<td>0.980</td>
</tr>
</tbody>
</table>

¹ P value for the t-tests testing the null hypothesis that the expected value of ln(Treatment 1) - ln(Treatment 2) = 0
² Degrees of freedom
³ Estimate of the average difference ln(Treatment 1)-ln(Treatment 2). Estimate < 0 indicates that the GUS activity of Treatment 1 is lower than the GUS activity of Treatment 2. Estimate > 0 indicates that the GUS activity of Treatment 1 is higher than the GUS activity of Treatment 2.
⁴ 95% Confidence Interval for the Estimate.
⁵ Fold Change of Treatment 1/Treatment 2. FC = e^Estimate. FC = 1 indicates the two treatments have equal GUS activity. FC > 1 indicates that Treatment 1 has higher GUS activity that Treatment 2. FC < 1 indicates that Treatment 2 has higher GUS activity that Treatment 1.
**Supplementary Table S2. Continued.**

<table>
<thead>
<tr>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>P-Value $^1$</th>
<th>T Statistic</th>
<th>DF $^2$</th>
<th>Estimate $^3$</th>
<th>95% CI $^4$</th>
<th>FC $^5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UptPthXo1 (G) + UptPthXo1 (T) + PthXo1 W232P</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.469</td>
<td>-0.734</td>
<td>30</td>
<td>-0.109</td>
<td>-0.411 to 0.194</td>
<td>0.897</td>
</tr>
<tr>
<td>UptPthXo1 (G) + PthXo1 W232R</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.011</td>
<td>-2.913</td>
<td>14</td>
<td>-0.439</td>
<td>-0.763 to -0.116</td>
<td>0.644</td>
</tr>
<tr>
<td>UptPthXo1 (G) + PthXo1 W232T</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.022</td>
<td>-2.425</td>
<td>30</td>
<td>-0.447</td>
<td>-0.823 to -0.071</td>
<td>0.640</td>
</tr>
<tr>
<td>UptPthXo1 (T) + UptPthXo1 (T) + PthXo1 W232N</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.003</td>
<td>-3.509</td>
<td>17</td>
<td>-0.532</td>
<td>-0.852 to -0.212</td>
<td>0.587</td>
</tr>
<tr>
<td>UptPthXo1 (T) + UptPthXo1 (T) + PthXo1 W232Q</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.001</td>
<td>-3.645</td>
<td>30</td>
<td>-0.635</td>
<td>-0.990 to -0.279</td>
<td>0.530</td>
</tr>
<tr>
<td>UptPthXo1 (T) + UptPthXo1 (T) + PthXo1 W232R</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.012</td>
<td>-2.789</td>
<td>19</td>
<td>-0.644</td>
<td>-1.128 to -0.838</td>
<td>0.525</td>
</tr>
<tr>
<td>UptPthXo1 (T) + UptPthXo1 (T) + PthXo1 W232T</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.023</td>
<td>-2.556</td>
<td>14</td>
<td>-0.456</td>
<td>-1.195 to -0.073</td>
<td>0.634</td>
</tr>
<tr>
<td>UptPthXo1 (A) + UptPthXo1 (T) + PthXo1 W232N</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.580</td>
<td>-0.564</td>
<td>18</td>
<td>-0.129</td>
<td>0.352 to 0.879</td>
<td></td>
</tr>
<tr>
<td>UptPthXo1 (C) + UptPthXo1 (T) + PthXo1 W232N</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.397</td>
<td>-0.867</td>
<td>18</td>
<td>-0.206</td>
<td>0.292 to 0.814</td>
<td></td>
</tr>
<tr>
<td>UptPthXo1 (G) + UptPthXo1 (T) + PthXo1 W232N</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.019</td>
<td>2.574</td>
<td>18</td>
<td>0.370</td>
<td>0.068 to 0.673</td>
<td>1.448</td>
</tr>
<tr>
<td>UptPthXo1 (A) + UptPthXo1 (T) + PthXo1 W232P</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.001</td>
<td>3.713</td>
<td>30</td>
<td>0.568</td>
<td>0.255 to 0.880</td>
<td>1.764</td>
</tr>
<tr>
<td>UptPthXo1 (C) + UptPthXo1 (T) + PthXo1 W232P</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.002</td>
<td>3.360</td>
<td>30</td>
<td>0.521</td>
<td>0.204 to 0.838</td>
<td>1.684</td>
</tr>
<tr>
<td>UptPthXo1 (G) + UptPthXo1 (T) + PthXo1 W232P</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.000</td>
<td>5.775</td>
<td>30</td>
<td>0.526</td>
<td>0.340 to 0.712</td>
<td>1.692</td>
</tr>
<tr>
<td>UptPthXo1 (A) + UptPthXo1 (T) + PthXo1 W232Q</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.270</td>
<td>1.137</td>
<td>19</td>
<td>0.229</td>
<td>0.041 to 0.649</td>
<td>1.257</td>
</tr>
<tr>
<td>UptPthXo1 (C) + UptPthXo1 (T) + PthXo1 W232Q</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.047</td>
<td>2.128</td>
<td>19</td>
<td>0.321</td>
<td>0.005 to 0.637</td>
<td>1.379</td>
</tr>
<tr>
<td>UptPthXo1 (G) + UptPthXo1 (T) + PthXo1 W232Q</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.005</td>
<td>3.165</td>
<td>19</td>
<td>0.577</td>
<td>0.195 to 0.958</td>
<td>1.780</td>
</tr>
<tr>
<td>UptPthXo1 (A) + UptPthXo1 (T) + PthXo1 W232R</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.012</td>
<td>2.905</td>
<td>14</td>
<td>0.540</td>
<td>0.141 to 0.939</td>
<td>1.716</td>
</tr>
<tr>
<td>UptPthXo1 (C) + UptPthXo1 (T) + PthXo1 W232R</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.035</td>
<td>2.335</td>
<td>14</td>
<td>0.461</td>
<td>0.038 to 0.884</td>
<td>1.585</td>
</tr>
<tr>
<td>UptPthXo1 (G) + UptPthXo1 (T) + PthXo1 W232R</td>
<td>UptPthXo1 (T) + PthXo1</td>
<td>0.857</td>
<td>0.184</td>
<td>14</td>
<td>0.016</td>
<td>-0.175 to 1.017</td>
<td></td>
</tr>
</tbody>
</table>

---

1. P value for the $t$-tests testing the null hypothesis that ln(Treatment 1) - ln(Treatment 2) = 0
2. Degrees of freedom
3. Estimate of the average difference ln(Treatment 1)-ln(Treatment 2). Estimate < 0 indicates that the GUS activity of Treatment 1 is lower than the GUS activity of Treatment 2. Estimate > 0 indicates that the GUS activity of Treatment 1 is higher than the GUS activity of Treatment 2.
4. 95% Confidence Interval for the Estimate.
5. Fold Change of Treatment 1/Treatment 2. FC = $e^{\text{Estimate}}$. FC = 1 indicates the two treatments have equal GUS activity. FC > 1 indicates that Treatment 1 has higher GUS activity than Treatment 2. FC < 1 indicates that Treatment 2 has higher GUS activity than Treatment 1.
Supplementary Table S2. Continued.

<table>
<thead>
<tr>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>P-Value&lt;sup&gt;1&lt;/sup&gt;</th>
<th>T Statistic</th>
<th>DF&lt;sup&gt;2&lt;/sup&gt;</th>
<th>Estimate&lt;sup&gt;3&lt;/sup&gt;</th>
<th>95% CI&lt;sup&gt;4&lt;/sup&gt;</th>
<th>FC&lt;sup&gt;5&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>UptPthXo1 (A) + PthXo1 W232T</td>
<td>UptPthXo1 (T) + PthXo1 W232T</td>
<td>0.255</td>
<td>1.161</td>
<td>30</td>
<td>0.180</td>
<td>-0.137 to 0.497</td>
<td>1.197</td>
</tr>
<tr>
<td>UptPthXo1 (C) + PthXo1 W232T</td>
<td>UptPthXo1 (T) + PthXo1 W232T</td>
<td>0.606</td>
<td>0.521</td>
<td>30</td>
<td>0.080</td>
<td>-0.235 to 0.396</td>
<td>1.084</td>
</tr>
<tr>
<td>UptPthXo1 (G) + PthXo1 W232T</td>
<td>UptPthXo1 (T) + PthXo1 W232T</td>
<td>0.000</td>
<td>4.686</td>
<td>30</td>
<td>0.419</td>
<td>0.237 to 0.602</td>
<td>1.521</td>
</tr>
</tbody>
</table>

<sup>1</sup>P value for the t-tests testing the null hypothesis that the expected value of ln(Treatment 1) - ln(Treatment 2) = 0

<sup>2</sup>Degrees of freedom

<sup>3</sup>Estimate of the average difference ln(Treatment 1) - ln(Treatment 2). Estimate < 0 indicates that the GUS activity of Treatment 1 is lower than the GUS activity of Treatment 2. Estimate > 0 indicates that the GUS activity of Treatment 1 is higher than the GUS activity of Treatment 2.

<sup>4</sup>95% Confidence Interval for the Estimate.

<sup>5</sup>Fold Change of Treatment 1/Treatment 2. FC = e<sup>Estimate</sup>. FC = 1 indicates the two treatments have equal GUS activity. FC > 1 indicates that Treatment 1 has higher GUS activity that Treatment 2. FC < 1 indicates that Treatment 2 has higher GUS activity that Treatment 1.

Supplementary Figure S1. GUS activity for PthXo1 variants with all 19 possible single amino acid substitutions for W232. W232 is the wild-type PthXo1 with W at position 232. Substitutions to other amino acids are indicated as W232X. UptPthXo1 (A), UptPthXo1 (C), UptPthXo1 (G), and UptPthXo1 (T) indicate that the target UptPthXo1 is preceded by an adenine, cytosine, guanine, or thymine, respectively. To facilitate comparison across multiple experiments, activity within each experiment was normalized so that
the activity of PthXo1 + UptPthXo1 (T) is equal to 1.0. Values are the mean of three replicates. Error bars indicate s.d. Horizontal lines at bottom group constructs that were tested on the same day in the same experiment.

Supplementary Figure S2. ClustalW multiple sequence alignment of N terminal sequences of Ralstonia TAL-like effectors (RTLs) and PthXo1. RSc1825, Genbank ID CAD15517.1 from Ralstonia solanacearum strain GMI1000(1). hpx17, Genbank ID AB178011.1 from strain RS1085(2). RscCAQ18687, Genbank ID CAQ18687.1 from strain MolK2 (direct Genbank submission by Genoscope -C.E.A.). PthXo1, Genbank ID ACD58243.1 from Xanthomonas oryzae strain PXO99A (3).
Supplementary Figure S3. Secondary structure predictions for PthXo1 and Rsc1815 N terminal regions. Left: PthXo1, ACD58243.1 from Xanthomonas oryzae strain PXO99A (3). Right: RSc1815 (GenBank protein CAD15517.1) from Ralstonia solanacearum strain GMI1000 (1).

Supplementary references


CHAPTER 7. GENERAL CONCLUSIONS.

General Discussion

In this dissertation, I describe computational and experimental approaches for studying TAL effector-DNA binding, and I have apply the results to TAL effector design and target prediction and to answer questions about the role of TAL effectors in plant disease. In Chapter 3, I analyzed known TAL effector-DNA binding site pairs to develop a series of design guidelines for custom TAL effectors and TAL effector nucleases (TALENs). I reasoned that naturally occurring TAL effectors would have evolved to have high levels of activity on their targets. Out of 15 TALEN pairs designed according to these guidelines, 14 were active, suggesting that following the guidelines aided in designing active TALENs.

However, more recent research suggests that conformation to the design guidelines did not boost TAL effector/TALEN activity. For example, the guidelines advise selecting targets with a high percentage of A’s (corresponding to the RVD NI) and relatively few G’s (corresponding to the RVD NN). However, the solved structures of two TAL effectors bound to their respective DNA targets showed that the NI-A interaction was weaker than other RVD-nucleotide pairs (1,2). Experimental data has subsequently confirmed NI has significantly lower affinity for A than HD for C or NN for G (3,4). Although we found evidence supporting our design guidelines in the activity of intermediate constructs, close review of those data show that the constructs with the highest activity also had high percentages of HD-C pairs. Additionally, the TALEN pair that was not active had one TALEN monomer in which 9 of 23 positions were NI-A’s (5).

Analysis by Reyon et al. (6) found that adherence to any of the guidelines did not significantly improve TALEN activity in the context of human cell-based reporter assay (they also did not test the guidelines in the context of gene activation by single TAL effectors). It is possible that the positional biases observed at the ends of TAL effector RVD sequences are due to lack of recombination at the ends of the repeat region rather than an evolutionary adaptation for binding affinity or activity. Therefore, although the design guidelines have not been systematically tested in a variety of cell types and using TALENs targeting endogenous genes, designing TALENs to target GC-rich sites is likely the best way to ensure their activity..

I subsequently updated the TALEN design software tool described in Chapter 3 to reflect this changing knowledge. The newer version, described in Chapter 4, was made available as part of the TAL Effector-Nucleotide Targeter website (TALE-NT 2.0). Instead of design
guidelines, the updated TALEN design tool allowed users to design TALENs optimized for their specific application and to specify the approximate location of the TALEN cut site.

In addition to the updated TALEN design tool, I also created a software tool for TAL effector target prediction, called Target Finder. Because RVD-nucleotide binding specificity is not absolute (7), custom TAL effectors and TALENs may bind to sequences other than their intended target. Custom TAL effector transcription factors can activate the transcription of non-target genes with promoter sequences similar to the intended target (8). The potential for such off-target activity is a critical problem in TALEN-based applications, where off-target binding could result in the cleavage and disruption of non-target genes. Target Finder uses a mathematical scoring function to predict candidate binding sites for a TAL effector in a user-selected genome. In addition to allowing researchers to quickly assess the off-target binding potential of a TAL effector, it can be used to identify candidate TAL effector-targeted genes for natural TAL effectors in plant disease (an example of this is described in detail in the Appendix).

Because the tool only searches for sites for one TAL effector at a time, identifying off-target sites for TALENs requires running searches for each TALEN monomer and then identifying opposing sites with the proper orientation and spacing (this has since been updated, see Significance, below). Another weakness of the tool is the score cutoff used to predict binding. Although the cutoff is based on experimental data for predicted and verified TAL effector target sites (see Appendix), it likely misses some sites that are just above the cutoff and also identifies a high number of false positive sites scoring below the cutoff. Nevertheless, the Target Finder tool provides a relatively simple way for researchers to estimate and compare the off-target potential of custom TAL effectors or TALENs. Improvements to the tool will require better understanding the importance of binding site context and of the individual contributions RVD-nucleotide to binding affinity (see Recommendations for Future Research, below).

In Chapter 5, I present a case study in engineering disease resistance in rice by stably inserting multiple TAL effector EBEs into the promoter of a naturally occurring $R$ gene. One potential pitfall to this mode of engineered resistance is that inserted EBEs might overlap cis-regulatory elements. The insertion of such elements into the promoter of the executor $R$ gene might cause the gene to be activated during development, in certain tissues, or under certain environmental conditions in the absence of any TAL effector, ultimately leading to cell death and the death of the plant. To assess this possibility, I analyzed the inserted TAL effector EBEs for evidence of potential regulatory elements. I discovered that the EBEs contained sequences under selection, suggesting that TAL effectors might have evolved specifically to target these
regulatory sequences in host plants. It is possible that the positional biases observed in TAL effector targets discussed in Chapter 3 reflect this evolution.

Finally, in Chapter 6 I explored the role of the cryptic -1st repeat in TAL effector DNA binding. Although the role of RVDs in specifying the binding site has been well established, it is not clear how the TAL effector specifies the thymine (T) directly preceding the repeat-specified binding sites of almost all TAL effectors in nature (7). Replacing this T with another nucleotide preceding the binding site dramatically lowers or even eliminates TAL effector activity (9,10), so custom TAL effectors and TALENs have typically been designed to target sites preceded by T. Understanding how this 0th position T is specified could enable engineering of TAL effectors with an altered or relaxed specificity for this position. The solved crystal structure for TAL effector PthXoI bound to its DNA target revealed that tryptophan 232 (W232), located in the “cryptic” -1st repeat appeared to make base-specific, energetically favorable contacts with the 0th position T (1). Although we found that single amino acid substitutions for W232 altered TAL effector activity and specificity for the 0th position T, the effects of individual substitutions varied depending on the assay, the TAL effector repeat sequence, and other portions of the N terminus. These context-based effects suggest that other portions of the TAL effector, such as the N terminus and the cryptic 0th repeat, may also be involved in specifying the 0th position T.

A weakness in our approach was that we did not test the effects of the W232 substitutions on stability of the TAL effector N terminus. Loss of N terminal stability could affect the N terminal’s ability to nucleate TAL effector DNA binding. The observed lower TAL effector activity would be due to the reduced binding affinity and not to changes in specific interactions between the 0th position T and the amino acid at position 232. Therefore, the effect of W232 substitutions on N terminal stability must be determined before it can be definitively stated that W232 is or is not involved in specifying the preference for the 0th position T. It is also feasible that W232 does specify this preference, and that W232 substitutions also destabilize the N terminus. If this proves to be the case, the creation of enhanced TAL effectors with broadened targeting capacity will require substantial reengineering of the entire N terminus. In addition, understanding the role of TAL effector context on specificity for the 0th position T, especially the role of the RVD sequence and mismatches, may help to improve the target prediction tools described in Chapter 3 (for example, in cases where bases other the T might be tolerated at the 0th position.)
Significance

The design of maximally efficient and highly specific custom TAL effectors and TALENs is becoming increasingly important as TAL effector-based constructs are adapted for a variety of research and biotechnology applications. Although TAL effector repeats appear to behave modularly, meaning they can be assembled in any order with no effects on activity or specificity, custom TAL effectors and TALENs exhibit varying levels of activity. As noted above, differences in activity level may be due to RVD sequence composition, with RVD sequences containing more strong-binding RVDs having higher activity. However, other factors, such as binding site context may also play a role (see Recommendations for Future Research). The software tools for TAL effector/TALEN design described in Chapters 3 and 4 were some of the first TAL effector/TALEN design tools available. More importantly, they were the first TALEN design tools to recognize that designing active TAL effector-based constructs might depend on selecting RVD and target sequences with specific features, rather than simply assembling a string of RVDs with no attention to context or composition. The manuscript in Chapter 3, which describes an early version of these tools, as well as a method for TAL effector/TALEN construction, has been cited 94 times.

The TAL effector target prediction tool Target Finder, which was described in Chapter 4, is also unique. It is the only existing tool for predicting binding sites that allows mismatches between the TAL effector and target sequence. Since the publication of that manuscript, the TALE-NT 2.0 website has been updated to include a new target prediction tool, based on Target Finder, which searches for paired TAL effector binding sites for TALENs. Paired Target Finder allows users to perform off-target site searches for TALENs in one search instead of searching for sites for each monomer individually. Recently, we integrated Paired Target Finder with TALEN Targeter, allowing users to design TALENs and receive predicted off-target site counts in a single step. Addition of these features to the TALEN 2.0 website dramatically increased user traffic, with the website as a whole processing 1300+ queries per week. Currently, the TALEN design tool (with integrated off-target site counting) processes an average of 760+ queries per week, and the manuscript presented in Chapter 4 has already been cited 12 times. Usage of these tools will likely remain high, because the flexibility of the TALE-NT 2.0 website interface and code makes it simple to update these tools in response to changing knowledge of TAL effector-DNA binding. Additionally, the use of these tools to predict targets of natural TAL effectors in host plants will continue to grow as more Xanthomonas genome sequences become available and their TAL effectors are inventoried. Publication of the study reported in the Appendix, which describes the first successful attempt at in silico genome-wide prediction of
TAL effector targets, will increase interest in using target prediction tools to study the role of TAL effectors in plant disease by providing a roadmap for future similar studies.

The work described in Chapter 5 addresses the role of TAL effectors in plant pathology from the angle of disease resistance. Although it had been previously shown that introducing an EBE into a gene’s promoter made that gene responsive to the corresponding TAL effector (9), this study, spearheaded by Aaron Hummel, was the first to show that such a mechanism could be used to engineer and broaden resistance by making an executor R gene responsive to multiple TAL effectors from multiple Xanthomonas strains. This was also the first study to report that TAL effectors may target sequences under selection, including potential cis-regulatory elements (my contribution). The introduction of TAL effector binding sites into gene promoters might introduce regulatory sequences that could cause activation the R gene under specific conditions and lead to the death of the plant. This potential pitfall to engineering TAL effector-responsive executor R genes had not been discussed in the literature.

Finally, the manuscript in Chapter 6 adds to the understanding of the basic biology of TAL effector-DNA interactions. Although the crystal structure of TAL effector PthXo1 bound to its target indicates that amino acid W232 in the cryptic -1\textsuperscript{st} repeat interacts with and specifies the required T preceding the TAL effector binding site (1), the -1\textsuperscript{st} repeat region is not well resolved in the other published structure of artificial TAL effector dHAX3 (2). Previously, it was thought that the cryptic -1\textsuperscript{st} repeat specified the 0\textsuperscript{th} position T through interactions between a single amino acid (W232) and a single nucleotide, much like the interactions between the canonical repeats and binding site nucleotides. However, our study of W232 substitutions found numerous context effects, suggesting that the cryptic repeats may rely on interactions with each other, other regions of the protein, and even multiple nucleotides in the target DNA to specify the 0\textsuperscript{th} repeat and nucleate DNA binding. Therefore, this study revealed previously unknown complexities in TAL effector DNA binding, and will pave the way for additional studies of the structure and function of the TAL effector N terminus. Overall better understanding of the TAL effector-DNA binding mechanism may also enhance custom TAL effector design and inform target prediction models.

**Recommendations for Future Research**

Better understanding of TAL effector DNA-binding will be needed to continue to improve custom TAL effector design and target prediction. Currently, the effect of binding site context on TAL effector-DNA binding is not well understood. Other site-specific DNA binding proteins are sensitive to local features such as chromatin status and methylation (11), and TAL effector
transcriptional activators are not able to overcome epigenetic gene silencing (12). Other features like regulatory elements may also affect TAL effector binding and activity. Understanding the effects of such features will allow researchers to target sites that will not be negatively affected by local context, or to design TALENs that overcome such context-imposed limitations. For example, TAL effectors can be designed to target methylated cytosines by using non-standard RVDs (13). Similarly, the discovery of new, highly specific RVDs, such as NH to target G, may improve TAL effector specificity (14).

Designing highly specific TAL effectors also depends on accurately predicting off-target binding sites. The method of target prediction presented in Chapter 4 uses as scoring function that relies on observed RVD-nucleotide association frequencies to approximate the individual contributions RVD-nucleotide pairs to binding affinity. Replacing these frequencies with actual binding affinity data would likely improve prediction accuracy. In addition, the effects of mismatches on binding affinity are not well understood. There is some evidence that mismatches at the 5' end of the binding site are more detrimental to binding than mismatches at the 3' end (3), and the distribution of mismatches may also be important. Systematic study of the effects of mismatches on binding would enable the development of a target prediction function that accurately accounts for these complexities. The scoring function might also be improved by giving higher weights to strong-binding RVDs and lower weights to weak binding RVDs (3,4). Such improvements would enhance our ability to predict off-target binding sites for custom TAL effectors /TALENs and to identify TAL effector-targeted genes involved in plant disease.

Successful deployment of the method of engineered disease resistance presented in Chapter 5 depends on the selection of appropriate TAL effectors and their corresponding EBEs. It is important to choose TAL effectors that are present in a large number of Xanthomonas strains to ensure broad resistance. Therefore, it will be necessary to inventory the TAL effector complement of as many Xanthomonas strains as possible. This may require the development of new sequencing technologies, as current technologies are unable to sequence through the repetitive DBD. Identification of novel R and S genes as described in the Appendix will also require sequencing TAL effectors from diverse populations of Xanthomonas.

Finally, understanding the complexities involved in TAL effector-DNA binding, as described in Chapter 6, will be crucial for optimizing TAL effectors and TALENs for biotechnology applications and for studying the role of TAL effectors in plant disease. The current understanding of TAL effector-DNA binding is based on the crystal structures of two TAL effectors bound to their targets. Additional structures for TAL effectors with a variety of RVD
sequences and target mismatches will be necessary to begin sorting out the complex interactions between the TAL effector N terminal and the DNA target. Structure-function studies and a solved structure for a Ralstonia TAL-like effector, with its slightly different amino acid sequence, would also allow for comparisons of specific interactions between the two proteins and their targets. In light of the complexities in TAL effector-DNA binding and binding site specificities, future attempts to alter TAL effector target specificity or activity levels by engineering the N terminus will likely require mutations at multiple positions. These efforts will need to be guided by structure modeling and binding affinity predictions to pinpoint useful combinations of mutations to characterize.

References


APPENDIX: DE NOVO PREDICTION OF TAL EFFECTOR TARGETS IDENTIFIES A NOVEL SUSCEPTIBILITY GENE FOR BACTERIAL LEAF STREAK OF RICE

Summary

This appendix contains data from a larger study entitled “High-throughput identification of TAL effector targets in bacterial leaf streak of rice reveals striking differences from bacterial blight and a new class of susceptibility gene” that is currently in preparation for publication. The data presented here represent a case study in combining TAL effector target prediction with transcriptomics data for the high-throughput identification of previously unknown TAL effector targets and the discovery of previously unknown plant disease susceptibility (S) genes.

Transcription-activator like (TAL) effectors of Xanthomonas spp. induce expression of targeted S genes which aid in bacterial spread and multiplication and symptom development during host plant infection. Following injection into the host plant cell, TAL effectors localize to the nucleus where they bind to specific DNA sequences and activate target gene expression. In this work we focused on X. oryzae pv. oryzicola (Xoc), the causal agent of rice bacterial leaf streak (BLS). Xoc is closely related to the rice bacterial blight (BB) pathogen X. oryzae pv oryzae (Xoo). Although the role of many Xoo TAL effectors in disease has been characterized, the role of most Xoc TAL effectors remains unknown. Therefore, we combined microarray analyses, computational predictions and experimental validations to pinpoint a novel set of Xoc TAL effectors targets. Follow-up experiments confirmed that 19 of 36 predicted candidate targets of Xoc TAL effectors were induced by TAL effectors. Our list of confirmed targets included two genes targeted by Xoc TAL effector Tal2g, which we identified as an important virulence factor for BLS. To determine which Tal2g target was the biologically relevant S gene, we used designer TAL effectors (dTALEs) to specifically activate each target. We found that complementation with the dTALE activating the sulfate transporter gene was able to restore virulence to a mutant strain of BLS256 lacking the tal2g gene, and identified the sulfate transporter as an S gene. Thus, combining TAL effector target predictions with transcriptomics data allowed us to discover a previously unknown set of host genes induced by Xoc TAL effectors, and identify a key S gene for BLS of rice.

1 Authors Raul A. Cernadas, Erin L. Doyle, David O. Niño-Liu, Clarice L. Schmidt, Li Wang, Timothy Bancroft, Rico Caldo, Bing Yang, Frank F. White, Dan Nettleton, Roger P. Wise, and Adam J. Bogdanove
Introduction

*Xanthomonads* are widely distributed plant pathogens that affect numerous plant species, including the economically important crops rice, cotton, citrus, cassava, and tomato (1). Bacterial leaf streak (BLS) and bacterial blight (BB) diseases of rice (*Oryza sativa*) are caused by *Xanthomonas oryzae* pv. oryzicola (*Xoc*) and *Xanthomonas oryzae* pv. oryzae (*Xoo*), respectively. Although BLS is less important and more sporadic than BB, both cause considerable reductions of crop yield in the major rice-producing countries. During infection, *Xoo* colonizes the vascular system of rice leaves, while *Xoc* is limited to the non-vascular mesophyll tissue. The determinants for this difference in host tissue specificity are unknown (2).

The virulence of both *Xoc* and *Xoo* depends on a type-three secretion system (T3SS), which injects a variety of effectors from the bacterium into the host plant cells. Included in this T3SS-delivered effector cocktail are transcription-activator like (TAL) effectors (3,4). Upon injection into the plant cell, TAL effectors localize to the nucleus and induce expression of specific host genes by binding to effector-specific DNA sequences called effector binding elements (EBEs). In BB, induction of *susceptibility* (S) genes by *Xoo* TAL effectors is necessary for pathogen multiplication and spreading as well as the occurrence of disease symptoms (5-10). Recent reports have shown that many of the major TAL effector-targeted genes on which *Xoo* pathogenicity relies are sugar transporters that are presumably involved in sucrose phloem loading (7,11-13). Despite the fact that the sequenced genome of *Xoc* strain BLS256 encodes an unusually high number of TAL effectors, 26 (14), the role of TAL effectors in *Xoc* virulence have not been characterized, and the majority of BLS256 TAL effectors have unknown targets.

Each TAL effector’s specific DNA target is determined by its central repeat domain, which is composed of a variable number of nearly identical, tandem repeats of 34-35 amino acids in length. Repeat-to-repeat variation occurs primarily at residues 12 and 13 of each repeat (known as the repeat-variable diresidue or RVD), with each RVD forming base-specific contacts and specifying a single binding site nucleotide (15-17). The most commonly occurring RVDs are HD, NG, NI, and NN, and they associate preferentially with the DNA bases C, T, A, and G/A, respectively. This simple TAL effector-DNA binding code has made it possible to identify the specific EBEs of *Xoo* TAL effectors in the promoters of TAL effector-targeted genes (18). However, the tolerance for multiple mismatches (positions where an RVD is aligned with a non-preferred nucleotide) between TAL effectors and their targets, as well as the existence of rare RVDs whose specificities have not been characterized, make the *de novo* prediction of TAL effector targets a challenging task.
To test whether it is possible to predict *in silico* the TAL effector targets of a non-characterized strain, we undertook whole genome transcriptional profiling and combined it with computational prediction and experimental validation to comprehensively identify TAL effector target genes that may function as S genes for BLS. By combining microarray analysis with computational predictions, we identified 34 candidate target genes for *Xoc* TAL effectors. Experimental validations later confirmed that 19 of the 34 were TAL effector targets. Using a library of BLS256 TAL effector mutants, we identified Tal2g as a major TAL effector virulence factor of *Xoc*. Our computational analysis and experimental verification revealed that two genes, a monocupper oxidase and a sulfate transporter, were directly targeted and induced by Tal2g. We used designer TAL effectors (dTALEs) to selectively activate the expression of each gene, revealing that the sulfate transporter was the relevant target for *Xoc* virulence. These results demonstrate the utility of high-throughput identification of TAL effector targets by combining computational predictions with transcriptomics. Such analysis could be easily extended to other important crops affected by *Xanthomonas* spp.

**Results**

*X. oryzae* pv. *oryzicola* and *X. oryzae* pv. *oryzae* induce very different gene expression changes in rice leaves

To examine host gene expression changes associated with BLS and BB, we compared transcript profiles in rice leaves over time after infection with *Xoc* BLS256 or *Xoo* PXO99^A^ (14,19). We focused our statistical analyses of gene expression on the kinetics of the infection instead of looking at a particular time point. Therefore, we identified genes whose average pattern of expression over time was different during the host interaction with *Xoo* or *Xoc* relative to the control. A total of 517 genes were differentially regulated (*q* ≤ 0.3) in the different treatment comparisons. Eighty-seven and ninety-nine genes were induced uniquely in response to *Xoc* or *Xoo*, respectively. Only five genes were commonly regulated by both *Xoc* and *Xoo* (Figure 1). Among the *Xoo*-induced genes, we identified genes previously reported to be TAL effector-targeted S genes for *Xoo* infection, including the sugar transporter, *Os8N3* (7), the transcription initiation factor *TFIIAγ*, and the bZIP transcription factor *OsTFXI* (20). Functional ontology enrichment of *Xoc*- and *Xoo*-upregulated genes revealed that different host signaling pathways appear to be involved in the plant susceptibility response to *Xoc* and *Xoo*. Our data indicated that the *Xoc*-induced genes are potentially involved in reactive oxygen species detoxification and redox status control. These experiments demonstrate that *Xoc* and *Xoo* induce broadly distinct sets of genes in rice.
Many up-regulated genes are predicted to be targets of TAL effectors

We used the TAL effector–target scoring function developed previously (17,21) to scan in silico nearly 60,000 rice gene promoters for candidate binding sites for Xoc and Xoo TAL effectors. For each TAL effector, we identified all sites in the promoterome scoring below a computed cutoff score for that TAL effector. Candidate TAL effector binding sites were required to be directly preceded by a T at the 5’ end. The lists of candidate sites scoring below the cutoff were correlated with the microarray data to identify putative effector binding elements (EBEs) in the promoters of genes induced during infection. We identified thirty-six genes significantly induced by Xoc that contain putative EBEs for twenty different Xoc TAL effectors (Table 1). In addition, our analysis identified thirty Xoo-induced genes (Table 2) containing putative EBEs for Xoo TAL effectors, including those previously reported for PthXo1, PthXo6 and PthXo7 (7,20,22). We refer to these as candidate TAL effector-targeted genes. Eight and four of the candidate TAL effector-targeted genes induced by Xoc and Xoo, respectively, contain putative EBEs for more than one TAL effector, suggesting a partial redundancy among effector specificity.

The number of identified candidate TAL effector-targeted genes for a particular TAL effector was variable. In the case of Xoc, we identified up to five putative EBEs on different promoters for Tal3b and Tal6 and only one putative EBE for each of Tal2c, Tal2f, Tal5a, Tal8, Tal9b and
Tal11b (Table 1). Exceptionally, we found three overlapping candidate EBEs of different Xoc TAL effectors (Tal2a, Tal11b and Tal1c) on the promoter of gene 06g14750. No other overlapping EBEs for Xoc TAL effectors were found in the analysis. For Xoo TAL effectors, we found seven potentially targeted genes for TAL effector PthXo6 and only one targeted gene for each of PthXo1, PthXo7, Tal6a and Tal7a/7b (Table 2). Overlapping of putative EBEs on the promoter of 07g11510 was also identified for Xoo TAL effectors (Table 2). Finally, we also identified potential EBEs on three genes (03g03034, 07g06970 and 02g15290) that are commonly regulated by Xoc and Xoo (Table 2), suggesting that some of the commonly regulated genes are targeted by different TAL effectors from both strains. The majority of genes identified as significantly up-regulated in the microarray experiments were not predicted to be TAL effector targets.

Experimental validation yields 19 confirmed TAL effector targets for X. oryzae pv. oryzicola TAL effectors

We chose to focus our subsequent investigations on targets of Xoc TAL effectors since nothing is yet known about their role in BLS. For each of the predicted Xoc TAL effector-targeted genes, we used semi-quantitative RT-PCR to test whether or not they were induced by an Xoc BLS256 mutant lacking the corresponding TAL effector. We also tested the ability of the corresponding cloned TAL effector gene to complement the mutation and restore induction. Additionally, we expressed each of the cloned TAL effectors in the non-host pathogen X. axonopodis pv. glycines (Xag), strain EB08, which elicits neither symptoms nor hypersensitive reaction when inoculated onto rice, and tested whether delivery of the individual TAL effector by Xag led to induction of the candidate target gene. In all, 19 out of the 45 predicted candidates were experimentally validated as TAL effector targets (Table 1).

Comparison of validated and un-validated predicted TAL effector binding sites reveal characteristics that may be useful for future TAL effector target prediction

Comparison of validated and un-validated candidate TAL effector binding sites revealed a few differing characteristics. Direct comparison of binding site scores for different TAL effectors was impossible, because TAL effectors with more repeats and longer binding sites will always have higher scores. To facilitate comparing scores across multiple TAL effectors, we developed the concept of a “score ratio”. The score ratio is calculated as the actual score for a TAL effector-target alignment divided by the perfect score (the hypothetical score for the TAL effector aligned with its perfect match target). Because score ratios account for the different numbers of repeats, they can be used to compare binding site scores for multiple TAL effectors. We found
that the validated binding sites typically had score ratios ranging from 1.22-2.8, while unverified sites had score ratios ranging from 1.7-3.18. When the binding sites for each TAL effector were ranked by score, verified sites typically had lower (better) ranks, with 15 of 19 verified sites having ranks less than 150. Conversely, 20 of 27 unverified sites had ranks higher than 200, suggesting that rank may be more useful than score or score ratio for separating functional and nonfunctional sites. Finally, we compared the contexts of the functional and non-functional sites, and found that 18 of the 19 verified sites were located within 150bp of the native transcriptional start site (TXS), and many were near TATA boxes. For un-verified sites, only 6 were within 150bp of the TXS. Because our initial lists of predicted sites were quite long, filtering out sites that do not meet some of the characteristics described above may prove to be a useful way to filter the predicted binding site lists.

The sulfate transporter gene targeted by Tal2g is a major S gene for bacterial leaf streak

We previously identified mutant M27 from a library of BLS256 TAL effector mutants as having a mutation in its tal2g gene. The M27 mutant showed a significant loss in virulence (as measured by lesion length) relative to the wild type bacteria which was restored by complementation with the tal2g gene (Figure 2). In comparisons of rice inoculated with wild type BLS256 and mutant M27, the rice inoculated with the mutant strain showed less bacterial exudate on the inoculated leaf surfaces (data not shown). Similarly, isolation and quantification of surface bacteria 7 days after inoculation found 400-fold fewer bacteria on the surface of the M27-inoculated leaves (Figure 2). Thus Tal2g is a major virulence factor in BLS that contributes

\[\text{Appendix Figure 2. Virulence contribution of } X. \text{ oryzae pv. oryzicola BLS256 TAL effector Tal2g. (A)} \]

\[\text{Lengths of lesions caused by } X. \text{ oryzae pv. oryzicola BLS256 (WT), the tal2g knockout derivative M27 carrying an empty plasmid vector (ev), and M27 carrying the vector with the cloned tal2g gene, measured 10 days after infiltration. The asterisk indicates a significant difference relative to WT (P<0.01). Error bars represent standard deviation (N≥10). (B) Total and surface (exudate) bacterial populations of leaves 7 days after inoculation with the strains in panel A. The asterisk indicates a significant difference relative to WT (P<0.01). (P<0.01) between the treatments. Error bars represent standard deviation (N≥6). Experiments were repeated three times with consistent results.}\]
both to lesion expansion and bacterial egress to the leaf surface.

Our microarray and target prediction analysis identified two candidate TAL effector-targeted genes for Tal2g, Os06g46500 and Os01g52130 (Table 1). Both genes were found to be genuinely induced by Tal2g. To distinguish which of these is a biologically relevant target, i.e., an S gene, we engineered designer TAL effectors (dTALes) to specifically activate each of the targets and tested whether the dTALEs restore virulence to M27. dTALE434 was designed to specifically induce Os06g46500, which encodes a monocopper oxidase. Both dTAL436 and dTAL437 were designed to induce Os01g52130, a sulfate transporter (Figure 3A). When transformed individually into M27, all three dTALEs restored induction of their targets to levels similar to induction by Tal2g (Figure 3B). Virulence assays conducted with each dTALE-transformed M27 strain revealed that the induction Os01g52130, but not Os06g46500, is sufficient to restore the wild type virulence in M27 (Figure 3C). These data point to Os01g52130, which encodes a sulfate transporter induced by Tal2g, as a principal susceptibility gene candidate for BLS.

**Discussion**

The work presented here represents the first study to attempt high-throughput, *in silico* prediction of TAL effector targets. We show that by combining TAL effector target site predictions with microarray data and experimental validation, we were able to identify a previously unknown set of TAL effector target genes. Significantly, our approach led to the discovery of a previously unknown S gene for BLS.

Although our method was successful at identifying a set of TAL effector targeted genes, it is unclear how many TAL effector-targeted genes were missed. To construct our initial lists of candidate target genes, we relied on a somewhat arbitrary scoring cutoff. Although this cutoff seemed reasonable, it is unlikely that it captured all TAL effector targets. Additionally we relied on microarray analysis to filter our initial lists of predicted targets, which limited us to discovering TAL effector targets that were represented on the Gene Chip used. Similar analysis using RNA-seq transcriptomics data would likely have yielded more candidate targets.

The TAL effector-target scoring function used in this analysis also yields high rate of false positives, identifying hundreds of candidate sites per TAL effector in the rice genome. Therefore, filtering the predicted lists based on transcriptomics or other biological data is essential. Filtering the lists of predicted sites by characteristics such as score, rank, and
Appendix Figure 3. Determination of Os01g52130 as the relevant target of Tal2g using designer TAL effectors. (A) DNA sequence of the promoter regions of Tal2g induced genes Os06g46500 and Os01g52130 in rice cv. Nipponbare. The effector binding elements (EBEs) for Tal2g are in bold. The EBEs for designer TAL effectors dT434 targeting Os06g46500 and dT436 or dT437 targeting Os01g52130 are underlined and labeled above. Short arrows indicate transcriptional start sites and italics indicates the translational start site per the Rice Genome Annotation Project (Release 7, http://rice.plantbiology.msu.edu). (B) Activation of Os06g46500 and Os01g52130 by Tal2g and specific activation of Os06g46500 and Os01g52130 by dT434, and dT436 or dT437, respectively. Shown are the results of semi-quantitative RT-PCR of leaf RNA isolated 48 h after inoculation by infiltration with X. oryzae pv. oryzicola BLS256 (WT), the tal2g knockout derivative M27 carrying an empty plasmid vector (ev), M27 carrying the vector with the cloned tal2g gene, or M27 carrying the vector with coding sequences for dT436, dT436, or dT437 as indicated. An actin gene was used to equalize cDNA amounts. (C) Rescue of the virulence defect of M27 by dT436 or dT437 but not dT434, revealing Os01g52130 as the relevant target of Tal2g. Lesion lengths were measured 10 days after inoculation with the indicated strains. Values labeled with the same letter are not significantly different (P>0.01). Error bars represent standard deviation (N=10). Experiments were repeated twice with consistent results.

proximity to the transcriptional start site may also help separate functional TAL effector binding sites from false positives.

A weakness in our comparison of functional and non-functional TAL was that we did not distinguish between non-functional sites at which the TAL effector failed to bind the DNA and non-functional sites at which the TAL effector bound the DNA target but did not activate transcription. Therefore, dividing predicted sites into functional or non-functional categories based on binding site characteristics may be useful for applications in which TAL effectors are used to activate transcription or for identifying TAL effector targets in nature. However, these characteristics might not extend to identifying TAL effector targets in other contexts, such as...
TAL effector nucleases or TALENs, where proximity to a TATA box or transcriptional start site is less important. It is unclear if features such as rank or score ratio will be useful in other applications.

Overall, we anticipate that our approach will be useful for studying the role of TAL effectors in a variety of *Xanthomonas*-host plant systems. One bottleneck in this approach is the characterization and sequencing of TAL effectors from *Xanthomonas* populations. Sequencing of TAL effector repeat regions is difficult, as next-generation sequencing technologies do not produce reads long enough to sequence through a TAL effector’s repeats. However, as sequencing improves and more complete *Xanthomonad* genomes become available, the approach described here should prove extremely valuable for identifying TAL effector targets and understanding host-bacterial disease interactions.

**Materials and Methods**

**Plant materials**

Rice (*Oryza sativa* L.) plants were grown in LC-1 soil mixture (SunGRO, Bellevue, WA, U.S.A.) in a growth chamber under cycles of 12 h of light (28°C) and 12 h of dark (25°C). Fertilizer (Peters Professional, St. Louis) and iron chelate micronutrient (Becker Underwood, Ames, IA, U.S.A.) were applied twice a week during watering at 0.25 and 4.5 g/l, respectively.

**Bacterial strains and preparation of inoculum**

*Xoo* strain PXO99 and *Xoc* strain BLS256 were used. For each, a single colony from a glucose yeast extract (GYE; 20 g glucose, 10 g yeast extract per liter) agar plate was transferred to 5 ml of GYE liquid medium and incubated for 24 h at 28°C with constant shaking at 250 r.p.m. Subsequently, 2 ml of this culture were transferred to 300 ml of fresh GYE liquid medium and incubated as above for an additional 18 h. Cells were pelleted by centrifugation at 4000 r.p.m. for 10 min, washed twice and resuspended in sterile 10 mm MgCl₂ to an OD₆₀₀ of 0.5, yielding 7 liters of inoculum. Tween-20 was added to a final concentration of 0.5%.

**Bacterial inoculation: vacuum infiltration method**

Fourteen days after sowing, and 2 h after the beginning of the light period, flats of seedlings were inverted into inoculum and vacuum infiltrated at 500 mm Hg for two minutes, two consecutive times, in a custom-made chamber. A total of three flats were used; each flat was inoculated with one of *Xoo*, *Xoc* or mock inoculum (sterile 10 mm MgCl₂, 0.5% Tween-20). Following inoculation, plants were returned to the growth chamber. During growth, plants were cycled weekly in the same order through three growth chambers. In this way all plants could be
incubated in the same final chamber following inoculation. For each replicate, position of the flats in the growth chamber (from left to right, used also for the order of inoculation and harvest) was determined by a split-plot design.

**Tissue collection**

Within flats, assignment of pots for the harvest of inoculated tissue at 2, 4, 8, 24 and 96 h was made using a separate randomized complete block design. Tissue was harvested from plants in four pots (approximately 2 g fresh weight) for each time point per inoculum per replication. Harvested tissue was immediately frozen in liquid nitrogen and stored at -80°C until processing.

**RNA Isolation and processing for microarray experiments**

Total RNA was isolated from frozen whole plant tissue, purified using RNeasy spin columns (Qiagen), and adjusted to a final concentration of 1 µg/µL. Probe synthesis, labeling, and hybridization were performed at the Iowa State University GeneChip Core facility, according to their standard protocols. GeneChips (the Affymetrix Rice Genome Array) were washed and stained with streptavidin-phycoerythrin using the fluidics script EukGE-WS2v5_450 in the Affymetrix GeneChip Fluidics Station 450. Stained chips were immediately scanned with the Affymetrix GeneChip Scanner 3000 7G.

**Microarray data scaling and normalization**

The hybridization data were processed using Affymetrix GeneChip Operating Software (GCOS v1.4.0.036). Scanned images were examined for any visible defects and satisfactory image files were analyzed to generate raw data files using the default settings of GCOS from Affymetrix. Signal measures on each GeneChip were scaled to a target intensity of 500, a normalization value of 1, and the default parameter settings for the Rice Genome Array were used. The raw microarray data were uploaded to the Plant Expression Database (PLEXdb, http://plexdb.org/); experiment name OS3.

**Microarray analysis**

The microarray experiment was analyzed as a split-plot design with four replications. The four replications were obtained on four separate occasions at 1-week intervals using a single growth chamber. Within each replication, one tray of plants was used for each of the three treatments (Mock, Xoc, and Xoo). The five collection timepoints were randomly assigned to rows of plants within each tray. Thus, an experiment’s treatment and time serve as the whole-plot and split-plot treatment factors, respectively.
A mixed linear model was fit separately to RMA-normalized data from each gene. Each mixed linear model included fixed effects for replications, treatments, times, and treatment-by-time interaction, as well as random effects for the trays serving as whole-plot experimental units. As part of our mixed linear model analysis for each probe set, we tested for a difference in the pattern of expression over time between Mock and Xoc, between Mock and Xoo, and between Xoc and Xoo. These tests are designed to detect genes whose changes in expression across time differed between treatments. The 57,381 p-values from each of these three tests were converted to q-values (23). Genes with q-values less than 0.30 were identified as significantly differentially expressed. This method controls that false discovery rate (FDR) at approximately 30% for each comparison.

**Cloning TAL effector genes of X. oryzae pv. oryzicola**

Two micrograms of genomic DNA of X. oryzae pv. oryzicola strain BLS256 were digested with BamHI and separated in 1% agarose by electrophoresis. DNA fragments corresponding from 2 to 5 kb were gel purified and ligated into pBlueScript SK(-) (Agilent Technologies) previously digested with BamHI. The ligation was transformed into E. coli TOP10 cells and colonies encoding TAL effector genes were identified by colony PCR using oligos P270 (5’-GCCAAGTCCTGCCCGCG-3’) and P271 (5’-CCTCCAGGGGCGGTGC-3’), which target a conserved N-terminus region of Xanthomonas oryzae TAL effectors. The corresponding SphI fragments of each Xoc TAL effector encoding the repeat region were cloned into the entry vector pCS466 and confirmed by 5’ and 3’ sequencing with oligos p235 (5’-GGAGGCCTTGCTCACGGATGC-3’) and p236 (5’-GGCCGGTGACAGCACGATCCG-3’), respectively. In the particular cases of tal2g, tal4a and tal8, we used BamHI fragments instead, because those genes have only one of the SphI sites. Xoc TAL effectors in pCS466 were recombined to the broad host-range destination vector pKEB31 using Gateway LR Clonase (Invitrogen) (24).

**Determination of bacterial virulence and growth in rice leaves**

Virulence tests of Xoc BLS256 and TAL effector mutants were conducted on 3 to 4 week-old leaves of rice cv. Nipponbare. The lesion lengths were recorded as the sum of water-soaked expansions in both directions from the infiltration site. Lesion lengths were measured 10 to 14 days after bacterial infiltration.

To measure the multiplication of wild type BLS256 and TAL effector mutants in planta, three leaves per treatment at defined time-points were collected. Leaf sections corresponding to five centimeters on either side of the initial infiltration site were sliced into small pieces and ground in
2 ml of water. Samples were diluted serially in sterile water and spread on peptone sucrose agar (PSA) supplemented with cephalexin at 20 mg/L. Plates were incubated at 28°C until appearance of single colonies, and colonies were counted. The results are displayed as the mean and standard deviation of 8 replicates for this experimental procedure. For bacterial counting on the leaf surface, the water-soaked leaf areas (approximately 1 cm²) were washed thoroughly two times with 50 μl of sterile water and then diluted into 1 ml of water. Viable counts were determined by serial dilution as described before. Bacterial counting experiments were repeated at least three times with consistent results.

**Computational prediction of TAL effector targets**

Promoter sequences (defined as 1000 bases upstream of the start codon) for approximately 60,000 rice genes were downloaded from the MSU Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/). For each of the approximately 40 Xoo and Xoc TAL effectors, we used our previously published TAL effector-target scoring model (21) to identify the best-scoring site for the TAL effector in each of the 60,000 promoter sequences. Based on the list of best-scoring sites, an outlier score cutoff was computed for each TAL effector. The cutoff was defined as the 25th percentile – 1.5*Interquartile range. Promoters were rescanned for each TAL effector to identify all sites scoring below the TAL effector’s cutoff (a lower score indicates better binding affinity). All identified candidate binding sites were required to be directly preceded by a T at the 5’ end. The lists of candidate binding sites for each TAL effector were cross-referenced to the lists of up-regulated genes from the microarray experiments. Sites scoring below the cutoff for an Xoc TAL effector in promoters of genes differentially expressed in the Xoc-mock comparison were identified as candidate targets for Xoc TALs. Similarly, sites scoring below the cutoff for an Xoo TAL effector in promoters of genes differentially expressed in the Xoo-mock comparison were identified as candidate targets for Xoo TALs. Genes possessing such sites were considered candidate TAL effector-targeted genes.

**RNA extraction and RT-PCR analyses**

RNA was prepared from leaf tissue collected at 48 h after infiltration using the RNeasy Mini Kit (Qiagen). The RNA was subjected to in-column DNase treatment with the RNase-Free DNase Set (Qiagen). 2μg of total RNA were reverse-transcribed using SuperScript III (Invitrogen) and standard oligo dT₂₀. For the semi-quantitative RT-PCR analysis, 1 μL (5%) of the RT-PCR reaction was used as a template with Phire Hot Start II DNA polymerase (Thermo Scientific) together with the gene-specific oligos. Cycling parameters were 30 s at 98°C; 23 cycles of 10 s at 98°C, 5 s at 60°C, and 10 s at 72°C.
**Designer TAL effectors**

To design TAL effectors targeting the promoter regions of Os01g52130 and Os06g46500, we used the TAL Effector Nucleotide Targeter 2.0 website (21). dTALes were first assembled into an entry vector (25) and subsequently transferred to pKEB31 destination vector by Gateway LR Clonase (Invitrogen) to generate dT434, dT436 and dT437.

**References**


8. Sugio, A., Yang, B., Zhu, T. and White, F.F. (2007) Two type III effector genes of *Xanthomonas oryzae pv. oryzae* control the induction of the host genes OsTFIIA(\(\gamma\))1 and OsTFX1 during bacterial blight of rice. Proc Natl Acad Sci U S A.


Tables

Appendix Table 1. Predicted X. oryzae pv. oryzicola BL256 TAL effector targets in rice and target verification.

<table>
<thead>
<tr>
<th>TAL effector</th>
<th>Locus ID</th>
<th>EBE location</th>
<th>EBE score</th>
<th>EBE score ratio</th>
<th>EBE rank</th>
<th>TAL knockout</th>
<th>Xag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tal1c</td>
<td>LOC_Os07g06970</td>
<td>-24</td>
<td>5.97</td>
<td>1.22</td>
<td>0</td>
<td>-</td>
<td>+</td>
<td>HEN1</td>
</tr>
<tr>
<td>Tal1c</td>
<td>LOC_Os03g03034</td>
<td>-114</td>
<td>19.55</td>
<td>1.83</td>
<td>0</td>
<td>-</td>
<td>+</td>
<td>Flavonol synthase/flavanone 3-hydroxylase</td>
</tr>
<tr>
<td>Tal2d</td>
<td>LOC_Os04g49194</td>
<td>-26</td>
<td>8.22</td>
<td>1.26</td>
<td>0</td>
<td>-</td>
<td>+</td>
<td>Naringenin,2-oxoglutarate 3-dioxygenase</td>
</tr>
<tr>
<td>Tal2g</td>
<td>LOC_Os01g52130</td>
<td>-58</td>
<td>15.60</td>
<td>1.78</td>
<td>77</td>
<td>-</td>
<td>+</td>
<td>Sulfate transporter</td>
</tr>
<tr>
<td>Tal2g</td>
<td>LOC_Os06g46500</td>
<td>-59</td>
<td>15.73</td>
<td>1.8</td>
<td>117</td>
<td>-</td>
<td>+</td>
<td>Monocopper oxidase</td>
</tr>
<tr>
<td>Tal3b</td>
<td>LOC_Os05g27590</td>
<td>-33</td>
<td>11.40</td>
<td>1.88</td>
<td>2</td>
<td>-</td>
<td>+</td>
<td>Wound-induced protein W112</td>
</tr>
<tr>
<td>Tal3b</td>
<td>LOC_Os02g34970</td>
<td>-29</td>
<td>15.33</td>
<td>2.53</td>
<td>75</td>
<td>-</td>
<td>+</td>
<td>No apical meristem protein</td>
</tr>
<tr>
<td>Tal3b</td>
<td>LOC_Os07g36430</td>
<td>-31</td>
<td>15.78</td>
<td>2.6</td>
<td>108</td>
<td>-</td>
<td>+</td>
<td>Expressed protein</td>
</tr>
<tr>
<td>Tal3c</td>
<td>LOC_Os02g47660</td>
<td>63</td>
<td>10.93</td>
<td>1.92</td>
<td>53</td>
<td>-</td>
<td>+</td>
<td>Basic helix-loop-helix</td>
</tr>
<tr>
<td>Tal3c</td>
<td>LOC_Os03g07540</td>
<td>-99</td>
<td>12.33</td>
<td>2.17</td>
<td>350</td>
<td>-</td>
<td>+</td>
<td>bHLH family protein</td>
</tr>
<tr>
<td>Tal4a</td>
<td>LOC_Os03g37840</td>
<td>-151</td>
<td>15.58</td>
<td>1.52</td>
<td>0</td>
<td>-</td>
<td>+</td>
<td>Potassium transporter</td>
</tr>
<tr>
<td>Tal4b</td>
<td>LOC_Os09g32100</td>
<td>-84</td>
<td>8.15</td>
<td>1.5</td>
<td>16</td>
<td>-</td>
<td>+</td>
<td>Expressed protein</td>
</tr>
<tr>
<td>Tal4c</td>
<td>LOC_Os06g37080</td>
<td>-39</td>
<td>14.64</td>
<td>1.78</td>
<td>0</td>
<td>-</td>
<td>+</td>
<td>L-ascorbate oxidase precursor</td>
</tr>
<tr>
<td>Tal5a</td>
<td>LOC_Os02g15290</td>
<td>-30</td>
<td>23.32</td>
<td>1.88</td>
<td>107</td>
<td>-</td>
<td>+</td>
<td>VQ domain containing protein</td>
</tr>
</tbody>
</table>

1 Location relative to the start codon
2 Score calculated as in (21)
3 Score divided by Best Possible Score (the score for the TAL effector on its perfect target)
4 EBE rank is out of the collection of all sites in the promoterome scoring below the cutoff for the TAL effector
5 + and – indicate that the corresponding TAL effector mutant induced or failed to induce the gene, respectively. ND, no data.
6 + and – indicate that delivery of the corresponding TAL effector by Xag induced or failed to induce gene expression. ND, no data.
Appendix Table 1. Continued.

<table>
<thead>
<tr>
<th>TAL effector</th>
<th>Locus ID</th>
<th>EBE location</th>
<th>EBE score</th>
<th>EBE score ratio</th>
<th>EBE rank</th>
<th>TAL knockout</th>
<th>Xag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tal6</td>
<td>LOC_Os12g42970</td>
<td>-30</td>
<td>16.84</td>
<td>2.14</td>
<td>139</td>
<td>-</td>
<td>+</td>
<td>GATA zinc finger domain containing protein</td>
</tr>
<tr>
<td>Tal6</td>
<td>LOC_Os09g29100</td>
<td>83</td>
<td>17.00</td>
<td>2.13</td>
<td>167</td>
<td>-</td>
<td>+</td>
<td>Cyclin</td>
</tr>
<tr>
<td>Tal6</td>
<td>LOC_Os01g31220</td>
<td>19</td>
<td>18.75</td>
<td>2.38</td>
<td>685</td>
<td>-</td>
<td>+</td>
<td>Expressed protein</td>
</tr>
<tr>
<td>Tal9b</td>
<td>LOC_Os01g51040</td>
<td>-47</td>
<td>14.07</td>
<td>2.81</td>
<td>275</td>
<td>-</td>
<td>+</td>
<td>Transmembrane protein 16K</td>
</tr>
<tr>
<td>Tal1c</td>
<td>LOC_Os06g14750</td>
<td>-521</td>
<td>12.00</td>
<td>2.44</td>
<td>256</td>
<td>+</td>
<td>-</td>
<td>Phosphatidylinositol I-4-phosphate 5-Kinase family protein</td>
</tr>
<tr>
<td>Tal2a</td>
<td>LOC_Os02g43760</td>
<td>22</td>
<td>15.87</td>
<td>1.75</td>
<td>21</td>
<td>+</td>
<td>+</td>
<td>Ubiquitin carboxyl-terminal hydrolase, family 1</td>
</tr>
<tr>
<td>Tal2a</td>
<td>LOC_Os06g14750</td>
<td>-527</td>
<td>17.59</td>
<td>1.94</td>
<td>88</td>
<td>+</td>
<td>-</td>
<td>Phosphatidylinositol I-4-phosphate 5-Kinase family protein</td>
</tr>
<tr>
<td>Tal2d</td>
<td>LOC_Os01g51040</td>
<td>-315</td>
<td>14.32</td>
<td>2.19</td>
<td>324</td>
<td>+</td>
<td>-</td>
<td>Transmembrane protein 16K</td>
</tr>
<tr>
<td>Tal3a</td>
<td>LOC_Os05g12450</td>
<td>-818</td>
<td>16.07</td>
<td>2.01</td>
<td>294</td>
<td>+</td>
<td>-</td>
<td>Hydroquinone glucosyltransferase</td>
</tr>
<tr>
<td>Tal3a</td>
<td>LOC_Os07g06970</td>
<td>-567</td>
<td>16.21</td>
<td>2.03</td>
<td>354</td>
<td>+</td>
<td>-</td>
<td>HEN1</td>
</tr>
<tr>
<td>Tal3b</td>
<td>LOC_Os03g03034</td>
<td>-142</td>
<td>16.73</td>
<td>2.76</td>
<td>258</td>
<td>+</td>
<td>-</td>
<td>Flavonol synthase/flavanone 3-hydroxylase</td>
</tr>
<tr>
<td>Tal4a</td>
<td>LOC_Os09g20220</td>
<td>-152</td>
<td>28.93</td>
<td>2.83</td>
<td>280</td>
<td>+</td>
<td>-</td>
<td>Glutathione S-transferase</td>
</tr>
<tr>
<td>Tal4a</td>
<td>LOC_Os01g27210</td>
<td>-310</td>
<td>29.22</td>
<td>2.85</td>
<td>341</td>
<td>+</td>
<td>-</td>
<td>Glutathione S-transferase</td>
</tr>
<tr>
<td>Tal4c</td>
<td>LOC_Os02g47660</td>
<td>-30</td>
<td>24.73</td>
<td>3.01</td>
<td>434</td>
<td>+</td>
<td>-</td>
<td>Basic helix-loop-helix</td>
</tr>
<tr>
<td>Tal4c</td>
<td>LOC_Os07g29750</td>
<td>-610</td>
<td>25.09</td>
<td>3.05</td>
<td>557</td>
<td>+</td>
<td>-</td>
<td>Glycosyl hydrolases family 16</td>
</tr>
<tr>
<td>Tal6</td>
<td>LOC_Os07g47790</td>
<td>8</td>
<td>13.38</td>
<td>1.7</td>
<td>8</td>
<td>+</td>
<td>-</td>
<td>AP2 domain containing protein</td>
</tr>
<tr>
<td>Tal6</td>
<td>LOC_Os02g14770</td>
<td>-30</td>
<td>18.48</td>
<td>2.35</td>
<td>569</td>
<td>+</td>
<td>-</td>
<td>Phosphoenolpyruvate carboxylase</td>
</tr>
<tr>
<td>Tal7</td>
<td>LOC_Os03g25490</td>
<td>-341</td>
<td>16.32</td>
<td>2.01</td>
<td>494</td>
<td>+</td>
<td>-</td>
<td>Cytochrome P450 72A1</td>
</tr>
<tr>
<td>Tal7</td>
<td>LOC_Os02g43760</td>
<td>-560</td>
<td>16.45</td>
<td>2.03</td>
<td>547</td>
<td>+</td>
<td>-</td>
<td>Ubiquitin carboxyl-terminal hydrolase, family 1</td>
</tr>
</tbody>
</table>

---

1 Location relative to the start codon
2 Score calculated as in (21)
3 Score divided by Best Possible Score (the score for the TAL effector on its perfect target)
4 EBE rank is out of the collection of all sites in the promoterome scoring below the cutoff for the TAL effector
5 + and – indicate that the corresponding TAL effector mutant induced or failed to induce the gene, respectively. ND, no data.
6 + and – indicate that delivery of the corresponding TAL effector by Xag induced or failed to induce gene expression. ND, no data.
Appendix Table 1. Continued.

<table>
<thead>
<tr>
<th>TAL effector</th>
<th>Locus ID</th>
<th>EBE location (^1)</th>
<th>EBE score (^2)</th>
<th>EBE score ratio (^3)</th>
<th>EBE rank (^4)</th>
<th>TAL knock-out (^5)</th>
<th>Xag (^6)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tal8</td>
<td>LOC_Os06g37080</td>
<td>-191</td>
<td>19.92</td>
<td>2.32</td>
<td>605</td>
<td>+</td>
<td>-</td>
<td>L-ascorbate oxidase precursor</td>
</tr>
<tr>
<td>Tal9a</td>
<td>LOC_Os12g01490</td>
<td>-621</td>
<td>19.71</td>
<td>2.56</td>
<td>205</td>
<td>+</td>
<td>-</td>
<td>MYB family transcription factor</td>
</tr>
<tr>
<td>Tal9a</td>
<td>LOC_Os11g01480</td>
<td>-653</td>
<td>19.71</td>
<td>2.56</td>
<td>206</td>
<td>+</td>
<td>-</td>
<td>MYB family transcription factor</td>
</tr>
<tr>
<td>Tal11a</td>
<td>LOC_Os03g05370</td>
<td>-422</td>
<td>17.39</td>
<td>2.74</td>
<td>307</td>
<td>+</td>
<td>-</td>
<td>Expressed protein</td>
</tr>
<tr>
<td>Tal11a</td>
<td>LOC_Os02g15290</td>
<td>-195</td>
<td>20.15</td>
<td>3.18</td>
<td>582</td>
<td>+</td>
<td>-</td>
<td>VQ domain containing protein</td>
</tr>
<tr>
<td>Tal11b</td>
<td>LOC_Os06g14750</td>
<td>527</td>
<td>12.44</td>
<td>2.77</td>
<td>129</td>
<td>+</td>
<td>-</td>
<td>Phosphatidylinositol I-4-phosphate 5-Kinase family protein</td>
</tr>
<tr>
<td>Tal11b</td>
<td>LOC_Os06g47950</td>
<td>-506</td>
<td>16.20</td>
<td>2.55</td>
<td>19</td>
<td>ND</td>
<td>ND</td>
<td>tetratricopeptide-like helical dehydrogenase</td>
</tr>
<tr>
<td>Tal2d</td>
<td>LOC_Os09g23560</td>
<td>-64</td>
<td>14.19</td>
<td>2.17</td>
<td>288</td>
<td>ND</td>
<td>ND</td>
<td>HSF-type DNA-binding domain containing protein</td>
</tr>
<tr>
<td>Tal3b</td>
<td>LOC_Os01g53220</td>
<td>-151</td>
<td>17.72</td>
<td>2.92</td>
<td>611</td>
<td>ND</td>
<td>ND</td>
<td>retrotransposon protein, unclassified</td>
</tr>
<tr>
<td>Tal4b</td>
<td>LOC_Os07g34510</td>
<td>-951</td>
<td>8.88</td>
<td>1.63</td>
<td>33</td>
<td>ND</td>
<td>ND</td>
<td>plastocyanin-like domain containing protein</td>
</tr>
<tr>
<td>Tal7</td>
<td>LOC_Os02g15710</td>
<td>-243</td>
<td>15.75</td>
<td>1.94</td>
<td>265</td>
<td>ND</td>
<td>ND</td>
<td></td>
</tr>
</tbody>
</table>

\(^1\) Location relative to the start codon  
\(^2\) Score calculated as in (21)  
\(^3\) Score divided by Best Possible Score (the score for the TAL effector on its perfect target  
\(^4\) EBE rank is out of the collection of all sites in the promoterome scoring below the cutoff for the TAL effector  
\(^5\) + and – indicate that the corresponding TAL effector mutant induced or failed to induce the gene, respectively. ND, no data.  
\(^6\) + and – indicate that delivery of the corresponding TAL effector by Xag induced or failed to induce gene expression. ND, no data.

Appendix Table 2. Predicted X. oryzae pv. oryzae PXO00A TAL effector targets in rice

<table>
<thead>
<tr>
<th>TAL effector</th>
<th>Locus ID</th>
<th>EBE location (^1)</th>
<th>EBE score (^2)</th>
<th>EBE score ratio (^3)</th>
<th>EBE rank (^4)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PthXo1</td>
<td>LOC_Os08g42350</td>
<td>-78</td>
<td>15.84</td>
<td>1.95</td>
<td>0</td>
<td>nodulin MN3 family protein precursor</td>
</tr>
<tr>
<td>PthXo6</td>
<td>LOC_Os01g52610</td>
<td>-64</td>
<td>22.66</td>
<td>1.65</td>
<td>197</td>
<td>peroxidase precursor</td>
</tr>
<tr>
<td>PthXo6</td>
<td>LOC_Os06g48010</td>
<td>-32</td>
<td>17.16</td>
<td>2.08</td>
<td>2</td>
<td>peroxidase precursor</td>
</tr>
</tbody>
</table>

\(^1\) Location relative to the start codon  
\(^2\) Score calculated as in (21)  
\(^3\) Score divided by Best Possible Score (the score for the TAL effector on its perfect target  
\(^4\) EBE rank is out of the collection of all sites in the promoterome scoring below the cutoff for the TAL effector
### Appendix Table 2. Continued.

<table>
<thead>
<tr>
<th>TAL effector</th>
<th>Locus ID</th>
<th>Location</th>
<th>EBE score</th>
<th>EBE score ratio</th>
<th>EBE rank</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PthXo6</td>
<td>LOC_Os07g11510</td>
<td>50</td>
<td>21.24</td>
<td>1.72</td>
<td>54</td>
<td>RAL6 - Seed allergenic protein RA5/RA14/RA17 precursor</td>
</tr>
<tr>
<td>PthXo6</td>
<td>LOC_Os08g44220</td>
<td>-654</td>
<td>22.50</td>
<td>1.66</td>
<td>169</td>
<td>GLTP domain containing protein bZIP transcription factor domain</td>
</tr>
<tr>
<td>PthXo6</td>
<td>LOC_Os09g29820</td>
<td>-30</td>
<td>15.61</td>
<td>2.33</td>
<td>1</td>
<td>containing protein calcium-transporting ATPase, plasma membrane-type</td>
</tr>
<tr>
<td>PthXo6</td>
<td>LOC_Os10g28240</td>
<td>-51</td>
<td>22.32</td>
<td>1.67</td>
<td>148</td>
<td>expressed protein</td>
</tr>
<tr>
<td>PthXo6</td>
<td>LOC_Os11g42950</td>
<td>-690</td>
<td>23.80</td>
<td>1.60</td>
<td>460</td>
<td>transcription initiation factor IIA gamma chain</td>
</tr>
<tr>
<td>PthXo7</td>
<td>LOC_Os01g73890</td>
<td>-29</td>
<td>15.34</td>
<td>2.04</td>
<td>1</td>
<td>transcription initiation factor IIA gamma chain</td>
</tr>
<tr>
<td>PthXo7</td>
<td>LOC_Os01g73890</td>
<td>-441</td>
<td>25.93</td>
<td>1.43</td>
<td>561</td>
<td>flavonol synthase/flavanone 3-hydroxylase</td>
</tr>
<tr>
<td>PthXo7</td>
<td>LOC_Os03g03034</td>
<td>-267</td>
<td>32.39</td>
<td>1.32</td>
<td>763</td>
<td>unknown</td>
</tr>
<tr>
<td>PthXo7</td>
<td>LOC_Os06g03080</td>
<td>-356</td>
<td>32.17</td>
<td>1.32</td>
<td>420</td>
<td>OsFBL36 - F-box domain and LRR containing protein helix-loop-helix</td>
</tr>
<tr>
<td>PthXo7</td>
<td>LOC_Os07g17460</td>
<td>-301</td>
<td>32.24</td>
<td>1.32</td>
<td>330</td>
<td>DNA-binding domain containing protein</td>
</tr>
<tr>
<td>Tal2a</td>
<td>LOC_Os01g01840</td>
<td>-709</td>
<td>11.12</td>
<td>2.18</td>
<td>142</td>
<td>RAL6 - Seed allergenic protein RA5/RA14/RA17 precursor</td>
</tr>
<tr>
<td>Tal2a</td>
<td>LOC_Os07g11510</td>
<td>46</td>
<td>10.04</td>
<td>2.50</td>
<td>32</td>
<td>RAL6 - Seed allergenic protein RA5/RA14/RA17 precursor</td>
</tr>
<tr>
<td>Tal2a</td>
<td>LOC_Os08g26620</td>
<td>-99</td>
<td>12.38</td>
<td>1.95</td>
<td>482</td>
<td>OsSub6 - Putative Subtilisin homologue</td>
</tr>
<tr>
<td>Tal4</td>
<td>LOC_Os01g58240</td>
<td>-449</td>
<td>11.25</td>
<td>2.36</td>
<td>72</td>
<td>cytochrome b5-like Heme/Steroid binding domain containing protein P450</td>
</tr>
<tr>
<td>Tal4</td>
<td>LOC_Os02g43360</td>
<td>-366</td>
<td>11.94</td>
<td>2.19</td>
<td>166</td>
<td>RALF28 - Rapid ALkalinization Factor RALF family protein precursor</td>
</tr>
<tr>
<td>Tal4</td>
<td>LOC_Os04g27020</td>
<td>-598</td>
<td>12.14</td>
<td>2.15</td>
<td>213</td>
<td>RALF28 - Rapid ALkalinization Factor RALF family protein precursor</td>
</tr>
<tr>
<td>Tal4</td>
<td>LOC_Os06g29730</td>
<td>-369</td>
<td>10.47</td>
<td>2.63</td>
<td>26</td>
<td>RALF28 - Rapid ALkalinization Factor RALF family protein precursor</td>
</tr>
</tbody>
</table>

1. Location relative to the start codon
2. Score calculated as in (21)
3. Score divided by Best Possible Score (the score for the TAL effector on its perfect target)
4. EBE rank is out of the collection of all sites in the promotorome scoring below the cutoff for the TAL effector
Appendix Table 2. Continued.

<table>
<thead>
<tr>
<th>TAL effector</th>
<th>Locus ID</th>
<th>EBE location</th>
<th>EBE score</th>
<th>EBE score ratio</th>
<th>EBE rank</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tal5a</td>
<td>LOC_Os07g09190</td>
<td>162</td>
<td>13.98</td>
<td>1.66</td>
<td>578</td>
<td>transketolase</td>
</tr>
<tr>
<td>Tal5a</td>
<td>LOC_Os07g11510</td>
<td>65</td>
<td>14.43</td>
<td>1.62</td>
<td>803</td>
<td>RAL6 - Seed allergenic protein RA5/RA14/RA17 precursor</td>
</tr>
<tr>
<td>Tal5a</td>
<td>LOC_Os09g27580</td>
<td>-303</td>
<td>13.73</td>
<td>1.68</td>
<td>501</td>
<td>potassium transporter</td>
</tr>
<tr>
<td>Tal5a</td>
<td>LOC_Os11g42950</td>
<td>-38</td>
<td>11.21</td>
<td>1.98</td>
<td>32</td>
<td>expressed protein</td>
</tr>
<tr>
<td>Tal6a</td>
<td>LOC_Os09g29210</td>
<td>-31</td>
<td>16.86</td>
<td>1.97</td>
<td>91</td>
<td>purine permease</td>
</tr>
<tr>
<td>Tal7a/8a</td>
<td>LOC_Os01g68740</td>
<td>409</td>
<td>9.14</td>
<td>5.33</td>
<td>0</td>
<td>keratin, type I cytoskeletal 9</td>
</tr>
<tr>
<td>Tal7b/8b</td>
<td>LOC_Os01g40290</td>
<td>-31</td>
<td>12.55</td>
<td>1.95</td>
<td>2</td>
<td>expressed protein</td>
</tr>
<tr>
<td>Tal7b/8b</td>
<td>LOC_Os10g32700</td>
<td>232</td>
<td>18.49</td>
<td>1.49</td>
<td>279</td>
<td>hypersensitive-induced response protein</td>
</tr>
<tr>
<td>Tal9a</td>
<td>LOC_Os06g39120</td>
<td>246</td>
<td>17.01</td>
<td>1.50</td>
<td>103</td>
<td>expressed protein</td>
</tr>
<tr>
<td>Tal9a</td>
<td>LOC_Os07g06970</td>
<td>2</td>
<td>7.81</td>
<td>3.71</td>
<td>0</td>
<td>HEN1</td>
</tr>
<tr>
<td>Tal9b</td>
<td>LOC_Os01g58194</td>
<td>-239</td>
<td>27.52</td>
<td>1.56</td>
<td>259</td>
<td>protein disulfide isomerase flavonol synthase/flavanone</td>
</tr>
<tr>
<td>Tal9b</td>
<td>LOC_Os03g03034</td>
<td>16</td>
<td>26.59</td>
<td>1.59</td>
<td>130</td>
<td>3-hydroxylase O-methyltransferase</td>
</tr>
<tr>
<td>Tal9d</td>
<td>LOC_Os04g01470</td>
<td>-696</td>
<td>9.36</td>
<td>1.92</td>
<td>691</td>
<td>unknown</td>
</tr>
<tr>
<td>Tal9d</td>
<td>LOC_Os11g28740</td>
<td>-637</td>
<td>9.33</td>
<td>1.92</td>
<td>626</td>
<td>VQ domain containing protein</td>
</tr>
<tr>
<td>Tal9e</td>
<td>LOC_Os02g15290</td>
<td>-30</td>
<td>23.52</td>
<td>2.14</td>
<td>104</td>
<td>VQ domain containing protein</td>
</tr>
</tbody>
</table>

1 Location relative to the start codon
2 Score calculated as in (21)
3 Score divided by Best Possible Score (the score for the TAL effector on its perfect target)
4 EBE rank is out of the collection of all sites in the promoterome scoring below the cutoff for the TAL effector
ACKNOWLEDGEMENTS

I would like to thank my dissertation advisor, Adam Bogdanove, for his continual mentoring and support throughout my graduate experience. Our regular meetings, conversations about this work and life outside the lab, and his assistance in preparing these manuscripts have been invaluable. I also greatly appreciate his efforts to involve me collaborations with other labs and his support to present my work at conferences and meetings. I am especially grateful to my dissertation committee members Dan Nettleton, Sue Lamont, Leonor Leandro, and Steve Whitham, for suggestions and advice throughout my time at Iowa State. Sue Lamont has also been an honest and inspiring mentor, both personally and professionally. Without Clarice Schmidt’s technical skills and assistance, I would not have been able to complete my research. Aaron Hummel, Andres Cernadas, and Li Wang have also given me help on countless occasions. Nick Booher and John VanDyk provided all of the needed technical support for the computational work presented in this dissertation and for the TALE-NT 2.0 website, often at the last minute. Volker Brendel, Daniel Standage, and Divya Mystry provided additional support for developing the website. I would also like to acknowledge collaborators Dan Voytas, Barry Stoddard, and Phil Bradley, who were involved in the design and execution of much of the work presented here. Their comments and suggestions have helped to improve and shape these manuscripts. Discussions with Matt Moscou early in my graduate school experience have strongly influenced my approach to solving computational problems. Past and current members of the Boglab have made my time at Iowa State an enjoyable experience. Finally, my husband Jesse has been a supportive and understanding partner throughout this journey. Without his constant patience and encouragement I would not have been able to complete this work.