Sequence diversity and functional conformity: A comparative molecular characterization of TALE-like proteins

Kumulative Dissertation

Der Fakultät für Biologie an der
Ludwig-Maximillians-Universität München
Zur Erlangung des
Doktorgrades der Naturwissenschaften (Dr. rer. Nat.)

Vorgelegt von
Orlando de Lange
Aus London, Vereinigtes Königreich

München, den 13. Juli 2015

Erstgutachter: Prof. Dr. Thomas Lahaye
Zweitgutachter: Prof. Dr. Martin Parniske

Dissertation eingereicht am: 13.07.2015
Tag der mündlichen Prüfung: 16.10.2015

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt ist.

München, den ....(Orlando de Lange)

Erklärung

Hiermit erkläre ich, dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen
Prüfungskommission vorgelegt worden ist. Ich habe nicht versucht eine Dissertation einzureichen oder mich der Doktorprüfung zu unterziehen.

München, den ....(Orlando de Lange)

Table of Contents

Figure list ii
List of abbreviations iii
Publications iv
Contributions of the author v

Summary 1

1 Introduction
  1.1 Molecular plant pathology in the context of food security for the 21st century 5
  1.2 Introduction to TALE Biology 6
  1.3 The DNA binding properties of TALEs 7
  1.4 TALEs in the context of plant disease 12
  1.5 Introduction to TALE technology 14
  1.6 TALE-like proteins have so far been identified in three sources outside of *Xanthomonas* bacteria 17
    1.6.1 Prior knowledge about *Ralstonia* TALE-likes
    1.6.2 Prior knowledge about *Burkholderia* TALE-likes
    1.6.7 Prior knowledge about Marine Organism TALE-likes

2 Aims of this study
  2.1: Conformity to the TALE paradigm? 22
  2.2 Unifying and defining features of the TALE-likes? 24
  2.3 What is known about TALE evolution? 24
  2.4 New frontiers of TALE technology 25

3 Breaking the DNA-binding code of *Ralstonia solanacearum* TAL effectors provides new possibilities to generate plant resistance genes against bacterial wilt disease 28

4 Programmable DNA-binding proteins from *Burkholderia* provide a fresh perspective on the TALE-like repeat domain 43

5 DNA-binding proteins from marine bacteria make novel contributions to the sequence diversity of TALE-like repeats 57

6 Discussion
  6.1 Overview 85
  6.2 Placing this work into the context of similar studies 85
  6.3 Addressing the goals of this work
    6.3.1 Partially conformity to the TALE paradigm 88

6.3.2 A unifying definition of the TALE-likes 93
6.3.3 Insights into TALE and TALE-like evolution 94
6.3.4 Practical and conceptual contributions to TALE technology 95

7 Concluding remarks 97

8 Bibliography 98

9 Appendices

9.1 Supplementary data – Chapter 3 106
9.1 Supplementary data – Chapter 4 154
9.1 Supplementary data – Chapter 5 198

10 Acknowledgements 219

11 Curriculum Vitae 220

Figure List

1.1 An overview of TALE functional domains and TALE-DNA interactions

1.2 AvrBs3 canonical repeat alignment

1.3 Structure of TALE repeat array bound to DNA

1.4 Structures of individual TALE repeats bound to DNA

1.5 TALE complements of three *Xanthomonas* strains

1.6 Overview of TALE fusion protein designs

1.7 Brg11 canonical repeat alignment

1.8 Domain composition (a) and consensus repeat sequences (b) of representatives of each TALE-like group addressed in this work.

2.1 TALE repeat sequence logo

Acronyms and Abbreviations Used In This Thesis

TALE – Transcription activator like effector

TALEN – TALE nuclease

dTALE – designer TALE

BSR – Base specifying residue

RVD – Repeat variable diresidue

AD – Activation domain

AAD – Acidic activation domain

NLS – Nuclear localization signal

NTR – N-terminal non-repetitive region

CTR – C-terminal non-repetitive region

G_{SL} – Glycine short loop TALE repeat

RipTAL – *Ralstonia* injected protein transcription activator like

Bat – *Burkholderia* TALE-lie

MOrTL – Marine organism TALE-like

T3SS – Type III secretion system

T3E – Type III secreted effector

CRISPR - Clustered regularly interspaced short palindromic repeats

DNA – Deoxyribonucleic acid

EBE – Effector binding element

BE – binding element

bp – base pair

nt - nucleotide

pv – pathovar

Publications

1. de Lange, Orlando, Schreiber, Tom, Schandry, Niklas, Radeck, Jara, Braun, Karl Heinz, Koszinowski, Julia, Heuer, Holger, Strauß, Annett & Lahaye, Thomas

Breaking the DNA-binding code of *Ralstonia solanacearum* TAL effectors provides new possibilities to generate plant resistance genes against bacterial wilt disease

New Phytologist (2013) **3**: 773-786.

2. de Lange, Orlando, Wolf, Christina, Dietze, Jörn, Elsaesser, Janett, Morbitzer, Robert & Lahaye, Thomas

Programmable DNA-binding proteins from *Burkholderia* provide a fresh perspective on the TALE-like repeat domain.

Nucleic Acids Research (2014) **42** (11): 7436-7449.

3. de Lange, Orlando, Wolf, Christina, Thiel, Phillip, Krueger, Jens, Kohlbacher, Oliver & Lahaye, Thomas

DNA-binding proteins from marine bacteria make novel contributions to the sequence diversity of TALE-like repeats

Nucleic Acids Research (submitted 17.04.2015).

Declaration of contribution as coauthor

1. Breaking the DNA-binding code of *Ralstonia solanacearum* TAL effectors provides new possibilities to generate plant resistance genes against bacterial wilt disease

Joint conception of the study together with T. Schreiber, A. Strauß and T. Lahaye.
Molecular cloning of domain swap constructs, for figures 2-6, and S6.
Carrying out experiments shown in figures 2-7, 9 and S6 and S11.
Preparation of the manuscript, with assistance from N. Schandry, A. strauß and T. Lahaye.

………………………… …………………………
(Orlando de Lange) (Prof. Thomas Lahaye)

2. Programmable DNA-binding proteins from *Burkholderia* provide a fresh perspective on the TALE-like repeat domain.

Conception of the study with guidance from T. Lahaye and support from C. Wolf.
Design and molecular cloning of all constructs used in the study.
Carrying out experiments shown in figures 4, 8 and S9. C. Wolf carried out all EMSA, MST and flow cytometry experiments in this study.
Preparation of the manuscript with assistance from C. Wolf and T. Lahaye.

………………………… ..…………………………… …………………………
(Orlando de Lange) (Christina Wolf) (Prof. Thomas Lahaye)

3. DNA-binding proteins from marine bacteria make novel contributions to the sequence diversity of TALE-like repeats

Conception of the study with guidance from T. Lahaye and support from C. Wolf.
Design and molecular cloning of all constructs used in the study.
Carrying out experiments shown in figures 3 and 5. C. Wolf carried out all EMSA, MST and protein stability experiments in this study.
Preparation of the manuscript with assistance from C. Wolf and T. Lahaye.

………………………… ..…………………………… …………………………
(Orlando de Lange) (Christina Wolf) (Prof. Thomas Lahaye)

Summary

At least four phylogenetically distinct groups of bacteria encode repeat proteins with the common ability to bind specific DNA sequences with a unique but conserved code. Each repeat binds a single DNA base, and specificity is determined by the amino acid residue at position 13 of each repeat. Repeats are typically 33-35 amino acids long. Comparing repeat sequences across all groups reveals that only three positions are hyper-conserved. Repeats are in most cases functionally compatible such that they can be assembled together into a single chimeric array. This functional conformity and inter-compatibility is a result of structural conservation. Repeat arrays of these proteins have been demonstrated or predicted to form almost identical tertiary structures: a right-handed super helix that wraps around the DNA double strand with the base specifying residue of each repeat positioned in the major groove next to its cognate target base. The mechanism of DNA binding is conserved.

The first discovered group, providing the name for the rest, are the Transcription Activator Like Effectors (TALEs) of plant-pathogenic *Xanthomonas* bacteria. The eukaryotic transactivation domain, which lends this group their name, allows them to activate specifically targeted host genes for the benefit of the bacterial invader. The other groups, discovered after the TALEs, are the RipTALs of *Ralstonia solanacearum*, the Bats of *Burkholderia rhizoxinica,* and MOrTL1 and MOrTL2 of unknown marine bacteria. Together they are designated TALE-likes. Each designation contains some allusion to the TALEs. The term RipTAL stands for *Ralstonia* injected proteins TALE-like, the Bats are *Burkholderia* TALE likes, and the MOrTLs Marine Organism TALE-likes. This unity of terminology belies disunity in the lifestyles of these different bacteria, and the biological roles fulfilled by these proteins.

The TALEs have already been researched extensively. The code that describes the relationship between the base specifying residues and their cognate bases is often referred to as the TALE code. This code was deciphered by two groups independently and published in 2009, a year before I began my doctoral work. Since then research into TALEs has not slowed and a great deal has been learnt both about the native biology and biotechnological uses of TALEs. My work has been focused on the other TALE-like groups, none of which had been previously characterized in terms of DNA recognition properties, before I began my work.

RipTALs are effector proteins delivered during bacterial wilt disease caused by *R. solanacearum* strains. This devastating disease affects numerous crop species worldwide. Characterizing the molecular properties of the RipTALs provides a first step towards uncovering their role in the disease. The Bats and MOrTLs are primarily of interest as comparison groups to the TALEs and RipTALs and as

sources of sequence diversity for future efforts into TALE repeat engineering.

In the introduction of this dissertation, which explores TALE biology, a particular focus will be placed on the DNA binding properties of TALEs and how this can be put to use in TALE technology. After this the RipTALs, Bats and MOrTLs are each introduced, explaining what is known about their provenance and sequence features. The aims of my doctoral work are then listed and expounded in turn. The proximal goal of my doctoral work was to carry out a comparative molecular characterization of each group of non-TALE TALE-likes. In doing so we hoped to gain insights into the principles of TALE-like DNA-binding properties, evolutionary history of the different groups and their potential uses in biotechnology. In the case of the RipTALs this work should begin to unravel the role these proteins play in bacterial wilt disease, as a means to fight this devastating pathogen.

The articles I have worked on covering the molecular characterizations of RipTALs, Bats and MOrTLs are then presented in turn. Working together with others I was able to show that repeats from each group of TALE-likes mediate sequence specific DNA binding, revealing a conserved code in each case. This code links position 13 of any TALE-like repeat to a specific DNA base preference in a reliable fashion.

I will argue that the TALE-likes represent a fascinating case of conserved structure and function in a diverse sequence space. In addition the TALEs and RipTALs may simply represent one face of the TALE-likes, a protein family mediating as yet unknown biological roles as bacterial DNA binding proteins.

Zusammenfassung

Gene die ähnlichen repeat-enthaltende Proteine kodieren finden sich in den Genomen von mindestens vier phylogenetischen Gruppen von Bakterien. Alle diese verschiedene Proteine nutzen den gleichen, einzigartigen Code um bestimmte DNS Basen zu binden. Repeats bestehen in der Regel aus 33-35 Aminosäuren und jeder Repeat bindet eine einzige Base, wobei die Spezifität durch die Aminosäure an Position 13 im Repeat bestimmt wird. Ein Vergleich der Repeatsequenzen aus den verschiedenen Proteinen zeigt nur drei hyperkonservierte Aminosäuren, aber chimäre Repeat-domänen sind meistens dennoch Funktionsfähig. Funktionseinigkeit und Kompatibilität sind Folge struktureller Konservation. Es wurde schon gezeigt oder vorhergesagt dass die Repeat-Domäne dieser Proteine fast identische Strukturen bilden: Eine rechtsgängige Superhelix die sich um die DNS wickelt, wobei die basenbestimmenden Aminosäuren in der großen Furche neben der verwandte Base positioniert werden. Dieser Bindungsmechanismus ist innerhalb dieser Gruppe von Proteinen konserviert.

Die erste beschriebene Gruppe sind die Transcription Activator Like Effectors (TALEs) aus pflanzenkrankheitserregenden *Xanthomonas* Bakterien, und ist namensgebend für diese Klasse von Proteinen. Der Name stammt von der eukaryotischen Aktivierungsdomäne durch die ermöglicht wird, dass die TALEs spezifische Wirtsgene aktivieren zum Vorteil der Bakterien. Die anderen Gruppen sind RipTALs aus Ralstonia solanacearum, Bats aus Burkholderia rhizoxinica, und MOrTL1 und MOrTL2 aus bis jetzt unbekannten Meeresbakterien. Kollektivbegriff dafür ist TALE-likes. Der Name jeder Gruppe deutet darauf hin, dass diese Proteine TALEs ähnlich sind. RipTAL bedeutet *Ralstonia* injected proteins TALE-like; Bats sind *Burkholderia* TALE likes; und die MOrTLs Marine Organism TALE-likes. Diese Namenseinigkeit verbirgt die unterschiedliche Lebensweisen dieser Bakterien und die verschiedenen natürliche Rollen dass die TALE-likes annehmen könnten.

TALEs werden schon seit langem erforscht. Der TALE-Code beschreibt die Beziehung zwischen bestimmten Resten und passenden Basen. Er wurde in 2009 entschlüsselt und von zwei unabhängigen Forschungsgruppen veröffentlicht. Das war ein Jahr bevor ich mit meiner Doktorarbeit anfing. Seitdem hat die TALE-Forschung stark weiterentwickelt und vieles wurde über die natürliche Biologie und biotechnische Applikation dieser Proteine gelernt. Meine Arbeit fokussiert sich auf die anderen TALE-likes, von denen zuvor kein einziges charakterisiert war.

RipTALs sind Effektor-Proteine, die von *R. solanacearum* im Laufe der Entwicklung der bakterieller Schleimkrankheit sekretiert werden. Diese Krankheit betrifft weltweit viele wichtige Nutz- und Kulturpflanzen. Eine Charakterisierung der molekularen Eigenschaften der RipTALs ist der erste Schritt um die Rolle dieser Proteine in der Pathogenese zu verstehen. Die Bats und MOrTls sind als Vergleichsgruppe zu den

TALEs und RipTALs relevant. Ausserdem könnten die Polymorphismen die einen leichten Einfluss auf DNS-Bindung als eine Ressource genutzt werden um TALE-Technologie zu verstärke.

Die Einleitung dieser Dissertation erklärt das Wesentliche über TALE Biologie mit Schwerpunkt auf DNA-Bindung und wie diese zur TALE-Technologie entwickelt wurde. Anschließend werden die RipTALs, Bats und MOrTLs der Reihe nach vorgestellt, und das Wissen über Herkunft und Sequenzmerkmale präsentiert. Die Ziele meiner Doktorarbeit werden dann ausgeführt. Das Hauptziel meiner Arbeit war ein Vergleich der molekularen Eigenschaften der einzelnen TALE-like Gruppen anzustellen. Davon erhofften wir neue Einblicke in die Interaktion zwischen TALE-likes und DNS, die Evolution der TALE-likes sowie eventuelle Biotechnologische Anwendungen zu gewinnen. Im Falle der RipTALs ist dieses Wissen direkt einsetzbar um die bakterielle Schleimkrankheit zu bekämpfen.

Artikel zur Charakterisierung von RipTALs, Bats und MOrTls sind Resultate meiner Doktorarbeit und kommen als Ergebnisteil der Dissertation vor. Zusammen mit Kollegen konnte ich zeigen dass die Repeats aus jeder TALE-like Gruppe Basen-spezifisch DNS-Binden. Der TALE Code, aus dem hervorgeht welche Aminosäurereste an Position 13 Bindung an welche Base vermitteln, ist in allen Gruppen konserviert.

Ich beschreibe auch wie ich die TALE-likes als faszinierender Fall von struktureller Konservation in einem vielfaltigem Sequenzraum verstehe. Überdies könnte es sein dass die TALE-likes, trotz dem akzeptierten Bild als eukaryotische Transkriptionsfaktoren, eine viel diversere Gruppe sind mit Exemplaren die bis jetzt unerklärte Rollen erfüllen.

1 Introduction

1.1 Molecular plant pathology in the context of food security for the 21st century.

In the age of cloud computing, 3D-printing and humanoid robotics it is easy to feel that inorganic technologies will be the major shaping force in humanities development for the coming century. Yet we are still biological beings. We are born, we grow old, we get sick and we die. Along the way we grow and we reproduce and to do so we require sustenance. This sustenance comes in the form of organic matter, and specifically it comes from plant tissue, either directly or to feed livestock. The advent of the lab grown burger may challenge this paradigm somewhat but no one has yet seriously challenged the notion that crop plants will continue to sustain human life on this planet for the next centuries, much as they have done for the previous millennia of human existence.

The major challenge will be the scale of crop production required: latest UN projections put the human population of the earth at 9.5 billion by 2050 (UN Department of Economic and Social Affairs: 2012 Revision of World Population Prospects). Moreover, this population needs to be fed without critically depleting the earth's natural resources or worsening our already catastrophic impact on climate and biodiversity.

There are three main ways to increase agricultural productivity: increased land utilization (plant more), yield maximization (grow more from what you plant), and reducing waste during storage and transportation (lose less of what you grow). The first of these, additional land use is undesirable if sustainability and a reduced environmental impact are to be achieved. Yield maximization, depending on the method used, offers, in contrast, the promise of more for less. That is more food, with fewer inputs and a smaller ecological imprint. This strategy should also be coupled to infrastructure optimization, to reduce losses after harvesting, since a third of food is currently wasted worldwide (1).

The modern population explosion, which began in the 18th century, has at each point been made possible through agricultural yield increases, with a few specific innovations standing out. Artificial Nitrogen-based fertilizer production, allowed huge yield boosts at the start of the twentieth century. In the latter half of the twentieth century, agronomy, the science of crop improvement, came into its own in the form of the green revolution lead by Norman Borlaug. Whilst humans had been domesticating and gradually selecting for improved crop varieties for millennia the rational exploration of crop traits allowed for the selection of high-yield, dwarf varieties of rice and wheat, which together with fertilizer application and use of improved agricultural techniques allowed yield increases of 208% and 109% for wheat and rice respectively in the developing world (2).

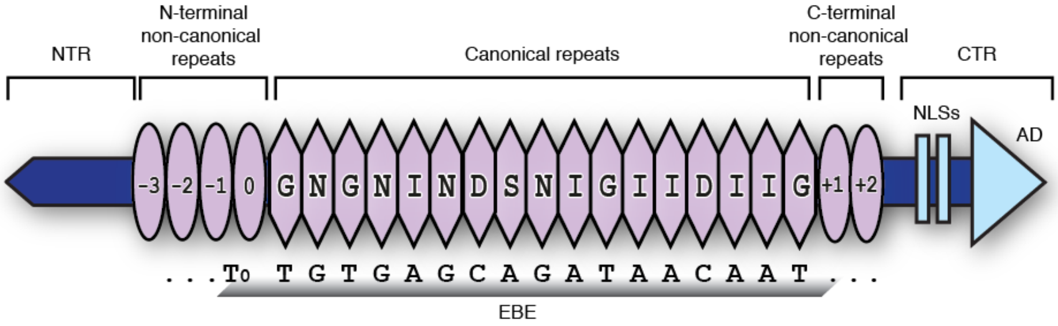
Among the traits that could benefit from additional research for crop improvement, pathogen and pest resistance are surely among the most vital. One study records that 5-25% of staple crop yield was lost to plant pathogens at the start of the millennium (3). Farmers can apply a range of anti-microbials to reduce the impact of plant pathogens, with an associated cost and environmental impact. Alternatively, resistant crop lines can be developed, either by breeding for resistance or direct genetic modification (GM) allowing the plant to produce anti-microbials or otherwise defend itself. The latter approach has already proved successful in the fight against insect pests. The insecticidal Cry proteins of *Bacillus thuringiensis* have been transformed into various crops (Bt-maize, Bt-cotton and most recently Bt-brinjal, among others) rendering them resistant to certain insect pests and reducing the need for insecticide applications (4). The most successful implemented GM approach for pathogen resistance project is the Rainbow papaya resistant to *Papaya Ringspot Virus* due to expression of the viral coat protein gene (5). Yet in all these cases one or a few genes have been inserted, to achieve simple dominant phenotypes. This trait limitation is due to a limitation in the tools available to create transgenics. All the GM crops currently grown were developed by the insertion of one or a few genes via *Agrobacterium*-mediated transformation or biolistic bombardment. In both methods genes are inserted at random producing an array of transgenic lines with different properties depending on the position of the transgenes. The lines must be subsequently screened to identify those transgenic insertion events resulting in suitable gene expression levels. In the last five years tools have become available to manipulate specific genomic loci allowing the insertion of transgenes into predetermined locations (6). Additionally, instead of being limited to the insertion of genes, molecular tools are now available to activate, suppress or modify endogenous loci (7). The nascent field of plant synthetic biology aims to harness these new tools to substantially redesign plant traits to human benefit (8). Thus tools have become available to produce GM crop lines faster than ever before and with qualitatively novel modifications. Biotechnology should be high on the agenda when devising strategies for global food security over the coming century. Of course GM approaches to tackle plant disease will rely on our understanding of plant diseases and immune responses. The research I have undertaken during my doctoral work combines these two important areas, bringing together plant disease research and biotechnology through the node of TALE biology.

1.2 Introduction to TALE biology

TALEs are secreted by *Xanthomonas* bacteria through the needle-like Type III secretion system (T3SS) during infection. Species of the genus *Xanthomonas* fill three of the top ten slots in a list of most important bacterial plant pathogens for economic and scientific impact (9). The TALEs are unusual among pathogen effectors in working as eukaryotic transcription factors, manipulating host gene expression to the benefit of the pathogen. Most effectors produced by plant pathogenic bacteria function by blocking elements of the immune response pathway (10). Considering that till now no TALEs have been found to suppress host immunity they stand out amongst characterized effectors. Another intriguing feature of TALEs is the diversity of TALE DNA binding domains, and thus of host target preferences, within and between strains (11). This allows a level of adaptability that makes them a fascinating subject of study for pathogen effector biology, and is also the key to their popularity as tools for biotechnology.

The key to the adaptability of the TALEs is in their domain structure. TALE-DNA binding is mediated by a tandem repeat array forming the largest and central region of a given TALE protein (Figure 1.1). The repeat array can be divided into the core group of near-identical canonical repeats, and then flanking them several sequence-degenerate non-canonical repeats. The N-terminal non-canonical-repeats are particularly important for DNA binding. The repeat domain is itself flanked N- and C-terminally by non-repetitive domains encoding other necessary functions. An N-terminal secretion signal allows transportation of TALE proteins from the bacterial to the host cell through the T3SS. C-terminal of the binding domain nuclear localization signals (NLSs) allow TALEs to penetrate the host nucleus. An activation domain (AD), rich in acidic residues and therefore alternatively referred to as the acidic activation domain (AAD) mediates induction of host promoters after binding (12). The domain structure of a TALE is illustrated in Figure 1.1.

Figure 1.1: An overview of TALE functional domains and TALE-DNA interaction


NTR
N-terminal non-canonical repeats
Canonical repeats
C-terminal non-canonical repeats
CTR
NLSs
AD
-3
-2
-1
0
G N G N I N D S N I G I I D I I G
+1
+2
. . . T0 T G T G A G C A G A T A A C A A T . . .
EBE

A) A schematic of a TALE and its functional domains: canonical DNA binding repeats are displayed as purple polygons, non-canonical repeats as purple ovals. NLSs as blue bars and the AS a blue triangle. B) Canonical repeats each bind a single base in the corresponding EBE. Base specificity is dertermined by position 13 of each repeat. This residue is referred to as the base specifying residue (BSR). BSRs of each repeat are given using the single letter amino acid code. The BSR composition illustrated here is arbitrary. Non-canonical repeats are numbered, and the thymine bound by repeat -1 is indicated (T_0).

1.3 The DNA binding properties of TALEs

Sequence specific DNA binding is mediated by the canonical repeats according to the TALE code

Examining the canonical repeats of any TALE reveals that sequence diversity is extremely limited (Figure 1.2). Position 12 and position 13, are by far the most variable. In fact, the variability of these two residues was the initial inspiration behind the hypothesis that TALE repeats each bind one base with a simple cipher. Residues 12 and 13 were termed the repeat variable diresidue (RVD) and were indeed found to co-vary with target bases (13, 14). Yet residue 12 in TALE repeats is almost always occupied by a His or Asn residue, and is therefore not as variable as position 13. Structural and functional studies later showed that only position 13 interacts with the DNA base and determines specificity (15, 16). For historical reasons the term RVD is, however, still widely used but the term base specifying residue (BSR) referring to position 13 is preferable.

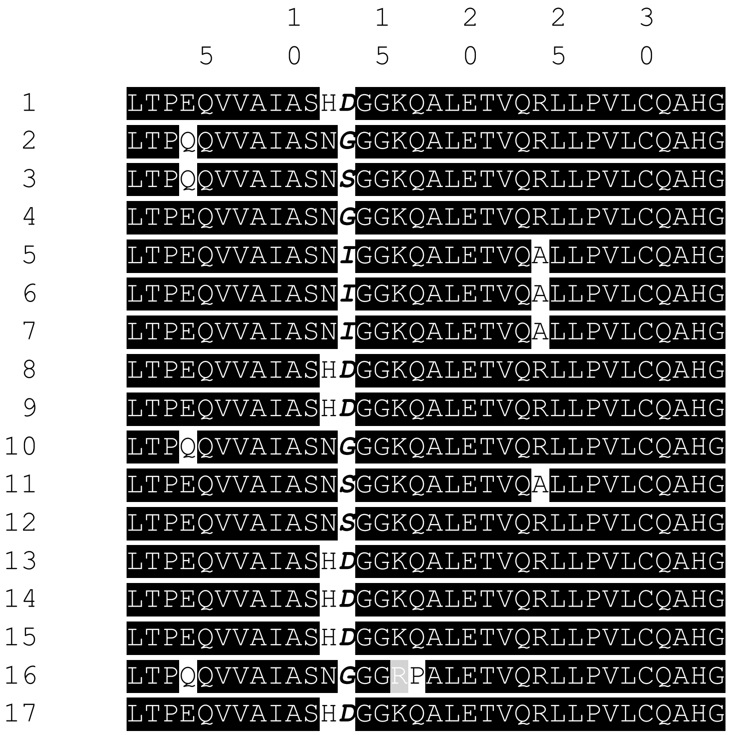
The canonical DNA binding repeats of TALEs are modular with each repeat binding a single nucleotide (Figure 1.1). In addition to non-specific interactions with the phosphate backbone, each repeat can make a single base specific interaction. This is mediated by the BSR. BSR-base interactions lend TALEs their sequence specificity. In natural TALEs only a few residues are found at the BSR (overwhelmingly Asp, Gly, Asn, Ile and Ser (17)), and each mediates a different base specificity. Some BSRs are highly selective for a single base (e.g. Asp-C or Ile-A), whilst others are more promiscuous (e.g. Asn – G/A, Ser – A/C/G/T). Since other positions within the repeat remain largely constant in a given array TALE repeats can be defined based on their BSRs. Thus one can talk of a Gly repeat or an Asp repeat, meaning a repeat with position 13 occupied by residues Gly or Asp respectively. The full complement of repeats with different BSRs and their orders within the array can be referred to as the BSR composition.

The reliable relationships between repeats with a certain BSR and their corresponding DNA base partners form what is known as the TALE code. This code allows researchers to search for potential host target genes bioinformatically. Several such search algorithms are available (18, 19) and when combined with transcriptomics allow for candidate target promoters to be rapidly identified. Additionally the TALE code lays the foundation for the use of TALEs in biotechnology, explored below.

N-terminal non-canonical repeats encode a fixed T_0 preference

Beyond the canonical repeats, the other key element of the sequence specificity of TALE DNA-binding is the preference for a thymine at the position immediately 5’ of the base bound by the first canonical repeat. This thymine is referred to as $T_{0.}$ This T_0 preference is so strong that till now only one case has been discovered of a natural TALE-promoter interaction where the T_0 is violated and a C found at this position in the target box instead (20). Studies examining the sequence specificity of TALEs using unbiased approaches have consistently returned T as the preferred base at the zero position (21, 22).

Figure 1.2: An alignment of the canonical repeats of TALE AvrBs3.



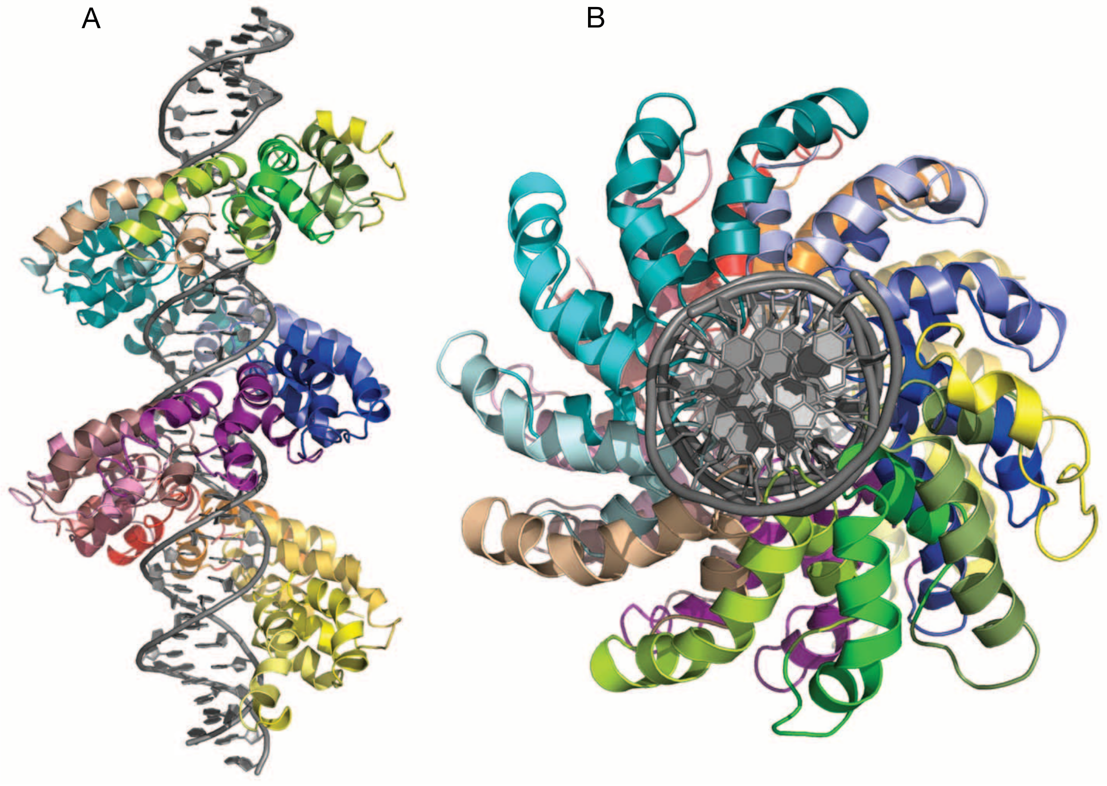
```
                 1    1    2    2    3
             5   0    5    0    5    0
 1      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
 2      LTPQQVVAIASNGGGKQALETVQRLLPVLCQAHG
 3      LTPQQVVAIASNSGGKQALETVQRLLPVLCQAHG
 4      LTPEQVVAIASNGGGKQALETVQRLLPVLCQAHG
 5      LTPEQVVAIASNIGGKQALETVQALLPVLCQAHG
 6      LTPEQVVAIASNIGGKQALETVQALLPVLCQAHG
 7      LTPEQVVAIASNIGGKQALETVQALLPVLCQAHG
 8      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
 9      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
10      LTPQQVVAIASNGGGKQALETVQRLLPVLCQAHG
11      LTPEQVVAIASNSGGKQALETVQALLPVLCQAHG
12      LTPEQVVAIASNSGGKQALETVQRLLPVLCQAHG
13      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
14      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
15      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
16      LTPQQVVAIASNGGGRPALETVQRLLPVLCQAHG
17      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
```

Amino acids that are identical between the repeat units are displayed as white letters on a black background. A white background indicates residues differing from the consensus at that position. Grey background indicates residues with similar chemical properties at the same position. Numbers above the sequence (5, 10, 15, 20, 25, 30) indicate the position of the given amino acid within the repeat.

An appreciation of the three-dimensional structure helps explain the DNA binding properties of TALEs

Structures have been solved for canonical-repeat regions of several natural and constructed TALEs alone and together with their cognate DNA targets. The first structures were published in tandem in the journal *Science* in 2012. These were of natural TALE PthXo1 (15), and dHax3, derived from TALE Hax3 (16). Following these were crystal structures of AvrBs3 (23) and various derivatives of dHax3 as well as a constructed TALE designated dTALE2 (24, 25). These studies all returned the same structure for the canonical repeat domain (Figure 1.3). The whole repeat array forms a right-handed super helix with 11 repeats per turn and a pitch of 60Å, contracting to 35Å in contact with the DNA. The repeats are arranged such that BSR-loops trace along the major groove. Each repeat is a pair of alpha-helices with the BSR-loops interpolating into the major groove allowing close proximity to target bases.

Figure 1.3: Crystal structure of TALE PthXo1 bound to DNA



The crystal structure of the DNA binding domain of TALE PthXo1 bound to DNA shown from longitudinal (A) and transverse (B) perspectives. This figure is taken with permission from Mak *et al*., Science 2012 (15). The repeat array forms a right-handed super helix wrapping around the DNA. Each individual repeat is constituted by a pair of alpha-helices, a BSR loop pointing into the major groove of the double-helix and a flexible inter-repeat loop. Repeats are individually and arbitrarily coloured, with green shades for the N-terminal and yellow shades for the C-terminal repeats.

As expected each canonical repeat forms an almost identical structure, a helix-turn-helix motif with a flexible linker region to join to the next repeat (Figure 1.4A). Position 1 of each canonical repeat, as traditionally defined, is in the middle of the flexible region. The first alpha-helix, termed the short-helix, then runs from positions 2 to 10. The BSR-loop at positions 11-15 is framed by helix-breaking serine and glycine residues. The long-helix is formed by residues 16-32, with a kink caused by Pro27.

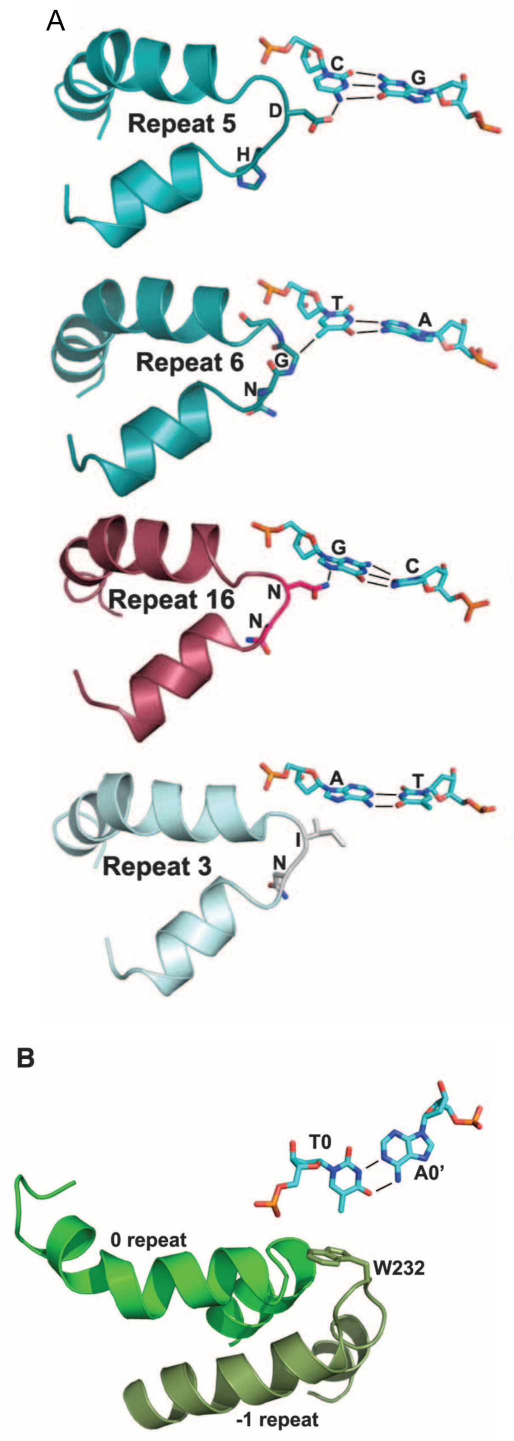
The crystal structures provided evidence that only residue 13 interacts with the specific base (Figure 1.4A), corroborating studies showing that only base 13 determines specificity (26). Another important finding was that different BSRs mediate qualitatively different interactions with their target bases. Asp and Asn are able to form strong hydrogen bonds, making sense of the higher affinity interactions found to be mediated by such repeats (22, 27). Gly is able to form van der Waals interactions with T, whilst the Ile-A interaction is promoted only by desolvation energy. These different interaction forms can be seen in Figure 1.4A. Further modeling, based on the available crystal structures, revealed that negative discrimination of sterically incompatible bases by the BSR is an important part of base specificity (28).

It should be noted that most of the binding energy for the DNA-canonical repeat interaction comes from non-base-specific interactions between the DNA phosphate backbone and positively charged residues 16-17, forming an electropositive stripe running alongside the DNA (16, 28).

Structures of the non-canonical-repeats 0 and -1 have been resolved for PthXo1 (15) and AvrBs3 (23). In addition structures are available for repeats -3 to 0 for dTALE2 bound to DNA (25). It seems that the same overall structure unites the non-canonical with the canonical repeats, but that they differ as to the degree of interpolation into the major groove, with an impact on DNA binding properties. What sets apart canonical repeats is not sequence similarity alone but the presence of a loop region from positions 11-15, at the apex of which lies the BSR, penetrating close to the DNA base at and forming base specific interactions. It is thus a functional distinction. Of the N-terminal non-canonical repeats of TALEs only repeat -1 displays a BSR loop, and Trp232 in this loop region mediates the T_0 interaction (Figure 1.4B). However, canonical TALE repeats bearing Trp as the BSR are non-functional (29, 30) likely reflecting differences in the angle and distance of approach for repeat -1 and the T_0 compared to canonical BSR-base interactions (Figure 1.4) (15). It is thus justifiable to assert that the -1 to T_0 interaction is non-canonical. Positively charged residues, found in abundance, across all the N-terminal non-canonical-repeats mediate a strong non-base-specific DNA affinity, crucial for successful TALE-DNA interactions (25).

In addition to these static views there is some information available about conformational changes taking place during binding. Most notably there is a contraction along the longitudinal axis of the super helix during DNA binding. This insight was first made by comparing crystal structures of bound and unbound TALEs (15). However, a recent study on the movement of TALEs along DNA to identify target boxes has revealed further insights (31). It seems that during one-dimensional scanning along the DNA target the TALE is in its extended confirmation and only the N-terminal non-canonical repeats contact the DNA. Once a potential target sequence is encountered contraction occurs leading to a tight and highly stable (32) interaction.

Figure 1.4: Structures of individual repeats from the crystal structure of TALE PthXo1



Canonical repeats, indicating the common BSR-base interactions (A) and the special case of repeat 0 and -1 interacting with the T_0 base (B). This figure is taken with permission from Mak *et al*., Science 2012 (15). Repeats are shown as arbitrarily coloured ribbon-spirals (alpha-helices) and loop regions, with the side-chain structures of residues 12 and 13 shown. Numbers indicate the position of the repeat within the PthXo1 repeat array. The single letter code is used for amino acids. Nucleotides and base-base hydrogen bonds are also shown in each case.

The Asp-C and the Asn-N interactions are each mediated by a hydrogen bond (dashed line). Gly-T is mediated by hydrophobic van der Waals interactions, while Ile-A forms instead weak interface promoted by the displacement of solvent molecules.

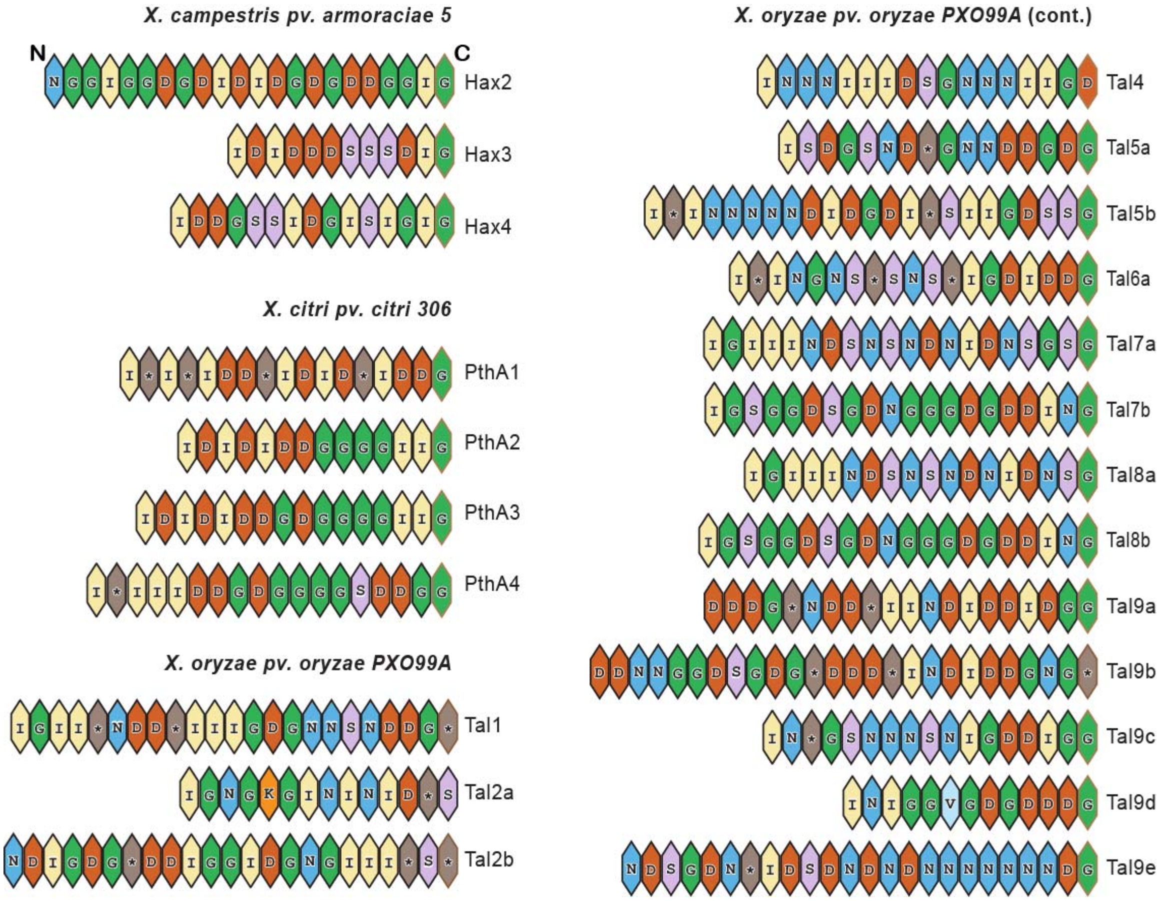
In the case of repeats 0 and -1 (B), Trp232 of repeat -1 forms non-polar van der Waals interactions with T_0.

1.4 TALEs in the context of plant disease

Although the sequence identity of individual TALE repeats and non-repeat regions is high, the BSR compositions of TALEs are diverse both within and between strains (Figure 1.5). Whatever the mechanism that controls this, it seems that repeat number and BSR polymorphisms arise frequently in *Xanthomonas* strains providing raw material for TALE evolution. This raw material has been guided by natural selection to create TALEs with DNA binding domains targeting diverse host genes. Amongst the *S*-genes so far identified are sugar transporters, transcription factors, and RNA processing factors (33). The genes targeted by TALEs provide an insight into the lifestyle of the pathogen. For example, two sucrose exporters of the rice SWEET family (*OsSWEET11* and *OsSWEET14*) have been convergently targeted by at least six *X. oryzae* pv *oryzae* TALEs (33). TALEs of closely related *X. oryzae* pv *oryzicola* do not target *SWEET* genes, yet $\mathrm{TAL20}_{\mathrm{Xam668}}$ of more distantly related *X. axonopodis* pv *manihotis* was found to target *MeSWEET10a*, an ortholog of *OsSWEET11* and *OsSWEET14* in cassava. Since *X. oryzae pvoryzae* and *X. axonopodis* pv *manihotis* both proliferate in the xylem, unlike *X. oryzae* pv *oryzicola* this may reflect a requirement of an apoplastic sugar source for xylem-dwelling xanthomonads. Understanding how the pathogen operates within the host can also inform resistance breeding or engineering approaches. For instance, it has been shown that rice *SWEET* gene knockouts are resistant to bacterial streak (34).

TALEs not only activate S-genes, but also in some cases trigger the expression of resistance genes (R-genes), which trigger a hypersensitive response preventing pathogen growth. Several such R-genes have been discovered in cultivars amongst otherwise susceptible hosts (Bs3 and Bs4C in pepper (35, 36), Xa10 and Xa23 in rice (37, 38)) indicating that R-genes have evolved as a response to pathogen pressure. Evidence is already available to show that TALE repeat domain evolution is a mechanism to avoid host recognition (39). The impressive diversity of TALE BSR compositions (Figure 1.5) thus needs to be considered in the context of competing pressures to induce S-genes and avoid R-genes. This evolutionary dynamism makes TALEs a fascinating case of pathogen effectors at the frontline of the arms race between host and pathogen.

Figure 1.5: TALE complements of three *Xanthomonas* strains


X. campestris pv. armoraciae 5
N
C
N G G I G G D G D I D I D G D G D D G G I G
Hax2
I D I D D D S S S D I G
Hax3
I D D G S S I D G I S I G I G
Hax4
X. citri pv. citri 306
I * I * I D D * I D I D * I D D G
PthA1
I D I D I D D G G G G I I G
PthA2
I D I D I D D G D G G G G I I G
PthA3
I * I I I D D G D G G G G S D D G G
PthA4
X. oryzae pv. oryzae PXO99A
I G I I * N D D * I I I G D G N N S N D D G *
Tal1
I G N G K G I N I N I D * S
Tal2a
N D I G D G * D D I G G I D G N G I I I * S *
Tal2b
X. oryzae pv. oryzae PXO99A (cont.)
I N N N I I I D S G N N N I I G D
Tal4
I S D G S N D * G N N D D G D G
Tal5a
I * I N N N N N D I D G D I * S I I G D S S G
Tal5b
I * I N G N S * S N S * I G D I D D G
Tal6a
I G I I I N D S N S N D N I D N S G S G
Tal7a
I G S G G D S G D N G G G D G D D I N G
Tal7b
I G I I I N D S N S N D N I D N S G
Tal8a
I G S G G D S G D N G G G D G D D I N G
Tal8b
D D D G * N D D * I I N D I D D I D G G
Tal9a
D D N N G G D S G D G * D D D * I N D I D D G N G *
Tal9b
I N * G S N N N S N I G D D I G G
Tal9c
I N I G G V G D G D D D G
Tal9d
N D S G D N * I D S D N D N D N N N N N N N D G
Tal9e

Canonical repeat arrays and repeat +1 are shown for complete TALE complements of three *Xanthomonas* strains, as listed in (17). Repeats are shown as polygons coloured by BSR, which is given in single letter amino-acid code inside each repeat. An asterisk indicates a shorter repeat with a single amino acid missing at position 13 when compared to other TALE repeats (termed G_{SL} repeats; (7)). A brown edge indicates repeat +1, always the C-terminal most repeat of each array. Repeats are shown N to C terminal as indicated above the repeats of Hax2.

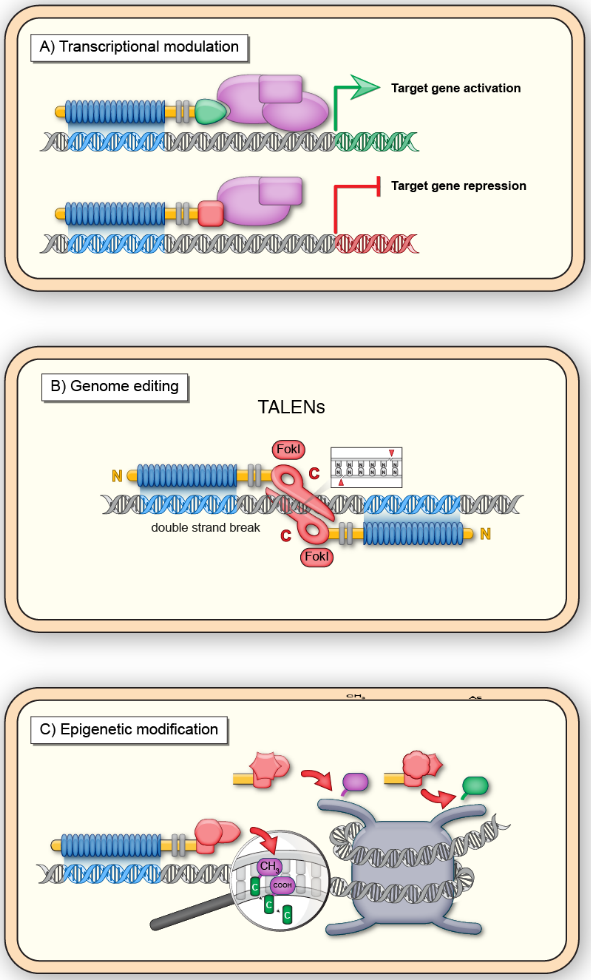
1.5 Introduction to TALE technology

TALE technology was born with the publication of the TALE code. One of the two publications describing the TALE code made use of synthesized TALEs with novel BSR compositions to test the code (13). This laid the foundation for the creation of what is generally referred to as a designer TALE (dTALE). The key trait of a dTALE is having a predetermined BSR composition and therefore a predetermined sequence preference. Tests on dTALE specificity have shown that their activity is limited to target sequences only, even in complex genomes (40). This makes dTALEs a straightforward and reliable molecular tool to target specific DNA sequences.

The tandem repeat composition, which makes dTALE design so straightforward also makes assembly complicated. PCR based methods and classical restriction enzyme based methods are unsuitable for assembling long arrays of near-identical sequences. One of the solutions arrived at independently by several research groups is to create a library of individual segments encoding TALE repeats with different RVDs and to use the golden gate cloning method to assemble these segments in a desired order in only a few steps (18, 41). Golden gate cloning uses type IIS restriction enzymes, which cleave outside of their binding sites and generate four base-pair sticky ends, to allow reliable user-defined multi-fragment assembly.

Whilst user-defined plant transcription factors are a desirable tool for many researchers, the success of TALE technology rides on the compatibility of the DNA binding domain with other functional domains. The AD and much of the rest of the NTR and CTR can be truncated without impairing DNA binding, and other functional domains can then be fused in their place. This has resulted in a panoply of TALE-based tools (Figure 1.6). The C-terminal fusion of the FokI nuclease domain to create a TALE nuclease (TALEN; Figure 1.6B) has been the most studied and most used. The generation of knock-out lines is a key step in most reverse genetics approaches. In the past such knock-outs had to be first randomly generated and then screened and characterized, TALENs allow researchers to create specific knock-outs. This was already possible to some extent thanks to Zinc-Finger Nucleases (ZFN), which can reasonably be called the predecessors to TALENs, since this technology is now little used but the basic design is analogous and was directly copied to create TALENs. Functional TALENs and ZFNs are both paired-sets of DNA binding domains with C-terminal FokI nuclease domain fusions. Whilst ZFNs are modular DNA binding proteins there is no particular code that can be used to predict DNA binding specificity from primary protein structure. For these reasons the popularity of ZFNs as a tool declined with the advent of TALENs. Today the popularity of TALENs is challenged by an RNA-guided programmable nuclease system known as CRISPR-Cas9, explored in more detail in 1.2.7

Figure 1.6: Overview of TALE fusion protein designs



TALE technology for manipulating transcriptome (A), genome (B) and epigenome (C). Source: de Lange *et al*., *Plant J*, 2014. TALE DNA binding domains are illustrated with blue ovals, binding their DNA target sites, shown as blue regions on otherwise grey DNA. Yellow bars indicate NTR and CTR regions. C-terminally fused functional domains are shown as a green triangle (AD), red box (repressor domain), red scissors (nuclease activity), red hooks or rings (recombination) or red polygons (epigenetic modifications).

All of these construct designs have been created and confirmed for the desired function. In particular TALE activators and repressors (a) have been employed in the creation of synthetic genetic circuitry and as tools for fundamental biology. The paired C-terminal FokI fusion protein design is what is commonly referred to as TALENs and has been employed extensively for the creation of gene-knockouts and gene-targeting. Numerous epigenetic modifier domains have all been demonstrated in a proof of concept study (42) but have yet to be employed beyond that context.

1.6 TALE-like proteins have so far been identified in three sources outside of *Xanthomonas* bacteria

My doctoral work has been focused on characterizing TALE-like proteins from non-*Xanthomonas* bacteria. In studying TALE-like proteins I hoped to reveal unexpected insights about the TALEs themselves, which might prove useful in biotechnology and tackling plant disease. In addition, one of the TALE-like producing bacteria, *R. solanacearum*, is itself a devastating plant pathogen causing bacterial wilt disease. I hope that the molecular characterization of *R. solanacearum* TALE-like proteins described in this thesis may prove useful in the fight against bacterial wilt.

What follows is first an introduction to the TALE-like proteins so far identified and their host organisms. After that I will introduce and expound on each of the central questions of this thesis. Do non-*Xanthomonas* TALE-likes adhere to the TALE paradigm in terms of functional domain structure and DNA-binding properties? What are the unifying features of the TALE-likes? Can they reveal anything about TALE-evolution? Finally, could TALE-likes be useful additions to TALE technology?

Three groups of TALE-like proteins of non-*Xanthomonas* origins have been identified to date. They are the RipTALs of *R. solanacearum*, The Bats of *Burkholderia rhizoxinica* strain HKI 454, and the MOrTLs of uncertain marine bacterial origin. The latter group actually contains two TALE-like repeat-proteins each as different from one another at the sequence level as they are from any other TALE-like group, and therefore considered two groups: MOrTL1 and MOrTL2. All of these different TALE-likes were first recognized in the results of genome sequencing projects. In the case of the RipTALs some functional information was available prior to the beginning of this work, reviewed below. In addition, molecular characterizations of the RipTALs and Bat1 have been carried out by other research groups and were published shortly before those presented in this thesis. Information from those publications is integrated into the discussion. The next paragraphs provide information on the providence of each group and the existing literature prior to the initiation of this work.

1.6.1 Prior knowledge about *Ralstonia* TALE-likes

Discovery and early work: R. solanacearum *TALE-likes are T3Es found across much of the species complex*

The first report of a TALE-like protein comes from the early years of TALE research. In 2002 the genome of *R. solanacearum* strain GMI1000 was sequenced (43). Amongst the newly identified putative effector genes was one, Rsc1815, likely homologous to TALEs. This assessment was based on sequence similarity: 40% identical to AvrBs3 at the protein level; and on domain structure: 16 tandem repeats of 35 amino acids (Figure 1.7) flanked by more sequence degenerate repeats and then NTR and CTR domains. The term RipTAL was coined for these TALE-likes in 2013 as part of an integrated approach to *R. solanacearum* T3E nomenclature (44). Rip stands for *Ralstonia* injected protein, and is common to all, whilst the TAL alludes to their TALE-like nature.

A screen for genes regulated by the T3SS master regulator HrpB pulled out Rsc1815, providing it the designation *brg11* (HrpB-regulated 11; (45)). Around the same time the homolog from closely related strain RS100 was pulled out from a screen for T3E's based on tagging genes with adenylate cyclase and observing spikes in cAMP after infection of plant leaves (46). The homolog in RS100 was named Hpx17 (hrpB dependent expression 17). Unlike *brg11* this gene lacks the bulk of the repeat region, encoding a protein with only a single canonical repeat. Together, these studies confirmed that the RipTALs are indeed type-III-secreted.

A descriptive study has been carried out on RipTALs from a set of, mostly Asian, *R. solanacearum* strains. The strains were tested by PCR for the presence of a Brg11 or Hpx17 homolog. *ripTALs* were detected in 285 out of 319 strains. Repeat regions were amplified and characterized based on length and restriction digest pattern (47). This revealed that the vast majority of *ripTALs* were the same length *as brg11*. The other major isoform was the hpx17-likes. A few *ripTALs* were found with a length matching neither brg11 nor hpx17, but these mostly differed in size by just one or two repeats.

In 2010 a report into virulence contributions of GMI1000 T3Es found Brg11 to be amongst the most influential (48). The study compared the growth of wild type and effector knock-out strains infiltrated together into aubergine leaf (*Solanum melongena* cv. Zebrina). The *brg11* knock-out strain grew at 20% of the rate of the wild-type indicating compromised *in planta* growth properties, such that Brg11 must provide some virulence contribution in wild-type strains. Several *R. solanacearum* genomes had already been sequenced in addition to GMI1000 prior to our work on the molecular characterization of Brg11 and other phylotype I RipTALs. Specifically strain Molk2 from Phylotype II, Psi07 of Phylotype IV and closely related Banana Blood Disease Bacterium (BDB) strain R229 had all been sequenced and found to encode RipTALs. However, *ripTALs* were not found in other sequenced strains such as K60, CFB2957 and many of the strains studied by Heuer *et al*..

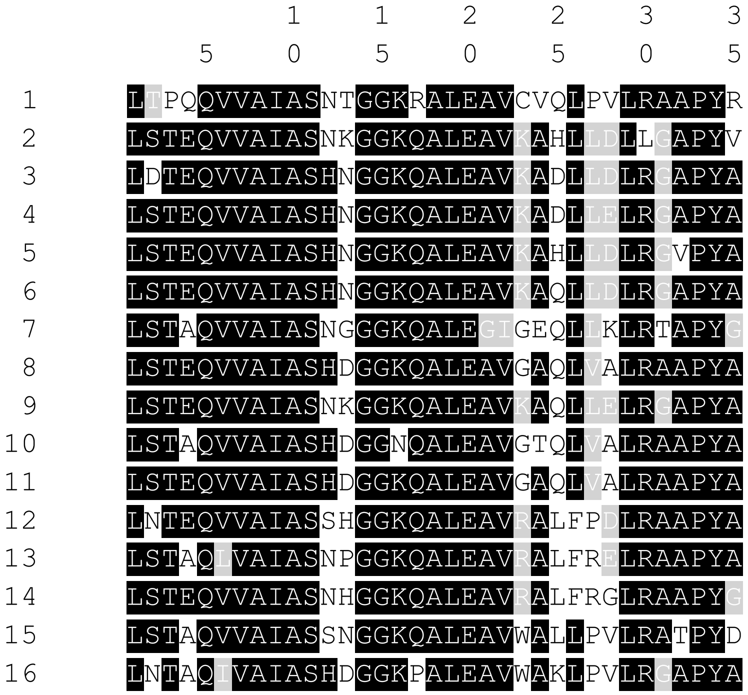
Thus the state of knowledge as to the RipTALs prior to the work presented in this thesis was that many, but not all, *R. solanacearum* strains possess RipTALs; sequence similarity to TALEs is high enough to be able to make a confident assignment of homology; at least some individual strains were shown to regulate/deliver their RipTALs with the type III secretion system; and Brg11 had been found to contribute to *in planta* growth in at least one natural host species. It was therefore already assumed within the field that RipTALs, like TALEs, are T3Es and mediate some virulence contribution by acting as eukaryotic transcription factors and inducing target host genes. Yet besides translocation, none of these molecular details had been studied.

RipTAL domain structure is similar to that of TALEs but the repeats display a level of inter-repeat polymorphism far exceeding that of TALEs

The domain composition of Brg11 seems at first glance rather like that of TALEs (Figure 1.8A). A tandem array of 16 repeats of high sequence similarity (average of

75%; Figure 1.7) are flanked by more sequence degenerate repeats and the whole repeat region is framed by large sections of NTR and CTR. The repeats of RipTALs and TALEs are similar enough to suggest a conserved molecular function, yet the most striking observation from a comparison of TALE and RipTAL repeat arrays is the number of inter-repeat polymorphisms. RipTAL repeats vary from one another not only at positions 12 and 13, but also across a section of the C-terminal end of each repeat, as well as a few other positions (Figure 1.7).

Figure 1.7: Brg11 canonical repeat alignment



```
                 1    1    2    2    3    3
             5   0    5    0    5    0    5
 1      LTPQQVVAIASNTGGKRALEAVCVQLPVLRAAPYR
 2      LSTEQVVAIASNKGGKQALEAVKAHLLDLLGAPYV
 3      LDTEQVVAIASHNGGKQALEAVKADLLDLRGAPYA
 4      LSTEQVVAIASHNGGKQALEAVKADLLELRGAPYA
 5      LSTEQVVAIASHNGGKQALEAVKAHLLDLRGVPYA
 6      LSTEQVVAIASHNGGKQALEAVKAQLLDLRGAPYA
 7      LSTAQVVAIASNGGGKQALEGIGEQLLKLRTAPYG
 8      LSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYA
 9      LSTEQVVAIASNKGGKQALEAVKAQLLELRGAPYA
10      LSTAQVVAIASHDGGNQALEAVGTQLVALRAAPYA
11      LSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYA
12      LNTEQVVAIASSHGGKQALEAVRALFPDLRAAPYA
13      LSTAQLVAIASNPGGKQALEAVRALFRELRAAPYA
14      LSTEQVVAIASNHGGKQALEAVRALFRGLRAAPYG
15      LSTAQVVAIASSNGGKQALEAVWALLPVLRATPYD
16      LNTAQIVAIASHDGGKPALEAVWAKLPVLRGAPYA
```

Amino acids that are identical between the repeat units are displayed as white letters on black background. A white background indicates residues differing from the consensus at that position. Grey background indicates residues with similar chemical properties at the same position. Numbers above the sequence (5, 10, 15, 20, 25, 30, 35) indicate the position of the given amino acid within the repeat.

1.6.2 Prior knowledge about *Burkholderia* TALE-likes

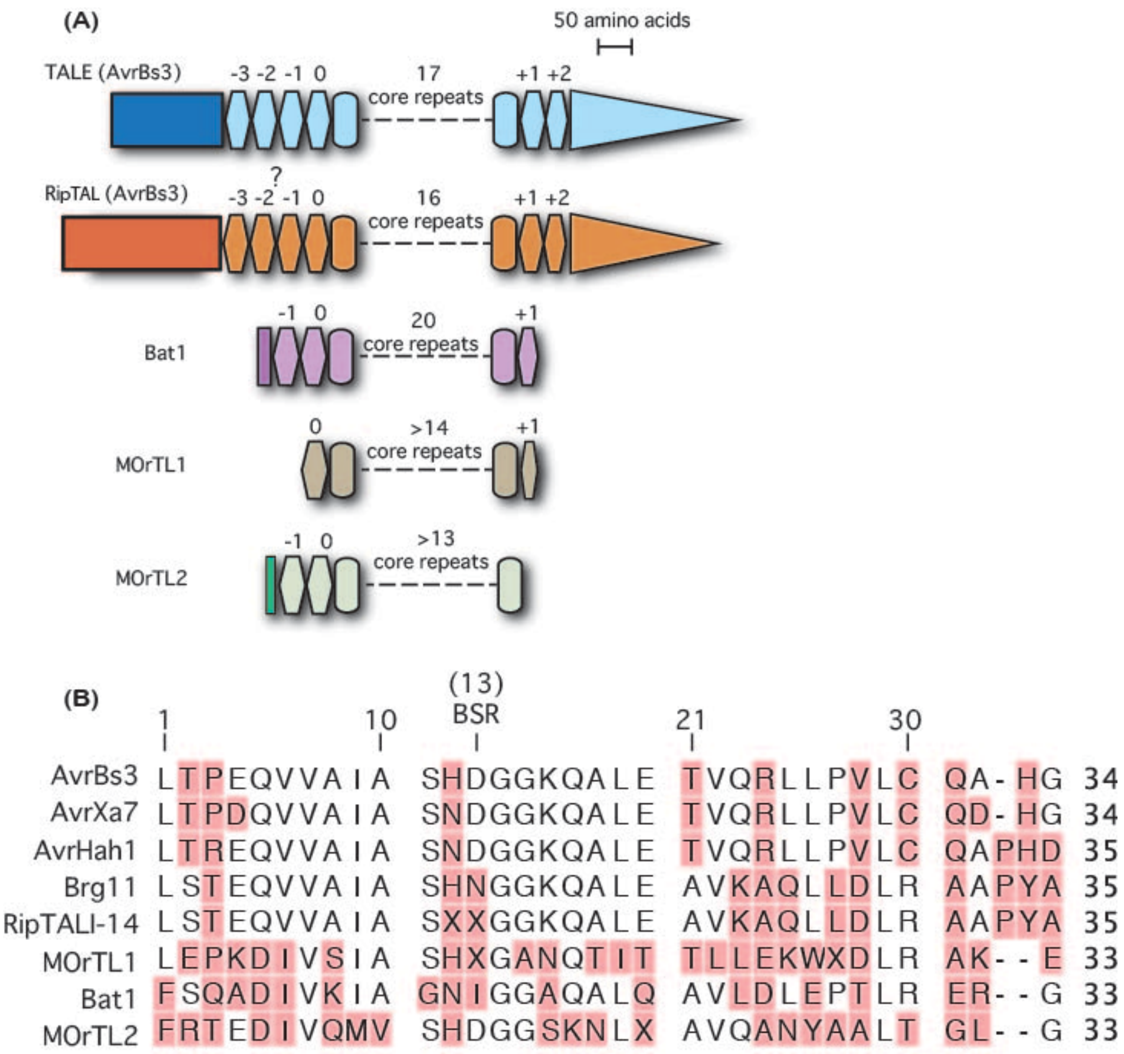
The genome of the rather esoteric fungal endobacterium *Burkholderia rhizoxinica* (strain HKI 454) was sequenced in 2011 (49). Amongst the many novel predicted proteins in the genome were a set of three genes similar enough to *TALEs* to be picked up by BLAST or HMM based searches using any TALE as a template (many thanks to Diana Horvath for initially alerting us to this fact). No information was or is available as to the expression of these predicted proteins within the bacterium, and the interesting lifestyle of this bacterium adds to the mystery shrouding the biological function of these proteins. *B. rhizoxinica* is naturally an obligate endosymbiont of zygomycete fungus *Rhizopus microsporus*. This fungus is a plant pathogen causing seedling blight (50), and in rare cases it has been found to cause nosocomial infections in humans with compromised immune systems. It should be noted though, that whilst plant pathogenesis of *R. microsporus* requires the bacterial endosymbiont, human pathogenesis proceeds in the absence of the endosymbiont (51). *R. microsporus* and *B. rhizoxinica* are of greater interest for the toxin produced during pathogenesis, rhizoxin, which has been shown to have anti-tumorigenic properties (52). This toxin is actually produced by *B. rhizoxinica* and delivered to the fungus (53). It is also possible, though far from facile, to culture *B. rhizoxinica* independently of its fungal host, raising the possibility of a fermentation system for rhizoxin production (53). There is thus cause to research the biology of this bacterium though it cannot be said to pose a great threat to agriculture or human health.

The *Burkholderia* TALE-likes (Bats) consist only of tandem repeat arrays divisible into canonical repeats and a few flanking sequence-degenerate repeats and a short (17 AA in the case of Bat1) N-terminal non-repetitive region (Figure 1.8A). Moreover, the repeats are less than 40% identical to consensus repeats of AvrBs3 or Brg11 (Figure 1.8B). Like RipTAL repeats, Bat repeats also display a lot of inter-repeat polymorphism. Yet despite this similar or identical residues are found in the region around position 13 in repeats of Bats, RipTALs and TALEs (Figure 1.8B) providing encouragement to test the hypothesis that the repeat arrays of the Bat proteins mediate sequence specific DNA binding.

1.6.3 Prior knowledge about Marine Organism TALE-likes

The TALEs, RipTALs and Bats all hail from terrestrial plant-pathogenic or plant-pathogen-associated bacteria. The final two TALE-likes addressed in this thesis are from as yet unspecified marine bacteria from the waters of the Gulf of Mexico. They were picked up as part of the J. Craig Venter Institute Global Oceanic Survey (54, 55). The Marine Organism TALE-likes (MOrTLs) come from orphan contigs from this dataset but based on size filtering of environmental samples prior to genomic library construction they are almost certainly bacterial. Similar to the Bats they seem to be formed only of tandem repeats with some flanking sequence degenerate repeats. However, the available sequences likely do not fully cover the complete coding sequences and it is not possible to make definitive statements about gene structure. The two genes *mortl1* and *mortl2* encode repeat arrays with lower inter-repeat variability than found among Bats. However, whilst repeats of MOrTL1 are all similar to one another and the same for MOrTL2, the repeats differ greatly between the two proteins. In fact they differ as much from another as they do from repeats of the TALEs, Bats and RipTALs (consensus repeat sequences in figure 1.8B). However, repeat length (33 amino acids) and certain motifs seen in other TALE-like repeats was enough to mark these repeats out as similar to TALE-like repeats and worthy of further investigation (Figure 1.8B). The MOrTLs were first suggested to be TALE-likes in a publication from another group in 2013 (56) based on sequence similarities but they were not further investigated.

Figure 1.8: Domain composition and repeat sequence differences between representative TALE-likes


(A)
50 amino acids
TALE (AvrBs3)
-3 -2 -1 0
17
core repeats
+1 +2
RipTAL (AvrBs3)
?
-3 -2 -1 0
16
core repeats
+1 +2
Bat1
-1 0
20
core repeats
+1
MOrTL1
0
>14
core repeats
+1
MOrTL2
-1 0
>13
core repeats
(B)
1
10
(13)
BSR
21
30
AvrBs3 LTPEQVVAIA SHDGGKQALE TVQRLLPVLC QA-HG 34
AvrXa7 LTPDQVVAIA SNDGGKQALE TVQRLLPVLC QD-HG 34
AvrHah1 LTREQVVAIA SNDGGKQALE TVQRLLPVLC QAPHD 35
Brg11 LSTEQVVAIA SHNGGKQALE AVKAQLLDLR AAPYA 35
RipTALI-14 LSTEQVVAIA SXXGGKQALE AVKAQLLDLR AAPYA 35
MOrTL1 LEPKDIVSIA SHXGANQTIT TLLEKWXDLR AK--E 33
Bat1 FSQADIVKIA GNIGGAQALQ AVLDLEPTLR ER--G 33
MOrTL2 FRTEDIVQMV SHDGGSKNLX AVQANYAALT GL--G 33

Domain composition (A) and consensus repeat sequences (B) of representatives of each TALE-like group addressed in this work. Different TALE-likes are given arbitrary colour groups. Ovals represent canonical repeats, and polygons non-canonical. Rectangles are used for N-terminal non-repetitive regions and triangles for C-terminal non-repetitive regions. All domain graphics are sized within the same scale and the scale bar allows graphic lengths to be related to amino acid sequence length of each functional domain. Below an alignment is shown with representative consensus repeats made form canonical repeat alignments for different TALE-likes, with the number taken from each group reflecting to some extent the number of known representatives of that group: TALEs (AvrBs3, AvrXa7, AvrHah1), RipTALs (Brg11, RipTALI-14), Bats (Bat1), MOrTL1 and MOrTL2. Pink background is used to indicate polymorphisms. Repeat position is indicated above and repeat length to the side. Alignments and consensus repeat calculations were carried out in CLC Main Workbench 7.

2 Aims of this work

My goal was to characterize the functional domains of all identified TALE-like groups, with a particular focus on DNA-binding properties. Four specific questions guided this analysis:

1. Do other TALE-likes conform to the TALE paradigm as pertains to (a) functional domains and (b) DNA binding properties?
2. What are the unifying features of TALE-likes and which of these could be said to be defining?
3. Can we gain insights into TALE-like evolution from the comparison of TALE-like protein sequences?
4. Could TALE-likes be used to make novel contributions to TALE technology?

These are introduced in turn and placed in their relevant context within the existing body of TALE research.

2.1 Conformity to the TALE paradigm?

Question one: Do TALE-like proteins conform to the TALE paradigm as pertains to functional domains and DNA binding properties?

(a) Functional domains and biological role

Whilst TALEs are highly diverse in the BSR compositions of their repeat arrays they are uniform in their domain composition. The particular domain structure of TALEs can be understood in terms of their biological function as pathogen effectors working as eukaryotic transcription factors. TALEs can be divided into three regions the N-terminal non-Repeat region (NTR), with a length of around 150 amino acids (157; AvrBs3), the repeat region, and the C-terminal non-repeat region (CTR), around 250 amino acids (229; AvrBs3). The repeat region can be further subdivided into canonical and non-canonical repeats, of which there are four N-terminal and two C-terminal of the canonical repeats. In addition the NTR and CTR contain smaller functional domains: the T3SS in the NTR and the NLSs and AAD in the CTR.

Some conclusions can already be made on the relative functional domain structures of the non-*Xanthomonas* TALE-likes. As shown in Figure 1.8A the RipTALs share a similar domain structure to that of the TALEs and although sequence identity is highest across the canonical repeats there are patches of clearly homologous sequences across the NTR and CTR. The functional domains of the RipTALs may well fulfill a similar biological function to those of TALEs. The same cannot be said for the Bats and MOrTLs. In the case of the MOrTLs it is not clear whether all the domains of the full length coding sequences are represented in the available DNA sequences. The Bats, however, are predicted from a fully assembled genome and more confident statements can be made as to their domain structure. In every case the Bats are formed of a tandem canonical repeat array flanked by two degenerate N-terminal repeats and one degenerate C-terminal repeat. In addition there is a 17 or

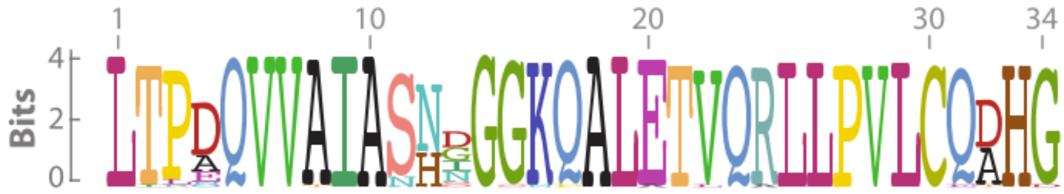
18 amino acid Non repetitive region forming the complete NTR. There are thus no long stretches of NTR and CTR in the Bats to mediate the translocation, and transcriptional activation functions of TALEs. The same seems to be true for the MOrTLs, though again caution must be taken here. These contrasts suggest differences in functional domain composition and biological function between TALE-like groups.

(b) DNA binding properties

Is there conservation of the TALE code in the face of non-BSR polymorphisms?

As described above, the TALE code is a reliable means to predict TALE target sequences or design dTALEs for a target of choice. Knowing the BSR composition of a particular TALE is typically enough to know all there is to know about its base preference. Yet could the simplicity of the TALE code simply reflect the lack of sequence diversity at non-BSR positions? Boch & Bonas created a sequence logo by pooling repeats from 113 TALEs from across the xanthomonads (Figure 2.1) and found that only positions 4, 12, 13 and 32 show any considerable polymorphism.

Figure 2.1: A sequence logo of TALE repeats based on 113 TALEs



Canonical repeat sequences of 113 TALEs were aligned by Boch & Bonas to create this sequence logo (Annual Reviews, 2010). Repeat positions are individually coloured. Alternative residues are stacked and the total height for each stack is correlated to the percentage conservation at that position across the aligned repeats. The BSR is the most variable position. Postions 4, 12 and 32 are variable to the extent that two alternative residues commonly occupy those positions.

It has been demonstrated that non-BSR residues can influence the activity and to a lesser extent the specificity of TALE repeats. As mentioned already position 12 is a naturally variable residue in TALE repeats and is generally referred to as co-determinant of base preference despite clear evidence to the contrary. Studies analyzing all possible combinations of residues 12 and 13 found that residue 12 seems to regulate activity to a much greater extent than specificity (29, 30). Few natural polymorphisms occur beyond this and studies are therefore lacking.

As mentioned, TALE and RipTAL repeats differ at a number of positions. Additionally RipTAL repeat arrays show far greater diversity in terms of inter-repeat non-BSR polymorphisms (Figures 1.3 and 1.4). The consensus repeats of TALEs and RipTALs are still conserved at just over half of all repeat positions (Figure 1.8B). Yet when including the Bat and MOrTL consensus repeats the polymorphism outweighs the conservation (Figure 1.8B). It is thus a major aim of my work to understand the impact this has on DNA binding properties. Is the TALE code conserved among TALE-likes and if so do non-BSR polymorphisms exert any influence?

How will differences in the non-canonical repeats affect DNA-binding?

In addition to the canonical repeats there are reasons to suspect differences in DNA binding properties arising from differences in the non-canonical repeats. Most relevant are the N-terminal non-canonical repeats, which, in TALEs, work as the nucleation point for DNA target site recognition (31) contributing decisively to DNA binding (25) and also mediating the fixed T_0 preference (57). N-terminal non-canonical repeats with similar sequences to those of TALEs are discernible in the RipTALs. The same cannot be said of the Bats, though non-canonical repeats N and C-terminal of the canonical repeats can be identified. In the MOrTLs it is not clear if the full gene sequences are available but based on the information at hand non-canonical repeats are presented but do not resemble those of TALEs. The impact of these repeats, particular the N-terminal non-canonical repeats is addressed for each of these groups in the articles in section 2.

2.2 Unifying and defining features of the TALE-likes

Question two: What are the unifying features of TALE-likes and which of these could be said to be defining?

The term TALE-like has been used only rarely till now and when used it is used inconsistently. The RipTALs have been referred to as TALE-likes or simply as the TALEs of *R. solanacearum* (58, 59). The Bat proteins have received little attention in the literature till now. Thus the term TALE-like in the current literature is a vaguely defined measure of sequence similarity not a precisely defined protein family. Prior to any functional characterizations or structural predictions, the term just means sequence-similar to TALEs.

2.3 What is known about TALE evolution?

Question three: Can we gain insights into TALE-like evolution from the comparison of TALE-like protein sequences?

Although my work has focused on a few representatives of each group, at the protein level, and is therefore ill designed for evolutionary enquiry, some interesting observations were made in the course of study. Therefore I consider it relevant to review the most germane research regarding TALE evolution to date, particularly to provide some context of the current assumptions within the field as to the evolutionary processes shaping the BSR composition diversity of TALEs.

Gene copy number differs between strains and Tn3 transposons may be involved

By comparing the TALE phylogeny of Ferreira *et al*. to a *Xanthomonas* phylogeny based on whole genomes (60), it appears that TALEs are a basal feature of the genus and have diverged with the major clades within the genus. In general the TALEs of rice infecting (*X. oryzae*) and non-rice-infecting xanthomonads show different patterns of diversity. The rice infecting strains tend to bear TALEs in an excess of 15 per strain. In addition TALEs of *X. oryzae* pv *oryzae*, though not *X. oryzae* pv *oryzicola*, strains tend to be arranged in tandem gene repeats flanked by transposon inverted repeats sequences (11, 61). The two may well be linked since not only are *TALE* genes in many cases flanked by Tn3 transposon inverted repeat sequences, but these are flanked by 5 bp direct repeats, an indicator of a transposon event (61). However, that *X. oryzae* pv *oryzicola* strains also bear *TALE* genes in multicopy without evidence of transposition speaks against this.

In the non-rice-infecting strains TALEs are often plasmid localized (61) The significance of this for *TALE* evolution and distribution has not been studied. However, it is known that xanthomonads can transfer plasmids between strains (62) so plasmid localized *TALEs* may be mobile, at least among closely related strains.

TALE repeat domain evolution involves repeat deletion, and repeat recombination

In an experimental set-up selecting for mutants of *X. oryzae* pv *oryzae* TALE *AvrXa7* Yang *et al*. (39) were able to recover *TALE* mutants with deletions of large sections of canonical repeats, replicating a previous finding of Yang and Gabriel, 1995 (63) for *X. citri* TALE *PthA*. They also found evidence of repeat recombination between AvrXa7 and some other *TALE* gene from the genome of this strain.

Potential molecular mechanisms allowing for changes in repeat number and composition in *TALE* genes include intra-genic recombination, inter-genic recombination and replication slippage. Ferreira *et al*. prefer the latter without providing supporting evidence for this (61) and it is clear from Yang *et al*. that recombinatorial processes can also direct TALE repeat domain evolution (39).

2.4 New frontiers in TALE technology

Question four: Could TALE-likes be used to make novel contributions to TALE technology?

TALE technology is the use of TALE-derived molecular tools to carry out desired functions. This requires modification of natural TALEs to create the appropriate tools. This of course involves the assembly of dTALEs with different BSRs and also testing different NTR/CTR truncations coupled to domain fusions. Furthermore, *TALE* gene codon optimization and the use of non-natural promoters and terminators could also qualify. I will not devote further space here to dTALE assembly or domain fusions, but rather look toward what the current challenges in TALE engineering. In order to make judgments on the logical next steps for TALE technology it is important to mention the recently arrived rival system: CRISPR/Cas9.

RNA-binding protein Cas9, together with a guide RNA supplied from a CRISPR locus, form a sequence specific DNA binding complex based on Watson-Crick base pairing (64). The specificity of the system can be reprogrammed simply by modifying the 20 bp spacer region of the guide RNA gene, which makes CRISPR/Cas9 an incredibly simple and therefore attractive system. This system was adapted from bacterium *Streptococus pyogenes* but is widespread among prokaryotes. The natural Cas9 protein acts a DNA endonuclease, allowing the complex to carry out its function of identifying and destroying viral DNA within the cell. Endonuclease defective versions have been created and additional domains fused to Cas9 to adapt CRISPR/Cas9 for use as a targeted transcription factor for example (65).

Considering the efficacy and simplicity of the CRISPR/Cas9 it is worth questioning what the value is of continuing to use TALE technology. Indeed the initially identified drawbacks of the CRISPR/Cas9 platform have indeed been largely overcome already. Two limitations that we identified previously (7) are that guide RNA expression is typically driven by constitutive Pol I promoters and therefore can't be regulated, and the Cas9 protein will accept any guide RNA expressed in the cell, limiting orthogonal construct deployment. It has now been shown that guide RNAs can be expressed from pol II promoters if they are flanked by ribozymes to remove 5' and 3' caps which would lead to nuclear export of the guide RNA (66). Also, several Cas9 proteins from different bacterial species, each with their own specific guide RNA upstream recognition sequence requirements have been characterized (67). It is therefore possible to have several sets of orthogonal Cas9s and guide RNAs in one system. In addition using RNA scaffolds allows specific functional domains to be linked to specific guide RNAs (68). Therefore, by now the available regulatory nodes for CRISPR/Cas9 are diverse and unlikely to be a real limitation for implementation, even for complex applications.

The more serious drawback of the CRISPR/Cas9 system is that the DNA binding interface cannot be easily modified. The number of possible binding interactions is limited to Watson-crick base pairing. By contrast there are 20 natural amino acids which could occupy the BSR position of a TALE repeat, as well as the commonly occurring G_{SL} repeats, missing the BSR residue (7). In addition polymorphisms at other positions may be modified to alter the protein-DNA interface, which has been demonstrated at least for position 12 (30). This could prove advantageous in certain applications where the precise adjustment of affinity and specificity between protein and DNA partner (7) is desirable. It could also be possible to make dTALEs targeting the same target sequence with different binding strengths. Such a library of dTALEs for a single DNA element could be advantageous in synthetic biology in particular, where they could be used as modular resistor elements of varying strength in synthetic genetic circuitry, as has already been partially demonstrated (69).

The other advantage of the dTALE system is the potential to design repeat arrays with very low mismatch tolerance. A comparison to CRISPR/Cas9 found that TALE repeat arrays had lower mismatch tolerance (65). In addition using less common BSRs and non-naturally occurring residues at 12 allowed the creation of dTALEs with even lower mismatch tolerance than a standard dTALE, without compromising on-target activity (30). Such engineering options are not open to the CRISPR/Cas9 sequence, where base-pairing properties and target length are fixed, and till now studies have found that the off-target activity of CRISPR/Cas9 in complex genomes is too high to be useful for applications such as gene therapy, requiring high fidelity targeting (70). Another study found that using two modified CRISPR/Cas9 nickases (analogous to a TALEN pair) reduced off-target activity to the same level as a TALEN pair within a human cell targeted cleavage set-up (40). However, for applications such as transcriptional modification, where single dTALEs and not pairs are used, the potentially greater targeting stringency could make dTALEs the more attractive tool.

The possibility of higher-fidelity, tunable targeting with TALEs may provide an advantage in the areas of gene therapy and synthetic biology. In the former case off-target mutations must be minimized. In the later case synthetic genetic circuitry requires a toolbox of tunable transcriptional regulators which can function orthologously in a synthetic system, which has been demonstrated for dTALEs (71).

If, as I believe, the future of applied TALE technology lies in synthetic biology and high fidelity targeting applications, then certain engineering goals can be outlined. I think that more attention should be paid to the repeat-DNA interface, modulating it through the careful selection of BSR and non-BSR polymorphisms. This could allow fine-tuned control of DNA binding domain properties by modifying the protein-DNA interface. The exploration of non-BSR polymorphisms may also aide in the creation of dTALEs with even higher sequence specificity without compromising activity; the main issue for high fidelity applications such as gene therapy.

Finally, an unresolved problem of TALE technology, relevant to all applications, is that *TALE* genes seem to be rather unstable. Specifically repeats are lost over time, with the magnitude of the effect dependent on the particular cell environment or delivery system (72, 73). Repeat loss could totally inactivate the dTALE or lead to novel target specificities if the loss is small in scale. This would be devastating for gene therapy, where high fidelity is required; and for synthetic biology applications where dTALEs might need to be stably expressed in a synthetic biological system potentially indefinitely. It may be possible to alleviate this problem by reducing average pairwise repeat similarity at the DNA level. Only so much can be achieved with alternative codon usage and thus studies into the potential for modifying repeat amino acid sequence without impairing activity would be desirable.

3 Breaking the DNA-binding code of *Ralstonia solanacearum* TAL effectors provides new possibilities to generate plant resistance genes against bacterial wilt disease

This chapter is identical to the publication:

de Lange, Orlando, **Schreiber, Tom, Schandry, Niklas, Radeck, Jara, Braun, Karl Heinz, Koszinowski, Julia, Heuer, Holger, Strauß, Annett & Lahaye, Thomas**

Breaking the DNA-binding code of *Ralstonia solanacearum* TAL effectors provides new possibilities to generate plant resistance genes against bacterial wilt disease

New Phytologist (2013) **3**: 773-786.

New Phytologist

Research

Breaking the DNA-binding code of *Ralstonia solanacearum* TAL effectors provides new possibilities to generate plant resistance genes against bacterial wilt disease

Orlando de Lange[1], Tom Schreiber[2], Niklas Schandry[1], Jara Radeck[1], Karl Heinz Braun[1], Julia Koszinowski[1], Holger Heuer[3], Annett Strauß[1] and Thomas Lahaye[1]

[1]Genetics, Department of Biology I, Ludwig-Maximilians-University Munich, D-82152, Martinsried, Germany; [2]Institute of Biology, Department of Genetics, Martin-Luther-University Halle-Wittenberg, D-06099, Halle (Saale), Germany; [3]Julius Kühn Institute, Messeweg 11/12, D-38104, Braunschweig, Germany

Author for correspondence:
Thomas Lahaye
Tel: +49 (0)89 2180 74740
Email: lahaye@bio.lmu.de

Received: *5 March 2013*
Accepted: *3 April 2013*

New Phytologist (2013) **199:** 773–786
doi: 10.1111/nph.12324

Key words: AvrBs3 (avirulence protein triggering *Bs3* resistance), Brg11(*hrpB*-regulated 11), designer transcription activator-like effector (TALE), GMI1000, *Ralstonia solanacearum*, *Ralstonia* transcription activator-like effector (TALE)-like (RTL), repeat variable diresidue (RVD), transcription activator-like (TAL) effectors.

Summary

- *Ralstonia solanacearum* is a devastating bacterial phytopathogen with a broad host range. *Ralstonia solanacearum* injected effector proteins (Rips) are key to the successful invasion of host plants. We have characterized Brg11(*hrpB*-regulated 11), the first identified member of a class of Rips with high sequence similarity to the transcription activator-like (TAL) effectors of *Xanthomonas* spp., collectively termed RipTALs.
- Fluorescence microscopy of *in planta* expressed RipTALs showed nuclear localization. Domain swaps between Brg11 and *Xanthomonas* TAL effector (TALE) AvrBs3 (avirulence protein triggering *Bs3* resistance) showed the functional interchangeability of DNA-binding and transcriptional activation domains. PCR was used to determine the sequence of *brg11* homologs from strains infecting phylogenetically diverse host plants.
- Brg11 localizes to the nucleus and activates promoters containing a matching effector-binding element (EBE). Brg11 and homologs preferentially activate promoters containing EBEs with a 5′ terminal guanine, contrasting with the TALE preference for a 5′ thymine.
- Brg11 and other RipTALs probably promote disease through the transcriptional activation of host genes. Brg11 and the majority of homologs identified in this study were shown to activate similar or identical target sequences, in contrast to TALEs, which generally show highly diverse target preferences. This information provides new options for the engineering of plants resistant to *R. solanacearum*.

Introduction

Ralstonia solanacearum is the causal agent of the crop diseases most commonly referred to as bacterial wilt (Genin, 2010). Known for its broad host range, this pathogen infects many economically important crop and ornamental species, including solanaceous crops potato, tomato, tobacco, pepper and aubergine) but also nonsolanaceous dicots (geranium and peanut) and monocots (ginger and banana). Endemic in much of Asia and the Americas (Elphinstone, 2005), *R. solanacearum* is also considered a quarantine pest in the European Union (EU) as well as a potential bioterrorism agent in the USA (Lambert, 2002).

A range of type III effectors are known to be translocated into plant host cells by *R. solanacearum* and contribute to the extraordinarily broad host range of this plant pathogen (Genin & Boucher, 2004; Remenant *et al.*, 2010). Here we provide a molecular characterization of the *R. solanacearum* type III effector Brg11 (*hrpB*-regulated 11; strain GMI1000; Cunnac *et al.*, 2004), previously also designated RSc1815 (*Ralstonia solanacearum chromosomal* gene 1815; Salanoubat *et al.*, 2002) or *Ralstonia* transcription activator-like (TAL)-like 2 (RTL2; Li *et al.*, 2013), as well as homologs from diverse strains of *R. solanacearum*, including Hpx17 (*hrpB*-dependent expression 17; strain RS1000; Mukaihara *et al.*, 2010). Brg11 was the first identified member of this class of effectors, notable for their sequence similarity to the transcription activator-like effectors (TALEs) of *Xanthomonas* spp. and thus termed *Ralstonia* injected protein TALs (RipTALs).

Members of the *Xanthomonas* TALE family share a common modular architecture: an N-terminal type III secretion signal marking TALEs for translocation from the pathogen into the host cell, C-terminal nuclear localization signals (NLSs), tandemly arranged repeats that facilitate sequence-specific DNA binding at promoters of target genes and a C-terminal transcriptional activation domain (AD) that induces transcription of the target gene (Bogdanove *et al.*, 2010). Distinct TALEs vary predominately in their DNA-binding specificity, allowing them to target diverse genes within the host genome. Examples of TALEs and corresponding target genes include the well-studied AvrBs3 protein of *Xanthomonas campestris* pv. *vesicatoria* (*Xcv*; also referred to as *X. euvesicatoria* or *X. axonopodis* pv. *vesicatoria*

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

(Vauterin *et al.*, 1995; Jones *et al.*, 2004)), which is believed to promote disease by transcriptional activation of the pepper *UPA20* (upregulated by AvrBs3, gene 20) gene (Kay *et al.*, 2007) but also triggers defense in pepper genotypes that contain the matching resistance (*R*) gene *Bs3* (Bacterial spot *R* gene 3; Römer *et al.*, 2007).

TALE DNA-binding repeats are typically 33–35 amino acids long and, within a given TALE, inter-repeat polymorphism is found predominantly in positions 12 and 13, the repeat variable diresidue (RVD). It was found, through analysis of TALE target boxes within promoters, that there is a simple code linking RVDs and target sequences (Boch *et al.*, 2009; Moscou & Bogdanove, 2009). Each repeat binds a single DNA base and the amino acids occupying the RVD determine base specificity. This 'TALE code' can be used to create designer TALEs (dTALEs) with desired DNA-binding specificities (Bogdanove & Voytas, 2011). As made clear by the elucidation of the crystal structures of TALEs bound to DNA (Deng *et al.*, 2012; Mak *et al.*, 2012), the binding domain forms a right-handed superhelix around the DNA, tracking along the sense strand, with each repeat making contact with a single DNA base. Contact occurs between the DNA base and TALE repeat residue 13, while residue 12 of each repeat forms a stabilizing hydrogen bond with residue 8 of the same repeat (Bochtler, 2012).

As has previously been noted, the *R. solanacearum* Brg11 protein shares 40% homology with the *Xanthomonas* TALE protein AvrBs3 (Schornack *et al.*, 2006) and possesses a central region of tandemly arranged 35-amino acid repeats, with each repeat bearing an average similarity of 55% to the repeats of AvrBs3. For this reason, Brg11 was annotated as a putative TALE within the GMI1000 genome (Salanoubat *et al.*, 2002). Subsequently homologs have been detected by PCR in over 300 *R. solanacearum* strains (Heuer *et al.*, 2007). However, despite their prevalence, the functional characterization of these proteins, beyond evidence of type III delivery and nuclear localization (Mukaihara *et al.*, 2010; Li *et al.*, 2013), remains minimal. Mutant GMI1000 derivatives lacking this effector showed reduced virulence relative to the wild-type strain in aubergine (Macho *et al.*, 2010), demonstrating that Brg11 contributes to pathogenicity.

We report here on the functional characterization of the RipTAL Brg11, showing that it contains functional NLSs, and a C-terminal AD. Furthermore, while RipTAL RVDs have previously been tested in the context of *Xanthomonas* TALE repeats (Streubel *et al.*, 2012) and a RipTAL consensus repeat (Li *et al.*, 2013), we show here for the first time predicted target sequences that are transcriptionally activated by wild-type RipTALs *in planta*. We demonstrate that repeats of the RipTAL Brg11 recognize DNA in a manner analogous to that of TALEs, with a code governed by the RVDs of consecutive repeats. In addition, however, non-RVD residues, which are highly variable in RipTALs but not TALEs, have a significant impact on the DNA recognition properties of repeats. Furthermore, Brg11 is functional only when the binding box is preceded by a 5′ guanine (G_0), in contrast to TALEs which have a preference for a 5′ thymine (T_0). Homologs of Brg11 from strains of *R. solanacearum* isolated from phylogenetically diverse host plants have > 98% homology to Brg11 and most are identical to Brg11 in repeat composition and DNA specificity.

Materials and Methods

Bacterial strains and growth conditions

Escherichia coli TOP10 and DB3.1 (Life Technologies, Carlsbad, CA, USA) were cultivated at 37°C in LB (Lysogeny Broth) and *Agrobacterium tumefaciens* GV3101 pMP90 (Koncz & Schell, 1986) at 30°C in YEB (Yeast Extract Broth) medium.

Plant material

Nicotiana benthamiana plants were grown in a glasshouse at 60–70% humidity, at 22°C during the day (16-h light) and 18°C at night. Six- to eight-wk-old plants were used for inoculation.

Generation of *brg11* and *hpx17* pENTR-D constructs

Using TOPO-BluntII-hpx17 (T. Mukaihara, Research Institute for Biological Sciences, Okayama, Japan) as a template we amplified *hpx17* with the primers brg11_CACC-ATG and brg11_no-stop (Supporting Information Table S1) and cloned the PCR product into pENTR-D (Life Technologies), generating pENTR-D_*hpx17* containing wild-type *hpx17*. To create an ENTRY clone for *brg11*, the *Xho*I/*Bgl*II fragment of pCDN2.1_*brg11* containing *brg11* (S. Genin, Laboratoire des Interactions Plantes-Microorganismes, Castanet-Tolosan, France) was cloned into the *Xho*I/*Bgl*II-cleaved pENTR-D_*hpx17*.

Creation of *hpx17* truncations and *NLS* mutants

NLS mutations and domain truncations were made to *hpx17* in pENTR-D via site- directed PCR mutagenesis utilizing Phusion polymerase (New England Biolabs, Beverly, MA, USA) with its GC buffer and PCR additive PreCES-I (Ralser *et al.*, 2006). *NLSI* (sequence HRKR) was replaced with a single *Hin*dIII site (resulting sequence QAYW) via PCR mutagenesis with primers brg11_mut-NLSI_fwd and brg11_mut-NLSI_HindIIIrev (Table S1); *NLSII* (sequence RRKR) was replaced with a single *Bam*HI site (resulting sequence PDPW) with primers brg11_mut-NLSII_BamHIfwd and brg11_mut-NLSII_rev (Table S1). *hpx17*-ΔN-Δrep was created with the removal of N-terminal- and repeat region-encoding sequences with the primers hpx17_Cterm_fwd and hpx17_Cterm_rev (Table S1), and *hpx*-17-Δrep-ΔC with hpx17_Nterm_fwd and hpx17_Nterm-rev (Table S1). Nucleotide sequences for all *hpx17* derivatives are displayed in Fig. S2.

Creation of C- and N-terminal AvrBs3/Brg11 chimeras

To generate the C-terminal chimera (CTC) and the ΔAD derivative, a fragment encoding the C-terminus of Brg11 (*brg11C*919-1245), and flanked by *Bsa*I sites, was amplified using the primers listed in Table S1. As a template, a *brg11* full-length gene, codon optimized for *in planta* expression (synthesized by GenScript;

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

Piscataway, NJ, USA), was used. The primers CTC-ΔAD Fwd and Rev (Table S1) were used to remove the 3′-most 35 codons to create *CTC-ΔAD*. The *CTC* and *CTC-ΔAD* were each combined with a *Bsa*I site-flanked fragment encoding the N-terminus of AvrBs3 into a modified pENTR-D derivative containing *Bsa*I sites and a chloramphenicol acetyltransferase (*cat*)-*ccdB* cassette via *Bsa*I cut-ligation.

N-terminal Brg11/AvrBs3 chimeras were created by replacing repeats –1 and 0 of AvrBs3 with the corresponding Brg11 region (N-terminal chimera 1 (NTC1)) or by replacing the complete N-terminus of AvrBs3 with that of Brg11 (NTC2) (Fig. S8). Both chimeras were generated by PCR mutagenesis with the primers listed in Table S1. As templates, wild-type *avrBs3* as well as *brg11* codon optimized for *in planta* expression (synthesized by GenScript) were used. Together with the *avrBs3* fragment encoding the C-terminus, the generated *NTC* fragments were cloned in a modified pENTR-D derivative containing *Bsa*I sites and a *cat-ccdB* cassette via *Bsa* I cut-ligation (see above). In each case, *Bpi*I sites between the N- and C-terminal encoding regions allowed the integration of the *avrBs3* repeats via *Bpi*I cut-ligation (according to Morbitzer *et al.*, 2011). All chimeras were recombined into the T-DNA binary vector pGWB641 (Nakamura *et al.*, 2010) via LR recombination (Life Technologies).

Creation of $\mathrm{dTALE}_{\mathrm{EBE\ Brg11}}$ and derivatives containing contiguous Brg11 repeats

Binding domains of $\mathrm{dTALE}_{\mathrm{EBE\ Brg11}}$ and derivatives were assembled as described in Morbitzer *et al.* (2011), using either the *TALE* repeat modules previously described or repeat modules encoding the repeats of Brg11 (Fig. S7) synthesized using GenScript. All constructs were recombined into the T-DNA binary vector pGWB641 (Nakamura *et al.*, 2010) via LR recombination (Life Technologies).

Creation of AvrBs3-Brg11 derivatives and promoter constructs used in the trimer test

The previously described *Bs3* promoter derivatives containing three cytosines (Cs), guanines (Gs) or thymines (Ts) instead of three adenines (As) within the upregulated by TALE $(UPT)_{\mathrm{AvrBs3}}$ box of the *Bs3* promoter (Römer *et al.*, 2007; Morbitzer *et al.*, 2010) were cloned upstream of the *uidA* (*GUS*) reporter gene into the T-DNA vector pGWB3* via *Bsa*I cut-ligation. pGWB3* was generated by LR recombination of a small linker fragment containing two *Bsa*I sites with *CACC* and *AAGG* overhangs into the pGWB3 vector (Nakagawa *et al.*, 2007). A *cat-ccdB* cassette was then cloned into the modified pGWB3 vector using an *Age*I restriction site within the linker fragment resulting in pGWB3*.

Binding domains of all *avrBs3* derivatives used in the trimer test were assembled via *Bsa*I cut-ligation of *TALE* repeats matching the RVD composition of *avrBs3*, as described in Morbitzer *et al.* (2011), along with trimers of *brg11*-derived repeats synthesized (synthesized by GenScript) with the codon usage of *Xanthomonas campestris* pv. *vesicatoria* (Fig. S5). Binding domains assembled in this fashion were cloned into a pENTR-D derivative containing fragments of *avrBs3* encoding the N- and the C-terminal parts of the TALE via *Bpi*I cut-ligation (according to Morbitzer *et al.*, 2011). *avrBs3* derivatives were transferred into the T-DNA binary vector pGWB641 (Nakamura *et al.*, 2010) via LR recombination (Life Technologies).

Creation of target promoters for Brg11 and its homologs

The predicted RipTAL binding box was introduced in each case via PCR mutagenesis in place of the UPT_{AvrBs3} box in the *Bs3* promoter (Römer *et al.*, 2007). In the case of $\mathrm{EBE}_{\mathrm{Brg11}}$, the forward primer $Bs3p_{\mathrm{EBE\ Brg11}}$ Fwd was used in combination with the reverse primers $Bs3p_{\mathrm{EBE\ Brg11\ A0}}$ for A_0, $Bs3p_{\mathrm{EBE\ Brg11\ C0}}$ for C_0, $Bs3p_{\mathrm{EBE\ Brg11\ G0}}$ for G_0 and $Bs3p_{\mathrm{EBE\ Brg11\ T0}}$ for T_0 (Table S1). Equivalent primers were used to introduce binding boxes for RipTALI-6, -9, -11 and -14 (Table S1). As a template, a pUC57 (GenScript) derivative was used containing a 360-bp fragment of the pepper *Bs3* promoter. The promoter constructs were cloned via *Bsa*I cut-ligation into the T-DNA vector pGWB3*.

Isolation and sequencing of *brg11* homologs

Genomic DNA was isolated from *Ralstonia solanacearum* (Smith 1896) (Yabuuchi *et al.*, 1995) strains as follows: cell pellets of 30-ml overnight cultures were resuspended in 50 mM Tris-HCl (pH 8), 50 mM EDTA (pH 8.0), 0.5% (v/v) Tween 20 and 0.5% (v/v) Triton X-100 buffer containing RNAse A (0.02% (w/v)), lysozym (0.27% (w/v)) and protease (0.5 AU) (Qiagen). DNA extraction and precipitation were carried out as described previously (Grover *et al.*, 2012) and the pellet was redissolved in 0.1 TE (Tris EDTA) (pH 8.0). Amplification of *RipTALs* from genomic DNA was carried out with Phusion polymerase (New England BioLabs), the PCR additive preCES-I (Ralser *et al.*, 2006) and the primers brg11_CACC-ATG and brg11_no-stop (Table S1). The PCR products were sequenced using the primers listed in Table S1. Furthermore, the PCR products for *RipTALI-1* to *-3* and *RipTALI-5* to *-7* were cloned and sequenced in the vector pENTR-D (Life Technologies). RipTALI-9, -11 and -14 were PCR-amplified from genomic DNA using primers inF RipTALIs F and R (Table S1) and cloned via inFusion (Clontech, Palo Alto, CA, USA) into a pENTR-Brg11 digested with *Xho*I and *Bgl*II. Subsequent sequencing showed an identical result to the original PCR product sequencing results for these homologs. All constructs were recombined into pGWB641 (Nakamura *et al.*, 2010) via LR recombination (Life Technologies).

Agrobacterium tumefaciens-mediated transient transformation of *N. benthamiana* plants and GUS assays

For subcellular localization and GUS assays, *A. tumefaciens* strains were grown overnight in YEB medium containing rifampicin and kanamycin (each 100 $\mu g\ ml^{-1}$; for pGWB3-, pGWB3*- and pGWB5-containing strains) or rifampicin and spectinomycin (each 100 $\mu g\ ml^{-1}$; for pGWB641-containing strains), collected by centrifugation, resuspended in inoculation medium (10 mM $MgCl_2$, 5 mM MES, pH 5.3, and 150 μM acetosyringone), and adjusted to an optical density at 600 nm (OD_{600nm}) of 0.8. For localization studies, equal amounts of *A. tumefaciens* strains containing *brg11*, *hpx17* or its derivatives as *green fluorescent protein* (*GFP*) fusions and the silencing inhibitor *p19* (Voinnet *et al.*, 2003) were mixed and inoculated into *N. benthamiana* leaves by blunt-end syringe infiltration. For GUS assays, equal amounts of *A. tumefaciens* strains containing *35S*-promoter-driven *RipTAL* genes, *avrBs3* or *avrBs3/brg11* chimeras and *Bs3* promoter constructs containing corresponding binding boxes fused to the reporter gene *uidA* (*GUS*) were mixed before inoculation. GUS staining was carried out as described previously (Strauß *et al.*, 2012). Wherever shown in the figures, stained leaf discs were selected as representative for results obtained in at least two independent experiments each including three separate plants. For quantification of GUS activity, two leaf discs were taken from each of three separate plants, pooled and homogenized in a tissue lyzer. Two hundred micromole GUS extraction buffer (50 mM sodium phosphate (pH 7), 10 mM DTT, 10 mM EDTA, 0.1% N-lauryl-sarcosine, 0.1% Triton-X100, and 1× protease inhibitor cocktail (Roche; complete mini, EDTA-free)) was then added and quantification of GUS activity carried out as described previously (Kay *et al.*, 2007). Samples for each time-point were taken in duplicate with a plate reader (TECAN, Maennedorf, Switzerland) and averaged, and pmol 4-MU min^{-1} μg^{-1} protein extract values were calculated and averaged for the three biological replicates.

Microscopy

For subcellular localization, *brg11*, *hpx17* and derivatives were cloned into pGWB5 (Nakagawa *et al.*, 2007) by LR recombination (Life Technologies) creating C-terminal GFP fusions, transformed into *A. tumefaciens*, and transiently expressed in *N. benthamiana* leaves as already described in the section '*Agrobacterium tumefaciens* mediated transformation of *N. benthamiana* plants and GUS assays'. Two to 3 d post inoculation, epidermal cells of *N. benthamiana* were inspected with a confocal laser-scanning microscope (TCS SP5; Leica Microsystems, Bensheim, Germany) equipped with a Leica HCX PL APO 20× water immersion objective. Images were processed using the Leica AF software and IMAGEJ (Schneider *et al.*, 2012).

Results

RipTALs Brg11 and Hpx17 are nuclear localized, with contributions from N- and C-terminal regions

Brg11 has previously been shown to localize to the plant nucleus (Li *et al.*, 2013). Having confirmed the nuclear localization of Brg11 in leaves of *N. benthamiana* (Fig. 1a) and demonstrated that Hpx17 also localizes to the nucleus (Fig. 1b), we tested truncation and mutation variants to identify the corresponding NLSs within these proteins. AvrBs3 is known to possess two functional C-terminal NLSs whose removal prevents nuclear localization (Van den Ackerveken *et al.*, 1996). Sequence comparisons identified similar motifs in the C-terminus of Brg11 and Hpx17 (Figs S1–S3), yet mutation of these potential NLSs in Hpx17 did not impair nuclear localization (mut-NLS; Fig. 1c). For this analysis *hpx17* was used as it lacks the vast majority of the repeats and is therefore more amenable to PCR-based modifications. NLS prediction software (e.g. NLSTRADAMUS; Nguyen Ba *et al.*, 2009) identified an additional putative NLS in the N-terminal domain of Brg11 and Hpx17 (Figs S1, S2) and, consistent with this prediction, removal of the N-terminal and repeat domains in addition to the aforementioned mutations led to nuclear and cytoplasmic localizations (Hpx17-ΔN-Δrep-mut-NLS; Fig. 1d). The same truncation without removal of the C-terminal NLSs was not sufficient to ablate nuclear localization (Hpx17-ΔN-Δrep; Fig. 1e), indicating that it is not simply the lower molecular weight that allows passive diffusion into and out of the nucleus. The N-terminus of Hpx17 alone (Hpx17-Δrep-ΔC) localizes exclusively to the nucleus (Fig. 1f). These observations show that N- and C-terminal sequences within Hpx17 contribute in a functionally redundant manner to nuclear localization. Given that Brg11 and Hpx17 are almost identical in the N- and C-terminal nonrepeat regions (Figs S1, S2), it is likely that the same sequences contribute to the nuclear targeting of Brg11 (Fig. 1a).

The C-terminus of Brg11 contains a functional eukaryotic AD

Key to the disease-promoting function of TALEs is their ability to activate plant promoters (Bogdanove *et al.*, 2010). In the case of the TALE AvrBs3, an AD is found in the far C-terminus (Szurek *et al.*, 2001; Figs S1, S3a). Prediction software (Piskacek *et al.*, 2007) identified a eukaryotic nine-amino acid transactivation domain in the C-terminal-most nine amino acids of Brg11 (Figs 1, S3). We thus created a construct encoding a CTC in which the C-terminal 295 amino acids of AvrBs3 were replaced with the equivalent region of Brg11 (Fig. S3b) and tested its ability to activate the pepper *Bs3* promoter (*Bs3p*), which contains a well-characterized AvrBs3 binding box (Römer *et al.*, 2009). The *35S*-promoter-driven *avrBs3* or *CTC* and a *Bs3p-uidA* reporter were co-delivered via *A. tumefaciens* into *N. benthamiana* leaves and GUS assays were performed (Fig. 2). In this assay, blue staining of leaf discs is caused by expression of the GUS reporter and indicates TALE-mediated promoter activation. The CTC chimera was indeed able to activate the GUS reporter, indicating the presence of a functional AD in the C-terminal region of Brg11 (Fig. 2). The stronger reporter activation observed with AvrBs3 as compared with the chimeric protein may indicate a functional difference in the strength of the activation domains of AvrBs3 and Brg11. Alternatively, the Brg11 AD performs suboptimally in the chimeric context. We also tested a truncation variant of the chimera that lacks 33 amino acids at its C-terminus, including the predicted activation domain (CTC-ΔAD; Fig. S3b). CTC-ΔAD did not activate *Bs3p* (Fig. 2), indicating that the predicted Brg11 AD is necessary to the predicted function of Brg11 as a transcriptional activator.

The repeats of Brg11 mediate sequence-specific DNA recognition with an RVD code similar to that of TALEs

Having ascertained that Brg11 localizes to the nucleus and possesses a functional AD, we sought to determine if its central

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

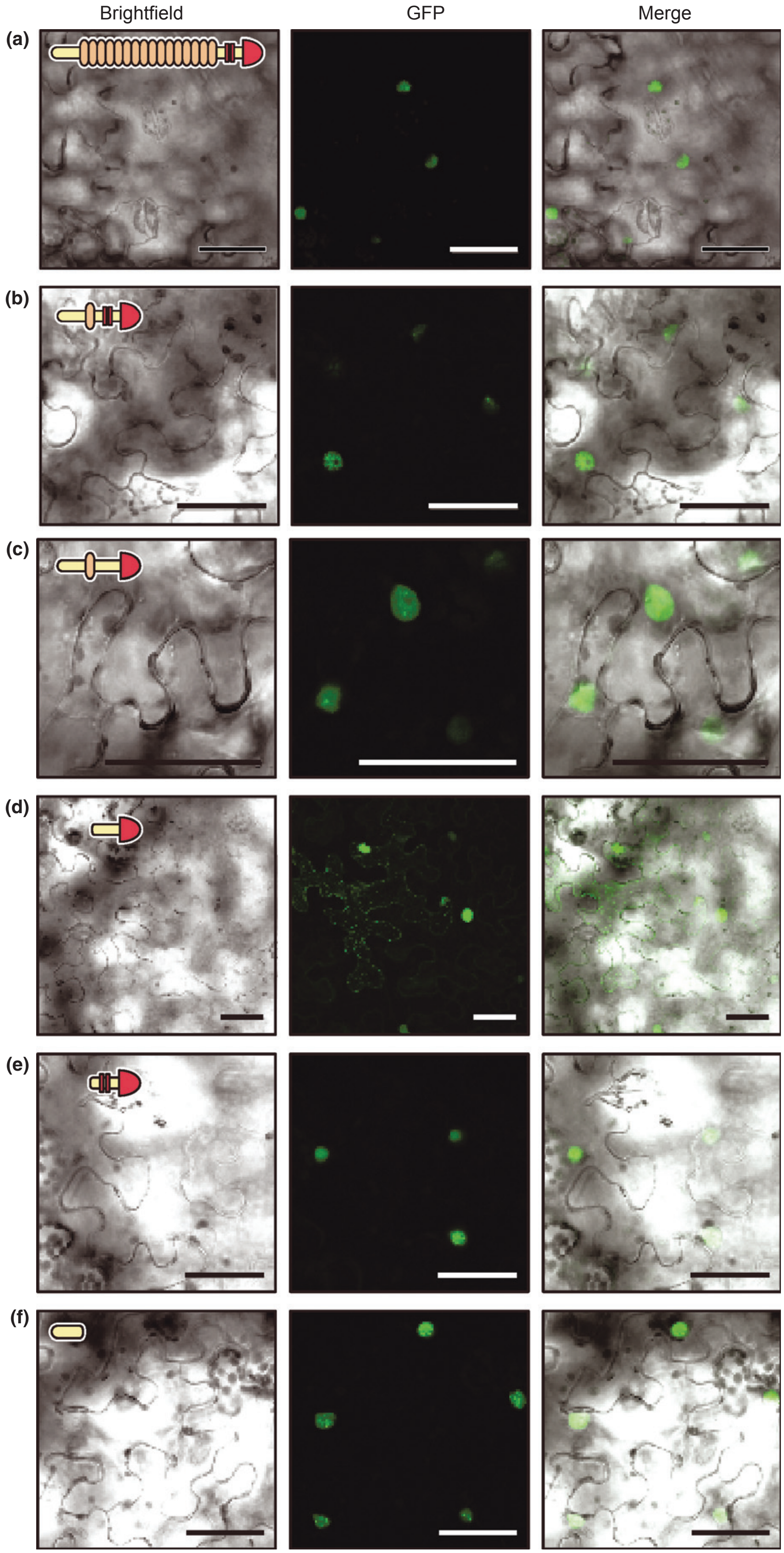


Fig. 1 Subcellular localization of *Ralstonia* injected protein transcription activator-like (RipTAL) proteins and deletion derivatives in *Nicotiana benthamiana* leaf cells. Leaves were coinfiltrated with *Agrobacterium tumefaciens* strains transformed with pGWB5 T-DNA vectors containing DNA sequences under control of the constitutive cauliflower mosaic virus *35S* promoter encoding C-terminal GFP fusion proteins of (a) Brg11, (b) Hpx17 (hrpB-dependent expression 17), (c) Hpx17 with both C-terminal nuclear localization signals (NLSs) mutated (mut-NLS), (d) the Hpx17 C-terminal region with putative NLSs mutated (ΔN-Δrep-mut-NLS), (e) the Hpx17 C-terminal region (ΔN-Δrep), and (f) the Hpx17 N-terminal region (Δrep-ΔC). In all cases an *A. tumefaciens* transformant containing a *35S*-promoter-driven *p19* gene that encodes a viral silencing suppressor was coinfiltrated. Leaf discs were harvested 2–3 d after inoculation and imaged using a confocal laser-scanning microscope. Bars, 50 μm. Cartoon representations of RipTALs are shown next to the corresponding microscope pictures: ellipses indicate putative DNA-interacting repeats; a semicircle (Brg11) indicates a putative transcriptional activation domain; rectangles indicate confirmed or putative NLSs.

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

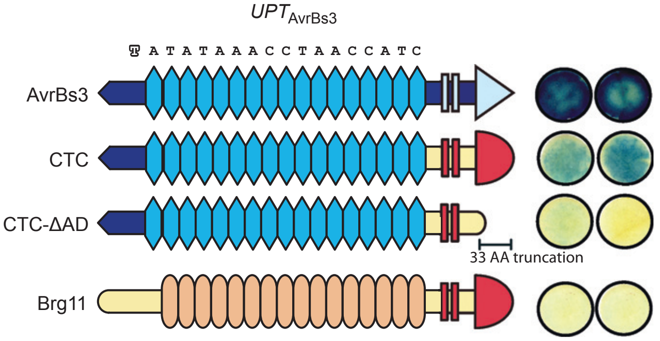


Fig. 2 Transcriptional activation of *uidA* reporter gene constructs by AvrBs3 (avirulence protein triggering *Bs3* resistance)and derivatives bearing C-terminal regions of Brg11 (*hrpB*-regulated gene 11). Gene constructs encoding the depicted proteins were co-delivered into *Nicotiana benthamiana* leaves via *Agrobacterium tumefaciens*-mediated T-DNA transfer along with a pepper *Bs3* (Bacterial spot *R* gene 3) promoter fragment, bearing an AvrBs3 target box (UPT_{AvrBs3}), upstream of a *uidA* reporter gene. Leaf discs were harvested 26 h post infiltration and stained for GUS activity. Brg11 serves as a negative control for background reporter activity. In the C-terminal chimera (CTC), the C-terminal-most 295 amino acids of AvrBs3 were replaced with the equivalent region of Brg11 (Supporting Information Fig. S3). CTC-ΔAD is identical to the CTC but lacks the C-terminal-most 35 amino acids, including the predicted transcriptional activation domain (AD) (Fig. S1). Cartoon representations of AvrBs3, Brg11 and chimeras are shown next to the corresponding leaf discs: blue, AvrBs3-derived domains; yellow/orange/red, Brg11 domains; polygons (AvrBs3) or ellipses (Brg11), repeats; a triangle (AvrBs3) or semicircle (Brg11), a confirmed or putative transcriptional AD; rectangles, confirmed or putative nuclear localization signals. The UPT_{AvrBs3} sequence is shown at the top of the figure with the 5′ thymine shown as a white letter with a black frame.

repeat domain also mediates sequence-specific DNA binding and, if so, whether a code operates similar to that of TALEs. The key features of the TALE code are the one-to-one repeat to DNA base relationship, and the determination of base specificity by the RVDs (Boch *et al.*, 2009; Moscou & Bogdanove, 2009). While an alignment of all the core repeats of AvrBs3 (Fig. S4) shows that inter-repeat polymorphisms are limited to the RVDs and few other residues, polymorphisms between Brg11 repeats are not restricted to the RVDs and are far higher in frequency (Fig. S4). In light of the many non-RVD polymorphisms among Brg11 repeats, we tested the DNA recognition properties of each repeat individually. To do so, we utilized a system, herein referred to as the trimer test, adapted from a previous study (Morbitzer *et al.*, 2010) where Brg11-derived test repeats are set into the repeat array of AvrBs3 (Fig. S5). In AvrBs3, repeats 5–7 are occupied by three repeats with the RVD NI, corresponding to a run of three adenines in the AvrBs3 target box within the *Bs3* promoter (referred to as $Bs3p_{3xA}$; Römer *et al.*, 2007). In the trimer test, AvrBs3 repeats 5–7 were replaced by three identical, tandemly arranged repeats of Brg11 (Fig. S5). To determine base specificity of each of the AvrBs3-embedded Brg11 repeats, the three adenines in the AvrBs3 target site of the *Bs3* promoter were replaced by three guanines ($Bs3p_{3xG}$), thymines ($Bs3p_{3xT}$) or cytosines ($Bs3p_{3xC}$) and these promoter derivatives were fused upstream of a *uidA* reporter gene. Subsequently we tested each of the AvrBs3 derivatives for their ability to activate each of the four distinct promoter constructs. *35S*-promoter-driven genes encoding the AvrBs3 derivatives were co-delivered together with each of the four distinct promoter-reporter constructs into *N. benthamiana* leaves via *A. tumefaciens*-mediated T-DNA transfer, and GUS activity was quantified 48 h post infiltration. This process was carried out for each of the 16 core repeats of Brg11. Repeats 8 and 11, identical in all positions at the amino acid level (Fig. S4), were tested as one construct (referred to as repeat 8/11). As specificity controls for this assay, we included trimers of *Xanthomonas* TALE repeats with RVDs HD, NG, NN or NK. The results show that in each case at least one reporter construct was activated (Fig. 3, Table S2). In most cases, the RVD code for Brg11 repeats matched expectations based on known TALE repeat specificities, although some notable exceptions were discovered (summarized in Table 1): RVDs NT and NH showed more relaxed base specificity than previously reported for the same RVDs in TALE repeats (Streubel *et al.*, 2012), while NK showed the same base specificity but higher activity (activation of $Bs3p_{3xG}$ with derivatives of Brg11 NK repeats 2 and 9 was 15–18 times higher than for corresponding TALE-derived NK repeats; Fig. 3). Overall, however, our data suggest that base specificity of Brg11 repeats is determined by repeat residues 12 and 13, as is the case for *Xanthomonas* TALE repeats. To confirm this, we reciprocally exchanged the RVDs of Brg11 repeats with clearly contrasting specificities, 8/11 (RVD HD; pairs preferentially to cytosine) and 12 (RVD SH; pairs preferentially to guanine). As would be expected if the RVDs determine specificity, we observed a reciprocal switch in DNA recognition specificities in the trimer test (Fig. S6).

In order to investigate whether differences in reporter activation by derivatives bearing repeats with the same RVD could be attributed to specific non-RVD residues, we compared repeats 10 and 8/11. The derivative bearing the repeat 10 trimer gave 25 times poorer activation of the $Bs3p_{3xC}$ reporter compared with repeat 8/11 (Fig. 3, Table S2), which shares the same RVD. Alignment of these repeats showed that they differ at three residues (Fig. 4). Of these polymorphic residues, the asparagine in position 16 is found only in repeat 10, while all other Brg11 repeats have lysine at this position (Fig. S4). Analysis in the trimer test found that replacing the asparagine at position 16 of repeat 10 with lysine (defined as repeat 10 N16K) led to reporter activation to the same level as that recorded for repeat 8/11 (Fig. 4). The converse mutation (defined as repeat 8/11 K16N), while not abolishing activity entirely, did render 8/11 K16N much weaker in its activation of the reporter as compared with the wild-type (Fig. 4). In summary, our data support the conclusion that non-RVD residues can have a profound effect on the DNA recognition capabilities of repeats.

The N-terminus of Brg11 mediates recognition of a G_0 base necessary for activation of a promoter bearing the Brg11 target box

Based on results obtained for each repeat in isolation, we predicted a 16-bp effector-binding element (EBE) for Brg11 (Fig. 3), which we inserted in place of the AvrBs3 binding site in the *Bs3* promoter ($Bs3p_{EBE\ Brg11}$) upstream of a *uidA* reporter

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

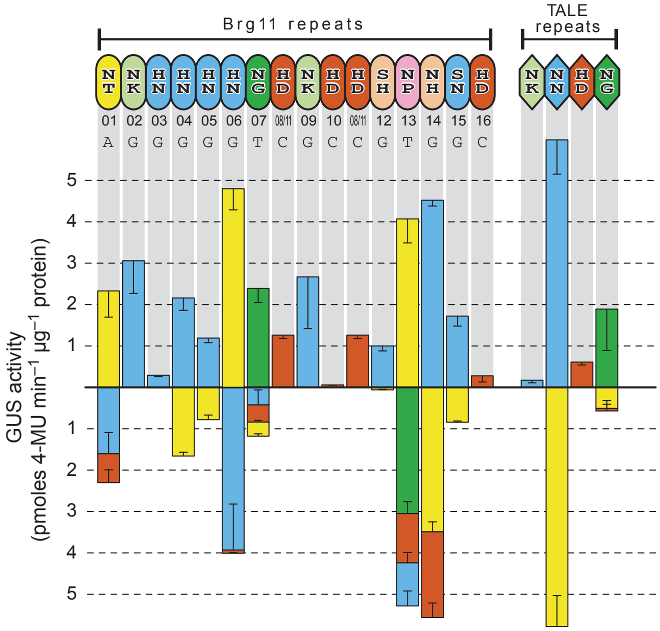


Fig. 3 Analysis of the base specificity of Brg11 (*hrpB*-regulated gene 11) repeats using the trimer test. AvrBs3 (avirulence protein triggering *Bs3* resistance) derivatives, bearing a trimer of identical Brg11 repeats, were co-delivered into *Nicotiana benthamiana* leaves via *Agrobacterium tumefaciens*-mediated T-DNA transfer along with a *uidA* reporter gene that is under transcriptional control of the pepper *Bs3* (Bacterial spot *R* gene 3) promoter or derivatives thereof. The AvrBs3 target site in the *Bs3* promoter contains three adenines (3xA) at the position corresponding to the Brg11 trimer embedded in AvrBs3. In the *Bs3* promoter derivatives, the three adenines were replaced with three guanines (3xG), cytosines (3xC) or thymines (3xT). Leaf discs were harvested 48 h post infiltration, total protein was extracted and GUS activity was quantified. GUS activity is given in picomoles 4-MU min^{-1} µg^{-1} protein. The size of the bar indicates the level of GUS activity. The color of the bars indicates the reporter construct used (yellow for 3xA, blue for 3xG, orange for 3xC and green for 3xT). The reporter with the highest GUS activity is shown above the *x*-axis, and the results for any other activated reporters are shown below this line. Data shown are based on three biological replicates. Standard errors are indicated. Results are shown for each repeat in order of appearance in the polypeptide. Brg11 repeats, displayed as ovals, are shown in order of appearance in the polypeptide as a guide and repeat numbers are given underneath accordingly. The Brg11 target sequence that was predicted and experimentally studied is displayed in black font below the repeat numbers. Repeats with different repeat variable diresidues (RVDs) are displayed in distinct colors. Trimers of AvrBs3-derived repeats bearing NK, NN, HD or NG as the RVD are shown on the right-hand side as specificity controls and cartoon representations of the corresponding repeats, shown as colored polygons, are given. The same color code is used for these repeats as for the Brg11 repeats. Results for AvrBs3$_{\mathrm{Brg11\ rep8/11}}$ are duplicated in the figure (see repeat positions 8 and 11) as this construct was used to test both repeats 8 and 11 of Brg11, which are identical at the amino acid level. The data corresponding to the graphical display of this figure given in Table S2.

Table 1 DNA recognition properties of Brg11 repeat variable diresidues (RVDs)

RVD	Best match	Additional match(es)	Differences in activity compared with same RVD in TALE repeats
NT	A	G/C	Additional C nucleotide recognition[1]
NK	G	–	Stronger reporter activation [2]
SH	G	–	Stronger reporter activation[1]
NH	G	A/C	Less specific [1,3]
NG	T	A/C/G	Additional G nucleotide recognition
NP	A	T/C/G	Different best match nucleotide [1]
HD	C	–	Repeats 10 and 16, weaker reporter activation [2]
HN	G	A	–
SN	G	A	Stronger reporter activation[1]

Results are according to Fig. 3 and Table S2.

[1] Streubel *et al.* (2012).

[2] Comparison to transcription activator-like effector (TALE) repeat controls in this study.

[3] Cong *et al.* (2012).

gene. *Agrobacterium tumefaciens*-mediated transformation of *N. benthamiana* leaves with a *35S*-promoter-driven *brg11* in combination with the *Bs3p*$_{\mathrm{EBE\ Brg11}}$ promoter, however, led to no activation above background level (Fig. 5). A dTALE designed to bind the Brg11 EBE (dTALE$_{\mathrm{EBE\ Brg11}}$) did, however, activate *Bs3p*$_{\mathrm{EBE\ Brg11}}$ (Fig. 5), thereby demonstrating that the reporter construct was functional. In order to confirm that the repeats of the Brg11 core binding domain could successfully recognize the predicted EBE$_{\mathrm{Brg11}}$, we tested the entire binding domain of Brg11 as contiguous blocks of five to six repeats each embedded into the binding domain of dTALE$_{\mathrm{EBE\ Brg11}}$ (Fig. S7). Chimeras containing these Brg11 repeat blocks were found to be functional to a level comparable to the positive control dTALE$_{\mathrm{EBE\ Brg11}}$ (Fig. 5), suggesting that the lack of reporter activation by Brg11 was attributable not to the core binding repeats but rather to some other region(s) of the protein. We postulated that an as yet undefined, strict base preference, mediated by putative N-terminal degenerate repeats, analogous to those of TALEs (Bochtler, 2012), made Brg11 incompatible with the predicted target sequence. To investigate whether the Brg11 N-terminal domain contributes to DNA recognition specificity, we created two NTCs for which either repeats 0 and −1 (NTC1) or the entire N-terminus (NTC2) of AvrBs3 was replaced with the equivalent sequence from Brg11 (Fig. S8). The *in planta* functionality of the Brg11-AvrBs3 chimeras was then tested, against *Bs3p* (T_0) and three derivatives with a 5′ terminal adenine (A_0), cytosine (C_0) or guanine (G_0) upstream of *uidA*. Both chimeras, NTC1 and NTC2 displayed a strong preference for G_0, activating other promoters very weakly or not at all (Fig. 6). Thus, the N-terminal regions of Brg11 and *Xanthomonas* TALEs both contribute to DNA recognition but differ with respect to their base specificity.

Accordingly, A_0, G_0 and C_0 derivatives of *Bs3p*$_{\mathrm{EBE\ Brg11}}$ were created. Brg11 and dTALE$_{\mathrm{EBE\ Brg11}}$ were then tested against each of these promoters in addition to the original T_0 version with AvrBs3 as a negative control (Fig. 7). Indeed, Brg11 activated the G_0 promoter, and only this promoter (Fig. 7). dTALE$_{\mathrm{EBE\ Brg11}}$ was able to activate all promoters irrespective of the identity of the N_0 base (Fig. 7). The lack of preference for T_0 in this dTALE is possibly a result of the high affinity across the rest of the core repeats where every RVD is paired with its best match base. This is not the case for AvrBs3 and the UPT_{AvrBs3} (Fig. 6) box, where some RVDs are mismatched, which might explain why in this TALE-EBE combination the N-terminal TALE repeats exert a stronger discriminating effect than is the case for dTALE$_{\mathrm{EBE\ Brg11}}$

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

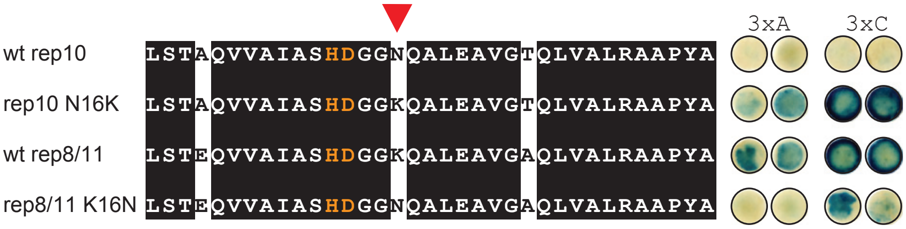

3xA
3xC
wt rep10 LSTAQVVAIASHDGGNQALEAVGTQLVALRAAPYA
rep10 N16K LSTAQVVAIASHDGGKQALEAVGTQLVALRAAPYA
wt rep8/11 LSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYA
rep8/11 K16N LSTEQVVAIASHDGGNQALEAVGAQLVALRAAPYA

Fig. 4 Functional impact of an asparagine versus lysine polymorphism at the non-repeat variable diresidue (RVD) repeat position 16. The alignment shows differences between the Brg11 (*hrpB*-regulated 11) wild-type (wt) repeats 10 and 8/11. Polymorphic residues are highlighted with black letters on a white background. Of these, position 16 (indicated with a red triangle) has been exchanged for the corresponding residue found in the other repeat in each case (rep10 N16K; rep8/11 K16N). RVDs are highlighted in orange font. The functionality of the repeat derivatives was defined using the trimer test as described in Fig. 3, except that *Nicotiana benthamiana* leaf discs were stained to visualize reporter activity. The intensity of the staining corresponds to promoter activity and is used as a proxy indicator of binding at the promoter.

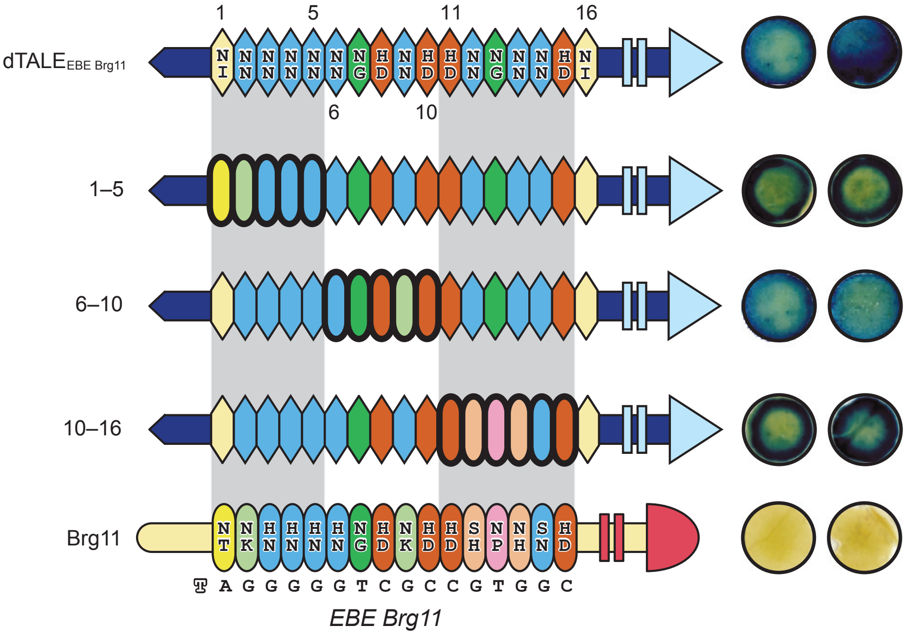

1
5
11
16
dTALE$_{\text{EBE Brg11}}$
N N N N N N N H N H H N N N N H N
I N N N N N G D N D D N G N N D I
6
10
1–5
6–10
10–16
Brg11
N N H H H H N H N H H S N N S H
T K N N N N G D K D D H P H N D
T A G G G G G T C G C C G T G G C
EBE Brg11

Fig. 5 Functional analysis of Brg11 (*hrpB*-regulated 11) repeat subarrays. The AvrBs3 (avirulence protein triggering *Bs3* resistance) binding site in the *Bs3* promoter was replaced by the predicted Brg11 binding site (effector-binding element (EBE)$_{\text{Brg11}}$) preceded by a 5′-terminal thymine to allow activation by a matching designer transcription activator-like effector (dTALE$_{\text{EBE Brg11}}$). We also generated genes encoding dTALE$_{\text{EBE Brg11}}$ derivatives, depicted on the left, in which the first five (1–5), the second five or the last six (11–16) repeats of dTALE$_{\text{EBE Brg11}}$ were replaced by corresponding *brg11* repeat subarrays. *35S*-promoter-driven *dTALE$_{EBE\ Brg11}$*, *brg11* and the above-described *dTALE$_{EBE\ Brg11}$* derivatives were cotransformed into *Nicotiana benthamiana* leaves via *Agrobacterium tumefaciens*-mediated T-DNA delivery, along with a *uidA* reporter gene that was downstream of the *Bs3* promoter derivative containing the predicted EBE$_{\text{Brg11}}$ box. Leaf discs were harvested 48 h post infiltration and stained to visualize GUS reporter activity. The predicted EBE$_{\text{Brg11}}$ box is displayed underneath the repeats of Brg11 in black letters. The nucleotide 5′ of the box is shown by a white letter with a black frame. Cartoon representations that display the domain architecture of the studied proteins are shown next to the corresponding leaf disc: the blue horizontal bar and the colored polygons represent dTALE$_{\text{EBE Brg11}}$ and its repeats. The yellow horizontal bar and the colored ellipses represent Brg11 and its repeats. Repeats framed with a bold outline indicate Brg11 repeats embedded into dTALE$_{\text{EBE Brg11}}$. A triangle (dTALE$_{\text{EBE Brg11}}$) or semicircle (Brg11) indicates a confirmed or putative transcriptional activation domain; rectangles indicate confirmed or putative nuclear localization signals. Colors of repeats (ellipses for Brg11 repeats or polygons for TALE repeats) correspond to different residues at position 13 of each repeat. Gray bars are used to demarcate repeats 1–5, 6–10 and 11–16.

and the matching EBE Brg11. This accords with a recent finding that a dTALE perfectly matching *UPT*$_{\text{AvrBs3}}$ showed less N$_0$ discrimination than the wild-type AvrBs3 (Meckler *et al.*, 2013).

Brg11 homologs isolated from strains infecting diverse hosts target identical or similar DNA sequences

TALEs from *Xanthomonas* strains infecting diverse hosts (e.g. pepper (*Capsicum* sp.) and rice (*Oryza sativa*)) show high conservation in N- and C-terminal nonrepeat regions and within the repeats at non-RVD positions but are highly diverse with respect to repeat number and RVD composition (Schornack *et al.*, 2006; Boch & Bonas, 2010), thus targeting distinct nucleotide sequences. We studied RipTALs from different *R. solanacearum* strains (Table S3) in order to learn more about the natural diversity of these effectors and how they compare to TALEs in this respect. The selected strains have been previously analyzed for possession of *TALE* homologs via PCR (Heuer *et al.*,

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

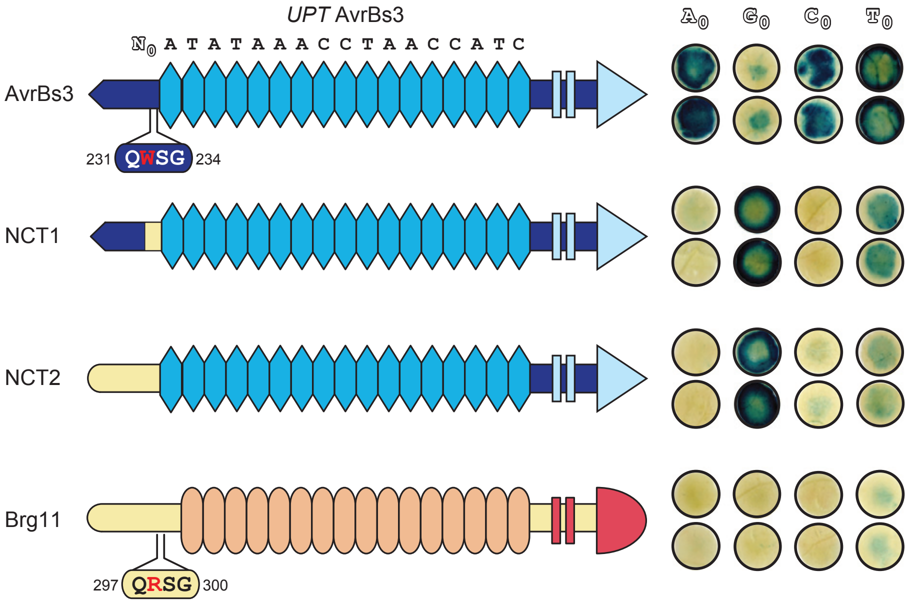


Fig. 6 Functional analysis of the N-terminal region of Brg11 (*hrpB*-regulated 11). N-terminal chimeras (NTCs) were created either bearing the full Brg11 N-terminal region (residues 1–354) in place of the corresponding AvrBs3 (avirulence protein triggering *Bs3* resistance) region (NTC2) or where only repeats 0 and –1 were replaced (NTC1). AvrBs3, Brg11, NTC1 and NTC2 were tested for their ability to activate either the *Bs3* promoter which contains a thymine at the 5′ terminus of the AvrBs3 target site (T_0) or derivatives that contain a guanine (G_0), cytosine (C_0) or adenine (A_0) at the corresponding position. These promoter-*uidA* fusions were co-transformed along with *35S*-promoter-driven genes encoding the proteins depicted on the left in *Nicotiana benthamiana* leaf tissue via *Agrobacterium tumefaciens*-mediated transient transformation. Leaf discs were harvested after 48 h and stained to visualize promoter activity. Amino acid sequences for the tryptophan (W) in repeat -1 that was shown to interact with T_0 (Mak *et al.*; 2012) are shown along with surrounding residues for AvrBs3 and the corresponding region for Brg11 inside ellipses, with positions inside the polypeptide indicated with numbers. Blue, AvrBs3-derived sequences; yellow, orange and red, Brg11-derived sequences.

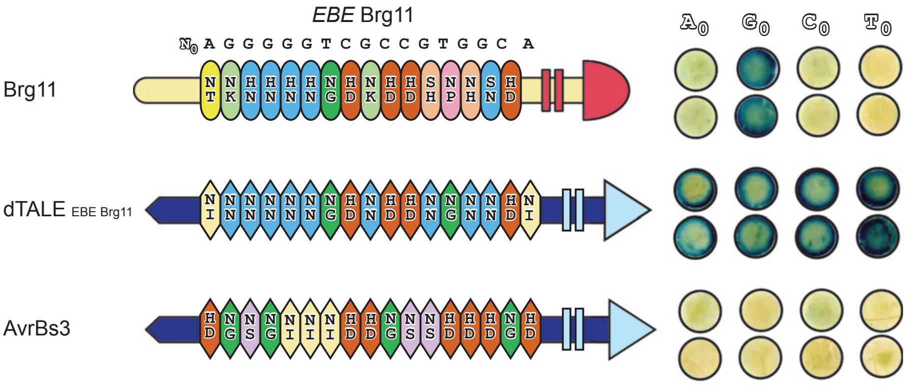

EBE Brg11
N0 A G G G G G T C G C C G T G G C A
A0
G0
C0
T0
Brg11
NT NK HN HN HN HN NG HD NK HD HD SH NP NH SN HD
dTALE EBE Brg11
NI NN NN NN NN NN NG HD NN HD HD NN NG NN NN HD NI
AvrBs3
HD NG NS NG NI NI NI HD HD NG NS NS HD HD HD NG HD

Fig. 7 Testing of Brg11 (*hrpB*-regulated 11) target boxes preceded by distinct 5′ bases. The AvrBs3 (avirulence protein triggering *Bs3* resistance) binding site in the *Bs3* promoter was replaced by the predicted Brg11 binding site (effector-binding element $(\mathrm{EBE})_{\mathrm{Brg11}}$) preceded by a 5′-terminal adenine (A_0), guanine (G_0), cytosine (C_0) or thymine (T_0). *35S*-promoter-driven genes encoding AvrBs3, Brg11 or designer transcription activator-like effector $(\mathrm{dTALE})_{\mathrm{EBE\ Brg11}}$ were each cotransformed along with the four *Bs3* promoter derivatives driving a *uidA* reporter into *Nicotiana benthamiana* leaves via *Agrobacterium tumefaciens*-mediated transient transformation. Leaf discs were harvested 48 h post infiltration and stained to visualize reporter activity. The sequence of $\mathrm{EBE}_{\mathrm{Brg11}}$ is displayed in black font above the repeats of Brg11. A white letter with a black frame shows the zero base in each promoter. Cartoon representations that display the domain architecture of the given protein are shown next to the corresponding leaf disc: blue horizontal bar and colored polygons, AvrBs3 and its repeats; yellow horizontal bar and the colored elipses, Brg11 and its repeats; triangle (AvrBs3) or semicircle (Brg11), a confirmed or putative transcriptional activation domain; rectangles, confirmed or putative nuclear localization signals. Repeats (ellipses for Brg11 repeats or polygons for TALE repeats) with the same repeat variable diresidues (RVDs) are displayed in the same color.

2007) and the amplicons subsequently analyzed via restriction-based fingerprinting. From that strain collection, representative members were selected covering 10 out of 13 *Alu*I restriction profiles (Table S3). All strains in this study are members of *R. solanacearum* phylotype I, and corresponding RipTALs were named RipTALI-X, where the Roman numeral ‘I’ indicates strain phylotype and X is an Arabic numeral denoting the given homolog (Figs 8, S9, S10, Table S3). After PCR amplification and sequencing of homologs, alignments were made of translation products (Fig. S10), and polymorphisms with a potential impact

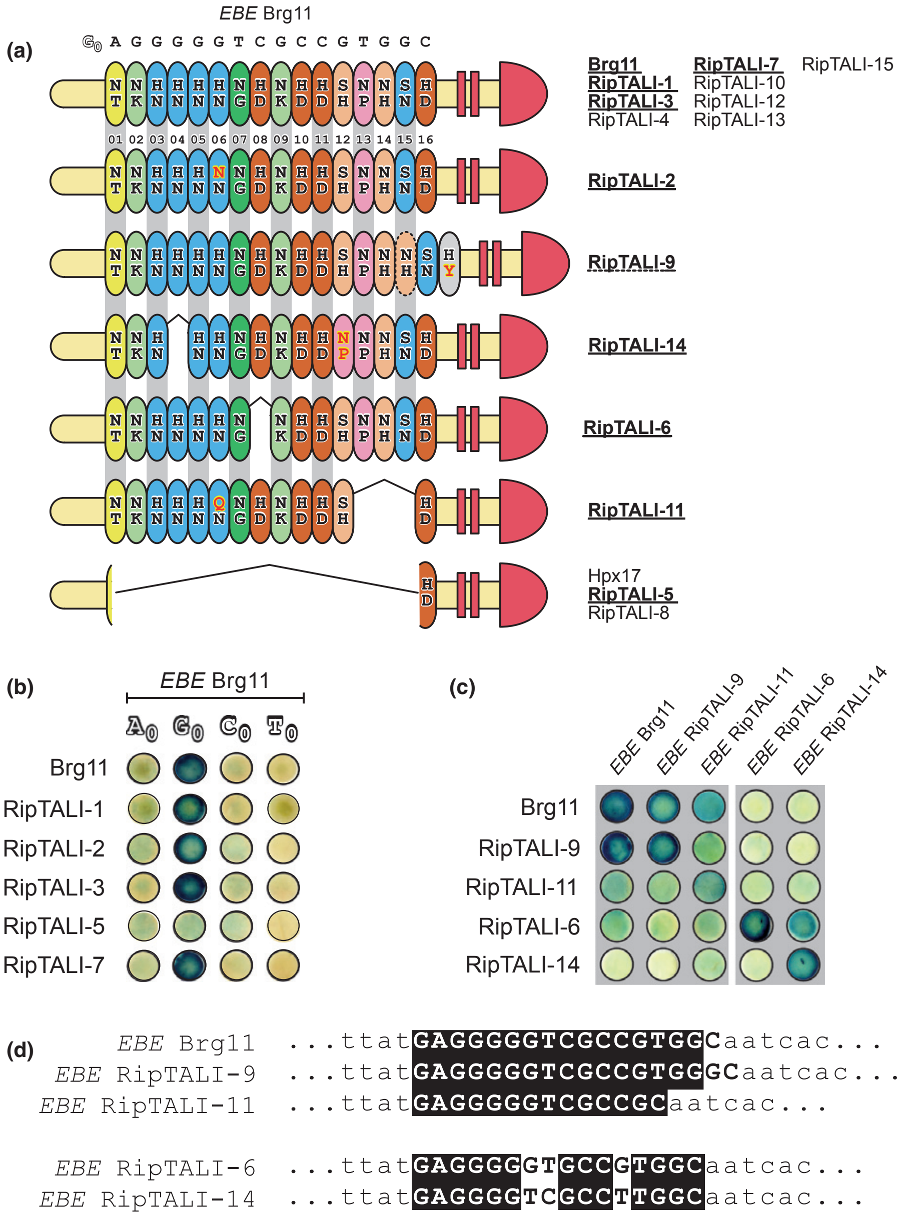


Fig. 8 Repeat variable diresidue (RVD) composition and target specificity of *Ralstonia* injected protein transcription activator-likes (RipTALs) isolated from distinct strains of *Ralstonia solanacearum*. (a) Cartoon representations (left) indicate the RVD composition deduced from PCR-amplified *RipTALs* (designation given on the right). Note that, where multiple RipTALs are listed next to one structure, these RipTALs have an identical composition of RVDs but differ in their amino acid sequence in other residues (see Figs S9, S10). Ellipses, the repeats of the core binding domain; semicircle, a putative transcriptional activation domain; rectangles, putative nuclear localization signals. Repeat positions are indicated below the core repeats of Brg11, (*hrpB*-regulated 11) with gray and white stripes showing the beginning and end of each repeat, respectively. Colors of repeats correspond to RVDs. RVD residues differing from their Brg11 equivalents are indicated with red letters with a yellow frame. Brackets indicate deletions with respect to Brg11. The ellipse that is framed with a dashed line indicates a repeat that probably occurred by duplication of an adjacent repeat. The predicted target sequence of Brg11 is displayed in black font above the repeats of Brg11. A white letter with a black frame shows the G_0 found to be preferentially activated by the Brg11 N-terminus. (b) Four RipTALs with a predicted binding element identical to that of Brg11 were tested alongside Brg11 for their ability to activate pepper *Bs3* promoter (*Bs3p*) derivatives bearing effector-binding element $(EBE)_{Brg11}$ with an A_0, G_0, C_0 or T_0 upstream of a *uidA* reporter gene. (c) Four RipTALs with a predicted binding element differing from that of Brg11 were tested for their ability to activate *Bs3p* derivatives bearing G_0 EBEs for Brg11 and each of the four other homologs. Two groups of EBEs for which the corresponding RipTALs seem to have cross-reactivity are indicated with gray boxes. *35S*-driven genes encoding the RipTALs were coinfiltrated along with the promoter constructs into *Nicotiana benthamiana* leaves via *Agrobacterium tumefaciens*-mediated transient transformation. Leaf discs were harvested 48 h post infiltration and stained to visualize reporter activity. Names of RipTALs functionally tested in (b) and (c) are underlined in (a). A sequence comparison of the EBEs used in (c) is shown in (d). In each case the EBE, including the G encoded by the zero position, is shown in bold font, uppercase letters. Lowercase letters indicate the *Bs3p* context in which the distinct EBEs were embedded. White lettering on a black background indicates identical bases between at least two EBEs.

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

on the predicted DNA target sites are shown in Fig. 8(a). Non-RVD polymorphisms facilitated correct alignment of RipTAL repeat domains even when mutations were discovered in the RVDs.

Similar to TALEs, an overall conservation of $> 97\%$ was found between different RipTAL proteins when repeat number polymorphisms were excluded (Fig. S10). Polymorphisms were distributed across N- and C-terminal and repeat regions but, in contrast to TALEs, the studied RipTALs showed a rather limited range of polymorphisms with respect to their RVD composition: eight of 15 newly isolated RipTALs showed not only an identical number of repeats, but also the same RVD composition as Brg11; two of 15 RipTALs showed the same number of repeats as Hpx17; four of 15 homologs showed a similar repeat domain to Brg11 but, relative to Brg11, seemed to have either lost repeats (RipTALI-14, deletion of repeat 4; RipTALI-6, deletion of repeat 8; RipTALI-11, deletion of repeats 13–15) or gained a repeat (RipTALI-9, repeat 14 possibly duplicated) (Fig. 8a). Given that the majority share their repeat architecture with Brg11, homologs with distinct architecture are possibly derivatives of a Brg11-like progenitor. We also observed that some RipTALs, compared with Brg11, contained substitutions in one or both RVD residues of certain repeats. For example, RipTALI-2 and -11 contain NN and QN RVDs, respectively, in their sixth repeat where Brg11 has an HN RVD (Figs 8a, S10). TALEs bind DNA via repeat residue 13 (Bochtler, 2012) and thus the observed RVD variations in residue 12 of repeat 6 probably have no significant impact on the base specificity of these repeats. By contrast, repeat 11 of RipTALI-14 and repeat 17 of RipTALI-9 both showed variation in repeat residue 13 (Fig. 8a), suggesting a change in base specificity for these repeats relative to their putative progenitor repeat in Brg11.

The lack of RVD polymorphisms between Brg11 and homologs with an identical number of repeats suggests that these RipTALs would bind to identical DNA targets. Indeed, when a subset of these were tested in GUS reporter assays, RipTALI-1, -3 and -7, differing at many non-RVD residues but identical in repeat number and RVD composition (Figs 8a, S10), were all found to be able to activate *Bs3p*$_{\mathrm{EBE\ Brg11}}$ (Fig. 8b). Furthermore, like Brg11, they were only able to activate the *Bs3p*$_{\mathrm{EBE\ Brg11}}$ with a G_0 (Fig. 8b). Additionally, RipTALI-2, which contains an RVD polymorphism in residue 12 of the sixth repeat, also activated *Bs3p*$_{\mathrm{EBE\ Brg11}}$, as anticipated (Fig. 8b). While RipTALI-5 and -8, like Hpx17, would be predicted to bind any guanine-cytosine dinucleotide, of which there are two within EBE$_{\mathrm{Brg11}}$, TALEs bearing only one to three repeats are unable to activate target promoters (Boch *et al.*, 2009). Consistent with this, RipTALI-5 was unable to activate any of the *Bs3p*$_{\mathrm{EBE\ Brg11}}$ reporters (Fig. 8b). RipTALs I-6, -9, -11 and -14 have predicted target boxes differing from that of Brg11 (Fig. 8d) and were tested against *Bs3p* derivatives bearing corresponding G_0 EBEs (*Bs3p* $_{\mathrm{EBE\ RipTALI\text{-}X}}$; Figs 8c, d). Each promoter was activated to the highest level by its corresponding RipTAL, although the strength of staining indicated differences in activity even between the optimal RipTAL-promoter pairs (Fig. 8c). We also observed some cross-reactivity where EBEs had high sequence similarity. EBE$_{\mathrm{RipTALI\text{-}14}}$ differs in only three positions from EBE$_{\mathrm{RipTALI\text{-}6}}$ and was activated by both RipTALI-6 and -14, although to differing degrees. Surprisingly, the converse was not observed (Fig. 8c). Brg11 and RipTALI-9 were able to activate promoters bearing each other's EBEs to similar levels, which is unsurprising as these target sites differ from one another by only one nucleotide in each case (Fig. 8c,d). Both RipTALs were also able to weakly activate EBE$_{\mathrm{RipTALI\text{-}11}}$, with which each has three mismatches in the 3′ end, and RipTALI-11 likewise showed some cross-reactivity with EBE$_{\mathrm{Brg11}}$ and EBE$_{\mathrm{RipTALI\text{-}9}}$ (Figs 8c, d).

RipTALs I-6, -9, -11 and -14 were also tested for their N_0 preferences (Fig. S11), and in each case the G_0 reporter was the most strongly activated, although in the case of RipTALI-6 and -9 there was some activation of the A_0 reporter, indicating a relaxed specificity. This difference in base specificity, however, could not be correlated to any specific amino acid polymorphisms shared between these homologs.

Discussion

We have been able to demonstrate an overall functional similarity between TALEs and RipTALs. RipTALs Brg11 and Hpx17, like TALEs, localize to the nucleus (Fig. 1) and bear a functional C-terminal AD (Fig. 2). The repeats of Brg11 were shown to mediate one-to-one base-specific DNA recognition with a code determined by RVDs (Fig. 3) very similar to that of *Xanthomonas* TALEs. Indeed, the Brg11 target sequence used in this study is among those previously predicted from analysis of Brg11 RVDs inside a TALE repeat scaffold (Fig. S12; Streubel *et al.*, 2012). This suggests that the scaffold formed by the tandem arranged repeats is similar in TALE and RipTAL proteins. As Brg11 has been shown to promote virulence in the interaction with eggplant (*Solanum melongena*; Macho *et al.*, 2010), it is conceivable that Brg11 contributes to disease similarly to *Xanthomonas* TALEs by transcriptional activation of host susceptibility genes. The now known DNA base preferences of Brg11 and other RipTALs will simplify the identification of potential susceptibility genes in the host.

The functional characterization of RipTALs provides useful information for the optimization of existing TALE-based DNA-binding domains. While RipTAL RVDs have already been used within the context of dTALE repeat arrays (Streubel *et al.*, 2012) and arrays of identical repeats based on one consensus RipTAL repeat (Li *et al.*, 2013), the functional impact of non-RVD residues in combination with specific RVDs remains unexploited. Non-RVD polymorphisms are rare among TALEs (Boch & Bonas, 2010) but abundant in RipTAL repeats (Fig. S4), representing a valuable resource for the optimization of TALE-based DNA-binding modules.

In the case of repeats 10 and 8/11, we were able to demonstrate the functional impact of a specific non-RVD polymorphism (Fig. 4), but further inferences can be made based on data from the trimer test (Fig. 3, Table S2). For example, we found that the RVD NK in Brg11 repeats 2 and 9 gave 15–18 times stronger reporter activation than was the case for NK in a TALE repeat (Fig. 3, Table S2). The observed functional differences are

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

probably attributable to the nine residues common to Brg11 repeats 2 and 9, and not found in TALE repeats (Fig. S13). Further examples are repeats 3–6 of Brg11, all possessing HN RVDs and differing from each other at only four non-RVD positions (Fig. S4). Pronounced differences in both the level of activity (Fig. 3, Table S2) and specificity were detected for derivatives bearing trimers of these repeats. These data support the conclusion that non-RVD polymorphisms can have profound effects on both repeat activity and specificity, and that further analysis of RipTALs from across the *R. solanacearum* species complex may reveal further examples of use to those working with TALE-based DNA-binding domains.

The successful engineering of a TALE derivative that specifically activates a reporter with a target box preceded by a G nucleotide (NTC1 and NTC2; Fig. 6) is also of relevance to those working with TALE-based custom DNA-binding modules. All known TALEs show a preference for a T_0, while sequences with G_0 are generally poorly activated (e.g. AvrBs3; Fig. 6). dTALEs bearing Brg11 N-terminal regions may be of use to those wishing to create DNA-binding domains for GC-rich regions where the requirement for a T_0 can be constraining. Moreover, the high base specificity for targets with a G_0, which we observed for the RipTAL N-terminal DNA-binding domain (Fig. 6), may allow for the construction of custom TALEs with higher specificity than previously attainable.

The sequencing of *brg11* homologs from phylotype I *R. solanacearum* strains, in combination with functional data, provides insights into the evolution of the encoded RipTALs. *Xanthomonas* TALEs are highly flexible with respect to repeat number and RVD composition, and even closely related strains target identical host genes via repeat arrays with distinct binding specificities. One such example is the rice gene *Os11N3*, which is transcriptionally activated by the distinct *Xanthomonas oryzae* pv. *oryzae* TALEs PthXo3 (pathogenicity *Xanthomonas oryzae 3*), AvrXa7 (avirulence protein triggering *Xa7* resistance) (strains PXO61 and PXO86, respectively; Antony *et al.*, 2010) and TalC (strain BAI3; Yu *et al.*, 2011). Notably, each TALE targeting the *Os11N3* promoter has a unique binding site, suggesting convergent evolution. In marked contrast, the majority of RipTALs in this study, isolated from geographically diverse locations and phylogenetically diverse hosts (Table S3), have the same predicted DNA target site (Fig 8). This suggests that RipTALs, in contrast to TALEs, target sequences that are conserved even in phylogenetically diverse host genomes. Alternatively, RipTALs may be functional in only a few host species that have matching RipTAL-binding sites.

Our study also provides a first glimpse into the mechanisms that drive evolution of RipTALs. The few differences in RipTAL RVD composition identified in this study are a result only of loss, duplication or mutations of individual repeats. By contrast, genes encoding TALE repeat arrays are highly variable in their RVD composition and seem to evolve diversity by inter- and intragenic recombination (Yang *et al.*, 2005), allowing for major rearrangements of the repeat array. Thus, the mechanistic principles in TALE and RipTAL evolution might be fundamentally different.

Insights into the molecular mechanisms of RipTAL activity could also be of value for plant breeding purposes. *Ralstonia solanacearum* is a devastating pathogen and the few well-defined instances of genetic resistance to this disease are limited to certain bacterial strains or host plants, such as the effector PopP2 (*Pseudomonas* outer protein P2), eliciting a resistance response in *A. thaliana* but with restricted distribution among *R. solanacearum* strains (Genin, 2010). By contrast, a number of plant *R* genes that mediate resistance to *Xanthomonas* via recognition of TALEs have been isolated, with transcriptional activation of *R* gene promoters being the most common mechanism (Bogdanove *et al.*, 2010; Strauß *et al.*, 2012). Knowledge of RipTAL-binding preferences can assist in the identification or creation of transcriptionally activated *R* genes conferring resistance to *R. solanacearum.* Plant germplasm collections can be screened for resistance responses induced by RipTALs. The predicted binding sequence can then be used in combination with transcriptomics to rapidly isolate the *R* gene responsible, as recently described for the identification of an *R* gene from pepper which is activated by and mediates recognition of the *Xcv* TAL effector AvrBs4 (Strauß *et al.*, 2012). Alternatively, RipTAL-binding sites could be inserted upstream of known executor type *R* genes (Bogdanove *et al.*, 2010) that provide a defense response upon transcriptional activation. We also envision that multiple RipTAL-binding sites could be inserted into one promoter that would be activated by and confer resistance to strains bearing RipTAL homologs with distinct target sequences (Morbitzer *et al.*, 2010; Hummel *et al.*, 2012). In summary, the elucidation of the RipTAL code provides new possibilities for classical as well as molecular breeding of wilt-resistant crop plants.

Acknowledgements

We thank D. Horvath for helpful comments on the manuscript. We would also like to thank J. Elsaesser, S. Leisz, R. Morbitzer, S. Solle and T. Strauß for technical support. O.d.L. and N.S. received funding through grants from the Deutsche Forschungsgemeinschaft to T.L. (SFB924). A.S. received funding through grants from the Two Blades Foundation. Work on RipTALs has also been supported by grants from the Exzellenznetzwerk Biowissenschaften (Ministry of Culture of Saxonia-Anhalt).

References

Antony G, Zhou J, Huang S, Li T, Liu B, White F, Yang B. 2010. Rice *xa13* recessive resistance to bacterial blight is defeated by induction of the disease susceptibility gene *Os-11N3*. *Plant Cell* **22**: 3864–3876.

Boch J, Bonas U. 2010. *Xanthomonas* AvrBs3 family-type III effectors: discovery and function. *Annual Review of Phytopathology* **48**: 419–436.

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

Bochtler M. 2012. Structural basis of the TAL effector-DNA interaction. *Biological Chemistry* **393**: 1055–1066.

Bogdanove AJ, Schornack S, Lahaye T. 2010. TAL effectors: finding plant genes for disease and defense. *Current Opinion in Plant Biology* **13**: 394–401.

Bogdanove AJ, Voytas DF. 2011. TAL effectors: customizable proteins for DNA targeting. *Science* **333**: 1843–1846.

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

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

Cunnac S, Occhialini A, Barberis P, Boucher C, Genin S. 2004. Inventory and functional analysis of the large Hrp regulon in *Ralstonia solanacearum*: identification of novel effector proteins translocated to plant host cells through the type III secretion system. *Molecular Microbiology* **53**: 115–128.

Deng D, Yan C, Pan X, Mahfouz M, Wang J, Zhu J, Shi Y, Yan N. 2012. Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **335**: 720–723.

Elphinstone J. 2005. The current bacterial wilt situation: a global overview. In: Allen C, Prior P, Hayward AC, eds. *Bacterial wilt disease and the Ralstonia solanacearum species complex*. St Paul, MN, USA: APS Press, 9–28.

Genin S. 2010. Molecular traits controlling host range and adaptation to plants in *Ralstonia solanacearum*. *New Phytologist* **187**: 920–928.

Genin S, Boucher C. 2004. Lessons learned from the genome analysis of *Ralstonia solanacearum*. *Annual Review of Phytopathology* **42**: 107–134.

Grover A, Chakrabarti SK, Azmi W, Khurana SMP. 2012. Rapid method for isolation of PCR amplifiable genomic DNA of *Ralstonia solanacearum* infested in potato tubers. *Advances in Microbiology* **2**: 441–446.

Heuer H, Yin Y-N, Xue Q-Y, Smalla K, Guo J-H. 2007. Repeat domain diversity of *avrBs3*-like genes in *Ralstonia solanacearum* strains and association with host preferences in the field. *Applied and Environmental Microbiology* **73**: 4379–4384.

Hummel AW, Doyle EL, Bogdanove AJ. 2012. 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. *New Phytologist* **195**: 883–893.

Jones JB, Lacy GH, Bouzar H, Stall RE, Schaad NW. 2004. Reclassification of the xanthomonads associated with bacterial spot disease of tomato and pepper. *Systematic and Applied Microbiology* **27**: 755–762.

Kay S, Hahn S, Marois E, Hause G, Bonas U. 2007. A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science* **318**: 648–651.

Koncz C, Schell J. 1986. The promoter of T_L-DNA gene 5 controls the tissue-specific expression of chimaeric genes carried by a novel type of *Agrobacterium* binary vector. *Molecular and General Genetics* **204**: 383–396.

Lambert C. 2002. Agricultural bioterrorism protection act of 2002: possession, use, and transfer of biological agents and toxins; interim and final rule. *Federal Register 2002* **67**: 76907–76938.

Li L, Atef A, Piatek A, Ali Z, Piatek M, Aouida M, Sharakuu A, Mahjoub A, Wang G, Khan S *et al.* 2013. Characterization and DNA-binding specificities of *Ralstonia* TAL-like effectors. *Molecular Plant.* doi: 10.1093/mp/sst006.

Macho AP, Guidot A, Barberis P, Beuzón CR, Genin S. 2010. *Ralstonia solanacearum* type III effector mutant strains with reduced fitness in host plants. *Molecular Plant-Microbe Interactions* **23**: 1197–1205.

Mak A, Bradley P, Cernadas R, Bogdanove AJ, Stoddard BL. 2012. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335**: 716–719.

Meckler JF, Bhakta MS, Kim M-S, Ovadia R, Habrian CH, Zykovich A, Yu A, Lockwood SH, Morbitzer R, Elsäesser J *et al.* 2013. Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Research.* **41**: 4118–4128.

Morbitzer R, Elsaesser J, Hausner J, Lahaye T. 2011. Assembly of custom TALE-type DNA binding domains by modular cloning. *Nucleic Acids Research* **39**: 5790–5799.

Morbitzer R, Römer P, Boch J, Lahaye T. 2010. Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors. *Proceedings of the National Academy of Sciences, USA* **107**: 21617–21622.

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

Mukaihara T, Tamura N, Iwabuchi M. 2010. Genome-wide identification of a large repertoire of *Ralstonia solanacearum* type III effector proteins by a new functional screen. *Molecular Plant-Microbe Interactions* **23**: 251–262.

Nakagawa T, Kurose T, Hino T, Tanaka K, Kawamukai M, Niwa Y, Toyooka K, Matsuoka K, Jinbo T, Kimura T. 2007. Development of series of gateway binary vectors, pGWBs, for realizing efficient construction of fusion genes for plant transformation. *Journal of Bioscience and Bioengineering* **104**: 34–41.

Nakamura S, Mano S, Tanaka Y, Ohnishi M, Nakamori C, Araki M, Niwa T, Nishimura M, Kaminaka H, Nakagawa T *et al.* 2010. Gateway binary vectors with the bialaphos resistance gene, *bar*, as a selection marker for plant transformation. *Bioscience, Biotechnology, and Biochemistry* **74**: 1315–1319.

Nguyen Ba AN, Pogoutse A, Provart N, Moses AM. 2009. NLStradamus: a simple hidden Markov model for nuclear localization signal prediction. *BMC Bioinformatics* **10**: 202.

Piskacek S, Gregor M, Nemethova M, Grabner M, Kovarik P, Piskacek M. 2007. Nine-amino-acid transactivation domain: establishment and prediction utilities. *Genomics* **89**: 756–768.

Ralser M, Querfurth R, Warnatz H-J, Lehrach H, Yaspo M-L, Krobitsch S. 2006. An efficient and economic enhancer mix for PCR. *Biochemical and Biophysical Research Communications* **347**: 747–751.

Remenant B, Coupat-goutaland B, Guidot A, Cellier G, Wicker E, Allen C, Fegan M, Pruvost O, Elbaz M, Calteau A *et al.* 2010. Genomes of three tomato pathogens within the *Ralstonia solanacearum* species complex reveal significant evolutionary divergence. *BMC Genomics* **11**: 379.

Römer P, Hahn S, Jordan T, Strauß T, Bonas U, Lahaye T. 2007. Plant pathogen recognition mediated by promoter activation of the pepper *Bs3* resistance gene. *Science* **318**: 645–648.

Römer P, Strauss T, Hahn S, Scholze H, Morbitzer R, Grau J, Bonas U, Lahaye T. 2009. Recognition of AvrBs3-like proteins is mediated by specific binding to promoters of matching pepper *Bs3* alleles. *Plant Physiology* **150**: 1697–1712.

Salanoubat M, Genin S, Artiguenave F, Gouzy J, Mangenot S, Arlat M, Billault A, Brottier P, Camus JC, Cattolico L *et al.* 2002. Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* **415**: 497–502.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**: 671–675.

Schornack S, Meyer A, Römer P, Jordan T, Lahaye T. 2006. Gene-for-gene-mediated recognition of nuclear-targeted AvrBs3-like bacterial effector proteins. *Journal of Plant Physiology* **163**: 256–272.

Strauß T, Poecke RV, Strauß A, Römer P, Minsavage GV, Sing S, Wolf C, Strauß A, Kim S, Lee HA *et al.* 2012. RNA-seq pinpoints a *Xanthomonas* TAL-effector activated resistance gene in a large-crop genome. *Proceedings of the National Academy of Sciences, USA* **109**: 19480–19485.

Streubel J, Blücher C, Landgraf A, Boch J. 2012. TAL effector RVD specificities and efficiencies. *Nature Biotechnology* **30**: 593–595.

Szurek B, Marois E, Bonas U, Van den Ackerveken G. 2001. Eukaryotic features of the *Xanthomonas* type III effector AvrBs3: protein domains involved in transcriptional activation and the interaction with nuclear import receptors from pepper. *Plant Journal* **26**: 523–534.

Van den Ackerveken G, Marois E, Bonas U. 1996. Recognition of the bacterial avirulence protein AvrBs3 occurs inside the host plant cell. *Cell* **87**: 1307–1316.

Vauterin L, Hoste B, Kersters K, Swings J. 1995. Reclassification of *Xanthomonas*. *International Journal of Systematic Bacteriology* **45**: 472–489.

Voinnet O, Rivas S, Mestre P, Baulcombe D. 2003. An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus. *The Plant Journal* **33**: 949–956.

Yabuuchi E, Kosako Y, Yano I, Hotta H, Nishiuchi Y. 1995. Transfer of two *Burkholderia* and an *Alcaligenes* species to *Ralstonia* gen. nov. proposal of *Ralstonia pickettii* (Ralston, Palleroni and Doudoroff 1973) comb. nov., *Ralstonia solanacearum* (Smith 1896) comb. nov. and *Ralstonia eutropha* (Davis 1969) comb. nov. *Microbiology and Immunology* **39**: 897–904.

Yang B, Sugio A, White F. 2005. Avoidance of host recognition by alterations in the repetitive and C-terminal regions of AvrXa7, a type III effector of *Xanthomonas oryzae* pv. *oryzae*. *Molecular Plant-Microbe Interactions* **18**: 142–149.

Yu Y, Streubel J, Balzergue S, Champion A, Boch J, Koebnik R, Feng J, Verdier V, Szurek B. 2011. Colonization of rice leaf blades by an african strain of *Xanthomonas oryzae pv. oryzae* depends on a new TAL effector that induces the

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

rice nodulin-3 *Os11N3* gene. *Molecular Plant-Microbe Interactions* **9**: 1102–1113.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Annotated sequences of Brg11 and AvrBs3.

Fig. S2 Amino acid sequences of Hpx17 and corresponding truncation derivatives used in subcellular localization studies.

Fig. S3 Alignment of C-terminal regions of AvrBs3 and Brg11 (a) and the sequences of the C-terminal chimeras CTC and CTC-ΔAD (b).

Fig. S4 Alignments of Brg11 and AvrBs3 core repeats.

Fig. S5 Representative amino acid sequences used in the trimer test.

Fig. S6 Reciprocal exchange of RVDs between Brg11 repeats 12 and 8/11 leading to exchange of DNA recognition specificity.

Fig. S7 Nucleotide sequences of *brg11* repeats used to assemble three identical, tandem-arranged repeat blocks that were tested in the context of $\mathrm{dTALE}_{\mathrm{EBE\ Brg11}}$.

Fig. S8 Alignment of the N-terminal regions of AvrBs3 and Brg11 (a) and sequence of chimeras NTC1 and NTC2 (b).

Fig. S9 Individual amino acid sequences of RipTALs analyzed in this study.

Fig. S10 Amino acid sequence alignment of RipTALs analyzed in this study.

Fig. S11 Zero base preferences of RipTALI-6, -9, -11 and -14.

Fig. S12 Comparison of predicted Brg11 binding sequences from this study and from Streubel *et al.* (2012).

Fig. S13 Alignment of Brg11 repeats 2 and 9 and a consensus AvrBs3 repeat.

Table S1 Sequences of primers used in this study

Table S2 Origin of RipTALs analyzed in this study

Table S3 GUS activities determined in the trimer test

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

About *New Phytologist*

- *New Phytologist* is an electronic (online-only) journal owned by the New Phytologist Trust, a **not-for-profit organization** dedicated to the promotion of plant science, facilitating projects from symposia to free access for our Tansley reviews.
- Regular papers, Letters, Research reviews, Rapid reports and both Modelling/Theory and Methods papers are encouraged. We are committed to rapid processing, from online submission through to publication 'as ready' via *Early View* – our average time to decision is <25 days. There are **no page or colour charges** and a PDF version will be provided for each article.
- The journal is available online at Wiley Online Library. Visit **www.newphytologist.com** to search the articles and register for table of contents email alerts.
- If you have any questions, do get in touch with Central Office (np-centraloffice@lancaster.ac.uk) or, if it is more convenient, our USA Office (np-usaoffice@ornl.gov)
- For submission instructions, subscription and all the latest information visit **www.newphytologist.com**

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

4 Programmable DNA-binding proteins from *Burkholderia* provide a fresh perspective on the TALE-like repeat domain

This chapter is identical to the publication:

de Lange, Orlando, Wolf, Christina, Dietze, Jörn, Elsaesser, Janett, Morbitzer, Robert & Lahaye, Thomas

Programmable DNA-binding proteins from *Burkholderia* provide a fresh perspective on the TALE-like repeat domain.

Nucleic Acids Research (2014) **42** (11): 7436-7449.

Nucleic Acids Research Advance Access published May 3, 2014

doi: 10.1093/nar/gku329

Programmable DNA-binding proteins from *Burkholderia* provide a fresh perspective on the TALE-like repeat domain

Orlando de Lange†, Christina Wolf†, Jörn Dietze, Janett Elsaesser, Robert Morbitzer and Thomas Lahaye*

Genetics, Department of Biology I, Ludwig-Maximilians-University Munich, Martinsried, Bavaria, 82152, Germany

Received December 13, 2013; Revised April 04, 2014; Accepted April 4, 2014

ABSTRACT

The tandem repeats of transcription activator like effectors (TALEs) mediate sequence-specific DNA binding using a simple code. Naturally, TALEs are injected by *Xanthomonas* bacteria into plant cells to manipulate the host transcriptome. In the laboratory TALE DNA binding domains are reprogrammed and used to target a fused functional domain to a genomic locus of choice. Research into the natural diversity of TALE-like proteins may provide resources for the further improvement of current TALE technology. Here we describe TALE-like proteins from the endosymbiotic bacterium *Burkholderia rhizoxinica*, termed Bat proteins. Bat repeat domains mediate sequence-specific DNA binding with the same code as TALEs, despite less than 40% sequence identity. We show that Bat proteins can be adapted for use as transcription factors and nucleases and that sequence preferences can be reprogrammed. Unlike TALEs, the core repeats of each Bat protein are highly polymorphic. This feature allowed us to explore alternative strategies for the design of custom Bat repeat arrays, providing novel insights into the functional relevance of non-RVD residues. The Bat proteins offer fertile grounds for research into the creation of improved programmable DNA-binding proteins and comparative insights into TALE-like evolution.

INTRODUCTION

When the DNA binding code of transcription activator like effectors (TALEs) was published in 2009 (1,2), a doorway was opened for researchers to build custom DNA-binding proteins. In nature, TALE proteins are injected by members of the plant pathogenic bacterial genus *Xanthomonas* into host cells. They act as eukaryotic transcription factors, inducing expression of targeted host genes that promote bacterial disease. This relies on a set of functional domains within the protein (3). Upon injection into host cells, nuclear localisation signals (NLSs) target TALEs to the plant nucleus. There the central domain of the protein, composed of tandem-arranged repeats, mediates sequence-specific binding to the promoters of target genes. A C-terminal transcriptional activation domain (AD) mediates promoter activation. The unique repeat array, mediating interaction of TALEs with DNA, has received great attention in the past years. Functional arrays are typically composed of 10–30 repeats, each 33–35 amino acids in length (3). Within repeats, variation is almost exclusively limited to positions 12 and 13, termed the repeat variable di-residue (RVD; 2). One repeat binds one base with specificity determined by the RVD. The TALE code refers to this 1-to-1 correlation and the base preferences defined by the distinct RVDs, providing a simple guide for users. By modifying repeat number and RVD composition users can design custom TALE repeat arrays that target nucleotide sequences of desired length and base composition.

Since the inter-repeat polymorphisms of TALE repeat arrays are almost solely limited to the RVDs, reprogramming of base specificity is straightforward. As a consequence of the almost identical amino acid composition, each TALE repeat forms a near identical structure irrespective of its position in the array (4,5). Accordingly, each repeat is competent to make almost exactly the same inter-repeat interactions regardless of the residues occupying the RVD positions (4,5). Thus, changes to repeat number or position do not perturb the network of inter-repeat interactions that stabilize the superhelical structure formed by tandem-arranged repeats. This allows each repeat to be treated as a

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

---

[*] To whom correspondence should be addressed. Tel: +49 7071 29 7 8745; Fax: +49 7071 29 50 42; Email: thomas.lahaye@zmbp.uni-tuebingen.de

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Present address:
Orlando de Lange, Christina Wolf, Robert Morbitzer and Thomas Lahaye, Department of General Genetics, Centre for Plant Molecular Biology, University of Tuebingen, Auf der Morgenstelle 32, Tuebingen, Baden-Wuerttemberg, 72076, Germany.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

functionally independent module and isolates the RVD as the only position within the repeat of interest to the user.

Functional domains of choice can be fused to the TALE DNA binding domain and targeted to a predefined DNA sequence. By now TALE-activators, repressors and nucleases have been used extensively (6) and more recently TALE fusions mediating targeted epigenetic modifications have also been described (7–9).

Work on the TALE-like proteins of *Ralstonia solanacearum,* termed RipTALs, has revealed that they too act as eukaryotic transcription factors and that RipTAL target specificity is linked to RVDs as in TALEs (10). Comparative analysis of TALE and RipTAL repeat arrays also revealed functional differences, due to non-RVD polymorphisms, which could be used to improve custom TALE repeat arrays. Considering the ever-increasing use of TALEs across fundamental and applied biology, it seems sensible to further explore the natural diversity of this protein class in order to identify new functional features of benefit to users.

Burkholderia rhizoxinica is an obligate endosymbiotic bacterium of the fungal plant pathogen *Rhizopus microsporus* (11). The genome of *B. rhizoxinica* strain HKI 0454 has been sequenced (12) and among the predicted proteins are three with similarity to TALEs that we have termed Bat (*Burkholderia* TALE-like) proteins. The gene encoding the predicted Bat1 protein (Uniprot E5AV36, GenBank RBRH_01844) is located on megaplasmid pBRH01 while the predicted Bat2 (Uniprot E5AW45, GenBank RBRH_01776) and Bat3 proteins (Uniprot E5AW43, GenBank RBRH_01777) are encoded on neighbouring, non-overlapping open reading frames within plasmid pBRH02. Evidence for DNA binding activity and use as a programmable DNA binding domain has been demonstrated recently for Bat1 (alternatively designated BurrH; 13,14). We investigated DNA binding properties of the three Bat proteins, showing that Bat2 as well as Bat1 binds DNA with the same code as TALEs. We quantified the interaction of Bat1 with its predicted target DNA bearing the four possible zero bases and found that, unlike TALEs and RipTALs, Bat1 has no sequence preference at this position. Bat proteins share limited sequence identity with TALEs and also show greater inter-repeat diversity than TALEs or the recently described RipTALs. However, alignments between repeats of these different proteins reveal a core set of conserved residues that might be of use to identify further members of this class. We show that the Bat proteins can be used as modular DNA binding domains to mediate targeted transcriptional activation or site-directed DNA cleavage. However, in contrast to TALEs, no two repeats of any Bat proteins are identical, with inter-repeat similarity dropping below 50% in some cases. Because of this alternative approaches are possible for the customisation of the DNA binding repeats. We explored two options: exchanging whole repeats along with their RVDs or exchanging RVDs only. We found that while one strategy seems preferable, both are viable. In the process we gained evidence to suggest that polymorphisms at non-RVD positions affect binding domain function. Our observations suggest that the Bat proteins may offer a more compact alternative to the TALE platform for programmable DNA binding.

MATERIALS AND METHODS

Assembly of Bat1 and TALE expression constructs

Genes encoding the three Bat proteins were synthesized with *Escherichia coli* codon usage (GenScript) in separate BsaI-site flanked subunits (Supplementary Figure S4). For *E. coli* protein production, these modules were assembled via BsaI cut-ligation into a pENTR/D-TOPO (Life Technologies) derivative bearing BsaI sites (overlaps *CACC-AAGG*) within the LR recombination sites, created using primers listed in Supplementary Table S2. The genes were then transferred into pDEST-17 (Life Technologies). For human cell transfection and *in vitro* cleavage assays *bat* encoding modules were assembled along with BsaI-site-flanked modules encoding HA-NLS and NLS-3xFLAG-VP64 AD domains (acBat1, human cell reporter) or 3xHA/HA-NLS and HA-FokI (*in vitro* cleavage). Sequences are in Supplementary Figure S5. These were assembled into a modified pVAX vector (Life Technologies) with combined Cytomegalovirus (CMV)/Sp6 promoter and BsaI sites (*AATG-GCTT*), details and sequences for mutational primers given in Supplementary Table S2.

The acBat1 truncation derivatives tested in Figure 5 were carried out using polymerase chain reaction (PCR) on the individual synthesized blocks of Bat1 prior to assembly, using the primers listed in Supplementary Table S2. To create the acBat1 derivatives tested in Figure 6 modified assembly blocks were synthesized with the same codon usage as wild-type *acbat1* (GenScript). To create the *pSOX2* targeted dBats tested in Figure 7, a DNA fragment encoding the N- and C-terminal non-core-repeat sections of Bat1 was synthesized with the same codon usage as wild-type *acbat1* (GenScript; Supplementary Figure S11) and assembled into the pVAX vector along with HA-NLS, NLS-3xFLAG-VP64 constructs. The repeats were ordered as two blocks for each dBat (Supplementary Figure S11) and added into the expression vector between N- and C-terminally encoding regions via BpiI cut-ligation.

The repeat domains of dTALE$_{\mathrm{Bat1mimic}}$ and dTALE$_{\mathrm{SOX2}}$ were created using a previously described method (15). The assembly of dTALE$_{\mathrm{Bat1mimic}}$ required modifications to the toolkit. These included a novel level 2 vector, pUC57-CD-DEST, to allow assembly of more than 17 core repeats. This was created using PCR mutagenesis of pUC57 to insert the BsaI sites using primers listed in Supplementary Table S2. Repeats 4_NT, 5B_NN, 4_ND, 7C_NT, 1C_NR, 3_ND, 7D_NS and D $\frac{1}{2}$ N* were created via PCR mutagenesis on described repeat modules (15) or amplification from the repeats of *avrbs3* using the primers listed in Supplementary Table S2.

dTALE$_{\mathrm{UPT\ AvrBs3\ 3x\ Bat1\ rep2/6/8\ /17}}$ were created as previously described (10) with trimers synthesized by GenScript with the sequences listed in Supplementary Figure S14, while dTALE$_{\mathrm{UPT\ AvrBs3\ 3x\ NI/NN/NG}}$ were created with the aforementioned *TALE* assembly toolkit (15). Repeat domains were assembled into pENTR-D-TALE Δrep *Bpi*I-AC (15) and then *dTALEs* transferred into T-DNA binary vector pGWB641 (16) via LR recombination (Life Technologies).

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

Protein purification

Genes encoding the three N-terminally His tagged Bat proteins (Supplementary Figures S4 and S5) were expressed in *E. coli* Rosetta (DE31) pLaqI (Novagen) as previously described (17). In short, cells were induced at 30°C with IPTG for 3 h. After purification from cell lysate via TALON resin (Clontech), proteins were dialysed against storage buffer (480 mM KCl, 1.6 mM EDTA, 1 mM DTT, 12 mM Tris-Cl, pH 7.5; Slide-A-Lyzer, Thermo Scientific) and concentrated (Amicon Ultra, Millipore).

Electrophoretic mobility shift assay

Equal amounts of 100 μM 5’ Cy5 labelled forward strand and unlabelled reverse strand oligonucleotides (Metabion) were mixed 1:4 with annealing buffer (TALE storage buffer without DTT or Sodium Azide). After heating to 100°C for 10 min the mixture was allowed to cool to room temperature, then diluted 1/20 in annealing buffer. 2 μl of 1 μM Bat protein was mixed with 16 μl electrophoretic mobility shift assay (EMSA) buffer (15 mM Tris-Cl, 75 mM KCl, 2.5 mM DTT, 0.063% NP-40, 62.5 ng/μl dI.dC, 0.125 mg/ml BSA, 6.25% glycerol, 6.25 mM MgCl, 0.125 mM EDTA) and incubated 5 min at room temperature. 2 μl of target DNA were added followed by a further 30 min incubation. Total binding reactions were run on a 6% native polyacrylamide TBE-gel for 1 h at 100V, 4°C. Cy5 labelled DNA was visualized with the FMBIO III Multi View (Hitachi).

Microscale thermophoresis

Binding affinity was measured using the Monolith NT.115 from Nanotemper Technologies. Bat1 was labelled with the protein labelling kit RED (Nanotemper) according to the manufacturer’s instructions. Differing concentrations of unlabelled Bat1 target DNA (prepared as above) were incubated with 100 nM Bat1 protein in microscale thermophoresis (MST) buffer (Tris 20 mM [pH 7.4], NaCl 150 mM, 10 mM MgCl2 and 0.05% Tween). Samples were loaded into NT.115 Hydrophilic Capillaries. Measurements were performed at room temperature, using 40% LED and 20% IR-laser power. Data analysis and Kd calculations were performed using Nanotemper Analysis software, v.1.4.17 and Origin 9.1.

Assembly of target plasmids *in vivo* and *in vitro* reporters

For the analysis of reporter activation in human cells target sites were assembled into a BsaI-digested pUC57 derivative with BsaI sites (*TAGA-GGAT*) preceding a minimal CMV promoter followed by a *dsEGFP* reporter gene (18; Supplementary Figure S6). Target sites were introduced as annealed primers (Metabion, annealing as for EMSAs), with matching four base pair overlaps, and were ligated into the BsaI cleaved vectors.

To create the target for the *in vitro* cleavage assay, $\mathrm{BE}_{\mathrm{Bat1}}$ was introduced into the transcriptionally silent *Capsicum annuum Bs3* promoter, previously cloned into pUC57, via mutagenesis PCR (see Supplementary Table S2 for primers and Supplementary Figure S6 for target sequences). The *Bs3* promoter derivatives used in Figure 8 were delivered in modified binary vector pGWB3* upstream of a *uidA* (GUS) reporter gene as previously described (10).

Transfection of HEK293T cells

HEK293T cells were grown in Dulbecco’s modified Eagle’s medium—high glucose (Sigma-Aldrich) supplemented with 10% FBS (Sigma-Aldrich), penicillin (100 U/ml) and streptomycin (100 μg/ml) in a 10% CO_2 atmosphere. 5 × 10^5 cells were transiently transfected using Fugene (Promega) according to the manufacturer’s instructions. Cells were transfected with 3 μg of Bat/TALE expression vector and 300 ng of the *dsEGFP* reporter plasmid.

Immunohistochemistry and microscopy

For microscopic analysis HEK293T cells were mounted on poly-L-lysine coated glass slides. Forty-eight hours after transfection, the cells were fixed with 4% formaldehyde in phosphate buffered saline (PBS) for 10 min. After permeabilisation with 0.5% Triton X-100 for 10 min, the cells were incubated with 3% bovine serum albumin (BSA) in PBS for 30 min. After 1 h incubation with the primary antibody (1/200 diluted mouse monoclonal antibody ANTI-FLAG M2 (Sigma-Aldrich) in PBS supplemented with 0.05% Tween-20 (PBS-T) and 3% BSA), cells were washed three times with PBS-T. Cells were then incubated with 1/600 Alexa Fluor 594 rabbit Anti-Mouse IgG (Invitrogen) in PBS-T with 3% BSA for 1 h. After washing three times with PBS-T, nuclei were counterstained with 4,6-diamidino-2-phenylindole (DAPI) and stored in 90% Glycerol in PBS with 0.25% DABCO. Images were acquired and processed using a Leica TCS SP5 confocal microscope equipped with an HCX PL APO CS 63x 1.2 Water objective. Images were processed using Leica AF and ImageJ (14).

FACS analysis of transfected HEK293T cells

Flow cytometry measurements of GFP and Alexa Fluor 594 were performed with a Becton-Dickinson FACS-Aria II. HEK293T cells were harvested, pelleted by centrifugation at 500 x g for 5 min at room temperature and gently washed with PBS. Cells were fixed with 4% formaldehyde in PBS for 10 min, pelleted by centrifugation at 500 x g for 5 min and permeabilized with 0.5% Triton X-100 for 10 min. After pelleting, the cells were incubated in 3% BSA for 30 min and then with mouse monoclonal antibody ANTI-FLAG M2 (Sigma-Aldrich, 1/100 dilution in PBS-T with 3% BSA) for 1 h. Subsequently, the cells were pelleted and washed three times with PBS-T and incubated with Alexa Fluor 594 rabbit anti-mouse IgG (Invitrogen, 1/500 dilution with PBS-T with 3% BSA) for 1 h. The cells were then pelleted and washed three times with PBS-T, stored in 500 μl PBS and analysed with FACS. Data were analysed using FlowJo V 10.0.6 (Tree Star). dsEGFP values for cells with above-threshold (Supplementary Figure S13) Alexa Fluor 594 fluorescence were used in Figures 3, 5–7.

In vitro nuclease assay

bat1-FokI and *TALE-FokI* genes were expressed *in vitro* using the Sp6 Quick coupled Transcription/Translation

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

system (Promega) as per manufacturer's instructions. Target DNA was PCR amplified from the previously assembled *Bs3p* derivatives using primers listed in Supplementary Table S2 and purified (GeneJET Gel extraction and DNA clean up Microkit, Life Technologies). Two hundred nanogram of PCR product was incubated with 5 μl transcription/translation product for 3 h at 37°C in cleavage buffer (1x restriction enzyme buffer 4, New England Biosciences, 1 ml/ml BSA, 500 nM NaCl). Reactions were terminated by heating to 60°C and DNA was separated (with kit as above). One hundred nanogram of DNA purified from the cleavage reaction was run on a 2% agarose gel. DNA was visualized via ethidium bromide staining under UV light. Size estimation was made in comparison to a standard ladder (GeneRuler 100 bp plus, Fermentas) and band intensities were measured with ImageJ (14).

GUS assays

dTALE or reporter constructs were transformed into *Agrobacterium tumefaciens* (GV3101) via electroporation. Strains were grown overnight in YEB medium containing rifampicin and kanamycin (each 100 μg/ml; for pGWB3* containing strains) or rifampicin and spectinomycin (each 100 μg/ml; for pGWB641 containing strains), collected by centrifugation, resuspended in inoculation medium (10 mM $MgCl_2$, 5 mM MES, pH 5.3, 150 μM acetosyringone) and adjusted to an OD_{600nm} of 0.8. For GUS assays equal amounts of *A. tumefaciens* strains containing *35S*-promoter driven *dTALE* genes and reporter constructs containing corresponding binding boxes fused to the reporter gene *uidA* (*GUS*) were mixed prior to inoculation. Leaf tissue was harvested after 48 h and GUS quantification was carried out as described (10).

RESULTS

Three TALE-like proteins are encoded in the genome of *B. rhizoxinica* strain HKI-0454

The Bat polypeptides are formed entirely of repetitive sequences with similarity to those of TALEs (Figures 1A, Supplementary Figures S1 and S2), excluding 17–18 amino acids at the very N-terminus (non-repetitive N-terminal domain; NND). This contrasts from all known TALEs and RipTALs, which possess N-terminal and C-terminal non-repetitive domains of between 100 and 300 amino acids each (Supplementary Figure S2) that are crucial to translocation and their *in planta* function as transcriptional activators (3,10). The Bat proteins can be divided into a set of core repeats all >45% identical to each other at the amino acid level and cryptic repeats not reaching this threshold (Figure 1B, Supplementary Figures S1 and S3; alignments generated with Clusal Omega 19,20). Core repeats are so named as they form the central, and largest, section of the studied polypeptides. Bat1, Bat2 and Bat3 have 20, 26 and 6 core repeats, respectively. The core repeats are framed by two N-terminal (−1, 0) and one C-terminal (+1) cryptic repeat in each Bat protein. The sequence identities of the various domains of the Bat proteins to each other are given in Supplementary Table S1.

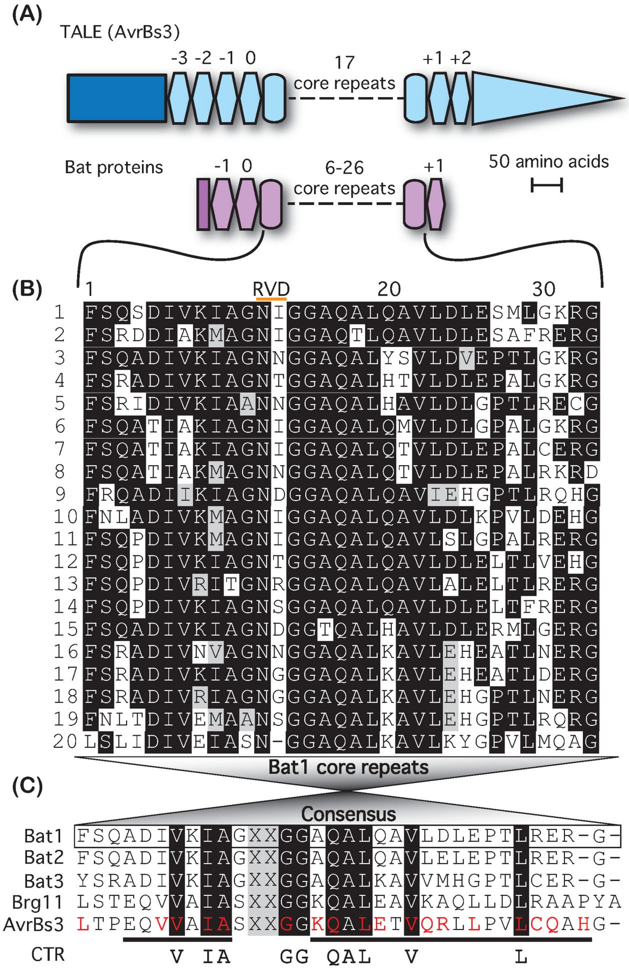

(A)
TALE (AvrBs3)
-3 -2 -1 0
17
core repeats
+1 +2
Bat proteins
-1 0
6-26
core repeats
+1
50 amino acids
(B)
1
RVD
20
30
1 FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
2 FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
3 FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
4 FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
5 FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
6 FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
7 FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
8 FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
9 FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14 FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15 FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16 FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17 FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18 FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
Bat1 core repeats
(C)
Consensus
Bat1 FSQADIVKIAGXXGGAQALQAVLDLEPTLRER-G-
Bat2 FSQADIVKIAGXXGGAQALQAVLELEPTLRER-G-
Bat3 YSRADIVKIAGXXGGAQALKAVVMHGPTLCER-G-
Brg11 LSTEQVVAIASXXGGKQALEAVKAQLLDLRAAPYA
AvrBs3 LTPEQVVAIASXXGGKQALETVQRLLPVLCQAHG-
CTR V IA GG QAL V L

Figure 1. Sequence-based comparison of TALE-like proteins. (**A**) Comparison of TALE (AvrBs3) and Bat architecture. The lengths of all domains are drawn to the indicated scale, except the dashes representing core repeats. TALE domains are shown in blue and Bat domains in purple. Rectangles indicate the N-terminal non-repetitive domain of each while a triangle indicates the non-repetitive C-terminal domain of TALEs including the transcriptional AD. Ovals represent core repeats, hexagons represent cryptic repeats (repeat number is indicated above). (**B**) Alignment of Bat1 core repeats, generated with Clustal Omega and Boxshade. Repeats are shown in order of appearance in the polypeptide. Repeat numbers are given on the left and positions within the repeat, including the RVD (indicated by an orange bar) above. (**C**) A consensus repeat generated from this alignment is compared to similarly generated consensus repeats from Bat2, Bat3, Brg11 (RipTAL) and AvrBs3 (TALE). From these a set of 10 hyper-conserved residues termed the consensus TALE-like repeat (CTR) was generated. The RVD positions are excluded from this. Repeat residues previously identified as involved in stabilising intra-molecular interactions from structural studies in TALEs (4) are highlighted with red lettering in the AvrBs3 consensus repeat, while the residues forming the first and second alpha helices (4) are underlined.

Consensus core repeats were deduced for each of the three Bat proteins (Figure 1B and Supplementary Figures S3). Bat1, 2 and 3 consensus repeats are 73–94% identical (Figure 1C, Supplementary Table S1). Each of the three Bat core repeat consensus sequences is less than 40% identical to equivalent consensus repeats of AvrBs3 and Brg11 (AvrBs3 from *X. campestris pv. vesicatoria* and Brg11 from *R. solanacearum* GMI1000 are used here as the represen-

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

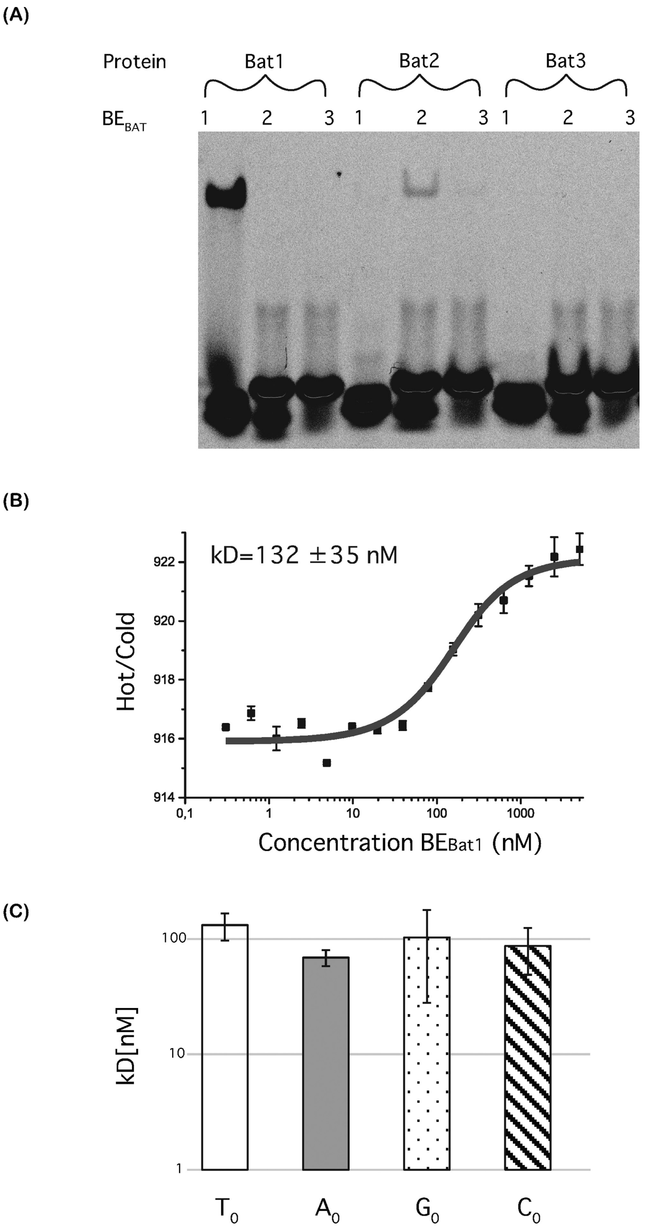


Figure 2. *In vitro* interaction studies of Bat proteins with predicted DNA targets. (**A**) Electrophoretic mobility shift assays were carried out for Bat1, 2 and 3 using 5'Cy5 labelled double-stranded DNA, bearing target sequences deduced from the TALE code. Each protein (100 nM) was tested against each target DNA (10 nM). Cy5 fluorescence was visualized after running through a native polyacrylamide gel. A shifted band, running slower on the gel, indicates the protein–DNA complex. (**B**) The interaction between Bat1 and its target (BE_{Bat1}) was quantified using microscale thermophoresis. The fluorescence ratio over the thermophoretic jump is shown on the y-axis against DNA concentration. Standard deviation for four repetitions is indicated. Measurements were made with 40% LED and 20% laser power. The dark grey line indicates the Kd fit. (**C**) This was repeated for BE_{Bat1} derivatives bearing A (grey bar), C (filled stripes) or G (spotted) at the zero position. The Kd was calculated in each case and is shown compared to that with BE_{Bat1} (T_0, empty bar).

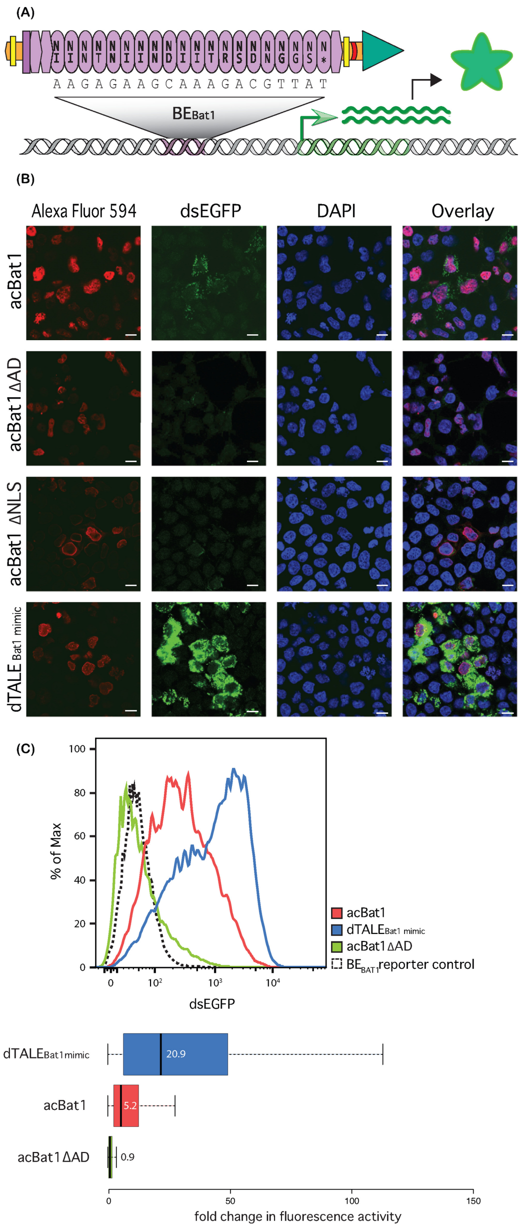


Figure 3. A Bat1 derived transcriptional activator (acBat1) is functional in a human cell reporter assay. (**A**) Schematic drawing showing the domain composition of acBat1. NLSs (yellow bars), a 3xFLAG tag (red crescent line) and a VP64 AD (green triangle) were fused onto Bat1 (purple) via flexible linkers (orange). This was introduced into HEK293T cells via transfection alongside a DNA reporter (grey) bearing BE_{Bat1} (purple) upstream of a dsEGFP coding sequence (green). Transcriptional activation of the reporter (green arrow) follows binding to BE_{Bat1}, leading to production of dsEGFP protein (green star). acBat1 is detected via the 3xFLAG epitope with use of an Alexa Fluor 594 labelled secondary antibody. (**B**) Alexa Fluor 594, dsEGFP and DAPI fluorescence are shown for transfected cells. acBat1 is compared to derivatives lacking AD (acBat1ΔAD) or NLSs (acBat1ΔNLSs) and to a dTALE created with the same NLSs and AD and with the same core repeat number and RVD composition as Bat1 ($dTALE_{Bat1mimic}$). The scale bar indicates 10 μm. (**C**) FACS analysis was used to quantify dsEGFP fluorescence for transfected cells expressing acBat1, ΔAD derivative or $dTALE_{Bat1mimic}$ as well as cells transfected with the reporter only. dsEGFP values are shown for the whole population (curves) as well as boxplots showing fold changes in fluorescence intensity compared to the reporter control. Boxplot whiskers represent the 2.5% and 97.5% data limits. Median values are written next to or inside each box plot and shown graphically with thick black lines.

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

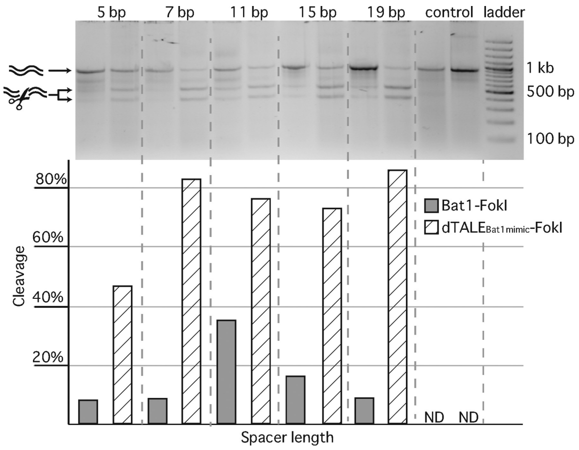


Figure 4. *In vitro* assessment of Bat1-FokI nuclease activity. Bat1- and TALE-FokI fusion proteins were expressed *in vitro* and equal volumes of transcription-translation product were incubated with a purified PCR product bearing two copies of $\mathrm{BE}_{\mathrm{Bat1}}$ in reverse complement, separated by 5–19 base pairs. A target with a control sequence replacing the Bat1 target boxes was also used. After 3 h incubation at 37°C DNA was purified from the nuclease reactions and run on a 2% agarose gel to discriminate cleaved and uncleaved DNA (indicated with arrows and illustrations on left side). Cleavage efficacy was calculated from the ratio of cleaved to uncleaved DNA band intensities in each lane with ImageJ (14). Full and striped bars indicate activities of the Bat1-FokI and TALEN constructs respectively. ND = none detected.

tative TALE and RipTAL, respectively). The Bat proteins thus form a highly diverged subgroup of the protein class referred to throughout this publication as ‘TALE-likes’ to mean TALEs, RipTALs and Bat proteins. Despite the high sequence diversity of repeats among TALE-like proteins, 10 residues are conserved in almost all TALE-like repeats and form what we term the ‘consensus TALE-like repeat’ (CTR; Figure 1C). The CTR includes residues clustering around the RVD as well as other residues, such as V22 and L29, able to form stabilising intra-molecular bonds in the crystal structure of DNA-bound TALE dHAX3 (Figure 1C; 4). Given their sequence conservation, the CTR residues are likely to make key contributions to the structure and function of the TALE repeat.

Bat1 and 2 mediate sequence-specific DNA binding with a code matching the TALE code

TALEs and RipTALs mediate sequence-specific DNA recognition with each core repeat recognising one DNA base and specificity determined by RVDs (the TALE code). We tested whether Bat proteins function similarly. In Bat proteins inter-repeat variability is not limited to the RVDs (positions 12 and 13), in fact position 12 varies very little and the diversity peaks between positions 23–30 (Figure 1B and Supplementary Figure S3). However, we continue to refer to positions 12 and 13 in Bat repeats as the RVD for consistency. The base specificities of most RVDs found in the Bat proteins are known from studies on TALEs and RipTALs allowing us to predict target sequences in each case. The single NR repeat (RVDs and their corresponding repeats are referred to with the single letter amino-acid code throughout) of Bat1 and the three repeats of Bat2 lacking both RVD residues were paired to Guanine and Thymine, based on presumed molecular similarities to NK and N* repeats, respectively.

Genes encoding His-tagged versions of the three Bat proteins were synthesized, expressed in *E. coli* (see Supplementary Figures S4 and S5 for sequences), purified and assayed for binding capabilities in EMSAs against their predicted binding elements ($\mathrm{BE}_{\mathrm{Bat1}}$, $\mathrm{BE}_{\mathrm{Bat2}}$ and $\mathrm{BE}_{\mathrm{Bat3}}$) (Figure 2A; sequences in Supplementary Figure S6). Bat1 and 2 both produced clear shifts in combination with their predicted target DNAs only (Figure 2A). Bat3, which has only six core repeats, was unable to produce a clear shift with any of the target DNAs (Figure 2A). Previous tests with TALEs have shown little activity with TALEs possessing fewer than 10 core repeats (1). It thus seems likely that Bat3 is either non-functional as a DNA-binding protein or mediates very weak interactions, not detectable in this assay.

Bat1 and 2, those displaying DNA binding with a clear sequence preference, are more similar to each other than either is to Bat3 (Supplementary Table S1). The Bat1 and 2 consensus core repeats are 94% identical. Considering the close homology of Bat1 and 2, DNA binding properties are likely conserved and only Bat1 was further characterized.

Bat1 binds its predicted target with an affinity within the upper boundary of TALE–DNA interactions and without base discrimination at the zero position

MST experiments were carried out to measure the binding strength of Bat1 with $\mathrm{BE}_{\mathrm{Bat1}}$. We found a disassociation constant (Kd) of 132 nM (Figure 2B). Affinities of TALEs with their target DNAs have been measured at 0.3 to >1000 nM (17), depending on the RVD composition. Yet, stronger interactions than that shown in Figure 2B are thought to be necessary for the *in vivo* function of TALEs. For example, the interaction of TALE AvrBs4 with its target site in the promoter of the pepper *Bs4C* resistance gene was previously measured by MST to have a Kd of 18.1 nM while the interaction with the homologous sequence from the non-activated *bs4C* allele had a Kd of 181.5 nM (21). Given that the affinity of Bat1 to $\mathrm{BE}_{\mathrm{Bat1}}$ is similar to the affinity of AvrBs4 to the non-activated *bs4C* allele, it is too low to suggest a strong interaction when assuming near-identical physiological conditions. This assumption may not be valid as, for example, the concentration of Bat proteins at the native site of action may differ from that of TALEs on delivery by *Xanthomonas* bacteria. Alternatively, $\mathrm{BE}_{\mathrm{Bat1}}$ may not represent the optimal binding sequence or additional endogenous factors may promote interaction *in vivo*.

$\mathrm{BE}_{\mathrm{Bat1}}$ was created in accordance with the TALE requirement for a thymine at the zero position (T_0 preference). However, the RipTALs do not share the T_0 preference and instead activate only G_0 targets (10). Therefore, we carried out further MST experiments with the different N_0, bases to clarify whether the T_0 preference holds for Bat1 or if another base is preferred. We found that in fact no significant differences were seen in the Kds of the different N_0 base target DNAs (Figure 2C and Supplementary Figure S7 and Table S3). This accords with the results of Juillerat *et al.* (22) using an *in vivo* reporter system. All further experiments

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

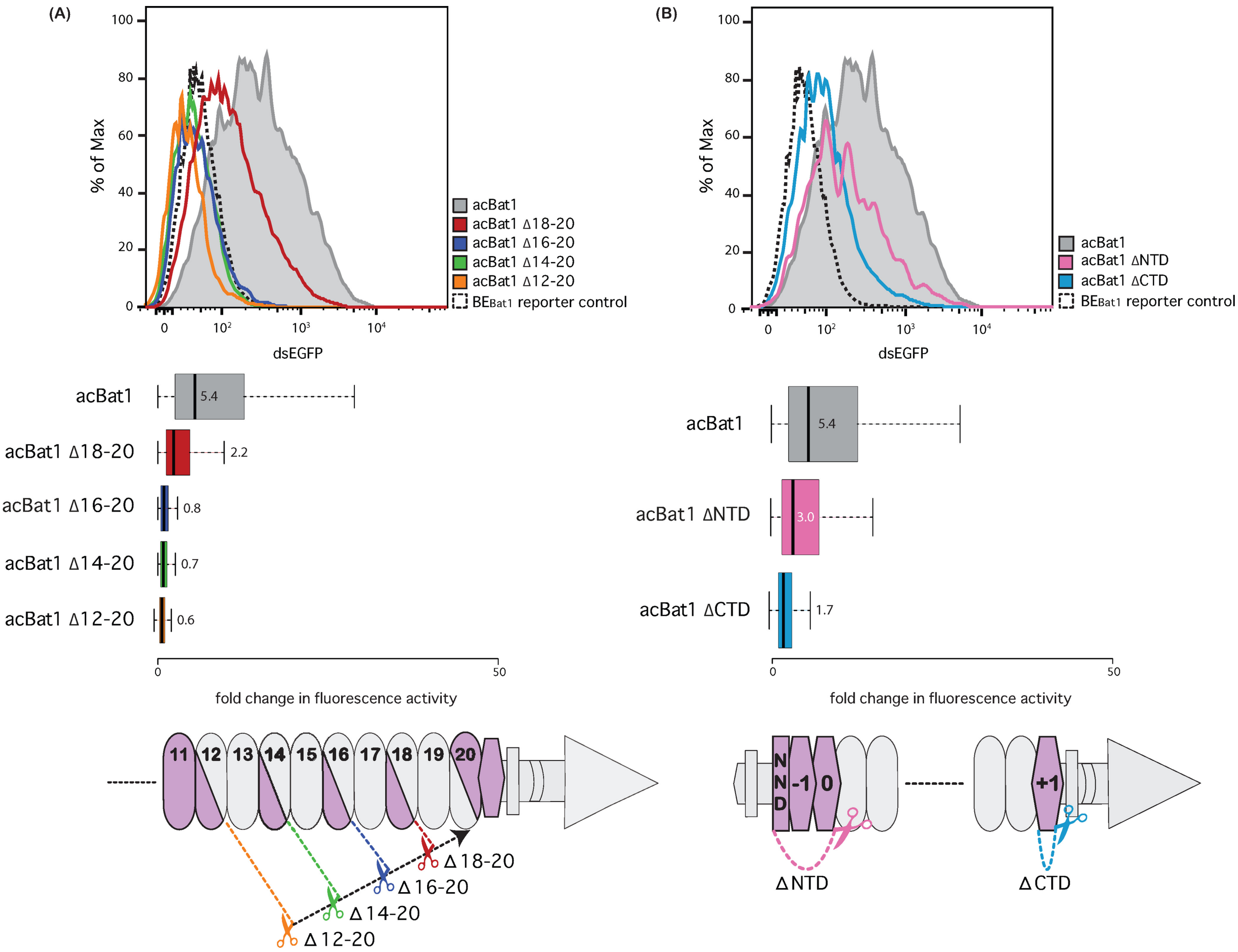


Figure 5. Functional analysis of acBat1 repeat truncations. Tests were carried out as described (Figure 3). Flow cytometry measurements of dsEGFP fluorescence are displayed as population distributions (top) or box plots (centre). Distinct colour codes are used throughout the whole figure and correspond to indicated constructs. Boxplots show fold changes in fluorescence intensity compared to the reporter control with whiskers representing the 2.5% and 97.5% data limits. Median values are written next to or inside each box plot and shown graphically as thick black lines. Cartoon representations of the tested truncations are shown below. Dashed lines with scissors indicate fixed (black) and variable (coloured) truncation points. Bat repeats and fused domains of acBat1 are represented as in Figure 3A. (**A**) Within the repeats grey or purple indicate truncated or retained regions, respectively. (**B**) N- (ΔNTD) or C- (ΔCTD) terminal truncations were tested. NND is the short non-repetitive N-terminal domain at the N-terminus of Bat1.

were carried out using T_0 targets to allow optimal conditions for comparison to TALE controls.

The fusion of NLSs and AD are sufficient to convert Bat1 into a targeted transcription factor in human cells and *in planta*

Having demonstrated that Bat1 binds its predicted target sequence *in vitro*, we developed a Bat1 derivative to function *in vivo* as a transcriptional activator and tested this with reporter assays. A Bat1 transcriptional activator (acBat1) was created through translational fusion of a viral NLS and a VP64 AD. A 3xFLAG epitope tag between NLS and VP64 domain (Supplementary Figure S5) allowed for antibody-based protein detection using an Alexa Fluor 594-tagged secondary antibody. We measured the ability of acBat1 to activate a dsEGFP-based reporter gene (18) in human cells (HEK293T; Figure 3A). A custom TALE-activator construct was tested in parallel. Termed $\mathrm{dTALE_{Bat1mimic}}$, it has the same repeat number and RVD composition as Bat1 and the same fused domains (Figure 3A, Supplementary Figures S5 and S8). Immunostaining showed that the acBat1 and $\mathrm{dTALE_{Bat1mimic}}$ both localized to the nucleus, while acBat1-ΔNLS, lacking the NLSs, did not localize to the nucleus. This demonstrates that NLSs must be added to Bat1 in order to target it to the nucleus in human cells (Figure 3B). dsEGFP expression in cells expressing acBat1 showed that it is able to activate the reporter. By contrast, cells expressing a derivative lacking the AD (acBat1-ΔAD) showed only Alexa Fluor 594 fluorescence, but did not show dsEGFP fluorescence indicating that the reporter was not activated (Figure 3B). Fusion of an AD is thus necessary to convert Bat1 into a functional transcriptional activator in human cells.

acBat1 induced the reporter 5-fold, while the $\mathrm{dTALE_{Bat1mimic}}$ induced the reporter 20-fold (Figure

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

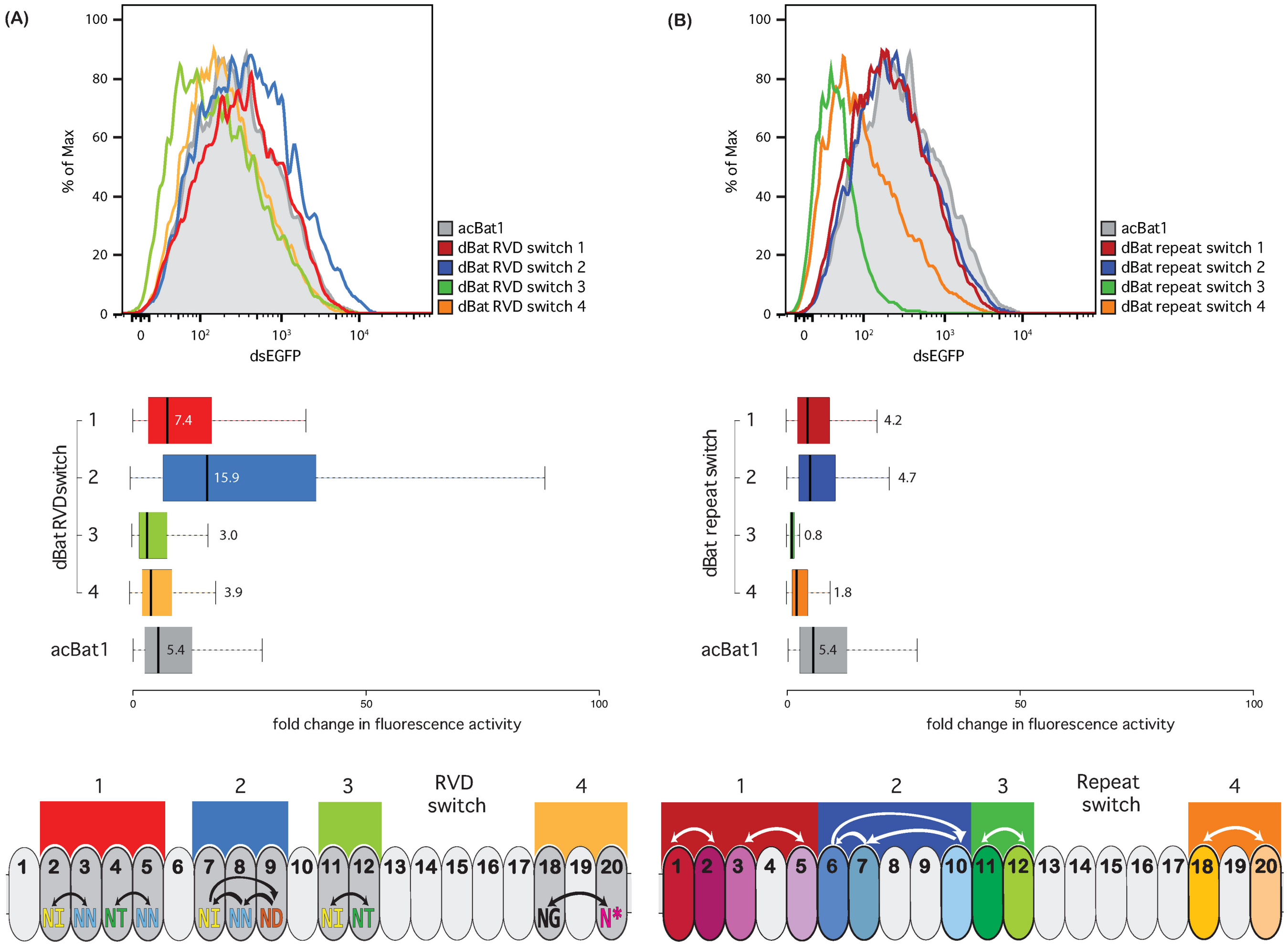


Figure 6. Functional analysis of designer (d)Bat constructs generated by RVD (**A**) or repeat switch (**B**). dBats were tested using flow cytometry with a transcriptional activation reporter as described (Figure 3). dsEGFP fluorescence values are displayed as population distributions (top) or boxplots (centre). dsEGFP values are normalized to the reporter only control (Supplementary Figure S13), which was BE_{Bat1} for all constructs except RVD switch 1 and 2 (Supplementary Figure S6). Boxplots show fold changes in fluorescence intensity compared to the reporter control with whiskers representing the 2.5% and 97.5% data limits. Median values are written next to or inside each box plot and shown graphically as thick black lines. dBat design is outlined below in each case. Coloured boxes indicate the repeats (ovals) modified in a given dBat. In the case of the RVD switch (A) modified repeats are highlighted with darker grey. RVDs are shown and colour coded by type. Arrows indicate the rearrangement of RVDs between repeats. In the case of the repeat switch (B) repeats are coloured to indicate that each has a unique set of non-RVD residues. Arrows indicate movement of whole repeats within the array.

3C). This may indicate that $dTALE_{Bat1mimic}$ has a higher affinity for BE_{Bat1} than acBat1 does. Alternatively, the activity of the C-terminally fused VP64 AD may be differentially affected by the architecture of each fusion protein.

To study functionality of acBat1 *in planta*, a corresponding T-DNA construct was delivered via *A. tumefaciens* into *Nicotiana benthamiana* leaves. In this assay, constitutively expressed acBat1 activated a co-delivered *uidA* reporter gene downstream of a promoter bearing BE_{Bat1} (Supplementary Figure S9). In analogy to the results observed in human cells, the $dTALE_{Bat1mimic}$ control was able to activate the reporter in plant cells to 3-fold higher levels than acBat1. In sum, we were able to show that acBat1 can transcriptionally activate a promoter with its target sequence in both human and plant cells.

Fusion of a FokI domain to the C-terminus of Bat1 creates a sequence-specific DNA nuclease

The most common approach for the creation of TALE-nucleases (TALENs) is a C-terminal translational fusion to a FokI endonuclease domain. Since the FokI endonuclease is active only as a dimer, interaction of two FokI domains is achieved by placing neighbouring TALEN target sites on opposite strands in reverse orientation promoting interaction of the FokI monomers after DNA binding. The FokI dimer catalyses formation of a double-strand break in the DNA spacer region between the two TALEN target sites. We created an analogous architecture using Bat1 to confer DNA binding specificity and compared its activity in an *in vitro* cleavage assay against the corresponding TALEN ($dTALE_{Bat1mimic}$-FokI; sequences given in Supplementary Figure S5). As target DNA we used a PCR product bearing two copies of BE_{Bat1} in reverse orientation on oppo-

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

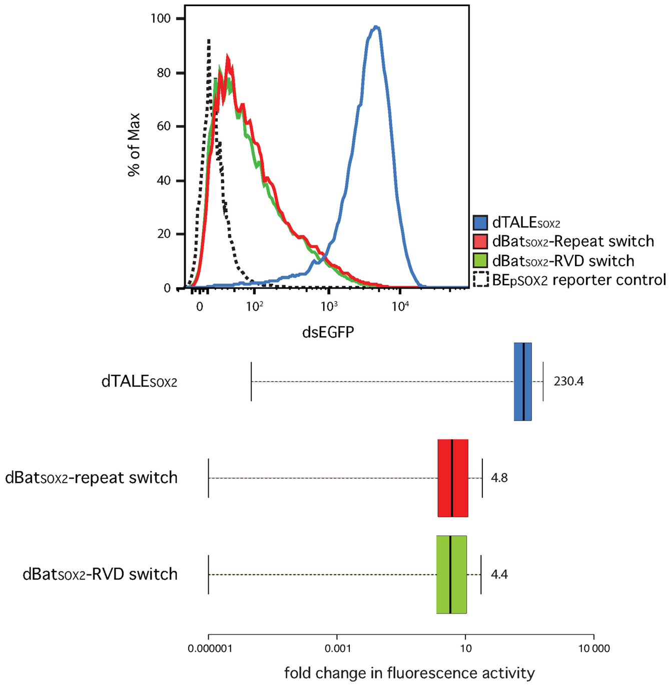


Figure 7. Functional analysis of designer (d)Bat constructs targeting the human SOX2 promoter. dBats were tested using flow cytometry with a transcriptional activation reporter as described (Figure 3). Population curves for dsGFP fluorescence are shown (top) as well as boxplots of fluorescence intensities (bottom) compared to the reporter control (logarithmic scale). Boxplots show fold changes in fluorescence intensity compared to the reporter control with whiskers representing the 2.5% and 97.5% data limits. Median values are written next to each box plot and shown graphically as thick black lines. Two dBats, designed based on the RVD ($\text{dBat}_{\text{SOX2 RVD switch}}$) or repeat switch ($\text{dBat}_{\text{SOX2 repeat switch}}$), and an equivalent dTALE were tested.

site strands. We generated derivatives differing only in the length of the DNA spacer separating the targets (Supplementary Figure S6) in order to determine the spacing between the two target sites that would result in the highest activity of the Bat-FokI fusion proteins. As a negative control, we tested a template with a control sequence instead of the Bat1 target sites.

Bat1-FokI and $\text{dTALE}_{\text{Bat1mimic}}$-FokI were expressed *in vitro* and equal volumes of reaction product were incubated with the target DNA. After 3 h at 37°C the DNA was size fractionated on a 2% agarose gel (Figure 4). Both Bat1 and TALE nucleases were able to cleave the target constructs. By contrast, the controls lacking target sites were not cleaved, indicating that Bat1-FokI, like the TALEN, is target specific in its DNA cleavage. The highest efficacy shown by Bat1-FokI was 35% cleavage (11 bp spacer) while $\text{dTALE}_{\text{Bat1mimic}}$-FokI had a maximum efficacy of 86% cleavage (19 bp spacer; Figure 4). That $\text{dTALE}_{\text{Bat1mimic}}$-FokI showed greater flexibility with respect to spacer length may relate to the previously optimized architecture employed (18). TALEN architecture is known to play a decisive role in spacer preference (23). Similarly, alternative Bat1 truncations or peptide linkers might allow for the creation of Bat1 nucleases with greater flexibility in spacer length.

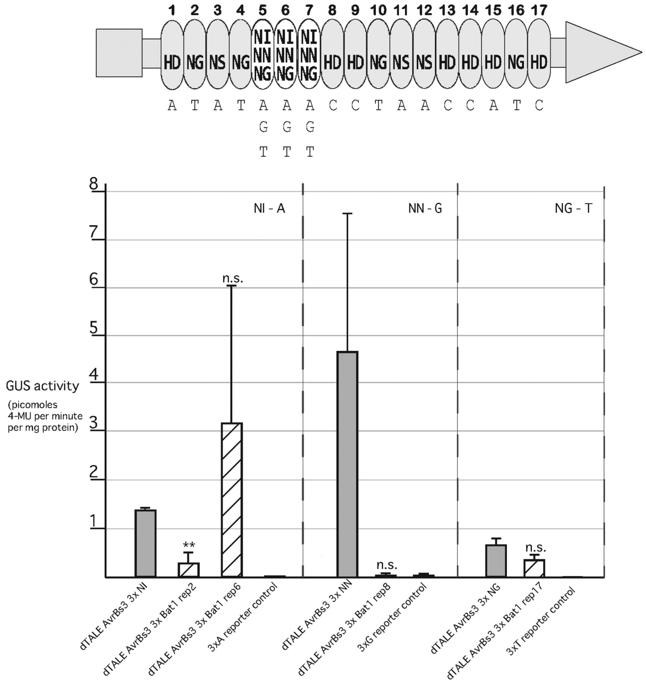


Figure 8. Functional analysis of Bat1 repeats within the context of a TALE repeat array. Trimers of identical Bat1 repeats or TALE repeats with the same RVDs as the Bat1 repeats were embedded into the repeat domain of the 17-repeat TALE AvrBs3 that targets the pepper *Bs3* promoter (*Bs3p*). Repeats 5–7 (3xRVD NI in AvrBs3) where replaced either by TALE repeat trimers with the RVDs NN or NG or by trimers of Bat1 repeats 2, 6, 8 and 17. This is shown in cartoon form with dTALE regions shown in light grey with the trimer of Bat1 repeats or dTALE repeats shown as white ovals. The grey rectangle and triangle indicate the native N- and C-terminal regions of AvrBs3, respectively. RVDs are given in each case and the matching bases in the target box underneath. The resulting chimeras (striped bars) were tested for their ability to activate a *Bs3p* derivative bearing the matching binding site upstream of a *uid*A (GUS) reporter gene and compared to non-chimeric dTALEs (filled bars) with the same RVDs. Dashed lines separate groups of constructs all with the same RVDs and tested against the same reporter. Barred lines indicate standard deviation. Two-tailed t-tests were used to compare chimeric and non-chimeric dTALEs for each reporter. A double asterisk indicates a P-value of below 0.02 and n.s. indicates a P-value of above 0.05.

The paradigm underlying the modification of core and cryptic TALE repeats cannot be applied to Bat1

In both natural (3) and custom TALEs, the number of core repeats is flexible, within a certain range. The number and position of cryptic N- and C-terminal repeats are typically inflexible, though alternative repeat −1 modules have recently been described (24,25). We tested acBat1 deletion derivatives to test if this paradigm applies to Bat1.

First, we tested variants of acBat1 lacking 2 (Δ18–20), 4 (Δ16–20), 6 (Δ14–20) or 8 (Δ12–20) core repeats (Figure 5A and Supplementary Figure S10). The later half of repeat 20 and repeat +1 were retained in each case. These truncations were tested against the BE_{Bat1} reporter and produced varied levels of reporter activation (Figure 5A). acBat1-Δ18–20 was able to activate the reporter more than 2-fold, corresponding to 40% activity of wild-type acBat1. The other truncation derivatives were unable to activate the reporter to levels above background. If we assume that each repeat contributes a certain amount of affinity to the Bat1–

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

BE_{Bat1} interaction then fewer than 17 repeats may simply be insufficient for an interaction strong enough to lead to reporter activation. This is in accordance with results from TALE repeat arrays showing that a certain number of core repeats are necessary for downstream reporter gene activation (1). Alternatively, the novel interface formed within the last repeat in each truncation derivative may create unfavourable intramolecular interactions, reducing protein activity. This explanation would not apply to TALEs where repeats are near identical and repeat order does not change the interface between repeats. Given the numerous non-RVD polymorphisms between Bat1 repeats, deletion or insertion of core repeats will always create novel repeat interfaces and should be experimentally validated before use in downstream applications.

We next tested acBat1 derivatives where the 82 residues N-terminal of core repeat 1 (acBat1ΔNTD; lacking repeats 0 and −1), or the 30 residues C-terminal of core repeat 20 (acBat1ΔCTD, lacking repeat +1) were deleted (Figure 5B and Supplementary Figure S10). Whilst acBat1ΔNTD showed a modest reduction in activity (56% of acBat1), acBat1ΔCTD was barely able to activate the reporter above background (Figure 5B). This does not match expectations based on TALEs where only the cryptic N- but not the cryptic C-terminal repeats are essential for DNA binding (26). By contrast, our results suggest that the cryptic C-terminal Bat1 repeat +1, in contrast to the corresponding cryptic TALE repeat +1, makes an unexpectedly strong contribution to activity and thus should be retained for the creation of active Bat1-based transcriptional activators.

Despite high inter-repeat diversity designer Bat1 proteins (dBats) with wild-type levels of activity can be assembled

The non-RVD residues of Bat1 repeats are highly polymorphic. This provides a means to study the functional relevance of non-RVD polymorphism in the native Bat1 as well as being relevant for the creation of Bat1 derivatives with novel specificity (dBats). We hypothesize that non-RVD polymorphisms may have two functionally relevant, non-mutually-exclusive, effects. (i) The formation of unique but functionally equivalent repeat interfaces that stabilize the superhelical structure formed by tandem-arranged repeats (4,5) (superstructural hypothesis). (ii) The creation of unique scaffolds optimized for the native RVD residues in each case (RVD scaffold hypothesis).

We used two different dBat design methods to test our hypotheses. These are the repeat switch and the RVD switch. Sequences of the dBats created can be found in Supplementary Figure S11. In the repeat switch whole repeats, including their native RVDs, were exchanged. This creates new interfaces between repeats but leaves RVDs in their native repeat context. If the superstructural hypothesis is correct then the repeat switch is likely to modify evolved repeat interfaces possibly yielding less active DNA-binding proteins. In the RVD switch it is only the RVDs that are changed while all non-RVDs remain unchanged. This design will not change repeat interfaces but will place RVDs in non-native repeat scaffolds. If the RVD scaffold hypothesis is correct then the RVD switch will reduce activity due to RVDs being sub-optimally oriented in relation to the paired DNA bases.

RVD composition and target sequence are key parameters determining affinity of TALE–DNA interactions and these were kept constant in our dBat tests as far as possible. For the repeat switch tests, we exchanged repeats with RVDs paired to the same base in BE_{Bat1} allowing the wild-type target construct to be used in each case. For the RVD switch constructs, where possible we exchanged RVDs with the same target base (dBat RVD switch 3 and 4) and tested these constructs against BE_{Bat1}. Where this was not possible exchanges were made between repeats in close proximity to one another to reduce any influence from an N- to C-terminal polarity effect as known for TALEs (17,27–29). These were then tested against BE_{Bat1} derivatives with the appropriate minor modifications in base composition. Thus any differences we see in activity are likely to be linked to effects arising from manipulation of repeats and not to differences in RVD composition or target sequence.

We found that despite the minor modifications in each case the different dBat constructs mediated strikingly varied levels of reporter activation. Of the four RVD switch constructs two were superior in activation level compared to acBat1 (2.9x and 1.4x relative to acBat1; Figure 6A). The other two dBat derivatives were slightly reduced in their activity as compared to acBat1 (0.56x and 0.72x relative to acBat1; Figure 6A). Overall, the impact on activity of the RVD switch constructs showed no single trend with some superior and some inferior to the wild type. Of the four repeat switch constructs none reached the activation level of acBat1 (Figure 6B). Notably, dBat repeat switch 3, in which core repeats 11 and 12 were exchanged, was unable to induce the reporter above background levels. Thus the repeat switch constructs all showed reduced activity compared to the wild type, and some dramatically so.

These data support that inter-repeat interfaces are unique and optimized (superstructural hypothesis) though whether the same is true for RVD scaffolds is not clear. That the RVD switch constructs performed differentially suggests that RVD scaffold can have a functional impact. However, the natural scaffold does not seem to be the optimal one in every case.

Custom dBats can be created to target a novel, user-defined sequence

We next tested whether the Bat1 repeat array could be fully customized to target a sequence of interest. Based on the two alternative strategies described above, $dBat_{SOX2}$-RVD switch and $dBat_{SOX2}$-repeat switch were created to activate a dsEGFP reporter driven from a minimal CMV promoter containing a binding element taken from the human *SOX2* promoter (Supplementary Figures S6, S8 and S11 for protein and reporter sequences). The SOX2 protein prevents determination in human neural stem cells and has previously been a target for dTALE studies (30). Both dBat repeat arrays were limited to 18 repeats instead of the wild-type 20 to bring them in line with the length of custom TALE repeat arrays commonly produced with our toolkit (15). The same NLS and VP64 fusions were used as for the assays displayed in Figure 3A. Both dBats were able to ac-

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

tivate the reporter to similar levels (Figure 7) suggesting that both the RVD and repeat switch strategies can yield successful constructs. dBat$_{\mathrm{SOX2}}$-repeat switch mediated 4.8x reporter activation and thus was slightly more active than the dBat$_{\mathrm{SOX2}}$-RVD switch (4.4x reporter activation). However, as seen previously (Figure 6), results can be surprisingly varied even between very similar dBat constructs and any potential design should be tested first in a reporter system before further application. Cross-reactivity assays testing the *SOX2* dBats on the BE$_{\mathrm{Bat1}}$ reporter showed that they were unable to activate the non-target reporter above background (Supplementary Figure S12) indicating that target specificity is maintained in the dBats. Further work on the creation of Bat1-based arrays and fusion proteins may improve activity levels. In conclusion, we were successfully able to reprogram the Bat1 protein for the creation of transcriptional activators with novel specificity.

TALE-Bat1 chimeras show varied activity but may be a means to harness the sequence diversity of Bat1 repeats

While the activation achieved with the SOX2 dBats was encouraging a custom TALE-activator for the *SOX2* promoter (dTALE$_{\mathrm{SOX2}}$) activated the reporter more than 200-fold (Figure 7). It may be possible to improve the activation levels achieved with dBats through further work on construct design and indeed Bat1 nuclease activities matching the corresponding activities of corresponding TALE nucleases were previously reported (22). However, another possibility is to create chimeric proteins to combine desirable features of both the Bat and TALE repeat scaffold.

We tested the principle of creating TALE-Bat chimeric repeat domains utilising a simple assay approach previously used in our lab to test chimeric TALE-RipTAL repeat arrays (10). Three identical copies of different Bat repeats were used to replace three repeats in a dTALE targeting the pepper *Bs3* promoter (*Bs3p*). These were then tested *in planta* against a reporter construct bearing a *Bs3p* fragment upstream of a *uidA* (GUS) gene. Three different reporters were used with triple A, G or T at the position that should be bound by the inserted Bat repeats in order to test repeats with different RVDs. In each case comparison was made to a dTALE assembled using only TALE repeats with the same RVD as the Bat repeats. As with earlier dBat tests we found strikingly different results for different constructs (Figure 8).

dTALE$_{\mathrm{AvrBs3\_3xBat1\_rep2,}}$ a dTALE bearing three copies of Bat1 repeat 2 (RVD NI) at the test positions, gave a significantly weaker induction of the reporter compared to the control with TALE repeats only (dTALE $_{\mathrm{AvrBs3\_3xNI}}$). dTALE$_{\mathrm{AvrBs3\_3xBat1\_rep8}}$ (RVD NN) was barely able to elicit any detectable activation, unlike its TALE repeat equivalent (dTALE$_{\mathrm{AvrBs3\_3xNN}}$). In contrast, dTALE$_{\mathrm{AvrBs3\_3x\_Bat1\ rep6}}$ (RVD NI) and dTALE$_{\mathrm{AvrBs3\_3xBat1\_rep17}}$ (RVD NG) activated their reporters to a level not significantly different from the TALE repeat control constructs. It is not possible to clarify whether differences in functionality arise from performance differences between Bat or TALE repeats in their native confirmations or if the differences arise due to the formation of novel and likely unfavourable inter-repeat interactions in these chimeric constructs (see superstructural hypothesis above). The functionality of any potential chimeric binding domain is likely to depend on both the particular repeats utilized and their arrangement within the repeat domain. However, we have demonstrated that such chimeric repeat domains containing some Bat1 repeats can be functional to the same level as TALE repeat equivalents paving the way for further development and applied uses.

DISCUSSION

The Bat proteins, together with the TALEs and RipTALs, form the TALE-like protein class. Like the other TALE-likes, Bat proteins mediate sequence-specific DNA binding with specificity predicted from the established TALE code. This functional similarity likely correlates to a structural similarity since DNA recognition proceeding via the TALE code relies on a particular structure that places position 13 of each repeat in close proximity to a single DNA base (4,5). Indeed modelling the structure of Bat1 based on the known structure of TALE Pthxo1 binding to its target DNA (5) suggests that the whole Bat1 polypeptide would form a sequence aligning closely to the TALE core repeat domain (Supplementary Figure S15).

Comparison of the core repeats of distinct TALE-likes enabled us to define a set of conserved residues, the CTR, as a unifying feature of the TALE-like proteins (Figure 1C). The CTR could be a useful tool to scan databases for further TALE-likes. In addition, the conservation of CTR residues suggests that they have an important functional relevance. Intriguingly, the CTR residues do not include some repeat residues such as K16, which have been shown to provide a large contribution to non-base-specific DNA binding, or H33, suggested as key to stabilisation of the TALE repeat (31). Conversely, some CTR residues such as L29 cannot currently be linked to a certain key function. Thus, investigation of the TALE-likes provides an interesting window into the opportunities for and constraints on sequence diversification whilst maintaining protein function.

We have demonstrated that the Bat1 protein itself can be taken as a targeting module for transcriptional activation (Figure 3) and nuclease function (Figure 4). The repeat array can also be reprogrammed to target a sequence of interest (Figure 7). Unlike the reprogramming of TALEs, alternative design strategies must be considered to generate Bat1 repeat arrays with desired base specificity and we have successfully employed two conceptually distinct design approaches (Figure 6). However, Bat1 and derivative fusion proteins were outperformed by equivalent TALE fusions (Figures 3, 4 and 7). This may relate to the relatively low affinity of Bat1 for BE$_{\mathrm{Bat1}}$ (Figure 2B) compared to known affinities of TALEs for their natural target boxes. However, the TALE platform has been optimized over several years. The creation of high activity TALE-nucleases, in particular, has been a focus of many labs. Thus, with further work to improve activity, the Bat platform may prove a more compact alternative to TALEs for targeted DNA binding without any zero base preference to be taken into account (Figure 2C). Alternatively, Bat repeats could be assembled along with TALE repeats to create chimeric DNA-binding proteins with novel properties. At the very least the inclusion of some Bat repeats into TALE repeat arrays would lower

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

sequence identity between repeats, useful for some cloning strategies, and possibly alleviating the previously reported problem of recombinatorial repeat loss (32). That Bat1 repeats can be integrated into a dTALE whilst retaining functionality is shown in Figure 8, but since no two Bat1 repeats are identical, so too must each Bat1-TALE chimera be treated as novel and requiring experimental validation before further use.

Functionally relevant differences between TALEs and Bat proteins were discovered upon attempting to modify the repeat domain. Bat1 showed surprisingly little tolerance to reductions in repeat number below 18 repeats (Figure 5A). These results seem to be in agreement with analysis of TALE proteins where a minimum number of repeats was needed to achieve *in vivo* function (1). The conclusion that has been drawn from such analysis is that each TALE repeat contributes something towards affinity and that a certain number of repeats are required to achieve the affinity necessary for *in vivo* function. However, the situation for Bat proteins is more complex. Due to the numerous non-RVD polymorphisms between each repeat (Figure 1B), a novel interface is formed when truncations are made within the repeat domain and these could have functionally deleterious consequences. Indeed the results of rearrangements within the repeat domain (Figure 6B) suggest that this is so.

A further difference between Bat1 and TALEs is the relative impact of truncations of the N- and C-terminal cryptic repeats. The N-terminal cryptic repeats of TALEs make a decisive contribution to DNA affinity such that their removal fully ablates DNA binding (26). By contrast, the limited evidence available suggests that the C-terminal cryptic repeats of TALEs contribute little to affinity and specificity. This includes the independently observed (17,27–29) N- to C-terminal reduction along the binding domain of contribution to base specificity. In addition, TALE fusion proteins with truncations in C-terminal cryptic repeat +2 (Supplementary Figure S2) are active (18) suggesting that any affinity contribution is not decisive. Thus in TALEs the N-terminal cryptic repeats seem to contribute more to DNA binding than the C-terminal cryptic repeats. This contrasts to our findings based on truncations of the N- and C-terminal cryptic repeats of acBat1. We found that the N-terminal truncation had a modest impact on reporter activation and did not contribute to specificity (Figures 2C, 5B and Supplementary Figure S7 and Table S3), whilst the truncation of the single C-terminal cryptic repeats almost entirely ablated activity (Figure 5B). This repeat may be important for DNA binding and the high proportion of positively charged residues (8/30; Supplementary Figure S1) is in agreement with a possible contribution to interaction with the negatively charged DNA phosphate backbone. Sequence comparison of the cryptic repeats of Bats and AvrBs3 (see Supplementary Figures S1 and S2) showed that the 0 repeats share a few residues (L1, L7 and K8) not found in the CTR (Figure 1C) but no such unique conserved residues can be found among the -1 or +1 repeats. Together with the results shown in Figure 5B it appears that, at both the sequence and functional level, at least the cryptic repeats -1 and +1 of Bats and TALEs are likely to be non-homologous.

Through the exploration of dBat assembly strategies, we gained insights into the functional significance of Bat1 non-RVD polymorphisms. These polymorphisms provided a molecular handle to question different models. The results of these experiments are possibly specific to Bat proteins but most likely are relevant to the non-RVD polymorphisms of other TALE-like proteins. The RVD switch constructs (Figure 6A) tested the importance of the RVD scaffold formed by all the non-RVD residues of a repeat, while the repeat switch constructs tested the importance of inter-repeat interactions (Figure 6B). We found that all repeat switch constructs were less active than the wild type (Figure 6B). This supports the hypothesis that the non-RVD polymorphisms of adjacent Bat1 repeats lead to the formation of unique but functionally equivalent interfaces between repeats. Our model for the structure of Bat1 bound to DNA suggests that unique bonds are indeed formed between varied residues of Bat1 repeats (Supplementary Table S4). Perturbation of these possibly co-evolved residues would likely impair protein function. The performances of the RVD-switch constructs (Figure 6A) were mixed, with some activating the reporter better than the wild-type acBat1. This speaks against the idea that each repeat scaffold has co-evolved with its RVD for optimal activity. The data do, however, support previous findings from RipTALs (10) and TALEs (33) that certain non-RVD polymorphisms can have profound effects on repeat activity. These effects can be negative or positive and must be investigated individually. The quantity of non-RVD polymorphisms in Bat1 repeats compared to TALEs (3) or RipTALs (10) thus complicates the creation of designer DNA binding domains but also represents an as yet unexploited pool of potentially beneficial repeat variants.

Comparing the diversity of Bat and TALE repeats also raises evolutionary questions. The consensus core repeats or TALEs and Bats are less than 40% conserved (Figure 1C) at the sequence level, but at the functional level Bat and TALE repeats are apparently very similar. This shows that the sequence composition of TALE-like repeats is not heavily constrained by functional requirements. If most polymorphisms are functionally equivalent we would expect that, over time, inter-repeat polymorphisms would accumulate. The high levels of inter-repeat polymorphism in the Bat proteins (Figure 1B and Supplementary Figure S3) are consistent with this assumption. What is surprising is the relative sequence uniformity of TALE repeats. This suggests that TALE repeats are under the influence of a selective pressure to maintain sequence conservation, not felt by Bat proteins. However, while the non-RVDs of each TALE repeat are highly uniform the RVD composition and repeat number are highly diverse (3). These observations may be mutually explanatory. It is known that repeat regions of *TALE* genes can evolve via intra- and inter-molecular recombination (34,35). It may be, therefore, that the sequence conservation between individual *TALE* repeats promotes this recombination and subsequent diversification of repeat number and RVD composition. This property may be positively selected for in *TALE* genes. These assumptions and hypotheses require further testing, but comparison to non-*Xanthomonas* TALE-likes will likely prove a helpful one. Indeed the RipTALs, which show intermediate sequence di-

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

versity and limited structural diversity (10), provide an interesting third group for comparison.

We have shown that the Bat proteins are a highly divergent subgroup within a class referred to as the TALE-likes, which they help to define. Moreover, Bat specificity can be programmed with a code matching to known TALE and RipTAL repeat specificity (Figure 2A). Bat proteins thus represent an alternative platform for programmable sequence-specific DNA targeting. In addition, the highly diverse Bat repeats may prove a valuable reservoir for novel residue combinations with beneficial properties. More than this they provide an out-group for comparative analysis into function and evolution of RipTALs and TALEs. Further research into the Bat proteins is thus likely to reap rewards for both fundamental and applied research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

T. Strauß for assistance in plant growth. K. H. Braun for assistance in assembly of constructs used in this publication.

FUNDING

Two Blades foundation; Deutsche Forschungsgemeinschaft (SFB924). Source of Open Access funding: Deutsche Forschungsgemeinschaft (SFB924) and the Open Access Publishing Fund of Tuebingen University.

Conflict of interest statement. T.L. is a partial owner of a patent application regarding the use of TALEs.

REFERENCES

1. Boch,J., Scholze,H., Schornack,S., Landgraf,A., Hahn,S., Kay,S., Lahaye,T., Nickstadt,A. and Bonas,U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509–1512.
2. Moscou,M.J. and Bogdanove,A.J. (2009) A simple cipher governs DNA recognition by TAL effectors. *Science*, **326**, 1501.
3. Boch,J. and Bonas,U. (2010) *Xanthomonas* AvrBs3 family-type III effectors: discovery and function. *Annu. Rev. Phytopathol.*, **418**, 419–436.
4. Deng,D., Yan,C., Pan,X., Mahfouz,M., Wang,J., Zhu,J.-K., Shi,Y. and Yan,N. (2012) Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science*, **335**, 720–723.
5. Mak,A.N.-S., Bradley,P., Cernadas,R.A., Bogdanove,A.J. and Stoddard,B.L. (2012) The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science*, **335**, 716–719.
6. Doyle,E.L., Stoddard,B.L., Voytas,D.F. and Bogdanove,A.J. (2013) TAL effectors: highly adaptable phytobacterial virulence factors and readily engineered DNA-targeting proteins. *Trends Cell Biol.*, **23**, 390–398.
7. Mendenhall,E.M., Williamson,K.E., Reyon,D., Zou,J.Y., Ram,O., Joung,J.K. and Bernstein,B.E. (2013) Locus-specific editing of histone modifications at endogenous enhancers. *Nat. Biotechnol.*, **31**, 1133–1136.
8. Maeder,M.L., Angstman,J.F., Richardson,M.E., Linder,S.J., Cascio,V.M., Tsai,S.Q., Ho,Q.H., Sander,J.D., Reyon,D., Bernstein,B.E. *et al.* (2013) Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat. Biotechnol.*, **31**, 1137–1142.
9. Konermann,S., Brigham,M.D., Trevino,A.E., Hsu,P.D., Heidenreich,M., Cong,L., Platt,R.J., Scott,D.A., Church,G.M. and Zhang,F. (2013) Optical control of mammalian endogenous transcription and epigenetic states. *Nature*, **500**, 472–476.
10. de Lange,O., Schreiber,T., Schandry,N., Radeck,J., Braun,K.H., Koszinowski,J., Heuer,H., Strauß,A. and Lahaye,T. (2013) Breaking the DNA binding code of *Ralstonia solanacearum* TAL effectors provides new possibilities to generate plant resistance genes against bacterial wilt disease. *New Phytol.*, **199**, 773–786.
11. Lackner,G., Moebius,N., Partida-Martinez,L.P., Boland,S. and Hertweck,C. (2011) Evolution of an endofungal lifestyle: deductions from the *Burkholderia rhizoxinica* genome. *BMC Genomics*, **12**, 210.
12. Lackner,G., Moebius,N., Partida-Martinez,L. and Hertweck,C. (2011) Complete genome sequence of *Burkholderia rhizoxinica*, an endosymbiont of *Rhizopus microsporus*. *J. Bacteriol.*, **193**, 783–784.
13. Stella,S., Molina,R., Bertonatti,C., Juillerrat,A. and Montoya,G. (2014). Expression, purification, crystallization and preliminary X-ray diffraction analysis of the novel modular DNA-binding protein BurrH in its apo form and in complex with its target DNA. *Acta Crystallogr. F*, **70**, 87–91.
14. Schneider,C.A., Rasband,W.S. and Eliceiri,K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671–675.
15. Morbitzer,R., Elsaesser,J., Hausner,J. and Lahaye,T. (2011) Assembly of custom TALE-type DNA binding domains by modular cloning. *Nucleic Acids Res.*, **39**, 5790–5799.
16. Nakamura,S., Mano,S., Tanaka,Y., Ohnishi,M., Nakamori,C., Araki,M., Niwa,T., Nishimura,M., Kaminaka,H., Nakagawa,T. *et al.* (2010) Gateway binary vectors with the *bialaphos* resistance gene, *bar*, as a selection marker for plant tranformation. *Biosci. Biotechnol. Biochem.*, **74**, 1315–1319.
17. Meckler,J.F., Bhakta,M.S., Kim,M.-S., Ovadia,R., Habrian,C.H., Zykovich,A., Yu,A., Lockwood,S.H., Morbitzer,R., Elsäesser,J. *et al.* (2013) Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Res.*, **41**, 4118–4128.
18. Mussolino,C., Morbitzer,R., Lütge,F., Dannemann,N., Lahaye,T. and Cathomen,T. (2011) A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res.*, **39**, 9283–9293.
19. Goujon,M., McWilliam,H., Li,W., Valentin,F., Squizzato,S., Paern,J. and Lopez,R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695–W699.
20. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Soding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
21. Strauß,T., Van Poecke,R., Strauß,A., Römer,P., Minsavage,G.V., Singh,S., Wolf,C., Strauß,A., Kim,S., Lee,H.-A. *et al.* (2012) RNA-seq pinpoints a *Xanthomonas* TAL-effector activated resistance gene in a large crop genome. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 19480–19485.
22. Juillerat,A., Bertonati,C., Dubois,G., Guyot,V., Thomas,S., Valton,J., Beurdeley,M., Silva,G.H., Daboussi,F. and Duchateau,P. (2014). BurrH: a new modular DNA binding protein for genome engineering. *Sci. Rep.*, **4**, 3831.
23. Miller,J.C., Tan,S., Qiao,G., Barlow,K.A., Wang,J., Xia,D.F., Meng,X., Paschon,D.E., Leung,E., Hinkley,S.J. *et al.* (2011) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.*, **29**, 143–148.
24. Lamb,B.M., Mercer,A.C. and Barbas,C.F. III. (2013) Directed evolution of the TALE N-terminal domain for recognition of all 5' bases. *Nucleic Acids Res.*, **41**, 9779–9785.
25. Tsuji,S., Futaki,S. and Imanishi,M. (2013) Creating a TALE protein with unbiased 5'-T binding. *Biochem. Biophys. Res. Commun.*, **1**, 262–265
26. Gao,H., Wu,X., Chai,J. and Han,Z. (2012) Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res.*, **22**, 1716–1720.
27. Garg,A., Lohmueller,J.J., Silver,P.A. and Armel,T.Z. (2012) Engineering synthetic TAL effectors with orthogonal target sites. *Nucleic Acids Res.*, **40**, 7584–7595.
28. Perez-Quintero,A.L., Rodriguez,R.L., Dereeper,A., Lopez,C., Koebnik,R., Szurek,B. and Cunnac,S. (2013) An improved method for TAL effectors DNA-binding sites prediction reveals functional convergence in TAL repertoires of *Xanthomonas oryzae* strains. *PLoS One*, **8**, e68464.
29. Mali,P., Aach,J., Stranges,P.B., Esvelt,K.M., Moosburner,M., Kosuri,S., Yang,L. and Church,G.M. (2013) CAS9 transcriptional

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.*, **31**, 833–838.

30. Zhang,F., Cong,L., Lodato,S., Kosuri,S., Church,G.M. and Arlotta,P. (2011) Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat. Biotechnol.*, **29**, 149–153.
31. Wicky,B.I., Stenta,M. and Dal Peraro,M. (2013) TAL effectors specificity stems from negative discrimination. *PLoS One*, **8**, e80261.
32. Holkers,M., Maggio,I., Liu,J., Janssen,J.M., Miselli,F., Mussolino,C., Recchia,A., Cathomen,T. and Gonçalves,M.A.F.V. (2013) Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic Acids Res.*, **41**, e63.
33. Sakuma,T., Ochiai,H., Kaneko,T., Mashimo,T., Tokumasu,D., Sakane,Y., Suzuki,K. I., Miyamoto,T., Sakamoto,N., Matsuura,S. *et al.* (2013) Repeating pattern of non-RVD variations in DNA-binding modules enhances TALEN activity. *Sci. Rep.*, **3**.
34. Yang,Y. and Gabriel,D.W. (1995) Intragenic recombination of a single plant pathogen gene provides a mechanism for the evolution of new host specificities. *J. Bacteriol.*, **177**, 4963–4968.
35. Yang,B., Sugio,A. and White,F.F. (2005) Avoidance of host recognition by alterations in the repetitive and C-terminal regions of AvrXa7, a type III effector of *Xanthomonas oryzae* pv. *oryzae*. *Mol. Plant-Microbe Interact.* **18**, 142–149.

Downloaded from http://nar.oxfordjournals.org/ at Istituto di Neurologia/ on June 4, 2014

5 DNA-binding proteins from marine bacteria make novel contributions to the sequence diversity of TALE-like repeats

This chapter is identical to the publication:

de Lange, Orlando, Wolf, Christina, Thiel, Phillip, Krueger, Jens, Kohlbacher, Oliver & Lahaye, Thomas

DNA-binding proteins from marine bacteria make novel contributions to the sequence diversity of TALE-like repeats

Nucleic Acids Research (submitted 17.04.2015).

DNA-binding proteins from marine bacteria make novel contributions to the sequence diversity of TALE-like repeats

Orlando de Lange[1†], Christina Wolf[1†], Philipp Thiel[2], Jens Krüger[2], Oliver Kohlbacher[2,3] & Thomas Lahaye[1*]

[†] These authors contributed equally to this work
[1] Department of General Genetics, Centre for Plant Molecular Biology, University of Tuebingen, Auf der Morgenstelle 32, Tuebingen, Baden-Wuerttemberg, 72076, Germany
[2] Department of Computer Science and Centre for Bioinformatics, University of Tuebingen, Sand 14, Tuebingen, Baden-Wuerttemberg, 72076, Germany
[3] Quantitative Biology Centre and Faculty of Medicine, University of Tuebingen, Sand 14, Tuebingen, Baden-Wuerttemberg, 72076, Germany
* To whom correspondence should be addressed. Tel: +49-7071 / 29 7 8745
Fax: +49-7071 / 29 50 42 Email: thomas.lahaye@zmbp.uni-tuebingen.de

Abstract

Transcription Activator-Like Effectors (TALEs) of *Xanthomonas* bacteria are programmable DNA binding proteins with unprecedented target specificity. Efforts have been made to engineer TALE DNA-binding repeats with novel properties, but are made difficult by the very limited pool of known TALE repeat sequence variations. More sequence-diverse TALE-like proteins are known from *Ralstonia solanacearum* (RipTALs) and *Burkholderia rhizoxinica* (Bats), but their repeats are conserved with those of TALEs around the DNA-binding residue. We have assessed the structure and DNA-binding properties of repeats of novel marine-organism TALE-like proteins (MOrTLs), the first to date of non-terrestrial origin. We found that they function analogously to, and are compatible with, TALE and Bat repeats despite low sequence similarity around the base specifying residue (BSR). MOrTL repeat sequences could be used in the future to augment existing TALE-technology. Repeat residues around the BSR were found to be conserved within but not between TALE-like groups, and in fact only three residues spread around the repeat are highly conserved across all groups. This conserved motif could prove useful as an identifier for future TALE-likes. Additionally, comparing MOrTL repeats with those of other TALE-likes suggests a common evolutionary origin for the TALEs, RipTALs and Bats.

Introduction:

TALEs are effector proteins injected by plant pathogenic bacteria of the genus *Xanthomonas* into host plants where they mimic eukaryotic transcription factors (1). An array of DNA-binding repeats mediate TALE-promoter interaction and an acidic activation domain at the C-terminus of the protein mediates host gene expression. TALEs hijack the host's transcriptional machinery (2, 3) to activate plant genes beneficial to the pathogen.

It was the 2009 publication (4, 5) of the TALE code for sequence-specific DNA binding that launched these proteins into the limelight. The code in its refined form describes the base preferences of each possible residue at repeat position 13 (base specifying residue; BSR) (6). Each repeat pairs with one base, generally without interference from adjacent repeats. With this information one can predict the DNA binding element (BE) for any given TALE and what is more, design a TALE to match any DNA sequence of interest. Designer (d)-TALE DNA-binding domains, coupled to a functional domain of choice are invaluable tools for precision manipulation of genome (7), transcriptome (8) and even epigenome (9).

One of the potential advantages of the TALE system over the alternative CRISPR/Cas9 system is that TALE repeat arrays are highly flexible with respect to the length and sequence composition of the desired target site. Furthermore BSRs bind their cognate bases with a range of different affinities and specificities as inferred from reproducible effects on the binding properties of arrays containing such repeats (10). This diversity of molecular interactions found in TALEs contrasts with the restricted Watson-Crick base pairing of the CRISPR/Cas9 platform. The diversity of base recognition characteristics found among BSRs offers the possibility of creating finely tuned DNA-binding proteins. This possibility could be particularly useful in synthetic biology where parts libraries covering a range of parameter values for a property of interest are highly desirable. A library of dTALEs with a range of binding strengths for the same DNA element could be useful for creating synthetic genetic circuitry.

The range of DNA-binding properties mediated by TALE repeats might be further expanded by exploiting non-BSR polymorphisms. Till now, the only non-BSR position to receive much attention for TALE repeat engineering has been the neighbouring residue 12 (commonly referred to together with position 13 as the RVD (5)). By now all 400 possible RVDs have been assessed for interaction strength and specificity (11, 12). This confirmed that residue 13 is the major base-specificity determinant but some residue 12 polymorphisms were found to affect DNA recognition properties, supporting a role for non-BSR polymorphisms in defining the DNA recognition properties of TALE repeats. Furthermore, repeat arrays with novel RVDs were found to outcompete those based on commonly occurring RVDs, displaying increased on-target and reduced off-target activity (11). However, non-BSR polymorphisms beyond position 12 are scarce in natural TALEs, and the studies examining the functional impact of this diversity are similarly scarce. Only 18 non-RVD polymorphisms are found between the full repeat arrays of TALEs AvrBs3 and AvrBs4, and when these were exchanged en masse they were found to have no significant functional impact on reporter gene activation (8). Another study found that TALE nuclease repeat arrays formed of repeats differing from one another at residues 4 and 32 had slightly enhanced activity *in vivo* compared to similar constructs made of fully identical repeat arrays (13). Overall, however, the paucity of natural diversity among TALE proteins limits the raw material available for repeat engineering.

TALEs are distributed widely among *Xanthomonas* species but sequence diversity is very low (14). TALE-like proteins from other bacterial groups offer a far greater pool of sequence diversity. An effector protein called Brg11, 41% identical to AvrBs3 , is produced by plant pathogen *Ralstonia solanacearum,* strain GMI1000 (15). Further *brg11* homologs were later identified in a large number of *R. solanacearum* strains (16, 17). These proteins are termed RipTALs (*Ralstonia* injected protein TALE-like). In 2013 we and others published molecular characterisations of these RipTAL proteins (16, 18) showing that they function analogously to TALEs as plant transcription factors. TALE and RipTAL repeats are polymorphic at numerous positions and, in addition, the polymorphism among RipTAL repeats is greater than that among TALE repeats. We looked at the DNA recognition properties of each of the repeats of Brg11 and found differences in reporter activation strength even when comparing repeats with identical BSRs (16). This suggested an impact on repeat-DNA interactions from non-BSR polymorphisms. RipTAL repeats could be useful as a pool of natural sequence diversity for TALE repeat engineering.

The pool of sequence polymorphisms amongst functionally validated TALE-like repeats was further expanded by the molecular characterisation of TALE-like proteins from bacterium *Burkholderia rhizoxinica* (19–21). We demonstrated, that two of these proteins, Bat1 and Bat2, bind DNA with a sequence specificity matching the TALE code (19). Furthermore, the solved crystal structures of Bat1 (21) and AvrBs3 are strikingly similar. The structure of individual repeats is also highly conserved between TALEs and Bats with BSRs in a loop between paired α-helices. This high structural similarity is remarkable given only 32% amino acid sequence similarity across the repeat region. Since TALE and Bat repeats are functionally compatible (19) the pool of available sequences for TALE repeat redesign is greatly enlarged by the Bats.

However, residues clustered around the BSR (positions 7 to 19) are largely invariant across all known TALEs, RipTALs and Bats (19). The conservation of these residues not just between but also within these groups of TALE-likes and their position next to the binding base suggests that these residues are important for repeat structure. It seems conceivable that residues adjacent to the BSRs have a major impact on the placement of the BSR with respect to the paired base. Accordingly, these residues may also be those most interesting for engineering attempts aimed at changing DNA binding properties.

We describe here molecular characterisations of two novel proteins with weak sequence similarity to TALEs predicted from marine bacterial metagenomics data (22, 23). We refer to these predicted proteins as MOrTL1 and MOrTL2 (Marine Organism TALE-Likes) to reflect the very limited information we have regarding their provenance. We show that repeats of both MOrTLs mediate sequence-specific DNA binding in accordance with the TALE code. To gain information on the structure of MOrTL1 and MOrTL2 we built homology models that were subsequently analysed using molecular dynamics (MD) simulations. The final models show a striking structural similarity to TALE and Bat repeats. Yet MOrTL repeats bear sequence motifs unknown from TALEs, RipTALs and Bats. The MOrTLs are as distant from one another at the sequence level as they are from any of the other TALE-likes. Furthermore, sequence similarity between TALEs and both MOrTLs is low, even in the otherwise highly conserved region around the BSR. This all makes the MOrTLs a fascinating addition to the growing family of TALE-likes and the pool of functionally validated repeat sequences for TALE repeat re-engineering.

Materials and Methods

MOrTL construct creation

Genes encoding EBN1909 and ECG96326, codon optimized for *E. coli* and with additional 5' and 3' BsaI recognition sites, were synthesized (GenScript). Sequences in Figure S1. Genes were cloned into a modified pENTR D-TOPO (Life Technologies) vector rendered Golden Gate compatible with the replacement of the native gateway cassette and Att sites with a gateway cassette flanked by BsaI recognition sites with the digest-overhangs TATG-GGTG.

To create Bat1 chimeras 5-mer subunits of the synthesized MOrTL genes were PCR amplified with the primers listed in Table S2 bearing BsaI sites corresponding to Block 2 of the previously described Bat1 cloning system (19). The MOrTL blocks, along with Bat1 blocks 1 and 3-5, were assembled into either pENTR or pBT102* CACC-AAGG (see below) via BsaI cut-ligation. Chimera sequences are given in the supplementary material.

To create TALE chimeras 5-mer subunits of the synthesized MOrTL genes were PCR amplified with the primers listed in Table S2 bearing BsaI sites corresponding to the 5B level 2 repeat blocks of the designer TALE assembly toolkit as previously described (24) but using Level 2 vectors pUC57-A5-DEST and pUC57-5B-DEST instead of pUC57-AB-DEST, to allow different A5 and 5B repeat blocks to be combined. A5 and BC Blocks to target $\mathrm{BE}_{\mathrm{Bat1}}$ were made with the same TALE toolkit. A5, 5B and BC blocks were assembled together via BpiI cut-ligation into pENTR 3xHA-TALE N/C-3xFlag-NLS-STOP (19) or pBT102* TALE Δ356/+90-GFP (See below).

Protein expression and purification

Genes were transferred from pENTR into pDEST-17 using the Gateway recombinase system (Life Technologies). Proteins were expressed and purified as previously described (19). In short, *E. coli* Rosetta cells were induced at 30°C with a final concentration of 0.1 mM IPTG for 3 h. His-tagged proteins were purified by affinity chromatography with an ÄKTA Protein Purification System (GE Life Sciences) using a HisTrap TALON crude column (GE Life Science).

EMSAs

EMSAs were performed as described previously (19). Complementary pairs of labelled or corresponding unlabelled oligonucleotides were annealed. Binding reactions contained 1 pmol of labelled probe, 0 pmol, 25 pmol, 50 pmol or 200 pmol of unlabelled probe and, if not otherwise stated, 4 pmol of protein. Binding reactions were incubated at room temperature for 30 mins and resolved on a 6% native polyacrylamide Gel for one hour at 100 V, 4°C. Labelled DNA was visualized with a Typhoon FLA 9500 (GE healthcare).

E. coli repressor reporter system

The repressor reporter system we used is an adaptation of the TALE-based bacterial NOT gate created by Politz *et al.* (25), who kindly provided us with plasmids pCherry (mCherry reporter) and the TALE expression plasmid pBT102 with a designer TALE they created to target the Lac operon cloned inside.

pCherry was modified by the insertion of target boxes 3' of the Lac Operon (see Figures S6 and S7) via PCR amplification of the whole plasmid, using primers listed in Table S2. The sequences of the promoter derivatives created can be found in Figure S7.

We created a golden-gate compatible version of the *E. coli* TALE expression vector pBT102. This was done by PCR amplifying the backbone of the vector, excluding the TALE gene, and ligating this together with a PCR amplicon of a gateway cassette flanked by BsaI recognition sites with overhangs 5' TATG – 3' GGTG (pBT102* TATG-GW-GGTG; Figure S6) or 5' CACC- 3' AAGG (pBT102* CACC-GW-AAGG). pBT102* TATG-GW-GGTG was then made into a level 3 dTALE vector through the addition of several subunits via BsaI-cutligation, 5' to 3': *Δ356 TALE N-terminal region, +90 TALE C-terminal region*, *gfp* (pBT102* TALE Δ356/+90-GFP; see Figure S8).

The assay was carried out by co-transforming 0.25 μl of pCherry and 0.5 μl plasmids into 25 μl of chemically competent *E. coli* Top10 cells and plating onto LB Agar plates containing 12.5 μg/ml Kanamycin, 50 μg/ml Ampicillin and 0.1mM IPTG. The IPTG was added to prevent interference from the endogenous lac inhibitor since pCherry has a lac operator. Plates were incubated 36 hours at 37°C to achieve stationary phase colonies. Single colonies were then picked into 150 μl of liquid LB medium with the same antibiotic/IPTG concentrations as above, in wells of a 96 well Greiner plate with black sides but a transparent bottom (Vision plate, 4ttitude). Cultures were shaken 3.5 hours at 37°C, 180 rpm. OD 600 was measured in a plate reader (TECAN) as well as mCherry fluorescence with the following parameters: Excitation 587 nm, Emission 610 nm, bandwidth ± 12nm, Gain 90, Z-position 6300 μm.

Fluorescence data were divided by OD600 to correct for bacterial density and these values were used as inputs for the generation of boxplots in RStudio (v. 0.98.501).

Structure modelling

Homology models of Bat1$_{M1\ 6\text{-}10}$ and Bat1$_{M2\ 6\text{-}10}$ were built using Schrödinger Prime (version 3.5; Schrödinger, LLC, New York, NY, 2014). For both chimeras we used PDB entry 4cja as template structure for modelling the protein. The template DNA structures were mutated *in silico* using the software package 3DNA (version 2.1) (26) in order to match the optimal bases for both constructs and merged into the homology models. To investigate the quality and reliability of the generated models we conducted MD simulations of both models using the software package GROMACS (version 4.6.7) (27). The protocol that was applied to both models used the CHARMM27 all-atom force field (version 2.0) with CMAP(28, 29) and TIP3P as the water model. In order to neutralize the solvated systems water molecules were replaced by sodium as counter-ions to adjust a zero net charge. The models were energy minimized in two steps using steepest descent and subsequent conjugate gradient. A total of 50 ns were simulated for each system with a time step of 2 fs. Neighbour searching was performed every 10 steps. The PME algorithm was used for electrostatic interactions with a cut-off of 1 nm. A reciprocal grid of 72 x 64 x 72 cells was used with 4th order B-spline interpolation. A single cut-off of 1 was used for van der Waals interactions. Temperature coupling was done with the v-rescale algorithm, while the Berendsen algorithm was used for pressure coupling. The results were analysed using tools from the GROMACS package. Figures and videos were generated using VMD(30) (version 1.9.2) and R (R Core Team: A language and environment for statistical computing. 2013. http://www.r-project.org). Potential energy and RMSD plots are shown in Figures S10 and S9. Input files and parameter settings for both simulations given in supplementary data files 3-7. PDB files with the final frames of each MD simulation with and without solvent molecules are provided as supplementary data files 8-10.

Results

MOrTL1 and MOrTL2 are predicted proteins from a marine metagenomics database

The term MOrTLs is throughout used to refer to two predicted proteins from marine microbial genomic DNA, sequenced as part of the Global Ocean Sampling Expedition (22). Host organisms bearing the MOrTLs sequences were sampled from the Gulf of Mexico/Yucatan Channel and are most likely bacterial based on size filtration of the biological material that was used for recovery of DNA (0.1 to 0.8 micron) (22, 23). MOrTL 1 and 2 are encoded on two separate contigs and each contig bears an additional repeat protein ORF. Further details can be found in Figure S1; MOrTL1 is synonymous with GenBank protein ID ECG96326 and MOrTL2 with EBN91409. These sequences have been previously suggested to encode modular DNA binding repeats (20) but no functional analysis has been reported till now.

Both predicted MOrTL proteins are tandem-repeat arrays, with each repeat 33 amino acids in length (Figure 1). However, comparing consensus repeat sequences (Figure S2) shows that MOrTLs 1 and 2 are as diverged from one another as they are from all other TALE-like groups. We therefore analysed both MOrTL1 and 2 because their repeats differ considerably at the sequence level.

Database sequences are likely incomplete

MOrTL1 is formed of 8 repeats, and MOrTL2 of 10 repeats (Figure 1b). Thus, the number of repeats in both MOrTLs is fewer than in any functionally-validated natural TALE-like protein examined to date. In addition, it has been shown for TALEs and Bats that sequence divergent repeats in the N- and/or C-terminal region of the protein make a decisive contribution to DNA binding (19, 31). Such sequences may also exist in the full-length MOrTL proteins but are not found in the DNA sequences available. Indeed CDSs of both MOrTLs 1 and 2 begin in what appears to be the middle of a repeat (Figures S1d and S1h) supporting this idea. We therefore considered it likely that the reference sequences would not yield functional proteins. We nevertheless had genes encoding the predicted MOrTLs 1 and 2 synthesized. We were able to express and purify MOrTL1 from *Escherichia coli*, while MOrTL2 formed protein aggregates preventing purification (Figure S3a). MOrTL1 was tested in electrophoretic mobility shift assays (EMSAs) at a range of concentrations against a fluorescently labelled oligonucleotide probe bearing a predicted DNA binding element (BE; $\mathrm{BE}_{\mathrm{MOrTL1}}$; Figure 1) based on the TALE code (Table S1). A shift was detectable only with a MOrTL1 concentration of 822 nM or greater (Figure S3b). Such weak DNA binding is inconsistent with expectations based on other TALE-likes (19, 32). In addition laddering was observed in the gel shift indicating the formation of higher order protein-DNA complexes (Figure S3b) again inconsistent with TALE-likes, which bind their targets in a 1-to-1 ratio with high sequence specificity.

As previously mentioned, both MOrTLs 1 and 2 are found in contigs with additional, highly similar, MOrTL ORFs immediately upstream (Figures S1b and

S1f), suggestive of a larger but incompletely sequenced repeat protein in each case. The unexpectedly weak and multi-species binding behaviour of MOrTL1 may be the result of working with a protein fragment only, since array length and N- and C-terminal degenerate repeats are known to be crucial for high affinity DNA binding in other TALE-likes (4, 16, 19). Since it has been shown for other TALE-likes that in many cases repeats can be rearranged without impairing function (19, 21), we created a fusion protein combining the repeats of both reads from *EN814823.1* (*EBN19408-MOrTL2;* Figure S4). We had the *EBN19408-MOrTL2* fusion gene synthesized but were once again unable to obtain soluble protein from *E. coli* preventing functional analysis.

MOrTL repeats embedded in Bat and TALE repeat domains bind predicted BEs *in vitro* and *in vivo*

We next decided to explore a repeat domain chimera approach that has proved highly informative in the past for the functional analysis of Bat and RipTAL repeats (16, 19). We tested blocks of five repeats from the central part of each MOrTL embedded within the repeat domain of Bat1 at positions 6-10 (Bat1$_{M1\ 6\text{-}10}$ and Bat1$_{M2\ 6\text{-}10}$; Figure 2a) or in a dTALE designed to target the same DNA sequence as Bat1 (dTALE-Bat1, dTALE-Bat1$_{M1\ 6\text{-}10}$, dTALE-Bat1$_{M2\ 6\text{-}10}$). The MOrTL repeats used for each of these chimeras are depicted in Figure 1. We assembled genes encoding the desired chimeras and expressed them in *E. coli*. Proteins were purified and tested in EMSAs against predicted BEs. In each case the integrated MOrTL repeats differ in their BSR composition from the Bat1 or dTALE repeats they replace, which should lead to a modified DNA sequence preference. Cognate binding elements were predicted based on the established TALE code (Table S1). Previous work has shown that even one or two mismatches can have a serious effect on DNA binding of TALE likes (16, 20, 33). Thus if the chimeras containing five MOrTL repeats mediate clear 1-to-1 DNA binding with their predicted BEs this strongly supports the hypothesis that MOrTL repeats are actually binding their cognate DNA bases in the BEs. This is indeed what we observed. Clear single shifts, of similar intensity, were observed for every chimeric and non-chimeric TALE-like at 200 nM with its cognate BE (Figures 2b and 2c).

To study DNA recognition properties of MOrTL repeats *in vivo* we also developed an *E. coli*-based repressor reporter. This system was an adaptation of a previously described TALE-based bacterial NOT gate (25). In this system binding of a TALE-like protein at a predicted BE inserted into a constitutively active *Trc* promoter inhibits expression of the downstream fluorophore (mCherry). Thus reduction in fluorescence levels should directly relate to the strength of the interaction between the given BE and the tested TALE-like protein.

In this assay *mcherry* reporter plasmid and *TALE* expression plasmid are co-transformed into *E. coli*. Expression of both *TALE* and *mCherry* is driven from synthetic constitutive promoters. A lac operator sequence (*LacO)* is contained in the *Trc* promoter, which drives *mcherry* expression. In the original system a dTALE made to bind this operon (dTALE-LacO) shuts off mCherry expression. We modified this set up by inserting an additional BE of choice just 3’ of *LacO* (see Fig. S6). The TALE expression plasmid was also modified to allow insertion of a *bat1*, *dTALE-bat1* or *MOrTL*-chimera gene. Repression is measured relative to a negative control plasmid (*gfp*) and is used as a proxy for DNA binding. The set-up is illustrated in Figure 3a. We found that Bat1$_{M1\ 6\text{-}10}$ and Bat1$_{M2\ 6\text{-}10}$, the Bat1 derivatives bearing five MOrTL repeats, were able to repress their cognate reporters to levels similar to that of wild type Bat1 (Figure 3b). Similarly, the TALE-MOrTL1 chimera, dTALE-Bat1$_{M1\ 6\text{-}10}$, performed well, able to mediate an 11.6-fold repression of its target reporter, compared to 9.5-fold for dTALE-Bat1 wild type (Figure 3c). However, dTALE-Bat1$_{M2\ 6\text{-}10}$ was only able to mediate a weak repression, 1.6-fold relative to the GFP control. Overall, however, the strong repression observed for three out of four tested chimeras corroborate the *in vitro* binding data, supporting the hypothesis that MOrTL repeats function as TALE-like DNA binding repeats with the expected properties.

The poor performance of dTALE-Bat1$_{M2\ 6\text{-}10}$ in the *in vivo* assay (Figure 3c) is seemingly inconsistent with the results of the *in vitro* EMSA approach (Figure 2c). This may, however, relate to fundamental differences in the nature of the two approaches. In the EMSA approach the protein is added in fourfold molar excess to the perfect-match (on-target) probe DNA, whilst in the *in vivo* approach the matching DNA binding element is just one of millions of potential DNA target sites within the bacterial cell. Thus a certain degree of sequence specificity is necessary to see any repression using the *in vivo* approach. Thus if the sequence specificity of the protein is low we might expect a poor performance in the repressor assay due to off-target binding, while the EMSA is still expected to give a clear shift since competing off-targets are absent in this assay. Thus the differences we see for dTALE-Bat1$_{M2\ 6\text{-}10}$ in the *in vitro* and *in vivo* assays may relate to differences in specificity, as discussed below.

Off-target binding tests confirm sequence specificity *in vitro* and *in vivo*

To carry out EMSA competition assays we designed off-target BEs using the TALE-code (Table S1) to test if the MOrTL repeats conform to the sequence specificity expectations of TALE-like repeats. Only bases 6-10 in the BEs (corresponding to MOrTL repeats in the chimeras) were altered, while all others were kept identical. We chose the least favoured base in each case for corresponding BEs: G used for Gly at the BSR and T used for Arg, Asp or Ile at the BSR. GGTTG (BE$_{Bat1\text{-}GGTTG}$) was used for all constructs except dTALE-Bat1$_{M2\ 6\text{-}10}$, for which the off-target sequence is TTGGT (BE$_{Bat1\text{-}TTGGT}$). In the EMSA competition assays the labelled on-target probe is mixed with an excess of either on- or off-target competitor DNA to study sequence specificity of protein-DNA interactions. If MOrTL repeats are indeed sequence-specific in their DNA binding, an excess of the on-target competitor should outcompete the on-target probe, leading to a loss of shifted signal while an excess of off-target competitor should have a less pronounced impact on probe-protein interaction. As seen in Figure 4 this was indeed observed in every case except for dTALE-Bat1$_{M2\ 6\text{-}10}$ (Figure 4e). Quantifications of the intensity of shifted and free probe for these gels are shown in Figure S5. In the case of dTALE-Bat1$_{M2\ 6\text{-}10}$ the on- and off-target competition assays gave very similar levels of shift depletion indicating poor sequence specificity of the chimeric protein. This observation is consistent with the poor performance of this chimeric protein in the *in vivo* reporter (see above). In addition Bat1$_{M1\ 6\text{-}10}$ and Bat1$_{M2\ 6\text{-}10}$ (Figures 4a and 4b; S5) showed some limited shift depletion with the off-target DNA. This was however far weaker than the depletion seen in combination with the on-target DNA. However, in the case of Bat1 the shift band (Figure 4c; S5), was hardly reduced in intensity even in presence of 200-fold excess of off-target competitor DNA. It may be that, at least in this specific chimeric context, MOrTL repeats are slightly less discriminating in their DNA-binding sequence specificity than Bat1 repeats.

in vivo assays confirmed that reporters bearing off-target boxes were not repressed by the tested DNA binding proteins (Figure 6). This confirms that the strong repression seen in on-target boxes (Figure 3) is dependent on the five MOrTLs repeats pairing base-specifically with the five cognate nucleotides of their BEs. Overall these data show that MOrTL repeats are able to mediate DNA binding in accordance with the TALE code.

Functional conservation is likely a consequence of structural conservation

We were able to show that MOrTL1 and 2 repeats mediate DNA binding with a sequence specificity matching the TALE code when embedded in a Bat1 repeat array. The same could be confirmed for MOrTL1 in a TALE repeat array. DNA binding properties are thus conserved among repeats of TALEs, RipTALs, Bats, MOrTL1 and MOrTL2. They are also functionally interchangeable as we have shown in this study and previous studies showing that chimeric TALE repeat arrays containing RipTAL and Bat repeats are functional (16, 19). We suggest that this functional conservation justifies the use of the term TALE-likes to refer to proteins bearing a tandem array of 33-35 amino acid repeats mediating 1-to-1 DNA binding with position 13 determining DNA binding specificity at least largely in accordance with the conserved TALE code. This functional conservation is suggestive also of a structural conservation allowing each repeat to contact a single nucleotide and for position 13 to mediate base specific interactions. Yet the sequence similarity between TALE-like repeats can be below 40% (Figure S2). Structural conservation can hardly be taken for granted for such dissimilar proteins.

There is however, already evidence in support of a high degree of structural similarity among TALEs and Bats. Crystal structures for Bat1, with and without its DNA target, have been solved (21) and are very similar to analogous structures for TALEs PthXo1, AvrBs3 and dTALE dHax3 (34–36). All examples possess the same super helical structure that contracts tightly around the B-form DNA double helix. Additionally, position 13 residues are located in loops that point into the major groove of the target DNA. Assuming these features form structural prerequisites for the DNA-binding properties of TALE-likes, we expect the MOrTL repeats, for which no experimentally derived structure is available yet, to adopt a similar structure. To support this hypothesis, we generated models of the functionally validated chimeras $\text{Bat1}_{\text{M1 6-10}}$ and $\text{Bat1}_{\text{M2 6-10}}$. Both models show structural properties similar to those described earlier for TALE-like repeats (Figure 6; supplementary data files 1-2). While these homology models resulted in a plausible protein structure, they do not provide functional information. To get information about the stability of the predicted protein-DNA interaction interfaces over time we conducted molecular dynamics (MD) simulations of the MOrTL homology models bound to DNA. Both simulations revealed highly stable complexes between the proteins and their target DNA, seen in the values for atomic distances between protein and DNA partners (Figure S9) and potential free energy of the complexes (Figure S10). Measuring base-BSR distances during MD simulations showed that interactions were stable and comparable for Bat1 and MOrTL repeats (Tables S3 and S4). This is in line with our *in vitro* and *in vivo* DNA-binding data. Taken together, it seems likely that, despite amino acid sequence similarities below 40%, repeats of TALEs, RipTALs, Bats, MOrTL1 and MOrTL2, adopt similar structures, facilitating a conserved DNA-binding mechanism. We suggest therefore that the TALE-like designation should not refer to proteins conforming to a particular amino-acid sequence but to any proteins bearing a repeat array conserved with those of TALEs both functionally and structurally.

MOrTL repeats differ from all other TALE-likes in residues around the BSR

The structural similarities between TALE-like repeats are surprising considering the low sequence similarity. To illustrate the variation among TALE-like repeats we created amino acid alignments of core repeats from representatives of each TALE-like group so far described (13 TALEs, 5 RipTALs, Bat1/Bat2 and both MOrTLs; Figure 7; see Table S5 for list of TALE-likes used). These alignments show first that TALE repeats are somewhat exceptional for their very low sequence diversity. In all other TALE-like groups more than one third of repeat positions are highly polymorphic. More specifically TALEs are highly polymorphic only at five positions: 4, 12, 13, 32 and 35. Bat and RipTAL repeats, in contrast, are polymorphic across much of the long helix (positions 15-32) and inter-repeat loop (positions 33-2) regions.

Percentage conservation at each repeat position was calculated separately for each TALE-like group. We then took the average conservation for each repeat position across the five TALE-like groups, thus giving equal weight to the conservation found in each of TALEs, RipTALs, Bats, MOrTL1 and MOrTL2. This average within-group conservation is shown in the line graph underneath the alignments (Figure 7b) With the exception of residue 12 and the BSR, conservation is 90% or more across residues 5-20. The conservation of these residues is logical considering their proximity to the crucial BSR position (see Figure 6).

Examining the sequences for each TALE-like group reveals that some of these residues are conserved not just within but also between groups. Between TALEs, RipTALs and Bats there is almost no polymorphism at positions 7, 9, 10, 14, 15 and 17-19. This is limiting for repeat engineering efforts because these residues cluster around the BSR and are therefore especially likely to exert significant influence over DNA binding properties. Furthermore, any effort to use natural diversity to create sequence-diverse TALE-likes less prone to repeat recombination (37, 38) will be held back by the lack of sequence diversity in this region.

Repeats of MOrTL1 and 2, however, have a number of residues in this region around the BSR not found in repeats of any other TALE-like (Figure 7a; red lettering). At positions, 10, 15, and 17-19 there is little to no sequence diversity among TALE-likes except that found in MOrTLs 1 and 2. Thus MOrTLs 1 and 2 make a substantial contribution to the sequence diversity of TALE-like repeats in residues around the BSR; a contribution that may prove crucial for future efforts to engineer TALE-like repeat arrays.

Discussion

We have been able to show that repeats from MOrTL1 and 2 (Figure 1) recognise DNA with a sequence specificity matching the TALE-code (Figures 2-5), despite more than 60% of residues differing from previously characterised TALE-like repeats on average (Figure S2). Blocks of five MOrTL repeats, embedded in Bat1 or designer TALE repeat arrays, were competent to discriminate TALE-code-predicted on-target BEs (Figures 2b, 2c, and 3), from off-target sequences (Figures 4 and 5). Sequence specificity was highly stringent for all tested constructs except for dTALE-Bat1$_{M2\ 6\text{-}10}$. The superior performance of the Bat1-MOrTL2 chimera compared to the dTALE-MOrTL2 chimera suggests functional differences in the MOrTL2 repeats used for the two chimeras, or some functional interference between the given TALE and MOrTL2 repeats. Since all on-target and mismatch sequences were predicted based on the TALE-code (Table S1) our results overall suggest that the MOrTL repeats within the tested chimeras were able to mediate DNA recognition with a sequence specificity matching that of previously characterised TALE-likes.

Based on previous knowledge of TALE-like repeat-DNA interactions our observations strongly suggest direct binding of the cognate DNA bases by MOrTL repeats. An alternative explanation is that the MOrTL repeats passively tolerate on-target bases but sterically clash with off-target bases, similar to the G$_{SL}$ repeats of some TALEs (6, 35). However, such TALE repeats are relatively un-discriminating (39), inconsistent with our observation of high stringency of DNA target recognition for TALE- and Bat1-MOrTL chimeras (Figures 4 and 5). Furthermore, homology models combined with MD analysis suggest that MOrTL and Bat1 repeats are conserved in overall structure (Figures 6, S9 and S10) and in proximity between BSRs and cognate bases (Tables S3 and S4). Our findings support the hypothesis that despite low sequence similarity the repeat arrays of TALEs, RipTALs, Bats and both MOrTLs are able to bind DNA with a largely conserved mechanism.

The repeats of MOrTLs 1 and 2 substantially increase the total variability of TALE-like repeats (Figure 7). Because MOrTL repeats have unique sequences in the otherwise highly conserved BSR cluster of residues (Figure 7a; red lettering) they represent a qualitatively different variability-contribution from that of all other functionally validated groups. As we have demonstrated, MOrTL1 repeats seem to be readily compatible with TALE and Bat repeats whilst MOrTL2 repeats are at least compatible with Bat repeats (Figures 2-4). This suggests that MOrTL repeats or repeat subdomains can be combined with other TALE-like repeats for the creation of novel TALE-like repeat arrays. The sequence diversity of MOrTL repeats thus presents a unique and potentially valuable contribution to the pool of available sequences for TALE-repeat engineering.

Even without using MOrTL repeats directly in engineered TALE-like repeats the comparison of TALE-like repeat sequences may improve understanding of TALE-like repeat structure and its relation to DNA binding properties. This improved understanding will in turn benefit TALE repeat engineering efforts. Till now descriptions of the roles of different TALE or TALE-like repeat residues, apart from the RVD, have come only from predictions based on structural models (21, 34, 35). Hypotheses about residue roles remain largely untested in a wet lab setting though molecular dynamics simulations have provided some insights (40). Using data from the natural experiment of evolution can help answer some questions or provide a starting point for hypothesis testing, complementing other methods. For example, positively charged residues Lys16 and Gln17 of TALE repeats have been suggested to form an electropositive stripe along the TALE superhelix and to form hydrogen bonds to the phosphate backbone of the DNA (34). In Bat and MOrTL repeats, position 16 is generally occupied by an uncharged residue speaking against the importance of Lys16 for repeat array function, unless the effect is elsewhere compensated. Gln17 in contrast is conserved across all groups, except for MOrTL2 where a Lysine is found at this position. This would support an important role for the electropositive strip formed from positive residues at position 17 only. To take another example, it seems logical that the highly conserved double Glycine at positions 14-15 in TALEs, RipTALs, Bats and MOrTL2 is necessary for the flexibility of the repeat loop. MOrTL1 repeats have either Alanine or Serine at position 15; does this effect flexibility of the BSR loop and consequently the interaction between BSR and base? Other positions are surprisingly conserved. Leu29 is one of only three residues highly conserved between all the TALE-like groups. Till now the only function attributed to this residue is a role in hydrophobic interactions that bring together neighbouring repeats as the TALE structure contracts upon DNA binding (21), yet other hydrophobic residues seem not to be tolerated at this position. Since MOrTL repeats are polymorphic at otherwise highly conserved positions in all other TALE-likes they may be especially useful for such comparative approaches to understanding the interplay of sequence, structure and function in the TALE-like repeat.

MOrTL1 and 2 also make useful outgroups for asking questions about the evolutionary history of other TALE-likes. As mentioned previously TALE and RipTAL repeats are conserved at many positions, while the Bats show greater sequence divergence. However some residues around the BSR are conserved among TALE, RipTAL and Bat repeats (Figure 7). So far it has remained an open question as to whether these sequence similarities are an indicator of common evolutionary origin or are rather the result of convergent evolution of similar proteins with a constrained sequence-structure space. The diversity of MOrTL1 repeat sequences in this region shows that several alternative sequences are tolerated within this structure. Therefore, that the TALEs, RipTALs and Bats are conserved in this region suggests that the share a common ancestor. Another insight can be gained from comparing the level of diversity found among TALE repeats to those of other TALE-likes. RipTAL, Bat and MOrTL repeats are highly diverse across much of the long-helix and inter-repeat loop region, while there is little repeat diversity when comparing repeats of all known TALEs (Figure 7). Furthermore, there is almost no non-RVD repeat polymorphism between the repeats of any single TALE repeat array (for example, compare repeat array alignments of TALE AvrBs3 and RipTAL Brg11, Figure S12). The majority of sequence diversity among TALE repeat comes from positions polymorphic between repeat arrays but conserved within each particular array, contrasting with the other TALE-likes. The within-array sequence conservation of TALE repeats could be the result of a selective pressure against sequence diversification. Such hypotheses required further investigation and we hope that the MOrTLs will prove useful as an outgroup for future studies into sequence-function relations and into the evolution of the TALE-likes.

Considering the full sequence diversity of TALE-like repeats may even assist with the identification of future functional homologs and, through these, novel evolutionary insights. Whilst TALE repeats are highly conserved across most positions only three residues are conserved across all TALE-like groups (Figure 7b, yellow shaded): Val7, Gly14 and Leu29. That these positions are so highly conserved suggests functional importance as discussed above, but in addition these conserved residues allow us to provide a broad sequence definition of TALE-like repeats as conforming to the motif $X_6\mathbf{V}X_6\mathbf{G}X_{13}\mathbf{L}X_{4\text{-}6}$. This motif may be useful as a basis for identifying additional TALE-likes from genome sequences.

By demonstrating that MOrTL repeats mediate the same DNA binding behaviour as other TALE-like repeats (Figures 2-4) we have gained insights into the nature of the whole TALE-like family and we hope this will enable further research into the distribution and functions of these fascinating DNA binding proteins.

Funding

This work was supported by Deutsche Forschungsgemeinschaft (SFB924); and the Two Blades foundation.

Conflict of interest statement. T.L. is a partial owner of a patent application regarding the use of TALEs.

Acknowledgements

We would like to thank Prof. S. Yooseph at the J. Craig Venter Institute, USA for assistance in accessing metadata relating to the MOrTLs sequence data. We would like to thank Prof. B. Pfleger at the University of Wisconsin-Madison for provision of constructs used to create the *E. coli* repressor reporter.

Author contributions

O.D.L. and C.W. conceived the study in consultation with T.L., and designed and carried out DNA binding experiments. P.T. and J.K. designed and carried out modelling and MD simulations in consultation with O.K. O.D.L. wrote the manuscript with input from all other authors.

References:

1. Szurek, B., Marois, E., Bonas, U. and Van den Ackerveken, G. (2001) Eukaryotic features of the *Xanthomonas* type III effector AvrBs3: protein domains involved in transcriptional activation and the interaction with nuclear import receptors from pepper. *Plant J.*, **26**, 523–534.

2. Kay, S., Hahn, S., Marois, E., Hause, G. and Bonas, U. (2007) A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science*, **318**, 648–651.

3. Römer, P., Hahn, S., Jordan, T., Strauss, T., Bonas, U. and Lahaye, T. (2007) Plant pathogen recognition mediated by promoter activation of the pepper *Bs3* resistance gene. *Science*, **318**, 645–648.

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

5. Moscou, M.J. and Bogdanove, A.J. (2009) A simple cipher governs DNA recognition by TAL effectors. *Science*, **326**, 1501.

6. de Lange, O., Binder, A. and Lahaye, T. (2014) From dead leaf, to new life: TAL effectors as tools for synthetic biology. *Plant J.*, **78**, 753–771 (2014).

7. Miller, J.C., Tan, S., Qiao, G., Barlow, K. a, Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J., et al. (2011) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.*, **29**, 143–148.

8. Morbitzer, R., Römer, P., Boch, J. and Lahaye, T. (2010) Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 21617–21622.

9. Konermann, S., Brigham, M.D., Trevino, A.E., Hsu, P.D., Heidenreich, M., Cong, L., Platt, R.J., Scott, D. a, Church, G.M. and Zhang, F. (2013) Optical control of mammalian endogenous transcription and epigenetic states. *Nature*, **500**, 472–476.

10. Meckler, J.F., Bhakta, M.S., Kim, M.-S., Ovadia, R., Habrian, C.H., Zykovich, a., Yu, a., Lockwood, S.H., Morbitzer, R., Elsaesser, J., et al. (2013) Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Res.*, **41**, 4118-4128

11. Miller, J.C., Zhang, L., Xia, D.F., Campo, J.J., Ankoudinova, I. V, Guschin, D.Y., Babiarz, J.E., Meng, X., Hinkley, S.J., Lam, S.C., et al. (2015) Improved

specificity of TALE-based genome editing using an expanded RVD repertoire. *Nat. Methods*, 10.1038/nmeth.3330.

12. Yang, J., Zhang, Y., Yuan, P., Zhou, Y., Cai, C., Ren, Q., Wen, D., Chu, C., Qi, H. and Wei, W. (2014) Complete decoding of TAL effectors for DNA recognition. *Cell Res.*, **24**, 628-631.

13. Sakuma, T., Ochiai, H., Kaneko, T., Mashimo, T., Tokumasu, D., Sakane, Y., Suzuki, K., Miyamoto, T., Sakamoto, N., Matsuura, S., et al. (2013) Repeating pattern of non-RVD variations in DNA-binding modules enhances TALEN activity. *Sci. Rep.*, **3**, 3379.

14. Schornack, S., Meyer, A., Römer, P., Jordan, T. and Lahaye, T. (2006) Gene-for-gene-mediated recognition of nuclear-targeted *avrbs3*-like bacterial effector proteins. *J. Plant Physiol.*, **163**, 256–72.

15. Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., Billault, a, Brottier, P., Camus, J.C., Cattolico, L., et al. (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature*, **415**, 497–502.

16. de Lange, O., Schreiber, T., Schandry, N., Radeck, J., Braun, K.H., Koszinowski, J., Heuer, H., Strauß, A. and Lahaye, T. (2013) Breaking the DNA-binding code of *Ralstonia solanacearum* TAL effectors provides new possibilities to generate plant resistance genes against bacterial wilt disease. *New Phytol.*, **199**, 773–786.

17. Heuer, H., Yin, Y.-N., Xue, Q.-Y., Smalla, K. and Guo, J.-H. (2007) Repeat domain diversity of avrBs3-like genes in *Ralstonia solanacearum* strains and association with host preferences in the field. *Appl. Environ. Microbiol.*, **73**, 4379–4384.

18. Li, L., Atef, A., Piatek, A., Ali, Z., Piatek, M., Aouida, M., Sharakuu, A., Mahjoub, A., Wang, G., Khan, S., et al. (2013) Characterization and DNA-binding specificities of *Ralstonia* TAL-like effectors. *Mol. Plant*, **6**, 1318–1330.

19. de Lange, O., Wolf, C., Dietze, J., Elsaesser, J., Morbitzer, R. and Lahaye, T. (2014) Programmable DNA-binding proteins from *Burkholderia* provide a fresh perspective on the TALE-like repeat domain. *Nucleic Acids Res.*, **42**, 7436-7449.

20. Juillerat, A., Bertonati, C., Dubois, G., Guyot, V., Thomas, S., Valton, J., Beurdeley, M., Silva, G.H., Daboussi, F. and Duchateau, P. (2014) BurrH: a new modular DNA binding protein for genome engineering. *Sci. Rep.*, **4**, 1–6.

21. Stella, S., Molina, R., López-Méndez, B., Juillerat, A., Bertonati, C., Daboussi, F., Campos-Olivas, R., Duchateau, P. and Montoya, G. (2014) BuD, a helix-loop-helix DNA-binding domain for genome modification. *Acta Crystallogr. D. Biol. Crystallogr.*, **70**, 2042–2052.

22. Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. a., Hoffman, J.M., Remington, K., et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, 0398–0431.

23. Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J. a., Heidelberg, K.B., Manning, G., Li, W., et al. (2007) The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biol.*, **5**, 0432–0466.

24. Morbitzer, R., Elsaesser, J., Hausner, J. and Lahaye, T. (2011) Assembly of custom TALE-type DNA binding domains by modular cloning. *Nucleic Acids Res.*, **39**, 5790–5799.

25. Politz, M.C., Copeland, M.F. and Pfleger, B.F. (2013) Artificial repressors for controlling gene expression in bacteria. *Chem. Commun. (Camb).*, **49**, 4325–4327.

26. Lu, X.-J. and Olson, W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, **3**, 1213–1227.

27. Hess, B., Kutzner, C., Van Der Spoel, D. and Lindahl, E. (2008) GRGMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435–447.

28. Mackerell, A.D., Jr, Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, **102**, 3586–3616.

29. Mackerell, A.D., Feig, M. and Brooks, C.L. (2004) Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulation. *J. Comput. Chem.*, **25**, 1400–1415.

30. Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: Visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.

31. Gao, H., Wu, X., Chai, J. and Han, Z. (2012) Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res.*, **2**, 1–5.

32. Römer, P., Strauss, T., Hahn, S., Scholze, H., Morbitzer, R., Grau, J., Bonas, U. and Lahaye, T. (2009) Recognition of AvrBs3-like proteins is mediated by specific binding to promoters of matching pepper *Bs3* alleles. *Plant Physiol.*, **150**, 1697–1712.

33. Strauss, T., van Poecke, R.M.P., Strauss, A., Römer, P., Minsavage, G. V, Singh, S., Wolf, C., Strauss, A., Kim, S., Lee, H.-A., et al. (2012) RNA-seq pinpoints a

Xanthomonas TAL-effector activated resistance gene in a large-crop genome. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 19480–19485.

34. Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J.-K., Shi, Y. and Yan, N. (2012) Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science*, **335**, 720–723.

35. Mak, A.N.-S., Bradley, P., Cernadas, R.A., Bogdanove, A.J. and Stoddard, B.L. (2012) The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science (80-.).*, **335**, 716–719.

36. Stella, S., Molina, R., Yefimenko, I., Prieto, J., Silva, G., Bertonati, C., Juillerat, A., Duchateau, P. and Montoya, G. (2013) Structure of the AvrBs3-DNA complex provides new insights into the initial thymine-recognition mechanism. *Acta Crystallogr. D. Biol. Crystallogr.*, **69**, 1707–16.

37. Holkers, M., Maggio, I., Liu, J., Janssen, J.M., Miselli, F., Mussolino, C., Recchia, A., Cathomen, T. and Gonçalves, M. a F. V (2012) Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic Acids Res.*, **41**, e63.

38. Lau, C.-H., Zhu, H., Tay, J.C.-K., Li, Z., Tay, F.C., Chen, C., Tan, W.-K., Du, S., Sia, V.-K., Phang, R.-Z., et al. (2014) Genetic rearrangements of variable di-residue (RVD)-containing repeat arrays in a baculoviral TALEN system. *Mol. Ther. Methods Clin. Dev.*, **1**, 14050.

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

40. Wan, H., Hu, J.-P., Li, K.-S., Tian, X.-H. and Chang, S. (2013) Molecular dynamics simulations of DNA-free and DNA-bound TAL effectors. *PLoS One*, **8**, e76045.

Figure 1

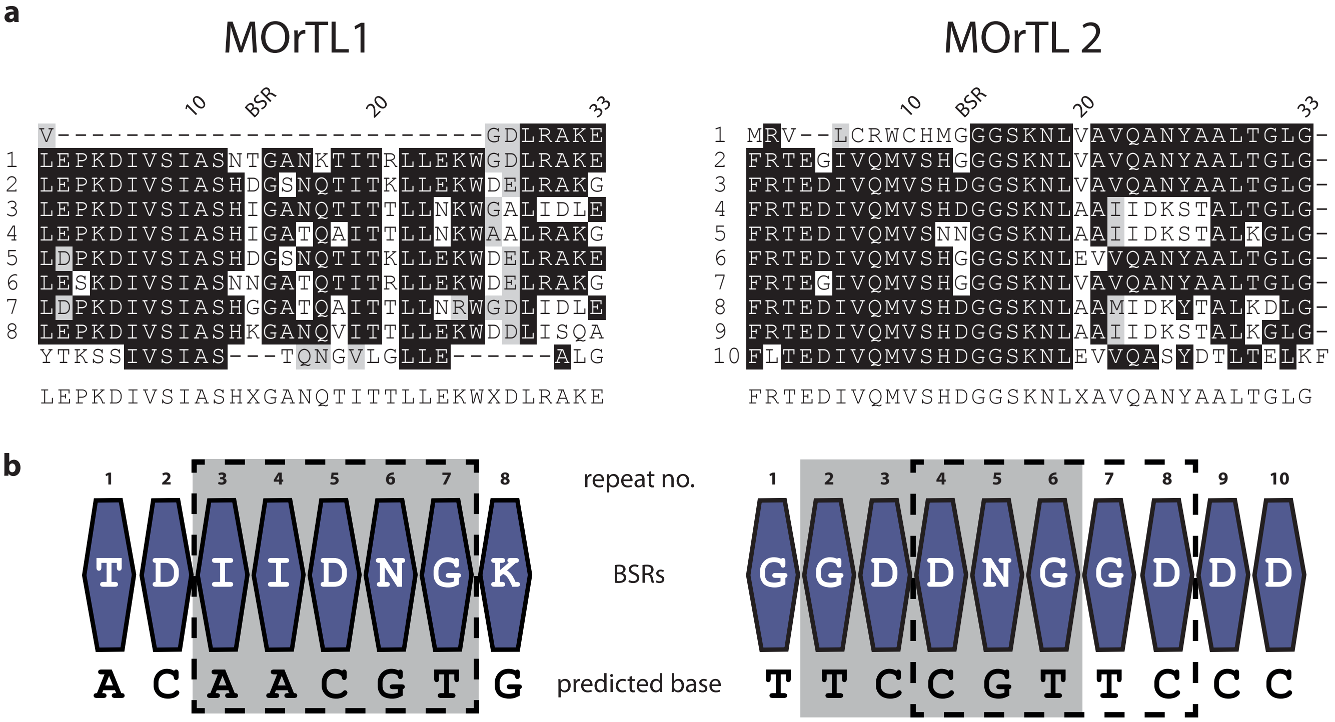

a
MOrTL1
10
BSR
20
33
V------------------------GDLRAKE
1 LEPKDIVSIASNTGANKTITRLLEKWGDLRAKE
2 LEPKDIVSIASHDGSNQTITKLLEKWDELRAKG
3 LEPKDIVSIASHIGANQTITTLLNKWGALIDLE
4 LEPKDIVSIASHIGATQAITTLLNKWAALRAKG
5 LDPKDIVSIASHDGSNQTITKLLEKWDELRAKE
6 LESKDIVSIASNNGATQTITRLLEKWDELRAKG
7 LDPKDIVSIASHGGATQAITTLLNRWGDLIDLE
8 LEPKDIVSIASHKGANQVITTLLEKWDDLISQA
YTKSSIVSIAS---TQNGVLGLLE------ALG
LEPKDIVSIASHXGANQTITTLLEKWXDLRAKE
MOrTL 2
10
BSR
20
33
1 MRV--LCRWCHMGGGSKNLVAVQANYAALTGLG-
2 FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG-
3 FRTEDIVQMVSHDGGSKNLVAVQANYAALTGLG-
4 FRTEDIVQMVSHDGGSKNLAAIIDKSTALTGLG-
5 FRTEDIVQMVSNNGGSKNLAAIIDKSTALKGLG-
6 FRTEDIVQMVSHGGGSKNLEVVQANYAALTGLG-
7 FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG-
8 FRTEDIVQMVSHDGGSKNLAAMIDKYTALKDLG-
9 FRTEDIVQMVSHDGGSKNLAAIIDKSTALKGLG-
10 FLTEDIVQMVSHDGGSKNLEVVQASYDTLTELKF
FRTEDIVQMVSHDGGSKNLXAVQANYAALTGLG
b
1 2 3 4 5 6 7 8
T D I I D N G K
A C A A C G T G
repeat no.
BSRs
predicted base
1 2 3 4 5 6 7 8 9 10
G G D D N G G D D D
T T C C G T T C C C

Figure 1

Amino acid sequences and cartoon graphics of predicted proteins MOrTL1 (ECG96326) and MOrTL2 (EBN19409). (a) The amino acid sequences are given as a series of aligned tandem repeats prepared with Boxshade. Repeat positions 10, 20 and 33 are indicated above each alignment as is residue 13, designated BSR based on our assumption that this is the base specifying residue. Core repeats are numbered down the left-hand side in each case, excluding those repeats less than 50% identical to the consensus, shown below. **(b)** Cartoon representations of the two proteins with core repeats (dark blue polygons) ordered N to C terminal with BSRs indicated (white font). Predicted binding bases are indicated under repeats, with predictions made using the TALE code. The MOrTL repeats used for chimera creation are indicated with grey shading (Bat1 chimeras) or a dashed-line box (dTALE chimeras).

Figure 2

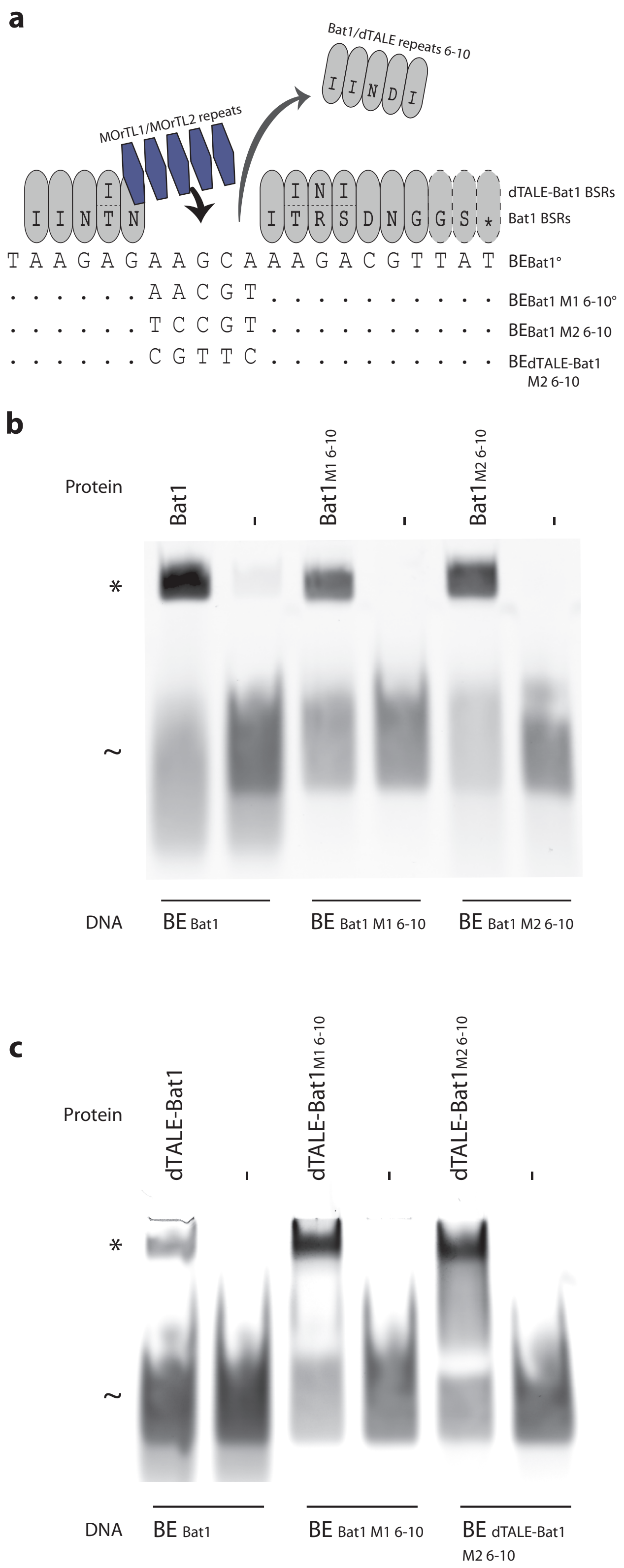
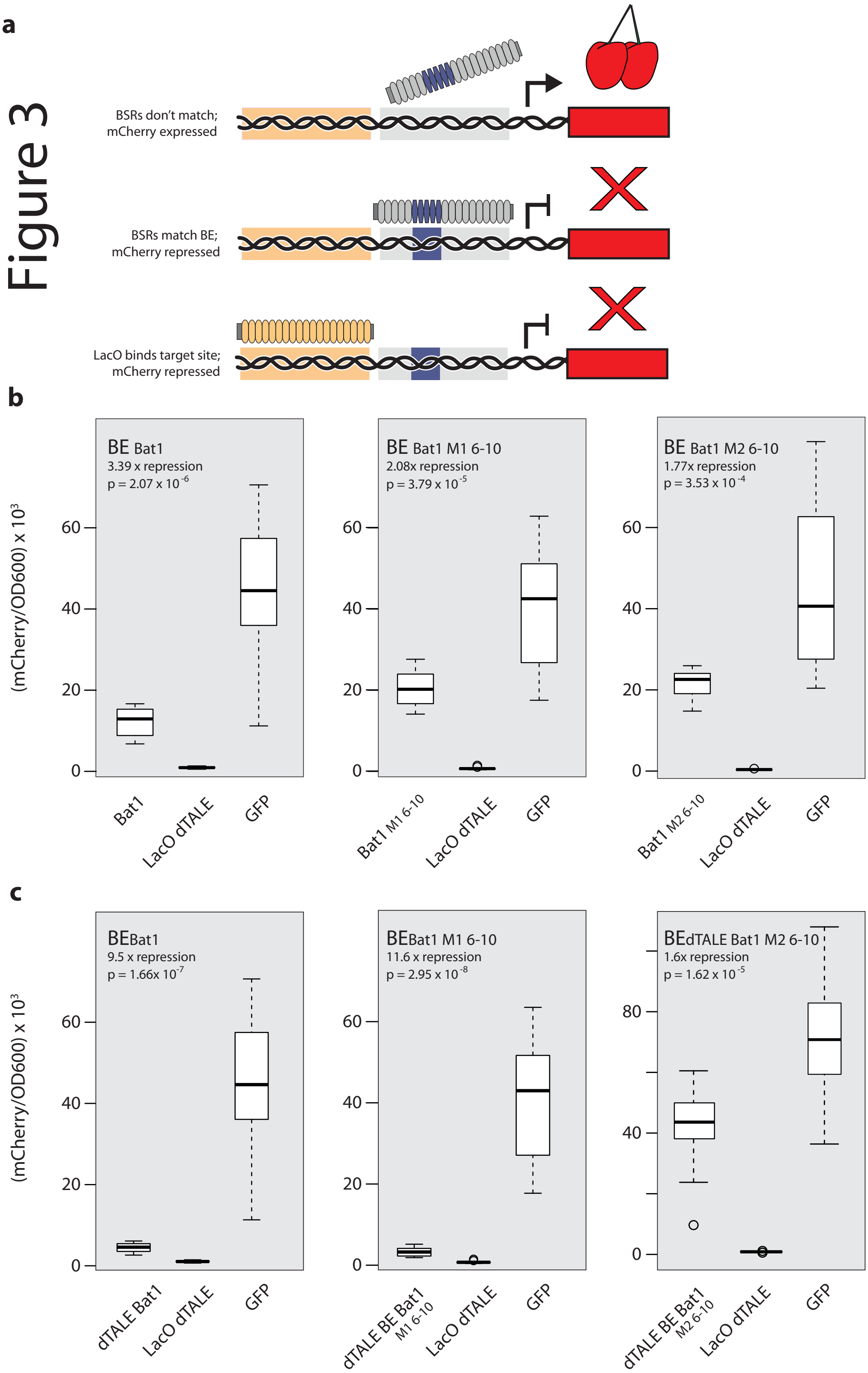


Figure 2

Bat1-MOrTL and TALE-MOrTL chimera proteins bind predicted target sequences *in vitro*. (a) Schematic display of chimeras containing five MOrTL1 or MOrTL2 repeats (dark blue polygons) in place of repeats 6-10 of Bat1 or dTALE-Bat1 (grey ovals). BSRs of the Bat1 repeats are given in each case, with an asterisk (*) for repeat 20 of Bat1 which lacks an amino acid at position 13, in reference to the consensus sequence. Where dTALEBat1 BSRs differ from Bat1 they are given above the dotted line in the relevant repeat. The last three repeats of Bat1 are outlined with dashes to reflect that dTALE-Bat1 is three repeats shorter. Binding elements (BEs) for each TALE-like chimera were predicted using the TALE code and are given below the cartoon display with dots indicating identical bases. Circles (°) indicate BEs used for more than one protein (BE$_{\text{Bat1}}$ for Bat1 and dTALE-Bat1; BE$_{\text{Bat1 M1 6-10}}$ for Bat1$_{\text{M1 6-10}}$ and dTALE-Bat1$_{\text{M1 6-10}}$. Electrophoretic mobility shift assays **(b, c)** were carried out using 5' Cy5-labelled double-stranded DNA probes at a final concentration of 50 nM and 200 nM for all proteins indicated. Shifted bands corresponding to the DNA:protein complexes are indicated with asterisks (*) and free probes with tildes (~). Each probe (DNA) was incubated in presence (+) or absence (-) of its cognate protein and run in a 6% polyacrylamide gel.

Figure 3



An *in vivo* reporter confirms that MOrTL repeats recognize predicted binding targets. (a) Schematic display of repressor assay: mCherry reporter (red symbols) and expression plasmids encoding TALE-likes are co-transformed into *E. coli*. TALE-like chimeras consist of TALE/Bat-repeats (grey ovals) and MOrTL-repeats (dark blue ovals). If the TALE-like is able to bind the given BE and repress the mCherry promoter (DNA black, BE grey/dark blue rectangle) this is observed as a reduction in mCherry fluorescence (red cherries). A dTALE that binds the Lac Operon (LacO) within the mCherry promoter provides a positive control for each reporter (BE shown with orange box). **(b)** Box and whisker plots showing mCherry fluorescence values for Bat1, Bat1$_{M1\ 610}$ and Bat1$_{M2\ 6\text{-}10}$ tested against reporters bearing corresponding BEs (designation across the top of each plot), normalized to cell density (OD600) and compared to positive (LacO TALE) and negative (GFP) control expression plasmids. Fold repression, based on median values, and p-values of a 2-tailed t-test with unequal variances comparing test and GFP samples are given in the top left corner of each plot. **(c)** dTALE-Bat1 and its two MOrTL chimera derivatives tested in the same system. Bat1 and dTALE-Bat1 have the same binding elements, as do Bat1$_{M1\ 610}$ and dTALE-Bat1$_{M1\ 610}$, so the same reporters are used for these constructs and the corresponding data for LacO and GFP are shown in both sets of box and whisker plots. N=16 in every case.

Figure 4

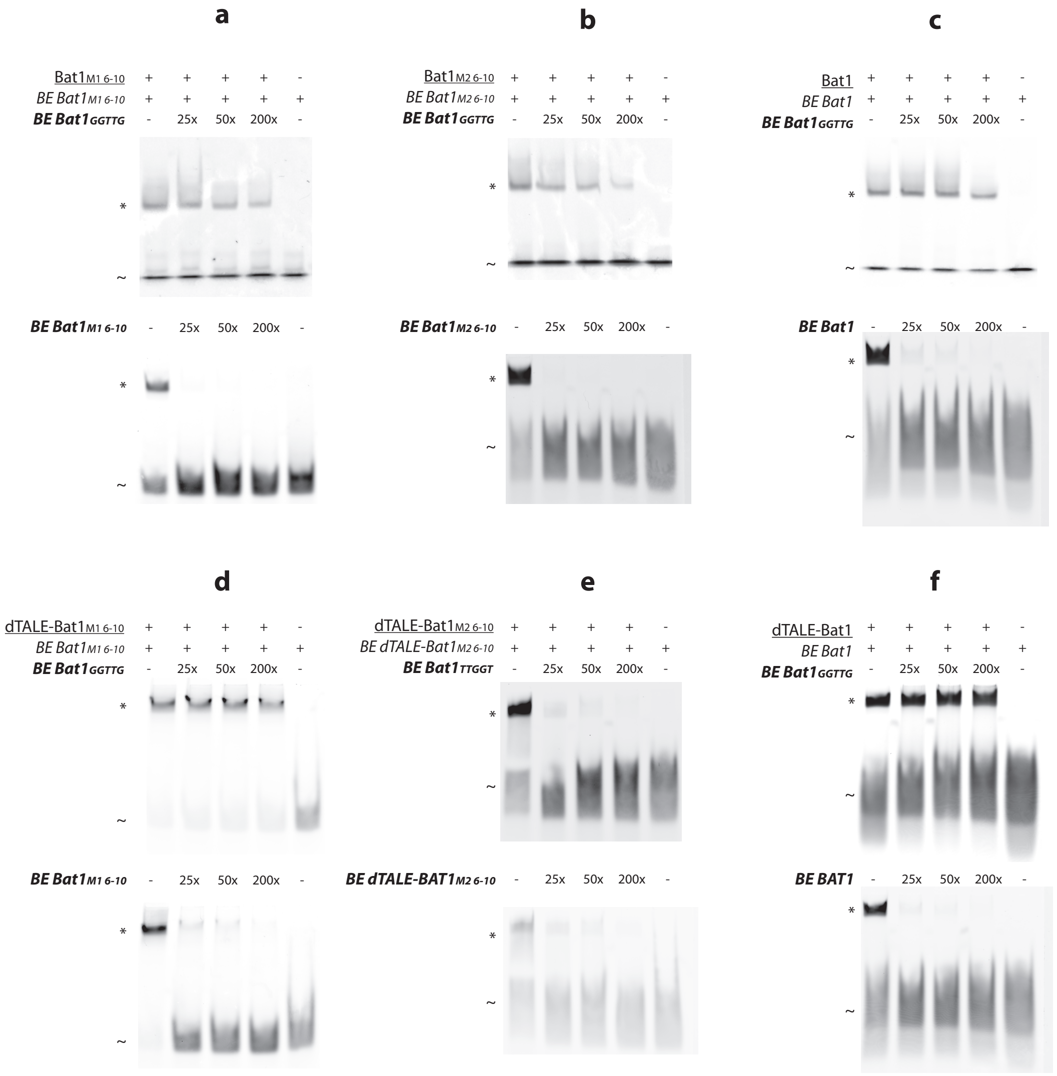

a
Bat1M1 6-10 + + + + -
BE Bat1M1 6-10 + + + + +
BE Bat1GGTTG - 25x 50x 200x -
*
~
BE Bat1M1 6-10 - 25x 50x 200x -
*
~
b
Bat1M2 6-10 + + + + -
BE Bat1M2 6-10 + + + + +
BE Bat1GGTTG - 25x 50x 200x -
*
~
BE Bat1M2 6-10 - 25x 50x 200x -
*
~
c
Bat1 + + + + -
BE Bat1 + + + + +
BE Bat1GGTTG - 25x 50x 200x -
*
~
BE Bat1 - 25x 50x 200x -
*
~
d
dTALE-Bat1M1 6-10 + + + + -
BE Bat1M1 6-10 + + + + +
BE Bat1GGTTG - 25x 50x 200x -
*
~
BE Bat1M1 6-10 - 25x 50x 200x -
*
~
e
dTALE-Bat1M2 6-10 + + + + -
BE dTALE-Bat1M2 6-10 + + + + +
BE Bat1TTGGT - 25x 50x 200x -
*
~
BE dTALE-BAT1M2 6-10 - 25x 50x 200x -
*
~
f
dTALE-Bat1 + + + + -
BE Bat1 + + + + +
BE Bat1GGTTG - 25x 50x 200x -
*
~
BE BAT1 - 25x 50x 200x -
*
~

Figure 4

Bat1- and TALE-MOrTL chimeras tested for sequence specificity using EMSA competition assays. Purified protein was incubated with 5 nM 5′Cy5 labelled double-stranded DNA probes. For the competition tests, the concentration of unlabelled competitor DNA was 25-200x that of the labelled probe. The off-target sequence was designed to bear mismatch bases for repeats 6-10 of each construct based on the TALE code ($\mathrm{BE}_{\mathrm{Bat1\ TTGGT}}$ for dTALE-$\mathrm{Bat1}_{\mathrm{M2\ 6\text{-}10}}$ and $\mathrm{BE}_{\mathrm{Bat1\ GGTTG}}$ for all other constructs). Proteins used were **(a)** $\mathrm{Bat1}_{\mathrm{M1\ 6\text{-}10}}$, **(b)** $\mathrm{Bat1}_{\mathrm{M2\ 6\text{-}10}}$, **(c)** Bat1, **(d)** dTALE-$\mathrm{Bat1}_{\mathrm{M1\ 6\text{-}10}}$, **(e)** dTALE-$\mathrm{Bat1}_{\mathrm{M2\ 6\text{-}10}}$ and **(f)** dTALE-Bat1. In each case the designation of the protein used is underlined, the probe italicized and the competitor bold and italicized. DNA-protein complexes and free probes are indicated with asterisks (*) and tildes (~), respectively. Plus/minus symbols indicate presence or absence of the respective DNA or protein.

Figure 5

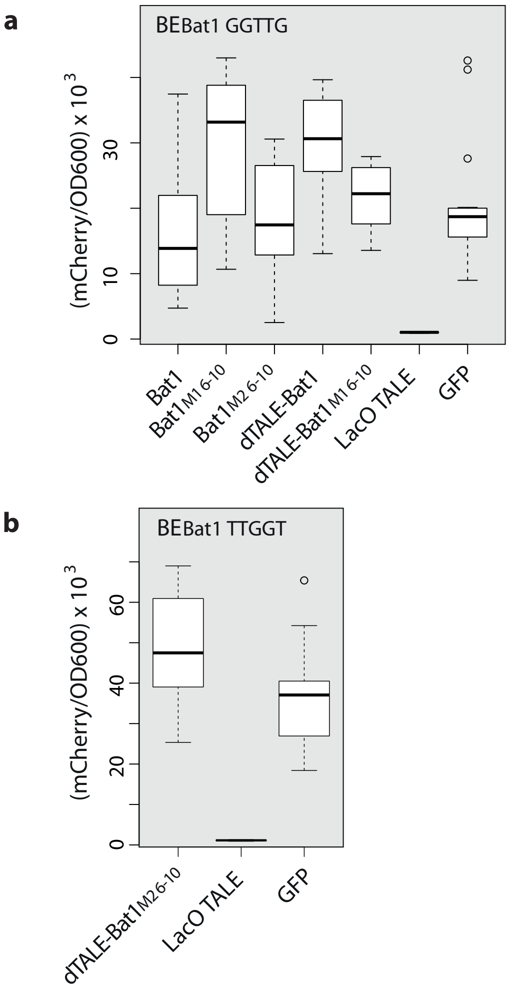


Figure 5

Bat1- and TALE-MOrTL chimeras do not activate off-target reporters *in vivo*. Off-target reporters were created with mismatch bases for repeats 6-10 of each construct based on the TALE code. These were tested against cognate TALE-likes and chimeras: **(a)** $\mathrm{BE}_{\mathrm{Bat1\ GGTTG}}$, all constructs except dTALE-$\mathrm{Bat1}_{\mathrm{M2\ 6\text{-}10}}$, **(b)** $\mathrm{BE}_{\mathrm{Bat1\ TTGGT}}$ with dTALE-$\mathrm{Bat1}_{\mathrm{M2\ 6\text{-}10}}$. Box and whisker plots show mCherry fluorescence values normalized to cell density (OD600) and compared to positive (LacO TALE) and negative (GFP) control expression plasmids for each reporter tested against all relevant TALE-likes and chimeras. N=16 in every case.

Figure 6

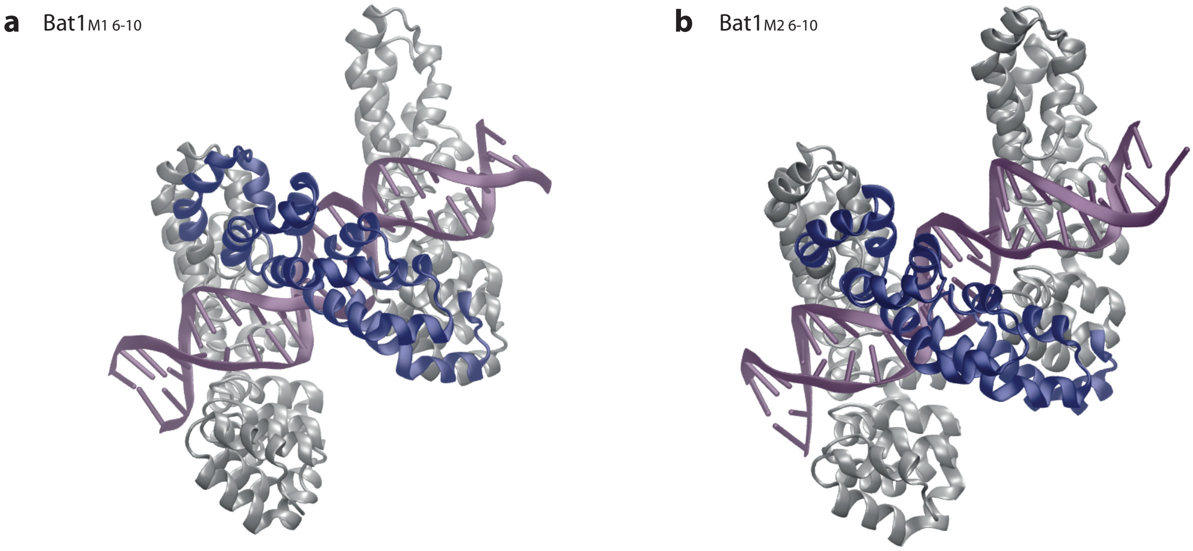


Figure 6
***In silico* homology models of Bat1-MOrTL chimeras bound to their cognate DNA targets correspond to known TALE-like structures.** Homology models of Bat1$_{M1\ 6-10}$ **(a)** and Bat1$_{M2\ 6-10}$ **(b)** were built using PDB entry 4cja as template structure with template DNA structures mutated *in silico* in order to match the optimal bases for both constructs. Single snapshots of the models bound to DNA (purple) are shown. Bat1 repeats are shown in grey. MOrTL repeats are highlighted in dark blue. Models are orientated with the N-terminus of each protein in the bottom left corner.

Figure 7

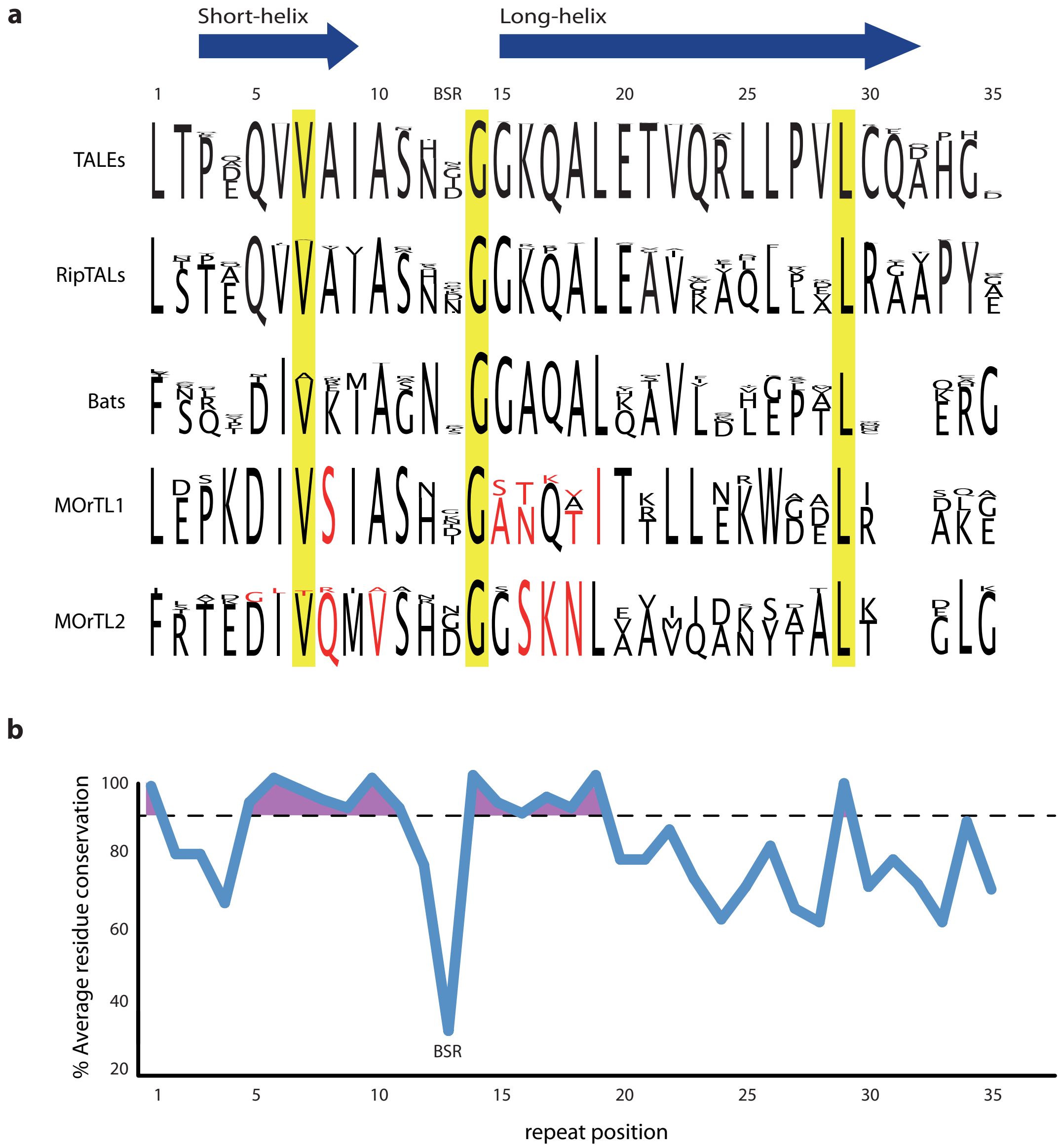


Figure 7
TALE-like repeat alignments show an underlying pattern of sequence conservation around the BSR position. (a) Repeat alignments and corresponding sequence logo were made from representative core repeat arrays from each TALE-like group characterized so far, using CLC Main Workbench 7. In the sequence logo the total height in each column correlates to conservation at that position. Repeat positions are indicated above the sequence logos. Dark blue arrows indicate likely alpha-helical regions based on crystal structures of TALE and Bat repeat arrays. Residues unique to MOrTL1/MOrTL2 at positions 5-19 are highlighted with red lettering. A yellow background highlights residues conserved across all TALE-like core repeats. **(b)** Graph shows within-group sequence conservation at each position, averaged across the five groups. Conservation ≥90% (dashed-line) is shaded purple and a distinct clustering of conserved residues can be seen around the BSR.

6 Discussion

6.1 Overview

The molecular characterization of RipTAL proteins (Chapter 3) was undertaken to reveal the role these proteins might play in bacterial wilt disease and to assist future efforts to breed or engineer resistance to bacterial wilt. It had been assumed by many that the RipTALs and TALEs function similarly, and particularly that RipTAL repeats, like those of TALEs, bind DNA in a sequence specific manner like TALE repeats (74). However so many positions are polymorphic between TALE and RipTAL repeats (Figure 1.8B) and between individual RipTAL repeats (Figure 1.7), that we considered such an assumption inadvisable. Therefore, each repeat was tested individually for base preference, but in each case results corresponded to results for TALE repeats with the same RVDs (3: Figure 3). The TALE code was indeed broadly conserved between TALEs and RipTALs. This code was thus taken as an assumption for further work, and found to hold for canonical repeats of Bat1, Bat2, MOrTL1 and MOrTL2 (chapters 3 and 4). This was surprising considering the sequence identities for canonical repeats of the different groups dropping well below 40% in many cases. Together this work provides evidence of functional conformity in the face of considerable sequence diversity.

With that said some notable differences were found in the DNA recognition properties of different TALE-likes. Non-BSR polymorphisms were found to have some impact on the DNA recognition properties of canonical repeats (3: Figures 3, 4 and S9) as well as on inter-repeat compatibility (4: Figure 5). TALEs require a T_0, RipTALs a G_0 (3: Figure 7) and Bat1 no preference at all (2.2: Figure 2). These and related observations are highly relevant for future work on TALE technology.

In addition, conformity stops at the canonical repeats with domain structure dividing the TALE-likes into at least two groups: the TALEs and RipTALs as contrasting with the Bats, MOrTL1 and MOrTL2. Comparing each group to the TALEs reveals that the TALE-like canonical repeat can be considered as a functionally equivalent domain in all groups though if it is homologous or the result of convergent evolution is unresolved. The canonical TALE-like repeat is, however, the only unifying feature of the TALE-likes.

6.2 Placing this work into the context of similar studies

In the case of both the RipTALs and Bats, molecular characterizations were published just prior to those presented in this thesis. Here I briefly outline those publications and how they compare to the findings presented in this thesis. These are discussed here before continuing on to address the questions posed in the introduction, so that conclusions can be placed into their appropriate context.

Characterization and DNA-binding specificities of Ralstonia TAL-Like effectors. Lixin Li et al., July 2013 (75)

The findings of this paper somewhat corroborated our own, but differ markedly in places, probably because they used a very different approach for DNA recognition analysis. The RipTALs (here designated RTLs) that they worked with were chimeric derivatives bearing native non-canonical-repeat regions and arrays composed of tandemly arranged consensus repeats, or with these consensus repeats embedded into TALE Hax3 in place of its canonical repeats. They showed that the consensus repeats flanked by native RipTAL regions localize to the nucleus. They were also able to show that consensus RipTAL repeats embedded in a Hax3 backbone activate reporters with predicted EBEs. They also modified the RVDs of the consensus repeats and tested specificity by using these chimeras to challenge reporters bearing different EBEs. In some cases their findings match our own: Asp repeats activate the C reporter only; Asn the A and G reporters. In other cases their results differ most strikingly Pro repeats activate all except for the T reporter, in their system. These differences may indicate the importance of non-BSR polymorphisms and/or repeat array context effects: they tested full arrays of consensus repeats, whereas we tested trimers of natural Brg11 repeats embedded in a TALE.

They were unable to show reporter activation by any full-length native RipTAL nor did they show activation by any of the consensus repeat arrays embedded in RipTAL non-canonical-repeat domains. Possibly such attempts were made but the G_0 preference of RipTALs (3: Figure 7) was not taken into account. They did in fact question the zero base specificity but used a RipTAL consensus repeat array embedded into a Hax3 backbone, including the Hax3 N-terminal non-canonical repeats. Unsurprisingly they discovered that T_0 is preferred.

Additional publications covering a molecular characterization of Bat1: Stella et al., January 2014 (76)*; Juillerat et al., January 2014* (56)*; Stella et al., May 2014* (77)

Three publications in quick succession came in 2014 from a collaboration of biomedical technology company Cellectis and the Spanish National Cancer Research Centre. Together they present a reasonably comprehensive report of Bat1-DNA binding interactions, and a number of demonstrations of the potential of Bat1 to act as a programmable DNA binding protein. Overall their findings agree with and complement our own. Bat1 is referred to in these publications as BurrH, or BuD when referring only to the canonical repeats.

The first of these publications reported the crystallization of Bat1 without presenting a structure (76). However it included an EMSA with Bat1 against predicted target sequence 5'-TT*AAGAGAAGCAAATACGTTA*A-3'. The italicized bases indicate cognate bases of the canonical repeats of Bat1, with the only base differing from BE_{Bat1} (4: Figure 1) underlined.

The second paper presented a further molecular characterization of Bat1 and explored the possibility of using Bat1 to create designer nucleases (56). They created Bat1-FokI fusions and demonstrated their function with a yeast single strand annealing reporter bearing two inverted repeats of BE_{Bat1}. They also tested versions of this reporter with all possible bases at positions 0, -1 and -2, finding no activity differences. This corroborates and extends our own finding that Bat1, unlike TALEs and RipTALs, has no fixed zero base preference, and that Bat1-FokI fusions function analogously to TALENs (4: Figure 4).

They then continued to test Bat1 nucleases in human cell culture targeted at endogenes, a necessary step if such nucleases are to be considered as an alternative for TALENs.

As part of their nuclease studies they created a designer repeat array of Bat1 by modifying BSRs only. That this approach proved successful supports our suggestion that the RVD-switch approach (4: Figure 5) is a good approach for modifying Bat1 repeat specificity.

The last of these publications provided not only a crystal structure for the Bat1-DNA complex but also, more fine-grained DNA binding analysis and a further exploration of designer nucleases (77).

Using isothermal calorimetry they tested the binding interactions of Bat1 and different species of nucleic acid single-stranded (ss) and double-stranded (ds) DNA and RNA, and combinations thereof bearing BE_{Bat1}. Of the tested species they found that the Bat1-dsDNA interaction gave the strongest affinity (Kd=25nM). The next best binding partner was a +DNA/-RNA duplex with an affinity half as strong as that for dsDNA. ssDNA, +RNA/-DNA, and also the random sequence DNAs they tested yielded no or extremely weak binding interactions. This all accords with a similar finding for TALEs (78), yet again demonstrating the analogous DNA binding properties of TALEs and Bats.

Binding kinetics were also addressed. They found first that AvrBs3 showed a twofold faster dissociation rate from its target DNA than Bat1 from its target. It would be interesting to know if this can be generalized for Bat vs. TALE DNA interactions. A study utilizing in-vivo fluorescence measurements after bleaching to estimate dissociation times for TALE-DNA interactions found that signal remained after over 7 minutes, indicating an extraordinarily stable interaction (32). Could Bat-DNA interactions be more stable still?

Importantly in this article they present a crystal structure of Bat1, alone and bound to its DNA target. They were able to resolve a structure for the whole protein and revealed a continuous solenoid of helix-turn-helix repeat structures. The super helix and individual repeat structures are highly reminiscent of the solved TALE crystal structures (15, 16) and our own structural prediction of Bat1 (4: Figure S15). Just as for TALEs, they found that, upon binding, the whole array contracts longitudinally to more closely match the pitch of B-form DNA.

One difference they uncovered was that two electropositive stripes run along the Bat1 repeat array, formed by Lys8 and Gln17 of each of repeats 1-18. This allows for an electrostatic interaction with the backbone phosphates of the + and – strands. TALE repeat arrays display only a single electropositive stripe (16).

A very striking observation was that the Arg BSR of repeat 13, in their structure, does not interact with the cognate T in the +strand of their target DNA. Instead they suggest interactions with two neighbouring bases in the –strand. This BSR-base pairing is the only one that differs from our BE_{Bat1}. We used a G at this position in accordance with findings from TALEs that Arg at the BSR position should pair best with G (29). Further molecular investigation isolating the binding contribution of repeat 13, and Arg repeats in general, would be beneficial in this case.

After addressing the structure of Bat1 they returned to the topic of Bat1 derived nucleases. In a yeast-based nuclease assay dBat-nucleases with BSRs for the EBE_{AvrBs3} showed greater target sequence stringency in their activity compared to TALENs. Two base pair mismatches were enough to ablate activity almost completely for the dBat-nucleases, whereas the TALENs performed comparably to the on-target pairing. This exciting finding deserves to be tested further, in different systems and with a range of dBats and dTALEs, to see if the principle can be generalized.

6.3 Addressing the goals of this work

As set out in chapter 2, there are three main questions underlying the work presented in this dissertation:

1. To what extent do TALE-likes conform to the established TALE paradigm in terms of domain structure and function?
2. What are the defining features of TALE-likes?
3. What can be inferred about TALE-like evolution from the comparison of TALE-likes?
4. Can TALE-likes augment TALE technology?

6.3.1 Partial conformity to the TALE paradigm

a) Domain structure and biological function

TALEs and RipTALs act in planta *as transcription factors*

Investigations were launched into the molecular properties of the RipTALs with a suspicion that these proteins, like TALEs, function *in planta* as sequence specific transcription factors. We were able to confirm that RipTALs do indeed function in this manner, localizing to the nucleus (3: Figure 1) and inducing expression of reporter genes bearing a cognate EBE (3: Figure 8).

Despite this functional similarity details of biological activity differ somewhat between TALEs and RipTALs. For instance we found that nuclear localization of RipTALs is mediated by signals in both the NTR and CTR (3: Figure 1). There are motifs in the RipTAL CTR corresponding to the TALE annotated NLSs, whereas the predicted NLSs in the RipTAL NTR (3: Figure S2) have no equivalent in the TALE NTR. The activation domains of TALEs and RipTALs are sequence related and functionally interchangeable. We were able to demonstrate not only that RipTALs function *in planta* as transcriptional activators (3: Figures 8 and 9), but also that the predicted RipTAL AD could functionally replace the AD of AvrBs3. Yet this chimeric protein gave a weaker activation of the reporter than found for wild type AvrBs3 (3: CTC, Figure 2). This may suggest that the activation domain of Brg11 mediates weaker gene induction, but may just as well indicate a structural incompatibility arising during the fusion of the Brg11 CTR to the binding domain and NTR of AvrBs3. Information is lacking on the levels to which RipTALs are able to activate host genes in a natural situation. It would also be interesting to know if the observed ability of TALEs to induce novel transcriptional start sites (79) holds for RipTALs. Homologous domains are found in the TALEs and RipTALs and they mediate similar but not necessarily identical functions.

Characterizing RipTAL functional domains was the first step in the long-term goal of this work was to understand the role these proteins play in bacterial wilt disease. They have been shown to contribute to bacterial growth in a competition assay in eggplant leaf tissue (48). RipTALs thus seem to fit the TALE paradigm closely. Yet TALEs differ greatly in their target sequences even between closely related strains (61) suggesting a co-evolution between host and pathogen, also mirrored in the *SWEET* gene targeting TALEs and corresponding promoter polymorphisms in rice (33). Contrastingly the phylotype I *R. solanacearum* strains studied as part of this thesis were isolated from distantly related host plants and most shared the same repeat array composition as Brg11 (3: Figure 9A). Even where differences were found, most were single repeat loss or gain and such RipTALs showed cross-reactivity on each other's target EBEs (3: Figure 9C). This speaks against a continuous co-evolution between RipTALs and hosts. It is also puzzling that the arrays of RipTALs are so conserved and their hosts so divergent. Are they in fact only relevant in certain hosts? Do they target a hyper-conserved promoter, or a commonly occurring promoter motif? There are also reasons to suspect different types of sequences targeted by TALEs and RipTALs, since the target sequences of all RipTALs we studied are GC rich, unlike TALEs which commonly target AT rich TATA elements of plant promoters. GC rich elements can be found in plant promoters, such as ethylene response elements (80), but there is reason at least to expect differences in what promoter elements are targeted by TALEs and RipTALs. To address these questions transcriptomic studies should be carried out using wild type and RipTAL knock out strains on a range of host plants.

Bats are sequence-specific DNA binding proteins but no further functional domains identified, and even less can be said of the biological functions of the MOrTLs.

Bats and MOrTLs do not carry out the same biological roles as TALEs. They are not delivered by plant pathogenic bacteria to induce virulence associated host genes. Bats are found in endosymbiotic bacteria within a fungus and the MOrTLs with their marine provenance cannot be associated with terrestrial disease causing plant pathogens. Furthermore, in the case of the Bats where we have a reliable idea of gene structure there is no evidence for the existence of the functional domains necessary for TALEs, and RipTALs, to act as *in planta* transcriptional activators. It is possible that Bats, like the toxin rhizoxin, are delivered by *B. rhizoxinica* to its fungal host to be further secreted into the plant host. However, nuclear localization signals are not predicted for this protein and it does not nuclear localize in human cells (4: Figure 3). Furthermore there is no functional eukaryotic AD as inferred from reporter assays in human cells and *in planta* (4: Figures 3 and S9). We were later able to show that Bat1 works as a transcriptional repressor in *E. coli* (5: Figure 4), which suggests, at least, that it could act as a transcriptional regulator inside the bacteria themselves. Yet merely because such an activity is possible does not make it likely. Also feasible is an unpredicted functional domain or a binding partner in the natural situation, mediating an as yet unknown function after DNA binding. Data on the expression and localization of Bats within the natural system would provide first clues. As for the MOrTLs, their biological roles remain completely mysterious, as is the case for so much about the marine pico-plankton. DNA binding repeats are the only proven functional domains of the Bats and MOrTLs and the only basis for assumptions on their biological roles. They might function as transcriptional repressors, but this is only one possibility. It would be intriguing to know whether the Bats and MOrTLs are rare and esoteric or just the first representatives of a larger group of bacterial TALE-like proteins with an as yet unknown functional importance.

A clear demarcation appears between the TALEs and RipTALs on one side, and the Bats and MOrTLs on the other. The RipTALs broadly match the TALE paradigm in respect to functional domains. Are the disease associated TALE-likes, the TALEs and RipTALs, one element of a larger protein group?

b) DNA binding properties

Canonical repeats

Despite the apparent disunity of domain structure between the TALE-likes there is one functional domain, which is strikingly conserved: the canonical repeat. Observed repeat sequence similarity inspired my doctoral work but I found that, as well as sequence similarity, TALE-like canonical repeats also display a striking functional conservation.

The TALE code is conserved across canonical repeats of all TALE-likes

TALE-like repeat sequence diversity is considerable (5: Figure 7) but the code linking BSRs and bases holds across all TALE-like canonical repeat domains so far investigated. Yet non-BSR polymorphisms are not totally without an effect on DNA binding properties. First I will summarize the experimental basis for the assertion that the TALE code based on BSRs is consistent across all TALE-like groups.

The base preferences of individual Brg11 repeats were tested in AvrBs3 derivatives bearing three identical copies of individual Brg11 repeats (3: Figure 3). These were tested against reporters bearing the four possible bases at the cognate EBE position. Despite the non-BSR polymorphisms between TALE and RipTAL repeats results were consistent with the known BSR preferences of TALEs: Asp-C, Gly-T, Asn-G/A, Lys-G, Thr – A/G, Pro-A/T (29, 30). The results of this investigation were then used to predict a binding element for Brg11 ($\mathrm{EBE}_{\mathrm{Brg11}}$), which was indeed recognized by Brg11 *in planta* (3: Figure 7). This showed that the results of the trimer test held true for repeats in their native context. The results of the trimer test were also used to successfully design EBEs for seven other RipTALs with canonical repeat arrays differing in repeat number or BSR composition from Brg11 (3: Figure 9). These reporter assays form the basis of the statement that the TALE code is conserved between TALEs and RipTALs.

In the case of the Bats and MOrTLs the conservation of the code was taken as an assumption to design binding elements. These were then shown to be recognized *in vitro* (4: Figure 2, 5: Figure 2). Off-target oligonucleotides and reporters were not bound by the cognate TALE-likes confirming sequence specificity (4: Figure 2; 5: Figure 4). Additionally, in the case of the Bats, dBats with modified BSRs were created and shown to function in combination with BEs derived from the TALE code (4: Figure 6A; (56, 77)). It is therefore reasonable to affirm that the binding properties of Bat and MOrTL repeats conform to the TALE code.

The conservation of the TALE code strongly suggests structural similarity. That the interactions between BSRs and bases are broadly conserved suggests that repeats all form a similar helix-turn-helix structure with the BSR interpolated into the DNA double helix to make close contacts to DNA bases. It is conceivable that multiple distinct structures are able to achieve the same DNA recognition behavior. But in such a case we would not expect repeat arrays from different TALE-like groups to be inter-compatible. Yet in this thesis RipTAL-TALE (3: Figure 3), Bat-TALE (4: Figure 8), MOrTL-Bat and MOrTL-TALE (5 Figure 2) repeat mixtures have all been demonstrated to be functional. Thus the functional data regarding DNA recognition strongly supports a conserved structure. More convincing than this are the actual solved crystal structures of Bat1 with or without its DNA partner (77) strongly resembling TALE crystal structures, both the super helical arrangement of repeats and the repeats themselves. In both cases the repeat array contracts to closely match the turn of B-form DNA, tracing the major groove. This allows the BSRs of each repeat to come into close enough contact with the base to form hydrogen bonds or van der Waals interactions, depending on the particular pairing. Our structural models of MOrTL repeat blocks embedded in the Bat1 repeat array show the same overall structure (5 Figure 6). This is hardly surprising in itself as the sequence was modeled onto the known Bat1 structure, but molecular dynamics simulations showed that this structure is highly stable, showing that MOrTL repeats can at least stably adopt this confirmation when embedded into a TALE-like repeat array (5 Figures S9 and S10). Combined with the functional observations, the available structures and models make a compelling case for a common structure for TALE-like canonical repeat arrays supporting a conserved BSR code.

Non-BSR polymorphisms affect interaction strength for the preferred target base and inter-repeat compatibility but not base preference

Whilst the dominant finding of this thesis as regards the TALE code is that the code is reliably conserved there are important caveats. Some of these, such as the impact of repeat position within the array, have already been addressed in chapter 1.3. What has been very little addressed till now in the literature is the functional impact of non-BSR polymorphisms on the properties of canonical TALE repeats. This is because only positions 12 and 13 are substantially variable in TALEs. The natural experiment of TALE-likes provides a convenient tool since one is immediately provided with a whole set of repeats of different sequence but at least broadly the same functionality.

Non-BSRs seem to impact the overall strength of the repeat-DNA interaction, with no major effect on base preference. For example, we found that the two Lys repeats of Brg11 both mediate several-fold higher activation of a cognate reporter than equivalent AvrBs3 repeats, when embedded in a dTALE (3: Figures 3 and S13). Yet all are highly specific for the G-reporter. This functional impact most likely arises from the impact of one, several or all of the 9 residues conserved between the Brg11 Lys repeats, which are not found in AvrBs3 repeats. In addition, we found that Brg11 Asp repeats 10 and 8/11 both activate only the C-reporter in the trimer test as expected, but differed in the strength of the activation. The TALE derivative with the repeat 10 trimer was barely able to activate the reporter above background, while repeat 8/11 activated the C-reporter more than the equivalent TALE Asp repeat control (3: Figure 3). We demonstrated that this was linked to a single Asn in repeat 10, occupied by a Lys residue in all other repeats (3: Figure 5). The activities and specificities of a few Bat repeats were assessed in the trimer test (4: Figure 8). These results suggest that non-BSRs impacting on DNA binding non-sequence specifically, either by interacting with the DNA backbone or impacting the positioning of the BSR loop or the conformation of the whole repeat array. This may have implications for TALE technology as explored in 3.2.4.

Another idea, explored in chapter 4: Figure 6, is that non-BSR polymorphisms have co-evolved with BSRs. If this is true then Bat repeat arrays should function best when only the native BSR of each repeat is used. In two of the four dBats we tested, re-arranging BSRs did have a slight negative impact on reporter activation. Numerous Bat1 derivatives, where BSRs only were modified, have been reported by Juillerat *et al*., (56) and Stella *et al*., (77) and shown to function well. Non-BSR polymorphisms do not seem to co-evolve with BSRs based on the available evidence.

Obviously the observed polymorphisms may simply arise through evolutionary drift. It is the most striking of my observations that the sequence polymorphisms between TALE-like repeats seem to matter so little. Certain constraints are however clearly in place on the TALE-like repeat sequence space. Two alpha-helical regions of particular length, a short BSR loop and a longer flexible inter-repeat loop impose constraints on which sequences could form a TALE-like repeat. Furthermore, residues around the BSR conserved within groups and largely within groups too (5: Figure 7) possibly due to the necessity of preserving certain inter-repeat interaction across the array. Yet such constraints still leave an ample sequence space from which residues could be arbitrarily selected during evolution. Further study on the functional impact of non-BSR polymorphisms will help to tease apart the roles of selection and drift in shaping the pattern of sequence diversity seen in TALE-like canonical repeats.

Non-canonical repeats

The canonical repeat arrays of TALEs, RipTALs, and Bats are all framed by non-canonical repeats. Yet while the non-canonical repeats of TALEs and RipTALs seem to be homologous those of Bats are non-homologous and functionally distinct.

N-terminal non-canonical repeat regions of TALEs and RipTALs are homologous and functionally analogous

Sequence comparisons and functional analyses suggest that the N-terminal non-canonical repeats of TALEs and RipTALs are homologous and functionally equivalent. Comparing the amino acid sequences of the known TALE repeats -3 to 0 with the equivalent region of Brg11 (Figure 3) suggests homology. Furthermore, we were able to show that this region in TALEs can be replaced with the equivalent stretch from Brg11 without compromising function (3: Figure 6). The one clear functional effect of this domain swap was a shift away from the canonical T_0 preference of TALEs to a G_0 preference. This was also confirmed to hold for wild type Brg11 (3: Figure 8) and all other RipTALs tested (3: Figures 9 and S11), though two showed some activation of the A_0 reporter (RipTALI-6 and RipTALI-9). The T_0 binding of TALEs has been shown to be directly mediated by Trp232, with co-ordination by several other residues in repeat -1 (15, 57). In Brg11 an Arginine residue occupies the position equivalent to Trp232 (3: Figure 6). This may underlie the observed base preference switch, especially since engineered TALE derivatives with a Trp232Arg mutation displayed a G_0 preference (57). This would suggest that broadly the same structural conformation is adopted by TALE and RipTAL repeats 0 and -1. Overall, it seems that the N-terminal non-canonical repeats of TALEs and RipTALs are homologous, functionally conserved domains.

The C-terminal non-canonical repeats of TALEs seem to be unimportant for DNA binding but it is not known if this holds true for those of RipTALs

Two C-terminal non-canonical repeats can be identified immediately following the canonical repeats of TALEs and RipTALs based on weak sequence similarity to canonical repeats (3: Figures S1). Little is known about the function or structure of TALE or RipTAL repeats +1 and +2 so only tentative statements can be made. Repeat +1 has been shown to exert a sequence preference in accordance with the TALE code (21). However, one study found that the N-terminal (canonical) portion of repeat +1 can be deleted without impairing reporter activation abilities (81). Repeat +2 can be truncated from the binding domain component of a TALE nuclease without impairing function (82). In addition repeat +2 of TALEs harbors a predicted basic leucine zipper motif (12) but one study into the functional importance of this motif for *in planta* activity concluded that it had no impact based on reporter assays with derivatives bearing mutations in this region (58).The C-terminal non-canonical repeats of TALEs clearly await further research but it can already be concluded that any impact on DNA binding is minimal. This contrasts with what we found for Bat1 repeat +1, described below.

N- and C-terminal non-canonical repeats of Bats are important for DNA-binding

Bat1-3 all display a similar N-terminal domain structure. In each case two non-canonical repeats precede the canonical repeats. As in TALEs and RipTALs, repeat 0 lacks a predicted BSR loop but repeat -1 possesses one. But here end the similarities. Bat N-terminal non-canonical repeats share no derived sequence features with those of TALEs. Instead they are sequence related to the canonical repeats of Bats. This suggests that the N-terminal non-canonical repeat domains of Bats and TALEs/RipTALs are likely not homologous but may have each independently derived from canonical repeat sequences of the respective groups. Furthermore, there is no zero base preference for Bat1 (4: Figure 2; (56)). In addition, truncation of these repeats from a designer Bat transcriptional activator impaired but did not totally ablate activity (4: Figure 5B), in contrast to equivalent truncations in dTALEs (25). Overall, the N-terminal non-canonical repeat domains of Bats are non-homologous to and functionally distinct from those of TALEs/RipTALs, though they do contribute to DNA interactions.

Bat proteins also have a C-terminal non-canonical repeat (repeat +1), bearing many positively charged Lys, His and Arg residues, setting it apart from the canonical repeats. Truncation of repeat +1 from acBat1 led to a more dramatic decrease in activity than the N-terminal non-canonical-repeat domain truncation (4: Figure 5B). Structural, sequence and functional observations are consistent with repeat +1 of Bat1 making an important contribution to non-sequence-specific DNA binding and that this domain has no functional equivalent in TALEs and RipTALs.

6.3.2 A unifying definition of the TALE-likes

The TALEs, RipTALs, Bats, MOrTL1 and MOrTL2 are described throughout this thesis as TALE-likes. The definition of the term TALE-likes has never been formally addressed in the literature, and the term itself has been used sporadically and inconsistently. It has been used to describe the RipTALs in both molecular characterization papers (3, (75)). I have used it to refer to Bats and MOrTLs (chapters 3 and 4). It has also been used in a review on TALEs to include RipTALs and Bats (83). Yet in the same review these proteins are simply referred to as *Ralstonia* or *Burkholderia* TALEs, or nonclassical TALE homologs. A working definition of TALE-likes is clearly desirable to allow for more consistency in the literature.

An array of repeats conforming structurally and functionally to those of TALEs is the unifying and defining feature of the TALE-likes

Far more features unite the TALEs and RipTALs than unite these two groups to the Bats and MOrTLs. In addition the evolutionary relationships between Bats, MOrTLs and TALEs/RipTALS are still highly unclear (see below). How then to define the TALE-likes?

I would assign TALE-like status based solely on the repeat array because it is the only conserved domain between TALEs/RipTALs, Bats and MOrTLs, and the most pertinent feature of TALEs to the majority of the research community. The TALE-like canonical repeat domain is an array consisting of tandemly arranged 33-35 amino acid repeats. These repeats consist of two alpha helices, a BSR loop and an inter-repeat loop. These repeats are likely to be highly polymorphic at position 13 but highly conserved around this position (5: Figure 7). Deviation from any of these guidelines would be inconsistent with a functional TALE-like repeat array.

Structural studies carried out on TALEs and Bat1 (15, 16, 77) revealed that the repeat arrays and individual repeats form similar structures. Molecular dynamics simulations that we carried out on modeled structures of Bat1-MOrTL repeat chimeras (5: Figure 6) also suggest that repeats of MOrTL1 and MOrLT2 can form stable structures resembling TALE and Bat repeats. No structures are available for RipTAL repeats but considering the higher sequence similarity these have to TALE repeats anything other than the by now familiar structure would be highly surprising. Since structure and function are clearly related the two together form the working definition of the canonical repeats which define the TALE-likes.

It is not practical to provide a consensus sequence definition TALE repeat due to the sequence diversity present among and within groups (5: Figure 7). As explored in chapter 5, only three residues are conserved across all TALE-like repeats ($X_6\mathbf{V}X_6\mathbf{G}X_{13}\mathbf{L}X_{4\text{-}6}$). This could be useful as a pattern to identify future TALE-likes. Yet using full repeat arrays of TALEs, RipTALs, Bats, MOrTL1 and MOrTL2 separately is likely to provide a more comprehensive and reliable search template.

With this in mind, I would suggest keeping TALE-likes as an informal collective term for any proteins bearing a repeat array conforming to the structure and function of the TALE canonical repeat array. This covers all proteins studied in this thesis, but excludes, for example, pentatricopeptide repeat (PPR) proteins. Such proteins display tandem arrays of 35 amino acid repeats, like RipTALs and some TALEs and these repeats form two alpha-helices with two loop regions, similar to the canonical TALE-like repeat structure. They have been explicitly compared to TALEs (84) and have even been asserted to belong to the same superfamily of proteins (24). Yet PPRs bind ssRNA, one of the nucleic acid species which cannot be bound by TALEs or Bats (78, 77). Moreover the TALE code cannot be applied to PPR-RNA interactions. Nor is there any clear sequence similarity between known PPRs and TALE-likes (84). The term TALE-like, based on canonical repeats mediating DNA binding conforming to the TALE code provides a practical definition with sufficient precision to exclude proteins with apparently similar domain structure.

6.3.3 Insights into TALE and TALE-like evolution

The key observations made to date regarding TALE evolution were outlined in 2.3. Briefly, *TALE* genes are found in multi-copy in *X. oryzae* strains, and in *X. oryzae* pv *oryzae* strains are mostly flanked by inverted repeats indicative of Tn3 transposition; in many non-*oryzae* strains they are plasmid localized (61). TALE repeats are highly uniform in sequence, but repeat arrays are diverse in respect to repeat number, BSR composition and thus diverse with respect to their DNA targets (Figure 1.5).

Despite functional similarities what we know of the evolutionary history of the RipTALs contrasts with that of TALEs

RipTALs do not fit this paradigm in any respect. When present they are generally found in single copy, with the exception of banana infecting strain Molk2 (two copies). *Xanthomonas* strains bearing TALEs in single copy generally carry them on episomes (61), whereas RipTALs are borne within the bi-partite genome of *R. solanacearum* strains. In addition, our analysis of Phylotype I *R. solanacearum* strains found that repeat array polymorphisms between orthologs are mostly single repeat losses or duplications (3: Figure 9). The other major isoform is the Hpx17-like lacking all repeats but one. TALEs seem to display an evolutionary dynamism, likely made possible by mechanisms of intra- and intergenic recombination (85). Such mechanisms in turn likely rely on the high sequence identity of TALE repeats. In contrast, RipTALs show little repeat array diversity across the whole species complex, but the sequences of individual repeats differ greatly from one another. Thus the rates of inter and intra genic recombination thought to allow TALEs to achieve their astonishing BSR diversity (Figure 1.5) may be far lower for RipTALs. There may thus be a mechanistic connection between individual repeat sequence identity on one hand, and evolutionary dynamism of repeat arrays on the other. The molecular characterization of RipTAL function and diversity described in chapter 3 has provided hints into RipTAL and TALE evolution allowing the formulation of testable hypotheses.

Despite their unusual provenance the Bats are likely to be homologs of the TALEs and RipTALs

The Bats are known only from a single strain of a single *Burkholderia* species. Genome sequences for 47 *Burkholderia* genomes are available on NCBI but only *B. rhizoxinica* has so far been found to bear TALE-likes. Thus the evolutionary provenance of the Bats is somewhat puzzling. In particular it is unclear if they have arisen independently of the TALEs and RipTALs. To approach an answer to this it is necessary to address the MOrTLs. Though both MOrTLs were isolated from bacteria occupying similar marine environments the two MOrTLs differ from one another at the sequence level as much as they do from each of the other TALE-like groups (Figure 1.5). Like the Bats their repeats are 33 amino acids long, and from the available sequence data it seems, based on the admittedly limited data available, that they lack large non-repeat domains, like the Bats (5: Figure 7). MOrTL 1 and 2 repeats are polymorphic at certain positions with respect to the TALEs, RipTALs and the Bats too. That is to say that certain residues at certain positions are found consistently in TALE, RipTAL and Bat repeats but in repeats of MOrTL 1 and/or MOrTL2 some other residue is found there. Common sequence features, where alternatives are demonstrably possible within the structural and functional constrains, are an indication of common evolutionary origins. The residues in question are Ala10, Gly15, Gln17, Ala18 and Leu19. 5 residues out of 33-35 cannot be taken as a definitive proof but is at least an indication. Chance convergence cannot be excluded, especially if the structure and function or MOrTL1 or MOrTL2 repeats differ in some subtle way that relaxes some sequence constraint on them. And even if the canonical repeats of TALEs/RipTALs and Bats share a common ancestor, that ancestor may have been a single helix-turn-helix domain, co-opted twice independently to form a long array of sequence-specific DNA binding repeats. I hope that in the coming years further TALE-likes are discovered, which shed some further light on these questions. Till then the available evidence can be taken to draw a tentative conclusion that the DNA-binding domains of Bats, TALEs and RipTALs are homologous.

6.3.4 Practical and conceptual contributions to TALE technology

TALEs, RipTALs and Bats, with the addition of MOrTL repeats represent alternative protein platforms for sequence specific DNA binding. Furthermore, having shown that dTALEs bearing non-*Xanthomonas* TALE-likes are in most cases functional, this paves the way for chimeric TALE-like arrays. I envision three primary ways in which these findings could be used to augment existing TALE technology: chimeric repeat arrays due to lower inter-repeat sequence similarity are less likely to undergo rearrangements, dBats as a replacement for dTALEs, and finally exploiting repeat polymorphisms to create a library of arrays with the same target sequence but different affinities. I will explore each in turn.

TALE-like repeats could be used to stabilize repetitive stretches in TALE genes

TALE genes are unstable in some systems (72, 73). Whole repeat arrays are lost due to presumed intragenic recombination. This property has been noticed when the instability is so drastic that it completely prevents application, yet it may be a far more widespread phenomenon than acknowledged. It may be that organisms or populations stably transformed with *TALEs* will over time become genetic mosaics bearing a mixture of intact and compromised *TALE* genes. Such an outcome would stand in the way of any application requiring stable and continuously active TALEs. These could include gene therapy and synthetic genetic circuits operating within modified organisms (e.g. (71, 86)). In 2013 a team at Harvard Medical School created "reTALEs" by using alternative codon usage to prevent identical nucleotide strings longer than 11 bp occurring anywhere within the repeat arrays (87). reTALEs were functional in lentiviral delivery vectors, which had been found to be incompatible with standard dTALEs (72) suggesting that this was sufficient to stabilize the genes. Using the sequence diversity of TALE-like repeats is an easy way to further reduce the sequence similarity of *TALE* genes, and therefore improve stability.

dBats could be used as an alternative to dTALEs

An alternative to creating sequence diverse dTALE arrays is to use the naturally sequence diverse Bat1 repeat array as a scaffold. As explored in 4: Figure 6, there are two approaches for reprogramming the Bat1 repeat array to create dBats: the repeat switch and the RVD (BSR) switch. The first of these was hypothesized to lead to steric clashes caused by repeat polymorphisms in neighboring repeats. In accordance with this all repeat-switch constructs displayed reduced activity relative to the wild-type (4: Figure 6A). Similarly we hypothesized that BSRs and non-BSRs might co-evolve such that BSRs will not be properly positioned relative to the base in a non-native setting. Yet the BSR-switch constructs performed better than the analogous repeat switch constructs (4: Figure 6B) and others have successfully created several dBats using this approach (56, 77). This suggests that the creation of dBats by using the native repeat array of Bat1 and modifying BSRs could be a straightforward and reliable approach akin to dTALE assembly. What then are the advantages of such a system? As well as the intimated stability improvement, another advantage of the dBat system is the more compact DNA-binding domain: fewer non-canonical repeats and each repeat one amino acid shorter. In addition, the findings of Stella *et al.* suggest that dBats could be more sequence specific than dTALEs (77). If this were found to be a general principle this would give a clear advantage to the dBat scaffold over the dTALE. Conversely, we found that the Bat1-BE_{Bat1} interaction is rather weak (4: Figure 2) and again if this is a general principle, perhaps related to the fewer non-canonical repeats this could be disadvantageous. Also disadvantageous is the extra research that would be necessary to establish dBats as a dTALE replacement. Over the last five years a number of excellent studies have a systematically tested different TALE domain truncations and fusions (26, 42, 82, 88) and quantified the overall interaction strength and sequence specific of dTALE repeat arrays with different BSR compositions (21, 22, 29, 30). It is perhaps unrealistic, given the rising star of the CRISPR/Cas9 system, to expect these investments to be repeated for a system that is directly analogous to the dTALE system.

TALE-like non-BSR polymorphisms could be used to fine tune the TALE repeat

Finally, TALE-like repeat diversity could be used on a smaller-scale to alter the DNA recognition and interaction properties of TALE repeats. Miller *et al*. have shown that residue 12 polymorphisms, have a profound on interaction strength (30). Similarly, we found drastic differences in inferred DNA interaction properties between TALE and RipTAL repeats or between individual RipTAL repeats that seem to be due to non-BSR polymorphisms (3: Figures 5 and S13). Testing the total theoretical sequence space of the TALE repeat would entail functional tests on 20^{34} constructs, and a random walk is likely to lead mostly through non-functional constructs. The natural experiment of TALE-like repeat diversity can be taken as a more limited raw material for TALE repeat engineering. The benefit of such an approach would be the possibility of designing libraries of dTALEs with not only different target sequences but also those with differing affinities for the same target sequence. In synthetic biology such libraries would be useful to be able to create analogous circuits with different outputs or to regulate flow through different regulatory pathways. As described above this would again require further heavy research investments. Yet the product would be something qualitatively different from anything achieved or achievable till now.

7 Concluding remarks

In the introduction I painted a vision of plant disease research working together with green biotechnology to increase the output and reliability of agricultural production for the coming centuries. Of course reality is forever reminding us of its nuances and contingencies. Knowing this I had no illusion of providing some panacea for global food security with my work characterising TALE-like proteins. What I hoped to achieve at the outset, and what I have done together with others, is to lay out the groundwork for future efforts from which tangible rewards might be reaped.

This groundwork is twofold. Firstly, for the role of RipTALs in plant disease. Our molecular characterisation of native RipTAL proteins, showed that they function as sequence specific transcription factors *in planta.* All the functional domains of TALEs seem to be conserved with RipTALs, suggesting that RipTALs, like TALEs contribute to plant disease by activating host *S*-genes. If this information can be used to identify such *S*-genes, then this can be used to better understand bacterial wilt disease and to inform resistance-breeding programmes.

Secondly, by showing that canonical repeats are analogous and in most cases functionally compatible among TALE-like groups this may lay out a path to the next generation of TALE technology. With the advent of CRISPR/Cas9 the future for TALE-technology will lie in high-fidelity applications, and the creation of parts libraries for synthetic biology. These applications could benefit from the introduction of TALE-like repeat sequences by improving and diversifying DNA binding properties of canonical repeats by exploiting non-BSR polymorphisms, and by increasing the genetic stability of *TALE* genes.

Finally, the characterisations of Bats and MOrTLs have unexpectedly revealed that the TALE-likes might be a larger group than previously thought. Whether or not all TALE-likes have arisen from a common ancestor they clearly carry out biological roles beyond that of effectors inducing host gene induction. TALE-likes lacking translocation signals or ADs may regulate bacterial transcription by acting as repressors, or mediating functions through as yet unknown binding partners. As with all research, observation leads to new directions of inquiry, and I eagerly await further marine bacterial genomics projects to learn if the MOrTLs are widely distributed. As in the early days of ocean exploration, there are no maps to guide us.

8 Bibliography

1. Gustavsson, J., Cederberg, C., Sonesson, U., van Otterdijk, R. and Meybeck, A. (2011) Global food losses and food waste FAO.

2. Pingali, P.L. (2012) Green Revolution: Impacts, limits, and the path ahead. *Proc. Natl. Acad. Sci.*, **109**, 12302–12308.

3. Oerke, E.-C. (2006) Crop losses to pests. *J. Agric. Sci.*, **144**, 31.

4. Pray, C.E., Huang, J., Hu, R. and Rozelle, S. (2002) Five years of Bt cotton in China - the benefits continue. *Plant J.*, **31**, 423–430.

5. Gonsalves, D. (2004) Transgenic papaya in Hawaii and beyond. *AgBioForum*, **7**, 36–40.

6. Voytas, D.F. (2013) Plant genome engineering with sequence-specific nucleases. *Annu. Rev. Plant Biol.*, **64**, 327–50.

7. De Lange, O., Binder, A. and Lahaye, T. (2014) From dead leaf, to new life: TAL effectors as tools for synthetic biology. *Plant J.*, **78**, 753–771.

8. Fesenko, E. and Edwards, R. (2014) Plant synthetic biology: A new platform for industrial biotechnology. *J. Exp. Bot.*, **65**, 1927–1937.

9. Mansfield, J., Genin, S., Magori, S., Citovsky, V., Sriariyanum, M., Ronald, P., Dow, M.A.X., Verdier, V., Beer, S. V, Machado, M.A., et al. (2012) Top 10 plant pathogenic bacteria in molecular plant pathology. **13**, 614–629.

10. Block, A., Li, G., Fu, Z.Q. and Alfano, J.R. (2008) Phytopathogen type III effector weaponry and their plant targets. *Curr. Opin. Plant Biol.*, **11**, 396–403.

11. Boch, J. and Bonas, U. (2010) Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu. Rev. Phytopathol.*, **48**, 419–36.

12. Szurek, B., Marois, E., Bonas, U. and Van den Ackerveken, G. (2001) Eukaryotic features of the Xanthomonas type III effector AvrBs3: protein domains involved in transcriptional activation and the interaction with nuclear import receptors from pepper. *Plant J.*, **26**, 523–34.

13. Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A. and Bonas, U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509–12.

14. Moscou, M.J. and Bogdanove, A.J. (2009) A simple cipher governs DNA recognition by TAL effectors. *Science*, **326**, 1501.

15. Mak, A.N.-S., Bradley, P., Cernadas, R.A., Bogdanove, A.J. and Stoddard, B.L. (2012) The Crystal Structure of TAL Effector PthXo1 Bound to Its DNA Target. *Science (80-.).*, **335**, 716–719.

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

17. Boch, J. and Bonas, U. (2010) Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu. Rev. Phytopathol.*, **48**, 419–36.

18. Cermak, T., Doyle, E.L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J. a, Somia, N. V, Bogdanove, A.J. and Voytas, D.F. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.*, 10.1093/nar/gkr218.

19. Pérez-Quintero, A.L., Rodriguez-R, L.M., Dereeper, A., López, C., Koebnik, R., Szurek, B. and Cunnac, S. (2013) An improved method for TAL effectors DNA-binding sites prediction reveals functional convergence in TAL repertoires of Xanthomonas oryzae strains. *PLoS One*, **8**, e68464.

20. Yu, Y., Streubel, J., Balzergue, S., Champion, A., Boch, J., Koebnik, R., Feng, J., Verdier, V. and Szurek, B. (2011) Colonization of rice leaf blades by an African strain of Xanthomonas oryzae pv. oryzae depends on a new TAL effector that induces the rice nodulin-3 Os11N3 gene. *Mol. Plant. Microbe. Interact.*, **24**, 1102–13.

21. Miller, J.C., Tan, S., Qiao, G., Barlow, K. a, Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J., et al. (2011) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.*, **29**, 143–8.

22. Meckler, J.F., Bhakta, M.S., Kim, M.-S., Ovadia, R., Habrian, C.H., Zykovich, a., Yu, a., Lockwood, S.H., Morbitzer, R., Elsaesser, J., et al. (2013) Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Res.*, 10.1093/nar/gkt085.

23. Stella, S., Molina, R., Yefimenko, I., Prieto, J., Silva, G., Bertonati, C., Juillerat, A., Duchateau, P. and Montoya, G. (2013) Structure of the AvrBs3-DNA complex provides new insights into the initial thymine-recognition mechanism. *Acta Crystallogr. D. Biol. Crystallogr.*, **69**, 1707–16.

24. Deng, D., Yan, C., Wu, J., Pan, X. and Yan, N. (2014) Revisiting the TALE repeat. *Protein Cell*, 10.1007/s13238-014-0035-2.

25. Gao, H., Wu, X., Chai, J. and Han, Z. (2012) Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res.*, **2**, 1–5.

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

27. Streubel, J., Blücher, C., Landgraf, A. and Boch, J. (2012) TAL effector RVD specificities and efficiencies. *Nat. Biotechnol.*, **30**, 593–595.

28. Wicky, B.I.M., Stenta, M. and Dal Peraro, M. (2013) TAL Effectors Specificity Stems from Negative Discrimination. *PLoS One*, **8**, e80261.

29. Yang, J., Zhang, Y., Yuan, P., Zhou, Y., Cai, C., Ren, Q., Wen, D., Chu, C., Qi, H. and Wei, W. (2014) Complete decoding of TAL effectors for DNA recognition. *Cell Res.*, 10.1038/cr.2014.19.

30. Miller, J.C., Zhang, L., Xia, D.F., Campo, J.J., Ankoudinova, I. V, Guschin, D.Y., Babiarz, J.E., Meng, X., Hinkley, S.J., Lam, S.C., et al. (2015) Improved specificity of TALE-based genome editing using an expanded RVD repertoire. *Nat. Methods*, 10.1038/nmeth.3330.

31. Cuculis, L., Abil, Z., Zhao, H. and Schroeder, C.M. (2015) Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nat. Commun.*, **6**, 7277.

32. Thanisch, K., Schneider, K., Morbitzer, R., Solovei, I., Lahaye, T., Bultmann, S. and Leonhardt, H. (2014) Targeting and tracing of specific DNA sequences with dTALEs in living cells. *Nucleic Acids Res.*, **42**, e38.

33. Boch, J., Bonas, U. and Lahaye, T. (2014) Research review TAL effectors – pathogen strategies and plant resistance engineering.

34. Li, T., Liu, B., Spalding, M.H., Weeks, D.P. and Yang, B. (2012) High-efficiency TALEN-based gene editing produces disease-resistant rice. *Nat. Biotechnol.*, **30**, 390–2.

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

36. Strauss, T., van Poecke, R.M.P., Strauss, A., Römer, P., Minsavage, G. V, Singh, S., Wolf, C., Strauss, A., Kim, S., Lee, H.-A., et al. (2012) RNA-seq pinpoints a Xanthomonas TAL-effector activated resistance gene in a large-crop genome. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 19480–5.

37. Tian, D., Wang, J., Zeng, X., Gu, K., Qiu, C., Yang, X., Zhou, Z., Goh, M., Luo, Y., Murata-Hori, M., et al. (2014) The Rice TAL Effector-Dependent Resistance Protein XA10 Triggers Cell Death and Calcium Depletion in the Endoplasmic Reticulum. *Plant Cell*, 10.1105/tpc.113.119255.

38. Wang, C., Fan, Y., Zheng, C., Qin, T., Zhang, X. and Zhao, K. (2014) High-resolution genetic mapping of rice bacterial blight resistance gene Xa23. *Mol. Genet. Genomics*, 10.1007/s00438-014-0848-y.

39. Yang, B., Sugio, A. and White, F.F. (2005) Avoidance of host recognition by alterations in the repetitive and C-terminal regions of AvrXa7, a type III effector of Xanthomonas oryzae pv. oryzae. *Mol. Plant. Microbe. Interact.*, **18**, 142–9.

40. Wang, X., Wang, Y., Wu, X., Wang, J., Wang, Y., Qiu, Z., Chang, T., Huang, H., Lin, R. and Yee, J. (2015) Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. 10.1038/nbt.3127.

41. Morbitzer, R., Elsaesser, J., Hausner, J. and Lahaye, T. (2011) Assembly of custom TALE-type DNA binding domains by modular cloning. *Nucleic Acids Res.*, **39**, 5790–9.

42. Konermann, S., Brigham, M.D., Trevino, A.E., Hsu, P.D., Heidenreich, M., Cong, L., Platt, R.J., Scott, D. a, Church, G.M. and Zhang, F. (2013) Optical control of mammalian endogenous transcription and epigenetic states. *Nature*, **500**, 472–6.

43. Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., Billault, a, Brottier, P., Camus, J.C., Cattolico, L., et al. (2002) Genome sequence of the plant pathogen Ralstonia solanacearum. *Nature*, **415**, 497–502.

44. Peeters, N., Carrère, S., Anisimova, M., Plener, L., Cazalé, A.-C. and Genin, S. (2013) Repertoire, unified nomenclature and evolution of the Type III effector gene set in the Ralstonia solanacearum species complex. *BMC Genomics*, **14**, 859.

45. Cunnac, S., Occhialini, A., Barberis, P., Boucher, C. and Genin, S. (2004) Inventory and functional analysis of the large Hrp regulon in Ralstonia solanacearum: identification of novel effector proteins translocated to plant host cells through the type III secretion system. *Mol. Microbiol.*, **53**, 115–28.

46. Mukaihara, T., Tamura, N. and Iwabuchi, M. (2010) Genome-wide identification of a large repertoire of Ralstonia solanacearum type III effector proteins by a new functional screen. *Mol. Plant. Microbe. Interact.*, **23**, 251–62.

47. Heuer, H., Yin, Y.-N., Xue, Q.-Y., Smalla, K. and Guo, J.-H. (2007) Repeat domain diversity of avrBs3-like genes in Ralstonia solanacearum strains and association with host preferences in the field. *Appl. Environ. Microbiol.*, **73**, 4379–84.

48. Macho, A.P., Guidot, A., Barberis, P., Beuzón, C.R. and Genin, S. (2010) A competitive index assay identifies several Ralstonia solanacearum type III effector mutant strains with reduced fitness in host plants. *Mol. Plant. Microbe. Interact.*, **23**, 1197–1205.

49. Lackner, G., Moebius, N., Partida-Martinez, L.P., Boland, S. and Hertweck, C. (2011) Evolution of an endofungal lifestyle: Deductions from the Burkholderia rhizoxinica genome. *BMC Genomics*, **12**, 210.

50. Partida-Martinez, L.P., Groth, I., Schmitt, I., Richter, W., Roth, M. and Hertweck, C. (2007) Burkholderia rhizoxinica sp. nov. and Burkholderia endofungorum sp. nov., bacterial endosymbionts of the plant-pathogenic fungus Rhizopus microsporous. *Int. J. Syst. Evol. Microbiol.*, **57**, 2583–2590.

51. Partida-Martinez, L.P., Bandemer, S., Rüchel, R., Dannaoui, E. and Hertweck, C. (2008) Lack of evidence of endosymbiotic toxin-producing bacteria in clinical Rhizopus isolates. *Mycoses*, **51**, 266–9.

52. Iwasaki, S., Okuda, S., Shimizu, F., Sasagawa, K., Fukami, M. and Fukuda, K. (1986) Rhizoxin, a macrocylic lactone antibiotic, as a new antitumor agent against huamn and murine tumor cells and their vincristine-resistant sublines.

53. Partida-Martinez, L.P. and Hertweck, C. (2005) Pathogenic fungus harbours endosymbiotic bacteria for toxin production. *Nature*, **437**, 884–888.

54. Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J. a., Heidelberg, K.B., Manning, G., Li, W., et al. (2007) The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biol.*, **5**, 0432–0466.

55. Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. a., Hoffman, J.M., Remington, K., et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, 0398–0431.

56. Juillerat, A., Bertonati, C., Dubois, G., Guyot, V., Thomas, S., Valton, J., Beurdeley, M., Silva, G.H., Daboussi, F. and Duchateau, P. (2014) BurrH: a new modular DNA binding protein for genome engineering. *Sci. Rep.*, **4**, 3831.

57. Lamb, B.M., Mercer, a. C. and Barbas, C.F. (2013) Directed evolution of the TALE N-terminal domain for recognition of all 5' bases. *Nucleic Acids Res.*, 10.1093/nar/gkt754.

58. Schreiber, T., Sorgatz, A., List, F., Blüher, D., Thieme, S., Wilmanns, M. and Bonas, U. (2015) Refined Requirements for Protein Regions Important for Activity of the TALE AvrBs3. *PLoS One*, **10**, e0120214.

59. Mahfouz, M.M., Piatek, A., Neal, C. and Jr, S. (2014) Genome engineering via TALENs and CRISPR / Cas9 systems : challenges and perspectives. *Plant Biotechnol. J.*, **12**, 1006–1014.

60. Rodriguez-R, L.M., Grajales, A., Arrieta-Ortiz, M., Salazar, C., Restrepo, S. and Bernal, A. (2012) Genomes-based phylogeny of the genus Xanthomonas. *BMC Microbiol.*, **12**, 43.

61. Rafael Marini Ferreira, Amanda Carolina P. de Oliveira, Leandro M. Moreira, J.B., Gourbeyre, E., Siguier, P., Ferro, M.I.T., Ferro, J.A., Chandler, M. and Varania, A.M. (2015) A TALE of Transposition : Tn 3 -Like Transposons Play a Major Role in the Spread of Pathogenicity Determinants of Xanthomonas citri and. 10.1128/mBio.02505-14.Editor.

62. Canteros, B.I. (1995) Diversity of Plasmids in Xanthomonas campestris pv. vesicatoria. *Phytopathology*, **85**, 1482.

63. Yang, Y., Gabriel, D.W. and Gabriel, D.W. (1995) Intragenic recombination of a single plant pathogen gene provides a mechanism for the evolution of new host specificities . These include : Intragenic Recombination of a Single Plant Pathogen Gene Provides a Mechanism for the Evolution of New Host Specifi. *Microbiology*, **177**.

64. Qi, L.S., Larson, M.H., Gilbert, L. a, Doudna, J. a, Weissman, J.S., Arkin, A.P. and Lim, W. a (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173–83.

65. Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L. and Church, G.M. (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.*, **31**, 833–8.

66. Gao, Y. and Zhao, Y. (2014) Self-processing of ribozyme-flanked RNAs into guide RNAs in vitro and in vivo for CRISPR-mediated genome editing. *J. Integr. Plant Biol.*, **56**, 343–349.

67. Esvelt, K.M., Mali, P., Braff, J.L., Moosburner, M., Yaung, S.J. and Church, G.M. (2013) Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods*, 10.1038/nmeth.2681.

68. Zalatan, J.G., Lee, M.E., Almeida, R., Gilbert, L. a., Whitehead, E.H., La Russa, M., Tsai, J.C., Weissman, J.S., Dueber, J.E., Qi, L.S., et al. (2014) Engineering Complex Synthetic Transcriptional Programs with CRISPR RNA Scaffolds. *Cell*, **160**, 339–350.

69. Politz, M.C., Copeland, M.F. and Pfleger, B.F. (2013) Artificial repressors for controlling gene expression in bacteria. *Chem. Commun. (Camb).*, **49**, 4325–7.

70. Liang, P., Xu, Y., Zhang, X., Ding, C., Huang, R., Zhang, Z., Lv, J., Xie, X., Chen, Y., Li, Y., et al. (2015) CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes. *Protein Cell*, 10.1007/s13238-015-0153-5.

71. Blount, B. a, Weenink, T., Vasylechko, S. and Ellis, T. (2012) Rational diversification of a promoter providing fine-tuned expression and orthogonal regulation for synthetic biology. *PLoS One*, **7**, e33279.

72. Holkers, M., Maggio, I., Liu, J., Janssen, J.M., Miselli, F., Mussolino, C., Recchia, A., Cathomen, T. and Gonçalves, M. a F. V (2012) Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic Acids Res.*, 10.1093/nar/gks1446.

73. Lau, C.-H., Zhu, H., Tay, J.C.-K., Li, Z., Tay, F.C., Chen, C., Tan, W.-K., Du, S., Sia, V.-K., Phang, R.-Z., et al. (2014) Genetic rearrangements of variable di-residue (RVD)-containing repeat arrays in a baculoviral TALEN system. *Mol. Ther. — Methods Clin. Dev.*, **1**, 14050.

74. Schornack, S., Meyer, A., Ro, P., Jordan, T. and Lahaye, T. (2006) Gene-for-gene-mediated recognition of nuclear-targeted AvrBs3-like bacterial effector proteins. *J. Plant Physiol.*, **163**, 256–272.

75. Li, L., Atef, A., Piatek, A., Ali, Z., Piatek, M., Aouida, M., Sharakuu, A., Mahjoub, A., Wang, G., Khan, S., et al. (2013) Characterization and DNA-binding specificities of Ralstonia TAL-like effectors. *Mol. Plant*, **6**, 1318–1330.

76. Stella, S., Molina, R., Bertonatti, C., Juillerrat, A. and Montoya, G. (2014) Expression, purification, crystallization and preliminary X-ray diffraction analysis of the novel modular DNA-binding protein BurrH in its apo form and in complex with its target DNA. *Acta Crystallogr. Sect. F, Struct. Biol. Commun.*, **70**, 87–91.

77. Stella, S., Molina, R., López-Méndez, B., Juillerat, A., Bertonati, C., Daboussi, F., Campos-Olivas, R., Duchateau, P. and Montoya, G. (2014) BuD, a helix-loop-helix DNA-binding domain for genome modification. *Acta Crystallogr. D. Biol. Crystallogr.*, **70**, 2042–52.

78. Yin, P., Deng, D., Yan, C., Pan, X., Xi, J.J., Yan, N. and Shi, Y. (2012) Specific DNA-RNA hybrid recognition by TAL effectors. *Cell Rep.*, **2**, 707–13.

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

80. Ohme-Takagi, M. and Shinshi, H. (1995) Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *Plant Cell*, **7**, 173–182.

81. Zheng, C.K., Wang, C.L., Zhang, X.P., Wang, F.J., Qin, T.F. and Zhao, K.J. (2014) The last half-repeat of transcription activator-like effector (TALE) is dispensable and thereby TALE-based technology can be simplified. *Mol. Plant Pathol.*, **15**, 690–697.

82. Mussolino, C., Morbitzer, R., Lütge, F., Dannemann, N., Lahaye, T. and Cathomen, T. (2011) A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res.*, **39**, 9283–93.

83. Schornack, S., Moscou, M.J., Ward, E.R. and Horvath, D.M. (2013) Engineering Plant Disease Resistance Based on TAL Effectors. 10.1146/annurev-phyto-082712-102255.

84. Filipovska, A. and Rackham, O. (2012) Modular recognition of nucleic acids by PUF, TALE and PPR proteins. *Mol. Biosyst.*, **8**, 699–708.

85. Yang, B., Sugio, A. and White, F. (2005) Avoidance of host recognition by alterations in the repetitive and C-terminal regions of AvrXa7, a type III effector of Xanthomonas oryzae pv. oryzae. *Mol. plant-microbe …*, **18**, 142–149.

86. Lienert, F., Torella, J.P., Chen, J.-H., Norsworthy, M., Richardson, R.R. and Silver, P. a (2013) Two- and three-input TALE-based AND logic computation in embryonic stem cells. *Nucleic Acids Res.*, 10.1093/nar/gkt758.

87. Yang, L., Guell, M., Byrne, S., Yang, J.L., De Los Angeles, A., Mali, P., Aach, J., Kim-Kiselak, C., Briggs, A.W., Rios, X., et al. (2013) Optimization of scarless human stem cell genome editing. *Nucleic Acids Res.*, **41**, 9049–9061.

88. Miller, J.C., Tan, S., Qiao, G., Barlow, K. a, Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J., et al. (2010) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.*, 10.1038/nbt.1755.

9 Appendices

Supplementary information for publications presented in this thesis. These documents are presented as in the original publications, without additional interrupting pages

9.1 Supplementary information for de Lange *et al*., New Phytologist (2013).

9.2 Supplementary information for de Lange, Wolf *et al*., Nucl. Acids Res. (2014).

9.3 Supplementary information for de Lange, Wolf *et al*., Nucl. Acids Res. (2015).

Supplementary Information

Figures

S1 Annotated sequences of Brg11 and AvrBs3.

S2 Amino acid sequences of Hpx17 and corresponding truncation derivatives used in subcellular localisation studies.

S3 Alignment of C-terminal regions of AvrBs3 and Brg11 (a) and the sequence of the C-terminal chimera CTC (b).

S4 Alignments of Brg11 and AvrBs3 core repeats.

S5 Representative amino acid sequences used in the "trimer test".

S6 Reciprocal exchange of RVDs between Brg11 repeats 12 and 8/11 leading to exchange of DNA recognition specificity.

S7 Nucleotide sequences of *brg11* repeats used to assemble three identical, tandem-arranged repeat blocks that were tested in the context of $\text{dTALE}_{\text{EBE Brg11}}$.

S8 Alignment of N-terminal regions AvrBs3 and Brg11 regions (a) and sequence of chimeras NTC1 and NTC2 (b).

S9 Individual amino acid sequences of RipTALs analysed in this study.

S10 Amino acid sequence alignment of RipTALs analysed in this study.

S11 Zero base preferences of RipTALI-6, -9, -11 and -14.

S12 Comparison of predicted Brg11 binding sequences from this study and from Streubel *et al.*, 2012.

S13 Alignment of Brg11 repeats 2 and 9 and a consensus AvrBs3 repeat.

Tables

Supplementary Table 1: Sequences of primers used in this study.

Supplementary Table 2: Origin of RipTALs analysed in this study.

Supplementary Table 3: GUS activities as determined in the trimer test and shown in Figure 3.

Supplementary Figure 1

Annotated sequences of Brg11 and AvrBs3. N-terminal and C-terminal non-repeat regions and the central repeat array are displayed in separate paragraphs but are part of contiguous polypeptides. Consecutive repeats are numbered (left side). The repeats 0 and -1 constitute the N-terminal, 1-17 the core and +1 and +2 the C-terminal repeats. The RVDs (residues at repeat positions 12 and 13) are marked as boldface black letters on grey background. Previously predicted monopartite nuclear localization signals (NLSs; consensus K-K/R-X-K/R, Chelsky et al., 1989) of AvrBs3 shown to be relevant for protein function are highlighted in blue background (Van den Ackerveken et al., 1996), corresponding sequences in Brg11 are highlighted in green background (see also Fig. S3a). Further NLS sequences not tested here were predicted by NLStradamus (Nguyen et al., 2009) and cNLS Mapper (Kosugi et al., 2009) and are shown in boldface, blue font. Only those predicted NLSs that were experimentally tested are indicated as red rectangles in cartoon representations of Brg11 and homologs in Figs 1-2, 5-8 and S10. The annotated acidic activation domain of AvrBs3 (Szurek et al., 2001) and a predicted nine amino acid transactivation domain in Brg11 are underlined.

>Brg11 (from *R. solanacearum* strain GMI1000; GenBank NP_519936.1)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAAGDGGLAHLHRRLAVGYAEDTPRTEA
RSPAPRRPLPVAPASAPPAPSLVPEPPMPVSLPAVSSPRFSAGSSAAITDPFPSLPPTPVLY
AMARELEALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIASACPQAFASPPRAP**RS**
ARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSREQIRKLKQESLSEIAKYHTT
LTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQY GERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLELRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAHLLDLRGVPYA
6 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGNQALEAVGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LNTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAVRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
15 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
16 LNTAQIVAIAS**HD**GGKPALEAVWAKLPVLRGAPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQASHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIR**RIR**

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQATRPALLRLLSSVGVT
EIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGSPGMAGQSACS
PHRKRPAETAIAPRSIRRSPNNAGQPSEPWPDQLAWLQRRKRTARSHIRADSAASVPANLHL
GTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPDPGTPTSADLASFEAEPFGVG
PLDFHLDWLLQILET

>AvrBs3 (from *Xanthomonas campestris* pv. *vesicatoria* strain 71-21; GenBank CAA34257.1)

MDPIRSRTPSPARELLPGPQPDGVQPTADRGVSPPAGGPLDGLPARRTMSRTRLPSPPAPSP
AFSAGSFSDLLRQFDPSLFNTSLFDSLPPFGAHHTEAATGEWDEVQSGLRAADAPPPTMRVA
VTAA**RPPRAKPAPRRR**AAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGH
GFTHAHIVALSQHPAALGTVAVKYQDMIAALPE

-1 ATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQ
0 LDTGQLLKIAKR GGVTAVEAVHAWRNALTGAPLN
1 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
2 LTPQQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
3 LTPQQVVAIAS**NS**GGKQALETVQRLLPVLCQAHG
4 LTPEQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
5 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
6 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
7 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
8 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
9 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
10 LTPQQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
11 LTPEQVVAIAS**NS**GGKQALETVQALLPVLCQAHG
12 LTPEQVVAIAS**NS**GGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
15 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
16 LTPQQVVAIAS**NG**GGRPALETVQRLLPVLCQAHG
17 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
+1 LTPQQVVAIASNGGGRPALESIVAQLSRPDPALAA
+2 LTNDHLVALACL GGRPALDAVKKGLPHAPALIKRT

NRRIPERTSHRVADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGMSRHGLLQLFRRVGVTELE
ARSGTLPPASQRWDRILQASGMKRAKPSPTSTQTPDQASLHAFADSLERDLDAPSPMHEGDQ
TRASSRKRSRSDRAVTGPSAQQSFEVRVPEQRDALHLPLSWRVKRPRTSIGGGLPDPGTPTA
ADLAASSTVMREQDEDPFAGAADDFPAFNEEELAWLMELLPQ

Supplementary Figure 2

Amino acid sequences of Hpx17 and corresponding truncation derivatives used in subcellular localisation studies (Fig. 1). N-terminal, repeat and C-terminal regions of Hpx17 (a) and its derivatives (b - e) are displayed in distinct paragraphs but are part of contiguous polpypeptides. Consecutive repeats are numbered (left side). Repeats 0 and -1 constitute the N-terminal, 1 the core and +1 and +2 the C-terminal repeats. The RVD is marked in boldface black font on grey background. Putative NLSs (see also in Brg11 Figs S1, S3) are highlighted in green background, corresponding NLS mutations in red background. Further, not experimentally tested NLSs predicted by NLStradamus (Nguyen et al., 2009) and cNLS Mapper (Kosugi et al., 2009) are shown in boldface, blue font. Only those predicted NLSs that were experimentally tested are indicated as red rectangles in cartoon representations of Brg11 and homologs in Figs 1-2, 5-8 and S10. Below each amino acid sequence the corresponding nucleotide sequence is shown. Dashes (-) indicate amino acids in repeat +2 that have not been integrated in the given constructs (d and e).

(a) Sequence of Hpx17

>Hpx17 (from *R. solanacearum* strain RS1000; GenBank BAD42396.1)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGYAEDTPRTGA
RSPAPRRPLPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPFSSLPPTPVLY
AMARELKALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIASACPQAFASPPRAP**RS**
ARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSREQIRKLKQESLSEIAKYHTT
LTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
 0 LSASQIATVAQY GERPAIQALYRLRRKLTRAPLH
 1 LTPQQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIR**RIR**

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLSSVGVT
EIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGSPGMAGQSACS
PHRKRPAETAIAPRSIRRRPNNAGQPSEPWPDQLAWLQRRKRTARSHIRADSAASVPANLHL
GTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPDPGTPTSAASFEAEPFGVGPL
DFHLDWLLQILEA

>hpx17

```
ATGAGAATAGGCAAATCAAGCGGTTGGTTGAACGAGTCCGTGTCTCTTGAATATGAACACGT
GTCCCCACCGACACGGCCTCGAGACACCCGTCGCCGGCCTCGCGCCGCTAGCGACGGCGGGC
TCGCGCATCTGCATCGCCGGCTCGCGGTCGGCTACGCGGAGGACACGCCGAGAACCGGGGCT
CGGTCTCCGGCGCCGCGCCGCCCGCTCCCTGTGGCACCTGCATCCGCACCGCCTGCACCGTC
CCTCGTTCCGGAACCCCCTATGCCGGTCAGCCTTCCTGTCGTATCGAGCCCGCGCTTCTCTG
CCGGCAGCTCGGCAGCCATCACCGATCCTTTTTCGAGCCTTCCGCCCACGCCCGTGCTGTAT
GCGATGGCTCGCGAACTGAAGGCGCTGTCCGACGCTACCTGGCAGCCAGCCGTACCGTTGCC
CGCCGAGCCGCCTACTGATGCGCGGCGCGGCAACACGGTATTTGACGAAGCGTCTGCATCAT
CGCCGGTGATCGCCTCTGCCTGCCCTCAAGCGTTTGCCAGCCCACCGCGAGCACCGCGCTCG
GCGCGAGCCCGCAGGGCTCGGACAGGCGGTGATGCTTGGCCGGCCCCGACTTTTCTTAGCCG
CCCCTCGTCATCCCGCATCGGCCGTGACGTGTTCGGGAAACTGGTCGCACTCGGCTATTCCC
GTGAGCAGATCCGGAAGCTCAAGCAGGAGAGCCTGAGCGAAATTGCGAAGTATCACACCACC
TTGACAGGACAAGGGTTCACGCACGCCGACATCTGCCGGATCAGCCGCAGACGGCAGTCGCT
CCGGGTGGTCGCCAGGAACTACCCGGAGTTGGCTGCGGCGCTCCCTGAGCTGACCAGGGCCC
ACATCGTGGACATCGCTCGGCAGCGATCGGGCGACTTGGCGCTGCAAGCGCTGCTACCCGTG
GCGACCGCACTGACAGCGGCCCCCCTGAGATTGAGCGCCTCGCAAATCGCGACCGTTGCGCA
GTACGGCGAGCGGCCGGCCATCCAGGCCCTTTATCGGCTGCGCCGGAAGCTCACGCGAGCAC
CGCTGCATCTCACACCGCAGCAGGTGGTGGCCATCGCCAGCCACGATGGCGGTAAACCGGCG
CTGGAAGCGGTCTGGGCGAAATTGCCGGTATTGCGCGGGGTGCCCTATGCGCTGAGCACCGC
GCAAGTGGTGGCCATTGCCTGCATCAGTGGCCAGCAGGCGCTGGAGGCAATCGAGGCGCACA
TGCCTACATTGCGCCAAGCCCCCCACAGCCTGAGTCCCGAGCGGGTGGCGGCGATCGCGTGC
ATCGGCGGCCGATCAGCCGTGGAGGCCGTCAGGCAGGGGTTGCCGGTGAAGGCGATCCGGCG
GATACGGCGCGAGAAAGCCCCTGTAGCCGGGCCGCCACCAGCCTCTCTTGGCCCAACCCCGC
AGGAACTCGTGGCGGTCCTGCATTTCTTCCGTGCACATCAGCAGCCCAGACAGGCCTTTGTC
GACGCACTGGCAGCGTTCCAGACCACCAGACCGGCACTGTTGAGGTTGCTCAGCAGTGTTGG
GGTCACAGAAATCGAGGCGCTCGGCGGCACGATCCCCGACGCCACCGAGCGCTGGCAGCGCC
TGCTTGGCCGGCTGGGCTTCAGGCCGGCAACCGGCGCTGCCGCGCCTTCGCCTGATTCCCTG
CAAGGGTTCGCCCAGTCACTTGAGCGCACGCTCGGGTCTCCCGGTATGGCAGGGCAATCGGC
TTGCTCACCACATCGCAAGCGGCCTGCCGAGACGGCCATCGCACCGCGGTCGATACGACGCA
GACCCAACAATGCGGGCCAACCCTCCGAGCCATGGCCCGATCAACTGGCATGGCTCCAACGC
AGGAAACGTACCGCTCGTTCGCACATACGGGCCGATTCGGCGGCAAGCGTGCCGGCAAATCT
CCACTTGGGCACGCGAGCCCAGTTCACGCCAGATCGTCTTCGGGCCGAACCGGGACCCATCA
TGCAGGCTCACACATCGCCGGCATCGGTCAGCTTCGGTTCTCACGTTGCTTTCGAGCCTGGC
CTGCCGGACCCCGGTACGCCCACCTCAGCAGATCTTGCATCGTTCGAGGCTGAGCCGTTCGG
CGTCGGGCCGCTTGATTTCCACCTCGACTGGCTTCTGCAAATATTGGAAGCG
```

(b) Sequence of Hpx17-mut-NLSI/II

>Hpx17-mut-NLSI/II

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGYAEDTPRTGA
RSPAPRRPLPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPFSSLPPTPVLY
AMARELKALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIASACPQAFASPPRAP**RS**
ARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSREQIRKLKQESLSEIAKYHTT
LTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

Repeat -1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
Repeat 0 LSASQIATVAQY GERPAIQALYRLRRKLTRAPLH
Repeat 1 LTPQQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYA
Repeat +1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
Repeat +2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIR**RIR**

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLSSVGVT
EIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGSPGMAGQSACS
PQAYWPAETAIAPRSIRRRPNNAGQPSEPWPDQLAWLQPDPWTARSHIRADSAASVPANLHL
GTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPDPGTPTSADLASFEAEPFGVG
PLDFHLDWLLQILEA

>*hpx17-mut-NLSI/II*

ATGAGAATAGGCAAATCAAGCGGTTGGTTGAACGAGTCCGTGTCTCTTGAATATGAACACGT
GTCCCCACCGACACGGCCTCGAGACACCCGTCGCCGGCCTCGCGCCGCTAGCGACGGCGGGC
TCGCGCATCTGCATCGCCGGCTCGCGGTCGGCTACGCGGAGGACACGCCGAGAACCGGGGCT
CGGTCTCCGGCGCCGCGCCGCCCGCTCCCTGTGGCACCTGCATCCGCACCGCCTGCACCGTC
CCTCGTTCCGGAACCCCCTATGCCGGTCAGCCTTCCTGTCGTATCGAGCCCGCGCTTCTCTG
CCGGCAGCTCGGCAGCCATCACCGATCCTTTTTCGAGCCTTCCGCCCACGCCCGTGCTGTAT
GCGATGGCTCGCGAACTGAAGGCGCTGTCCGACGCTACCTGGCAGCCAGCCGTACCGTTGCC
CGCCGAGCCGCCTACTGATGCGCGGCGCGGCAACACGGTATTTGACGAAGCGTCTGCATCAT
CGCCGGTGATCGCCTCTGCCTGCCCTCAAGCGTTTGCCAGCCCACCGCGAGCACCGCGCTCG
GCGCGAGCCCGCAGGGCTCGGACAGGCGGTGATGCTTGGCCGGCCCCGACTTTTCTTAGCCG
CCCCTCGTCATCCCGCATCGGCCGTGACGTGTTCGGGAAACTGGTCGCACTCGGCTATTCCC
GTGAGCAGATCCGGAAGCTCAAGCAGGAGAGCCTGAGCGAAATTGCGAAGTATCACACCACC
TTGACAGGACAAGGGTTCACGCACGCCGACATCTGCCGGATCAGCCGCAGACGGCAGTCGCT
CCGGGTGGTCGCCAGGAACTACCCGGAGTTGGCTGCGGCGCTCCCTGAGCTGACCAGGGCCC
ACATCGTGGACATCGCTCGGCAGCGATCGGGCGACTTGGCGCTGCAAGCGCTGCTACCCGTG
GCGACCGCACTGACAGCGGCCCCCCTGAGATTGAGCGCCTCGCAAATCGCGACCGTTGCGCA
GTACGGCGAGCGGCCGGCCATCCAGGCCCTTTATCGGCTGCGCCGGAAGCTCACGCGAGCAC
CGCTGCATGGGTTGCCGGTGAAGGCGATCCGGCGGATACGGCGCGAGAAAGCCCCTGTAGCC
GGGCCGCCACCAGCCTCTCTTGGCCCAACCCCGCAGGAACTCGTGGCGGTCCTGCATTTCTT
CCGTGCACATCAGCAGCCCAGACAGGCCTTTGTCGACGCACTGGCAGCGTTCCAGACCACCA
GACCGGCACTGTTGAGGTTGCTCAGCAGTGTTGGGGTCACAGAAATCGAGGCGCTCGGCGGC
ACGATCCCCGACGCCACCGAGCGCTGGCAGCGCCTGCTTGGCCGGCTGGGCTTCAGGCCGGC
AACCGGCGCTGCCGCGCCTTCGCCTGATTCCCTGCAAGGGTTCGCCCAGTCACTTGAGCGCA
CGCTCGGGTCTCCCGGTATGGCAGGGCAATCGGCTTGCTCACCACAAGCTTACTGGCCTGCC
GAGACGGCCATCGCACCGCGGTCGATACGACGCAGACCCAACAATGCGGGCCAACCCTCCGA
GCCATGGCCCGATCAACTGGCATGGCTCCAACCGGATCCATGGACCGCTCGTTCGCACATAC
GGGCCGATTCGGCGGCAAGCGTGCCGGCAAATCTCCACTTGGGCACGCGAGCCCAGTTCACG
CCAGATCGTCTTCGGGCCGAACCGGGACCCATCATGCAGGCTCACACATCGCCGGCATCGGT
CAGCTTCGGTTCTCACGTTGCTTTCGAGCCTGGCCTGCCGGACCCCGGTACGCCCACCTCAG
CAGATCTTGCATCGTTCGAGGCTGAGCCGTTCGGCGTCGGGCCGCTTGATTTCCACCTCGAC
TGGCTTCTGCAAATATTGGAAGCG

(c) Sequence of Hpx17-Δrep-ΔC

>Hpx17-Δrep-ΔC

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGYAEDTPRTGA
RSPAPRRPLPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPFSSLPPTPVLY
AMARELKALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIASACPQAFASPPRAP**RS**
ARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSREQIRKLKQESLSEIAKYHTT
LTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
 0 LSASQIATVAQY GERPAIQALYRLRRKLTRAPLH
LTPQ

>*hpx17-Δrep-ΔC*

ATGAGAATAGGCAAATCAAGCGGTTGGTTGAACGAGTCCGTGTCTCTTGAATATGAACACGT
GTCCCCACCGACACGGCCTCGAGACACCCGTCGCCGGCCTCGCGCCGCTAGCGACGGCGGGC
TCGCGCATCTGCATCGCCGGCTCGCGGTCGGCTACGCGGAGGACACGCCGAGAACCGGGGCT
CGGTCTCCGGCGCCGCGCCGCCCGCTCCCTGTGGCACCTGCATCCGCACCGCCTGCACCGTC
CCTCGTTCCGGAACCCCCTATGCCGGTCAGCCTTCCTGTCGTATCGAGCCCGCGCTTCTCTG
CCGGCAGCTCGGCAGCCATCACCGATCCTTTTTCGAGCCTTCCGCCCACGCCCGTGCTGTAT
GCGATGGCTCGCGAACTGAAGGCGCTGTCCGACGCTACCTGGCAGCCAGCCGTACCGTTGCC
CGCCGAGCCGCCTACTGATGCGCGGCGCGGCAACACGGTATTTGACGAAGCGTCTGCATCAT
CGCCGGTGATCGCCTCTGCCTGCCCTCAAGCGTTTGCCAGCCCACCGCGAGCACCGCGCTCG
GCGCGAGCCCGCAGGGCTCGGACAGGCGGTGATGCTTGGCCGGCCCCGACTTTTCTTAGCCG
CCCCTCGTCATCCCGCATCGGCCGTGACGTGTTCGGGAAACTGGTCGCACTCGGCTATTCCC
GTGAGCAGATCCGGAAGCTCAAGCAGGAGAGCCTGAGCGAAATTGCGAAGTATCACACCACC
TTGACAGGACAAGGGTTCACGCACGCCGACATCTGCCGGATCAGCCGCAGACGGCAGTCGCT
CCGGGTGGTCGCCAGGAACTACCCGGAGTTGGCTGCGGCGCTCCCTGAGCTGACCAGGGCCC
ACATCGTGGACATCGCTCGGCAGCGATCGGGCGACTTGGCGCTGCAAGCGCTGCTACCCGTG
GCGACCGCACTGACAGCGGCCCCCCTGAGATTGAGCGCCTCGCAAATCGCGACCGTTGCGCA
GTACGGCGAGCGGCCGGCCATCCAGGCCCTTTATCGGCTGCGCCGGAAGCTCACGCGAGCAC
CGCTGCATCTCACACCGCAG

(d) Sequence of Hpx17-ΔN-Δrep

>Hpx17-ΔN-Δrep

Repeat +2 M-----------------------LPVKAIR**RIR**

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLSSVGVT
EIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGSPGMAGQSACS
PHRKRPAETAIAPRSIRRRPNNAGQPSEPWPDQLAWLQRRKRTARSHIRADSAASVPANLHL
GTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPDPGTPTSADLASFEAEPFGVG
PLDFHLDWLLQILEA

>*hpx17-ΔN-Δrep*

ATGTTGCCGGTGAAGGCGATCCGGCGGATACGGCGCGAGAAAGCCCCTGTAGCCGGGCCGCC
ACCAGCCTCTCTTGGCCCAACCCCGCAGGAACTCGTGGCGGTCCTGCATTTCTTCCGTGCAC
ATCAGCAGCCCAGACAGGCCTTTGTCGACGCACTGGCAGCGTTCCAGACCACCAGACCGGCA
CTGTTGAGGTTGCTCAGCAGTGTTGGGGTCACAGAAATCGAGGCGCTCGGCGGCACGATCCC
CGACGCCACCGAGCGCTGGCAGCGCCTGCTTGGCCGGCTGGGCTTCAGGCCGGCAACCGGCG
CTGCCGCGCCTTCGCCTGATTCCCTGCAAGGGTTCGCCCAGTCACTTGAGCGCACGCTCGGG
TCTCCCGGTATGGCAGGGCAATCGGCTTGCTCACCACATCGCAAGCGGCCTGCCGAGACGGC
CATCGCACCGCGGTCGATACGACGCAGACCCAACAATGCGGGCCAACCCTCCGAGCCATGGC
CCGATCAACTGGCATGGCTCCAACCGCAGGAAACGTACCGCTCGTTCGCACATACGGGCCGAT
TCGGCGGCAAGCGTGCCGGCAAATCTCCACTTGGGCACGCGAGCCCAGTTCACGCCAGATCG
TCTTCGGGCCGAACCGGGACCCATCATGCAGGCTCACACATCGCCGGCATCGGTCAGCTTCG
GTTCTCACGTTGCTTTCGAGCCTGGCCTGCCGGACCCCGGTACGCCCACCTCAGCAGATCTT
GCATCGTTCGAGGCTGAGCCGTTCGGCGTCGGGCCGCTTGATTTCCACCTCGACTGGCTTCT
GCAAATATTGGAAGCG

(e) Sequence of Hpx17-ΔN-Δrep-mutNLSI/II

>Hpx17-ΔN-Δrep-mut-NLSI/II

+2 M-----------------------LPVKAIR**RIR**

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLSSVGVT
EIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGSPGMAGQSACS
PQAYWPAETAIAPRSIRRRPNNAGQPSEPWPDQLAWLQPDPWTARSHIRADSAASVPANLHL
GTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPDPGTPTSADLASFEAEPFGVG
PLDFHLDWLLQILEA

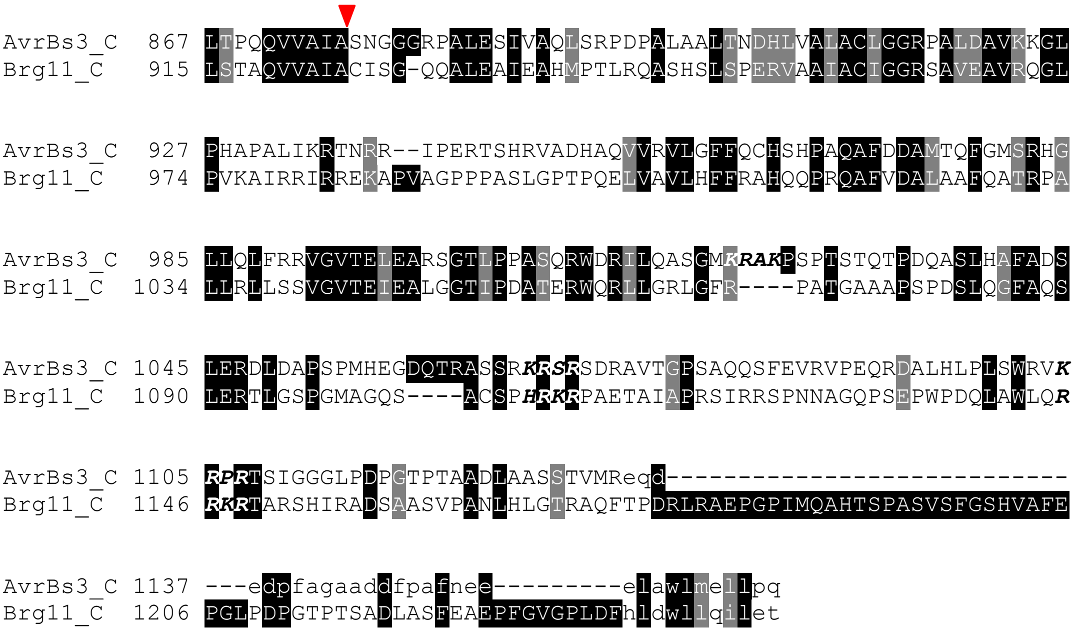
>hpx17-ΔN-Δrep-mut-NLSI/II

ATGTTGCCGGTGAAGGCGATCCGGCGGATACGGCGCGAGAAAGCCCCTGTAGCCGGGCCGCC
ACCAGCCTCTCTTGGCCCAACCCCGCAGGAACTCGTGGCGGTCCTGCATTTCTTCCGTGCAC
ATCAGCAGCCCAGACAGGCCTTTGTCGACGCACTGGCAGCGTTCCAGACCACCAGACCGGCA
CTGTTGAGGTTGCTCAGCAGTGTTGGGGTCACAGAAATCGAGGCGCTCGGCGGCACGATCCC
CGACGCCACCGAGCGCTGGCAGCGCCTGCTTGGCCGGCTGGGCTTCAGGCCGGCAACCGGCG
CTGCCGCGCCTTCGCCTGATTCCCTGCAAGGGTTCGCCCAGTCACTTGAGCGCACGCTCGGG
TCTCCCGGTATGGCAGGGCAATCGGCTTGCTCACCACAAGCTTACTGGCCTGCCGAGACGGC
CATCGCACCGCGGTCGATACGACGCAGACCCAACAATGCGGGCCAACCCTCCGAGCCATGGC
CCGATCAACTGGCATGGCTCCAACCGGATCCATGGACCGCTCGTTCGCACATACGGGCCGAT
TCGGCGGCAAGCGTGCCGGCAAATCTCCACTTGGGCACGCGAGCCCAGTTCACGCCAGATCG
TCTTCGGGCCGAACCGGGACCCATCATGCAGGCTCACACATCGCCGGCATCGGTCAGCTTCG
GTTCTCACGTTGCTTTCGAGCCTGGCCTGCCGGACCCCGGTACGCCCACCTCAGCAGATCTT
GCATCGTTCGAGGCTGAGCCGTTCGGCGTCGGGCCGCTTGATTTCCACCTCGACTGGCTTCT
GCAAATATTGGAAGCG

Supplementary Figure 3

Alignment of C-terminal regions of AvrBs3 and Brg11 (a) and the sequences of the C-terminal chimeras CTC and CTC-ΔAD (b).

(a) Alignment of the C-termini of AvrBs3 (AvrBs3_C) and Brg11 (Brg11_C). The alignment was constructed with ClustalW and Boxshade. Amino acids that are identical between the repeat units are displayed as white letters on black background. The fusion points for the creation of the C-terminal chimera is indicated with a red triangle. NLSs (for Brg11 those experimentally tested; Fig. 1) are highlighted in bold, italic font. The previously annotated acidic activation domain of AvrBs3 (Szurek et al., 2001) and a predicted nine amino acid transactivation domain in Brg11 are shown in lower case letters (see also Fig. S1).



```
AvrBs3_C  867 LTPQQVVAIASNGGGRPALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGL
Brg11_C   915 LSTAQVVAIACISG-QQALEAIEAHMPTLRQASHSLSPERVAAIACIGGRSAVEAVRQGL

AvrBs3_C  927 PHAPALIKRTNRR--IPERTSHRVADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGMSRHG
Brg11_C   974 PVKAIRRIRREKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQATRPA

AvrBs3_C  985 LLQLFRRVGVTELEARSGTLPPASQRWDRILQASGMKRAKPSPTSTQTPDQASLHAFADS
Brg11_C  1034 LLRLLSSVGVTEIEALGGTIPDATERWQRLLGRLGFR----PATGAAAPSPDSLQGFAQS

AvrBs3_C 1045 LERDLDAPSPMHEGDQTRASSRKRSRSDRAVTGPSAQQSFEVRVPEQRDALHLPLSWRVK
Brg11_C  1090 LERTLGSPGMAGQS----ACSPHRKRPAETAIAPRSIRRSPNNAGQPSEPWPDQLAWLQR

AvrBs3_C 1105 RPRTSIGGGLPDPGTPTAADLAASSTVMReqd----------------------------
Brg11_C  1146 RKRTARSHIRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFE

AvrBs3_C 1137 ---edpfagaaddfpafnee---------elawlmellpq
Brg11_C  1206 PGLPDPGTPTSADLASFEAEPFGVGPLDFhldwllqilet
```

(b) Annotated amino sequence of the C-terminal chimera (CTC). The sequence corresponding to AvrBs3 is shown in regular font, the Brg11 part is shown in boldface, italic font (for annotation of AvrBs3 and Brg11, see Fig. S1). The predicted transactivation domain of Brg11 is underlined. The sequences lacking from the CTC-ΔAD construct are indicated with green highlighting. Below the amino acid sequence the nucleotide sequence of the chimera is given starting with repeat 17 of *avrBs3*. To generate the chimera a *brg11* derivative codon optimized for *in planta* expression (*sbrg11,* synthesised by GenScript) was used.

>CTC

MDPIRSRTPSPARELLPGPQPDGVQPTADRGVSPPAGGPLDGLPARRTMSRTRLPSPPAPSP
AFSAGSFSDLLRQFDPSLFNTSLFDSLPPFGAHHTEAATGEWDEVQSGLRAADAPPPTMRVA
VTAARPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGH
GFTHAHIVALSQHPAALGTVAVKYQDMIAALPE

-1 ATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQ
0 LDTGQLLKIAKR GGVTAVEAVHAWRNALTGAPLN
1 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
2 LTPQQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
3 LTPQQVVAIAS**NS**GGKQALETVQRLLPVLCQAHG
4 LTPEQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
5 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
6 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
7 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
8 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
9 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
10 LTPQQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
11 LTPEQVVAIAS**NS**GGKQALETVQALLPVLCQAHG
12 LTPEQVVAIAS**NS**GGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
15 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
16 LTPQQVVAIAS**NG**GGRPALETVQRLLPVLCQAHG
17 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
+1 LTPQQVVAIA***CISGQQALEAIEAHMPTLRQASHS***
+2 ***LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR***

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQATRPALLRLLSSVGVT
EIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGSPGMAGQSACS
PHRKRPAETAIAPRSIRRSPNNAGQPSEPWPDQLAWLQRRKRTARSHIRADSAASVPANLHL
GTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPDPGTPTSADLASFEAEPFGVG
PLDFHLDWLLQILET

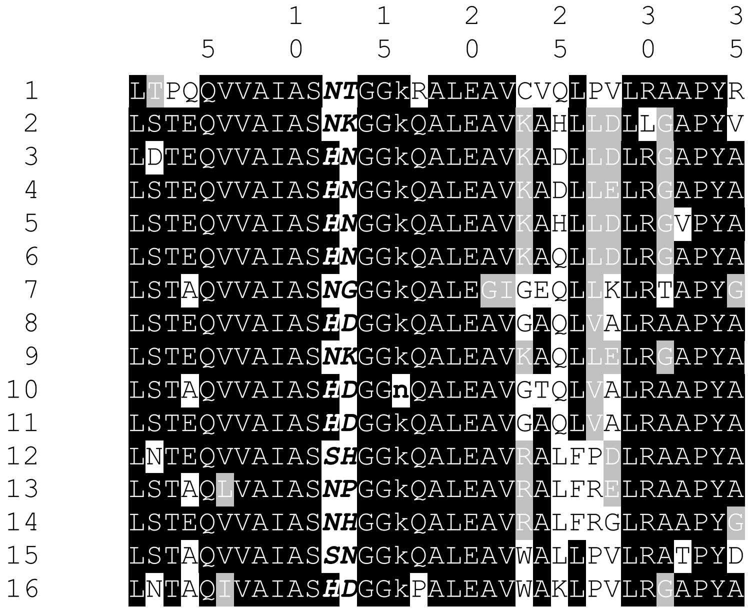
>CTC

…(*avrBs3 17th repeat*)CTGACCCCGCAGCAGGTGGTGGCCATCGCCAGCCATGATGG
CGGCAAGCAGGCGCTGGAGACGGTGCAGGCGCTGTTGCCGGTGCTGTGCCAGGCCCATGGCC
TGACCCCCCAGCAAGTCGTTGCAATCGCT(*sbrg11*→) ***TGCATCTCCGGTCAACAAGCTCTG***
GAGGCAATTGAAGCCCACATGCCAACTCTTAGACAGGCATCACATAGTCTTTCTCCTGAGAG
GGTTGCCGCTATTGCTTGTATTGGAGGAAGAAGTGCAGTTGAGGCTGTTAGGCAAGGTCTGC
CAGTGAAGGCTATTAGAAGAATAAGAAGGGAGAAGGCCCCCGTTGCTGGACCACCACCAGCA
TCACTTGGTCCAACACCTCAAGAGCTTGTAGCAGTTTTACACTTCTTCAGGGCTCACCAACA
GCCAAGACAGGCTTTTGTGGATGCACTAGCTGCATTTCAGGCAACTCGTCCTGCACTACTTA
GGCTACTTAGTTCTGTGGGAGTCACAGAGATAGAGGCTTTGGGCGGTACTATTCCAGACGCA
ACTGAAAGATGGCAGAGGTTGTTAGGAAGGCTTGGCTTTAGACCTGCAACTGGGGCTGCTGC
TCCATCTCCTGATTCCTTGCAAGGTTTCGCTCAATCTCTAGAACGTACACTTGGTTCACCAG
GAATGGCAGGTCAATCTGCATGTTCCCCACATAGGAAGAGGCCAGCTGAAACTGCTATCGCT
CCAAGGAGTATCAGGAGGTCCCCTAATAATGCTGGACAGCCTTCAGAGCCTTGGCCAGACCA
GCTGGCTTGGCTACAAAGGAGGAAACGAACAGCTCGTAGCCATATTAGGGCTGATTCTGCAG
CTTCTGTGCCAGCTAATCTCCATCTTGGTACAAGGGCACAATTTACTCCTGATAGGTTGAGA
GCAGAACCTGGACCCATTATGCAAGCACATACATCTCCTGCATCTGTTTCCTTCGGATCACA
CGTTGCATTTGAACCTGGTCTACCTGATCCAGGAACTCCTACTTCAGCAGATCTTGCTTCTT
TCGAGGCTGAACCATTCGGAGTGGGTCCATTGGACTTTCATCTGGACTGGCTCCTCCAAATT
CTTGAAACA

Supplementary Figure 4

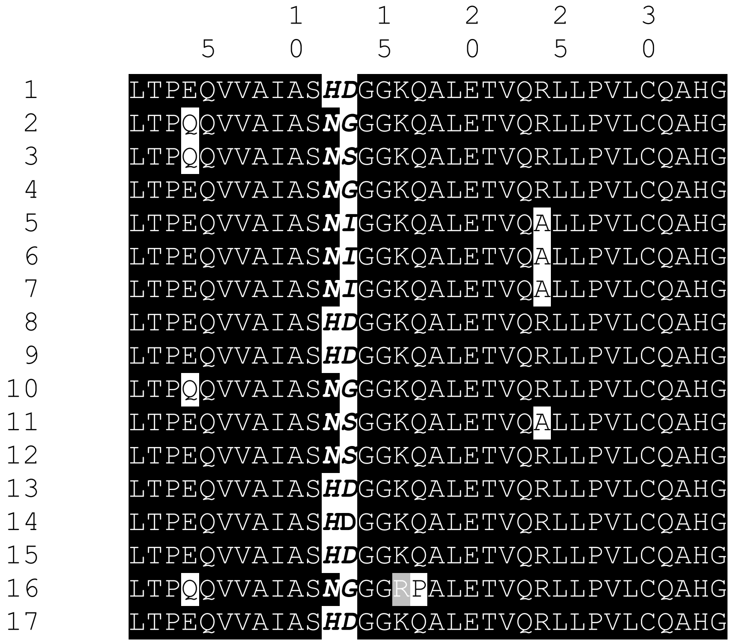
Alignments of Brg11 and AvrBs3 core repeats. Alignments were constructed with ClustalW and Boxshade. Amino acids that are identical between the repeat units are displayed as white letters on black background. The RVDs are highlighted in bold-italic font. The asparagine (n) in position 16 (displayed in lower case font) is found only in repeat 10, while all other Brg11 repeats have lysine (k) at this position. Numbers above the sequence (read from top to bottom; 5, 10, 15, 20, 25, 30, 35) indicate the position of the given amino acid within the repeat.

Brg11



```
                 1    1    2    2    3    3
            5    0    5    0    5    0    5
 1      LTPQQVVAIASNTGGkRALEAVCVQLPVLRAAPYR
 2      LSTEQVVAIASNKGGkQALEAVKAHLLDLLGAPYV
 3      LDTEQVVAIASHNGGkQALEAVKADLLDLRGAPYA
 4      LSTEQVVAIASHNGGkQALEAVKADLLELRGAPYA
 5      LSTEQVVAIASHNGGkQALEAVKAHLLDLRGVPYA
 6      LSTEQVVAIASHNGGkQALEAVKAQLLDLRGAPYA
 7      LSTAQVVAIASNGGGkQALEGIGEQLLKLRTAPYG
 8      LSTEQVVAIASHDGGkQALEAVGAQLVALRAAPYA
 9      LSTEQVVAIASNKGGkQALEAVKAQLLELRGAPYA
10      LSTAQVVAIASHDGGnQALEAVGTQLVALRAAPYA
11      LSTEQVVAIASHDGGkQALEAVGAQLVALRAAPYA
12      LNTEQVVAIASSHGGkQALEAVRALFPDLRAAPYA
13      LSTAQLVAIASNPGGkQALEAVRALFRELRAAPYA
14      LSTEQVVAIASNHGGkQALEAVRALFRGLRAAPYG
15      LSTAQVVAIASSNGGkQALEAVWALLPVLRATPYD
16      LNTAQIVAIASHDGGkPALEAVWAKLPVLRGAPYA
```

AvrBs3



```
                 1    1    2    2    3
            5    0    5    0    5    0
 1      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
 2      LTPQQVVAIASNGGGKQALETVQRLLPVLCQAHG
 3      LTPQQVVAIASNSGGKQALETVQRLLPVLCQAHG
 4      LTPEQVVAIASNGGGKQALETVQRLLPVLCQAHG
 5      LTPEQVVAIASNIGGKQALETVQALLPVLCQAHG
 6      LTPEQVVAIASNIGGKQALETVQALLPVLCQAHG
 7      LTPEQVVAIASNIGGKQALETVQALLPVLCQAHG
 8      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
 9      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
10      LTPQQVVAIASNGGGKQALETVQRLLPVLCQAHG
11      LTPEQVVAIASNSGGKQALETVQALLPVLCQAHG
12      LTPEQVVAIASNSGGKQALETVQRLLPVLCQAHG
13      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
14      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
15      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
16      LTPQQVVAIASNGGGRPALETVQRLLPVLCQAHG
17      LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
```

Supplementary Figure 5

Representative amino acid sequences used in the trimer test. (a) The amino acid sequence for Brg11 repeat 1 as a trimer is shown in bold font. The central repeat of the trimer is underlined to allow each repeat to be identified. Flanking sequences correspond to the amino acid sequences of AvrBs3. Below the amino acid sequence the corresponding nucleotide sequence is shown. Terminal BpiI recognition sites, indicated in lower case letter, facilitate compatibility with the TALE binding domain assembly toolkit (Morbitzer et al., 2011). In (b) the Brg11 trimer for repeat 1 (boldface black font) is shown in the context of the TALE AvrBs3. N-terminal, repeat and C-terminal regions of AvrBs3 are separated from each other. The repeats are numbered. The repeats 0 and -1 constitute the N-terminal, 1-17 the core and +1 and +2 the C-terminal repeats.

(a)

>Brg11 repeat 1 trimer

EDAETVQRLLPVLCQAHG**LTPQQVVAIASNTGGKRALEAVCVQLPVLRAAPYRLTPQQVVAIASNTGGKRALEAVCVQLPVLRAAPYRLTPQQVVAIASNTGGKRALEAVCVQLPVLRAAPYR**LTPEQVVAIASQS

>*brg11* repeat 1 trimer

*gaagac*GCTGAGACGGTGCAGCGCCTGCTGCCGGTGCTGTGCCAGGCCCATGGC**CTGACCCCGCAGCAGGTGGTGGCGATCGCGTCGAACACCGGCGGCAAGCGCGCCCTGGAAGCGGTGTGCGTGCAATTACCGGTGCTGCGCGCCGCGCCGTACCGTTTAACTCCTCAGCAGGTGGTGGCCATCGCGAGCAACACCGGCGGCAAGCGCGCCCTGGAGGCGGTGTGCGTGCAACTTCCGGTGCTGCGCGCGGCCCCGTACCGTTTAACTCCCCAACAAGTGGTGGCCATCGCCTCGAATACCGGCGGCAAGCGTGCGCTGGAGGCCGTGTGCGTGCAGCTGCCGGTGCT**GCGCGCCGCCCCGTATCGCCTGACCCCGGAACAGGTGGTGGCCATCGCAAGCCA*gtcttc*

(b)

>Brg11 repeat 1 trimer embedded in the AvrBs3 repeat array

MDPIRSRTPSPARELLPGPQPDGVQPTADRGVSPPAGGPLDGLPARRTMSRTRLPSPPAPSP
AFSAGSFSDLLRQFDPSLFNTSLFDSLPPFGAHHTEAATGEWDEVQSGLRAADAPPPTMRVA
VTAARPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGH
GFTHAHIVALSQHPAALGTVAVKYQDMIAALPE

-1 ATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQ
0 LDTGQLLKIAKR GGVTAVEAVHAWRNALTGAPLN
1 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
2 LTPQQVVAIASNGGGKQALETVQRLLPVLCQAHG
3 LTPQQVVAIASNSGGKQALETVQRLLPVLCQAHG
4 LTPEQVVAIASNGGGKQALETVQRLLPVLCQAHG
5 LTPQQVVAIASNTGGKRALEAVCVQLPVLRAAPYR
6 LTPQQVVAIASNTGGKRALEAVCVQLPVLRAAPYR
7 LTPQQVVAIASNTGGKRALEAVCVQLPVLRAAPYR
8 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
9 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
10 LTPQQVVAIASNGGGKQALETVQRLLPVLCQAHG
11 LTPEQVVAIASNSGGKQALETVQALLPVLCQAHG
12 LTPEQVVAIASNSGGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
15 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
16 LTPQQVVAIASNGGGRPALETVQRLLPVLCQAHG
17 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
+1 LTPQQVVAIASNGGGRPALESIVAQLSRPDPALAA
+2 LTNDHLVALACLGGRPALDAVKKGLPHAPALIKRT

NRRIPERTSHRVADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGMSRHGLLQLFRRVGVTELE
ARSGTLPPASQRWDRILQASGMKRAKPSPTSTQTPDQASLHAFADSLERDLDAPSPMHEGDQ
TRASSRKRSRSDRAVTGPSAQQSFEVRVPEQRDALHLPLSWRVKRPRTSIGGGLPDPGTPTA
ADLAASSTVMREQDEDPFAGAADDFPAFNEEELAWLMELLPQ

Supplementary Figure 6

Reciprocal exchange of RVDs between Brg11 repeats 12 and 8/11 leading to exchange of DNA recognition specificity. RVDs were swapped between repeats 12 and 8/11 of Brg11 and these novel repeats tested again in the trimer test to clarify the importance of RVDs in the base specificity of Brg11 repeats. The novel repeats are defined as Brg11 repeat 8/11 SH and Brg11 repeat 12 HD, with WT used to identify the wild-type repeats in each case. An alignment (Clustal Omega; Boxshade) of the wild-type and chimeric repeats is shown in (a). Orange font is used to indicate the RVDs. Amino acid positions are indicated with numbers above the corresponding positions in the alignment (read from top to bottom 1, 11, 21, 31). Trimers consisting of three identical copies of each repeat were created and embedded into TALE AvrBs3, replacing repeats 5-7, in the wild-type AvrBs3 a trimer of repeats bearing the RVD NI. *35S*-promoter driven genes encoding AvrBs3-embedded Brg11 repeats were co-delivered into *N. benthamiana* leaves via *A. tumefaciens* mediated T-DNA transfer along with an *uidA* reporter gene that is either under transcriptional control of the *Bs3* promoter (3xA) or of a derivative containing a triple guanine (3xG), cytosine (3xC) or thymine (3xT) at the position corresponding to the repeat trimer. Leaf discs were harvested forty-eight hours post-infiltration and stained to visualise reporter activity. The intensity of the staining corresponds to promoter activity and is used as a proxy indicator of binding at the promoter. Results are shown in (b). The repeat tested is indicated above the column of leaf discs in each case. The promoter derivate used is indicated with the abbreviation ‘3xN’ in each row.

Supplementary Figure 7

Nucleotide sequences of *brg11* repeats used to assemble three identical, tandem-arranged repeat blocks that were tested in the context of dTALE$_{\text{EBE Brg11}}$ (Fig. 5). BsaI sites used to assemble the repeats are distinguished with lower case letters.

>*brg11* repeat A1

```
ggtctcTTACGCCGCAGCAGGTGGTGGCCATTGCCAGCAACACTGGCGGCAAGCGGGCGTTG
GAGGCGGTCTGTGTGCAATTGCCCGTGCTGCGCGCGGCCCCCTATAGAgagacc
```

>*brg11* repeat 2

```
ggtctcTATAGACTGAGCACCGAGCAGGTGGTGGCGATCGCCAGCAACAAAGGCGGCAAACA
GGCGCTGGAGGCGGTCAAGGCGCACCTGCTGGATCTGCTCGGGGCACCCTATGTGAgagacc
```

>*brg11* repeat 3

```
ggtctcTTGTGCTCGACACCGAGCAAGTCGTGGCCATTGCCAGCCACAATGGCGGCAAGCAG
GCACTGGAGGCGGTCAAGGCGGACCTGCTGGATTTGCGCGGAGCACCCTAgagacc
```

>*brg11* repeat 4

```
ggtctcTCCCTATGCGTTGAGTACCGAGCAAGTCGTGGCCATTGCCAGCCACAATGGCGGCA
AACAGGCACTGGAGGCGGTCAAGGCGGACCTGCTGGAACTGCGCGGAGAgagacc
```

>*brg11* repeat 5

```
ggtctCTGGAGCACCCTATGCGTTGAGTACCGAGCAGGTGGTGGCCATCGCCAGCCACAATG
GCGGCAAGCAGGCTCTGAgagacc
```

>*brg11* repeat 6

```
ggtctcTTCTGGAAGCGGTCAAGGCGCACCTGCTGGATCTGCGTGGAGTGCCCTATGCGTTG
AGTACCGAGCAAGTCGTGGCCATCGCCAGCCACAATGGCGGCAAACAGGCGCTGGAGGCGGT
CAAGGCACAACAgagacc
```

>*brg11* repeat 7

```
ggtctcTCAACTGCTGGATCTGCGCGGAGCACCCTATGCGTTGAGTACCGCCCAGGTGGTCG
CCATCGCCAGTAACGGCGGCGGCAAACAGGCGCTGGAGGGGATTGGCGAACAGCTGCTGAAA
CTGCGGACTGCACCCTATGGGAgagacc
```

>*brg11* repeat 8

```
ggtctcTTGGGCTGAGTACCGAGCAGGTGGTCGCCATCGCCAGCCATGATGGCGGCAAACAA
GCGTTGGAAGCGGTCGGTGCGCAGTTGGTGGCACTGCGCAgagacc
```

>*brg11* repeat 9

ggtctcTGCGCGCGGCGCCTTATGCGCTGAGCACCGAGCAGGTGGTGGCCATCGCCAGCAAC
AAGGGTGGCAAGCAGGCACTGGAGGCGGTCAAGGCGCAACTGCTGGAGAgagacc

>*brg11* repeat 10B N16K

ggtctcTGGAGCTGCGCGGAGCACCCTATGCGTTGAGTACCGCCCAGGTGGTCGCCATCGCC
AGCCATGATGGCGGCAAGCAAGCACTGGAAGCGGTCGGTACGCAGTTGGTGGCGCTGCGCGC
GGCGCCTTATGCGCTGAGCACCGAGCAGGTGGTAGCGAgagacc

>*brg11* repeat 11B

ggtctctAGCGATCGCCAGTCATGATGGCGGCAAACAGGCATTGGAAGCGGTCGGGGCGCAG
TTGGTGGCGCTGCGCGCAGCGCCTTATGCGCTGAACAgagacc

>*brg11* repeat 12

ggtctcTGAACACCGAGCAGGTGGTGGCCATCGCCAGCAGCCATGGCGGCAAACAGGCGCTG
GAGGCGGTCCGGGCACTGTTTCCGGATCTGCGCGCGGCGCCTTATGCGCTGAGTACCAgaga
cc

>*brg11* repeat 13

ggtctcTTACCGCGCAACTGGTGGCCATCGCCAGCAACCCTGGCGGCAAACAAGCGCTGGAG
GCAGTCCGGGCACTGTTTCGGGAGCTGCGGGCGGCGCCCTATGAgagacc

>*brg11* repeat 14

ggtctcTTATGCGTTGAGCACCGAGCAAGTGGTAGCCATCGCCAGCAACCATGGCGGCAAAC
AGGCGCTGGAGGCCGTCCGGGCACTGTTTCGGGGTCTGCGAGCCGCGCCTTACGGAAgagac
c

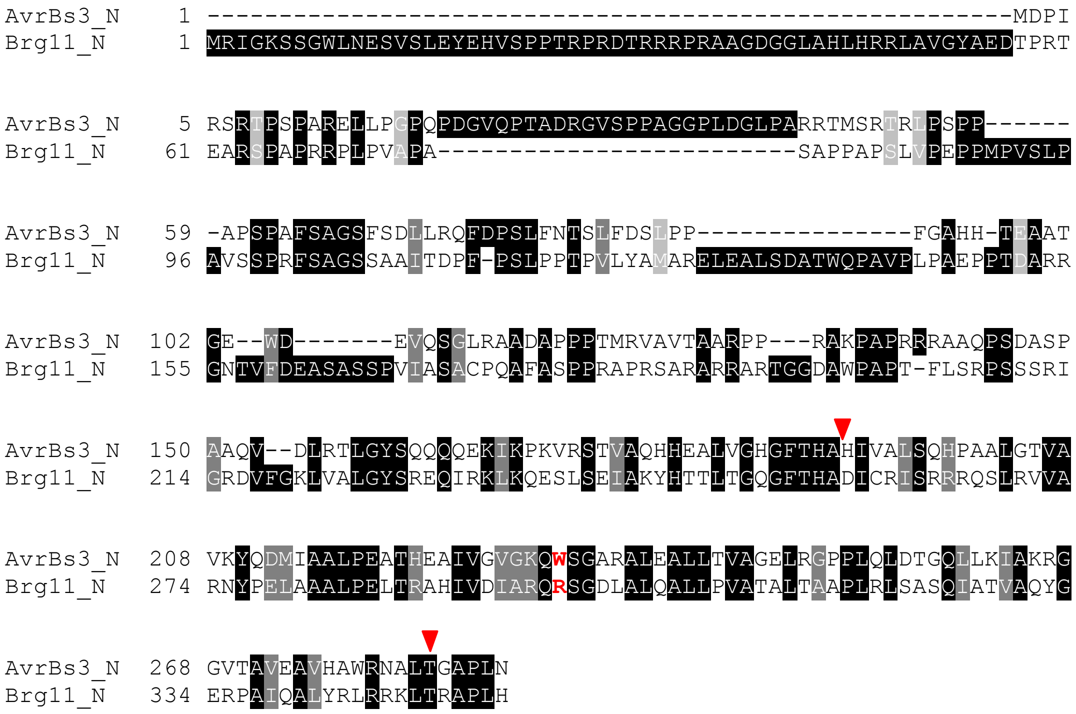
>*brg11* repeat 15/16

ggtctcTCGGACTCAGTACCGCGCAGGTGGTGGCCATCGCGAGCAGTAATGGCGGCAAACAG
GCGCTGGAGGCGGTCTGGGCGCTGCTACCGGTGTTGCGCGCCACACCCTACGATCTGAATAC
CGCGCAGATCGTGGCCATCGCCAGCCATGATGGGGGCAgagacc

Supplementary Figure 8

Alignment of N-terminal regions AvrBs3 and Brg11 regions (a) and sequence of chimeras NTC1 and NTC2 (b).

(a) Comparison of the N-terminal regions of AvrBs3 and Brg11, including repeats 0 and -1. The alignment was constructed with ClustalW and Boxshade. Amino acids that are identical between the repeat units are displayed as white letters on black background. Red triangles indicate the breakpoints for the creation of N-terminal chimera 1 (NTC1, see also Fig. S8b). Red font is used to highlight tryptophan 232 of AvrBs3 and the corresponding arginine of Brg11.



```
AvrBs3_N    1 -------------------------------------------------------MDPI
Brg11_N     1 MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAAGDGGLAHLHRRLAVGYAEDTPRT

AvrBs3_N    5 RSRTPSPARELLPGPQPDGVQPTADRGVSPPAGGPLDGLPARRTMSRTRLPSPP------
Brg11_N    61 EARSPAPRRPLPVAPA-------------------------SAPPAPSLVPEPPMPVSLP

AvrBs3_N   59 -APSPAFSAGSFSDLLRQFDPSLFNTSLFDSLPP---------------FGAHH-TEAAT
Brg11_N    96 AVSSPRFSAGSSAAITDPF-PSLPPTPVLYAMARELEALSDATWQPAVPLPAEPPTDARR

AvrBs3_N  102 GE--WD-------EVQSGLRAADAPPPTMRVAVTAARPP---RAKPAPRRRAAQPSDASP
Brg11_N   155 GNTVFDEASASSPVIASACPQAFASPPRAPRSARARRARTGGDAWPAPT-FLSRPSSSRI

                                                          ▼
AvrBs3_N  150 AAQV--DLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVA
Brg11_N   214 GRDVFGKLVALGYSREQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVA

AvrBs3_N  208 VKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRG
Brg11_N   274 RNYPELAAALPELTRAHIVDIARQRSGDLALQALLPVATALTAAPLRLSASQIATVAQYG

                             ▼
AvrBs3_N  268 GVTAVEAVHAWRNALTGAPLN
Brg11_N   334 ERPAIQALYRLRRKLTRAPLH
```

(b) Annotated amino sequences of the N-terminal chimeras NTC1 and NTC2. The sequences corresponding to AvrBs3 are shown in regular font, the Brg11 parts are shown in boldface, italics font (for annotation of AvrBs3 and Brg11 see Fig. S1). While NTC1 contains only a part of the Brg11 N terminus, chimera NTC2 contains the complete Brg11 N terminus, including repeats 0 and -1. Below each amino acid sequence the corresponding nucleotide sequence is given from the start codon up to and including the first repeat of the *avrBs3* core binding domain.

>NTC1

MDPIRSRTPSPARELLPGPQPDGVQPTADRGVSPPAGGPLDGLPARRTMSRTRLPSPPAPSP
AFSAGSFSDLLRQFDPSLFNTSLFDSLPPFGAHHTEAATGEWDEVQSGLRAADAPPPTMRVA
VTAARPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGH
GFTHA***DICRISRRRQSLRVVARNYPELAAALPE***

-1 ***LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR***
0 ***LSASQIATVAQY GERPAIQALYRLRRKLT***GAPLN
1 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
2 LTPQQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
3 LTPQQVVAIAS**NS**GGKQALETVQRLLPVLCQAHG
4 LTPEQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
5 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
6 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
7 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
8 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
9 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
10 LTPQQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
11 LTPEQVVAIAS**NS**GGKQALETVQALLPVLCQAHG
12 LTPEQVVAIAS**NS**GGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
15 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
16 LTPQQVVAIAS**NG**GGRPALETVQRLLPVLCQAHG
17 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
+1 LTPQQVVAIASNGGGRPALESIVAQLSRPDPALAA
+2 LTNDHLVALACL GGRPALDAVKKGLPHAPALIKRT

NRRIPERTSHRVADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGMSRHGLLQLFRRVGVTELE
ARSGTLPPASQRWDRILQASGMKRAKPSPTSTQTPDQASLHAFADSLERDLDAPSPMHEGDQ
TRASSRKRSRSDRAVTGPSAQQSFEVRVPEQRDALHLPLSWRVKRPRTSIGGGLPDPGTPTA
ADLAASSTVMREQDEDPFAGAADDFPAFNEEELAWLMELLPQ

>NTC1

(*avrBs3*)ATGGATCCCATTCGTTCGCGCACACCAAGTCCTGCCCGCGAGCTTCTGCCCGGA
CCCCAACCCGATGGGGTTCAGCCGACTGCAGATCGTGGGGTGTCTCCGCCTGCCGGCGGCCC
CCTGGATGGCTTGCCCGCTCGGCGGACGATGTCCCGGACCCGGCTGCCATCTCCCCCTGCCC
CCTCACCTGCGTTCTCGGCGGGCAGCTTCAGTGACCTGTTACGTCAGTTCGATCCGTCACTT
TTTAATACATCGCTTTTTGATTCATTGCCTCCCTTCGGCGCTCACCATACAGAGGCTGCCAC
AGGCGAGTGGGATGAGGTGCAATCGGGTCTGCGGGCAGCCGACGCCCCCCCACCCACCATGC
GCGTGGCTGTCACTGCCGCGCGGCCGCCGCGCGCCAAGCCGGCGCCGCGACGACGTGCTGCG
CAACCCTCCGACGCTTCGCCGGCCGCGCAGGTGGATCTACGCACGCTCGGCTACAGCCAGCA
GCAACAGGAGAAGATCAAACCGAAGGTTCGTTCGACAGTGGCGCAGCACCACGAGGCACTGG
TCGGCCATGGGTTTACACATGCC(*brg11*→)GATATATGTCGAATTTCTAGGAGAAGACAAA
GCCTTCGAGTAGTCGCTCGTAATTACCCTGAACTGGCAGCAGCTCTACCAGAATTGACTAGA
GCACATATCGTGGATATTGCAAGACAAAGGTCTGGAGATTTGGCATTACAAGCACTTCTCCC
TGTGGCAACAGCTCTCACAGCTGCACCTTTGAGGCTTTCTGCTAGCCAAATCGCTACAGTCG
CACAATATGGTGAGAGGCCAGCTATCCAAGCACTTTATAGACTTCGAAGGAAACTGACG(*avrBs3*
GAPLN Repeat 1 →)GGTGCCCCCCTGAACCTTACGCCGGAGCAGGTGGTGGCCA
TCGCCAGCCACGATGGCGGCAAGCAGGCGCTGGAGACGGTGCAGCGGCTGTTGCCGGTGCTG
TGCCAGGCCCATGGC...

>NTC2

***MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAAGDGGLAHLHRRLAVGYAEDTPRTEA
RSPAPRRPLPVAPASAPPAPSLVPEPPMPVSLPAVSSPRFSAGSSAAITDPFPSLPPTPVLY
AMARELEALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIASACPQAFASPPRAPRS
ARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSREQIRKLKQESLSEIAKYHTT
LTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE***

-1 ***LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR***
0 ***LSASQIATVAQY GERPAIQALYRLRRKLTRAPLH***
1 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
2 LTPQQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
3 LTPQQVVAIAS**NS**GGKQALETVQRLLPVLCQAHG
4 LTPEQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
5 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
6 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
7 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
8 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
9 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
10 LTPQQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
11 LTPEQVVAIAS**NS**GGKQALETVQALLPVLCQAHG
12 LTPEQVVAIAS**NS**GGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
15 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
16 LTPQQVVAIAS**NG**GGRPALETVQRLLPVLCQAHG
17 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
+1 LTPQQVVAIASNGGGRPALESIVAQLSRPDPALAA
+2 LTNDHLVALACL GGRPALDAVKKGLPHAPALIKRT

NRRIPERTSHRVADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGMSRHGLLQLFRRVGVTELE
ARSGTLPPASQRWDRILQASGMKRAKPSPTSTQTPDQASLHAFADSLERDLDAPSPMHEGDQ
TRASSRKRSRSDRAVTGPSAQQSFEVRVPEQRDALHLPLSWRVKRPRTSIGGGLPDPGTPTA
ADLAASSTVMREQDEDPFAGAADDFPAFNEEELAWLMELLPQ

>NTC2

*(brg11)*ATGAGAATTGGCAAGTCCTCAGGATGGTTAAATGAATCTGTGTCATTAGAATATG
AACATGTGTCACCACCAACTAGGCCTCGAGATACTCGAAGGCGACCTAGGGCAGCTGGAGAT
GGAGGTCTTGCACATTTGCACAGGAGGTTAGCTGTGGGCTATGCAGAGGATACTCCAAGAAC
CGAAGCTAGGAGTCCAGCTCCACGTAGGCCTCTTCCAGTGGCTCCAGCAAGCGCACCTCCAG
CTCCATCACTTGTTCCAGAGCCACCAATGCCAGTCTCCCTTCCCGCTGTTAGTTCTCCACGT
TTTAGCGCTGGTAGTAGTGCCGCTATCACAGATCCTTTTCCTTCACTTCCTCCAACACCAGT
GTTATATGCTATGGCTAGAGAGTTGGAGGCATTAAGTGACGCTACTTGGCAGCCAGCAGTGC
CATTACCTGCCGAGCCACCTACTGATGCAAGGAGAGGCAACACAGTTTTTGACGAAGCATCT
GCATCATCTCCAGTTATTGCATCCGCTTGCCCACAGGCTTTCGCTTCACCACCAAGAGCACC
TCGATCTGCTAGAGCTAGAAGAGCTAGGACAGGCGGAGATGCATGGCCAGCTCCTACTTTTC
TCAGCAGACCAAGCTCTTCTAGGATTGGTAGAGATGTATTTGGAAAGCTAGTCGCACTCGGG
TATAGTAGGGAACAGATCAGAAAGTTGAAACAAGAGTCTCTTTCTGAAATAGCAAAGTACCA
TACCACACTCACCGGTCAAGGATTCACCCATGCCGATATATGTCGAATTTCTAGGAGAAGAC
AAAGCCTTCGAGTAGTCGCTCGTAATTACCCTGAACTGGCAGCAGCTCTACCAGAATTGACT
AGAGCACATATCGTGGATATTGCAAGACAAAGGTCTGGAGATTTGGCATTACAAGCACTTCT
CCCTGTGGCAACAGCTCTCACAGCTGCACCTTTGAGGCTTTCTGCTAGCCAAATCGCTACAG
TCGCACAATATGGTGAGAGGCCAGCTATCCAAGCACTTTATAGACTTCGAAGGAAACTTACT
AGAGCCCCTCTTCAC(*avrBs3* Repeat 1 →)CTTACGCCGGAGCAGGTGGTGGCCATCG
CCAGCCACGATGGCGGCAAGCAGGCGCTGGAGACGGTGCAGCGGCTGTTGCCGGTGCTGTGC
CAGGCCCATGGC...

Supplementary Figure 9

Amino acid sequences of RipTALs (RipTALI-1 – RipTALI-15; Table S2) analysed in this study. N-terminal, repeat and C-terminal regions are separated from each other. The repeats are numbered. The repeats 0 and -1 constitute the N-terminal repeats and +1 and +2 the C-terminal repeats. The RVDs are marked as boldface black letters on grey background. RVDs that differ from that ones found in Brg11 are shown in boldface blue letters. For amplification of the *RipTALs* the primers brg11-CACC-ATG and brg11-no stop (Table S1) were used. Therefore all RipTALs are sequence identical in the corresponding N- and C-terminal amino acid sequences. The sequences can be found at NCBI under the GenBank accession numbers KC405066-KC405080.

>RipTALI-1 (from *R. solanacearum* strain JS53; GenBank KC405066)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGYAEDT
PRTGARSPAPRRSPPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPF
PSLPPTPVLYAMARELEALSDATWQPAVPLPSEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKADLLDLRGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
6 LSTAQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGKQALEAAGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LSTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAVRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
15 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
16 LNTAQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRRPNNAGQPSEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTPADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-2 (from *R. solanacearum* strain HB53; GenBank KC405067)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGYAEDT
PRTGARSPAPRRSLPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPF
PSLPPTPVLYAMARELEALSDATWQPAVPLPSEPPTDTRRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQLRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
+2 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
6 LSTEQVVAIAS**NN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGKQALEAVGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LSTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAIRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
15 LSTAQVVAVAS**SN**GGKQALEAVWALLPVLRATPYD
16 LNTAQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYE
+1 LSTAQVVAIACISGQQALEVIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRSPNNAGQPFEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTSADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-3 (from *R. solanacearum* strain GX53; GenBank KC405068)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAAGGGGLAHLPRRLAVGYAEDT
PRTGARSPAPRRSLPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPF
PSLPPTPVLDAMTRELEALSDATWQPAVPLPSEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFAHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLLGAPYV
4 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
5 LSTEQVVAIAS**HN**GGKLALEAVKAHLLDLRGAPYA
6 LSTEQVVAIAS**HN**GGKQALEAVKTQLLELRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGKQALEAVGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LSTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAVRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
15 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
16 LNTAQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIQ

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRSPNNAGQPFEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTPADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-4 (from *R. solanacearum* strain GX55; GenBank KC405069)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGYAEDT
PRTGARSPAPRRSLPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPF
PSLPPTPVLYAMARELEALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLELRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAHLLELRGAPYA
6 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAG**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGKQALEAVGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LSTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAIRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
15 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
16 LNTAQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRSPNNAGQPFEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTPADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-5 (from *R. solanacearum* strain HN515; GenBank KC405070)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGYAEDT
PRTGARSPAPRRPLPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPF
SSLPPTPVLYAMARELKALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRRPNNAGQPSEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
HGTPTPADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-6 (from *R. solanacearum* strain GD52; GenBank KC405071)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGYAEDT
PRTGARSPAPRRSPPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPF
PSLPPTPVLYAMARELEALSDATWQPAVPLPSEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLELRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAHLLDLRGAPYA
6 LSTAQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLRKLRTAPYG
8 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
9 LSTAQVVAIAS**HD**GGKQALEAVGTQLVALRAAPYA
10 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
11 LSTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
12 LSTAQLVAIAS**NP**GGKQALEAVRALFRELRAAPYA
13 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
14 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
15 LNTAQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRRPNNAGQPSEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTSADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-7 (from *R. solanacearum* strain FJ41; GenBank KC405072)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAAGDGGLAHLHRRLAVGYAEDT
PRTEARSPAPRRPLPVAPASAPPAPSLVPEPPMPVSLPAVSSPRFSAGSSAAITDPF
PSLPPTPVLYAMARELEALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLELRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAHLLDLRGVPYA
6 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGNQALEAVGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LNTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAVRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
15 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
16 LNTAQIVAIAS**HD**GGKPALEAVWAKLPVLRGAPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQATRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRSPNNAGQPSEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTSADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-8 (from *R. solanacearum* strain SD58; GenBank KC405073)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGYAEDT
PRTGARSPAPRRPLPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPF
SSLPPTPVLYAMARELKALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
 0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
 1 LTPQQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRRPNNAGQPSEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTSADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-9 (from *R. solanacearum* strain UW360; GenBank KC405074; a partial sequence that does not cover the complete repeat array has been deposited as EF435034)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAAGGGGLAHLHRRLAVDYAEDT
PRTGARSPAPRRPLPVAPASTPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPF
SSLPPTPVLYAMARELEALSDATWQPAVPLPSEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
6 LSTAQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGKQALEAAGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LSTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAVRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRELRAAPYA
15 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
16 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
17 LNTAQVVAIAS**HY**GGKPALEAVWAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLKS
PGMAGQSACSPHRKRPAETAIAPRSIRRRPNNAGQPSEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTSADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-10 (from *R. solanacearum* strain UW148; GenBank KC405075)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAAGDGGLAHLHRRLAVGYAEDT
PRTEARSPAPRRPLPVAPASAPPAPSLVPEPPMPVSLPAVSSPRFSAGSSAAITDPF
PSLPPTPVLYAMARELEALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLELRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAHLLDLRGVPYA
6 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGNQALEAVGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LNTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAVRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
15 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
16 LNTAQIVAIAS**HD**GGKPALEAVWAKLPVLRGAPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQATRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRSPNNAGQPSEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTSADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-11 (from *R. solanacearum* strain GX526; GenBank KC405076)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGYAEDT
PRTGARSPAPRRSLPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPF
PSLPPTPVLYAMARELEALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLELRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
6 LSTAQVVAIAS**QN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGKQALEAVGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LSTEQVVAIAS**SH**GGKQALEAVRALLPVLRATPYD
13 LNTAQVVAIAS**HD**GGKPALEAVRAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRRPNNAGQPSEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTSADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-12 (from *R. solanacearum* strain GX528; GenBank KC405077)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGFAEDT
PRTGARSPAPRRSLPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPF
PSLPPTPVLYAMARELEALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLELRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAHLLELRGAPYA
6 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAG**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGKQALEAVGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LSTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAIRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
15 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
16 LNTAQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRSPNNAGQPFEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTPADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-13 (from *R. solanacearum* strain GZ522; GenBank KC405078)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAAGGGGLAHLHRRLAVGYAEDT
PRTGARSPAPRRPLPVAPASAPPAPSLVPEPAMPVSLPVVSSPRFSAGSSAAITDPF
PSLPPTPVLYAMARELEALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLELRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAQLLELRGAPYA
6 LSTEQVVAIAS**HN**GSKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSAAQVVAIAS**HD**GGKQALEAVGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LSTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAIRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
15 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
16 LNTAQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYE
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRSPNNAGQPSEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGAIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTSADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-14 (from *R. solanacearum* strain GD45; GenBank KC405079)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAAGGGGLAHLHRRLAVGYAEDT
PRTGARSPAPRRPLPVAPASAPPAPSLVPEPAMPVSLPVVSSPRFSAGSSAAITDPF
PSIPPTPVLYAMARELEALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLELRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKAHLLDLRGVPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
6 LSTAQVVAIAG**NG**GGKQALEGIGEQLLKLRTAPYG
7 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
8 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
9 LSTAQVVAIAS**HD**GGNQALEAVGTQLVALRAAPYA
10 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
11 LNTEQVVAIAS**NP**GGKQALEAVRALFPDLRAAPYA
12 LSTAQLVAIAS**NP**GGKQALEAIRALFRELRAAPYA
13 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
14 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
15 LNTAQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKVPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRSPNNAGQPSEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTSADLASFEAEPFGVGPLDFHLDWLLQILET

>RipTALI-15 (from *R. solanacearum* strain ICPM11110; GenBank KC405080)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAASDGGLAHLHRRLAVGYAEDT
PRTGARSPAPRRSPPVAPASAPPAPSLVPEPPMPVSLPVVSSPRFSAGSSAAITDPF
PSLPPTPVLYAMARELEALSDATWQPAVPLPSEPPTDARRGNTVFDEASASSPVIAS
ACPQAFASPPRAPRSARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSR
EQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE

-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQYGERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKADLLDLRGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
6 LSTAQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGKQALEAAGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LSTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAVRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
15 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
16 LNTAQVVAIAS**HD**GGKPALEAVWAKLPVLRGVPYA
+1 LSTAQVVAIACISGQQALEAIEAHMPTLRQAPHS
+2 LSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIR

REKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQTTRPALLRLLS
SVGVTEIEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGS
PGMAGQSACSPHRKRPAETAIAPRSIRRRPNNAGQPSEPWPDQLAWLQRRKRTARSH
IRADSAASVPANLHLGTRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPD
PGTPTPADLASFEAEPFGVGPLDFHLDWLLQILET

Supplementary Figure 10

Alignment of all RipTALs analysed in this study. Only polymorphic positions are displayed. Amino acid position relative to Brg11 is indicated with numbers above the relevant column, to be read vertically (from the far left 27, 28, 29, 38 etc.). The amino acids found in Brg11 are distinguished in each case with bold font. Dashes indicate positions occupied by the identical amino acid as that found in Brg11. White boxes indicate that the equivalent position is missing from a certain RipTAL.

A colour code is used to distinguish the predicted structural regions of each polypeptide. Yellow corresponds to N- and C-terminal regions and all other colours to individual repeats of the repeat domain. Repeats bearing identical RVDs in Brg11 are identically coloured. White lines divide sequences corresponding to separate repeats. Grey horizontal bars and written descriptions below the alignment indicate sequences corresponding to the N-terminus, repeats and C-terminus of each RipTAL, while black lines and numbers below the alignment indicate the positions of individual repeats as they occur within the repeat region of Brg11. Where polymorphisms occur within the repeat region the relative position within the single repeat is indicated with numbers, to be read horizontally, above the appropriate column. Polymorphisms in the RVD positions (12 and 13) are highlighted with red lettering. A black triangle in between repeats 13 and 15 in the row corresponding to RipTALI-9 indicates the complete duplication of repeat 14, relative to Brg11, in this RipTAL.

A cartoon representation of Brg11 is shown below the alignment for reference. Ellipses indicate the repeats of the core binding domain; a red semi-circle indicates a putative transcriptional activation domain; red rectangles indicate putative nuclear localisation signals. Colouring for the N- and C-terminal and repeat regions is the same as in the alignment. Numbers above the ellipses indicate the position of the respective repeat within the repeat domain.

Supplementary Figure 11

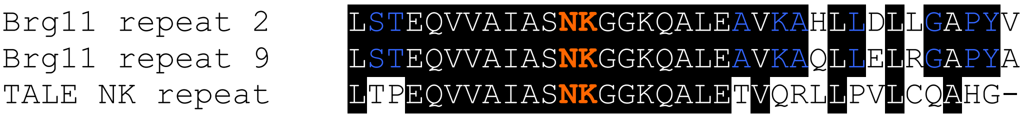
Zero base preferences of RipTALI-6, -9, -11 and -14. These four RipTALs were tested for ability to activate *Bs3p* promoter derivatives bearing their own predicted EBEs with an A_0, G_0, C_0 or T_0 upstream of an *uidA* reporter gene. *35-S* driven genes encoding the RipTALs were confiltrated along with the promoter constructs into *N. benthamiana* leaves via *A. tumefaciens*-mediated transient transformation. Leaf discs were harvested forty-eight hours post-infiltration and stained to visualise reporter activity. Sequences of the EBEs used are shown on the right in each case. White letters with a black border indicate the nucleotide occupying the zero position in each case. The RipTAL tested against each promoter is indicated on the left.

Supplementary Figure 12

Comparison of predicted Brg11 binding sequences from this study and from Streubel *et al.*, 2012. Numbers indicate repeats of the Brg11 core binding domain and the corresponding predicted binding partners are displayed in columns underneath as black letters on a grey background. A black line divides sequences from the two studies. In each case the row of letters closest to the line indicates the best predicted match and then alternative matches are displayed, in order of preference, extending above (this study) or below (Streubel *et al.*, 2012) the nucleotides of the best match sequence. Bold font is used to indicate the sequence used for $\mathrm{EBE}_{\mathrm{Brg11}}$ throughout this study.

Supplementary Figure 13

Alignment of Brg11 repeats 2 and 9 and a consensus AvrBs3 repeat. The repeats are displayed with NK as the RVD (orange font). Residues common to Brg11 repeats 2 and 9 and not found in an equivalent AvrBs3 repeat are shown in blue font. The alignment was constructed with ClustalW and Boxshade.



```
Brg11 repeat 2     LSTEQVVAIASNKGGKQALEAVKAHLLDLLGAPYV
Brg11 repeat 9     LSTEQVVAIASNKGGKQALEAVKAQLLELRGAPYA
TALE NK repeat     LTPEQVVAIASNKGGKQALETVQRLLPVLCQAHG-
```

Supplementary Tables

Table S1: Primers used in this study.

Name of Primer	**Application**	**Sequence**
brg11-CACC-ATG	Creation of *hpx17* and *brg11* pENTR-D constructs	CAC CAT GAG AAT AGG CAA ATC AAG CGG TTG GTT GAA C
brg11-no Stop	Creation of *hpx17* and *brg11* pENTR-D constructs	CGT TTC CAA TAT TTG CAG AAG CCA GTC G
brg11_mut-NLSI_fwd	Mutation of putative *hpx17* nuclear localisation signals	CCT GCC GAG ACG GCC ATC GCA CCG C
brg11_mut-NLSI_HindIIIrev	Mutation of putative *hpx17* nuclear localisation signals	CCA GTA AGC TTG TGG TGA GCA AGC CGA TTG CC
brg11_mut-NLSII_BamHIfwd	Mutation of putative *hpx17* nuclear localisation signals	CCG GAT CCA TGG ACC GCT CGT TCG CAC ATA CG
brg11_mut-NLSII_rev	Mutation of putative *hpx17* nuclear localisation signals	TTG GAG CCA TGC CAG TTG AT
hpx17_Cterm_fwd	Truncation of *hpx17* for localisation studies	TTG CCGGTG AAG GCG ATC CGG
hpx17_Cterm_rev	Truncation of *hpx17* for localisation studies	CAT GGT GAA GGG GGC GGC CGC
hpx17_Nterm_fwd	Truncation of *hpx17* for localisation studies	AAG GGT GGG CGC GCC GAC CCA
hpx17_Nterm-rev	Truncation of *hpx17* for localisation studies	CTG CGG TGT GAG ATG CAG CGG TGC
avrbs3 NtermCACCFwd	Subcloning of full *avrbs3* N-terminal encoding fragment	GGT CTC TCA CCA TGG ATC CCA TTC GTT CGC GCA C
avrbs3 NtermATAARev	Subcloning of full *avrbs3* N-terminal encoding fragment	GGT CTC ATT ATG GGA AGA CCG CGT AAG GTT CAG GGG G
sBrg11 CTC Fwd 1	First round of PCR for the subcloning of the CTC region of *brg11* C-terminal encoding fragment	CAG CAA GTC GTT GCA ATC GCT TGC ATC TCC
sBrg11 CTC Fwd 2	Second round of PCR for the subcloning of the CTC region of *brg11* C-terminal encoding fragment	GGT CTC TAT AAG GGA AGA AGA CCC ACC CCA GCA AGT CGT TGC
sBrg11 CTC Rev 1	First round of PCR for the subcloning of the CTC region of *brg11* C-terminal encoding fragment	AAG AAT TTG GAG GAG CCA GTC CAG ATG AAA GTC CAA TGG ACC CAC TCC G
sBrg11 CTC Rev 2	Second round of PCR for the subcloning of the CTC region of *brg11* C-terminal encoding fragment	GGT CTC ACC TTT GTT TCA AGA ATT TGG AGG AGC CAG TC
sBrg11 CTC-ΔAD Rev	Subcloning C-terminal region of *brg11* lacking C-terminal most 35 codons	TCC TGG ATC AGG TAG ACC AGG TTC AAA TGC
sBrg11 CTC- ΔAD Fwd	Subcloning C-terminal region of *brg11* lacking C-terminal most 35 codons	AAG GTG AGA CCC CTG CAT GCA AGC
sBrg11 780 Fwd	Subcloning *brg11* rep0 and -1 for NTC1	CAT GCC GAT ATA TGT CGA ATT TC
sBrg11 1040 Rev	Subcloning *brg11* rep0 and -1 for NTC1	TTT CCT TCG AAG TCT ATA AAG TGC
AvrBs3 LTG Fwd	Removal of rep0 and -1 from *avrbs3* N-terminal encoding region for NTC1 and removal of full N-terminal encoding region for NTC2	CTG ACG GGT GCC CCC CTG

Table S1: Primers used in this study.

Name of Primer	**Application**	**Sequence**
AvrBs3 VGH Rev	Removal of *avrbs3* rep0 and -1 for NTC1	TGT AAA CCC ATG GCC GAC C
AvrBs3 BsaI CACC Rev	Removal of *avrbs3* full N-terminal encoding region for NTC2	CAC GGT GAG AGA CCA AAG G
sBrg11 Nterm ATG Fwd	Subcloning of *brg11* N-terminal encoding fragment for NTC2	GGT CTC TCA CCA TGA GAA TTG GCA AGT CC
sBrg11 NtermBsaIBpiI Rev	Subcloning of *brg11* N-terminal encoding fragment for NTC2	GGT CTC ATT ATG GGA AGA CTT CAG GTG AAG AGG GGC TCT AG
sBrg11 Nterm TACG Fwd	Addition of BpiI site at 3' of subcloned *brg11* N-terminal encoding fragment for compatibility with golden-gate assembly method	CGA AGT CTT CCC ATA ATG AG
sBrg11 Nterm TACG	Addition of BpiI site at 3' of subcloned *brg11* N-terminal encoding fragment for compatibility with golden-gate assembly method	TAA GGT GAA GAG GGG CTC TAG
$Bs3p_{EBE\ Brg11}$ Fwd	Introducing Brg11 target box into the *Bs3* promoter in place of UPT_{AvrBs3}	GTG GCA ATC ACA ACT TCA AGT TAT CAT C
$Bs3p_{EBE\ Brg11\ A0}$	Introducing Brg11 target box with an A_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC GAC CCC CTT ATA AAA TTG GTC AG
$Bs3p_{EBE\ Brg11\ G0}$	Introducing Brg11 target box with a G_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC GAC CCC CTC ATA AAA TTG GTC AG
$Bs3p_{EBE\ Brg11\ C0}$	Introducing Brg11 target box with a C_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC GAC CCC CTG ATA AAA TTG GTC AG
$Bs3p_{EBE\ Brg11\ T0}$	Introducing Brg11 target box with a T_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC GAC CCC CTA ATA AAA TTG GTC AG
brg11_tow_prom_01	Primers used to sequence *RipTAL* homologs	GTG AGA TGC AGC GGT GCT CGC GTG AG
brg11_tow_stop_01	Primers used to sequence *RipTAL* homologs	CGG TGA AGG CGA TCC GGC GGA TAC
Hpx17N-termrev	Primers used to sequence *RipTAL* homologs	CGC TCG CCG TAC TGC GCA ACG
Hpx17C-termfwd	Primers used to sequence *RipTAL* homologs	GCC GCC ACC AGC CTC TCT TGG
Hpx17N- termIIfwd	Primers used to sequence *RipTAL* homologs	GAT CGC CTC TGC CTG CCC TCA AGC
Hpx17C- termIIfwd	Primers used to sequence *RipTAL* homologs	CCA TCG CAC CGC GGT CGA TAC G
Hpx17N- termIIrev	Primers used to sequence *RipTAL* homologs	GCT TGA GGG CAG GCA GAG GCG ATC
Hpx17C- termIIrev	Primers used to sequence *RipTAL* homologs	CGT ATC GAC CGC GGT GCG ATG G
Repeat region Seq Fwd[a)]	Primers used to sequence *RipTAL* homologs	CGC TGC ATC TGA CAC CGC AGC AGG T
Repeat region Seq Rev[a)]	Primers used to sequence *RipTAL* homologs	CCT TCA CCG GCA ACC CCT GCC TGA C
brg11 rep12 Seq	Primers used to sequence *RipTAL* homologs	ATA AGG CGC CGC GCG CAG ATC CGG
brg11_rep07_unique-	Primers used to sequence	CAG GCG CTG GAG GGG ATT GGC

Table S1: Primers used in this study.

Name of Primer	**Application**	**Sequence**
seq	*RipTAL* homologs	GA
brg11 rep17 Seq	Primers used to sequence *RipTAL* homologs	CTT GGC GCA ATG TGG GCA TGT GC
inF RipTALIs Fwd	Primers used to amplify RipTALI-9, -11 and -14	TCC ACC GAC ACG GCC TCG A
inF RipTALIs Rev	Primers used to amplify RipTALI-9, -11 and -14	GCC TCG AAC GAT GCA AGA TCT G
Bs3p $EBE_{RipTALI-6}$ Fwd	Introducing RipTALI-6 target box into the *Bs3* promoter in place of UPT_{AvrBs3}	GCA ATC ACA ACT TCA AGT TAT CAT C
Bs3p $EBE_{RipTALI-9}$ Fwd	Introducing RipTALI-9 target box into the *Bs3* promoter in place of UPT_{AvrBs3}	GTG GGC AAT CAC AAC TTC AAG TTA TCA TC
Bs3p $EBE_{RipTALI-11}$ Fwd	Introducing RipTALI-11 target box into the *Bs3* promoter in place of UPT_{AvrBs3}	GCA ATC ACA ACT TCA AGT TAT CATC
Bs3p $EBE_{RipTALI-14}$ Fwd	Introducing RipTALI-14 target box into the *Bs3* promoter in place of UPT_{AvrBs3}	TTG GCA ATC ACA ACT TCA AGT TAT CAT C
Bs3p $EBE_{RipTALI-6}$ A_0Rev	Introducing RipTALI-6 target box with an A_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC ACC CCC TTA TAA AAT TGG TCA G
Bs3p $EBE_{RipTALI-6}$ G_0Rev	Introducing RipTALI-6 target box with a G_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC ACC CCC TCA TAA AAT TGG TCA G
Bs3p $EBE_{RipTALI-6}$ C_0Rev	Introducing RipTALI-6 target box with a C_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC GAC CCC TGA TAA AAT TGG TCA G
Bs3p $EBE_{RipTALI-6}$ T_0Rev	Introducing RipTALI-6 target box with a T_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC GAC CCC TAA TAA AAT TGG TCA G
Bs3p $EBE_{RipTALI-14}$ A_0Rev	Introducing RipTALI-14 target box with an A_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC GAC CCC TTA TAA AAT TGG TCA G
Bs3p $EBE_{RipTALI-14}$ G_0Rev	Introducing RipTALI-14 target box with a G_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC GAC CCC TCA TAA AAT TGG TCA G
Bs3p $EBE_{RipTALI-14}$ C_0Rev	Introducing RipTALI-14 target box with a C_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC GAC CCC TGA TAA AAT TGG TCA G
Bs3p $EBE_{RipTALI-14}$ T_0Rev	Introducing RipTALI-14 target box with a T_0 into the *Bs3* promoter in place of UPT_{AvrBs3}	GGC GAC CCC TAA TAA AAT TGG TCA G

[a] Described in Heuer et al., 2007.

Table S2: Origin of RipTALs analysed in this study.

RipTAL	**repeat no.**	**Strain (source)**	**P[a)] AluI type[b)]**	**Host plant [a)] (family)**	**Geographic origin [a)]**
RipTALI-1	16	JS53	I d5	*Capsicum annuum (Solanaceae)*	China
RipTALI-2	16	HB53	I d3	*Capsicum annuum (Solanaceae)*	China
RipTALI-3	16	GX53	I d3	*Capsicum annuum (Solanaceae)*	China
RipTALI-4	16	GX55	I d1	*Capsicum annuum (Solanaceae)*	China
RipTALI-5	1	HN515	I a	*Capsicum annuum (Solanaceae)*	China
RipTALI-6	15	GD52	I c1	*Capsicum annuum (Solanaceae)*	China
RipTALI-7	16	FJ41	I d4	*Capsicum annuum (Solanaceae)*	China
RipTALI-8	1	SD58	I a	*Zingiber officinale (Zingiberaceae)*	China
RipTALI-9	17	UW360	I e	*Morus alba (Moraceae)*	China
RipTALI-10	16	UW148	I d4	*Rapistrum rugosum (Brassicaceae)*	Australia
RipTALI-11	13	GX526	I b2	*Arachis hypogaea (Fabaceae)*	China
RipTALI-12	16	GX528	I d1	*Arachis hypogaea (Fabaceae)*	China
RipTALI-13	16	GZ522	I d7	*Nicotiana tabacum (Solanaceae)*	China
RipTALI-14	15	GD45	I c2	*Solanum lycopersicum (Solanaceae)*	China
RipTALI-15	16	ICPM11110	I d5	*Casuarina equisetifolia (Casuarinaceae)*	China

P.: Phylotype; [a)] according to Heuer et al., 2007, Xue et al., 2011 and personal communication for RipTALI-10: A. Milling, Department of Plant Pathology, University of Wisconsin-Madison. [b)] according to Heuer et al., 2007; out of a total thirteen identified AluI types: a, b1-2, c1-2, d1-7 and e.

Table S3: GUS activities as determined in the trimer test and shown in Figure 3

Trimer test repeat	Bs3p 3xA	Bs3p 3xG	Bs3p 3xC	Bs3p 3xT
Brg11 repeat 1	2.33 ± 0.61	1.60 ± 0.51	0.70 ± 0.31	0.00 ± 0.00
Brg11 repeat 2	0.00 ± 0.00	3.06 ± 0.79	0.00 ± 0.00	0.00 ± 0.00
Brg11 repeat 3	0.01 ± 0.00	0.30 ± 0.03	0.00 ± 0.00	0.00 ± 0.00
Brg11 repeat 4	1.66 ± 0.09	2.16 ± 0.30	0.00 ± 0.00	0.00 ± 0.00
Brg11 repeat 5	0.78 ± 0.40	1.19 ± 0.11	0.00 ± 0.00	0.00 ± 0.00
Brg11 repeat 6	4.80 ± 0.51	3.93 ± 1.11	0.08 ± 0.02	0.00 ± 0.00
Brg11 repeat 7	0.34 ± 0.06	0.42 ± 0.36	0.42 ± 0.04	2.39 ± 0.34
Brg11 repeat 8/11	0.00 ± 0.00	0.00 ± 0.00	1.26 ± 0.08	0.00 ± 0.00
Brg11 repeat 9	0.00 ± 0.00	2.67 ± 1.25	0.00 ± 0.00	0.00 ± 0.00
Brg11 repeat 10	0.00 ± 0.00	0.00 ± 0.00	0.06 ± 0.01	0.00 ± 0.00
Brg11 repeat 12	0.06 ± 0.01	1.00 ± 0.12	0.00 ± 0.00	0.00 ± 0.00
Brg11 repeat 13	4.07 ± 0.58	1.04 ± 0.36	1.19 ± 0.25	3.05 ± 0.29
Brg11 repeat 14	3.49 ± 0.24	4.52 ± 0.14	2.07 ± 0.35	0.00 ± 0.00
Brg11 repeat 15	0.84 ± 0.03	1.72 ± 0.24	0.00 ± 0.00	0.00 ± 0.00
Brg11 repeat 16	0.00 ± 0.00	0.00 ± 0.00	0.28 ± 0.15	0.00 ± 0.00
TALE NK repeat	0.00 ± 0.00	0.17 ± 0.06	0.00 ± 0.00	0.00 ± 0.00
TALE HD repeat	0.00 ± 0.00	0.00 ± 0.00	0.61 ± 0.07	0.00 ± 0.00
TALE NN repeat	5.78 ± 0.75	5.98 ± 0.83	0.00 ± 0.00	0.00 ± 0.00
TALE NG repeat	0.51 ± 0.19	0.00 ± 0.00	0.06 ± 0.16	1.89 ± 1.00

GUS activitys are in picomoles 4-MU per minute per µg protein. Standard error of the mean is given after the ± symbol in each case. Results are shown to 3 significant figures. A description of the trimer test approach and details of the protocol for GUS activitiy quantification can be found in Materials and Methods. Briefly, *A. tumefaciens* strains bearing derivatives of AvrBs3 where repeats 5-7 have been replaced with three identical copies of the repeats indicated in the first column, were coinfiltrated into *N. benthamiana* along with strains bearing an *uidA* reporter downstream of one of four *Bs3p* derivatives. In these promoter derivatives the three nucelotides corresponding to repeats 5-7 of AvrBs3, which are normally three adenines, have been replaced by three gunanines, cytosines or thymines to create *Bs3p*$_{3\text{xA}}$, *Bs3p*$_{3\text{xG}}$, *Bs3p*$_{3\text{xC}}$ and *Bs3p*$_{3\text{xT}}$. Results are shown for protein extracts taken from leaf tissue forty-eight hours post infiltration and are averages based on three biological replicates.

de Lange *et al.* – Supplement

Supplementary Figures

S1 – Annotated amino acid sequences of Bat1, Bat2 and Bat3

S2 – Annotated amino acid sequences of AvrBs3 and Brg11

S3 – Amino acid alignments of the Bat2 and Bat3 core repeats.

S4 – Nucleotide sequences of synthesised *Bat1*, *Bat2* and *Bat3* genes.

S5 – Sequences of translational fusions for protein purification, transcriptional activation reporters and nuclease assay.

S6 – Target and reporter sequences used in this study.

S7 – MST results for Bat1 measured against $BE_{Bat1\ T\text{-}0,\ A\text{-}0,\ C\text{-}0\ and\ G\text{-}0}$.

S8 – Amino acid sequence of dTALEs used in this study.

S9 – *in planta* transcriptional activation mediated by acBat1.

S10 – Amino acid sequences of acBat1 derivatives (dBats) tested in Figures 5 and 6.

S11 – Nucleotide and amino acid sequences of $dBat_{SOX2}$-RVD switch and -repeat switch.

S12 – Specificity test with the BE_{pSOX2} targeted dBats.

S13 – Pseudocolour density blots of fluorescence and extended boxplots including outliers for experiments shown in Figures 3, 5-7.

S14 – Sequences of Bat1 repeat trimers used in Figure 8.

S15 – Structural predictions for Bat1 based on the structure of PthXo1 bound to DNA.

Supplementary Tables

S1 – Sequence identities of the Bat proteins sorted by domain.

S2 – Sequences of primers used in this study.

S3 – Hydrogen bonds predicted to be formed between repeats of Bat1 based on the structure shown in Figure S15.

Supplementary Figure 1: Annotated amino acid sequences of Bat1, Bat2 and Bat3

Annotated sequences of the three predicted Bat proteins. Each is formed of a short Non-repetitive N-terminal Domain (NND) followed by an array of cryptic (-1, 0, +1) and core repeats (1, 2, 3…). Consecutive repeats are numbered (left side). The RVDs (residues at repeat positions 12 and 13) are marked as boldface black letters on grey background. Blue lettering is used for the positively charged residues within repeat +1.

>Bat1 (from *Burkholderia rhizoxinica* strain HKI-0454 plasmid pBRH01, GenBank NC_014718.1, RBRH_01844; Uniprot E5AV36)

```
 NND MSTAFVDQDKQMANRLN
-1  LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0  LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1  FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2  FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3  FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4  FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5  FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6  FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7  FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
 8  FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9  FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10  FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11  FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12  FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13  FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14  FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15  FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16  FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17  FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18  FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
19  FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
20  LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1  RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>Bat2 (from *Burkholderia rhizoxinica* strain HKI-0454 plasmid pBRH02, GenBank NC_014723.1, RBRH_01776; Uniprot E5AW45)

```
   NND MPATSMHQEDKQSANGLN
-1 LSPLERIKIEKHYGGGATLAFISNQHDELAQV
 0 LSRADILKIASYDCAAQALQAVLDCGPMLGKRG
 1 FSRADIVRIAGNGGGAQALYSVLDVEPTLGKRG
 2 FSQVDVVKIAG--GGAQALHTVLEIGPTLGERG
 3 FSRGDIVTIAGNNGGAQALQAVLELEPTLRERG
 4 FNQADIVKIAGNGGGAQALQAVLDVEPALGKRG
 5 FSRVDIAKIAG--GGAQALQAVLGLEPTLRKRG
 6 FHPTDIIKIAGNNGGAQALQAVLDLELMLRERG
 7 FSQADIVKMASNIGGAQALQAVLNLEPALCERG
 8 FSQPDIVKMAGNSGGAQALQAVLDLELAFRERG
 9 FSQADIVKMASNIGGAQALQAVLELEPALHERG
10 FSQANIVKMAGNSGGAQALQAVLDLELVFRERG
11 VRQADIVKIVGNNGGAQALQAVFELEPTLRERG
12 FNQATIVKIAANGGGAQALYSVLDVEPTLDKRG
13 FSRVDIVKIAG--GGAQALHTAFELEPTLRKRG
14 FNPTDIVKIAGNKGGAQALQAVLELEPALRERG
15 FNQATIVKMAGNAGGAQALYSVLDVEPALRERG
16 FSQPEIVKIAGNIGGAQALHTVLELEPTLHKRG
17 FNPTDIVKIAGNSGGAQALQAVLELEPAFRERG
18 FGQPDIVKMASNIGGAQALQAVLELEPALRERG
19 FSQPDIVEMAGNIGGAQALQAVLELEPAFRERG
20 FSQSDIVKIAGNIGGAQALQAVLELEPTLRESD
21 FRQADIVNIAGNDGSTQALKAVIEHGPRLRQRG
22 FNRASIVKIAGNSGGAQALQAVLKHGPTLDERG
23 FNLTNIVKIAGNGGGAQALKAVIEHGPTLQQRG
24 FNLTDIVEMAGKGGGAQALKAVLEHGPTLRQRG
25 FNLIDIVEMASNTGGAQALKTVLEHGPTLRQRD
26 LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAL---LLERQ
```

>Bat3 (from *Burkholderia rhizoxinica* strain HKI-0454 plasmid pBRH02 GenBank NC_014723.1, RBRH_01777; Uniprot E5AW45)

```
   NND MPVTSVYQKDKPFGARLN
-1 LSPFECLKIEKHSGGADALEFISNKYDALTQV
 0 LSRADILKIACHDCAAHALQAVLDYEQVFRQRG
 1 FARADIIKITGNGGGAQALKAVVVHGPTLNECG
 2 FSQADIVRIADNIGGAQALKAVLEHGPTLNERD
 3 YSGADIVKIAGNGGGARALKAVVMHGPTLCESG
 4 YSGADIVKIASNGGGAQALEAVAMHGSTLCERG
 5 YCRTDIAKIAGNGGGAQALKAIVMHGPTLCERG
 6 YSRTDIVKIADNNGGAQALKAVFEHGPALTQAG
+1 RSNEDIVNMAARTGAAGQIRKMAAQ---LSGRQ
```

Supplementary Figure 2: Annotated amino acid sequences of AvrBs3 and Brg11

AvrBs3 and Brg11 are the first characterised TALE and RipTAL respectively (36, 10). Annotated amino-acid sequences are given for Brg11 and AvrBs3. N-terminal and C-terminal non-repeat regions and the central repeat array are displayed in separate paragraphs but are part of contiguous polypeptides. Consecutive repeats are numbered (left side). Repeats can be divided into cryptic (-1, 0, +1, +2) and core (1, 2, 3…). The RVDs (residues at repeat positions 12 and 13) are marked as boldface black letters on grey background.

>AvrBs3 (from *Xanthomonas campestris* pv. *vesicatoria* strain 71-21; GenBank CAA34257.1)

MDPIRSRTPSPARELLPGPQPDGVQPTADRGVSPPAGGPLDGLPARRTMSRTRLPSPPAPSP
AFSAGSFSDLLRQFDPSLFNTSLFDSLPPFGAHHTEAATGEWDEVQSGLRAADAPPPTMRVA
VTAARPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGH
GFTHAHIVALSQHPAALGTVAVKYQDMIAALPE
-1 ATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQ
0 LDTGQLLKIAKR-GGVTAVEAVHAWRNALTGAPLN
1 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
2 LTPQQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
3 LTPQQVVAIAS**NS**GGKQALETVQRLLPVLCQAHG
4 LTPEQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
5 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
6 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
7 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
8 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
9 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
10 LTPQQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
11 LTPEQVVAIAS**NS**GGKQALETVQALLPVLCQAHG
12 LTPEQVVAIAS**NS**GGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
15 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
16 LTPQQVVAIAS**NG**GGRPALETVQRLLPVLCQAHG
17 LTPEQVVAIAS**HD**GGKQALETVQRLLPVLCQAHG
+1 LTPQQVVAIASNGGGRPALESIVAQLSRPDPALAA
+2 LTNDHLVALACL-GGRPALDAVKKGLPHAPALIKRT
NRRIPERTSHRVADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGMSRHGLLQLFRRVGVTELE
ARSGTLPPASQRWDRILQASGMKRAKPSPTSTQTPDQASLHAFADSLERDLDAPSPMHEGDQ
TRASSRKRSRSDRAVTGPSAQQSFEVRVPEQRDALHLPLSWRVKRPRTSIGGGLPDPGTPTA
ADLAASSTVMREQDEDPFAGAADDFPAFNEEELAWLMELLPQ

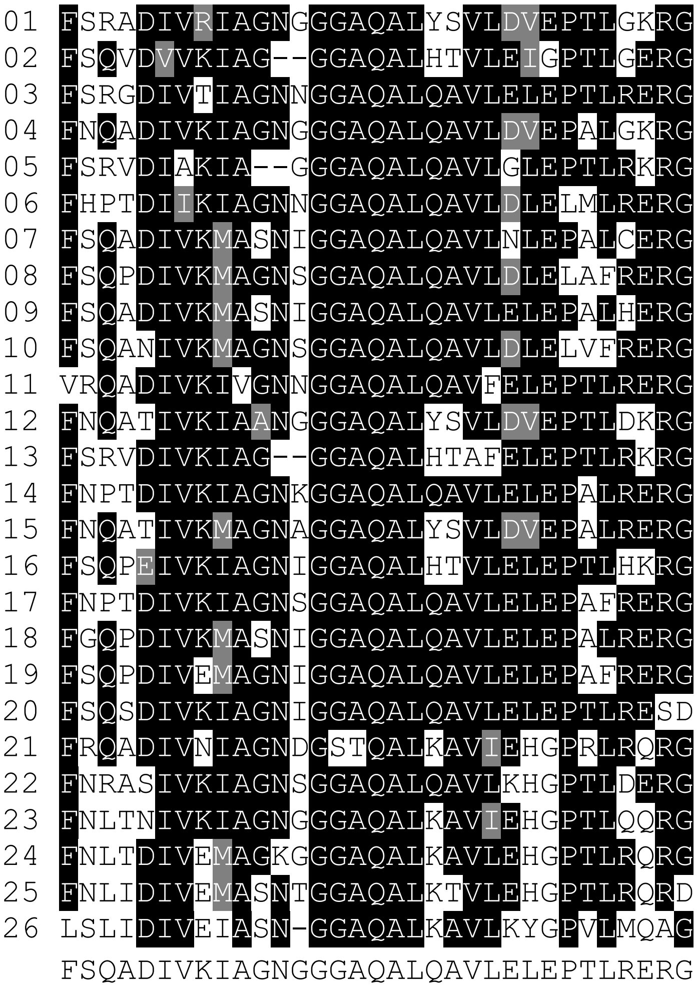
>Brg11 (from *Ralstonia solanacearum* strain GMI1000; GenBank NP_519936.1)

MRIGKSSGWLNESVSLEYEHVSPPTRPRDTRRRPRAAGDGGLAHLHRRLAVGYAEDTPRTEA
RSPAPRRPLPVAPASAPPAPSLVPEPPMPVSLPAVSSPRFSAGSSAAITDPFPSLPPTPVLY
AMARELEALSDATWQPAVPLPAEPPTDARRGNTVFDEASASSPVIASACPQAFASPPRAPRS
ARARRARTGGDAWPAPTFLSRPSSSRIGRDVFGKLVALGYSREQIRKLKQESLSEIAKYHTT
LTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE
-1 LTRAHIVDIARQRSGDLALQALLPVATALTAAPLR
0 LSASQIATVAQY GERPAIQALYRLRRKLTRAPLH
1 LTPQQVVAIAS**NT**GGKRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIAS**NK**GGKQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIAS**HN**GGKQALEAVKADLLDLRGAPYA
4 LSTEQVVAIAS**HN**GGKQALEAVKADLLELRGAPYA
5 LSTEQVVAIAS**HN**GGKQALEAVKAHLLDLRGVPYA
6 LSTEQVVAIAS**HN**GGKQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIAS**NG**GGKQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
9 LSTEQVVAIAS**NK**GGKQALEAVKAQLLELRGAPYA
10 LSTAQVVAIAS**HD**GGNQALEAVGTQLVALRAAPYA
11 LSTEQVVAIAS**HD**GGKQALEAVGAQLVALRAAPYA
12 LNTEQVVAIAS**SH**GGKQALEAVRALFPDLRAAPYA
13 LSTAQLVAIAS**NP**GGKQALEAVRALFRELRAAPYA
14 LSTEQVVAIAS**NH**GGKQALEAVRALFRGLRAAPYG
15 LSTAQVVAIAS**SN**GGKQALEAVWALLPVLRATPYD
16 LNTAQIVAIAS**HD**GGKPALEAVWAKLPVLRGAPYA
+1 LSTAQVVAIACI-SGQQALEAIEAHMPTLRQASHS
+2 LSPERVAAIACI-GGRSAVEAVRQGLPVKAIRRIRR
EKAPVAGPPPASLGPTPQELVAVLHFFRAHQQPRQAFVDALAAFQATRPALLRLLSSVGVTE
IEALGGTIPDATERWQRLLGRLGFRPATGAAAPSPDSLQGFAQSLERTLGSPGMAGQSACSP
HRKRPAETAIAPRSIRRSPNNAGQPSEPWPDQLAWLQRRKRTARSHIRADSAASVPANLHLG
TRAQFTPDRLRAEPGPIMQAHTSPASVSFGSHVAFEPGLPDPGTPTSADLASFEAEPFGVGP
LDFHLDWLLQILET

Supplementary Figure 3: Amino acid alignments of Bat2 and Bat3 core repeats.

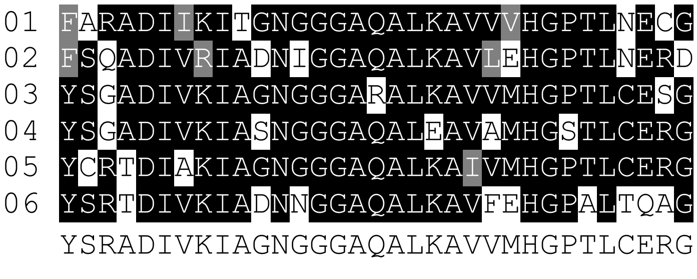
Alignments of the core repeats of Bat2 and Bat3 were created in Clustal Omega (34, 35) and Boxshade was used for formatting. White lettering on a black background indicates a consensus residue. Black lettering on a grey background indicates a residue similar to the consensus residue. Black lettering on a white background indicates a residue neither identical nor similar to the consensus residue. Repeats are shown in order of appearance in the polypeptide and numbered accordingly. The consensus repeat is shown below each alignment.

>Alignment of Bat2 core repeats



```
01 FSRADIVRIAGNGGGAQALYSVLDVEPTLGKRG
02 FSQVDVVKIAG--GGAQALHTVLEIGPTLGERG
03 FSRGDIVTIAGNNGGAQALQAVLELEPTLRERG
04 FNQADIVKIAGNGGGAQALQAVLDVEPALGKRG
05 FSRVDIAKIA--GGGAQALQAVLGLEPTLRKRG
06 FHPTDIIKIAGNNGGAQALQAVLDLELMLRERG
07 FSQADIVKMASNIGGAQALQAVLNLEPALCERG
08 FSQPDIVKMAGNSGGAQALQAVLDLELAFRERG
09 FSQADIVKMASNIGGAQALQAVLELEPALHERG
10 FSQANIVKMAGNSGGAQALQAVLDLELVFRERG
11 VRQADIVKIVGNNGGAQALQAVFELEPTLRERG
12 FNQATIVKIAANGGGAQALYSVLDVEPTLDKRG
13 FSRVDIVKIAG--GGAQALHTAFELEPTLRKRG
14 FNPTDIVKIAGNKGGAQALQAVLELEPALRERG
15 FNQATIVKMAGNAGGAQALYSVLDVEPALRERG
16 FSQPEIVKIAGNIGGAQALHTVLELEPTLHKRG
17 FNPTDIVKIAGNSGGAQALQAVLELEPAFRERG
18 FGQPDIVKMASNIGGAQALQAVLELEPALRERG
19 FSQPDIVEMAGNIGGAQALQAVLELEPAFRERG
20 FSQSDIVKIAGNIGGAQALQAVLELEPTLRESD
21 FRQADIVNIAGNDGSTQALKAVIEHGPRLRQRG
22 FNRASIVKIAGNSGGAQALQAVLKHGPTLDERG
23 FNLTNIVKIAGNGGGAQALKAVIEHGPTLQQRG
24 FNLTDIVEMAGKGGGAQALKAVLEHGPTLRQRG
25 FNLIDIVEMASNTGGAQALKTVLEHGPTLRQRD
26 LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
   FSQADIVKIAGNGGGAQALQAVLELEPTLRERG
```

> Alignment of Bat3 core repeats



```
01 FARADIIKITGNGGGAQALKAVVVHGPTLNECG
02 FSQADIVRIADNIGGAQALKAVLEHGPTLNERD
03 YSGADIVKIAGNGGGARALKAVVMHGPTLCESG
04 YSGADIVKIASNGGGAQALEAVAMHGSTLCERG
05 YCRTDIAKIAGNGGGAQALKAIVMHGPTLCERG
06 YSRTDIVKIADNNGGAQALKAVFEHGPALTQAG
   YSRADIVKIAGNGGGAQALKAVVMHGPTLCERG
```

Supplementary Figure 4: Nucleotide sequences of synthesised *Bat1, Bat2* and *Bat3* genes

Genes encoding the three predicted proteins were synthesised with *E. coli* codon usage (GenScript). Each was synthesised as a series of separate blocks flanked by BsaI sites allowing ordered assembly via BsaI cut-ligation into target vectors. BsaI recognition sites are underlined, while bold typeface marks the overlaps created upon digest. Start and stop codons are distinguished with the use of lowercase italics.

>Bat1 block 1

GGTCTCT**CACC***atg*AGCACCGCCTTCGTGGACCAAGATAAGCAAATGGCAAATCGCC
TGAACCTGTCACCGCTGGAACGTAGCAAAATTGAAAAACAATATGGCGGTGCAACCA
CGCTGGCTTTTATTAGCAACAAACAGAATGAACTGGCACAAATCCTGAGCCGTGCTG
ATATTCTGAAAATCGCGTCTTACGACTGCGCAGCACATGCACTGCAGGCTGTCCTGG
ATTGTGGCCCGATGCTGGGCAAACGCGGTTTTAGCCAGTCTGACATTGTCAAGATCG
CCGGTAACATTGGCGGTGCACAGGCACTGCAAGCAGTGCTGGATCTGGAAAGTATGC
TGGGCAAACGTGGTTTCTCCCGCGATGACATTGCGAAGATGGCCGGCAATATCGGCG
GTGCACAGACCCTGCAGGCCGTGCTGGATCTGGAATCAGCCTTTCGTGAACGCGGCT
TCTCGCAGGCCGACATTGTTAAAATCGCCGGTAACAATGGCGGTGCACAAGCTCTGT
ATAGTGTGCTGGATGTTGAACCGACCCTGGGTAAACGTGGTTTTTCACGCGCTGACA
TTGTTAAGATCGCCGGTAACACCGGCGGTGCCCAAGCACTGCACACGGTCCTGGATC
TGGAACCGGCCCTGGGCAAGCGTGGTTTCTCCCGCATTGATATCGTTAAGATCGCAG
CTAACAACGGTGGTGCTCAAGCCCTGCACGCTGTCCTGGATCTGGGTCCGACGCTGC
GCGAATG**TGGG**TGAGACC

>Bat1 block 2

GGTCTCT**TGGG**TTCTCGCAGGCAACCATCGCAAAAATCGCTGGCAATATCGGCGGTG
CTCAGGCTCTGCAAATGGTGCTGGATCTGGGTCCGGCTCTGGGCAAACGTGGTTTTA
GCCAGGCAACCATTGCTAAGATCGCCGGTAACATTGGCGGTGCACAGGCACTGCAAA
CGGTCCTGGATCTGGAACCGGCGCTGTGCGAACGCGGCTTCTCTCAGGCCACCATCG
CAAAAATGGCTGGTAACAATGGCGGTGCACAGGCTCTGCAAACGGTTCTGGATCTGG
AACCGGCCCTGCGTAAACGCGATTTTCGTCAGGCGGACATTATCAAGATTGCCGGTA
ATGACGGTGGCGCCCAGGCACTGCAAGCAGTGATCGAACATGGCCCGACCCTGCGCC
AACACGGTTTCAACCTGGCAGACATTGTTAAGATGGCTGGTAATATCGGTGGTGCTC
AAGCTCTGCAAGCGGTGCTGGACCTGAAGCCGGTGCTGGACGAACAT**GGTT**TGAGAC
C

>Bat1 block 3

GGTCTCT**GGTT**TCTCTCAACCGGATATCGTCAAGATGGCGGGCAACATTGGTGGTGC
TCAAGCCCTGCAAGCCGTCCTGTCACTGGGTCCGGCGCTGCGTGAACGTGGCTTTAG
CCAGCCGGATATTGTCAAAATCGCCGGTAACACCGGCGGTGCACAGGCACTGCAAGC
AGTGCTGGATCTGGAACTGACGCTGGTTGAACATGGCTTCTCTCAACCGGACATTGT
TCGCATCACCGGTAATCGTGGCGGTGCCCAAGCTCTGCAAGCGGTGCTGGCTCTGGA
ACTGACCCTGCGTGAACG**AGGA**TGAGACC

>Bat1 block 4

GGTCTCT**AGGA**TTTAGCCAACCGGACATCGTGAAAATCGCGGGCAATAGCGGCGGTG
CTCAAGCTCTGCAAGCGGTCCTGGATCTGGAACTGACGTTTCGTGAACGCGGCTTTA
GCCAGGCGGATATTGTCAAAATCGCCGGTAACGACGGCGGTACCCAAGCACTGCATG
CTGTGCTGGATCTGGAACGTATGCTGGGCGAACGTGGTTTCTCTCGCGCAGACATTG
TGAACGTTGCTGGCAACAATGGCGGTGCGCAGGCCCTGAAAGCCGTGCTGGAACACG
AAGCCACGCTGAATGAACGTGGCTTTAGTCGCGCAGATATTGTCAAGATCGCGGGTA
ACGGTGGCGGCGCACAAGCACTGAAGGCGGTTCTGGAACACGAAGCGACCCTGGATG
AACG**CGGC**TGAGACC

>Bat1 block 5

GGTCTCT**CGGC**TTTTCTCGTGCTGATATTGTCCGTATTGCGGGTAATGGTGGTGGTG
CCCAGGCTCTGAAGGCTGTGCTGGAACATGGTCCGACGCTGAACGAACGTGGCTTTA
ATCTGACCGATATTGTTGAAATGGCGGCCAACAGTGGCGGTGCACAGGCTCTGAAAG
CGGTCCTGGAACACGGCCCGACGCTGCGTCAACGTGGTCTGAGCCTGATTGACATCG
TGGAAATTGCATCTAACGGCGGTGCGCAGGCCCTGAAAGCTGTCCTGAAGTATGGTC
CGGTGCTGATGCAAGCAGGTCGTAGCAATGAAGAAATCGTGCACGTTGCCGCTCGTC
GTGGTGGTGCTGGCCGTATCCGTAAGATGGTTGCTCCGCTGCTGGAACGTCAG*tag***A**
AGGTGAGACC

>Bat2 block1

GGTCTCT**CACC***atg*CCGGCCACCTCGATGCACCAAGAAGATAAACAGTCCGCAAACG
GTCTGAACCTGAGCCCGCTGGAACGTATTAAAATTGAAAAACATTATGGCGGTGGCG
CGACCCTGGCCTTTATTAGTAACCAGCACGATGAACTGGCACAAGTGCTGAGCCGTG
CTGACATTCTGAAAATCGCCTCTTATGACTGTGCTGCTCAAGCTCTGCAAGCGGTGC
TGGACTGCGGCCCGATGCTGGGTAAACG**CGGC**TGAGACC

>Bat2 block2

GGTCTCT**CGGC**TTTTCCCGTGCTGATATTGTCCGTATTGCTGGTAATGGTGGTGGTG
CCCAAGCTCTGTATTCTGTCCTGGATGTTGAACCGACGCTGGGTAAACGTGGCTTTA
GCCAGGTTGATGTGGTTAAAATTGCGGGCGGTGGCGCACAAGCACTGCATACCGTCC
TGGAAATCGGTCCGACGCTGGGTGAACGTGGCTTCTCTCGCGGTGACATTGTTACCA
TCGCCGGCAACAATGGTGGCGCACAGGCTCTGCAAGCAGTTCTGGAACTGGAACCGA
CGCTGCGTGAACGCGGTTTTAACCAGGCGGATATTGTCAAAATCGCCGGTAATGGTG
GCGGTGCACAGGCACTGCAAGCAGTCCTGGATGTGGAACCGGCTCTGGGTAAACGTG
GCTTTTCCCGCGTGGACATTGCAAAAATCGCTGGCGGTGGCGCCCAAGCCCTGCAGG
CAGTTCTGGGTCTGGAACCGACCCTGCGTAAACGCGGCTTCCACCCGACGGACATTA
TCAAAATTGCGGGTAACAATGGTGGTGCCCAAGCACTGCAAGCAGTTCTGGATCTGG
AACTGATGCTGCGTGAACGCGGCTTTAGCCAGGCAGACATTGTGAAAATGGCTTCTA
ACATCGGTGGCGCCCAAGCTCTGCAAGCGGTTCTGAATCTGGAACCGGCCCTGTGCG
AACGCGGTTTCTCACAGCCGGATATCGTCAAAATGGCCGGTAACTCGGGTGGCGCCC
AAGCGCTGCAAGCAGTGCTGGATCTGGAACTGGCTTTTCGTGAACGCGGCTTCAGTC
AGGCGGACATTGTGAAAATGGCCTCCAATATCGGCGGCGCACAAGCACTGCAAGCTG
TCCTGGAACTGGAACCGGCTCTGCACGAACGCGGCTT**TAGT**TGAGACC

>Bat2 block3

GGTCTCA**TAGT**CAAGCAAATATCGTCAAAATGGCGGGTAATAGTGGTGGTGCCCAAG
CCCTGCAAGCGGTCCTGGATCTGGAACTGGTCTTTCGTGAACGTGGCGTGCGCCAGG
CGGATATTGTGAAAATCGTTGGTAACAATGGCGGTGCACAGGCTCTGCAAGCAGTCT
TTGAACTGGAACCGACCCTGCGTGAACGCGGCTTCAACCAGGCTACGATTGTTAAAA
TCGCAGCAAATGGCGGTGGCGCACAAGCACTGTATAGCGTCCTGGATGTGGAACCGA
CCCTGGACAAACGTGGTTTCTCTCGCGTTGATATTGTCAAAATCGCAGGTGGCGGTG
CCCAAGCTCTGCATACCGCTTTTGAACTGGAACCGACGCTGCGTAAACGCGGCTTCA
ACCCGACCGACATTGTCAAAATCGCCGGTAATAAAGGCGGTGCACAGGCACTGCAAG
CAGTGCTGGAACTGGAACCGGCTCTGCGTGAACGCGGCTTTAACCAGGCAACGATTG
TGAAAATGGCGGGTAATGCCGGCGGTGCACAAGCTCTGTACAGTGTGCTGGATGTTG
AACCGGCACTGCGTGAACGTGGTTTCTCCCAGCCGGAAATTGTTAAAATCGCCGGTA
ACATCGGCGGTGCGCAAGCCCTGCATACGGTTCTGGAGTTAGAACCGACCCTGCACA
AACGTGGCTTTAACCCGACCGATATTGTGAAAATCGCGGGTAATAGCGGCGGTGCCC
AGGCCCTGCAGGCGGTTCTGGAACTGGAACCGGCGTTTCGTGAACGCGGCTTCGGTC
AGCCGGACATTGTTAAAATGGCCAGCAATATCGGCGGTGCCCAAGCCCTGCAAGCCG
TCCTGGAACTGGAACCGGCCCTGCGTGAACGTGGTTTTAG**CCAG**TGAGACC

>Bat2 block4

GGTCTCT**CCAG**CCGGATATTGTGGAAATGGCGGGTAACATCGGCGGCGCTCAAGCCC
TGCAAGCTGTCCTGGAACTGGAACCGGCCTTTCGTGAACGCGGCTTTAGCCAGTCTG
ATATTGTTAAAATCGCGGGTAACATTGGCGGTGCACAGGCACTGCAAGCAGTTCTGG
AACTGGAACCGACCCTGCGCGAAAGCGATTTCCGTCAGGCAGACATTGTGAACATCG
CTGGCAATGACGGTTCTACCCAAGCGCTGAAAGCCGTTATTGAACATGGCCCGCGTC
TGCGCCAGCGTGGTTTTAACCGCGCGAGTATTGTCAAAATCGCCGGCAATTCCGGCG
GTGCACAGGCTCTGCAAGCAGTGCTGAAACACGGCCCGACCCTGGATGAACGTGGTT
TCAACCTGACGAATATTGTTAAAATCGCCGGTAACGGCGGTGGCGCACAGGCACTGA
AAGCTGTCATTGAACATGGCCCGACCCTGCAGCAACGCGGTTTTAATCTGACGGATA
TCGTGGAAATGGCGGGCAAAGGTGGCGGTGCACAAGCTCTGAAAGCAGTTCTGGAAC
ACGGTCCGACCCTGCGTCAGCGTGGTTTCAACCTGATTGACATCGTCGAAATGGCGT
CCAATACGGGCGGTGCGCAAGCCCTGAAAACCGTTCTGGAACATGGTCCGACGCTGC
GCCAGCGTGATCTGTCACTGATTGACATCGTGGAAATTGCATCGAATGGTGGTGCAC
AGGCTCTGAAAGCTGTCCTGAAATATGGCCCGGTGCTGATGCAGGCAGGTCGTAGCA
ATGAAGAAATCGTGCACGTTGCCGCTCGTCGTGGTGGTGCGGGCCGTATTCGTAAAA
TGGTTGCTCTGCTGCTGGAACGCCAA*t****aa*****GG**TGAGACC

>Bat3 block 1

GGTCTCT**CACC**ATGCCGGTCACCAGCGTCTACCAAAAAGATAAACCGTTCGGCGCAC
GTCTGAACCTGAGCCCGTTTGAATGTCTGAAAATTGAAAAACATAGCGGCGGTGCGG
ATGCCCTGGAATTTATTTCTAACAAATATGACGCCCTGACCCAGGTGCTGAGTCGTG
CAGATATTCTGAAAATCGCTTGCCACGACTGTGCCGCCCACGCTCTGCAAGCTGTGC
TGGACTATGAACAAGTTTTTCGCCAACG**CGGC**TGAGACC

>Bat3 block 2

GGTCTCT**CGGC**TTCGCTCGTGCAGATATTATTAAAATCACGGGTAACGGCGGTGGTG
CCCAAGCCCTGAAAGCAGTGGTTGTCCATGGTCCGACGCTGAACGAATGCGGTTTTT
CACAGGCGGATATTGTCCGTATCGCCGACAATATTGGCGGTGCGCAAGCCCTGAAAG

CGGTGCTGGAACATGGCCCGACCCTGAACGAACGTGATTATTCGGGTGCAGACATTG
TGAAAATCGCTGGTAATGGCGGTGGCGCACGTGCTCTGAAAGCAGTGGTTATGCACG
GTCCGACGCTGTGTGAAAGCGGTTACTCTGGCGCGGATATTGTTAAAATCGCAAGTA
ACGGTGGCGGTGCACAGGCACTGGAAGCAGTCGCTATGCATGGTTCCACCCTGTGCG
AACGTGGCTATTGTCGCACGGACATTGCGAAAATCGCCGGCAACGGCGGTGGCGCAC
AAGCACTGAAAGCAATTGTCATGCACGGTCCGACCCTGTGTGAACGCGGCTACAGCC
GCACGGATATTGTGAAAATCGCAGACAACAATGGTGGCGCACAGGCTCTGAAAGCTG
TTTTCGAACATGGTCCGGCACTGACCCAAGCTGGCCGCAGTAACGAAGATATCGTTA
ATATGGCCGCACGCACGGGCGCAGCGGGTCAGATTCGTAAAATGGCGGCACAACTGT
CGGGTCGTCAA*t****aa*****GG**TGAGACC

Supplementary Figure 5: Sequences of translational fusions for protein purification, transcriptional activation reporters and nuclease assay.

Only the sequences specific to each expression construct are shown. The sequence of the relevant Bat protein or derivative or TALE derivative fills the position indicated. Epitopes used for purification or antibody binding are indicated with a red background. NLSs are indicated with a yellow background. Green background marks an activation domain and mustard-brown a nuclease domain.

>Protein expression and purification

MSYYHHHHHHLESTSLYKKAGSAAAPFT – Bat1, Bat2 or Bat3 coding sequence – STOP

>Human cell transcriptional activation assay: Full construct

MGYPYDVPDYASRPKKKRKVGIHAM – Bat1, dBat or dTALE coding sequence – GGGGGGSGGGGSGGGGSDYKDHDGDYKDHDIDYKDDDDKGSSPKKKRKVEASGSGRADALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLINSR – STOP

>Human cell transcriptional activation assay: ΔAD

MGYPYDVPDYASRPKKKRKVGIHAM – Bat1 coding sequence – GGGGGGSGGGGSGGGGSDYKDHDGDYKDHDIDYKDDDDKGSSPKKKRKVEAS – STOP

>Human cell transcriptional activation assay: ΔNLSs

START – Bat1 coding sequence – GGGGGGSGGGGSGGGGSDYKDHDGDYKDHDIDYKDDDDKGSGRADALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLINSR – STOP

>*Planta* transcriptional activation assay: Full construct

START – Bat1 or dTALE coding sequence – GGGGGGSGGGGSGGGGSDYKDHDGDYKDHDIDYKDDDDKGSSPKKKRKVEASGSGRADALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLINSR – STOP

>*Planta* transcriptional activation assay: ΔAD

START – Bat1 coding sequence – GGGGGGSGGGGSGGGGSDYKDHDGDYKDHDIDYKDDDDKGSSPKKKRKVEAS – STOP

>*In vitro* nuclease assay

MGLINIFYPYDVPDYAGYPYDVPDYAGSYPYDVPDYAAQCSG – Bat1 coding sequence – GGQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

MGYPYDVPDYASRPKKKRKVGIHAS – TALE coding sequence – GSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

Supplementary Figure 6: Target and reporter sequences used in this study.

Sequences of binding elements used for electrophoretic mobility shift assays (Figure 2). Only forward strand shown, the binding element is highlighted with bold lettering

$\mathrm{BE}_{\mathrm{Bat1}}$/$\mathrm{BE}_{\mathrm{Bat1\ T\text{-}0}}$
TAGACT**AAGAGAAGCAAAGACGTTAT**ATGC

$\mathrm{BE}_{\mathrm{Bat2}}$
TAGACT**TTGTTGAAAAGTTGTAAAAACATTAT**ATGC

$\mathrm{BE}_{\mathrm{Bat3}}$
TAGACATAGATTAT**TATATTTG**TAACAAGTAAATGC

$\mathrm{BE}_{\mathrm{Bat1\ C\text{-}0}}$
TAGACC**AAGAGAAGCAAAGACGTTAT**ATGC

$\mathrm{BE}_{\mathrm{Bat1\ G\text{-}0}}$
TAGACG**AAGAGAAGCAAAGACGTTAT**ATGC

$\mathrm{BE}_{\mathrm{Bat1\ A\text{-}0}}$
TAGACA**AAGAGAAGCAAAGACGTTAT**ATGC

Sequences of reporters and binding elements used in assessments of transcriptional activation (Figures 3, 5-7 and S8)

pCMV-*BE-dsEGFP* – transcriptional activation reporter in human cells. (Figures 3, 5-7, S8). Green highlighting is used for the dsEGFP coding sequence and italics for the subsequence polyA signal. Grey highlighting for the minimal CMV promoter. The bold-N positions are filled by one of the four binding elements listed.

$\mathrm{BE}_{\mathrm{Bat1}}$	AAGAGAAGCAAAGACGTTAT
$\mathrm{BE}_{\mathrm{dBatRVDswitch1}}$	A**GAGA**AAGCAAAGACGTTAT
$\mathrm{BE}_{\mathrm{dBatRVDswitch2}}$	AAGAGA**GCA**AAAGACGTTAT
$\mathrm{BE}_{\mathrm{pSOX2}}$	TTTATTCCCTGACAGCCCC

CTAGACT**NNNNNNNNNNNNNNNNNNNNN**ATGCGGATCCACGTATGTCGAGGTAGGCG
TGTACGGTGGGAGGCCTATATAAGCAGAGCTCGTTTAGTGAACCGTCAGATCGCCTG
GAGGTACCGCCACCATGGGCTTAATTAATATAATTAATAATCCACTTAAGAATTCTT
TAAAGTGGATTATTAATTATAGGACCGGTATTACCCTGTTATCCCTAGTGAGCAAGG
GCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAA
ACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGC
TGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCG
TGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGC
AGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCT
TCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACA
CCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCC
TGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACA
AGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCA
GCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGC
TGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACG
AGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCG
GCATGGACGAGCTGTACAAGAAGCTTAGCCATGGCTTCCCGCCGGAGGTGGAGGAGC
AGGATGATGGCACGCTGCCCATGTCTTGTGCCCAGGAGAGCGGGATGGACCGTCACC
CTGCAGCCTGTGCTTCTGCTAGGATCAATGTGTAGCTAAGTAAGATCC*TTCGAGCAG*
ACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAA
AATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCT
GCAATAAACAAGTTAACAACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGG
AGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTAAAA

Bs3p-BE$_{\text{Bat1}}$-*uidA* for *in planta* assessment of transcriptional activation (Figure S8). Blue indicates the coding sequence of the *uidA* reporter gene, which is a part of the vector pGWB3* (10). BE$_{\text{Bat1}}$ is embedded within the pepper *Bs3* promoter (italics) and is distinguished with bold typeface. In this construct a guanine base is paired with the 20th repeat of acBat1 and dTALE$_{\text{Bat1mimic}}$.

TCATAGTCAAGCTAACGAAACTTATGCAAGGGAAATATGAAATTAGTATGCAAGTAA
ACTCAAAGAACTAATCATTGAACTGAAAGATCAATATATCAAAAAAAAAAAAAAAAC
AATAAAACCGTTTAACCGATAGATTAACCATTTCTGGTTCAGTTTATGGGTTAAACC
ACAATTTGCACACCCTGGTTAAACAATGAACACGTTTGCCTGACCAATTTTATTATA
*TAAACCTAACCATCCTCACAACT**AAGAGAAGCAAAGACGTTAG**GTTCAAGTTATCAT*
CCCCTTTCTCTTTTCTCCTCTTGTTCTTGTCACCCGCTAAATCTATCAAAACACAAG
*TAGTCCTAGTTGCACTATATTTC*AAGGGTGGGCGCGCCGACCCAGCTTTCTTGTACA
AAGTGGTTCGATCTAGAGGATCCCCGGGTGGTCAGTCCCTTATGTTACGTCCTGTAG
AAACCCCAACCCGTGAAATCAAAAAACTCGACGGCCTGTGGGCATTCAGTCTGGATC
GCGAAAACTGTGGAATTGATCAGCGTTGGTGGGAAAGCGCGTTACAAGAAAGCCGGG
CAATTGCTGTGCCAGGCAGTTTTAACGATCAGTTCGCCGATGCAGATATTCGTAATT
ATGCGGGCAACGTCTGGTATCAGCGCGAAGTCTTTATACCGAAAGGTTGGGCAGGCC
AGCGTATCGTGCTGCGTTTCGATGCGGTCACTCATTACGGCAAAGTGTGGGTCAATA
ATCAGGAAGTGATGGAGCATCAGGGCGGCTATACGCCATTTGAAGCCGATGTCACGC
CGTATGTTATTGCCGGGAAAAGTGTACGTATCACCGTTTGTGTGAACAACGAACTGA
ACTGGCAGACTATCCCGCCGGGAATGGTGATTACCGACGAAAACGGCAAGAAAAAGC
AGTCTTACTTCCATGATTTCTTTAACTATGCCGGAATCCATCGCAGCGTAATGCTCT
ACACCACGCCGAACACCTGGGTGGACGATATCACCGTGGTGACGCATGTCGCGCAAG

```
ACTGTAACCACGCGTCTGTTGACTGGCAGGTGGTGGCCAATGGTGATGTCAGCGTTG
AACTGCGTGATGCGGATCAACAGGTGGTTGCAACTGGACAAGGCACTAGCGGGACTT
TGCAAGTGGTGAATCCGCACCTCTGGCAACCGGGTGAAGGTTATCTCTATGAACTGT
GCGTCACAGCCAAAAGCCAGACAGAGTGTGATATCTACCCGCTTCGCGTCGGCATCC
GGTCAGTGGCAGTGAAGGGCGAACAGTTCCTGATTAACCACAAACCGTTCTACTTTA
CTGGCTTTGGTCGTCATGAAGATGCGGACTTGCGTGGCAAAGGATTCGATAACGTGC
TGATGGTGCACGACCACGCATTAATGGACTGGATTGGGGCCAACTCCTACCGTACCT
CGCATTACCCTTACGCTGAAGAGATGCTCGACTGGGCAGATGAACATGGCATCGTGG
TGATTGATGAAACTGCTGCTGTCGGCTTTAACCTCTCTTTAGGCATTGGTTTCGAAG
CGGGCAACAAGCCGAAAGAACTGTACAGCGAAGAGGCAGTCAACGGGGAAACTCAGC
AAGCGCACTTACAGGCGATTAAAGAGCTGATAGCGCGTGACAAAAACCACCCAAGCG
TGGTGATGTGGAGTATTGCCAACGAACCGGATACCCGTCCGCAAGGTGCACGGGAAT
ATTTCGCGCCACTGGCGGAAGCAACGCGTAAACTCGACCCGACGCGTCCGATCACCT
GCGTCAATGTAATGTTCTGCGACGCTCACACCGATACCATCAGCGATCTCTTTGATG
TGCTGTGCCTGAACCGTTATTACGGATGGTATGTCCAAAGCGGCGATTTGGAAACGG
CAGAGAAGGTACTGGAAAAAGAACTTCTGGCCTGGCAGGAGAAACTGCATCAGCCGA
TTATCATCACCGAATACGGCGTGGATACGTTAGCCGGGCTGCACTCAATGTACACCG
ACATGTGGAGTGAAGAGTATCAGTGTGCATGGCTGGATATGTATCACCGCGTCTTTG
ATCGCGTCAGCGCCGTCGTCGGTGAACAGGTATGGAATTTCGCCGATTTTGCGACCT
CGCAAGGCATATTGCGCGTTGGCGGTAACAAGAAAGGGATCTTCACTCGCGACCGCA
AACCGAAGTCGGCGGCTTTTCTGCTGCAAAAACGCTGGACTGGCATGAACTTCGGTG
AAAAACCGCAGCAGGGAGGCAAACAATGA
```

Sequences of the PCR templates used to create targets for the nuclease assays shown in Figure 4. Only the forward strand is shown. Grey highlighting shows the annealing sites for the amplification primers used in the PCR to create the target DNA for the nuclease assays. The two copies of $\mathrm{BE}_{\mathrm{Bat1}}$ in reverse orientation are underlined. The italicised bases are one of the five spacers listed below. The entire yellow-highlighted region is replaced by the given sequence in the case of the 'no target' control.

5bp CTAGC

7bp TCTAGAC

11bp TACGTCTAGAC

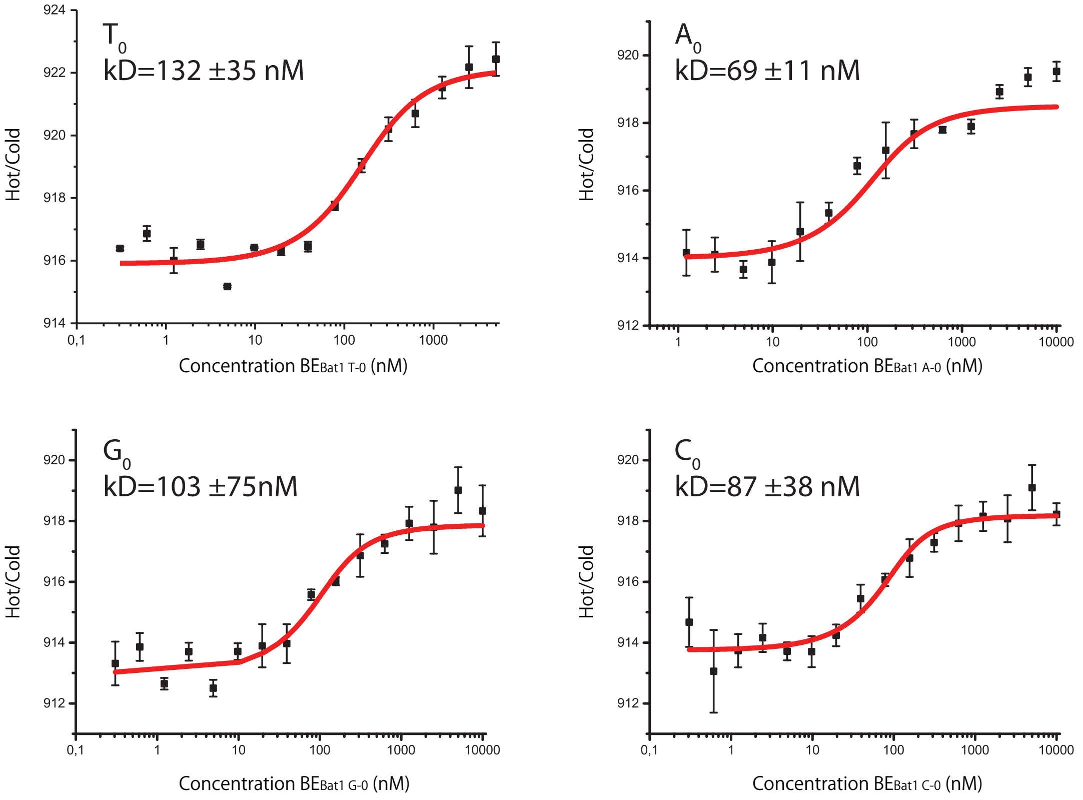
15bp TACGTACGTCTAGAC

19bp AAGCTACGTACGTCTAGAC

No target ATTGCCACGGCGACTCTCTTG

GCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGC
CCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCTGGCTTAACTATGCGGC
ATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATG
CGTAAGGAGAAAATACCGCATCAGGCGCCATTCGCCATTCAGGCTGCGCAACTGTTG
GGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATG
TGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAA
AACGACGGCCAGTGAATTCGAGCTCGGTACCTCGCGAATGCATCTAGATATCGGATC
CCGGGCCCGTCGACTGCAGAGGGGTCTCCCCTTGAAATATAGTGCAACTAGGACTAC
TTGTGTTTTGATAGATTTAGCGGGTGACAAGAACAAGAGGAGAAAAGAGAAAGGGGA
TGATAACTTGAATAAGAGAAGCAAAGACGTTAT*NNNNNNNNNNNNN*ATAACGTCTTT
GCTTCTCTTAGTTGTGAGGATGGTTAGGTTTATATAATAAAATTGGTCAGGCAAACG
TGTTCATTGTTTAACCAGGGTGTGCAAATTGTGGTTTAACCCATAAACTGAACCAGA
AATGGTTAATCTATCGGTTAAACGGTTTTATTGTTTTTTTTTTTTTTTGATATATTG
ATCTTTCAGTTCAATGATTAGTTCTTTGAGTTTACTTGCATACTAATTTCATATTTC
CCTTGCATAAGTTTCGTTAGCTTGACTATGAGGTGGGAGACCCCTGCATGCAAGCTT
GGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCC
ACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAG
CTAAC

Supplementary Figure 7 – MST results for Bat1 measured against $BE_{Bat1A\text{-}0}$, $\text{-}_{C\text{-}0}$, $\text{-}_{G\text{-}0}$, and $\text{-}_{T\text{-}0}$



Supplementary Figure 8: Amino acid sequence of dTALEs used in this study
Core and cryptic repeats are numbered. Grey background and bold typeface highlight the RVD residues. In all cases only the TALE-derived amino acids are shown. The sequences of fused domains are given in Figure S5.

>dTALE$_{\text{Bat1mimic}}$ (for transcriptional activation assays)

MDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGH
GFTHAHIVALSQHPAALGTVAVKYQDMIAALPE
-1 ATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQ
0 LDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLN
1 LTPQQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
2 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
3 LTPEQXVAIAS**NN**GGKQALXTVQRLLPVLCQAHG
4 LTPQQVVAIAS**NT**GGKQALXTVQRLLPVLCQAHG
5 LTPQQVVAIAS**NN**GGKQALETVQRLLPVLCQAHG
6 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
7 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
8 LTPEQVVAIAS**NN**GGKQALETVQRLLPVLCQAHG
9 LTPEQVVAIAS**ND**GGKQALETVQRLLPVLCQAHG
10 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
11 LTPEQVVAIAS**NI**GGKQALETVQRLLPVLCQAHG
12 LTPQQVVAIAS**NT**GGKQALETVQALLPVLCQAHG
13 LTPQQVVAIAS**NR**GGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIAS**NS**GGKQALETVQRLLPVLCQAHG
15 LTPEQVVAIAS**ND**GGKQALETVQRLLPVLCQAHG
16 LTPEQVVAIAS**NN**GGKQALETVQRLLPVLCQAHG
17 LTPEQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
18 LTPEQVVAIAS**NG**GGKQALETVQRLLPVLCQAHG
19 LTPQQVVAIAS**NS**GGKQALETVQALLPVLCQAHG
+1 LTPQQVVAIAS**N-**GGRPALESIVAQLSRPDPALAA
+2 LTNDHLVALACL-GGRPALDAVKKGLPHAPALIKR
TNRRIPERTSHRVA

>dTALE$_{\mathrm{SOX2}}$ (for human cell transcriptional activation assay)

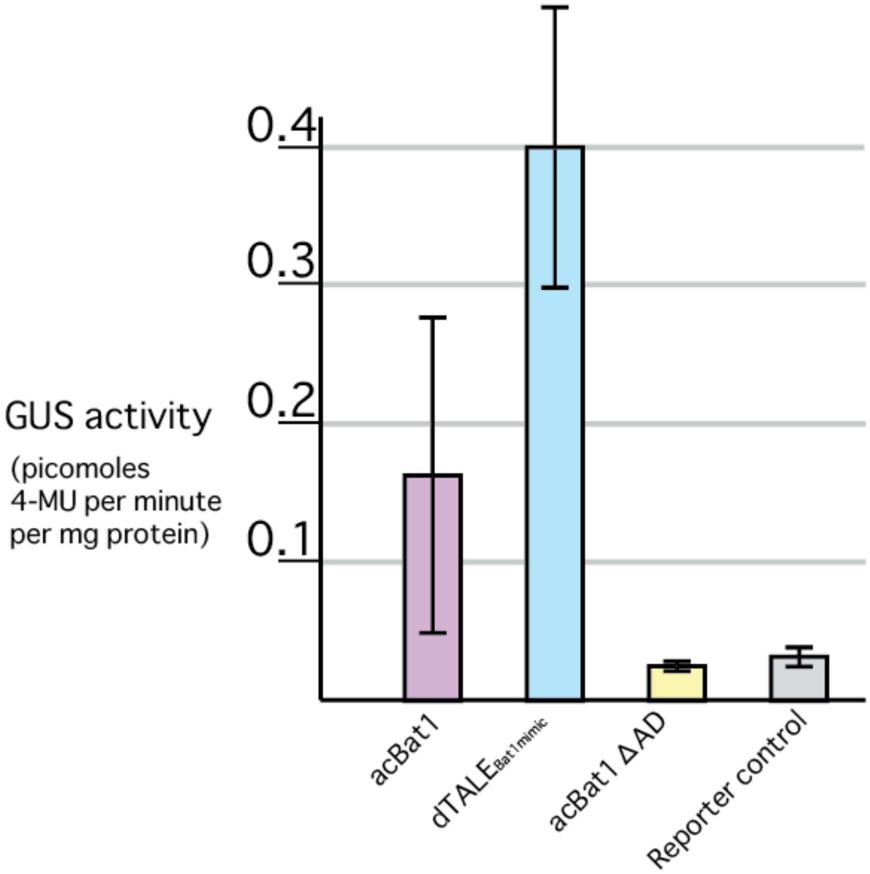
```
MDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGH
GFTHAHIVALSQHPAALGTVAVKYQDMIAALPE
-1 ATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQ
 0 LDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLN
 1 LTPQQVVAIASNIGGKQALETVQRLLPVLCQAHG
 2 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
 3 LTPEQXVAIASNNGGKQALXTVQRLLPVLCQAHG
 4 LTPQQVVAIASNTGGKQALXTVQRLLPVLCQAHG
 5 LTPQQVVAIASNNGGKQALETVQRLLPVLCQAHG
 6 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
 7 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
 8 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
 9 LTPEQVVAIASNDGGKQALETVQRLLPVLCQAHG
10 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
11 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
12 LTPQQVVAIASNTGGKQALETVQALLPVLCQAHG
13 LTPQQVVAIASNRGGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASNSGGKQALETVQRLLPVLCQAHG
15 LTPEQVVAIASNDGGKQALETVQRLLPVLCQAHG
16 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
17 LTPEQVVAIASNGGGKQALETVQRLLPVLCQAHG
+1 LTPQQVVAIASN-GGRPALESIVAQLSRPDPALAA
+2 LTNDHLVALACL-GGRPALDAVKKGLPHAPALIKR
   TNRRIPERTSHRVA
```

>dTALE$_{\mathrm{Bat1mimic}}$ (for nuclease assay)

```
MAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGH
GFTHAHIVALSQHPAALGTVAVKYQDMIAALPE
-1 ATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQ
 0 LDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLN
 1 LTPQQVVAIASNIGGKQALETVQRLLPVLCQAHG
 2 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
 3 LTPEQXVAIASNNGGKQALXTVQRLLPVLCQAHG
 4 LTPQQVVAIASNTGGKQALXTVQRLLPVLCQAHG
 5 LTPQQVVAIASNNGGKQALETVQRLLPVLCQAHG
 6 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
 7 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
 8 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
 9 LTPEQVVAIASNDGGKQALETVQRLLPVLCQAHG
10 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
11 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
12 LTPQQVVAIASNTGGKQALETVQALLPVLCQAHG
13 LTPQQVVAIASNRGGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASNSGGKQALETVQRLLPVLCQAHG
15 LTPEQVVAIASNDGGKQALETVQRLLPVLCQAHG
16 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
17 LTPEQVVAIASNGGGKQALETVQRLLPVLCQAHG
18 LTPEQVVAIASNGGGKQALETVQRLLPVLCQAHG
19 LTPQQVVAIASNSGGKQALETVQALLPVLCQAHG
+1 LTPQQVVAIASN-GGRPALESIVAQLSRPDPALAA
+2 LT
```

Supplementary Figure 9: *in planta* transcriptional activation mediated by acBat1.

BE_{Bat1} was embedded within a 360 base pair fragment of the silent pepper *Bs3* promoter, using the primers listed in Table S2. This promoter derivative was then inserted upstream of *uidA* in the binary vector pGWB3* as previously described (10). Bat1 and TALE derivatives were assembled via BsaI cut-ligation along with the NLSs and VP64 activation domain (Figure S5) into pENTR/D-TOPO (Life technologies) derivatives bearing BsaI sites. They were then transferred to binary vector pGWB442 via LR recombination (Life technologies). *Agrobacterium tumefaciens* strains carrying pGWB442acBat1, pGWB442acBat1ΔAD or $pGWB442dTALE_{Bat1mimic}$ were co-delivered into *Nicotiana benthamiana* leaves alongside a strain carrying the target reporter. In addition the reporter plasmid was delivered alone as a control. The target reporter was a promoter bearing BE_{Bat1} upstream of a *uidA* reporter gene (Figure S6). Leaf discs were harvested after 48 hours and GUS activity quantified (10). Results are shown for three biological replicates with error bars indicating standard deviation.



Supplementary Figure 10: Amino acid sequences of all acBat1 derivatives (dBats) tested in figures 5 and 6.

Dashes indicate truncated residues. Red font is used to highlight residues truncated or rearranged in each case. In all cases repeat numbering is used to identify repeats with those in the wild-type Bat1 protein. Grey background and bold typeface highlights the RVD residues. NND stands for non-repetitive N-terminal Domain. In all cases only the Bat1-derived amino-acids are shown. The sequences of fused domains are given in Figure S5.

>acBat1 Δ18-20

```
NND MSTAFVDQDKQMANRLN
-1  LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0  LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1  FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2  FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3  FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4  FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5  FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6  FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7  FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
 8  FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9  FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10  FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11  FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12  FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13  FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14  FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15  FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16  FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17  FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18  FSRADIVRIAGNGGGAQ----------------
19  ---------------------------------
20  -----------------ALKAVLKYGPVLMQAG
+1  RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>acBat1 Δ16-20

```
   NNDMSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1 FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2 FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3 FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4 FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5 FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6 FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7 FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
 8 FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9 FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14 FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15 FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16 FSRADIVNVAGNNGGAQ----------------
17 ---------------------------------
18 ---------------------------------
19 ---------------------------------
20 ------------------ALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>acBat1 Δ14-20

```
NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1 FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2 FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3 FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4 FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5 FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6 FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7 FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
 8 FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9 FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14 FSQPDIVKIAGNS--------------------
15 ---------------------------------
16 ---------------------------------
17 ---------------------------------
18 ---------------------------------
19 ---------------------------------
20 -------------GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>acBat1 Δ12-20

```
NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1 FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2 FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3 FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4 FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5 FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6 FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7 FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
 8 FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9 FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNT--------------------
13 ---------------------------------
14 ---------------------------------
15 ---------------------------------
16 ---------------------------------
17 ---------------------------------
18 ---------------------------------
19 ---------------------------------
20 -------------GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>acBat1 ΔNTD

```
NND
-1
 0
 1 FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2 FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3 FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4 FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5 FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6 FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7 FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
 8 FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9 FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14 FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15 FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16 FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17 FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18 FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>acBat1 ΔCTD

NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
1 FSQSDIVKIAG**NI**GGAQALQAVLDLESMLGKRG
2 FSRDDIAKMAG**NI**GGAQTLQAVLDLESAFRERG
3 FSQADIVKIAG**NN**GGAQALYSVLDVEPTLGKRG
4 FSRADIVKIAG**NT**GGAQALHTVLDLEPALGKRG
5 FSRIDIVKIAA**NN**GGAQALHAVLDLGPTLRECG
6 FSQATIAKIAG**NI**GGAQALQMVLDLGPALGKRG
7 FSQATIAKIAG**NI**GGAQALQTVLDLEPALCERG
8 FSQATIAKMAG**NN**GGAQALQTVLDLEPALRKRD
9 FRQADIIKIAG**ND**GGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAG**NI**GGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAG**NI**GGAQALQAVLSLGPALRERG
12 FSQPDIVKIAG**NT**GGAQALQAVLDLELTLVEHG
13 FSQPDIVRITG**NR**GGAQALQAVLALELTLRERG
14 FSQPDIVKIAG**NS**GGAQALQAVLDLELTFRERG
15 FSQADIVKIAG**ND**GGTQALHAVLDLERMLGERG
16 FSRADIVNVAG**NN**GGAQALKAVLEHEATLNERG
17 FSRADIVKIAG**NG**GGAQALKAVLEHEATLDERG
18 FSRADIVRIAG**NG**GGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAA**NS**GGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIAS**N**-GGAQALKAVLKYGPVLMQAG
+1

>dBat RVD switch 1

NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
1 FSQSDIVKIAG**NI**GGAQALQAVLDLESMLGKRG
2 FSRDDIAKMAG**NN**GGAQTLQAVLDLESAFRERG
3 FSQADIVKIAG**NI**GGAQALYSVLDVEPTLGKRG
4 FSRADIVKIAG**NN**GGAQALHTVLDLEPALGKRG
5 FSRIDIVKIAA**NT**GGAQALHAVLDLGPTLRECG
6 FSQATIAKIAG**NI**GGAQALQMVLDLGPALGKRG
7 FSQATIAKIAG**NI**GGAQALQTVLDLEPALCERG
8 FSQATIAKMAG**NN**GGAQALQTVLDLEPALRKRD
9 FRQADIIKIAG**ND**GGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAG**NI**GGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAG**NI**GGAQALQAVLSLGPALRERG
12 FSQPDIVKIAG**NT**GGAQALQAVLDLELTLVEHG
13 FSQPDIVRITG**NR**GGAQALQAVLALELTLRERG
14 FSQPDIVKIAG**NS**GGAQALQAVLDLELTFRERG
15 FSQADIVKIAG**ND**GGTQALHAVLDLERMLGERG
16 FSRADIVNVAG**NN**GGAQALKAVLEHEATLNERG
17 FSRADIVKIAG**NG**GGAQALKAVLEHEATLDERG
18 FSRADIVRIAG**NG**GGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAA**NS**GGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIAS**N**-GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ

>dBat RVD switch 2

```
   NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1 FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2 FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3 FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4 FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5 FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6 FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7 FSQATIAKIAGNNGGAQALQTVLDLEPALCERG
 8 FSQATIAKMAGNDGGAQALQTVLDLEPALRKRD
 9 FRQADIIKIAGNIGGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14 FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15 FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16 FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17 FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18 FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>dBat RVD switch 3

```
   NNDMSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1 FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2 FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3 FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4 FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5 FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6 FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7 FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
 8 FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9 FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAGNTGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNIGGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14 FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15 FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16 FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17 FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18 FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>dBat RVD switch 4

```
   NNDMSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1 FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2 FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3 FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4 FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5 FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6 FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7 FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
 8 FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9 FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14 FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15 FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16 FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17 FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18 FSRADIVRIAGN-GGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIASNGGGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>dBat Repeat switch 1

```
NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 2 FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 1 FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 4 FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5 FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 3 FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 6 FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7 FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
 8 FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9 FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14 FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15 FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16 FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17 FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18 FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>dBat Repeat switch 2

```
NND MSTAFVDQDKQMANRLN
-1  LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0  LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1  FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2  FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3  FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4  FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5  FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 7  FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
10  FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
 8  FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9  FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
 6  FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
11  FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12  FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13  FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14  FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15  FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16  FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17  FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18  FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
19  FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
20  LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1  RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>dBat Repeat switch 3

```
NND MSTAFVDQDKQMANRLN
-1  LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0  LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1  FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2  FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3  FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4  FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5  FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6  FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7  FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
 8  FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9  FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10  FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
12  FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
11  FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
13  FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14  FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15  FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16  FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17  FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18  FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
19  FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
20  LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1  RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

>dBat Repeat switch 4

```
NND MSTAFVDQDKQMANRLN
-1  LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0  LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1  FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 2  FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
 3  FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 4  FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
 5  FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6  FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
 7  FSQATIAKIAGNIGGAQALQTVLDLEPALCERG
 8  FSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
 9  FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
10  FNLADIVKMAGNIGGAQALQAVLDLKPVLDEHG
11  FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12  FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13  FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14  FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15  FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16  FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17  FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
20  LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
19  FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
18  FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
+1  RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

Supplementary Figure 11: Nucleotide and amino acid sequences of dBat$_{\mathrm{SOX2}}$-RVD switch and -repeat switch.

Genes encoding the dBats were synthesised with *E. coli* codon usage (GenScript). One block encodes the N- and C-terminal regions including the cryptic repeats, separated by BpiI sites, flanked by BsaI sites. This was assembled via BsaI cut-ligation into the pVAX destination vector. Repeats were encoded on two BpiI flanked modules assembled directly into the destination vector via BpiI cut-ligation. BsaI recognition sites are underlined and BpiI sites grey-highlighted, while bold typeface marks the overlaps created upon digest.

In the amino acid sequences consecutive repeats are numbered, corresponding to the repeats of wild type Bat1. The RVDs (residues at repeat positions 12 and 13) are marked as boldface black letters on grey background. The sections encoded by the BsaI-flanked N- and C- terminal module are underlined.

>Bat1 N-BpiI BpiI-C

GGTCTCT**TATG**AGCACCGCCTTCGTGGACCAAGATAAGCAAATGGCAAACCGCCTGA
ACCTGTCACCGCTGGAACGTAGCAAAATTGAAAAACAATATGGCGGTGCAACCACGC
TGGCTTTTATTAGCAACAAACAGAATGAACTGGCACAAATCCTGAGCCGTGCTGATA
TTCTGAAAATCGCGTCTTACGACTGCGCAGCACATGCACTGCAGGCTGTCCTGGATT
GTGGCCCGATGCTGGGCAAACGCGGTTT**TAGC**TAGTCTTCTAGAAGACTA**GGCG**GTG
CGCAGGCCCTGAAAGCTGTCCTGAAGTATGGTCCGGTGCTGATGCAAGCAGGTCGTA
GCAATGAAGAAATCGTGCACGTTGCCGCTCGTCGTGGTGGTGCTGGCCGTATCCGTA
AGATGGTTGCTCCGCTGCTGGAACGTCAG**GGTG**TGAGACC

>dBat$_{\mathrm{SOX2}}$ Repeat switch AB

GAAGACTT**TAGC**CGCGCAGATATTGTCAAGATCGCGGGTAACGGTGGCGGCGCACAA
GCACTGAAGGCGGTTCTGGAACACGAAGCGACCCTGGATGAAAGCGGCTTTAGTCGC
GCAGATATTGTCAAGATCGCGGGTAACGGTGGCGGCGCACAAGCACTGAAGGCGGTT
CTGGAACACGAAGCGACCCTGGATGAAAGCGGCTTCTCCCGCGATGACATTGCGAAG
ATGGCCGGCAATATCGGCGGTGCACAGACCCTGCAGGCCGTGCTGGATCTGGAATCA
GCCTTTCGTGAACGCGGCTTTTCTCGTGCTGATATTGTCCGTATTGCGGGTAATGGT
GGTGGTGCCCAGGCTCTGAAGGCTGTGCTGGAACATGGTCCGACGCTGAACGAACGT
GGCTTTTCTCGTGCTGATATTGTCCGTATTGCGGGTAATGGTGGTGGTGCCCAGGCT
CTGAAGGCTGTGCTGGAACATGGTCCGACGCTGAACGAACGTGGCTTTCGTCAGGCG
GACATTATCAAGATTGCCGGTAATGACGGTGGCGCCCAGGCACTGCAAGCAGTGATC
GAACATGGCCCGACCCTGCGCCAACACGGTTTTAGCCAGGCGGATATTGTCAAAATC
GCCGGTAACGACGGCGGTACCCAAGCACTGCATGCTGTGCTGGATCTGGAACGTATG
CTGGGCGAACGTGGTTTTCGTCAGGCGGACATTATCAAGATTGCCGGTAATGACGGT
GGCGCCCAGGCACTGCAAGCAGTGATCGAACATGGCCCGACCCTGCGCCAACACGGT
TTTAGTCGCGCAGATATTGTCAAGATCGCGGGTAACGGTGGCGGCGCACAAGCACTG
AAGGCGGTTCTGGAACACGAAGCGACCCTGGATGAAAGCG**GTTT**TAGTCTTC

>dBat$_{SOX2}$ Repeat switch BC

GAAGACTG**GTTT**CTCCCGCATTGATATCGTTAAGATCGCAGCTAACAACGGTGGTGC
TCAAGCCCTGCACGCTGTCCTGGATCTGGGTCCGACGCTGCGCGAATGTGGGTTCTC
GCAGGCAACCATCGCAAAAATCGCTGGCAATATCGGCGGTGCTCAGGCTCTGCAAAT
GGTGCTGGATCTGGGTCCGGCTCTGGGCAAACGTGGTTTTAGCCAGGCGGATATTGT
CAAAATCGCCGGTAACGACGGCGGTACCCAAGCACTGCATGCTGTGCTGGATCTGGA
ACGTATGCTGGGCGAACGTGGTTTTAGCCAGTCTGACATTGTCAAGATCGCCGGTAA
CATTGGCGGTGCACAGGCACTGCAAGCAGTGCTGGATCTGGAAAGTATGCTGGGCAA
ACGTGGTTTCTCGCAGGCCGACATTGTTAAAATCGCCGGTAACAATGGCGGTGCACA
AGCTCTGTATAGTGTGCTGGATGTTGAACCGACCCTGGGTAAACGTGGTTTTCGTCA
GGCGGACATTATCAAGATTGCCGGTAATGACGGTGGCGCCCAGGCACTGCAAGCAGT
GATCGAACATGGCCCGACCCTGCGCCAACACGGTTTTAGCCAGGCGGATATTGTCAA
AATCGCCGGTAACGACGGCGGTACCCAAGCACTGCATGCTGTGCTGGATCTGGAACG
TATGCTGGGCGAACGTGGTTTTCGTCAGGCGGACATTATCAAGATTGCCGGTAATGA
CGGTGGCGCCCAGGCACTGCAAGCAGTGATCGAACATGGCCCGACCCTGCGCCAACA
CGGTTTTAGCCAGGCGGATATTGTCAAAATCGCCGGTAACGAC**GGCG**AAGTCTTC

>dBat$_{SOX2}$ RVD switch AB

GAAGACTT**TAGC**CAGTCTGACATTGTCAAGATCGCCGGTAACGGTGGCGGTGCACAG
GCACTGCAAGCAGTGCTGGATCTGGAAAGTATGCTGGGCAAACGTGGTTTCTCCCGC
GATGACATTGCGAAGATGGCCGGCAATGGTGGCGGTGCACAGACCCTGCAGGCCGTG
CTGGATCTGGAATCAGCCTTTCGTGAACGCGGCTTCTCGCAGGCCGACATTGTTAAA
ATCGCCGGTAACATTGGCGGTGCACAAGCTCTGTATAGTGTGCTGGATGTTGAACCG
ACCCTGGGTAAACGTGGTTTTTCACGCGCTGACATTGTTAAGATCGCCGGTAACGGT
GGCGGTGCCCAAGCACTGCACACGGTCCTGGATCTGGAACCGGCCCTGGGCAAGCGT
GGTTTCTCCCGCATTGATATCGTTAAGATCGCAGCTAACGGTGGTGGTGCTCAAGCC
CTGCACGCTGTCCTGGATCTGGGTCCGACGCTGCGCGAATGTGGGTTCTCGCAGGCA
ACCATCGCAAAAATCGCTGGCAATGATGGCGGTGCTCAGGCTCTGCAAATGGTGCTG
GATCTGGGTCCGGCTCTGGGCAAACGTGGTTTTAGCCAGGCAACCATTGCTAAGATC
GCCGGTAACGATGGCGGTGCACAGGCACTGCAAACGGTCCTGGATCTGGAACCGGCG
CTGTGCGAACGCGGCTTCTCTCAGGCCACCATCGCAAAAATGGCTGGTAACGATGGC
GGTGCACAGGCTCTGCAAACGGTTCTGGATCTGGAACCGGCCCTGCGTAAACGCGAT
TTTCGTCAGGCGGACATTATCAAGATTGCCGGTAATGGTGGTGGCGCCCAGGCACTG
CAAGCAGTGATCGAACATGGCCCGACCCTGCGCCAACACG**GTTT**TAGTCTTC

>dBat$_{SOX2}$ RVD switch BC

GAAGACTA**GTTT**CAACCTGGCAGACATTGTTAAGATGGCTGGTAATAATGGTGGTGC
TCAAGCTCTGCAAGCGGTGCTGGACCTGAAGCCGGTGCTGGACGAACATGGTTTCTC
TCAACCGGATATCGTCAAGATGGCGGGCAACATTGGTGGTGCTCAAGCCCTGCAAGC
CGTCCTGTCACTGGGTCCGGCGCTGCGTGAACGTGGCTTTAGCCAGCCGGATATTGT
CAAAATCGCCGGTAACGACGGCGGTGCACAGGCACTGCAAGCAGTGCTGGATCTGGA
ACTGACGCTGGTTGAACATGGCTTCTCTCAACCGGACATTGTTCGCATCACCGGTAA
TATTGGCGGTGCCCAAGCTCTGCAAGCGGTGCTGGCTCTGGAACTGACCCTGCGTGA
ACGAGGATTTAGCCAACCGGACATCGTGAAAATCGCGGGCAATAACGGCGGTGCTCA
AGCTCTGCAAGCGGTCCTGGATCTGGAACTGACGTTTCGTGAACGCGGCTTTAGCCA
GGCGGATATTGTCAAAATCGCCGGTAACGACGGCGGTACCCAAGCACTGCATGCTGT
GCTGGATCTGGAACGTATGCTGGGCGAACGTGGTTTCTCTCGCGCAGACATTGTGAA
CGTTGCTGACAACAATGGCGGTGCGCAGGCCCTGAAAGCCGTGCTGGAACACGAAGC
CACGCTGAATGAACGTGGCTTTAGTCGCGCAGATATTGTCAAGATCGCGGGTAACGA
TGGCGGCGCACAAGCACTGAAGGCGGTTCTGGAACACGAAGCGACCCTGGATGAAAG
CGGCTTTTCTCGTGCTGATATTGTCCGTATTGCGGGTAATGAT**GGCG**AAGTCTTC

>dBat$_{SOX2}$ RVD switch

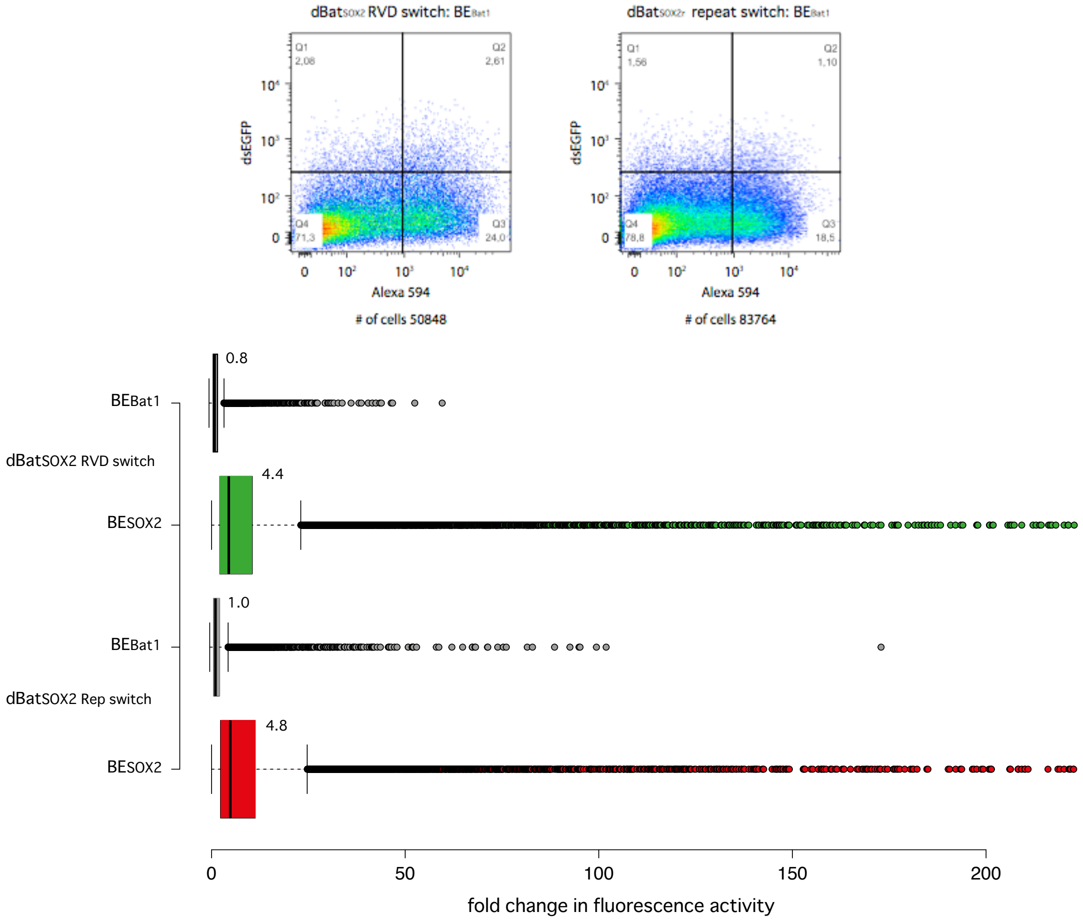
NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
1 FSQSDIVKIAG**NG**GGAQALQAVLDLESMLGKRG
2 FSRDDIAKMAG**NG**GGAQTLQAVLDLESAFRERG
3 FSQADIVKIAG**NI**GGAQALYSVLDVEPTLGKRG
4 FSRADIVKIAG**NG**GGAQALHTVLDLEPALGKRG
5 FSRIDIVKIAA**NG**GGAQALHAVLDLGPTLRECG
6 FSQATIAKIAG**ND**GGAQALQMVLDLGPALGKRG
7 FSQATIAKIAG**ND**GGAQALQTVLDLEPALCERG
8 FSQATIAKMAG**ND**GGAQALQTVLDLEPALRKRD
9 FRQADIIKIAG**NG**GGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAG**NN**GGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAG**NI**GGAQALQAVLSLGPALRERG
12 FSQPDIVKIAG**ND**GGAQALQAVLDLELTLVEHG
13 FSQPDIVRITG**NI**GGAQALQAVLALELTLRERG
14 FSQPDIVKIAG**NN**GGAQALQAVLDLELTFRERG
15 FSQADIVKIAG**ND**GGTQALHAVLDLERMLGERG
16 FSRADIVNVAD**NN**GGAQALKAVLEHEATLNERG
17 FSRADIVKIAG**ND**GGAQALKAVLEHEATLDESG
18 FSRADIVRIAG**ND**GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ

>dBat$_{SOX2}$ repeat switch

```
NND MSTAFVDQDKQMANRLN
-1  LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0  LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
17  FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
17  FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
 2  FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
18  FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
18  FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
 9  FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
15  FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
 9  FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
17  FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
 5  FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
 6  FSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
15  FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
 1  FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
 3  FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
 9  FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
15  FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
 9  FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHG
15  FSQADIVKIAGNDGGAQALKAVLKYGPVLMQAG
+1  RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

Supplementary figure 12: specificity test with the BE_{pSOX2} targeted dBats.

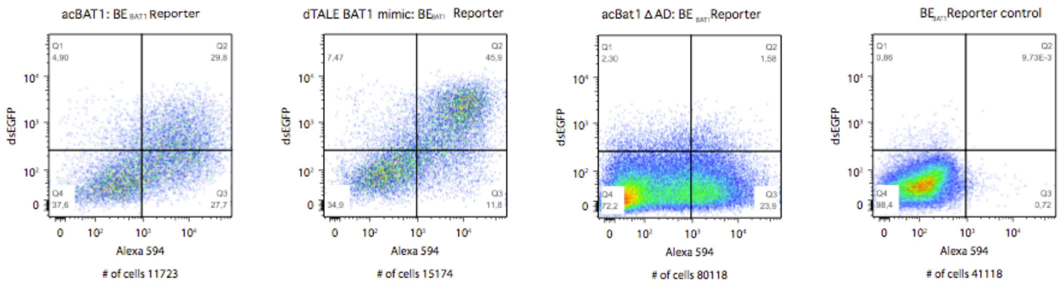
Both dBats were tested against the $\mathrm{BE}_{\mathrm{Bat1}}$ reporter as described in Materials and Methods. The number of cells analysed is indicated below each pseudodensity plot and the vertical bar indicates the threshold Alexa Fluor 594 level above which cells were considered as expressing the relevant Bat or TALE construct and included in downstream analysis. Colour from blue-green to yellow-red indicates increasing cell density. The box plots show fold-change in dsEGFP fluorescence intensity relative to the reporter only control for the two dBats against either the BE_{pSOX2} or $\mathrm{BE}_{\mathrm{Bat1}}$ reporters. Median values are given next to the boxes in each case.



Supplementary Figure 13: Pseudocolour density blots of fluorescence and extended boxplots including outliers for experiments shown in Figures 3, 5-7.

dsEGFP and Alexa Fluor 594 fluorescence levels are shown for all cells analysed for the preparation of figures 3 and 5-7. Data are sorted by figure and transfected constructs are written above the plot in each case. The number of cells analysed is indicated below. The vertical bar indicates the threshold Alexa Fluor 594 level above which cells were considered as expressing the relevant Bat or TALE construct and included in downstream analysis. The x-axis utilises a logical display. Colour from blue-green to yellow-red indicates increasing cell density. Boxplots are also sorted by figure and transfected constructs are given beside each plot. dsEGFP fluorescence is given relative to the reporter alone and is shown only for those cells with above-threshold Alexa Fluor 594 levels.

Figure 3

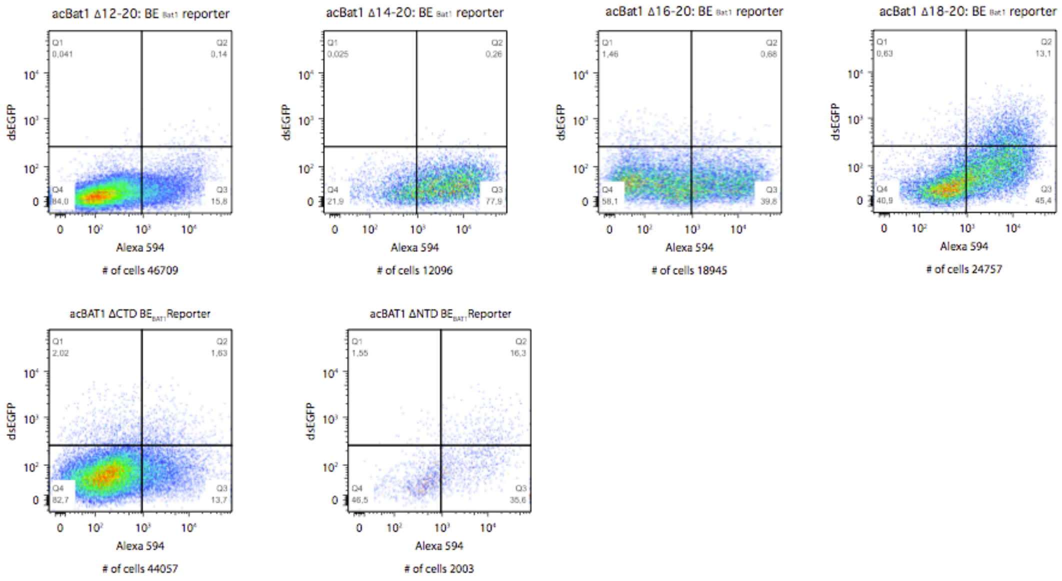

acBAT1: BE$_{BAT1}$ Reporter — # of cells 11723

dTALE BAT1 mimic: BE$_{BAT1}$ Reporter — # of cells 15174

acBat1 Δ AD: BE $_{BAT1}$Reporter — # of cells 80118

BE$_{BAT1}$Reporter control — # of cells 41118

Figure 5


acBat1 Δ12-20: BE $_{Bat1}$ reporter — # of cells 46709

acBat1 Δ14-20: BE $_{Bat1}$ reporter — # of cells 12096

acBat1 Δ16-20: BE $_{Bat1}$ reporter — # of cells 18945

acBat1 Δ18-20: BE $_{Bat1}$ reporter — # of cells 24757

acBAT1 ΔCTD BE$_{BAT1}$Reporter — # of cells 44057

acBAT1 ΔNTD BE$_{BAT1}$Reporter — # of cells 2003

Figure 6

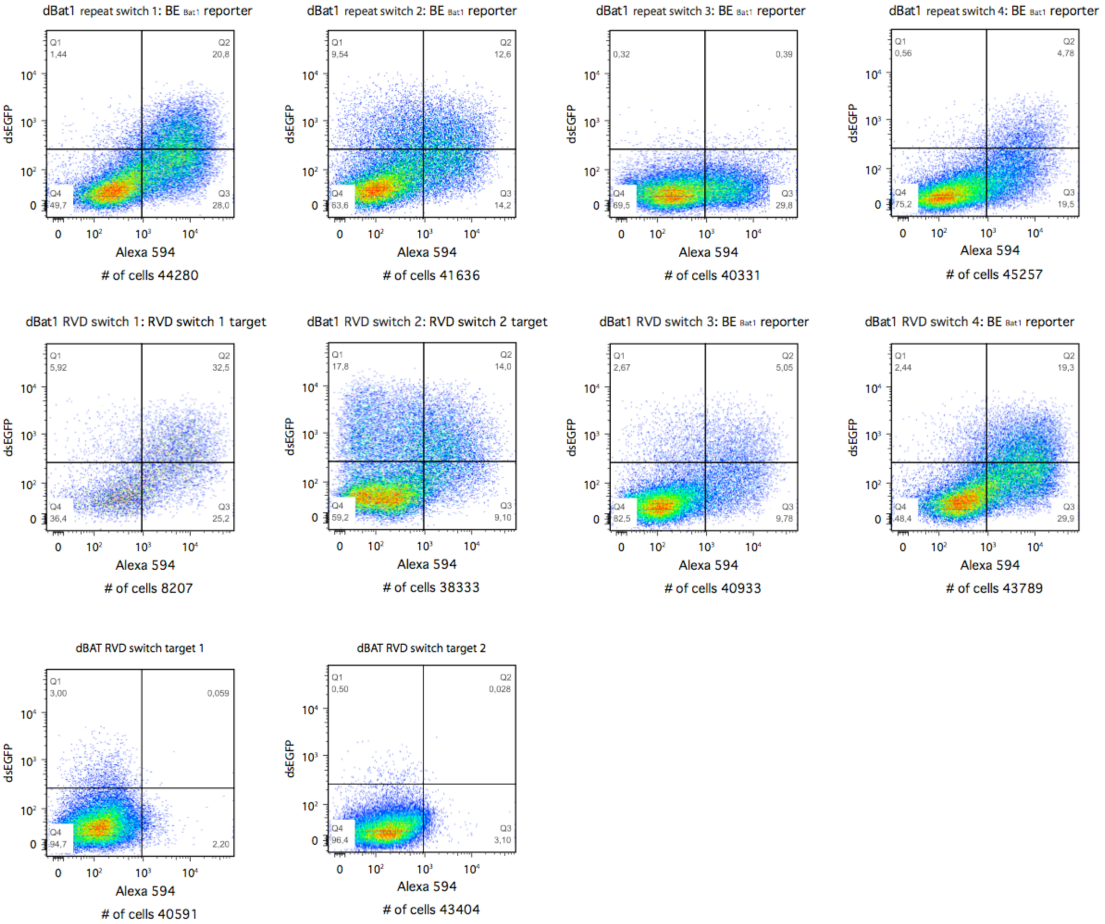


Figure 7

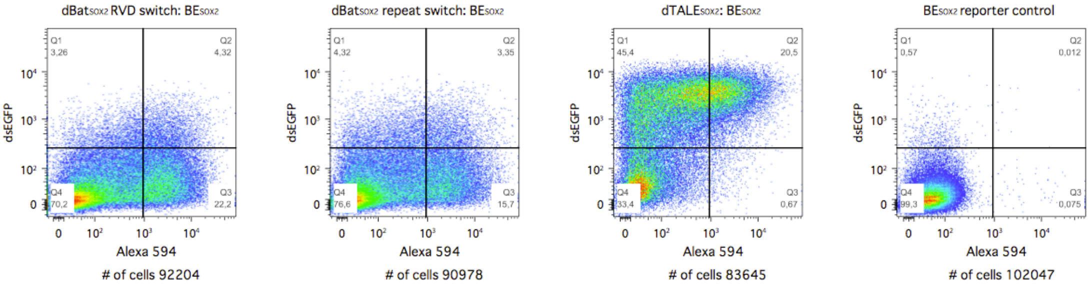


Figure 3

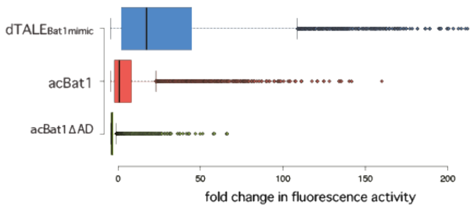


Figure 5

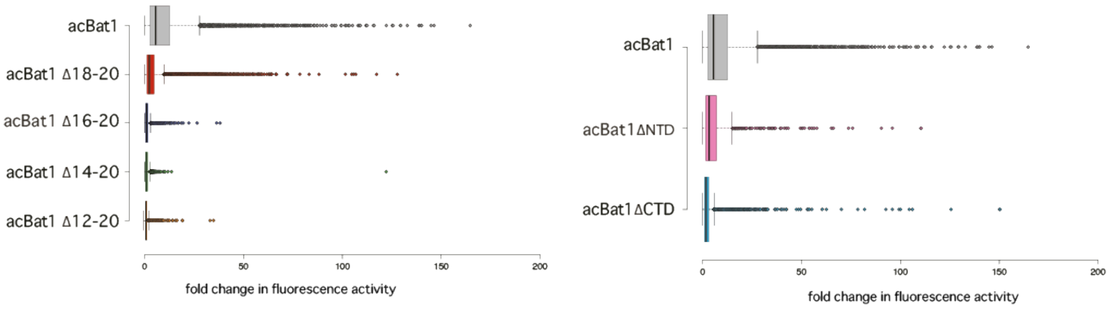


Figure 6

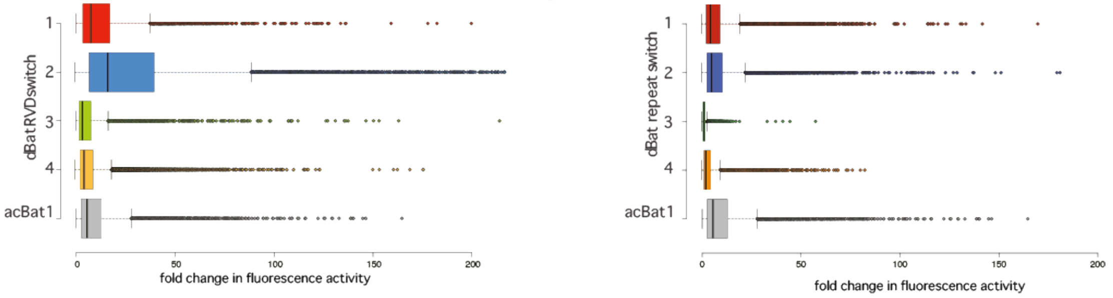
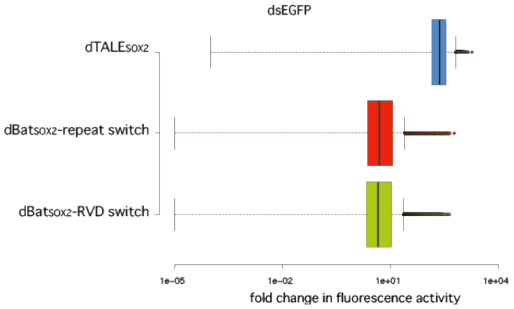


Figure 7



Supplementary Figure 14: Amino acid sequences of the Bat1 repeat trimers used in Figure 8. The sequences corresponding to the Bat repeats are shown in bold and the central repeat of the trimer is underlined to allow each repeat to be identified. Flanking sequences correspond to sections of AvrBs3 necessary for cloning via the previously established toolkit (15). Sequences corresponding to the terminal BpiI recognition sites facilitating compatibility with the TALE binding domain assembly toolkit are highlighted and are removed during cloning.

>Bat1 repeat 2 trimer

EDAETVQRLLPVLCQAHG**FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERGFSRDDIAKMAG**
NIGGAQTLQAVLDLESAFRERGFSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
LTPEQVVAIASQS

>Bat1 repeat 6 trimer

EDAETVQRLLPVLCQAHG**FSQATIAKIAGNIGGAQALQMVLDLGPALGKRGFSQATIAKIAG**
NIGGAQALQMVLDLGPALGKRGFSQATIAKIAGNIGGAQALQMVLDLGPALGKRG
LTPEQVVAIASQS

>Bat1 repeat 8 trimer

EDAETVQRLLPVLCQAHG**FSQATIAKMAGNNGGAQALQTVLDLEPALRKRDFSQATIAKMAG**
NNGGAQALQTVLDLEPALRKRDFSQATIAKMAGNNGGAQALQTVLDLEPALRKRD
LTPEQVVAIASQS

>Bat1 repeat 17 trimer

EDAETVQRLLPVLCQAHG**FSRADIVKIAGNGGGAQALKAVLEHEATLDERGFSRADIVKIAG**
NGGGAQALKAVLEHEATLDERGFSRADIVKIAGNGGGAQALKAVLEHEATLDERG
LTPEQVVAIASQS

Supplementary Figure 15: Structural predictions for Bat1 based on the structure of PthXo1 bound to DNA.

Homology Model: Created using SWISS-MODEL (38)

Template: 3UGM PthXo1

Sequence identity 38.20%

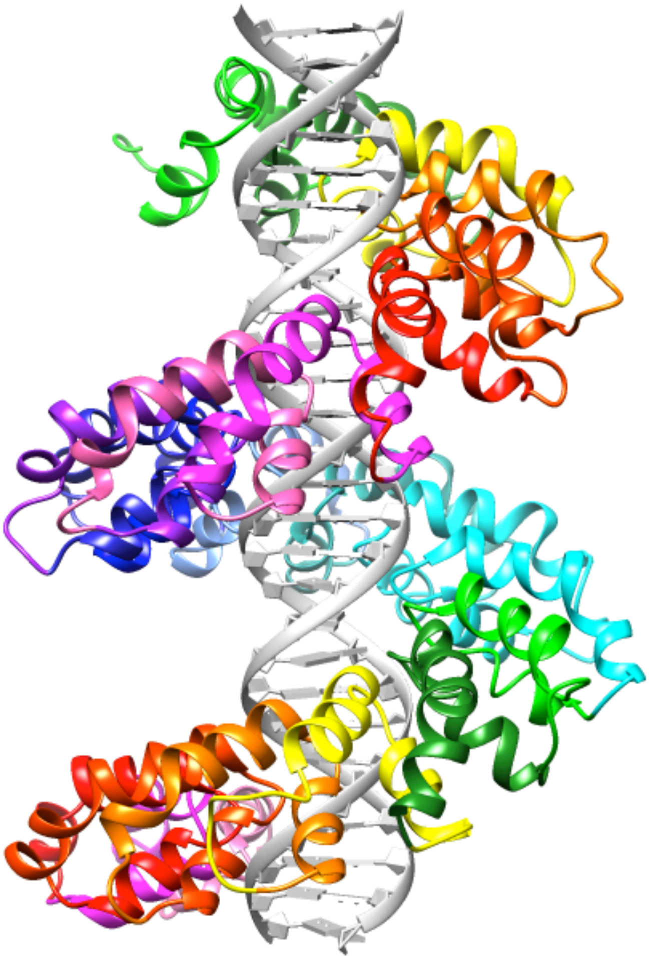
Range of Bat1 covered by the alignment: 11-767

GMQE: 0.70,

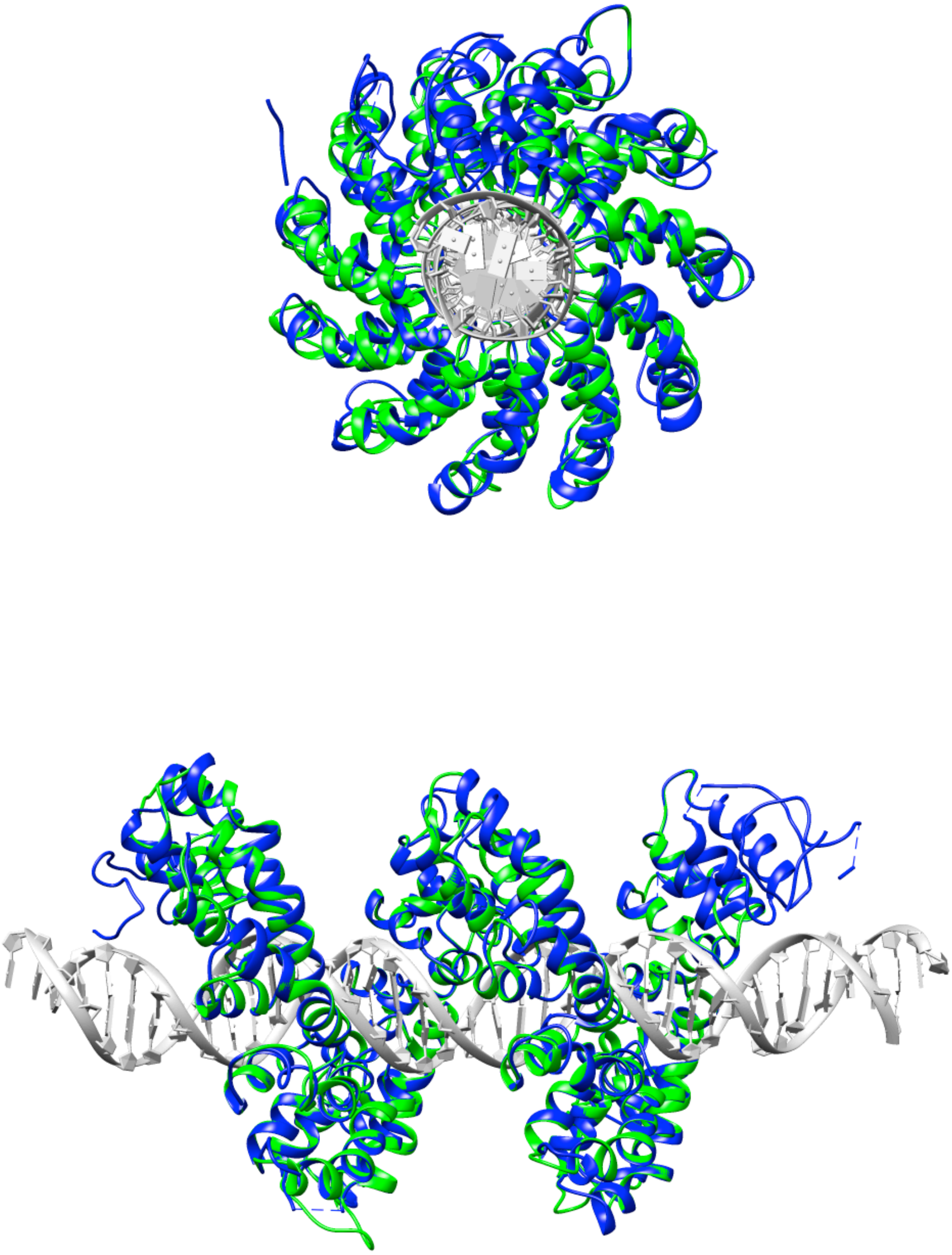
QMEANA4: -6.75,

Diameter of pore: 16.5-19 Angstroms

Average inter-repeat angle: 33°



Model of Bat1 wrapped around BE_{Bat1} (silver) based on the structure of PthXo1 bound to its target DNA shown from N- to C-terminus going down the page. Each repeat is coloured individually.



Longitudinal and transverse views of the Bat1 structural prediction (green) aligned to the structure of PthXo1 (blue). PthXo1 target DNA is shown (silver). Created in UCSF Chimera (39).

Table S1: Percentage sequence identities of the Bat proteins sorted by domain.

	NND	Repeats -1/0	Consensus core repeats	Repeat +1
Bat1 ⇔ Bat2	50	86	94	97
Bat1 ⇔ Bat3	39	66	73	67
Bat2 ⇔ Bat3	50	66	76	67

Consensus core refers to the consensus formed from an alignment of all the core repeats of a single Bat protein. Alignments were performed on CLC Main Workbench 6.1. (Gap open cost 10.0, Gap extension cost 1.0). Percentage identities shown to two significant figures.

Table S2: A list of primers used in this study

Primer name	**Sequence**	**Notes**
pUC57 BB D Fwd	GGG GTC TCT TAA CTA GTC TTC GGG CCC GTC GAC TG	Used to create modified *TALE* toolkit level 2 vector pUC57-CD-DEST. 5' phosphorylated
pUC57 BB C Rev	CCT TGG TCT CAG GGT TAG TCT TCC GAT ATC TAG ATG C	Used to create modified *TALE* toolkit level 2 vector pUC57-CD-DEST
Toolkit_N12_Rev	ATT GCT GGC GAT GGC CAC CAC C	5' Phosphorylated. Used to modify RVDs of *TALE* toolkit repeats
Rep7C_13T_Fwd	ACC GGT GGC AAG CAG GCG CTG	Used to create modified *TALE* toolkit repeat 7C_NT
Rep4_13T_Fwd	ACC GGC AAG CAG GCG CTT GAG	Used to create modified *TALE* toolkit repeat 4_NT
Rep3_ 13D_Fwd	GAC GGT GGC AAG CAG GCG CTG	Used to create modified *TALE* toolkit repeat 3_ND
Rep4_13D_Fwd	GAC GGC AAG CAG GCG CTT GAG	Used to create modified *TALE* toolkit repeat 4_ND
Toolkit_13R_Fwd	CGG GGT GGC AAG CAG GCG CTG	Used to create modified *TALE* toolkit repeat 1C_NR
1/2_13*_Fwd	GGC GGC AGG CCG GCG C	Used to create modified *TALE* toolkit repeat D1/2_N*
rep6_mut_6 ½ Fwd	CGA GAG ACC CCG GGA TCC GAT ATC TAG	Used to create B overlap on toolkit repeat 6 (6 ½ B)
rep_mut_rep ½ Rev	CTA CCA CCT GCT CCG GGG TCA GGC	Used to create B overlap on toolkit repeat 6 (6 ½ B). 5' phosphorylated.
Linker 5-6 ½ Fwd	CGG GTC TCT TGA GGG GGA GCG TGA GAC CTG	Used to create Linker 5-6 in pUC57 with two BsaI sites. Repeats 5_NN and 6 ½ B were then ligated into linker 5-6 to create 5B_NN
Linker 5-6 ½ Rev	CAG GTC TCA CGC TCC CCC TCA AGA GAC CCG	Used to create Linker 5-6 in pUC57 with two BsaI sites. Repeats 5_NN and 6 ½ B were then ligated into linker 5-6 to create 5B_NN
Toolkit D ½ BpiI Rev	GGG GAA GAC CCT AAC CCC GCA GCA GGT GG	Used to create flexible *TALE* toolkit half repeat modules with the D overlap.
pUC57 ½ BpiI Rev	CCC GAA GAC CCA GCG CCG GCC TGC	Used to create flexible *TALE* toolkit half repeat modules with the D overlap.
Rep7_D-overlap_Fwd	TAA CTG AGA CCT GGG CCC GTC GAC TGC AG	Used to create modified *TALE* toolkit repeat 7D_NS
Rep7_D-overlap_Rev	GGC CAT GGG CCT GGC ACA GCA CCG	Used to create modified *TALE* toolkit repeat 7D_NS
pVAX GoldenGate + Sp6 Fwd	ATC AAT GTG AGA CCT TTC CCG GGT TTG GTC TCT GCT TGG GCC CGT TTA AAC CCG CTG ATC AG	Used to remove the previous TALEN gene from a published TALEN expression vector (18), replace it with BsaI sites and introduce an Sp6 priming site into the CMV promoter.
pVAX GoldenGate + Sp6 Rev	ATC ACT AGC TTC TAT AGT GTC ACC TAA ATC AGC TTG AGT CTC CCT ATA GTG AGT CG	Used to remove the previous TALEN gene from a published TALEN expression vector (18), replace it with BsaI sites and introduce an Sp6 priming site into the

		CMV promoter.
HA-NLS GoldenGate AATG Fwd	TTG GTC TCT AAT GGG CTA CCC TTA CGA CGT GC	Used to amplify HA-NLS domain from a published TALEN construct (18) and introduce BsaI sites.
HA-NLS GoldenGate TATG Rev	AAT GGT CTC ACA TAG CGT GGA TGC CCA CTT TCC GC	Used to amplify HA-NLS domain from a published TALEN construct (18) and introduce BsaI sites.
3xHA goldengate Fwd	TTT GGT CTC TAA TGG GGT TAA TTA ACA TCT TTT ACC CAT ACG	Used to amplify 3xHA from binary vector pGWB13 (37) and introduce BsaI sites
3xHA goldengate Rev	TTT GGT CTC ACA TAC CGC TGC ACT GAG CAG CGT AAT C	Used to amplify 3xHA from binary vector pGWB13 (37) and introduce BsaI sites
FokI GGTG BpiI Fwd	TTT GGT CTC TGG TGG TCA GCT AGT GAA ATC TGA ATT GGA AGA G	Used to amplify FokI nuclease domain from a published TALEN construct (18) and introduce BsaI sites.
FokI GGTG BpiI Rev	AAT GGT CTC AAA GCT TAT CTC ACC GTT ATT AAA TTT CCT TCT CAC	Used to amplify FokI nuclease domain from a published TALEN construct (18) and introduce BsaI sites.
*Bat1*_Block 1 TATG Rev	CAT AAG AGA CCA TTG GGA TCG GAT C	Used to modify ‘Bat1 Block1’ (Figure S4) for cloning into the pVAX derived human cell expression vector and remove start codon (provided by N-terminal tag).
*Bat1*_Block 1 ATGless Fwd	AGC ACC GCC TTC GTG GAC CAA G	5’ Phosphorylated. Used to modify ‘Bat1 Block1’ (Figure S4) for cloning into the pVAX derived human cell expression vector and remove start codon (provided by N-terminal tag).
Block 5 GGTG Fwd phospho	GGT GTG AGA CCG ACC CAA TAT C	5’ Phosphorylated. Used to modify ‘Bat1 Block5’ (Figure S5) to remove stop codon and for cloning into the pVAX derived human cell expression vector.
Block5 Last codon Rev	CTG ACG TTC CAG CAG CGG AG	Used to modify ‘Bat1 Block5’ (Figure S5) to remove stop codon and for cloning into the pVAX derived human cell expression vector.
acBat1 AD out Rev phospho	GCT GGC CTC CAC CTT TCT C	Used to remove VP64 activation domain from acBat1 C-terminal domain.
acBat1 AD out Fwd	TAG GCT TTG AGA CCA CGA AG	Used to remove VP64 activation domain from acBat1 C-terminal domain.
acBat1 NLS out Rev	CTT GTC ATC GTC ATC CTT GTA GTC	Used to remove the NLS from the acBat1 C-terminal domain.
acBat1 NLS out Fwd	GGT TCC GGA CGG GCT GAC	5’ phosphorylated. Used to remove the NLS from the acBat1 C-terminal domain.
BAT1rep20 2nd Helix Fwd	GCC CTG AAA GCT GTC CTG AAG TAT G	Used to create acBat1Δ18-20 and acBat1Δ16-20
BAT1rep20 GG Fwd	GGC GGT GCG CAG GCC CTG AAA GCT GTC CTG	Used to create acBat1Δ14-20 and acBat1Δ12-20

	AAG	
BAT1 rep18 1st Helix Rev	CTG GGC ACC ACC ACC ATT ACC CGC	Used to create acBat1Δ18-20
BAT1 rep16 1st Helix Rev	CTG CGC ACC GCC ATT GTT GCC AGC AAC GTT C	Used to create acBat1Δ16-20
BAT1 rep14 1st Helix Rev	GCT ATT GCC CGC GAT TTT CAC GAT GTC CGG TTG	Used to create acBat1Δ14-20
BAT1 rep12 1st Helix Rev	GGT GTT ACC GGC GAT TTT GAC AAT ATC CGG CTG	Used to create acBat1Δ12-20
Bat1 NTD out Fwd	TTT AGC CAG TCT GAC ATT GTC AAG ATC GC	5' phosphorylated. Used to create acBat1ΔNTD
Bat1 NTD out Rev	CAT AAG AGA CCA TTG GGA TCG GAT C	Used to create acBat1ΔNTD
Bat1 CTD out Fwd	AAG GTG AGA CCG ACC CAA TAT C	5' phosphorylated. Used to create acBat1ΔCTD
Bat1 CTD out Rev	ACC TGC TTG CAT CAG CAC CG	Used to create acBat1ΔCTD
BE_{Bat1} into Bs3p Fwd	TGC TTC TCT TAG TTG TGA GGA TGG TTA GG	5' Phosphorylated. Used to create Bs3p BE_{Bat1} for GUS assays and for the creation of the Bat1-Fok1 target templates.
BE_{Bat1} into Bs3p Rev	AAGACGTTAGGTTCAAGT TATCATCCCC	Used to create Bs3p BE_{Bat1} for GUS assays and for the creation of the Bat1-Fok1 target templates.
Bat1-Fok1 target 5bp	CTA GCA TAA CGT CTT TGC TTC TCT TAG	Used to create the 5bp spacer target for the nuclease assays.
Bat1-Fok1 target 7bp	TCT AGA CAT AAC GTC TT GCT TCT C	Used to create the 7bp spacer target for the nuclease assays.
Bat1-Fok1 target 11bp	TAC GTC TAG ACA TAA CGT CTT TGC TTC TC	Used to create the 11bp spacer target for the nuclease assays.
Bat1-Fok1 target 15bp	TAC GTA CGT CTA GAC ATA ACG TCT TTG CTT CTC	Used to create the 15bp spacer target for the nuclease assays.
Bat1-Fok1 target 19bp	TAA GCT ACG TAC GTC TAG ACA TAA CGT C	Used to create the 19bp spacer target for the nuclease assays.
BE_{Bat1} TAGA Fwd	TAG ACT AAG AGA AGC AAA GAC GTT ATA TGC	To get BE_{Bat1} into dsEGFP reporter
BE_{Bat1} CCTA Rev	ATC CGC ATA TAA CGT CTT TGC TTC TCT TAG	To get BE_{Bat1} into dsEGFP reporter

Table S3: p-values for two-tailed t-tests without assuming equal variances to establish whether affinities differ between interactions of Bat1 with BE$_{Bat1}$ derivatives bearing A, C, G or T at the zero position.

	A_0	C_0	G_0	T_0
A_0				
C_0	**0.589**			
G_0	**0.860**	**0.860**		
T_0	**0.231**	**0.382**	**0.754**	

Sample size n=3. (A_0, G_0, T_0) or 5 (C_0). Results shown to three decimal places.

Table S4: Hydrogen bonds formed between repeat residues of Bat1 predicted with UCSF Chimera (39). Unless stated, interactions are between side chain and backbone atoms.

Repeats involved	AA 1	AA 2	Comment
-1 – 0	Gln 29	Ala 59	
0 – 1	Lys 57	Gly 93	
0 - 1	Tyr 61	Ala 92	
1 – 2	Gly 82	Arg 118	In the inter repeat loop region
3 – 4	Asn 160	Ala 191	
3 – 4	Gly 162	Thr 194	
2 – 4	Gly 148	Arg 184	In the inter repeat loop region
4 – 5	Thr 202	His 234	
5 – 6	Lys 222	Gly 258	
6 – 7	Asn 292	Ala 323	
7 – 8	Asn 325	Gly 357	
7 – 8	Gln 349	Arg 343	In the inter repeat loop region
7 - 8	Asn 326 (N)	Asp 359	
8 – 9	Asn 358	Gly 390	
8 – 9	Asn 358	Ala 389	
11 – 12	Lys 420	Gly 456	
11 – 12	Arg 442	Phe 446 (N)	Inter repeat connection
11 – 13	Arg 444	Ser 480 (OH)	
13 – 14	Arg 510	Leu 534 (O)	
13 – 14	Arg 491	Ser 524	
15 – 16	Asn 556	Gly 588	
16 – 17	Asn 589	Gly 621	
16 – 17	Glu 601	Lys 630	
17 – 18	Asp 615	Arg 646	Between two side chains
17 – 18	Glu 634	Lys 663	Between two side chains
19 – 20	Glu 700	Lys 728	Between two side chains
20 - +1	Asn 721	Arg 754 (O)	

36. Szurek, B., Marois, E., Bonas, U., Van den Ackerveken, G. (2001) Eukaryotic features of the Xanthomonas type III effector AvrBs3: protein domains involved in transcriptional activation and the interaction with nuclear import receptors from pepper. Plant J., 26, 523-534.

37. Nakagawa, T., Kurose, T., Hino, K., Tanaka, K., Kawamukai, M., Niwa, Y., Toyooka, K., Matsuoka, K., Jinbo, T., Kimuraf, T. (2007) Development of series of gateway binary vectors, pGWBs, for realizing efficient construction of fusion genes for plant transformation. J. Biosci. Bioeng., 104, 34-41.

38. Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res., 31,3381-3385.

39. Pettersen, E., Goddard, T., Huang, C., Couch, G., Greenblatt, D., Meng, E., Ferrin, T. (2004) UCSF Chimera - a visualization system for exploratory research and analysis. J. Comput. Chem., 25,1605-1612.

Supplementary Material – de Lange, Wolf *et al.*, 2015

Supplementary Figures

S1 - Annotated genomic loci bearing MOrTL ORFs
S2 - Heat map showing percentage pairwise sequence similarities of consensus repeats of different TALE-like groups
S3 - Protein expression gel for MOrTL1 and MOrTL2 and EMSA gel for MOrTL1 against BE_{MOrTL1}.
S4 - Sequence of EBN91408-MOrTL2
S5 - Quantifications of protein:DNA relative to free-DNA in EMSAs shown in Figure 6.
S6 - Maps of E. coli repressor reporter plasmids pCherry and pBT102*
S7 - Sequences of pCherry reporter constructs
S8 - Sequences of Bat1-MOrTL and TALE-MOrTL reporter constructs
S9 - MD Analysis (RMSD) showing spatial stability of modeled protein-DNA complexes.
S10 - Potential energy plots for protein-DNA complexes over course of MD simulation.
S11 - Core repeat alignments of a representative TALE and RipTAL

Supplementary Tables

S1 – The TALE code
S2 – Oligonucleotides used in this study (EMSA probes and PCR primers)
S3 – Averaged base-BSR distances from MD model of $Bat1_{M1\ 6\text{-}10}$
S4 – Averaged base-BSR distances from MD model of $Bat1_{M2\ 6\text{-}10}$
S5 – TALE likes used in the creation of repeat sequence logos shown in Figure 7

Supplementary Figure 1: Annotated genomic loci bearing MOrTL1 and 2 ORFs.

MOrTL1 (GenBank ECG96326) is a translation of a predicted ORF found in marine bacterial genomic contig *EM567463.1* (available at GenBank). This contig is an assembly of two reads both bearing ORFs encoding similar repeat array proteins: JCVI_READ_1093012032286 (a) encoding ECG96325 (b) and JCVI_READ_1092963399564 (c) encoding MOrTL1 (d). These sequences form part of environmental sample ID 1103283000023 from the Global Oceanic Survey. Sample Metadata are available via the CAMERA metagenomics data distribution centre: http://camera.crbs.ucsd.edu/projects/details.php?id=CAM_PROJ_GOS. Sequences from this dataset were obtained by paired-end Sanger sequencing of a plasmid library of sheared microbial DNA.

MOrTL2 (GenBank EBN91409) is a translation of a predicted ORF found in marine bacterial genomic contig *EN814823.1*. This contig is also an assembly of two reads each bearing similar repeat protein ORFs: JCVI_READ_1091143078068 (e) encoding EBN91408 (f) and JCVI_READ_1091143109172 (f) encoding MOrTL2 (h). These sequences form part of environmental sample ID 1103283000022 from the Global Oceanic Survey.

Synthesised coding sequences for MOrTL1 (i) and MOrTL2 (j) used in this study are also provided.

Note that In addition to the sequences presented here a further accession from the same metagenomics dataset, GenBank accession *EMO47375.1*, bears an ORF encoding repeats similar to those of MOrLT1. However, there are only three repeats in this ORF and it was not taken as a candidate for DNA binding assays in this study.

(a)
>JCVI_READ_1093012032286

```
GTAGGCTGAGGCTTAGATAGTTGGGACAAGTTAGTTGAAAAGGATTGGATAAGAACG
CCATTTTAAAGATTTCAATTTGTAACGGGGCTCATTTGGCGATTACCACGTTACTAG
AAAACTGGGATGCGTTAATAGATTTGGAACTGGAACCCAAAGATATTGTATCTATTG
CGTCTCATGGTGGGGCAACTCAGGCGATTACCACGTTACTAAACAAGTGGGATGACT
TAAGAGATAAGGGACTGGAACCCAAAGATATTGTATCCATTGCGTCTAATAATGGCG
CAACTCAGGCTATTGCTACGTTATTAGCAAAATGGGATTCCTTAATAGCTAAGGGAC
TGCAGCCCAAAGATATTGTATCCATTGCGTCTCATGGTGGGGCAACTCAGGCTATTA
CCACGTTACTAAACAGGTGGGGTGACTTAAGAGCTAAGGAACTGGAACCCAAAGATA
TTGTATCCATTGCGTCTCATGATGGGGCAACTCAGGCTATTACCACGTTACTAGAAA
AATGGGATGAGTTAAGAGCTAAGGGACTGGAACCCAAAGATATTGTATCCATTGCGT
CTCATATTGGCGCAAATCAGACTATTACTACGTTACTAAACAAGTGGGGTGCGTTAA
TAGATTTGGAACTGGAACCCAAAGATATTGTATCCATTGCGTCTCATGGTGGGGCAA
ATAAGGCTATTACCACGTTACTAGAAAAGTGGGCTGCCTTAAGAGCTAAGGAACTGG
AACCCAAAGATATTGTATCCATTGCGTCTCATAATGGAGCAACTCACGCTATTACTA
CGTTACTAAACAAGTGGGCTGCCTTAAGAGCTAAAGAACTGGAACCCAAAGATATTG
TATCCATTGCGTCTCATAATGGAGCAACTCACGCTATTACCATGTTATTAAACAAGT
GGGGTGACTTAAGAGCTAAGAACTGGAACCCAAAGATATTGTGTCCATTGCGTCACA
TGATGGGGCAACTCATGCTATTACTACGTTACTAGAAAATGGGATGAGTTAGAGCTA
```

```
ATGGTACTGCACCCAAAGATATTGTATCTATTGCGTCTATATGGCGCAAATCAGCGA
TTTCCACGTTACTAGAAAAGTGGGGTGCGTTATAG
```

(b)

```
>JCVI_READ_1093012032286 translation frame +3
RLRLR*LGQVS*KGLDKNAILKISICNGAHLAITTLLENWDALIDLE
LEPKDIVSIASHGGATQAITTLLNKWDDLRDKG
LEPKDIVSIASNNGATQAIATLLAKWDSLIAKG
LQPKDIVSIASHGGATQAITTLLNRWGDLRAKE
LEPKDIVSIASHDGATQAITTLLEKWDELRAKG
LEPKDIVSIASHIGANQTITTLLNKWGALIDLE
LEPKDIVSIASHGGANKAITTLLEKWAALRAKE
LEPKDIVSIASHNGATHAITTLLNKWAALRAKE
LEPKDIVSIASHNGATHAITMLLNKWGDLRAKN
WNPKILCPLRHMMGQLMLLLRY*KMG*VRANGTAPKDIVSIASIWRKSAISTLLEKW
GAL*
```

(highlighted section = ECG96325)

(c)

```
>Reverse complement of JCVI_READ_1092963399564
ATGGCGCAAATCCAGGCGATTTCCACGTTACTAGAAAAGTGGGGTGCGTTAATAGAT
TTGGAACTGGAACCCAAAGATATGTATCCATGCGTCTCATAATGAGCAAATCAGGCG
ATTACACGTTACTAAACAAGTGGGTGACTTAAGAGCTAAGGAACTGGAACCCAAAGA
TATTGTGTCCATTGCGTCTAATACTGGCGCAAATAAGACTATTACCAGGTTACTAGA
AAAGTGGGGTGACTTAAGAGCTAAGGAACTGGAACCCAAAGATATTGTATCCATTGC
GTCACATGATGGGTCAAATCAGACTATTACAAAGTTACTAGAAAAATGGGATGAGTT
AAGAGCTAAGGGACTGGAGCCCAAAGATATTGTATCCATTGCGTCTCATATTGGCGC
AAATCAGACTATTACTACGTTACTAAACAAGTGGGGTGCGTTAATAGATTTGGAACT
GGAACCCAAAGATATTGTATCCATTGCGTCTCATATTGGCGCAACTCAGGCTATTAC
TACGTTACTAAACAAGTGGGCTGCCTTAAGAGCTAAGGGACTGGACCCCAAAGATAT
TGTATCTATTGCGTCACATGATGGGTCAAATCAGACGATTACAAAGTTACTAGAAAA
ATGGGATGAGTTAAGAGCTAAGGAACTGGAATCCAAAGATATTGTATCCATTGCGTC
TAATAATGGCGCAACTCAGACTATTACCAGGTTACTAGAAAAATGGGATGAGTTAAG
AGCTAAGGGACTGGACCCCAAAGATATTGTATCCATTGCGTCTCATGGTGGTGCAAC
TCAGGCTATTACCACGTTACTAAACAGGTGGGGTGACTTAATAGATTTGGAACTGGA
ACCCAAAGATATTGTATCCATTGCGTCTCATAAAGGAGCAAATCAGGTTATTACTAC
GTTACTAGAAAAGTGGGATGACTTAATTAGTCAGGCATATACTAAGTCTAGCATTGT
GAGTATTGCTTCTACTCAGAATGGCGTATTAGGCCTATTGGAGGCGTTAGGTTAATA
ACATTATTTTCAAAGTAAAAAAGGGTTTATAAATACTGGAATATATTACTGATTATT
AAGTAAGGGAGTCTGCAATCCGTTAC
```

(d)

```
>Reverse complement of JCVI_READ_1092963399564 translation Frame +2
WRKSRRFPRY*KSGVR**IWN
WNPK ICIHASHNE QIRRLHVTKQVGDLRAKE
LEPKDIVSIASNTGANKTITRLLEKWGDLRAKE
LEPKDIVSIASHDGSNQTITKLLEKWDELRAKG
LEPKDIVSIASHIGANQTITTLLNKWGALIDLE
LEPKDIVSIASHIGATQAITTLLNKWAALRAKG
LDPKDIVSIASHDGSNQTITKLLEKWDELRAKE
LESKDIVSIASNNGATQTITRLLEKWDELRAKG
```

LDPKDIVSIASHGGATQAITTLLNRWGDLIDLE
LEPKDIVSIASHKGANQVITTLLEKWDDLISQA
YTKSSIVSIASTQNGVLGLLEALG**HYFQSKKGFINTGIYY*LLSKGVCNPL
(highlighted section = MOrTL1/ECG96326)

(e)
>JCVI_READ_1091143078068
GTGGCCCCGTCGGCTTGACCACATAACTAACTTTTGTTGAGTTTCAGGGTTCAAGCA
TTAACTAATTAGGATTGCATGGTGTGAGAACATATTATTAATTTATATTTTGCAAGG
AGTTTTGTATTTATGAGTAATCAAACAGAGCAAAAAATTCTAAAGTTTAAGCTAGAG
CTGCGCTATCCAACAGAATCAGCTCAATTAATACGTGCTGGATTTAATCGAGATCAA
GCGGATAGGATTATCTTAAGAGGCTCTTCACAACGTACCGTTGCAAAGTTACTGGAA
ATTCACAAGACGTTGTTAGCTCATCCCTATAGAATAACCTACGACGACCTCACTCGA
ATTGCAGCAAGAAATGGAGGCTCTAAAAACTTAGTGGCGGTGCAAGCAAACTATGCT
GCCTTAACAGAACTCGGGTTTAGTGCTAAGGATATTGTGCAGATGGTGTCACATGGT
GGAGGCTCTAAAAACTTAGAGGTGGTACAAGCAAACTATGCTGCCTTAACAGGACTC
GGGTTTCGTACTGAGGATATTGTGCAGATGGTGTCACATGATGGAGGCTCTAAAAAC
TTAGCGGCTATGATAGACAAGTCTACTGCCTTAAAAGACCTTGGGTTTCGTACTGAG
GATATTGTGCAGATGGTGTCACATGATGGAAGCTCTAAAAACTTAGCGGCTATGATA
GACAAGTCTACTGCCTTAAAAGGCCTCGGATTTCGTACTGAGGGTATTGTGCAGATG
GTGTCACATGGGTGGAGGCTCTAAAAACTTAGTGGCGGTGCAAGCAAACTATGCTGC
CTTAACAGGACTCGGATTTCGTACTGAGGGTATTGTGCAGATGGTGTCACATGGTGG
AGGCTCTAAAAACTTAGTGGCGGTGCAAGCAAACTATGCTGCCTTAACAGGACTCGG
GTTTCGTACTGAGGATATTGTGCAGATGGTGTCACATGATGGAGGCTCTAAAACTTA
GCGGCTATTATAGACAAGTCTACTGCCTTATAGGCCTTGGGTTTCGTACTGAGGATA
TTGTGCAGATGGTGTCTAACAATGGAGGCTCTAAAACTTAGCGGCTAGATAGACAAG
TCTACTGCCTTAAAAGGCGCCCGATTTCGTACTGAAGAGATTGTTGCCCATGGTGTC
CCATGGGTGGGAGGGCTCTTACAAACTATAAAGGGGGTGGGAGGGCGGAC

(f)
>JCVI_READ_1091143078068 translation frame +1
VAPSA*PHN*LLLSFRVQALTN*DCMV*EHIINLY
FARSFVFMSNQTEQKILKFKLELRYPTESAQLIRAG
FNRDQADRIILRGSSQRTVAKLLEIHKTLLAHPYR
ITYDDLTRIAARNGGSKNLVAVQANYAALTELG
FSAKDIVQMVSHGGGSKNLEVVQANYAALTGLG
FRTEDIVQMVSHDGGSKNLAAMIDKSTALKDLG
FRTEDIVQMVSHDGSSKNLAAMIDKSTALKGLG
FRTEGIVQMVSHGWRL*KLSGGASKLCCLNRTRISY*GYCADGVTWWRL*KLSGGAS
KLCCLNRTRVSY*GYCADGVT*WRL*NLAAIIDKSTAL*ALGFVLRILCRWCLTMEA
LKLSG*IDKSTALKGARFRTEEIVAHGVPWVGGLLQTIKGVGGR
(highlighted section = EBN19408)

(g)
>Reverse complement of JCVI_READ_1091143109172
CTTAGCGGCTATGATAGACAAGTCTACTGCCTTAAAAGACTTCGGGTTTCGTACTGA
GGATATGTGCAGATGGTGTCACATGATGGAGGCTCTAAAAACTTAGCGGCTATGATA
GACAAGTCTACTGCCTTAAAAGGCCTCGGATTTCGTACTGAGGGTATTGTGCAGATG
GTGTCACATGGTGGAGGCTCTAAAAACTTAGTGGCGGTGCAAGCAAACTATGCTGCC
TTAACAGGACTCGGATTTCGTACTGAGGGTATTGTGCAGATGGTGTCACATGGTGGA
GGCTCTAAAAACTTAGTGGCGGTGCAAGCAAACTATGCTGCCTTAACAGGACTCGGG

TTTCGTACTGAGGATATTGTGCAGATGGTGTCACATGATGGAGGCTCTAAAAACTTA
GCGGCTATTATAGACAAGTCTACTGCCTTAACAGGCCTTGGGTTTCGTACTGAGGAT
ATTGTGCAGATGGTGTCTAACAATGGAGGCTCTAAAAACTTAGCGGCTATTATAGAC
AAGTCTACTGCCTTAAAAGGCCTCGGATTTCGTACTGAGGATATTGTGCAGATGGTG
TCACATGGTGGAGGCTCTAAAAACTTAGAGGTGGTGCAAGCAAACTATGCTGCCTTA
ACAGGACTCGGATTTCGTACTGAGGGTATTGTGCAGATGGTGTCACATGGTGGAGGC
TCTAAAAACTTAGTGGCGGTGCAAGCAAACTATGCTGCCTTAACAGGACTCGGGTTT
CGTACTGAGGATATTGTGCAGATGGTGTCACATGATGGAGGCTCTAAAAACTTAGCG
GCTATGATAGACAAGTATACTGCCTTAAAAGACCTTGGGTTTCGTACTGAGGATATT
GTGCAGATGGTGTCACATGATGGAGGCTCTAAAAACTTAGCGGCTATTATAGACAAG
TCTACTGCCTTAAAAGGCCTCGGATTTCTTACTGAGGATATTGTGCAGATGGTGTCA
CATGATGGAGGCTCTAAAAACTTAGAGGTGGTGCAAGCAAGCTATGATACCTTAACA
GAACTCAAGTTTAGTGCTGAGCATCTCAGCCCTTC

(h)
>Reverse complement of JCVI_READ_1091143109172 translation frame +1
LSGYDRQVYCLKRLRVSY*GYVQMVSHDGGSKNLAAMIDKSTALKGLG
FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
FRTEDIVQMVSHDGGSKNLAAIIDKSTALTGLG
FRTEDIVQMVSNNGGSKNLAAIIDKSTALKGLG
FRTEDIVQMVSHGGGSKNLEVVQANYAALTGLG
FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
FRTEDIVQMVSHDGGSKNLAAMIDKYTALKDLG
FRTEDIVQMVSHDGGSKNLAAIIDKSTALKGLG
FLTEDIVQMVSHDGGSKNLEVVQASYDTLTELKFSAEHLSP
(highlighted section = MOrTL2/ EBN19409)

Note, the sequences at the N-terminus of MOrTL2 (RVLCRWCHM) differs from the sequence above. This is because MOrTL2 is a translation of the assembled sequence not the individual reads. Differences arise in the N-terminal section of MOrTL2 from reconciling polymorphic bases between this read and read 1091143078068. Looking at translations of the raw reads it seems likely that sequencing did not cover the whole insert and further MOrTL repeats separate the two reads.

(i) MOrTL 1/ECG96326 – synthesized CDS (GenScript). BsaI restriction enzyme binding sites underlined and overlaps italicized. Start and stop codons bold.

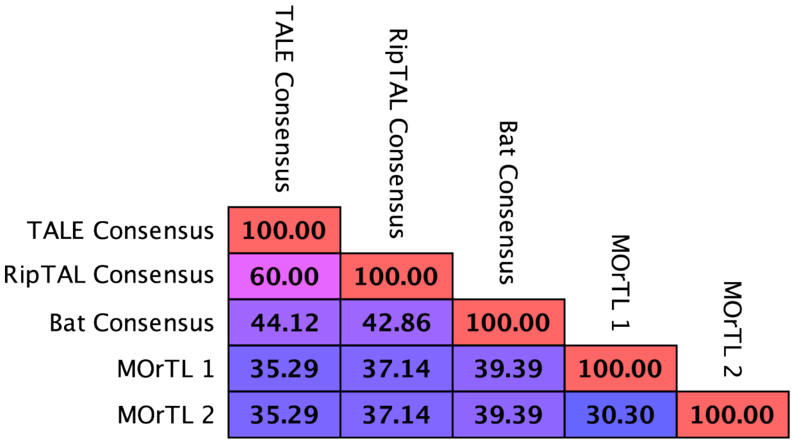
>MOrTL1_CDS_Genscript
GGTCTCA*T**ATG***GTTGGCGATCTGCGTGCGAAAGAACTGGAACCGAAAGACATTGTGA
GCATTGCCTCTAACACCGGCGCGAATAAAACGATTACCCGCCTGCTGGAAAAATGGG
GCGATCTGCGTGCCAAGGAGCTGGAACCGAAAGATATTGTCAGCATCGCCTCTCATG
ACGGCAGTAACCAGACCATTACGAAACTGCTGGAAAAATGGGATGAACTGCGCGCAA
AAGGTCTGGAACCGAAAGATATCGTGAGTATCGCATCCCACATTGGCGCTAACCAAA
CGATCACCACGCTGCTGAATAAATGGGGTGCACTGATTGATCTGGAATTAGAGCCGA
AAGATATCGTTTCAATCGCTTCGCATATTGGTGCAACCCAGGCTATCACCACGCTGC
TGAACAAATGGGCGGCCCTGCGTGCAAAAGGCCTGGATCCGAAAGACATTGTCAGCA
TCGCTTCTCACGATGGTTCTAATCAAACGATCACCAAGTTACTGGAAAAATGGGACG
AACTGCGCGCCAAAGAACTGGAAAGCAAAGACATTGTGAGTATCGCGTCCAACAATG
GCGCCACCCAGACGATCACCCGTCTGCTGGAGAAGTGGGACGAACTGCGCGCGAAAG

GTCTGGATCCGAAAGATATCGTGAGCATCGCATCGCATGGCGGTGCAACCCAGGCAA
TTACCACGCTGCTGAACCGTTGGGGCGATCTGATCGACCTGGAATTAGAACCTAAAG
ACATTGTGAGCATCGCATCTCACAAAGGTGCTAATCAGGTTATTACCACGCTGCTGG
AAAAATGGGACGACCTGATCAGTCAAGCGTATACCAAATCCTCAATCGTGTCAATCG
CATCAACGCAAAATGGTGTCCTGGGTCTGCTGGAAGCCCTGGGT**TAG***GGTG*AGAGAC
C

(j) MOrTL 2/EBN91409 – Synthesised CDS (GenScript). BsaI restriction enzyme binding sites underlined and overlaps italicized. Start and stop codons bold.

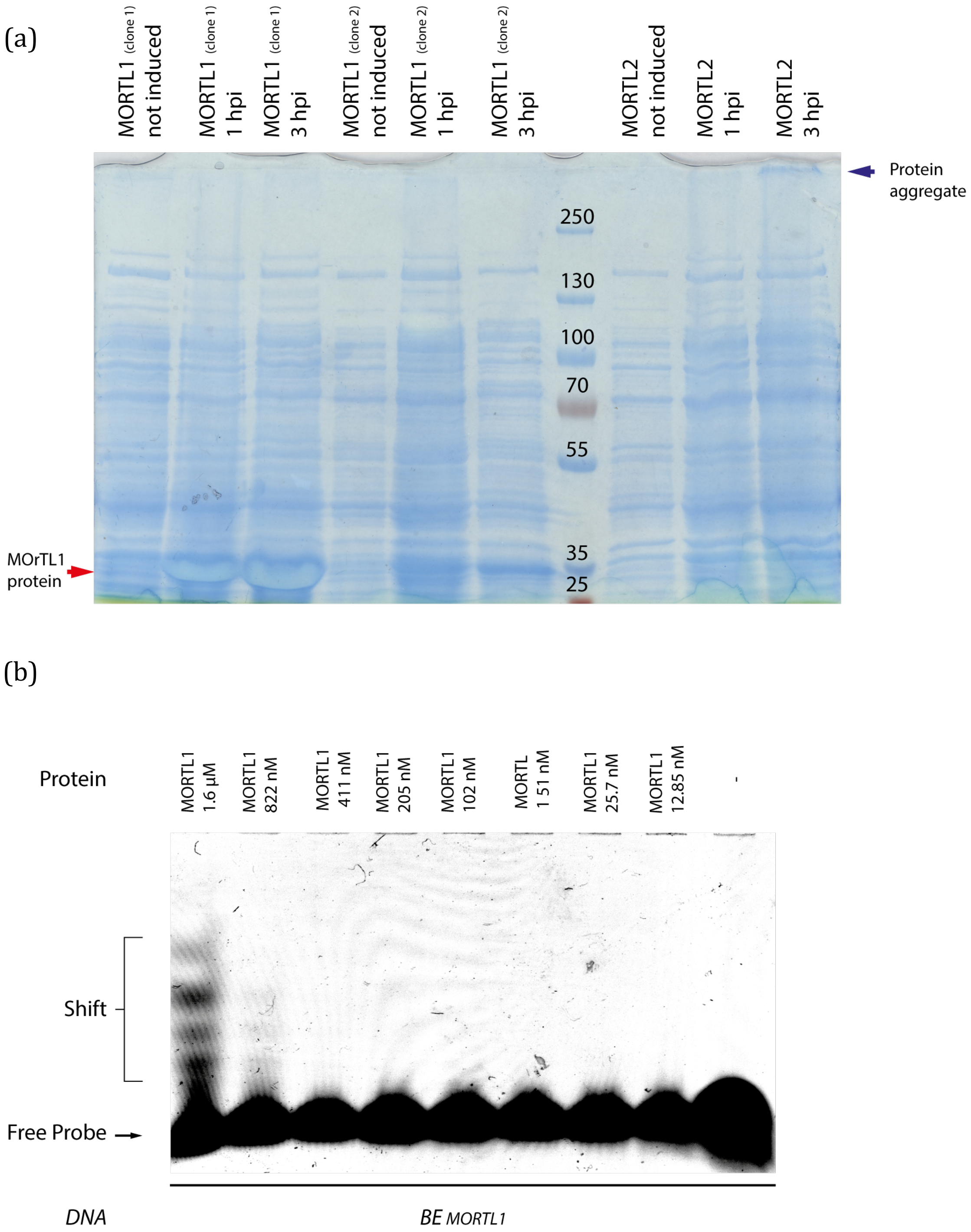
>MOrTL2_CDS_Genscript
GGTCTCA*T**ATG***ATGCGCGTTCTGTGTCGTTGGTGCCACATGGGCGGCGGCTCTAAAA
ATCTGGTTGCTGTTCAAGCTAACTATGCGGCTCTGACGGGCCTGGGTTTTCGTACCG
AAGGCATTGTCCAGATGGTGAGCCATGGCGGTGGCTCTAAAAACCTGGTCGCGGTGC
AAGCCAATTATGCAGCACTGACCGGTCTGGGCTTCCGTACGGAAGATATTGTTCAGA
TGGTCAGTCACGATGGTGGCTCCAAAAACCTGGTTGCAGTCCAAGCTAATTACGCAG
CTCTGACCGGTCTGGGCTTTCGTACGGAAGATATTGTGCAGATGGTTTCACATGATG
GTGGCTCGAAAAACCTGGCGGCCATTATCGACAAAAGTACCGCACTGACGGGTCTGG
GCTTCCGTACCGAAGATATCGTCCAAATGGTGAGCAACAATGGTGGCTCTAAAAATC
TGGCAGCTATTATCGATAAAAGCACCGCCCTGAAAGGTCTGGGCTTCCGCACCGAAG
ATATTGTCCAAATGGTCAGTCACGGTGGCGGTTCCAAAAATCTGGAAGTGGTGCAGG
CCAACTACGCCGCCCTGACGGGTCTGGGCTTTCGCACCGAAGGTATCGTTCAAATGG
TTTCACATGGCGGTGGCTCGAAAAATCTGGTGGCAGTTCAAGCGAACTATGCCGCCT
TAACGGGTCTGGGCTTTCGTACCGAAGATATTGTCCAGATGGTTAGCCACGATGGTG
GCTCTAAGAATCTGGCGGCCATGATTGATAAATATACCGCGCTGAAAGACCTGGGTT
TCCGCACGGAAGATATCGTGCAGATGGTTAGTCATGACGGTGGCTCCAAAAATCTGG
CCGCCATTATCGATAAATCTACGGCGCTGAAAGGTCTGGGCTTTCTGACCGAAGATA
TTGTTCAAATGGTGAGCCACGATGGCGGTAGCAAAAACCTGGAAGTGGTGCAAGCAT
CATACGACACGCTGACGGAACTGAAATTC**TAG***GGTG*AGAGACC

Supplementary Figure 2: Heat map of percentage pairwise sequence similarities of consensus repeats of different TALE-like groups. Based on alignments of representative TALEs and RipTALs, Bat1 and Bat2, MorTl1, and MOrTL2



	TALE Consensus	RipTAL Consensus	Bat Consensus	MOrTL 1	MOrTL 2
TALE Consensus	100.00				
RipTAL Consensus	60.00	100.00			
Bat Consensus	44.12	42.86	100.00		
MOrTL 1	35.29	37.14	39.39	100.00	
MOrTL 2	35.29	37.14	39.39	30.30	100.00

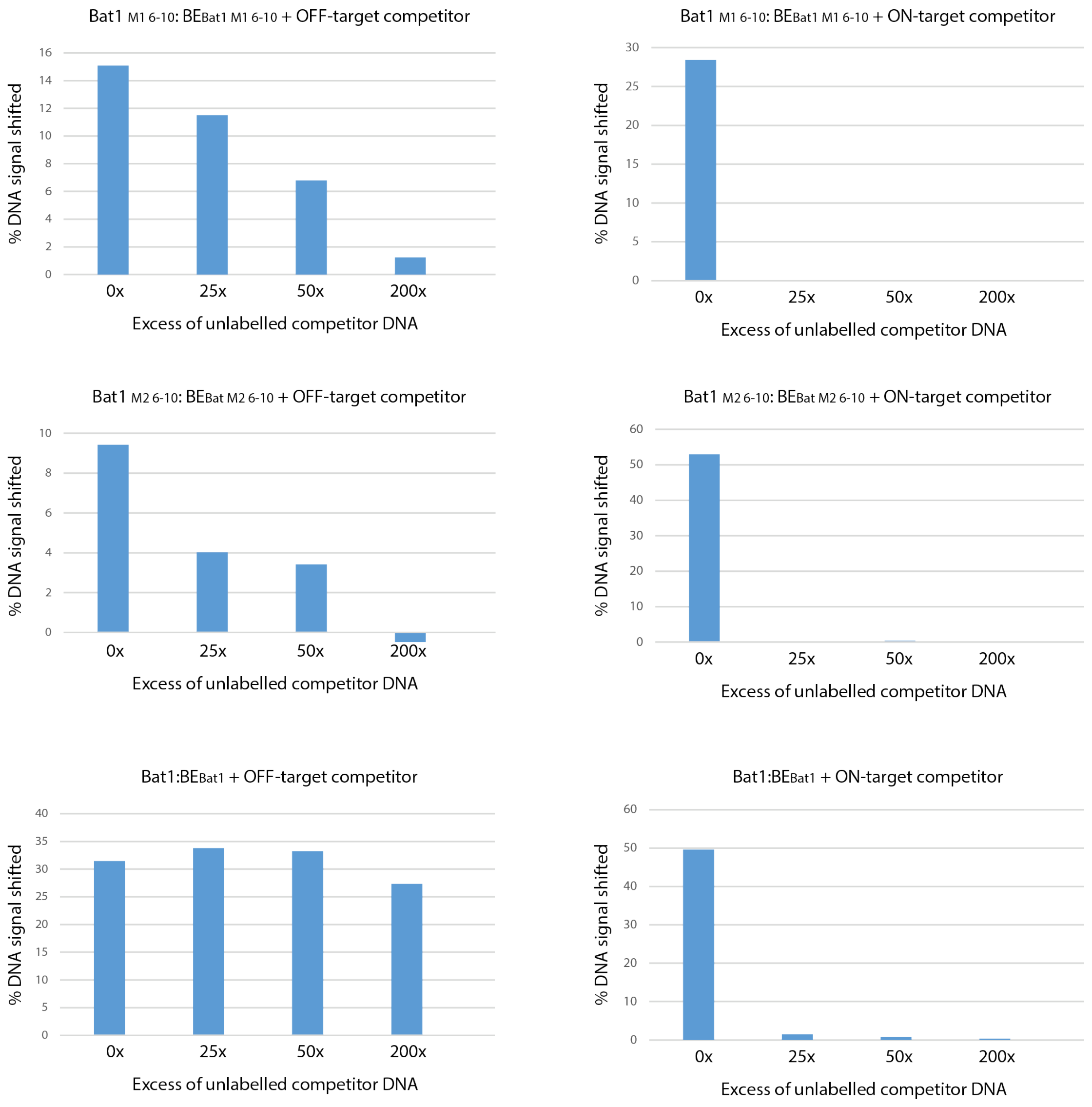
Supplementary Figure 3: (a) Protein expression gel for MOrTL1 and MOrTL2 and (b) EMSA gel for MOrTL1 against BE_{MOrTL1}.

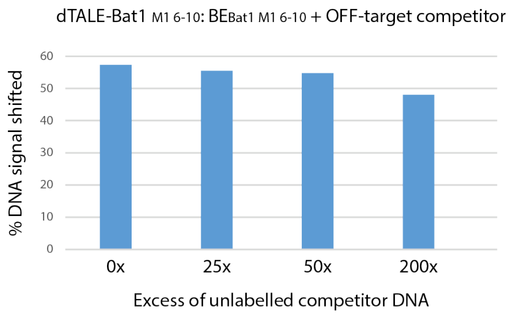


Supplementary Figure 4: Amino acid sequence of fusion protein EBN91408-MOrTL2. The two ORFs of genomic accession *EN814823.1* (see Figure S1) are separated by a frame-shift in the middle of MOrTL1 repeat 1. Removal of a single guanine base allows read through of a longer protein designated EBN91408-MOrTL2. Although, as noted in Figure S1, the true genomic locus likely contains further intervening repeats not covered in the assembly. EBN91408 is underlined. Repeats are numbered and 0 and -1 are uses to designate the sequence degenerate N-terminal repeats.

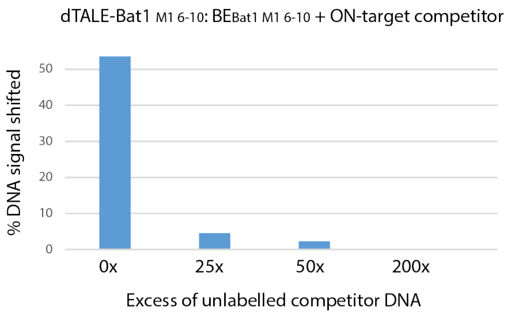
```
> EBN91408-MOrTL2.
   MSNQTEQKILKFKLELRYPTESAQLIRAG
-1 FNRDQADRIILRGSSQRTVAKLLEIHKTLLAHPYR
0  ITYDDLTRIAARNGGSKNLVAVQANYAALTELG
1  FSAKDIVQMVSHGGGSKNLEVVQANYAALTGLG
2  FRTEDIVQMVSHDGGSKNLAAMIDKSTALKDLG
3  FRTEDIVQMVSHDGSSKNLAAMIDKSTALKGLG
4  FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
5  FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
6  FRTEDIVQMVSHDGGSKNLVAVQANYAALTGLG
7  FRTEDIVQMVSHDGGSKNLAAIIDKSTALTGLG
8  FRTEDIVQMVSNNGGSKNLAAIIDKSTALKGLG
9  FRTEDIVQMVSHGGGSKNLEVVQANYAALTGLG
10 FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
11 FRTEDIVQMVSHDGGSKNLAAMIDKYTALKDLG
12 FRTEDIVQMVSHDGGSKNLAAIIDKSTALKGLG
13 FLTEDIVQMVSHDGGSKNLEVVQASYDTLTELKF
```

Supplementary Figure 5: Quantifications of protein : DNA relative to free-DNA in EMSAs shown in Figure 6.

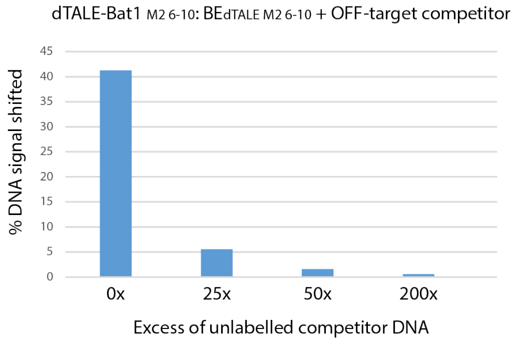




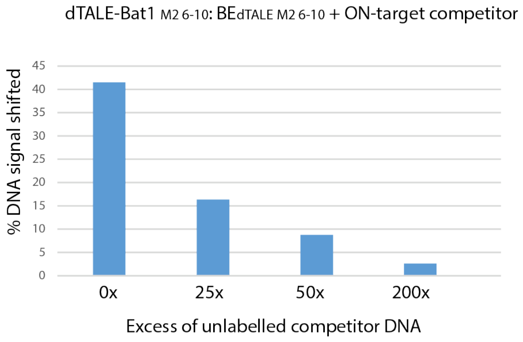
dTALE-Bat1$_{\text{M1 6-10}}$: BE$_{\text{Bat1 M1 6-10}}$ + OFF-target competitor



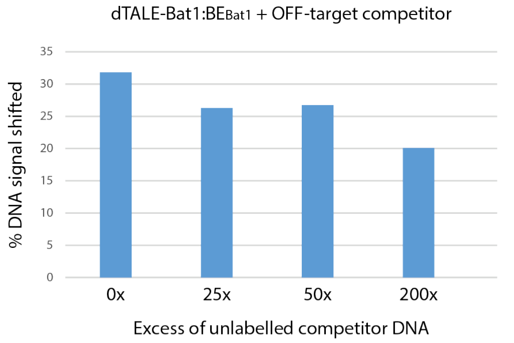
dTALE-Bat1$_{\text{M1 6-10}}$: BE$_{\text{Bat1 M1 6-10}}$ + ON-target competitor



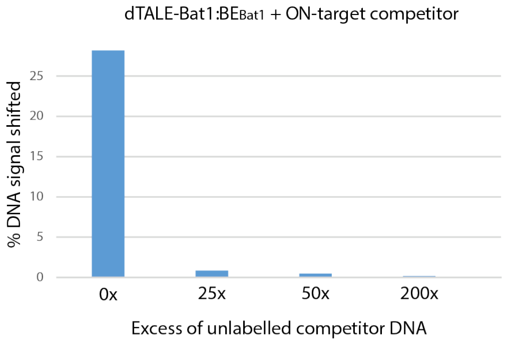
dTALE-Bat1$_{\text{M2 6-10}}$: BE$_{\text{dTALE M2 6-10}}$ + OFF-target competitor



dTALE-Bat1$_{\text{M2 6-10}}$: BE$_{\text{dTALE M2 6-10}}$ + ON-target competitor



dTALE-Bat1:BE$_{\text{Bat1}}$ + OFF-target competitor

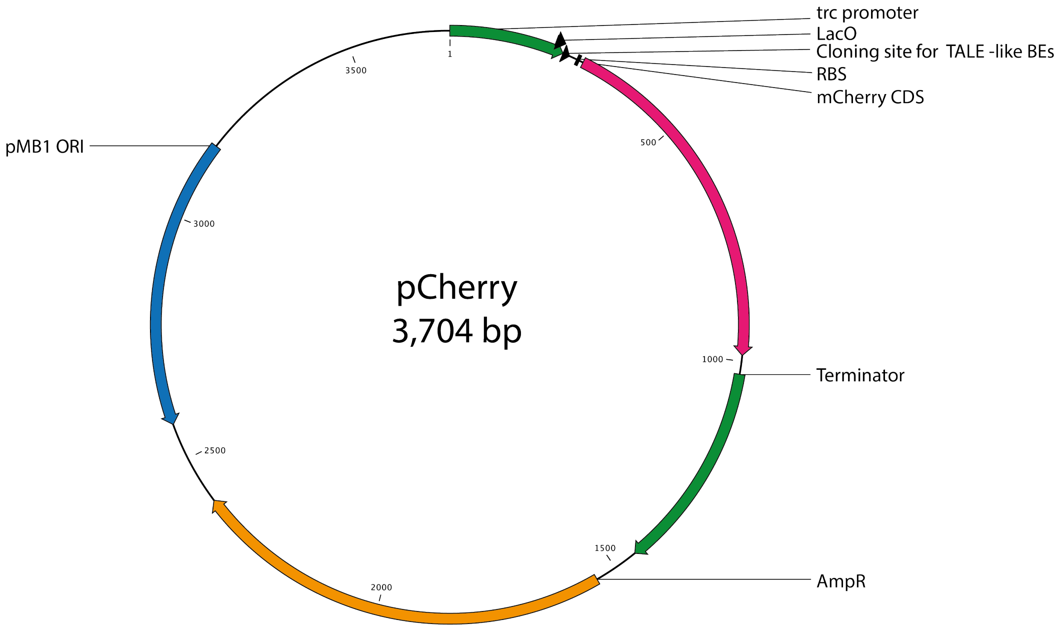


dTALE-Bat1:BE$_{\text{Bat1}}$ + ON-target competitor

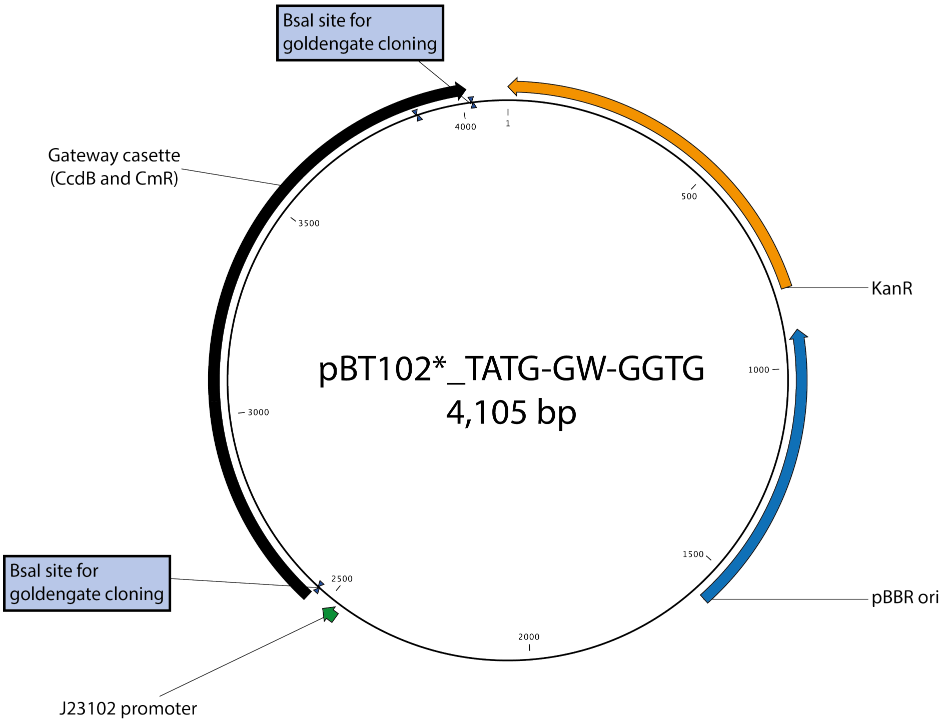
Supplementary Figure 6:

Maps of *E.coli* repressor reporter plasmids pMBS6 and pBT102*. TALE-like binding elements (BEs) are added to pCherry at the indicated position via PCR. TALE-like chimeras were added to pBT102* using BsaI cut-ligation. The pBT102 derivative with BsaI digest overlaps TATG (5') and GGTG (3') was used for the assembly of TALE chimeras. An additional derivative with overlaps CACC (5') and AAGG (3') but otherwise identical was created for the assembly of Bat1 chimeras. See materials and methods section for further details.

mCherry reporter plasmid



TALE-like expression plasmid


Bsal site for goldengate cloning
Gateway casette (CcdB and CmR)
KanR
pBT102*_TATG-GW-GGTG
4,105 bp
1
500
1000
1500
2000
2500
3000
3500
4000
pBBR ori
Bsal site for goldengate cloning
J23102 promoter

Supplementary Figure 7: Sequences of pCherry reporter constructs

pCherry – BE Bat1
mCherry CDS
Lac operator
Binding element Bat1

CGACTGCACGGTGCACCAATGCTTCTGGCGTCAGGCAGCCATCGGAAGCTGTGGTAT
GGCTGTGCAGGTCGTAAATCACTGCATAATTCGTGTCGCTCAAGGCGCACTCCCGTT
CTGGATAATGTTTTTTGCGCCGACATCATAACGGTTCTGGCAAATATTCTGAAATGA
GCTGTTGACAATTAATCATCCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACA
ATTTCTaagagaagcaaagacgttatGAATTCAAAAGATCTATCGATCGAGGATCCA
GGAGGTACAATCAATGGTGAGCAAGGGCGAGGAGGATAACATGGCCATCATCAAGGA
GTTCATGCGCTTCAAGGTGCACATGGAGGGCTCCGTGAACGGCCACGAGTTCGAGAT
CGAGGGCGAGGGCGAGGGCCGCCCCTACGAGGGCACCCAGACCGCCAAGCTGAAGGT
GACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTCCCCTCAGTTCATGTA
CGGCTCCAAGGCCTACGTGAAGCACCCCGCCGACATCCCCGACTACTTGAAGCTGTC
CTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAACTTCGAGGACGGCGGCGTGGT
GACCGTGACCCAGGACTCCTCCCTGCAGGACGGCGAGTTCATCTACAAGGTGAAGCT
GCGCGGCACCAACTTCCCCTCCGACGGCCCCGTAATGCAGAAGAAGACCATGGGCTG
GGAGGCCTCCTCCGAGCGGATGTACCCCGAGGACGGCGCCCTGAAGGGCGAGATCAA
GCAGAGGCTGAAGCTGAAGGACGGCGGCCACTACGACGCTGAGGTCAAGACCACCTA
CAAGGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTCAACATCAAGTTGGA
CATCACCTCCCACAACGAGGACTACACCATCGTGGAACAGTACGAACGCGCCGAGGG
CCGCCACTCCACCGGCGGCATGGACGAGCTGTACAAGTAA

Underlined sequence differs between the various reporters, with capital letters indicating sequences corresponding to MOrTL repeats

BE Bat1 M1 6-10
aagagAACGTaagacgttat

BE Bat1 M2 6-10
aagagTCCGTaagacgttat

BE dTALE-Bat1 M2 6-10
aagagCGTTCaagacgttat

BE Bat1 GGTTG
aagagGGTTGaagacgttat

BE Bat1 TTGGT
aagagTTGGTaagacgttat

<u>Supplementary Figure 8</u>: Sequences of (a) Bat1-MOrTL and (b) TALE-MOrTL reporter constructs

(a)
Bat1 constructs expressed form pDEST-17 are preceded by an N-terminal His-Tag of sequence: `MSYYHHHHHHLESTSLYKKAGSAAAPFM`

>Bat1
NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
1 FSQSDIVKIAGN**I**GGAQALQAVLDLESMLGKRG
2 FSRDDIAKMAGN**I**GGAQTLQAVLDLESAFRERG
3 FSQADIVKIAGN**N**GGAQALYSVLDVEPTLGKRG
4 FSRADIVKIAGN**T**GGAQALHTVLDLEPALGKRG
5 FSRIDIVKIAAN**N**GGAQALHAVLDLGPTLRECG
6 FSQATIAKIAGN**I**GGAQALQMVLDLGPALGKRG
7 FSQATIAKIAGN**I**GGAQALQTVLDLEPALCERG
8 FSQATIAKMAGN**N**GGAQALQTVLDLEPALRKRD
9 FRQADIIKIAGN**D**GGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAGN**I**GGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAGN**I**GGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGN**T**GGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGN**R**GGAQALQAVLALELTLRERG
14 FSQPDIVKIAGN**S**GGAQALQAVLDLELTFRERG
15 FSQADIVKIAGN**D**GGTQALHAVLDLERMLGERG
16 FSRADIVNVAGN**N**GGAQALKAVLEHEATLNERG
17 FSRADIVKIAGN**G**GGAQALKAVLEHEATLDERG
18 FSRADIVRIAGN**G**GGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAAN**S**GGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ

In PBT102 (extra C-terminal residues from cloning vector):
… GGTLIIPDLHSRKSKTSDRRLLT

>Bat1$_{\text{M1 6-10}}$
NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
1 FSQSDIVKIAGN**I**GGAQALQAVLDLESMLGKRG
2 FSRDDIAKMAGN**I**GGAQTLQAVLDLESAFRERG
3 FSQADIVKIAGN**N**GGAQALYSVLDVEPTLGKRG
4 FSRADIVKIAGN**T**GGAQALHTVLDLEPALGKRG
5 FSRIDIVKIAAN**N**GGAQALHAVLDLGPTLRECG
6 LEPKDIVSIASH**I**GANQTITTLLNKWGALIDLE
7 LEPKDIVSIASH**I**GATQAITTLLNKWAALRAKG
8 LDPKDIVSIASH**D**GSNQTITKLLEKWDELRAKE
9 LESKDIVSIASN**N**GATQTITRLLEKWDELRAKG
10 LDPKDIVSIASH**G**GATQAITTLLNRWGDLIDLG

11 FSQPDIVKMAGN**I**GGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGN**T**GGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGN**R**GGAQALQAVLALELTLRERG
14 FSQPDIVKIAGN**S**GGAQALQAVLDLELTFRERG
15 FSQADIVKIAGN**D**GGTQALHAVLDLERMLGERG
16 FSRADIVNVAGN**N**GGAQALKAVLEHEATLNERG
17 FSRADIVKIAGN**G**GGAQALKAVLEHEATLDERG
18 FSRADIVRIAGN**G**GGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAAN**S**GGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ

In PBT102 (extra C-terminal residues from cloning vector)
… GGTLIIPDLHSRKSKTSDRRLLT

>Bat1$_{\mathrm{M2\ 6\text{-}10}}$
NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
1 FSQSDIVKIAGN**I**GGAQALQAVLDLESMLGKRG
2 FSRDDIAKMAGN**I**GGAQTLQAVLDLESAFRERG
3 FSQADIVKIAGN**N**GGAQALYSVLDVEPTLGKRG
4 FSRADIVKIAGN**T**GGAQALHTVLDLEPALGKRG
5 FSRIDIVKIAAN**N**GGAQALHAVLDLGPTLRECG
6 FRTEGIVQMVSH**G**GGSKNLVAVQANYAALTGLG
7 FRTEDIVQMVSH**D**GGSKNLVAVQANYAALTGLG
8 FRTEDIVQMVSH**D**GGSKNLAAIIDKSTALTGLG
9 FRTEDIVQMVSN**N**GGSKNLAAIIDKSTALKGLG
10 FRTEDIVQMVSH**G**GGSKNLEVVQANYAALTGLG
11 FSQPDIVKMAGN**I**GGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGN**T**GGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGN**R**GGAQALQAVLALELTLRERG
14 FSQPDIVKIAGN**S**GGAQALQAVLDLELTFRERG
15 FSQADIVKIAGN**D**GGTQALHAVLDLERMLGERG
16 FSRADIVNVAGN**N**GGAQALKAVLEHEATLNERG
17 FSRADIVKIAGN**G**GGAQALKAVLEHEATLDERG
18 FSRADIVRIAGN**G**GGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAAN**S**GGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ

in PBT102 (extra C-terminal residues from cloning vector)
… GGTLIIPDLHSRKSKTSDRRLLT

(b)
dTALE-Bat1 In pDEST-17 (E.coli protein expression and purification construct): Underlined sequences represent peptide tags, N-terminal HA and C-terminal 3xflag with flexible linker.
MSYYHHHHHHLESTSLYKKAGSAAAPF

>dTALE-Bat1 In pDEST-17
MDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVK
YQDMIAALPE
-1 ATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQ
0 LDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLN
1 LTPQQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
2 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
3 LTPEQVVAIASN**N**GGKQALETVQRLLPVLCQAHG
4 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
5 LTPQQVVAIASN**N**GGKQALETVQRLLPVLCQAHG
6 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
7 LTPQQVVAIASN**I**GGKQALETVQALLPVLCQAHG
8 LTPEQVVAIASN**N**GGKQALETVQALLPVLCQAHG
9 LTPEQVVAIASH**D**GGKQALETVQRLLPVLCQAHG
10 LTPQQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
11 LTPQQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
12 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIASN**N**GGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
15 LTPQQVVAIASH**D**GGKQALETVQRLLPVLCQAHG
16 LTPEQVVAIASN**N**GGKQALETVQRLLPVLCQAHG
17 LTPQQVVAIASN**G**GGKQALETVQALLPVLCQAHG
+1 LTPQQVVAIASNGGGRPALESIVAQLSRPDPALAA
+2 LTNDHLVALACLGGRPALDAVKKGLPHAPALIKRT
NRRIPERTSHRVA
GGGGGGSGGGGSGGGGSDYKDHDGDYKDHDIDYKDDDDKGSSPKKKRKVEAS

In pBT102 (*E.coli* expression, repressor assay construct):

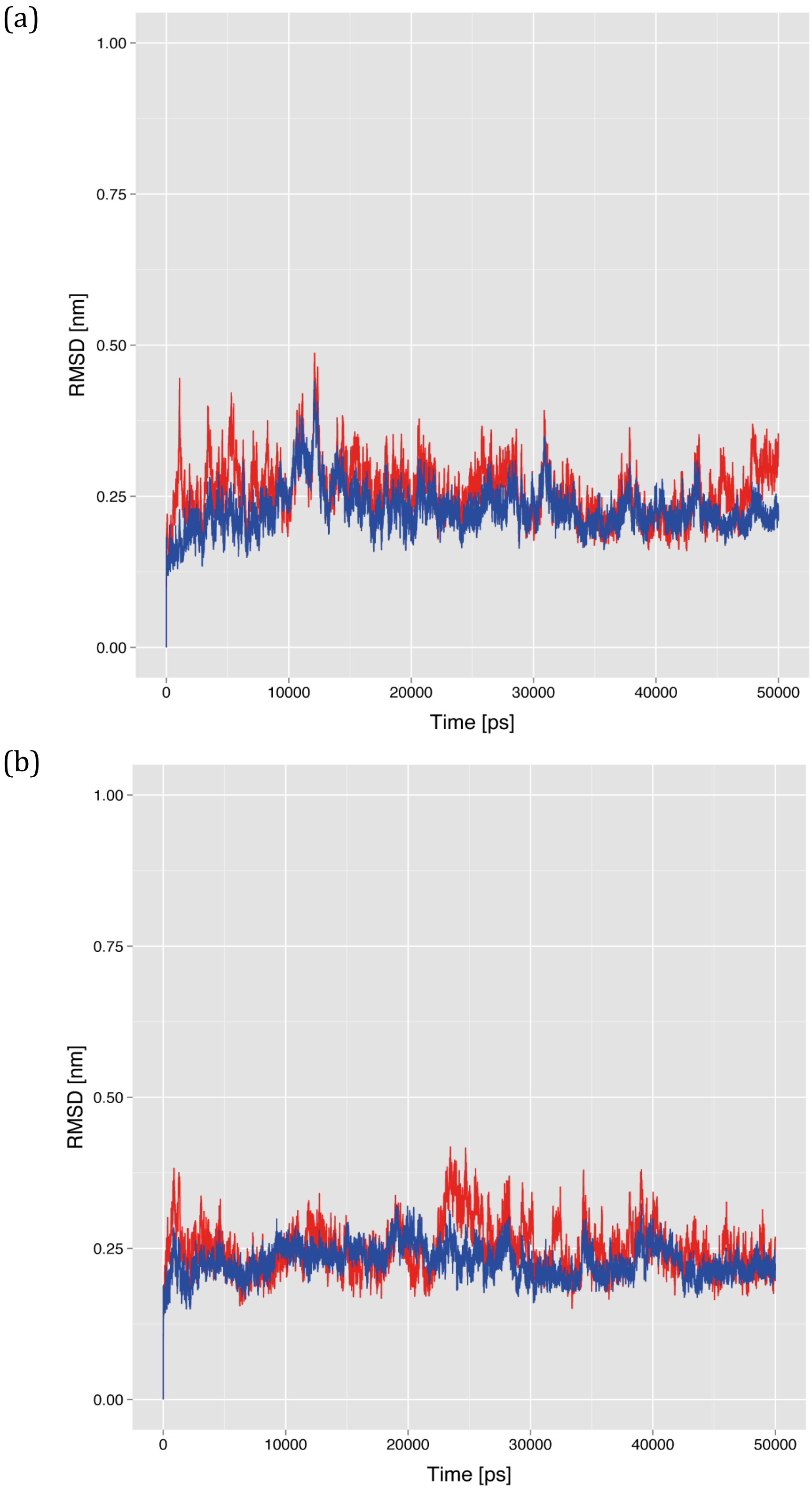
These constructs lack the N-terminal HA tag but otherwise are identical from the N-terminus until after the C-terminal degenerate repeats: the TALE-C-terminal section is longer and there is a C-terminal GFP.
…
+1 LTPQQVVAIASNGGGRPALESIVAQLSRPDPALAA
+2 LTNDHLVALACLGGRPALDAVKKGLPHAPALIKRT
NRRIPERTSHRVADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGMS
GSVSKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATIGKLTLKFICTTGKLPV
PWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTISFKDDGKYKTRAVV
KFEGDTLVNRIELKGTDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFTVRH
NVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQTVLSKDPNEKRDHMVLHEYVNA
AGIT

>dTALE-Bat1$_{M1\ 6\text{-}10}$

1 LTPQQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
2 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
3 LTPEQVVAIASN**N**GGKQALETVQRLLPVLCQAHG
4 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
5 LTPQQVVAFASN**N**GGKQALTKLLEKWDELRAKG
6 LEPKDIVSIASH**I**GANQTITTLLNKWGALIDLE
7 LEPKDIVSIASH**I**GATQAITTLLNKWAALRAKG
8 LDPKDIVSIASH**D**GSNQTITKLLEKWDELRAKE
9 LESKDIVSIASN**N**GATQTITRLLEKWDELRAKG
10 LDPKDIVSIASH**G**GATQAITTLLNRWGDLIDLE
11 LEPKDIVAIASN**I**GGKQALETVQRLLPVLCQAHG
12 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIASN**N**GGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
15 LTPQQVVAIASH**D**GGKQALETVQRLLPVLCQAHG
16 LTPEQVVAIASN**N**GGKQALETVQRLLPVLCQAHG
17 LTPQQVVAIASN**G**GGKQALETVQALLPVLCQAHG

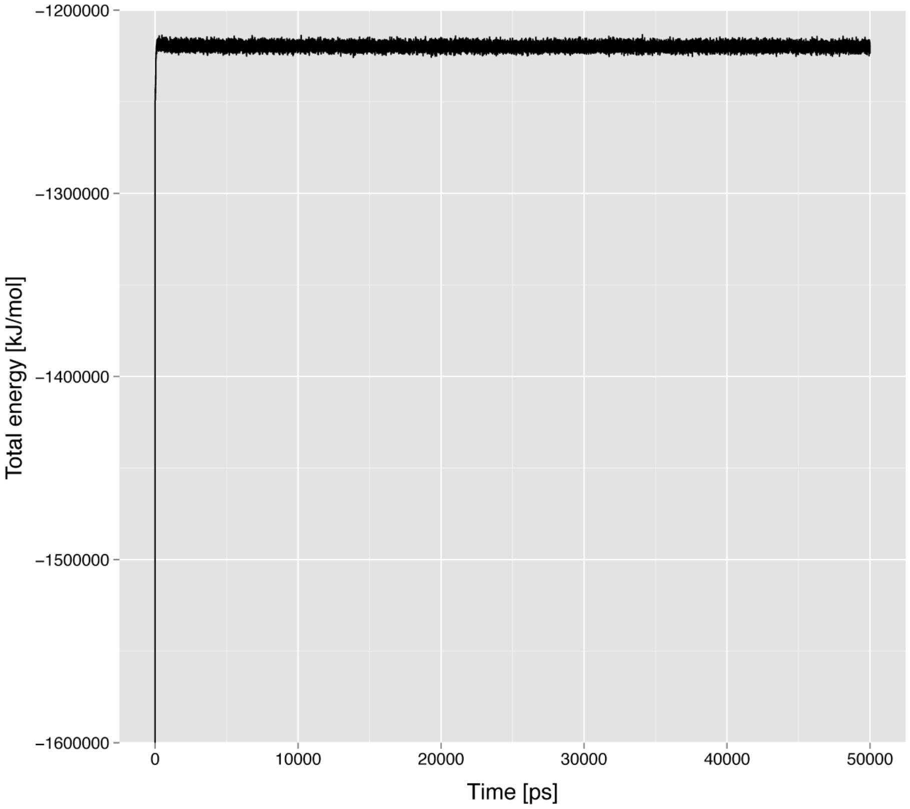
>dTALE-Bat1$_{M2\ 6\text{-}10}$
…
1 LTPQQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
2 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
3 LTPEQVVAIASN**N**GGKQALETVQRLLPVLCQAHG
4 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
5 LTPQQVVAIASN**N**GGKQALVAVQANYAALTGLG
6 FRTEDIVQMVSH**D**GGSKNLAAIIDKSTALTGLG
7 FRTEDIVQMVSN**N**GGSKNLAAIIDKSTALKGLG
8 FRTEDIVQMVSH**G**GGSKNLEVVQANYAALTGLG
9 FRTEGIVQMVSH**G**GGSKNLVAVQANYAALTGLG
10 FRTEDIVQMVSH**D**GGSKNLAAMIDKYTALKDLG
11 FRTEDIVAIASN**I**GGKQALETVQRLLPVLCQAHG
12 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIASN**N**GGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASN**I**GGKQALETVQRLLPVLCQAHG
15 LTPQQVVAIASH**D**GGKQALETVQRLLPVLCQAHG
16 LTPEQVVAIASN**N**GGKQALETVQRLLPVLCQAHG
17 LTPQQVVAIASN**G**GGKQALETVQALLPVLCQAHG
…

<u>Supplementary Figure 9:</u> MD Analyses (Root Mean Square Deviation of atomic positions) for models of Bat1-MOrTL chimeras with their target DNA over 50 ns. RMSD plots are shown for protein and DNA for Bat1$_{M2\ 6\text{-}10}$ (a) and Bat1$_{M1\ 6\text{-}10}$ (b). The plots show the protein C-alpha RMSDs after least squares fit to the protein C-alpha (red lines) as well as the DNA RMSDs after least squares fit to DNA (blue lines). That both protein and DNA in parallel over the course of the simulations indicates a stable interaction.

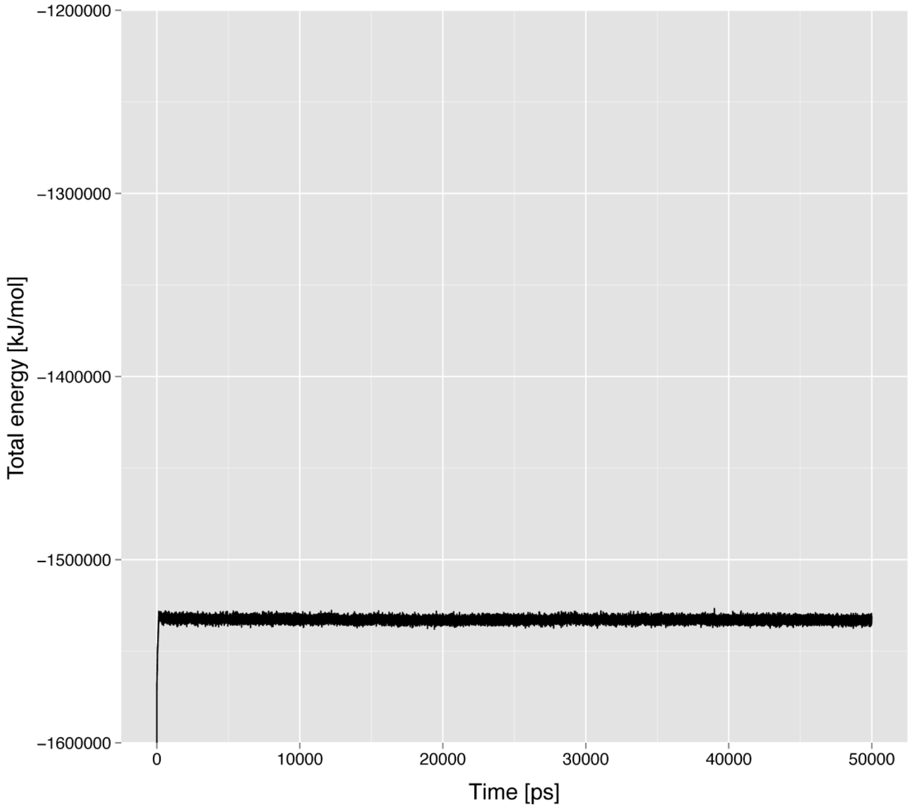


Supplementary Figure 10: Molecular Dynamics Analysis (energy):
Potential energy of Bat1$_{M2\ 6\text{-}10}$ (a) and Bat1$_{M1\ 6\text{-}10}$ (b) bound to their DNA targets, over a period of 50 nanoseconds.

(a)

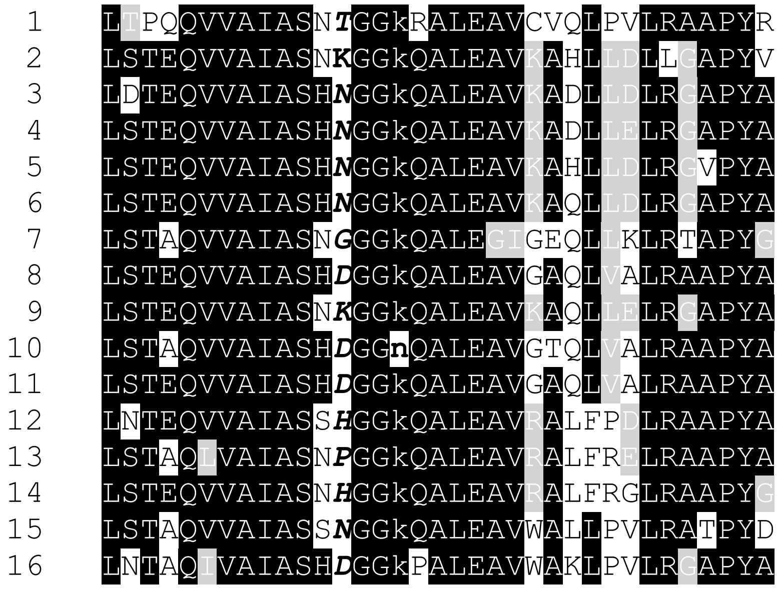


(b)



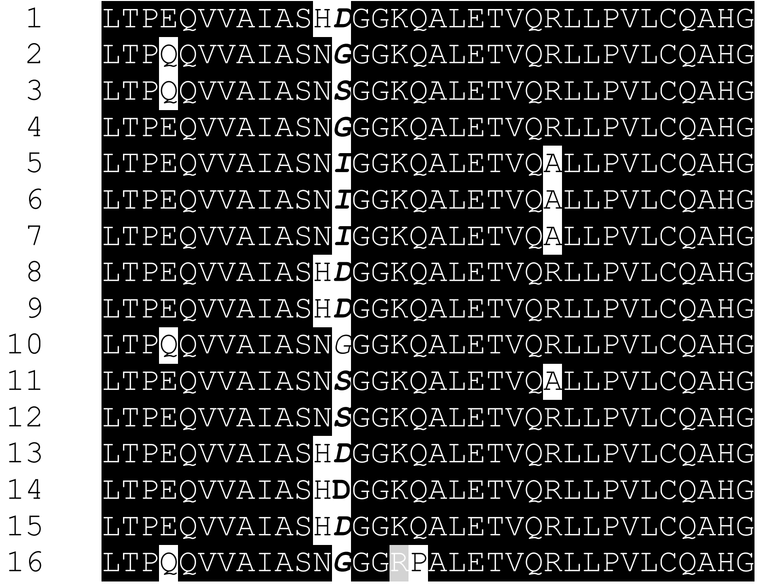
<u>Supplementary Figure 11:</u> Core repeat alignments of a representative RipTAL (Brg11) and TALE (AvrBs3). Alignments were constructed with CLC Main Workbench and images generated with Boxshade. Conserved residues are shown as white letters on a black background. The BSRs are highlighted in bold-italic font.

>RipTAL (Brg11)



1 LTPQQVVAIASN***T***GGkRALEAVCVQLPVLRAAPYR
2 LSTEQVVAIASN**K**GGkQALEAVKAHLLDLLGAPYV
3 LDTEQVVAIASH***N***GGkQALEAVKADLLDLRGAPYA
4 LSTEQVVAIASH***N***GGkQALEAVKADLLELRGAPYA
5 LSTEQVVAIASH***N***GGkQALEAVKAHLLDLRGVPYA
6 LSTEQVVAIASH***N***GGkQALEAVKAQLLDLRGAPYA
7 LSTAQVVAIASN***G***GGkQALEGIGEQLLKLRTAPYG
8 LSTEQVVAIASH***D***GGkQALEAVGAQLVALRAAPYA
9 LSTEQVVAIASN***K***GGkQALEAVKAQLLELRGAPYA
10 LSTAQVVAIASH***D***GG**n**QALEAVGTQLVALRAAPYA
11 LSTEQVVAIASH***D***GGkQALEAVGAQLVALRAAPYA
12 LNTEQVVAIASS***H***GGkQALEAVRALFPDLRAAPYA
13 LSTAQLVAIASN***P***GGkQALEAVRALFRELRAAPYA
14 LSTEQVVAIASN***H***GGkQALEAVRALFRGLRAAPYG
15 LSTAQVVAIASS***N***GGkQALEAVWALLPVLRATPYD
16 LNTAQIVAIASH***D***GGkPALEAVWAKLPVLRGAPYA

>TALE (AvrBs3)



1 LTPEQVVAIASH***D***GGKQALETVQRLLPVLCQAHG
2 LTPQQVVAIASN***G***GGKQALETVQRLLPVLCQAHG
3 LTPQQVVAIASN***S***GGKQALETVQRLLPVLCQAHG
4 LTPEQVVAIASN***G***GGKQALETVQRLLPVLCQAHG
5 LTPEQVVAIASN***I***GGKQALETVQALLPVLCQAHG
6 LTPEQVVAIASN***I***GGKQALETVQALLPVLCQAHG
7 LTPEQVVAIASN***I***GGKQALETVQALLPVLCQAHG
8 LTPEQVVAIASH***D***GGKQALETVQRLLPVLCQAHG
9 LTPEQVVAIASH***D***GGKQALETVQRLLPVLCQAHG
10 LTPQQVVAIASN*G*GGKQALETVQRLLPVLCQAHG
11 LTPEQVVAIASN***S***GGKQALETVQALLPVLCQAHG
12 LTPEQVVAIASN***S***GGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIASH***D***GGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASH**D**GGKQALETVQRLLPVLCQAHG
15 LTPEQVVAIASH***D***GGKQALETVQRLLPVLCQAHG
16 LTPQQVVAIASN***G***GGRPALETVQRLLPVLCQAHG

Supplementary Table 1: The TALE code

A reference guide for the specificities of commonly occurring BSRs, based on several publications:
1: Yang et al., 2014
2: de Lange et al., 2014
3: Boch et al., 2009
4: Cong et al., 2012
Boch et al., 2009; de Lange et al., 2013; Mak, Bradley, & *Cernadas*, 2012; Meckler et al., 2013; Moscou & Bogdanove, 2009; Yang et al., 2014).

BSR	Best-match	2nd best	Tolerated	Mismatch
Gly	T	-	A, G, C $_{(2, 4)}$	-
Gly$_{SL}$	C, C^{me}	-	T, A, G $_{(3, 4)}$	-
Asp	C	-	A$_{(3,4)}$	G, T
Ile	A	-	-	G, C, T
Ser	A, G, C		T $_{(1)}$	
Arg	G	A	-	C, T
His	G	-	A, C $_{(2)}$	C, T
Lys	G	-		A, C, T

Note: since specificity is only a measure of relative affinity the absolute affinities for BSRs to their best-match or mismatch bases can vary greatly. See Meckler et al., Nucl. Acids Res., 2013 for more detail on this.

Supplementary Table 2: A list of oligonucleotides used in this study

Primer name	**Sequence**	**Notes**
EMSA probes		
BE$_{\text{MOrTL1}}$ EMSA Fwd	`TAGCACAACGTGCTGAC`	EMSA Probe for non-chimeric MOrTL1 protein
BE$_{\text{MOrTL1}}$ EMSA Rev	`GTCAGCACGTTGTGCTA`	EMSA Probe for non-chimeric MOrTL1 protein
BE$_{\text{MOrTL2}}$ EMSA Fwd	`TAGCTTCCGTTCCCTGAC`	EMSA Probe for non-chimeric MOrTL2 protein
BE$_{\text{MOrTL2}}$ EMSA Rev	`GTCAGGGAACGGAAGCTA`	EMSA Probe for non-chimeric MOrTL2 protein
BEBat1 M1 6-10 EMSA Fwd	`TAGACTAAGAGAACGTAAGACGTTATATGC`	EMSA probe for Bat$_{\text{M1 6-10}}$ and dTALE-Bat1$_{\text{M1 6-10}}$
BEBat1 M1 6-10 EMSA Rev	`GCATATAACGTCTTACGTTCTCTTAGTCTA`	EMSA probe for Bat$_{\text{M1 6-10}}$ and dTALE-Bat1$_{\text{M1 6-10}}$
BEBat1 M2 6-10 EMSA Fwd	`TAGACTAAGAGTCCGTAAGACGTTATATGC`	EMSA probe for Bat$_{\text{M2 6-10}}$
BEBat1 M2 6-10 EMSA Rev	`GCATATAACGTCTTACGGACTCTTAGTCTA`	EMSA probe for Bat$_{\text{M2 6-10}}$
BEdTALE-Bat1 M2 6-10 EMSA	`TAGACTAAGAGCGTTCAAGACGTTATATGC`	EMSA probe for dTALE- Bat$_{\text{M2 6-10}}$

Fwd		
BEdTALE-Bat1 M2 6-10 EMSA Rev	GCATATAACGTCTTGAACGCTCTTAGTCTA	EMSA probe for dTALE- Bat$_{M2\ 6\text{-}10}$
BEBat1 GGTTG EMSA Fwd	TAGACTAAGAGGGTTGAAGACGTTATATGC	OFF-target EMSA probe for all except dTALE-Bat1$_{M2\ 6\text{-}10}$
BEBat1 GGTTG EMSA REv	GCATATAACGTCTTCAACCCTCTTAGTCTA	OFF-target EMSA probe for all except dTALE-Bat1$_{M2\ 6\text{-}10}$
BEBat1 TTGGT EMSA Fwd	TAGACTAAGAGTTGGTAAGACGTTATATGC	OFF-target EMSA probe dTALE-Bat1$_{M2\ 6\text{-}10}$
BEBat1 TTGGT EMSA Rev	GCATATAACGTCTTAACCACTCTTAGTCTA	OFF-target EMSA probe dTALE-Bat1$_{M2\ 6\text{-}10}$
Primers for PCR mutagenesis		
MOrTL1 Bat1 Block2 Mimic Fwd	GGTCTCTTGGGCTGGAACCGAAAGATATCGTG	To create MOrTL repeat blocks to insert into Bat1
MOrTL1 Bat1 Block2 Mimic Rev	GGTCTCAAACCCAGGTCGATCAGATCGCCCC	To create MOrTL repeat blocks to insert into Bat1
MOrTL2 Bat1 Block2 Mimic Fwd	GGTCTCTTGGGTTTCGTACCGAAGGCATTGTCCA	To create MOrTL repeat blocks to insert into Bat1
MOrTL2 Bat1 Block2 Mimic Rev	GGTCTCAAACCCAGACCCGTCAGGGCGGCGTAG	To create MOrTL repeat blocks to insert into Bat1
MOrTL1 dTALE 5B mimic Fwd	GAAGACTCTCTGACGAAACTGCTGGAAAAATG	To create MOrTL repeat blocks to insert into dTALE-Bat1
MOrTL1 dTALE 5B mimic Rev	GAAGACTCCGCTACAATGTCTTTAGGTTCTAATTC	To create MOrTL repeat blocks to insert into dTALE-Bat1
MOrTL2 dTALE 5B mimic Fwd	GAAGACTCTCTGGTTGCAGTCCAAGCTAATTACGC	To create MOrTL repeat blocks to insert into dTALE-Bat1
MOrTL2dTALE 5B mimic Rev	GAAGACTCCGCTACGATATCTTCCGTGCGGAAAC	To create MOrTL repeat blocks to insert into dTALE-Bat1
GFP-VS-Fwd	ATGGTGTCTAAGGGCGAAGAACTC	To create a GFP only pBT102 expression plasmid
TATG_BsaI_Rev	AAGAGACCCCTGCATGCAAGC	To create a GFP only pBT102 expression plasmid
pMBS6 BE$_{Bat1\ MX\ 6\text{-}10}$ Fwd	AAGACGTTATGAATTCAAAAGATCTATCGA	To get BE$_{Bat1\ M1\ or\ M2\ 6\text{-}10}$ into pMBS6
pMBS6 BE$_{Bat1\ M1\ 6\text{-}10}$ Rev	ACGTTCTCTTAGAAATTGTTACCGCTC	To get BE$_{Bat1\ M1\ 6\text{-}10}$ into pMBS6
pMBS6 BE$_{Bat1\ M2\ 6\text{-}10}$ Rev	ACGGACTCTTAGAAATTGTTATCCGCTC	To get BE$_{Bat1\ M2\ 6\text{-}10}$ into pMBS6
AvrBs3DeltaCTD Rev	GCTCATCCCGAACTGCGTCA	To create C-terminal TALE truncation derivate to match Politz *et al.* LacO dTALE
AvrBs3DeltaCTD Fwd	AAGGTGAGACCTTTGGGATCCGA	To create C-terminal TALE truncation derivate to match Politz *et al.* LacO dTALE

pMBS6 $BE_{dTALE\text{-}Bat1\ M2\ 6\text{-}10}$ Rev	GAACGCTCTTGAAATTG TTATCCGCTC	To get $BE_{dTALE\text{-}Bat1\ M2\ 6\text{-}10}$ into pMBS6
BsaI AAGG Rev	CCT TTG AGA CCG GTC GAC CTG C	To create a goldengate version of E.coli expression vector pBT102
BsaI GGTG Rev	CAC CTG AGA CCG GTC GAC CTG	To create a goldengate version of E.coli expression vector pBT102
BsaI TATG Fwd	TAT GTG AGA CCG CGG CCC CTC	To create a goldengate version of E.coli expression vector pBT102
Sequencing primers		
Sco5B MidSeqF	TATCGATAAAAGCACCG CCC	To sequence central section of pDEST17 dTALE-Bat1$M_{2\ 6\text{-}10}$
Sco5B MidSeqR	ACCGTGACTGACCATTT GGA	To sequence central section of pDEST17 dTALE-Bat1$M_{2\ 6\text{-}10}$

<u>Supplementary Table 3:</u> Averaged base-BSR distances from MD model of Bat1 M1 6-10. Average distances over all MD snapshots between BSR Cα-atom and the ring nitrogen that connects nucleobase and deoxyribose moieties. MOrTL pairs are highlighted in green.

BSR		*Nucleotide*		*Average distance (nm)*	*SD*
ILE	95	DA	1	0,742	0,044
ILE	128	DA	2	0,719	0,022
ASN	161	DG	3	0,666	0,026
THR	194	DA	4	0,706	0,027
ASN	227	DG	5	0,669	0,029
ILE	260	DA	6	0,753	0,040
ILE	293	DA	7	0,830	0,075
ASP	326	DC	8	0,853	0,088
ASN	359	DG	9	0,888	0,050
GLY	392	DT	10	0,815	0,044
ILE	425	DA	11	0,742	0,029
THR	458	DA	12	0,708	0,042
ARG	491	DT	13	0,854	0,047
SER	524	DA	14	0,696	0,043
ASP	557	DC	15	0,709	0,026
ASN	590	DG	16	0,671	0,025
GLY	623	DT	17	0,754	0,026
GLY	656	DT	18	0,782	0,030
SER	689	DA	19	0,661	0,032

Supplementary Table 4: Averaged base-BSR distances from MD model of Bat1$_{\mathrm{M2}\ 6\text{-}10}$. Average distances over all MD snapshots between BSR Cα-atom and the ring nitrogen that connects nucleobase and deoxyribose moieties. MOrTL pairs are highlighted in green.

BSR		*Nucleotide*		*Average distance (nm)*	*SD*
ILE	95	DA	1	0,728	0,028
ILE	128	DA	2	0,707	0,026
ASN	161	DG	3	0,674	0,029
THR	194	DA	4	0,716	0,028
ASN	227	DG	5	0,642	0,021
GLY	260	DT	6	0,738	0,030
ASP	293	DC	7	0,739	0,047
ASP	326	DC	8	0,763	0,068
ASN	359	DG	9	0,816	0,100
GLY	392	DT	10	0,815	0,056
ILE	425	DA	11	0,767	0,044
THR	458	DA	12	0,708	0,042
ARG	491	DT	13	0,869	0,051
SER	524	DA	14	0,715	0,042
ASP	557	DC	15	0,698	0,021
ASN	590	DG	16	0,685	0,033
GLY	623	DT	17	0,739	0,027
GLY	656	DT	18	0,773	0,035
SER	689	DA	19	0,661	0,022

Supplementary Table 5: TALE-likes used in the creation of repeat sequence logos shown in Figure 7. GenBank designations are given where relevant to avoid ambiguity.

TALEs	AvrBs3, AvrBs4, AvrXa27, AvrXa7, PthXo1 (ACD58243), PthB (NP_942641), AvrHah1, Hax2, Hax3, Hax4 TalC (AEK86668),AvrPth3, PthA (AAC43587)
RipTALs	Brg11, CCA82456,CAQ18687
Bats	BAT1_BURRH, Bat2 (E5AW45), Bat3(E5AW43), RipTALI_14, YP_003750492
MOrTL1	EBN91408, EBN91409, ECG96325, ECG96326

10 Acknowledgements

Turning up in Munich for an interview with Prof. Dr. Martin Parniske and finding myself instead opposite the newly minted Prof. Dr. Thomas Lahaye turned out to be the start of something wonderful. Thomas, thank you for taking a chance on me.

Despite not ending up in his group I would like to thank Prof. Parniske for numerous exciting and very useful discussion sessions during the first years of my PhD. I would also like to thank Dr. Ralf Kühn, who has also provided valuable suggestions and stimulating discussion over the course of my PhD as part of my thesis advisory committee. In addition I would like to thank Prof. Dr. Heinrich Leonhardt and Dr. Andreas Brachmann for advice, assistance and inspiration.

Within the Lahaye group I would like to give particular mention to the endlessly talented Christina Wolf with whom I worked very closely on the Bat and MOrTL characterizations, and Niklas Schandry for constant advice and entertainment.

I would like to thank the student assistants who provided me with both practical help and pleasant company during my PhD: Stephanie Solle, Annette Zehrer and Max Mattheuer. I was also very privileged to supervise the projects of a number of wonderful and talented people, who taught me more than they might have realized. Jara Radeck, Jörn Dietze, Isabel Blunck, and Dousheng Wu thank you.

Thank you to Barbara Dollrieß and Charley Rehm, I very literally could not have done it without your help.

Finally, I would like to thank my parents for encouraging me to pursue my goals, even if it takes me far from home; and Caro, Gloria, Jeffrey, Christina, Isra and for helping me make a new one in Munich.

11 Curriculum vitae

Orlando de Lange

Date of birth: 09.08.1988
Place of Birth: London, United Kingdom

Education

09/2000 - 07/2006 Alleyn's School, London
Final year exams taken in Biology, Chemistry, Mathematics and Latin.

10/2007 - 06/2010 BA Natural Sciences, The University of Cambridge
Bachelour's project undertaken in the group of Prof. David Balcoumbe: The role of 24 nucleotide siRNAs in regulation of protein coding genes in model plant *Arabidopsis thaliana*.

10/2010 – 07/2015 Doctoral studies at the Ludwig-Maximillians University
Sequence diversity and functional conformity: A comparative molecular characterization of TALE-like proteins.

Additional scientific work

03-05/2006 Curatorial assistance in the cryptogamic herbarium of the natural history museum, United Kingdom.

07 – 09/2009 Internship at the Millenium Seed Bank, United Kingdom.

International conference attendances

12/2011 VIPCA Gene Targeting; Vienna, Austria.

09/2013 Effectome Meeting VI; Lauret, France. Oral presentation.

09/2013 SFB924: Plant Biology of the Next Generation; Freising, Germany.

07/ 2014 XVI International Congress on Molecular Plant-Microbe Interactions; Rhodes, Greece. Poster presentation.

06/2015 Enabling Technologies for Eukaryotic Synthetic Biology; Heidelberg, Germany. Poster presentation.