Biochemical characterization of modular DNA-binding domains of novel TALE-like proteins

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl. Biol. Christina Wolf
aus Augsburg

Tübingen
2015

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	03.03.2016
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Thomas Lahaye
2. Berichterstatterin:	Prof. Dr. Ulrike Zentgraf

Dedico questa tesi alla mia cara nonna, che purtroppo non c’è più. Lei era una persona, che mi ha insegnato a guardare avanti, a credere in ció che facevo e a non arrendermi mai. Grazie.

Contents

Abbreviations

AD	activation domain
Bat	*Burkholderia* TALE-like
BE	binding element
bp	base pair
BSR	base specifying residue
CRISPR	clustered regulary interspaced short palindromic repeats
crRNA	CRISPR RNA
CTR	C-terminal region
DNA	desoxyribonucleic acid
dTALE	designer TALE
MOrTL	Marine organism TALE-like
NLS	nuclear localization signal
NTR	N-terminal region
PAM	protospacer adjacent motif
PDP	programmable DNA binding protein
pv	pathovar
RipTAL	*Ralstonia* injected protein transcription activator like
RVD	repeat variable diresidue
sgRNA	single guided RNA
T3SS	type III secretion system
TALE	transcription activator like effector
TALEN	TALE nuclease
tracrRNA	trans-activating CRISPR RNA
ZF	zinc finger
ZFN	zinc finger nuclease

Zusammenfassung

TALEs (Transcription Activator Like Effectors) sind Proteine, die von bakteriellen Pflanzenpathogenen der Gattung *Xanthomonas* über ein Typ III Sekretionssystem in Pflanzenzellen injiziert werden. In den Pflanzenzellen aktivieren TALEs Wirtsgene deren Expression das Wachstum oder die Verbreitung des Bakteriums begünstigen. Die TALE DNA-Bindedomäne ist aus 10-30 nahezu identischen, 33-35 Aminosäure-langen Modulen (‚Repeats') aufgebaut. Jeder Repeat bindet ein Nukleotid, wobei Aminosäure 13 die Basenspezifität definiert. Die bekannten Basen-paarungspräferenzen der verschiedenen Aminosäuren in Repeat Position 13 - der sogenannte TALE Code – ermöglicht es TALE Proteine mit gewünschter DNA-Bindungsspezifität zu generieren oder die DNA Zielsequenzen von TALEs vorherzusagen.

In der hier vorliegenden Arbeit und den dazugehörigen Fachartikeln werden zwei neue Klassen TALE-ähnlicher Proteine charakterisiert. Zum einen die Bats (*Burkholderia* TALE-likes), des Bakteriums *Burkholderia rhizoxinica. Z*um anderen die MOrTLs (marine organism TALE–likes), die in einer marinen Metagenomdatenbank identifiziert wurden. Es konnte gezeigt werden, dass die DNA Zielsequenzen der hier beschrieben TALE-ähnlichen Proteine mit dem TALE-Code vorhergesagt werden können. Detailstudien zeigten jedoch, dass die Repeats eines Vertreters der MOrTLs geringere Basenspezifitäten aufweisen als die der typischen TALEs. Zusätzlich unterscheiden sich MOrTLs und Bats durch ihre Affinität und höheren Proteinstabilität von TALEs. Durch funktionale Analysen von Proteinchimären konnte gezeigt werden, dass Repeats von TALEs und den TALE-ähnlichen Bat und MOrTL Proteinen untereinander kompatibel sind. Da Repeats von Bat und MOrTL Proteinen mit TALE Repeats funktional kompatibel sind, liefern diese neuen Proteinklassen auch wertvolles Rohmaterial für die Konstruktion von Chimären mit neuen Eigenschaften.

Summary

Members of the plant pathogenic bacterial genus *Xanthomonas* inject TALEs (Transcription Activator Like Effector) by a type III secretion system into host plant cells. Inside the plant cells TALEs bind and activate host genes thereby promoting bacterial disease. The DNA binding domain of TALEs is modular and consists of imperfect 33-35 long tandem-arranged amino acid repeats. Each repeat binds to a single nucleotide with position 13, determining base specificity. The base specificity of distinct residues in position 13 is known as the TALE code. This TALE code provides the possibility to create custom TALEs with desired DNA target specificity or to predict DNA targets of native TALEs.

This work characterizes two new members of the TALEs, called TALE-likes: (1) Bats, which derive from the bacterium *Burkholderia rhizoxinica* and (2) MOrTLs, whose DNA sequences were found in a marine metagenomics database. We demonstrate that DNA binding preferences of these two classes of TALE-likes can be predicted with the TALE-code. Yet, some of the repeats have a lower base specificity than TALEs. Additionally the TALE-likes have a different affinity to DNA and higher protein stability compared to TALEs. Analysis of protein chimeras showed that repeats of TALEs and TALE-like proteins are compatible and can be used to create protein chimeras. The TALE-likes have different DNA affinities and protein stabilities as compared to the TALEs. They can be adapted to create new proteins and protein chimeras with new useful properties.

List of publications

1. de Lange, Orlando*, **Wolf, Christina***, Dietze, Jörn, Elsaesser, Janett, Morbitzer, Robert & Lahaye, Thomas
   Programmable DNA-binding proteins from *Burkholderia* provide a fresh perspectiveon the TALE-like repeat domain.
   Nucleic Acids Research (2014) **42** (11): 7436-7449.

2. 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 (2015) **43** (20): 10065-10080.

* joint first authorship

Personal contribution

Erklärung nach § 5 Abs. 2 Nr. 7 der Promotionsordnung der Math.-Nat. Fakultät
Anteil an gemeinschaftlichen Veröffentlichungen

Declaration according to § 5 Abs. 2 No. 7 of the PromO of the Faculty of Science -
Share in publications done in team work

Publication Nr. 1 – shared first authorship

Christina Wolf

- Planned and designed experiments shown in figures 2, 3, 5, 6, 7, S7, S12 and S13 together with Orlando de Lange
- Performed experiments displayed in figure 2, 3, 5, 6, 7, S7, S12 and S13
- Analysed experiments related to figure 2, 3, 5, 6, 7, S7, S12 and S13, with support by Orlando de Lange and Thomas Lahaye
- Assisted in manuscript preparation

Publication Nr. 2 – shared first authorship

Christina Wolf

- Conceived the study together with Thomas Lahaye and Orlando de Lange

- Planned and designed the experiments related to figures 2, 4, table 1, figures S3, S5, S6 and S7, with support by Orlando de Lange
- Performed experiments displayed in figures 2, 4, table 1, figures S3, S5, S6 and S7
- Analysed and evaluated the experiments related to figure 2, 4, table 1, figures S3, S5, S6 and S7, with supported by Orlando de Lange and Thomas Lahaye
- Prepared manuscript together with Thomas Lahaye and Orlando de Lange

Introduction

TALEs are effector proteins of plant pathogenic Xantomonads

Members of the bacterial genus *Xanthomonas* infect a wide range of plants, including important crops such as pepper, tomato, rice and soybeans (Leyns et al., 1984). The gram-negative bacteria inject transcription activator like effectors (TALEs) into plant cells via the Type III secretion system (Rossier et al. 1999; Kay et al., 2007). TALEs are transported to the plant nucleus and transcriptionally activate the promoters of susceptibility genes (S-genes) to promote disease (Figure 1) (Bogdanove et al., 2010).

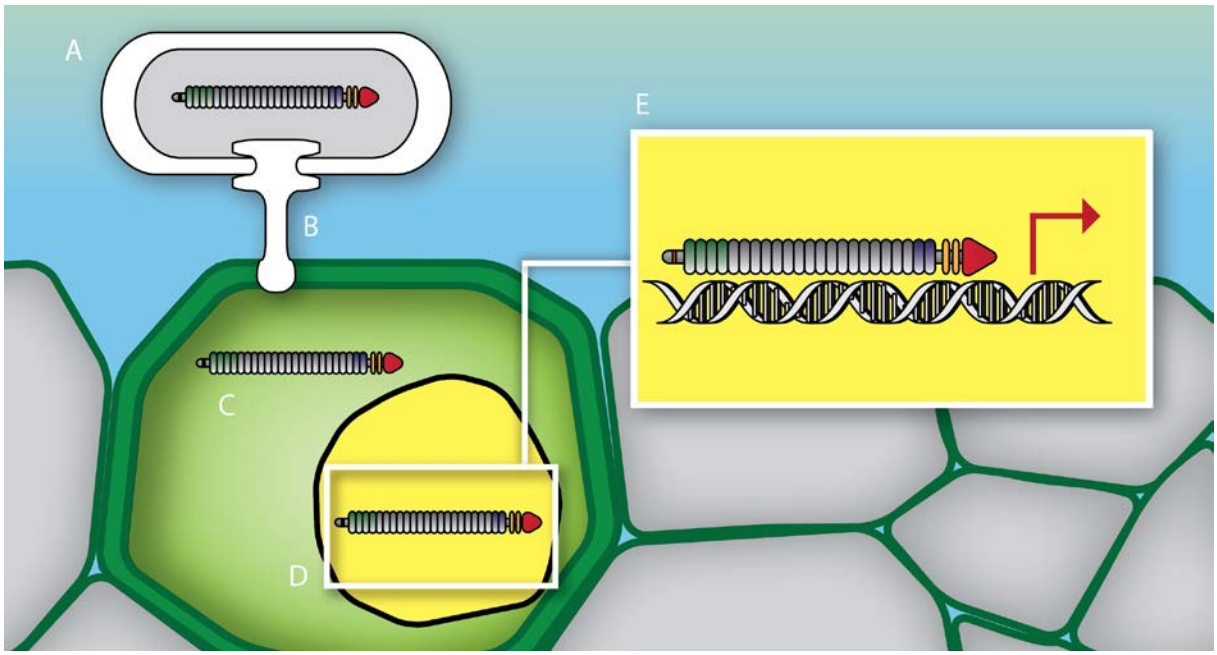


Figure 1: Model of the molecular function of TALEs. (A) After the infection of the plant by *Xanthomonas campestris pv. vesicatoria* (*Xcv*), TALEs are translocated by the Type III secretion system (B) into the plant cell (C). TALEs are transported into the nucleus (D), where they induce expression of specific target genes (E). Figure not to scale.

Many S-genes encode sugar transporters or transcription factors, which can facilitate bacterial growth (Chu et al. 2006; Zhou et al., 2015; Hu et al., 2014; Sugio et al., 2007). In resistant plants, plant immunity can be triggered by the detection of TALEs. There are at least four mechanisms for counteracting the pathogenic effect of TALEs:
(1) Direct detection of the TALE protein by the plant disease resistance proteins harbouring nucleotide binding site–leucine rich repeat (NBS-LRR) domain structures
(2) Mutation of a general transcription factor that interacts with the TALE
(3) Mutation of the binding site of the TALE in the promotor of the S-gene

(4) Activation of an executer gene by the TALE, whose promoter contains the TALE binding sequence (Gu et al., 2005; Römer et al., 2007; Ballvora et al., 2001; Schornack et al., 2004; Moscou & Bogdanove, 2009; Schornack et al., 2004).

TALEs are modular DNA binding transcription factors

The TALE protein is composed of 3 major structures: (1) The N-terminus, harbouring type III secretion signals and the N-terminal region (NTR) (Yang et al., 2000), the (2) central repeat region (CTR) that mediates DNA binding and consists of almost identical protein repeats (Bonas et al., 1989) and (3) the C-terminus, with nuclear localisation signals (NLS) (Yang et al., 2005b; Y. Yang & Gabriel, 1995b) and an acidic activation domain (AAD) (Yang et al., 2000; Zhu et al., 1998) (Figure 2).

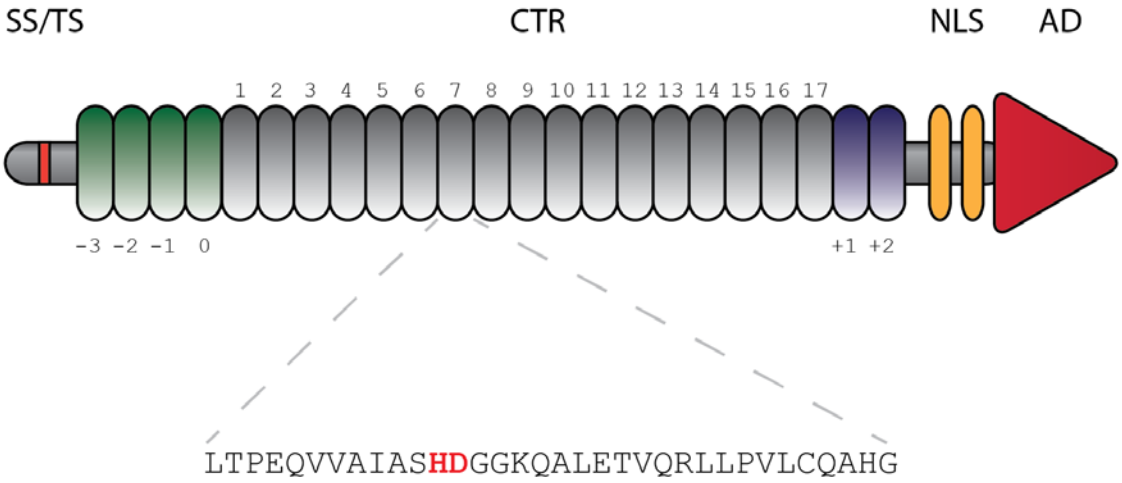


Figure 2: Model of the functional domains of a TALE. The TALE consists of the following domains (from N- to C-terminus): I) type III secretion signals (T3SS, red line). II) Four N-terminal cryptic repeats (green ovals, -3 to 0). III) central repeat region (CTR) with seventeen repeats (display of a typical repeat shown below, at position 12 and 13. The XX is a placeholder for amino acids called repeat variable diresidue (RVD), which varies throughout the central repeats). IV) Additional two cryptic repeats at the C-terminus (compare text). V) Nuclear localisation signals (NLS, yellow ovals). VI) A transcriptional activation domain (AD, red triangle). Model not to scale.

Experimental and computational data (Boch et al., 2009; Moscou & Bogdanove, 2009) supported by subsequent structural analysis (Deng et al., 2012; Mak et al., 2012) demonstrated that the TALE's nearly identical repeats in the CTR mediate sequence specific contact to the DNA. Each repeat recognizes a single base of the DNA. The TALE wraps around the DNA as a positive super helix, supported by its repeats. The repeats themselves are forming two alpha helices interrupted by a loop structure (Deng et al., 2012; Mak et al., 2012).

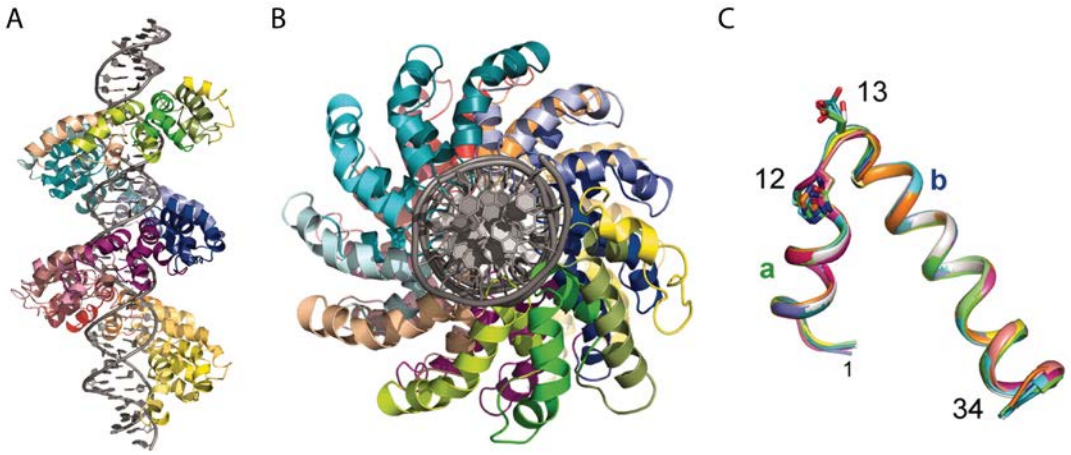


Figure 3: TALE structure. (A) and (B) Structure of the TALE PthXo1 DNA binding region in complex with its DNA target site**.** From Mak et al. 2009. Reprinted with permission from AAAS. (C) All TALE repeats have almost the same structure. Each repeat has a short (a) an and long (b) alpha helix connected by a loop where the two amino acids (position 12 and 13) of the RVD are located. From Deng et al. 2009. Reprinted with permission from AAAS.

The majority of the binding energy is derived from the TALEs' non-base specific interaction between the positively charged residues 16 and 17 of the repeats with the negatively charged DNA phosphate backbone (Deng et al., 2012). The TALE repeat array forms a single electropositive stripe around the DNA (Figure 3). While every repeat consists of almost identical amino acid sequences, the amino acids at position 12 and 13 of each repeat are variable. Hence, they are referred to as repeat variable residues (RVDs). The amino acid at position 12 is oriented away from the DNA and connects to the amino acid at position 8 in the first helix of the repeat. The hydrogen of the 12th amino acid makes a direct hydrogen bond to the carbonyl oxygen of the amino acid at the position 8 (Deng et al., 2012). This bond constrains the RVD-loop and stabilizes it. Earlier, it was believed, that the RVD defines the base specificity, but later on it was discovered that the amino acid at position 13 is responsible for it; in most cases by direct contact with the DNA (Mak et al., 2012). This amino acid is therefore called base specifying residue (BSR) (de Lange et al., 2014a). The base preference of the most common RVDs HD, NG, NN and NI is C, T, G and A respectively (Boch et al., 2009; Moscou & Bogdanove, 2009). The interactions between the protein and DNA are mediated by hydrogen bounds (e.g. NN to G), nonpolar interactions (e.g. NG to T) or van der Waals forces (e.g. HD to C) (Deng et al., 2012; Mak et al., 2012). The base specificity is often achieved by negative discrimination, due to a physical and electrostatic clash between the "non-perfect-match"-bases. For example the N* RVD (* depicts a missing 13th amino acid) allows

only pyrimidin bases, because there is a steric hindrance for purines (Cong et al., 2012).

Transcription factors typically activate only a limited number of genes, which is due to target specificity of their DNA binding domains. DNA specificity is defined by the affinity to the optimal sequence, compared to the affinities of all other possible sequences (off-targets). The base specificity of TALEs is generally very high, but not absolute. For instance, the interaction of amino acids I and D at position 13 of the repeat is highly specific to the bases adenine and cytosine respectively (Boch et al., 2009). On the other hand, a strong interaction can also be observed between an N at position 13th and bases guanine and adenine (Boch et al., 2009; Miller et al., 2015). Even though in most cases the BSR is in direct contact with the DNA, its binding is still influenced by the 12th amino acid of the repeat (Miller et al., 2015; Rogers et al., 2015). In case of H or N at this position, an overall higher protein activity could be observed (Miller et al., 2015). Furthermore, the second specifically bound DNA base of the TALE is in the most cases an adenine, whether it is supported by its BSR or not (Boch et al., 2009; Miller et al., 2015; Rogers et al., 2015). Other factors that can influence the specificity of TALEs are protein length (Meckler et al., 2013) and the identity of the adjacent RVDs. For instance it could be shown, that depending on the context, the RVD NN binds preferably to an adenine or guanine (Miller et al., 2015).

Parts of the N-terminus consist of so called non-canonical or cryptic repeats (NTR), with lower homologies to the canonical repeats of the CTR (Gao et al., 2012; Szurek et al., 2002). The -3 to 0 repeats have the same overall structure as the canonical repeats, but seem to interact with DNA in a functionally distinct way as compared to the canonical repeats. The cryptic repeat -1 connects with a thymine base of the DNA (position 0 of the binding motif) by interaction with the tryptophan at position 232 of TALE protein (Mak et al., 2012). Therefore most natural targets start with a thymine base (Boch et al., 2009; Moscou & Bogdanove, 2009). The other cryptic repeats do not show a base specificity. Instead of the BSR they contain a loop region spanning amino acids 11-15 (Mak et al., 2012). The NTR is important for the initial non-sequence-specific binding of the TALE to the DNA (Cuculis et al., 2015). This initial binding is almost independent of the DNA sequence (Gao et al., 2012) and facilitates sliding along the DNA. After this “nucleation event”, the CTR recognizes the DNA in a sequence specific fashion (Cuculis et al., 2015). Additional cryptic

repeats (+1, +2) are found at the C-terminus. It is known, that the +1 repeats exhibit base specificity (Boch et al. 2009).

Natural occurrences of TALEs and TALE-like proteins

Proteins with homology to TALEs (TALE-likes) are not restricted to the genus *Xanthomonas* but are also present in the bacterial plant pathogen *Ralstonia solanacearum (*de Lange et al., 2013; Hopkins et al., 1992) and *Burkholderia rhizoxinica*, a bacterial endosymbiont of a plant pathogenic fungus (Lackner et al., 2011a; Schornack et al., 2013). Similar to TALEs, the TALE-likes of *Ralstonia* (RipTALs) are transported into the nucleus and have an activation domain at the C-terminus to potentially activate genes in the host plant (de Lange et al., 2013). The function of the TALE-likes of *Burkholderia* (Bats) is unknown. While *Xanthomonas* TALE repeats differ mainly in their RVDs, RipTALs and Bats show amino acid sequence variation also in non-RVD residues of their repeats. TALE-likes use the same RVD code as TALEs, but the frequency of different RVDs varies. RipTAL repeats are functionally compatible with TALEs and interchangeable (de Lange et al., 2013; Li et al., 2013). Unlike TALEs, RipTALs have a preference for G at position 0 of the binding site (de Lange et al., 2013). In contrast to TALEs, the RVDs of RipTALs are quite similar. While RipTALs and TALEs are transcriptional activators, the function of Bats remains cryptic. The challenging model system (pathogenic fungal symbiont) has made the characterization of the proteins in their natural context difficult (Lackner et al., 2011b; Lackner et al., 2009; Silipo et al., 2012). The protein characterization of the Bats is part of this work.

The dawn of new genome engineering tools

During the last few years, the geneticist's toolbox consisting of DNA sequencing, restriction enzymes and PCR, got more powerful through the inclusion of programmable DNA binding proteins (PDPs). Some of them can be used to cut DNA and therefore act as a genome editing tool. With additional fusion proteins they can also mark loci, introduce epigenetic changes, activate or repress genes or exchange genomic regions (reviewed in de Lange et al. 2014a). To delete or modify a single gene and not to affect others is one of the major obstacles in modern biotechnology. It is therefore beneficial to have more than one technology of PDPs available as they exhibit different positive and negative characteristics depending on the application.

In the following chapters, three of the most used PDP systems are described: Zinc finger proteins, the CRISPR/Cas9 system and the TAL effectors.

Zinc finger proteins

Method of the year 2011 was the Zinc-finger Nuclease (ZFN) ("Method of the Year 2011," Nature Methods, 2012). It consists of a specific Zinc-finger (ZF) DNA binding protein and a nuclease. The composite protein is able to cut DNA *in vivo* and *in vitro*. Generally ZF proteins used for genome engineering belong to the Cys2-His2 group of ZFs which carry the motif X_2-Cys-$X_{2,4}$-Cys-X_{12}-His-$X_{3,4,5}$-His in their protein sequence (Kim et al., 1996; Papworth et al., 2006). The zinc ion in the ZF is coordinated between an alpha-helix and an antiparallel beta-sheet by the two histidine and two cysteine residues. A ZF binding protein consists of several single zinc finger domains which are organized in tandem (Gommans et al., 2005). For sufficient binding at least two fingers are required. For biotechnological applications typically three or more fingers are used in tandem to bind a 9 to 12 DNA base pair long motif (Pabo et al., 2001). Each of these fingers binds to 3 bases of the major groove of the DNA (Fairall et al., 1993; Pavletich & Pabo, 1991). The specificity is defined by the alpha-helix (“recognition helix”, especially the amino acids at position -1, 2, 3 and 6) of each finger that connects to the 3 or more DNA bases and produces an overlapping pattern of contacts with neighbouring zinc fingers (Fairall et al., 1993; Wolfe et al., 2000). Exchanging the amino acids at those positions will create a new ZF protein with a different binding specificity. Tandem arranged zinc fingers in the zinc finger binding protein are not independent functional units, but influence each other in their DNA specificity. This makes the design of a zinc finger binding protein non-trivial, especially for non-specialist researchers. Additionally, the overlapping contacts between adjacent fingers consist of Arginine residues, which enable a strong natural interaction with Guanines, limiting the flexibility of the target (Pavletich & Pabo, 1991). To facilitate the ZF design, a library of characterized ZFs is available from the “zinc finger consortium” (http://www.zincfingers.org/).

The most commonly used ZF construct, is the Zinc finger nuclease (ZFN). It consists of the nuclease FokI fused to a ZF protein. FokI cuts DNA as a dimer, therefore two ZFN are needed, whose binding sites face each other (Bitinaite et al., 1998; Y. G. Kim et al., 1996). This leads to a double strand break, which is repaired by the cells non-homology end joining (NHEJ) or homology directed repair. The efficiency of

ZFNs is low: the mutation rate of the desired target is around 10% (Gabriel et al., 2011; Kim et al., 2012; Kim et al., 2010). Double strand breaks are also prone to cause cell toxicity, if they occur in vital gene regions. Because they are highly mutagenic and the DNA binding specificity is not high, ZFN also exhibit off-target activity, which again can lead to cell toxicity (Beumer et al., 2006). Furthermore, to be useful in clinical applications zinc fingers have to be extensively screened (Carroll, 2011; Perez et al., 2008; Urnov et al., 2005).

Cas9/CRISPR

The primary function of the Cas9/CRISPR system *in vivo* is to provide sequence specific und acquired immune defence in bacteria and archaea (Sorek et al., 2008). Of the three known types of Cas9/CRISPR mechanisms, the type II system is the best studied one.

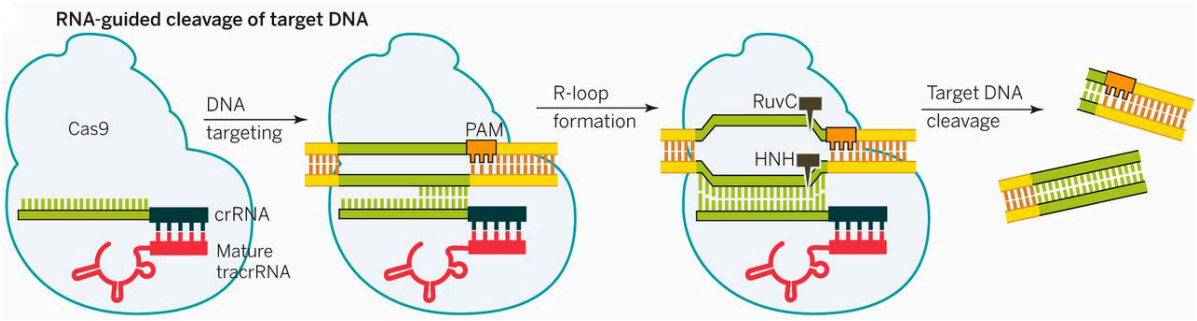


Figure 4: CRISPR/Cas9 system. A duplex of CRISPR RNA (crRNA) and the trans-activating CRISPR RNA (tracrRNA) recognizes the target DNA, which is then cleaved by the endonuclease domain of Cas9. From Doudna and Charpentier, 2014. Reprinted with permission from AAAS.

Cas9/CRISPR is composed of two main elements (Figure 4): (1) DNA repeats interspaced by sequences derived from viruses (Clustered regularly interspaced short palindromic repeats (CRISPR)) and (2) the protein Cas9. Foreign DNA is first detected and added to the CRISPR array. The CRISPR locus is then transcribed and spliced. The processed product is called CRISPR-RNA (crRNA). The crRNA retains a portion of the spacers and the repeats and carries a short sequence motif, termed protospacer adjacent motif (PAM), which is essential to distinguish the viral DNA from the bacterial one. Additionally another non-coding RNA, the trans-activating CRISPR RNA (tracrRNA), is transcribed from the CRISPR locus. Thirdly, the crRNA forms a complex with the tracrRNA and directs Cas9 to the viral target DNA, the so called protospacer (O'Connell et al., 2014). The specificity of this interaction stems from Watson-and-Crick base pairing of the crRNA and the protospacer (Ran et al.,

2013). Finally, the nuclease domain of the Cas9 protein makes double stranded breaks to the DNA with a two nucleotide overhang within the protospacer.
Changing the sequence of the spacer region to target different loci in a genome turns the system into a programmable DNA cleaving enzyme (Jinek et al., 2012). For biotechnology purposes the system was simplified by linking together the crRNA and the tracrRNA to create a single RNA, called single guided RNA (sgRNA).
At present the CRISPR/Cas9 is the preferred system for DNA editing as it does not because require protein engineering and is easily adapted. The sole required design element is the sgRNA, which is only limited by the short (3-5 bp) PAM DNA sequence. Furthermore “multiplexing”, that is introducing multiple sgRNAs, allows to simultaneously target different sites in the genome in the same cell/organism.
Because of the reliance on DNA-RNA base pairing, the CRISPR/Cas9 system can bind to other sequences than the desired ones (“off-targets”). Usually up to 5 mismatches are tolerated (Mali et al., 2013), sometimes even more (Wu et al., 2014). While binding of Cas9 can be observed to up to 6000 different targets in the genome, off-target cleavage is quite low (Wu et al., 2014). Several strategies are used to improve the specificity of the CRISPR/Cas9 system and thus to avoid or minimize these off-target effects. To optimize the sgRNA for less off-targets is not trivial, as demonstrated by several papers describing different conclusions about the design of the sgRNA (Fu et al., 2013; Hsu et al., 2013). Therefore, preferentially a nickase version of the Cas9 protein is used, which requires two sgRNAs to create a double strand break (Mali et al., 2013). Analogous to ZFN, the binding sequences have to face each other. While off-target prediction is not an easy task in itself, differences in expression levels can additionally affect the off-target occurrences (Mali et al., 2013).
A nuclease deficient Cas9 (dCas9) that is unable to cleave DNA, was used to fuse other functional domains, like activation domains or epigenetic modifiers to target genome sites analogous to the ZF protein (Qi et al., 2013). However, due to the high potential for off-target-bindings, the CRISPR/Cas9 system is not ideally suited for such purposes.

dTALEs

The strictly modular organisation of the TALE DNA binding domain provides the conceptual basis to compose designer TALEs (dTALEs) with the desired target specificity. Through different fusion proteins, dTALEs are able to activate, repress,

cut, modify, recombine and mark DNA (reviewed in de Lange et al. 2014a). Countless kits and tools are available for the researcher to compile a dTALE with the desired binding sequence (Cermak et al., 2011; Morbitzer et al., 2011; Reyon et al., 2012). Although only few reactions are necessary to assemble a dTALE construct, cloning and designing dTALEs is still not trivial for a non-specialist researcher. In theory, almost any sequence can be targeted (Meckler et al., 2013). The base preference for T in the N-terminus (T-0 limitation, see page 13) can be omitted by using different N-termini designed for different bases or strong binding RVDs and therefore the dTALE is able to bind any of the bases (Doyle et al., 2013; Lamb et al., 2013). Polarity effects (repeats at the N-terminus contribute more to the specificity than those at the C-terminus) and the influence of RVDs on adjacent repeats are the major obstacles in dTALE-design (Mali et al., 2013; Meckler et al., 2013; Miller et al., 2015; Rogers et al., 2015), as these can lead to off-target binding of TALEs (Lin et al., 2015; Mali et al., 2013; Mendenhall et al., 2013).

Viral vectors are often used to integrate foreign DNA into mammalian organisms. It was observed that by viral transport (adenoviral, lentiviral) into a mammalian system, dTALE DNA repeats can be recombined and lost (Cai et al., 2014; Holkers et al., 2013). The reason for this instability is probably that the almost sequence identical repeats are prone to polymerase slippage. This stability issue was tackled by synonymous codon exchanges (Yang et al., 2013), but could not be completely solved because of limitations due to codon usage.

Objectives

Previous research has discovered homologues to the TALE protein, called TALE-likes (de Lange, et al. 2013; Hopkins et al., 1992). While the TALE-likes from *Ralstonia* (RipTALs) were already described, the TALE-likes found in *Burkholderia* (Bats) and other organisms (MOrTLs) are not characterized yet. The aim of this study was to conduct a biochemical analysis of the Bat and MOrTL proteins using *in vitro* and *in vivo* approaches. These studies were intended to clarify if these TALE-like proteins act as functional DNA binding proteins and if they bind DNA according to the TALE code. A further goal was to localize Bat proteins in a human cell system and to investigate, if they can activate a reporter system in the human cells either by themselves or by creating a Bat fusion protein with an additional activation domain. A Bat activator was created in order to test if the Bat repeat array can be changed to target a new sequence of interest. Finally, we investigated if TALE-likes can be used to create chimeras with known TALE-likes.

The overarching goal of these analyses was to compile a pool of TALE-like sequences, which can be used to create custom DNA binding proteins for a variety of biotechnological applications.

Discussion

Bats and MOrTLs expand the group of known TALE-likes

A few years ago it was believed that TALEs are unique in their function as modular DNA-binding proteins, whereby one module can bind a base in a one to one fashion. With the characterization of RipTALs (de Lange et al., 2013; Li et al., 2013), we learned that this class of proteins is not limited to the genus *Xanthomonas*, but that other organisms have proteins with similar characteristics. Universally, these proteins were classified as TALE-likes (page 7). In our previous work, we identified at least two new groups of TALE-likes, named Bats and MOrTLs (de Lange and Wolf et al., 2014b; de Lange and Wolf et al., 2015)

Bats are able to bind DNA in a sequence specific fashion

Bats originate from the obligate endosymbiotic bacterium *Burkholderia rhizoxinica*, whose host *Rhizopus microspores* is a fungal plant pathogen *(*Lackner et al., 2009). In this mutualistic symbiosis, the bacterial endosymbiont produces an antimitotic macrolide called rhizoxin, which causes “rice seedling blight”, a severe plant disease that is affecting rice plants in Asia. Recently, the complete genome sequence of *B. rhizoxinica* was made available to the public by high-throughput sequencing (Lackner et al., 2011a; Lackner et al., 2011b). Three predicted proteins with homologies to TALEs were identified based on the sequenced genome. For two of them, designated Bat1 and Bat2, we demonstrated binding to their predicted DNA target *in vitro* whereas the third *B. rhizoxinica* protein, Bat3, was not able to bind DNA in an affinity assay (figure 2 (a) in de Lange and Wolf et al., 2014b). An explanation for that could be the reduced number of repeats of Bat3 (6) compared to Bat1 (20) or Bat2 (26), which might be insufficient to effectively bind DNA. In agreement with this assumption, TALEs with less than ten repeats did not activate promoter-reporter constructs that contain corresponding binding sites (Boch et al., 2009).

Although the Bat proteins share less than 40% identity with the consensus core repeat of TALEs from *Xanthomonas*, Bat1 and Bat2 bind to DNA targets predicted by the TALE code. As demonstrated in a competition assay, the base discrimination of Bat proteins is very stringent and appears similar to that of TALEs. In this competition assay, Bases 6-10 of a competitor DNA were replaced with the least favoured bases

for the corresponding binding element according to the TALE code. The competitor DNA containing the mismatches was not able to out-compete the TALE code-predicted DNA probe of Bat1, even if this off-target competitor was added in 200-fold excess (figure 2 (e) in de Lange and Wolf et al. 2015). We performed the same experiment with a TALE, which had the same binding site preference as Bat1 and it also couldn't be outcompeted with the same off-target sequence (figure 4 (e) in de Lange and Wolf et al., 2015). The measured specificity of Bats fits to previously the observed polarity effect for TALEs, which showed that altering DNA binding sequences at the 5' end has a stronger impact than mismatches in the 3' end (Mali et al., 2013; Meckler et al., 2013; Miller et al., 2015).

The specificity of the protein-DNA interaction in TALEs is predominantly defined by the 13th amino acid of the repeat. Despite the low sequence identity, between TALEs and Bats, our data indicate that the Bats discriminate nucleotides exactly like TALEs, with similar specificity. This is in agreement with the Bats crystal structure and their structural model, which revealed a high level of similarity to the three dimensional structure of TALEs (Stella et al., 2014).

The biological function of Bats is unknown

For a long time it was not known how the natural TALE is contributing to the pathogenicity of *Xanthomonas* species. Their nuclear localization and the presence of an acidic activation domain demonstrated that TALEs are *in planta* transcription factors that activate host genes to promote disease (Kay et al., 2007; Marois et al., 2002; Van den Ackerveken et al., 1996). At present, neither the localization nor the DNA targets of Bats are known. It is possible that Bats either act in the endosymbiotic bacterium, the fungal pathogen, the fungal host plant, or in several of these organisms. Given that Bats lack an obvious C-terminal activation domain it seems unlikely that they function as transcription factors. Yet, in the absence of molecular targets it will be difficult to substantiate this hypothesis experimentally.

In contrast to the other TALE-like proteins (TALEs and RipTALs), the N-terminus of Bats, which is followed by the repetitive DNA binding domain, consists of only 17-18 amino acids, with no predicted functional domains. The N-termini of TALEs and RipTALs have been shown to constitute functional type III translocation signals (de Lange et al., 2013; Mukaihara & Tamura, 2009; Rossier et al., 1999). These allow the bacteria to translocate effector proteins into plant cells, where they can activate

certain genes. For Bats we couldn't detect any obvious type III translocation signals correlated with the *Burkholderia* genus type of translocation signals with prediction programs based on known secretion systems (Rainbow et al., 2002; Samudrala et al., 2009). If there is indeed no type III secretion system sequence present, it is unlikely that Bats are translocated into eukaryotic systems, be it the symbiotic fungi or plants.

Supporting this hypothesis, no nuclear localisation signal (NLS) was predicted in Bats. It is possible, that Bats harbour a cryptic NLS, undetected by current algorithms, but most NLSs are conserved and can function across kingdoms (Chang et al., 2013; Guralnick et al., 1996; Lassner et al., 1991; Rhee et al., 2000; Vanderkrol & Chua, 1991). Therefore and because the natural system to study Bats, consisting of the bacterium *B*. *rhizoxinica*, the fungus *R. microspores* and the host plant, is not accessible in our lab, we decided to test the localization in a human cell system by transfection of HEK293T cells with the original Bat1 construct. However, no exclusive localisation to the nucleus or to other cell organelles was observed (figure 3 (b) in de Lange and Wolf et al., 2014b). This may suggest that Bats do not function in a eukaryotic cell system; instead it is possible that, if they are expressed, they act directly in *B. rhizoxinica*, where no translocation to the nucleus or to mitochondria would be required.

The affinity of Bats was demonstrated to be more than 10-times lower than that of most TALEs to their perfect DNA binding site (figure 2 (b) in de Lange and Wolf et al., 2014b; Meckler et al., 2013; Strauss et al., 2012). This could be another indicator that Bats are not transported to the outside of the bacterium, unlike TALEs. TALEs are secreted by a type III secretion system, which is limited in the amount of protein to be translocated (Enninga et al., 2005; Mills et al., 2008). The probability of binding a certain DNA site depends both on the affinity and the amount of proteins in the cell (Aurell et al., 2007; Bintu et al., 2005; Stormo, 1998; Stormo & Fields, 1998). Hence a high affinity can be important for proteins that are translocated via a type III secretion system, because of their limited abundance in the target cell. If the Bats are not translocated, but instead are acting as regulatory proteins in the bacterium itself, a lower binding affinity could actually be advantageous, as it could allow for regulation of Bat DNA binding activity by altering the available protein concentration in the cell.

By removing the activation domain, TALEs can function as transcriptional repressors (Blount et al., 2012). As no activation domain is predicted in the C-terminus of Bats, Bats could naturally act as repressors of gene expression. Alternatively, it is also possible that Bats recruit other proteins or signal molecules required for their native function. An example for this kind of regulation can be found in human cells, where a nuclear oncoprotein recruits factors to a DNA binding protein in order to repress transcriptional activation (Luo et al., 1999). It is possible that for the Bats' native function, additional bacterial, fungal or plant factors are required, which are absent from the assayed human cell system (figure 3 (b) in de Lange and Wolf et al., 2014b). Another option is that the transport of Bat proteins depends on the presence of helper proteins, which bind to the Bat and transport it to the plant or fungus. This so called "piggy-back" mechanism has been described for several other proteins (Wagstaff & Jans, 2009).
Further research, including the study of Bats in their natural context, will be required to untangle the native function of Bats' and to analyse if they are indeed expressed and active in the bacterial system.

acBats can be used to address functional differences between Bats and TALEs

In order to study the DNA binding properties of the Bats, we synthetically fused them to a transcriptional activation domain. These so called activator Bats (acBats) enabled us to study Bats *in vivo* by using simple promoter-reporter assays.
The analysis of TALEs by promoter-reporter assays was carried out predominantly using a GUS-reporter. A disadvantage of standard GUS assays is that the expression of TALEs is not monitored and thus the data are not normalized. We therefore decided to use two distinct fluorophores to measure reporter activity and expression of the TALE-like protein in a single cell assay. Fluorescent reporters are an invaluable tool to study gene expression and are also commonly used to assess functionality of TALEs. The fluorescent readout was quantified by flow cytometry and both protein expression and transcriptional activation could be monitored through two distinct fluorophores, thus rendering protein extraction and quantification unnecessary. Additionally with more than 10,000 single cells measured, the reproducibility of the assay was very high. A chimeric protein consisting of an activation domain, a NLS and Bat1 protein, was capable to drive expression of a GFP reporter in the human cell line HEK293T (figure 3 (b) in de Lange and Wolf et

al., 2014b). In contrast to the Bat wildtype protein these chimeras could be studied as transcriptional activators (acBats), thus simplifying experimental comparisons of Bats and TALEs. Compared to an equivalent dTALE construct, the fluorescence read out of the Bat-fusion protein was lower (figure 3 (c) in de Lange and Wolf et al., 2014b). The comparatively weak fluorescence of the reporter could indicate that Bat1 has a lower DNA binding affinity than most TALEs (Mali et al., 2013; Meckler et al., 2013). Meckler et al., 2013 demonstrated that TALEs with a lower affinity to a DNA sequence also showed reduced fluorescence output in a reporter assay. Our analysis using microscale thermophoresis (MST) confirmed that the binding affinity to its DNA target was lower for Bat1 than that generally reported for TALEs to their perfect binding sequence (figure 2 (b) in de Lange and Wolf et al., 2014b; Meckler et al., 2013). Another explanation for the low fluorescence of the reporter after the acBat induction could be due to a structural clash of the acBat with the C-terminal activation domain. It could be that the fusion of the activation domain hinders binding to the DNA or assembly of the transcription machinery.

In summary, the established fluorescence *in vivo* reporter system provides a convenient tool to study functional differences between Bats and TALEs. In the future this reporter system could be used to characterize even minor differences in DNA binding of TALE-likes in a fast and reliable fashion.

N- and C-terminal truncations of non-canonical repeats of Bat1 have an impact on reporter activation

Besides the almost identical canonical repeats of TALEs, which are responsible for direct specific DNA binding, TALEs contain non-canonical repeats in the N- and C-terminus, which are highly divergent at the sequence level. Due to the lower sequence homology in comparison to the core repeats they are also called cryptic repeats (Mak et al. 2012).

We studied the possible contribution of the Bat1 N- and C-termini for binding DNA by analysis of truncation derivatives using the established human cell reporter assay (see page 25). For the Bat1 protein both N-terminal and C-terminal truncations reduced the fluorescence of the reporter considerably (figure 5 (b) in de Lange and Wolf et al., 2014b), indicating that the deleted regions contribute to DNA binding. In the case of the N-terminal acBat deletion construct, the reporter activity was reduced by 56% in comparison to the full length acBat1. For TALEs, removal of the cryptic

repeats in the N-terminus has a more drastic impact: For instance N-terminal truncations of TALEs can abolish all detectable DNA affinity (Gao et al., 2012) or reduce the DNA affinity tenfold compared to a full length TALE (Kay et al., 2007).

Figure 5: Sequence alignment of the repeat -1 and 0 of the TALE AvrBs3 and Bat1. Identical amino acids are marked with red boxes.

One explanation for the reduced affinity is that the N-terminus of TALEs is necessary for positioning the protein on the DNA and sliding it to the target sequence (Cuculis et al., 2015). The N-terminus of TALEs makes a significant, but non-base specific contribution to the DNA binding, an observation that we couldn't confirm for Bats (figure 5 (b) in de Lange and Wolf et al., 2014b). The N-termini of Bats and TALEs differ both in size as well as in protein sequence. The comparison of the -1 and 0 repeat of AvrBs3 and Bat1 shows only a limited sequence homology of 20-30% (Figure 5), which is even lower than the overall protein homology. A homology below 30% generally indicates low conservation and potentially altered protein function (Koonin & Galperin, 2003; Pearson, 2013). It is therefore possible that the Bat-N-terminus does not contribute as much to the overall DNA binding affinity as the N-terminus of TALEs. The hypothesis that differences in the N-termini lead to changes in affinity could be tested by domain-swap experiments, where the N-terminus of Bat1 would be exchanged with the N-terminus of a TALE. Yet, it is possible that the Bat and TALE N-termini are functional only in their native context and thus the results of such experiments would have to be evaluated with caution.

The C-terminal truncation construct of Bat1 (the removal of the last +1 cryptic repeat) was barely able to activate the reporter above background level (33% of acBat)

(figure 5 (b) in de Lange and Wolf et al., 2014b). Again, this is in contrast to the observation made for TALEs, where removal of the C-terminal cryptic repeats did only have a minor or no effect on their activity (Mussolino et al., 2011; Römer et al., 2009).

Figure 6: Sequence alignment of the consensus repeat of Bats and the +1 repeat of Bat1. Same amino acids are marked with red boxes.

The primary structure of the +1 repeat of Bat1 is similar to that of the other canonical repeats, but their homology is very low (~33%) indicating a potentially different role in binding (Figure 6). It is possible that the highly positively charged residues at the C-terminus of Bat1 contribute more to the binding to the negatively charged DNA phosphate backbone than the C-termini of TALEs.

Bat1 +1 non canonical repeat
RSNEEIVHVAARRGGAGRIRKMVAPLLERQ

AvrBs3 +1 and +2 non canonical repeats
TPQQVVAIASNGGGRPALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRT

Figure 7: Comparison of the C-terminus non canonical repeats of Bat1 and AvrBs3. Seven of the thirty amino acids (~23%) of the C-terminus of Bat1 are charged, while only 7 of 69 amino acids (~10%) of the non-canonical C-terminus repeats of the TALE AvrBs3 are charged. Red letters indicate amino acids with charged residues.

A positively charged molecular surface is predicted to be crucial for DNA binding proteins to form an electrostatic interaction with the negatively charged DNA backbone (Marcovitz et al., 2015). Especially arginine is a common amino acid in DNA binding proteins that mediates the DNA – protein recognition (Gordan et al., 2013; Rohs et al., 2009; West et al., 2010). The N-terminus of TALEs, which is responsible for strong DNA binding, is highly positively charged. In comparison, the positively charged amino acids of the Bat1 C-terminus are, except for one, arginine residues, suggesting a strong contribution of the C-terminus in its interaction with DNA binding (Figure 7).

Truncations of canonical repeats of Bat1 reduce reporter activation

Efficient transactivation by TALEs required the presence of least 10 canonical repeats (Boch et al., 2009). In Bats, repeat truncations also reduced the activity of the Bat1 activator, (figure 5 (a) in de Lange and Wolf et al., 2014b) suggesting that Bat and TALE repeats collectively contribute to DNA binding. Loss of two C-terminal repeats (latter half of repeat 20 and repeat +1 were retained) caused a 40% reduction of reporter activity as compared to full length acBat1. Larger truncations (acBat1 lacking four, six or eight repeats) resulted in derivatives incapable of activating the fluorescent reporter above background levels (figure 5 (a) in de Lange and Wolf et al., 2014b). In contrast to TALEs, where deletions down to 15 canonical repeats usually do not alter the TALE activity *in vivo* Briggs et al., 2012; Cermak et al., 2011, Bats were affected more severely by the loss of repeats (figure 5a in de Lange and Wolf et al., 2014b). It is possible that due to the highly diverged repeats, Bats are not as modular as TALEs. Accordingly, truncations may cause more severe structural and functional disturbances in Bats than in TALEs. A loss of repeats could interfere with the protein's overall stability, thus leading to a drop in its activity. An alternate explanation could be that every repeat of the Bats contributes to the binding affinity of the full length protein. Losing repeats at the C-terminus could therefore result in an annihilation of the reporter's fluorescence signal (figure 5a in de Lange and Wolf et al., 2014b).

In conclusion, it seems that the structural characteristics of Bat repeats are different compared to those of TALEs and therefore the repeat structure of Bats probably follows a more complicated model.

MOrTLs are likely fragmented genes

The genetic diversity in marine microbial communities was assessed by the metagenomic Sorcerer II Ocean sampling expedition (Rusch et al., 2007). Two of the approximately 6000 open reading frames (ORFs), that couldn't be matched to any known genes, were later characterized as TALE-likes (Juillerat et al., 2014). Just like TALEs, RipTALs and Bats, it is thought that these sequences are bacterial in origin, because of the size exclusion of the harvesting filter (de Lange and Wolf et al., 2015). Therefore we named them MOrTLs (**m**arine-**or**ganism **T**ALE–**l**ikes).

The two TALE-likes ORFs, MOrTL1 and MOrTL2 share only about 37 % homology with TALEs and 33 % with each other (supplementary figure 1 in de Lange and Wolf

et al. 2015). It is possible that MOrTL1 and MOrTL2 stem from different marine microbial organisms, because of their low homology.

To characterize the proteins, we tried to express and purify MOrTL1 and MOrTL2 in an *E.coli* system. Because MOrTL2 precipitated in inclusion bodies and couldn't be purified (supplementary figure 3 (a) in de Lange and Wolf et al., 2015), DNA-protein binding experiments could only be performed with MOrTL1. By electro mobility shift assay (EMSA) we could demonstrate that MOrTL1 was able to bind DNA (supplementary figure 3 (b) de Lange et al., 2015). However, a very high amount of MOrTL1 protein (concentration over 0.8 µM of MOrTL1) was required for an electro mobility shift, indicating that the DNA binding affinity of the protein is very low. Interestingly, we could observe shifts with different migration patterns (supplementary figure 3 (b) in de Lange et al., 2015, potentially indicating a DNA to protein binding ratio other than the 1:1 ratio of TALEs to DNA. There are two possible explanations for this (Figure 8): (1) Multiple MOrTL1 proteins could attach to the target DNA via unspecific binding, resulting in the different shifts in the EMSA. (2) MOrTL1 could aggregate to form different multimers, which stick unspecifically to the DNA and produce different shift variants. It is known that TALEs in higher concentrations can aggregate and form dimers (Schreiber et al., 2015). The different shift species are therefore possibly an artificial effect resulting from the high protein concentrations, which are usually lower in living cells (Kim & O'Shea 2008, Sanguinetti et al. 2006, Liao et al. 2003, Gao et al. 2004). Additionally, as the MOrTL1 protein was purified from *E.coli*, the different shift species could be the result of interactions with native *E.coli* proteins of different molecular weight.

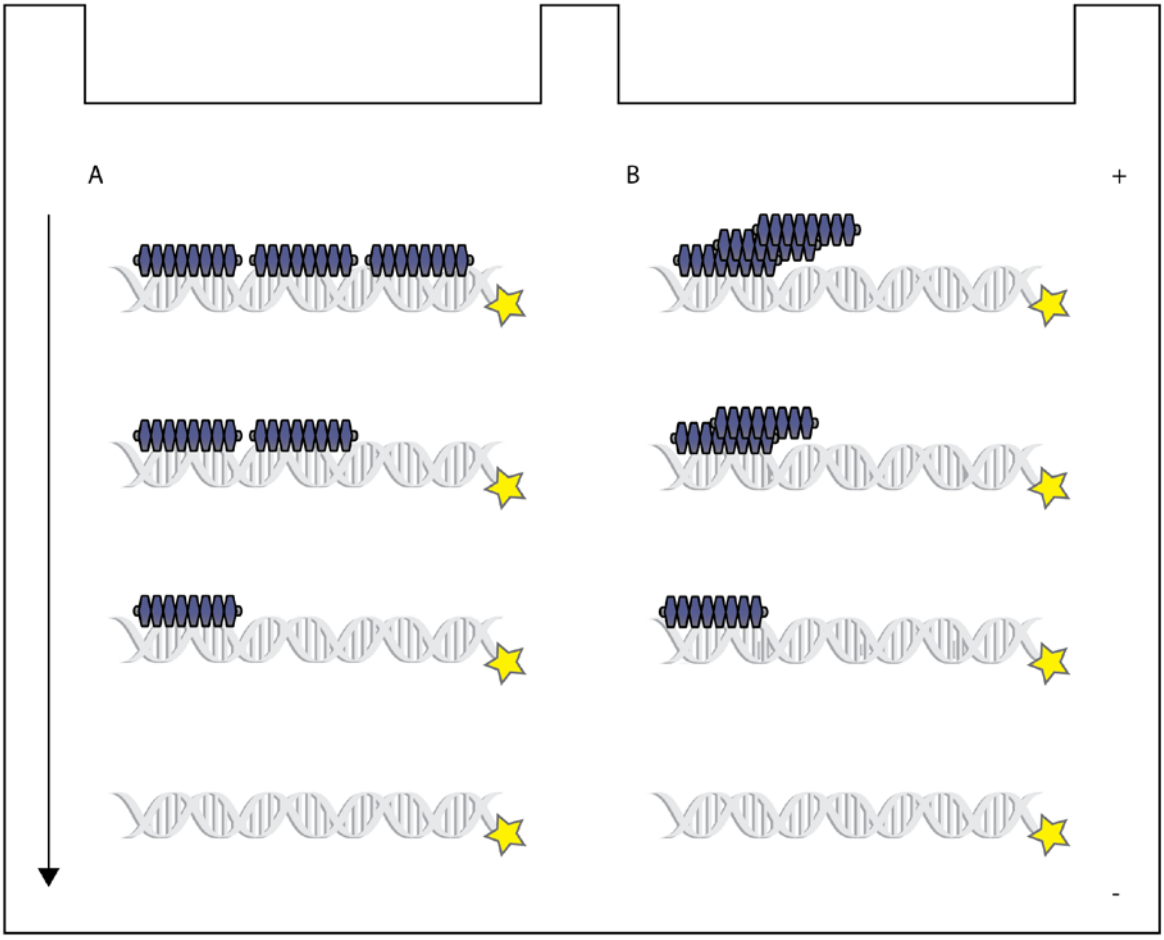


Figure 8: MOrTL1 EMSA hypothesis. DNA is detectable via fluorophore labelling. (A) the protein size of MOrTL1 is small enough to bind the DNA multiple times. Therefore different binding species can be observed. Alternatively (B), most TALEs tend to build aggregates. Therefore it is possible, that an aggregation product can produce different binding species, resulting in an observable laddering effect.

Efficient TALE DNA binding was demonstrated to require at least 10 repeats (Schreiber & Bonas, 2014; Streubel et al., 2012). As MOrTL1 and MOrTL2 consist of only 8 and 10 repeats respectively, a low DNA binding affinity does not seem surprising. It is therefore possible that binding to the target DNA is not efficient enough in physiological protein concentrations.

In TALEs, an important contribution to the DNA-binding ability is provided by the N-terminus (Gao et al., 2012) and the C-terminus seems to have a similar role in Bats (figure 3 (a) in de Lange and Wolf et al., 2014b, see page 26). The two MOrTL proteins contain only short non-repeat N- or C-termini, which is different compared to those of TALEs, RipTALs and Bats. Interestingly, both N- and C-termini of MOrTLs show a higher resemblance to repeat fragments than any other N- or C- termini of the known TALE-likes (de Lange and Wolf et al., 2015). It is therefore probable that the

available MOrTLs sequences are incomplete and fragmented. The native full length gene products could code for longer N- and/or C-termini, which would be needed for the full functionality of the proteins.
Because of the potentially incomplete sequence information, we decided to conduct follow-up experiments not with the predicted MOrTL1 and MOrTL2 proteins, but instead created chimeras containing repeats of MOrTL1 and MOrTL2 embedded in the repeat array of known TALE-Likes.

TALE-likes for biotechnological applications

Artificial dTALEs and their derived fusion proteins became one of the key tools in biology and biotechnology. With around 1000 publications since the year 2009, the interest in this technology is still high. The applications are manifold, ranging from cancer treatment to removing retro-viruses from the genome (reviewed in de Lange et al., 2014a). Nevertheless, many open questions and potential challenges regarding off-targets, affinities and specificities of TALEs remain (see page 19). The creation of new custom TALEs with different binding properties is a promising approach to generate optimized DNA binding proteins for many specific applications. A potential way to harness the advantages of both TALEs and TALE-likes and generate TALE variants with novel properties is to assemble chimeric versions containing sequences of both protein classes. An overview of the TALE like chimeras used in de Lange and Wolf et al. 2015 is shown in Figure 9.

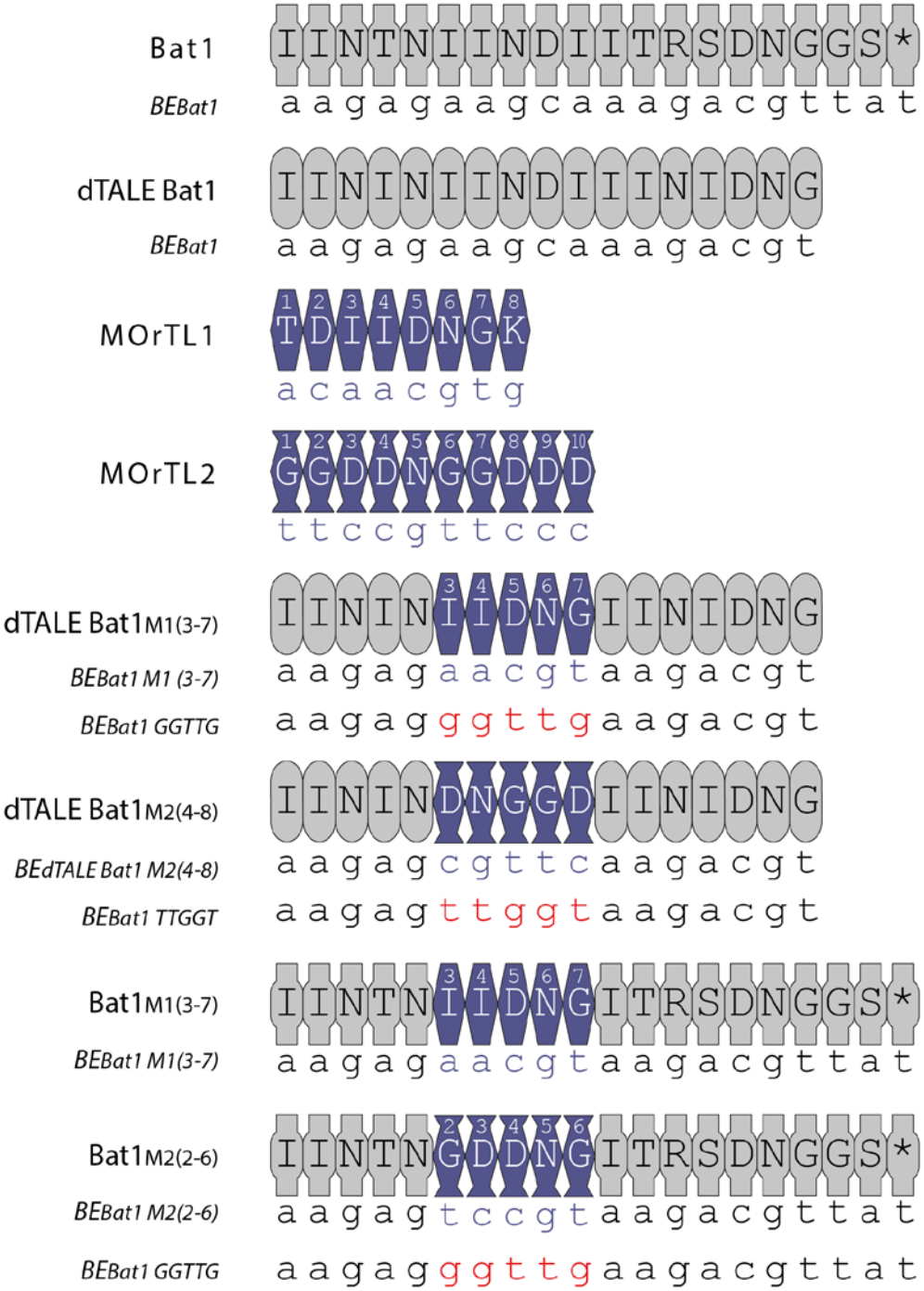


Figure 9: Overview of TALE-likes and TALE-like chimeras. The predicted binding sequence is depicted beneath every protein. dTALE Bat1 is a designer TALE with the same binding site as Bat1, but with 17 canonical repeats. dTALE $Bat1_{M1(3-7)}$ and dTALE $Bat1_{M2(4-8)}$ are based on the dTALE Bat1, but have the repeats 6-10 exchanged with MOrTL repeats (MOrTL1 repeats 3-7 and MOrTL2 repeats 4-8 respectively). $Bat1_{M1(3-7)}$ and $Bat1_{M2(2-6)}$ are based on the Bat1 protein, but repeats 6-10 are exchanged with MOrTL repeats (MOrTL1 repeats 3-7 and MOrTL2 repeats 2-6 respectively).

Repeats of TALE-likes are compatible among each other

We demonstrated that chimeric proteins combining Bats and TALEs are indeed functional. Using a T-DNA construct, DNA of TALE-Bat chimeras was delivered via *A. tumefaciens* into *Nicotiana benthamiana* and the expressed constructs successfully activated a GUS-reporter (figure 8 in de Lange and Wolf et al., 2014b).
The structure of Bats (Stella et al., 2014) and its striking similarity to TALEs can explain the high level of compatibility between TALEs and Bats, despite their overall low homology (figure 1 in de Lange and Wolf et al., 2014b). It is likely that Bat repeats in a TALE backbone can make direct contacts to the DNA despite differences in the repeat lengths.
We integrated an array of MOrTL repeats into Bat1 and TALE backbones, and analysed the chimeras' specific DNA binding properties by EMSAs, Thermophoresis and in an *in vivo* reporter system (Figure 9, figure 2-5 in de Lange and Wolf et al. 2015). All of the chimeric constructs, except for the one consisting of MOrTL2 repeats in the dTALE backbone, bound the target DNA with the expected specificity according to the RVD code of TALEs. The dTALE-Bat1$_{M2(4\text{-}8)}$ protein could also be outcompeted by an off-target DNA sequence (figure 4 (e,f) in de Lange and Wolf et al., 2015), however, this was not observed for any of the other chimeras. Additionally the affinity of the dTALE-Bat1$_{M2(4\text{-}8)}$ construct was lower (5.4 µM ± 1 µM) than of the other tested constructs and those of TALEs in general (figure 4 (d) in de Lange and Wolf et al., 2015). The affinity of dTALE chimeras (MOrTL repeats in a dTALE backbone) was lower than the affinity of Bat1 chimeras (MOrTL repeats in a Bat1 background). A reason for that could be an imperfect fit of the MOrTL repeats in the dTALE backbone, whereas the Bat backbone might provide a better structural match. The amino acid sequence of the MOrTL repeats is slightly more similar to Bat repeats than to TALE repeats (figure 1 in de Lange et al., 2015). A suboptimal interface between the repeats and the backbone could explain the observed reduced affinity. It was previously reported that in highly diverged repeats, the repeat order can influence binding specificity and measured activity (figure 6 in de Lange and Wolf et al., 2014b). This is in agreement with our observation that shuffling of Bat1 repeats in the Bat1 protein resulted in a loss of function (repeat order change) (figure 6 (a) in de Lange and Wolf et al., 2014b). It is conceivable, that the difference is caused by reduced protein stability, as a result of introducing foreign repeats into the TALE-likes. However, molecular dynamics (MD) simulations indicated that the proteins are

stable (figure 6 in de Lange and Wolf et al., 2015). Moreover, no change in the melting points between the MOrTL-Bat chimeras and the dTALE-MOrTL chimeras compared to the Bat1 and dTALE proteins could be observed, also suggesting similar levels of protein stability (table 1 in de Lange and Wolf et al., 2015).
In summary we were able to create functional chimeras with TALEs and both new TALE-like groups (Bats and MOrTLs) that could specifically bind to DNA according to the TALE-code. These chimeras can be used to create new DNA binding proteins with novel properties.

Programmability of TALE-likes

One possibility to use TALE-likes for biotechnological applications is to adapt them to bind new DNA sequences. As we showed de Lange et al., 2014b, Bat proteins have a high inter-repeat diversity, which allows the use of different reprogramming strategies, unlike for TALEs, where the sequence identity of each repeat is almost the same. We followed two designs approaches: (1) We retained the complete Bat backbone and only changed the RVDs (RVD-switch) and (2) we translocated whole repeats within the Bat proteins (Rep-switch) (figure 6 in de Lange and Wolf et al., 2014b). To reduce variables in the assay, we attempted to keep the binding sequence for the new acBats as close as possible to the predicted Bat1 DNA binding sequence; even though we changed repeats and RVDs. The activation of the constructs was not affected or at least not in a predictable pattern in comparison to the full length acBat (figure 6 in de Lange and Wolf et al., 2014b). Changing the RVDs of the repeats 7, 8 and 9 (RVD switch) (figure 6 (a) in de Lange and Wolf et al., 2014b, dBatRVDswitch2) increased the activity of the dBat in comparison to the acBat1 protein. For all the other dBatRVDswitch proteins, the activation was decreased or close to the activation observed for acBat1 protein (figure 6 (a) in de Lange and Wolf et al., 2014b). For the Rep-switch strategy none of our constructs showed a higher GFP activation as the acBat1 protein (figure 6 (b) in de Lange and Wolf et al., 2014b). The limited available permutations of the acBat1 construct (four Rep-switch and four RVD-switch constructs), however, were probably insufficient to detect a clear trend. In one Bat1 RVDswitch construct (dBatRVDswitch2) activation of the reporter was higher than in the original acBat1 construct. This indicates that the repeat scaffold is probably not co-evolved with its RVD for an optimal binding. Therefore adjustment of DNA binding properties is likely not as straightforward in

Bats as in TALEs. Previous other experiments also demonstrated (de Lange et al., 2013; Miller et al., 2015; Rogers et al., 2015) that sometimes the non RVD backbone and the position of the repeat in context of the protein can influence the binding specificity and general affinity.

We created designer Bats according to the RVD-switch and Rep-switch design strategies that bind a sequence in the promoter of the human Sox-2 gene (Zhang et al., 2011). These two dBats showed the same level of activation of the reporter as observed for the optimal sequence of the acBat1-construct. In order to compare the results, the dBat construct was designed with the same number of repeats as the corresponding dTALE construct. However, as native Bats contain more repeats than TALEs, the assembled dBat construct might not be long enough to bind the DNA efficiently. Concordantly, the truncation of repeats was correlated with a reduction of acBat activity (figure 5 (a) in de Lange and Wolf et al., 2014b) and discussed here (see page 29). A dBat construct with a higher number of repeats should probably result in increased activation of the reporter gene. Using an off-target reporter we could demonstrate sequence specific activation of the reporter gene, making it possible to design dBats to bind to new targets. This reprogrammability was also confirmed in other publications (Beurdeley et al., 2013; Juillerat et al., 2014).

MOrTLs are likely to have the same structural properties as the other TALE-likes and are predicted to form the same connections to the DNA as TALEs and Bats (figure 6 in de Lange and Wolf et al., 2015). It is therefore likely that MOrTLs are also programmable. However, the known MOrTL protein sequences will probably be insufficient for biotechnological applications, as we demonstrated that the identified MOrTLs act only as low affinity DNA binding proteins by themselves (supplementary figure 3 (b) in de Lange and Wolf et al., 2015). A valid alternative would be the use of chimeras composed of MOrTLs and TALE-like repeats (figure 2 and 4 in de Lange et al., 2015).

Bats have no zero base preference

The -1 repeat in most natural TALEs shows a preference for a thymidine base at position 0 of the binding site (Boch et al., 2009; Moscou & Bogdanove, 2009; Schreiber & Bonas, 2014). The affinity to the DNA target is mediated by a tryptophan in the -1 repeat at amino acid position 232 (Mak et al., 2012). This base preference was shown to sometimes hinder active binding of a customized dTALE to its desired

target (Römer et al., 2010), while other data indicated that if the perfect binding sequence is present (from the -1 repeat downstream), the -1 repeat can be tolerated by any nucleotide (Meckler et al., 2013). The TALE-likes from *Ralstonia*, RipTALs, have sequence specificity for guanine at the 0 position of their binding site (de Lange et al., 2013). Indiscriminate binding to all the four DNA bases was recently achieved by synthetically designed -1-repeats, which are based on the N-terminus of a *Xanthomonas* TALE (Doyle et al., 2013; Lamb et al., 2013).

While TALEs and RipTALs show different specificities at the base 0, we could not detect any statistically significant difference in the Bat1 construct in affinity measurements with different oligonucleotides altered at base 0 (figure 2 (c) in de Lange and Wolf et al., 2014b). This observation was confirmed by *in vitro* studies (Stella et al., 2014) and by an *in vivo* reporter system (Juillerat et al., 2014). It is possible that the affinity of position -1 to DNA is below the detection limit of the affinity measurement. As shown before, it is difficult to assess the affinity of the -1-repeat (Gao et al., 2012). The affinities of the N-terminus of TALEs at repeat position -1 to different nucleotide were indistinguishable, while structural evidence indicated a strong binding of thymine with the tryptophan at amino acid position 232 (Mak et al., 2012).

Nevertheless, at present, all the evidence indicates that no base preference at the repeat position -1 in the Bat proteins exists (de Lange and Wolf et al., 2014b; Juillerat et al., 2014). Therefore the Bat N-terminus could be used to create novel TALEs with a wobble position at the -1 repeat and therefore allow for flexible target site selection.

Engineering genetically stable repeats

Gene therapy with viral vectors has great potential in treating a range of human diseases like cancer and permanent viral infections. With the development of new tools for site-specific genome editing and transcriptional controlling, a viral transportation system seems to be one of the solutions to introduce DNA into the human cells to perform gene editing. TALEs with their DNA repeats are prone to recombination in nature and in retroviral vectors (Yang et al., 2005a; Yang & Gabriel, 1995a), likely because of their repetitive repeat sequence. For efficient control of genome editing it is necessary that the integrity of the TALE constructs is maintained, without loss or shuffling of repeats. One strategy to prevent recombination is the use of different synonymous codons (Yang et al., 2013). However, even with the

degenerated DNA code, it is almost impossible to generate highly diverged repeats. Another drawback is that an organism's codon usage can lead to inefficient protein production, wrongly folded proteins or even to amino acid changes in the protein (Novoa & de Pouplana, 2012; Plotkin & Kudla, 2011; Spencer & Barral, 2012). The natural variation in TALE repeats is probably insufficient to substantially increase their DNA sequence stability (about 99% sequence identity) (Boch et al., 2009; Moscou & Bogdanove, 2009). By using TALE-likes either as full length binding proteins as demonstrated for Bats (see page 35) or by using single repeats of other TALE-likes in a dTALE or Bat background, TALE-likes can contribute to the sequence variability of the repeats.. These adaptions could provide DNA sequences coding for functional DNA binding proteins that are more stable in viral assays.

Bats are more compact and stable than TALEs

Bats were proven to be fully functional as full length DNA binding proteins (de Lange and Wolf et al., 2014b; Juillerat et al., 2014; Stella et al., 2014). Because of their smaller size compared to TALEs (figure 1 in de Lange and Wolf et al., 2014b), they could be a preferred option for many medical and biotechnological applications. As a general rule, smaller proteins are translated more efficiently with fewer failures in their amino acid code. Additionally, shorter mRNAs are also transcribed more quickly. Moreover, only small proteins and other biomolecules are able to effectively penetrate membranes and therefore lead to high bioavailability (Gupta et al., 2005).
We demonstrated that Bat proteins are less temperature sensitive than TALEs. In an *in vitro* experiment, the measured melting point of Bat1 was around 44°C, while for the dTALE construct and its derivates it was around 30°C (table 1 in de Lange and Wolf et al., 2015). For medical applications in the human body under physiological conditions, this could negatively impact their usefulness. An assay in human cells demonstrated that cells grown at 37°C expressing TALENs showed decreased activity compared to cells grown at 20°C (Miller et al., 2015). A reason for the increased stability of the Bat1 protein could be the presence of two positive amino acid stripes along the DNA backbone (Stella et al., 2014) in comparison to one stripe in TALEs. Bat1 also dissociates two times slower as AvrBs3 from its DNA (Stella et al., 2014).

It is possible that TALEs may be more stable under native conditions. However, based on the experimental evidence, Bats could still offer an overall advantage in stability.

Outlook

TALE-likes offer great potential to improve the generation of customized and optimized DNA binding proteins for biotechnological applications. To that end, three different strategies can be employed:

(1) Using full length proteins such as dBats (see page 35) as new binding domains. dBats could offer a higher protein stability, as discussed on page 38.

(2) Creating repeat chimeras consisting of TALE and TALE-like repeats. Repeats of TALE-likes exhibit a lower repeat sequence identity compared to TALEs. Therefore, working chimeras (de Lange and Wolf et al., 2014b, de Lange and Wolf et al., 2015, discussed on page 34) could be viable options to increase the sequence diversity of repeats without loss of specificity, which is necessary for gene targeting. The assembly of chimeras is straightforward, as all the new TALE-likes bind the DNA according to the known TALE-code. By creating MOrTL1 and Bat1 chimeras, we could demonstrate that TALE-likes can have similar affinities and temperature stability.

(3) Integrating single amino acid polymorphisms into TALEs. TALE-like repeat polymorphisms could be used as a template to create new functional repeats by exchanging single amino acids in existing TALE repeats. One major advantage with this strategy is the avoidance of repetitive and therefore potentially unstable DNA sequences (see page 37). Moreover, potentially tuneable TALEs with new properties, including different affinities, kinetics and base specificities stemming from different repeat backbones can be created. Adjustment of these properties could be used to fine tune gene expression and to assemble more elaborate gene cascades.

The need to improve the current PDPs is there: Off-targets, bioavailability (protein size) and protein stability are some of the major challenges that have to be tackled before therapeutic genome editing will be feasible and save in humans using TALEs. Random testing of all possible combinations of amino acids in repeats is tedious. Our work on TALE-likes offers a starting point by providing a set of compatible, yet diverse repeat elements that can be used to improve programmable DNA binding proteins.

References

Aurell, E., d'Herouel, A. F., Malmnas, C., & Vergassola, M. (2007). Transcription factor concentrations versus binding site affinities in the yeast *S. cerevisiae*. *Phys Biol, 4*(2), 134-143. doi:10.1088/1478-3975/4/2/006

Ballvora, A., Pierre, M., van den Ackerveken, G., Schornack, S., Rossier, O., Ganal, M., Lahaye, T., & Bonas, U. (2001). Genetic mapping and functional analysis of the tomato Bs4 locus governing recognition of the *Xanthomonas campestris pv. vesicatoria* AvrBs4 protein. *Mol Plant Microbe Interact, 14*(5), 629-638. doi:10.1094/MPMI.2001.14.5.629

Beumer, K., Bhattacharyya, G., Bibikova, M., Trautman, J. K., & Carroll, D. (2006). Efficient gene targeting in *Drosophila* with zinc-finger nucleases. *Genetics, 172*(4), 2391-2403. doi:10.1534/genetics.105.052829

Beurdeley, M., Bietz, F., Li, J., Thomas, S., Stoddard, T., Juillerat, A., Zhang, F., Voytas, D. F., Duchateau, P., & Silva, G. H. (2013). Compact designer TALENs for efficient genome engineering. *Nature communications, 4*, 1762. doi:10.1038/ncomms2782

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., & Phillips, R. (2005). Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev, 15*(2), 125-135. doi:10.1016/j.gde.2005.02.006

Bitinaite, J., Wah, D. A., Aggarwal, A. K., & Schildkraut, I. (1998). FokI dimerization is required for DNA cleavage. *Proceedings of the National Academy of Sciences of the United States of America, 95*(18), 10570-10575. doi:DOI 10.1073/pnas.95.18.10570

Blount, B. A., Weenink, T., Vasylechko, S., & Ellis, T. (2012). Rational diversification of a promoter providing fine-tuned expression and orthogonal regulation for synthetic biology. *PloS one, 7*(3), e33279. doi:10.1371/journal.pone.0033279

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., & Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. *Science, 326*(5959), 1509-1512. doi:10.1126/science.1178811

Bogdanove, A. J., Schornack, S., & Lahaye, T. (2010). TAL effectors: finding plant genes for disease and defense. *Current opinion in plant biology, 13*(4), 394-401. doi:10.1016/j.pbi.2010.04.010

Bonas, U., Stall, R. E., & Staskawicz, B. (1989). Genetic and structural characterization of the avirulence gene *avrBs3* from *Xanthomonas campestris pv. vesicatoria*. *Mol Gen Genet, 218*(1), 127-136.

Briggs, A. W., Rios, X., Chari, R., Yang, L. H., Zhang, F., Mali, P., & Church, G. M. (2012). Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. *Nucleic acids research, 40*(15). doi:10.1093/nar/gks624

Cai, Y., Bak, R. O., & Mikkelsen, J. G. (2014). Targeted genome editing by lentiviral protein transduction of zinc-finger and TAL-effector nucleases. *Elife, 3*, e01911. doi:10.7554/eLife.01911

Carroll, D. (2011). Genome engineering with zinc-finger nucleases. *Genetics, 188*(4), 773-782. doi:10.1534/genetics.111.131433

Cermak, T., Doyle, E. L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J. A., Somia, N. V., Bogdanove, A. J., & Voytas, D. F. (2011). Efficient design

and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic acids research, 39*(12), e82. doi:10.1093/nar/gkr218

Chang, C. W., Counago, R. M., Williams, S. J., Boden, M., & Kobe, B. (2013). Distinctive conformation of minor site-specific nuclear localization signals bound to importin-alpha. *Traffic, 14*(11), 1144-1154. doi:10.1111/tra.12098

Chu, Z. H., Yuan, M., Yao, L. L., Ge, X. J., Yuan, B., Xu, C. G., Li, X. H., Fu, B. Y., Li, Z. K., Bennetzen, J. L., Zhang, Q. F., & Wang, S. P. (2006). Promoter mutations of an essential gene for pollen development result in disease resistance in rice. *Gene Dev, 20*(10), 1250-1255. doi:DOI 10.1101/gad.1416306

Cong, L., Zhou, R., Kuo, Y. C., Cunniff, M., & Zhang, F. (2012). Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature communications, 3*, 968. doi:10.1038/ncomms1962

Cuculis, L., Abil, Z., Zhao, H., & Schroeder, C. M. (2015). Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nature communications, 6*, 7277. doi:10.1038/ncomms8277

de Lange, O., Binder, A., & Lahaye, T. (2014a). From dead leaf, to new life: TAL effectors as tools for synthetic biology. *The Plant journal : for cell and molecular biology*. doi:10.1111/tpj.12431

de Lange, O., Schreiber, T., Schandry, N., Radeck, J., Braun, K. H., Koszinowski, J., Heuer, H., Strauss, A., & 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*(3), 773-786. doi:10.1111/nph.12324

de Lange, O., Wolf, C., Dietze, J., Elsaesser, J., Morbitzer, R., & Lahaye, T. (2014b). Programmable DNA-binding proteins from *Burkholderia* provide a fresh perspective on the TALE-like repeat domain. *Nucleic acids research, 42*(11), 7436-7449. doi:10.1093/nar/gku329

de Lange, O., Wolf, C., Thiel, P., Kruger, J., Kleusch, C., Kohlbacher, O., & Lahaye, T. (2015). DNA-binding proteins from marine bacteria expand the known sequence diversity of TALE-like repeats. *Nucleic acids research, 43*(20), 10065-10080. doi:10.1093/nar/gkv1053

Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J. K., Shi, Y., & Yan, N. (2012). Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science, 335*(6069), 720-723. doi:10.1126/science.1215670

Doudna, J. A. and E. Charpentier (2014). "Genome editing. The new frontier of genome engineering with CRISPR-Cas9." Science 346(6213): 1258096.

Doyle, E. L., Hummel, A. W., Demorest, Z. L., Starker, C. G., Voytas, D. F., Bradley, P., & Bogdanove, A. J. (2013). TAL effector specificity for base 0 of the DNA target is altered in a complex, effector- and assay-dependent manner by substitutions for the tryptophan in cryptic repeat -1. *PloS one, 8*(12). doi:10.1371/journal.pone.0082120

Enninga, J., Mounier, J., Sansonetti, P., & Van Nhieu, G. T. (2005). Secretion of type III effectors into host cells in real time. *Nature Methods, 2*(12), 959-965. doi:10.1038/Nmeth804

Fairall, L., Schwabe, J. W. R., Chapman, L., Finch, J. T., & Rhodes, D. (1993). The Crystal-Structure of a 2 Zinc-Finger Peptide Reveals an Extension to the Rules for Zinc-Finger DNA Recognition. *Nature, 366*(6454), 483-487. doi:DOI 10.1038/366483a0

Fu, Y. F., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., & Sander, J. D. (2013). High-frequency off-target mutagenesis induced by

CRISPR-Cas nucleases in human cells. *Nature biotechnology, 31*(9), 822-+. doi:10.1038/nbt.2623

Gabriel, R., Lombardo, A., Arens, A., Miller, J. C., Genovese, P., Kaeppel, C., Nowrouzi, A., Bartholomae, C. C., Wang, J., Friedman, G., Holmes, M. C., Gregory, P. D., Glimm, H., Schmidt, M., Naldini, L., & von Kalle, C. (2011). An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nature biotechnology, 29*(9), 816-823. doi:10.1038/nbt.1948

Gao, H., Wu, X., Chai, J., & Han, Z. (2012). Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell research, 22*(12), 1716-1720. doi:10.1038/cr.2012.156

Gommans, W. M., Haisma, H. J., & Rots, M. G. (2005). Engineering zinc finger protein transcription factors: The therapeutic relevance of switching endogenous gene expression on or off at command. *Journal of molecular biology, 354*(3), 507-519. doi:10.1016/j.jmb.2005.06.082

Gordan, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., & Bulyk, M. L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell reports, 3*(4), 1093-1104. doi:10.1016/j.celrep.2013.03.014

Gu, K., Yang, B., Tian, D., Wu, L., Wang, D., Sreekala, C., Yang, F., Chu, Z., Wang, G. L., White, F. F., & Yin, Z. (2005). R gene expression induced by a type-III effector triggers disease resistance in rice. *Nature, 435*(7045), 1122-1125. doi:10.1038/nature03630

Gupta, B., Levchenko, T. S., & Torchilin, V. P. (2005). Intracellular delivery of large molecules and small particles by cell-penetrating proteins and peptides. *Advanced drug delivery reviews, 57*(4), 637-651. doi:10.1016/j.addr.2004.10.007

Guralnick, B., Thomsen, G., & Citovsky, V. (1996). Transport of DNA into the nuclei of *Xenopus* oocytes by a modified VirE2 protein of *Agrobacterium*. *The Plant cell, 8*(3), 363-373.

Holkers, M., Maggio, I., Liu, J., Janssen, J. M., Miselli, F., Mussolino, C., Recchia, A., Cathomen, T., & Goncalves, M. A. (2013). Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic acids research, 41*(5), e63. doi:10.1093/nar/gks1446

Hopkins, C. M., White, F. F., Choi, S. H., Guo, A., & Leach, J. E. (1992). Identification of a family of avirulence genes from *Xanthomonas oryzae pv. oryzae*. *Mol Plant Microbe Interact, 5*(6), 451-459.

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y. Q., Fine, E. J., Wu, X. B., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G., & Zhang, F. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology, 31*(9). doi:10.1038/nbt.2647

Hu, Y., Zhang, J. L., Jia, H. G., Sosso, D., Li, T., Frommer, W. B., Yang, B., White, F. F., Wang, N. A., & Jones, J. B. (2014). Lateral organ boundaries 1 is a disease susceptibility gene for citrus bacterial canker disease. *Proceedings of the National Academy of Sciences of the United States of America, 111*(4), E521-E529. doi:10.1073/pnas.1313271111

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science, 337*(6096), 816-821. doi:10.1126/science.1225829

Juillerat, A., Bertonati, C., Dubois, G., Guyot, V., Thomas, S., Valton, J., Beurdeley, M., Silva, G. H., Daboussi, F., & Duchateau, P. (2014). BurrH: a new modular

DNA binding protein for genome engineering. *Scientific reports, 4*. doi:ARTN 3831
10.1038/srep03831
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*(5850), 648-651. doi:10.1126/science.1144956
Kim, E., Kim, S., Kim, D. H., Choi, B. S., Choi, I. Y., & Kim, J. S. (2012). Precision genome engineering with programmable DNA-nicking enzymes. *Genome research, 22*(7), 1327-1333. doi:10.1101/gr.138792.112
Kim, J. S., Lee, H. J., & Carroll, D. (2010). Genome editing with modularly assembled zinc-finger nucleases. *Nature Methods, 7*(2), 91. doi:10.1038/nmeth0210-91a
Kim, Y. G., Cha, J., & Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proceedings of the National Academy of Sciences of the United States of America, 93*(3), 1156-1160.
Koonin, E. V., & Galperin, M. Y. (2003). *Sequence - evolution - function : computational approaches in comparative genomics*. Boston: Kluwer Academic.
Lackner, G., Moebius, N., Partida-Martinez, L., & Hertweck, C. (2011a). Complete genome sequence of B*urkholderia rhizoxinica*, an Endosymbiont of *Rhizopus microsporus*. *Journal of bacteriology, 193*(3), 783-784. doi:10.1128/JB.01318-10
Lackner, G., Moebius, N., Partida-Martinez, L. P., Boland, S., & Hertweck, C. (2011b). Evolution of an endofungal lifestyle: Deductions from the *Burkholderia rhizoxinica* genome. *BMC Genomics, 12*. doi:10.1186/1471-2164-12-210
Lackner, G., Partida-Martinez, L. P., & Hertweck, C. (2009). Endofungal bacteria as producers of mycotoxins. *Trends in microbiology, 17*(12), 570-576. doi:10.1016/j.tim.2009.09.003
Lamb, B. M., Mercer, A. C., & Barbas, C. F., 3rd. (2013). Directed evolution of the TALE N-terminal domain for recognition of all 5' bases. *Nucleic acids research, 41*(21), 9779-9785. doi:10.1093/nar/gkt754
Lassner, M. W., Jones, A., Daubert, S., & Comai, L. (1991). Targeting of T7 RNA polymerase to tobacco nuclei mediated by an SV40 nuclear location signal. *Plant molecular biology, 17*(2), 229-234.
Leyns, F., De Cleene, M., Swings, J.-G., & De Ley, J. (1984). The Host Range of the Genus *Xanthomonas*. *The Botanical Review, 50*(3), 308-356. doi:10.1007/Bf02862635
Li, L., Atef, A., Piatek, A., Ali, Z., Piatek, M., Aouida, M., Sharakuu, A., Mahjoub, A., Wang, G., Khan, S., Fedoroff, N. V., Zhu, J. K., & Mahfouz, M. M. (2013). Characterization and DNA-binding specificities of *Ralstonia* TAL-like effectors. *Molecular Plant, 6*(4), 1318-1330. doi:10.1093/mp/sst006
Lin, J., Chen, H., Luo, L., Lai, Y., Xie, W., & Kee, K. (2015). Creating a monomeric endonuclease TALE-I-SceI with high specificity and low genotoxicity in human cells. *Nucleic acids research, 43*(2), 1112-1122. doi:10.1093/nar/gku1339
Luo, K. X., Stroschein, S. L., Wang, W., Chen, D., Martens, E., Zhou, S., & Zhou, Q. (1999). The Ski oncoprotein interacts with the Smad proteins to repress TGF beta signaling. *Gene Dev, 13*(17), 2196-2206. doi:DOI 10.1101/gad.13.17.2196
Mak, A. N., Bradley, P., Cernadas, R. A., Bogdanove, A. J., & Stoddard, B. L. (2012). The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science, 335*(6069), 716-719. doi:10.1126/science.1216211

Mali, P., Aach, J., Stranges, P. B., Esvelt, K. M., Moosburner, M., Kosuri, S., Yang, L., & Church, G. M. (2013). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology, 31*(9), 833-838. doi:10.1038/nbt.2675

Marcovitz, A., Naftaly, A., & Levy, Y. (2015). Water organization between oppositely charged surfaces: Implications for protein sliding along DNA. *J Chem Phys, 142*(8). doi:10.1063/1.4913370

Marois, E., Van den Ackerveken, G., & Bonas, U. (2002). The *Xanthomonas* type III effector protein AvrBs3 modulates plant gene expression and induces cell hypertrophy in the susceptible host. *Mol Plant Microbe Interact, 15*(7), 637-646. doi:10.1094/MPMI.2002.15.7.637

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., Lahaye, T., Segal, D. J., & Baldwin, E. P. (2013). Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic acids research, 41*(7), 4118-4128. doi:10.1093/nar/gkt085

Mendenhall, E. M., Williamson, K. E., Reyon, D., Zou, J. Y., Ram, O., Joung, J. K., & Bernstein, B. E. (2013). Locus-specific editing of histone modifications at endogenous enhancers. *Nature biotechnology, 31*(12), 1133. doi:10.1038/nbt.2701

Method of the Year 2011. (2012). *Nature Methods, 9*(1). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22312634

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., Paschon, D. E., Vincent, A. I., Dulay, G. P., Barlow, K. A., Shivak, D. A., Leung, E., Kim, J. D., Amora, R., Urnov, F. D., Gregory, P. D., & Rebar, E. J. (2015). Improved specificity of TALE-based genome editing using an expanded RVD repertoire. *Nature Methods, 12*(5), 465-471. doi:10.1038/nmeth.3330

Mills, E., Baruch, K., Charpentier, X., Kobi, S., & Rosenshine, I. (2008). Real-time analysis of effector translocation by the type III secretion system of enteropathogenic Escherichia coli. *Cell host & microbe, 3*(2), 104-113. doi:10.1016/j.chom.2007.11.007

Morbitzer, R., Elsaesser, J., Hausner, J., & Lahaye, T. (2011). Assembly of custom TALE-type DNA binding domains by modular cloning. *Nucleic acids research, 39*(13), 5790-5799. doi:10.1093/nar/gkr151

Moscou, M. J., & Bogdanove, A. J. (2009). A simple cipher governs DNA recognition by TAL effectors. *Science, 326*(5959), 1501. doi:10.1126/science.1178817

Mukaihara, T., & Tamura, N. (2009). Identification of novel *Ralstonia solanacearum* type III effector proteins through translocation analysis of hrpB-regulated gene products. *Microbiol-Sgm, 155*, 2235-2244. doi:10.1099/mic.0.027763-0

Mussolino, C., Morbitzer, R., Lutge, F., Dannemann, N., Lahaye, T., & Cathomen, T. (2011). A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic acids research, 39*(21), 9283-9293. doi:10.1093/nar/gkr597

Novoa, E. M., & de Pouplana, L. R. (2012). Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet, 28*(11), 574-581. doi:10.1016/j.tig.2012.07.006

O'Connell, M. R., Oakes, B. L., Sternberg, S. H., East-Seletsky, A., Kaplan, M., & Doudna, J. A. (2014). Programmable RNA recognition and cleavage by CRISPR/Cas9. *Nature, 516*(7530), 263-266. doi:10.1038/nature13769

Pabo, C. O., Peisach, E., & Grant, R. A. (2001). Design and selection of novel Cys2His2 zinc finger proteins. *Annual Reviews Biochemistry, 70*, 313-340. doi:10.1146/annurev.biochem.70.1.313

Papworth, M., Kolasinska, P., & Minczuk, M. (2006). Designer zinc-finger proteins and their applications. *Gene, 366*(1), 27-38. doi:10.1016/j.gene.2005.09.011

Pavletich, N. P., & Pabo, C. O. (1991). Zinc Finger DNA Recognition - Crystal-Structure of a Zif268-DNA Complex at 2.1-A. *Science, 252*(5007), 809-817. doi:DOI 10.1126/science.2028256

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics, Chapter 3*, Unit3 1. doi:10.1002/0471250953.bi0301s42

Perez, E. E., Wang, J. B., Miller, J. C., Jouvenot, Y., Kim, K. A., Liu, O., Wang, N., Lee, G., Bartsevich, V. V., Lee, Y. L., Guschin, D. Y., Rupniewski, I., Waite, A. J., Carpenito, C., Carroll, R. G., Orange, J. S., Urnov, F. D., Rebar, E. J., Ando, D., Gregory, P. D., Riley, J. L., Holmes, M. C., & June, C. H. (2008). Establishment of HIV-1 resistance in CD4(+) T cells by genome editing using zinc-finger nucleases. *Nature biotechnology, 26*(7), 808-816. doi:10.1038/nbt1410

Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics, 12*(1), 32-42. doi:10.1038/nrg2899

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell, 152*(5), 1173-1183. doi:10.1016/j.cell.2013.02.022

Rainbow, L., Hart, C. A., & Winstanley, C. (2002). Distribution of type III secretion gene clusters in *Burkholderia pseudomallei, B.thailandensis* and *B.mallei*. *J Med Microbiol, 51*(5), 374-384.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature protocols, 8*(11), 2281-2308. doi:10.1038/nprot.2013.143

Reyon, D., Tsai, S. Q., Khayter, C., Foden, J. A., Sander, J. D., & Joung, J. K. (2012). FLASH assembly of TALENs for high-throughput genome editing. *Nature biotechnology, 30*(5), 460-465. doi:10.1038/nbt.2170

Rhee, Y., Gurel, F., Gafni, Y., Dingwall, C., & Citovsky, V. (2000). A genetic system for detection of protein nuclear import and export. *Nature biotechnology, 18*(4), 433-437.

Rogers, J. M., Barrera, L. A., Reyon, D., Sander, J. D., Kellis, M., Joung, J. K., & Bulyk, M. L. (2015). Context influences on TALE-DNA binding revealed by quantitative profiling. *Nature communications, 6*, 7440. doi:10.1038/ncomms8440

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., & Honig, B. (2009). The role of DNA shape in protein-DNA recognition. *Nature, 461*(7268), 1248-1253. doi:10.1038/nature08473

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*(5850), 645-648. doi:10.1126/science.1144958

Römer, P., Recht, S., & Lahaye, T. (2009). A single plant resistance gene promoter engineered to recognize multiple TAL effectors from disparate pathogens. *Proceedings of the National Academy of Sciences of the United States of America, 106*(48), 20526-20531. doi:10.1073/pnas.0908812106

Römer, P., Recht, S., Strauss, T., Elsaesser, J., Schornack, S., Boch, J., Wang, S., & Lahaye, T. (2010). Promoter elements of rice susceptibility genes are bound and activated by specific TAL effectors from the bacterial blight pathogen, *Xanthomonas oryzae pv. oryzae*. *New Phytol, 187*(4), 1048-1057. doi:10.1111/j.1469-8137.2010.03217.x

Rossier, O., Wengelnik, K., Hahn, K., & Bonas, U. (1999). The *Xanthomonas* Hrp type III system secretes proteins from plant and mammalian bacterial pathogens. *Proceedings of the National Academy of Sciences of the United States of America, 96*(16), 9368-9373. doi:10.1073/pnas.96.16.9368

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D. Y., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J. E., Li, K., Kravitz, S., Heidelberg, J. F., Utterback, T., Rogers, Y. H., Falcon, L. I., Souza, V., Bonilla-Rosso, G., Eguiarte, L. E., Karl, D. M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M. R., Strausberg, R. L., Nealson, K., Friedman, R., Frazier, M., & Venter, J. C. (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *Plos Biol, 5*(3), 398-431. doi:10.1371/journal.pbio.0050077

Samudrala, R., Heffron, F., & McDermott, J. E. (2009). Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems. *PLoS pathogens, 5*(4), e1000375. doi:10.1371/journal.ppat.1000375

Schornack, S., Ballvora, A., Gurlebeck, D., Peart, J., Baulcombe, D., Ganal, M., Baker, B., Bonas, U., & Lahaye, T. (2004). The tomato resistance protein *Bs4* is a predicted non-nuclear TIR-NB-LRR protein that mediates defense responses to severly truncated derivatives of AvrBs4 and overexpressed AvrBs3 (vol 37, pg 46, 2004). *Plant Journal, 37*(5), 787-787. doi:10.1111/j.1365-313X.2004.02045.x

Schornack, S., Moscou, M. J., Ward, E. R., & Horvath, D. M. (2013). Engineering plant disease resistance based on TAL effectors. *Annual Reviews Phytopathology, 51*, 383-406. doi:10.1146/annurev-phyto-082712-102255

Schreiber, T., & Bonas, U. (2014). Repeat -1 of TAL effectors affects target specificity for the base at position zero. *Nucleic acids research, 42*(11), 7160-7169. doi:10.1093/nar/gku341

Schreiber, T., Sorgatz, A., List, F., Bluher, D., Thieme, S., Wilmanns, M., & Bonas, U. (2015). Refined requirements for protein regions important for activity of the TALE AvrBs3. *PloS one, 10*(3), e0120214. doi:10.1371/journal.pone.0120214

Silipo, A., Leone, M. R., Lanzetta, R., Parrilli, M., Lackner, G., Busch, B., Hertweck, C., & Molinaro, A. (2012). Structural characterization of two lipopolysaccharide O-antigens produced by the endofungal bacterium *Burkholderia sp*. HKI-402 (B4). *Carbohydrate Research, 347*(1), 95-98. doi:10.1016/j.carres.2011.10.038

Sorek, R., Kunin, V., & Hugenholtz, P. (2008). CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol, 6*(3), 181-186. doi:10.1038/nrmicro1793

Spencer, P. S., & Barral, J. M. (2012). Genetic code redundancy and its influence on the encoded polypeptides. *Computational and Structural Biotechnology Journal, 1*(1). doi:10.5936/csbj.201204006

Stella, S., Molina, R., Lopez-Mendez, B., Juillerat, A., Bertonati, C., Daboussi, F., Campos-Olivas, R., Duchateau, P., & Montoya, G. (2014). BuD, a helix-loop-

helix DNA-binding domain for genome modification. *Acta Crystallogr D, 70*, 2042-2052. doi:10.1107/S1399004714011183

Stormo, G. D. (1998). Information content and free energy in DNA--protein interactions. *Journal of theoretical biology, 195*(1), 135-137. doi:10.1006/jtbi.1998.0785

Stormo, G. D., & Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends in biochemical sciences, 23*(3), 109-113.

Strauss, T., van Poecke, R. M., Strauss, A., Romer, P., Minsavage, G. V., Singh, S., Wolf, C., Strauss, A., Kim, S., Lee, H. A., Yeom, S. I., Parniske, M., Stall, R. E., Jones, J. B., Choi, D., Prins, M., & Lahaye, T. (2012). RNA-seq pinpoints a Xanthomonas TAL-effector activated resistance gene in a large-crop genome. *Proceedings of the National Academy of Sciences of the United States of America, 109*(47), 19480-19485. doi:10.1073/pnas.1212415109

Streubel, J., Blucher, C., Landgraf, A., & Boch, J. (2012). TAL effector RVD specificities and efficiencies. *Nature biotechnology, 30*(7), 593-595. doi:10.1038/nbt.2304

Sugio, A., Yang, B., Zhu, T., & White, F. F. (2007). Two type III effector genes of *Xanthomonas oryzae pv. oryzae* control the induction of the host genes OsTFIIAgamma1 and OsTFX1 during bacterial blight of rice. *Proceedings of the National Academy of Sciences of the United States of America, 104*(25), 10720-10725. doi:10.1073/pnas.0701742104

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

Urnov, F. D., Miller, J. C., Lee, Y. L., Beausejour, C. M., Rock, J. M., Augustus, S., Jamieson, A. C., Porteus, M. H., Gregory, P. D., & Holmes, M. C. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature, 435*(7042), 646-651. doi:10.1038/nature03556

Van den Ackerveken, G., Marois, E., & Bonas, U. (1996). Recognition of the bacterial avirulence protein AvrBs3 occurs inside the host plant cell. *Cell, 87*(7), 1307-1316. doi:Doi 10.1016/S0092-8674(00)81825-5

Vanderkrol, A. R., & Chua, N. H. (1991). The Basic Domain of Plant B-Zip Proteins Facilitates Import of a Reporter Protein into Plant Nuclei. *The Plant cell, 3*(7), 667-675.

Wagstaff, K. M., & Jans, D. A. (2009). Importins and beyond: non-conventional nuclear transport mechanisms. *Traffic, 10*(9), 1188-1198. doi:10.1111/j.1600-0854.2009.00937.x

West, S. M., Rohs, R., Mann, R. S., & Honig, B. (2010). Electrostatic Interactions between Arginines and the Minor Groove in the Nucleosome. *J Biomol Struct Dyn, 27*(6), 861-866.

Wolfe, S. A., Nekludova, L., & Pabo, C. O. (2000). DNA recognition by Cys2His2 zinc finger proteins. *Annual Reviews Biophysics and Biomolecular Structure, 29*, 183-212. doi:10.1146/annurev.biophys.29.1.183

Wu, X., Scott, D. A., Kriz, A. J., Chiu, A. C., Hsu, P. D., Dadon, D. B., Cheng, A. W., Trevino, A. E., Konermann, S., Chen, S., Jaenisch, R., Zhang, F., & Sharp, P. A. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature biotechnology*. doi:10.1038/nbt.2889

Yang, B., Sugio, A., & White, F. F. (2005a). 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 In, 18*(2), 142-149. doi:10.1094/Mpmi-18-0142

Yang, B., Sugio, A., & White, F. F. (2005b). 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*(2), 142-149. doi:10.1094/MPMI-18-0142

Yang, B., Zhu, W., Johnson, L. B., & White, F. F. (2000). The virulence factor AvrXa7 of *Xanthomonas oryzae pv. oryzae* is a type III secretion pathway-dependent nuclear-localized double-stranded DNA-binding protein. *Proceedings of the National Academy of Sciences of the United States of America, 97*(17), 9807-9812. doi:10.1073/pnas.170286897

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., Huang, P. Y., Daley, G., & Church, G. (2013). Optimization of scarless human stem cell genome editing. *Nucleic acids research, 41*(19), 9049-9061. doi:10.1093/nar/gkt555

Yang, Y., & Gabriel, D. W. (1995a). Intragenic recombination of a single plant pathogen gene provides a mechanism for the evolution of new host specificities. *Journal of bacteriology, 177*(17), 4963-4968.

Yang, Y., & Gabriel, D. W. (1995b). *Xanthomonas* avirulence/pathogenicity gene family encodes functional plant nuclear targeting signals. *Mol Plant Microbe Interact, 8*(4), 627-631.

Zhang, F., Cong, L., Lodato, S., Kosuri, S., Church, G. M., & Arlotta, P. (2011). Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology, 29*(2), 149-U190. doi:10.1038/nbt.1775

Zhou, J. H., Peng, Z., Long, J. Y., Sosso, D., Liu, B., Eom, J. S., Huang, S., Liu, S. Z., Cruz, C. V., Frommer, W. B., White, F. F., & Yang, B. (2015). Gene targeting by the TAL effector PthXo2 reveals cryptic resistance gene for bacterial blight of rice. *Plant Journal, 82*(4), 632-643. doi:10.1111/tpj.12838

Zhu, W., Yang, B., Chittoor, J. M., Johnson, L. B., & White, F. F. (1998). AvrXa10 contains an acidic transcriptional activation domain in the functionally conserved C terminus. *Mol Plant Microbe Interact, 11*(8), 824-832. doi:10.1094/MPMI.1998.11.8.824

Danksagung

Zuerst möchte ich mich bei Herrn Prof. Dr. Thomas Lahaye für die Möglichkeit bedanken, meine Doktorarbeit in seiner Arbeitsgruppe durchzuführen, sowie für die Förderung und Unterstützung über die Jahre, Begutachtung meiner Arbeit und die Möglichkeit wissenschaftliche Selbstständigkeit zu erlangen.
Bedanken möchte ich mich auch bei Prof. Dr. Ulrike Zentgraf, für die Erstellung des Zweitgutachtens.
Innerhalb meiner Arbeitsgruppe möchte ich mich besonders bei Orlando de Lange bedanken, einem der intelligentesten, nettesten und talentiertesten Menschen, dem ich jemals begegnet bin. Es war mir eine große Freude mit ihm an den MOrTLs und Bats zu forschen. Er war eine Inspiration für mich.
Vielen Dank auch an Niklas Schandry, Christina Krönauer, Patrizia Ricca und Dousheng Wu für die wissenschaftliche Diskussion, die Unterstützung während meiner Doktorarbeit und die angenehme Zeit, die ich bei ihnen verbringen durfte.
Auch möchte ich mich bei meinen studentischen Hilfskräften, Bachelorstudenten und Praktikanten Sebastian, Daniel, Kipras, Jörn und Moritz für die zuverlässige Unterstützung bedanken.
Vielen Dank an meine Familie, die mich immer unterstützt.
Meinem Mann Andi danke ich für das Korrekturlesen meiner Arbeit, seine andauernde Unterstützung, Verständnis und Liebe. Vielen Dank, dass es Dich gibt.

Appendix – Publications

Publications presented in this thesis:

1. de Lange, Orlando*, **Wolf, Christina*****, Dietze, Jörn, Elsaesser, Janett, Morbitzer, Robert & Lahaye, Thomas
   Programmable DNA-binding proteins from *Burkholderia* provide a fresh perspectiveon the TALE-like repeat domain.
   Nucleic Acids Research (2014) **42** (11): 7436-7449.

2. 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 (2015) **43** (20): 10065-10080.

* joint first authorship

These documents are presented as in the original publications, without additional interrupting pages.

7436–7449 *Nucleic Acids Research, 2014, Vol. 42, No. 11*
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

*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.

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 $dTALE_{Bat1mimic}$ and $dTALE_{SOX2}$ were created using a previously described method (15). The assembly of $dTALE_{Bat1mimic}$ 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.

$dTALE_{UPT\ AvrBs3\ 3x\ Bat1\ rep2/6/8\ /17}$ were created as previously described (10) with trimers synthesized by GenScript with the sequences listed in Supplementary Figure S14, while $dTALE_{UPT\ AvrBs3\ 3x\ NI/NN/NG}$ 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).

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, BE_{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

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.

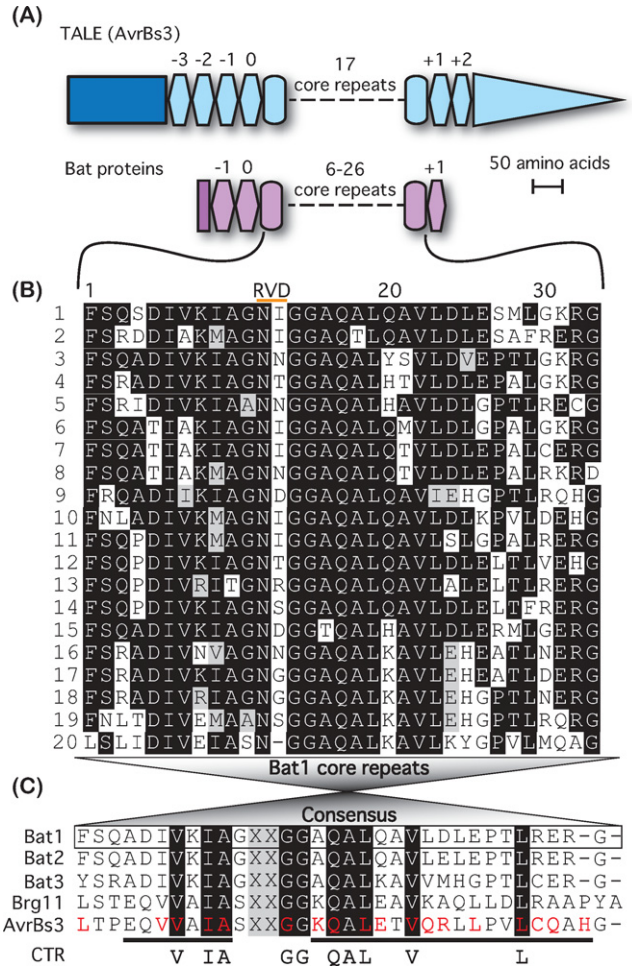


Figure 1. Sequence-based comparison of TALE-like proteins. (**A**) Comparison of TALE (AvrBs3) and Bat architecture. The lengths of all domains are drawn to the indicated scale, except the dashes representing core repeats. TALE domains are shown in blue and Bat domains in purple. Rectangles indicate 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-

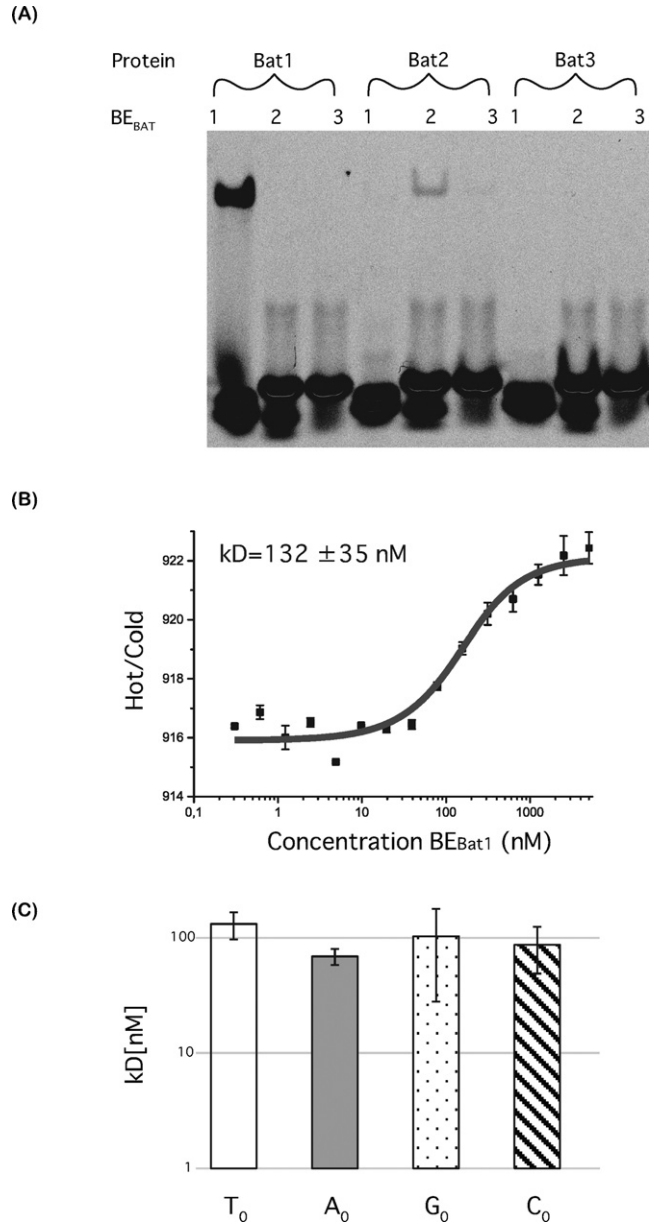


Figure 2. *In vitro* interaction studies of Bat proteins with predicted DNA targets. (**A**) Electrophoretic mobility shift assays were carried out for Bat1, 2 and 3 using 5'Cy5 labelled double-stranded DNA, bearing target sequences deduced from the TALE code. Each protein (100 nM) was tested against 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).

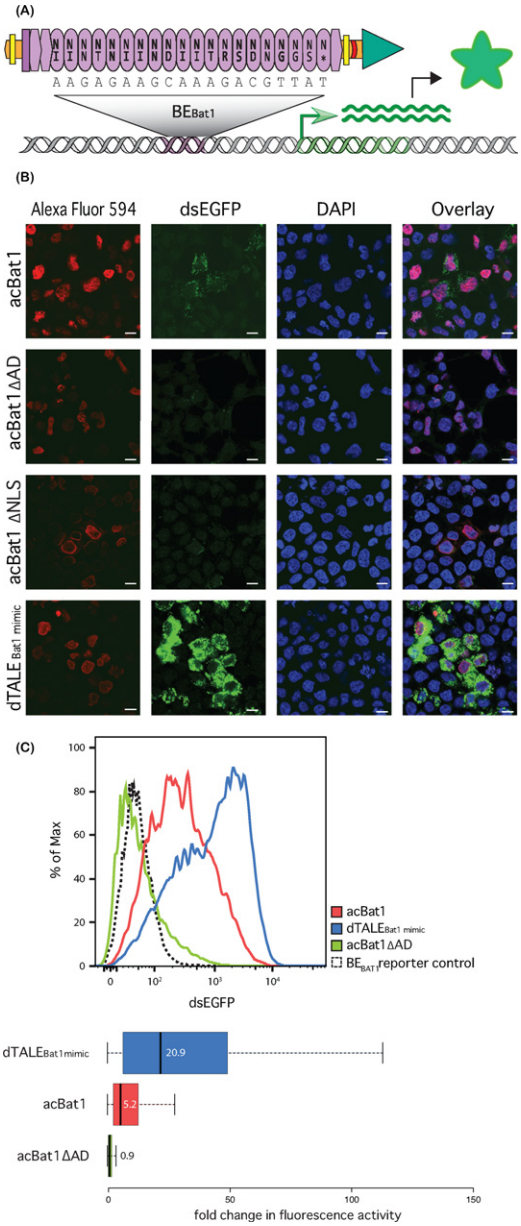


Figure 3. A Bat1 derived transcriptional activator (acBat1) is functional in a human cell reporter assay. (**A**) Schematic drawing showing the domain composition of acBat1. NLSs (yellow bars), a 3xFLAG tag (red crescent line) and a VP64 AD (green triangle) were fused onto Bat1 (purple) via flexible 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.

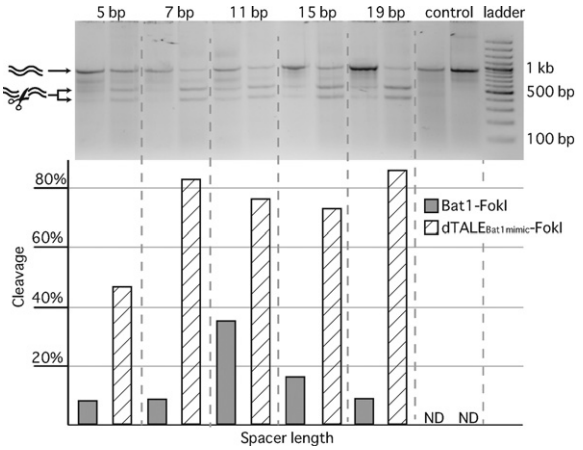


Figure 4. *In vitro* assessment of Bat1-FokI nuclease activity. Bat1- and TALE-FokI fusion proteins were expressed *in vitro* and equal volumes of transcription-translation product were incubated with a purified PCR product bearing two copies of BE_{Bat1} in reverse complement, separated by 5–19 base 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 (BE_{Bat1}, BE_{Bat2} and BE_{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 BE_{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 BE_{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, BE_{Bat1} may not represent the optimal binding sequence or additional endogenous factors may promote interaction *in vivo*.

BE_{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

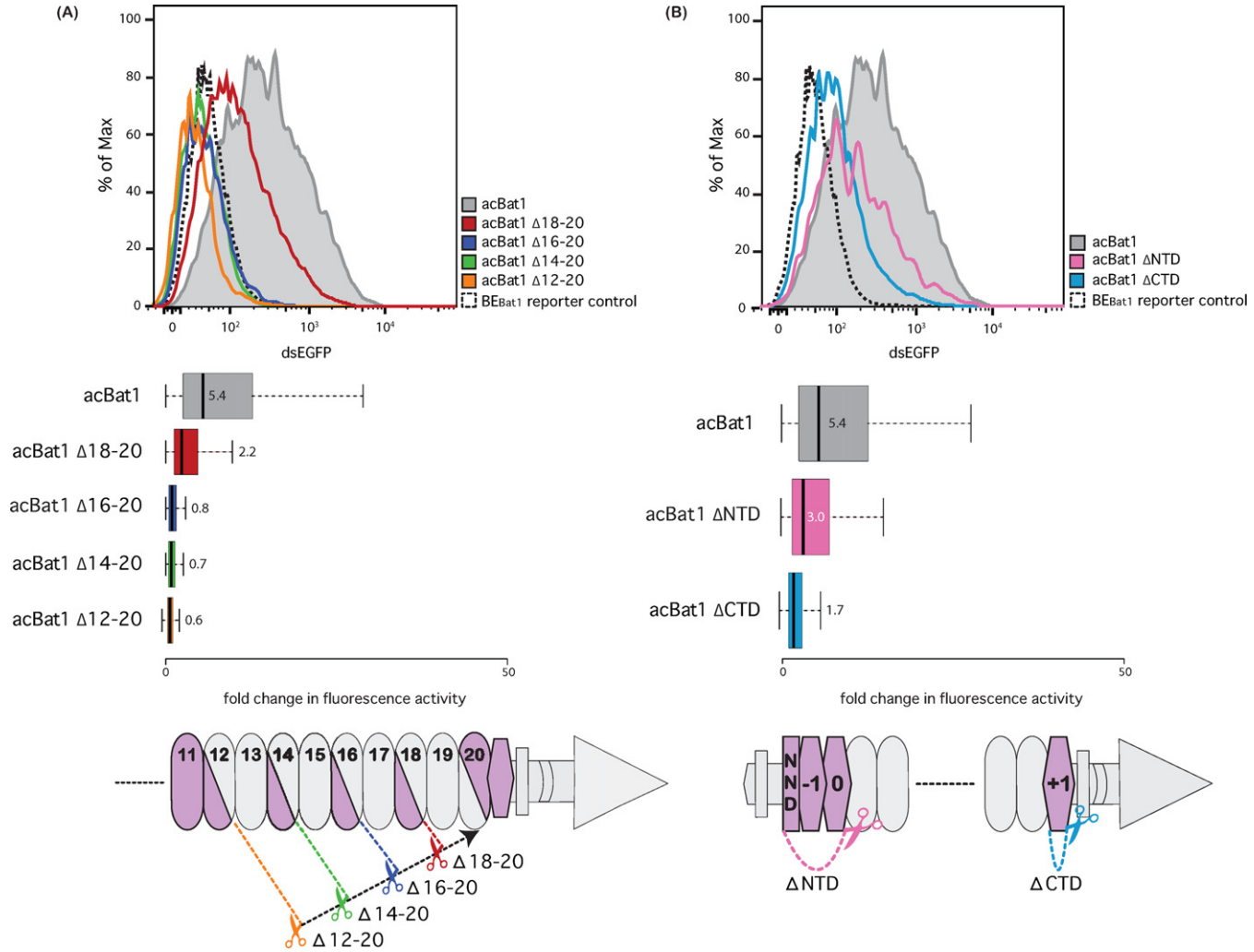


Figure 5. Functional analysis of acBat1 repeat truncations. Tests were carried out as described (Figure 3). Flow cytometry measurements of dsEGFP fluorescence are displayed as population distributions (top) or box plots (centre). Distinct colour codes are used throughout the whole figure and correspond 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 $dTALE_{Bat1mimic}$, 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 $dTALE_{Bat1mimic}$ 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 $dTALE_{Bat1mimic}$ induced the reporter 20-fold (Figure

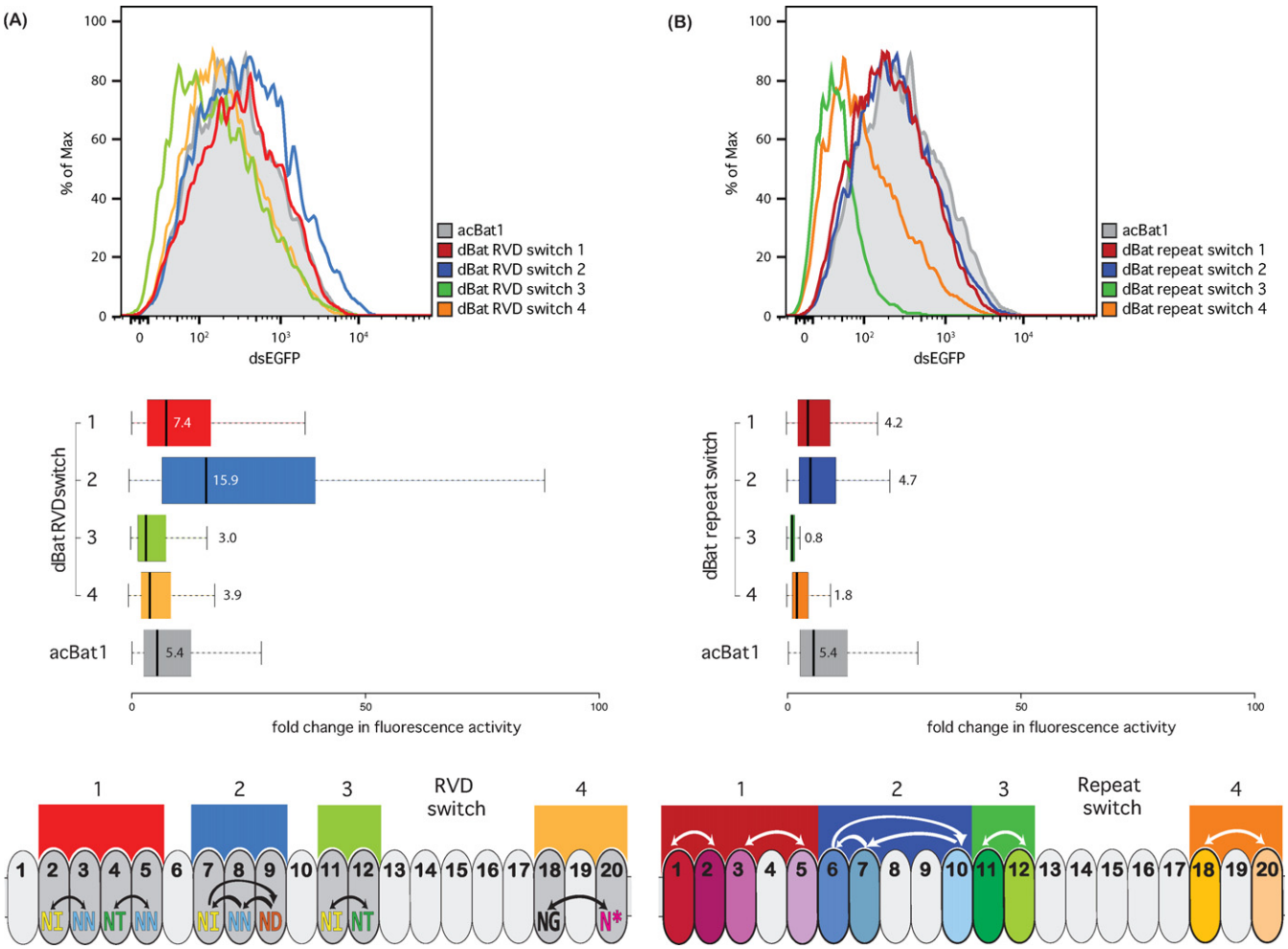


Figure 6. Functional analysis of designer (d)Bat constructs generated by RVD (**A**) or repeat switch (**B**). dBats were tested using flow cytometry with a transcriptional activation reporter as described (Figure 3). dsEGFP fluorescence values are displayed as population distributions (top) or boxplots (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-

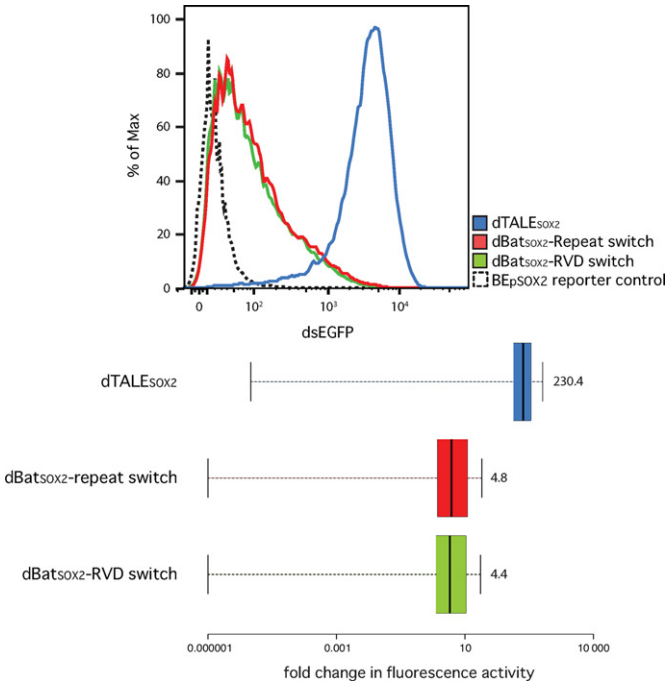


Figure 7. Functional analysis of designer (d)Bat constructs targeting the human SOX2 promoter. dBats were tested using flow cytometry with a transcriptional activation reporter as described (Figure 3). Population curves for dsGFP fluorescence are shown (top) as well as boxplots of fluorescence intensities (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 ($dBat_{SOX2\ RVD\ switch}$) or repeat switch ($dBat_{SOX2\ repeat\ switch}$), and an equivalent dTALE were tested.

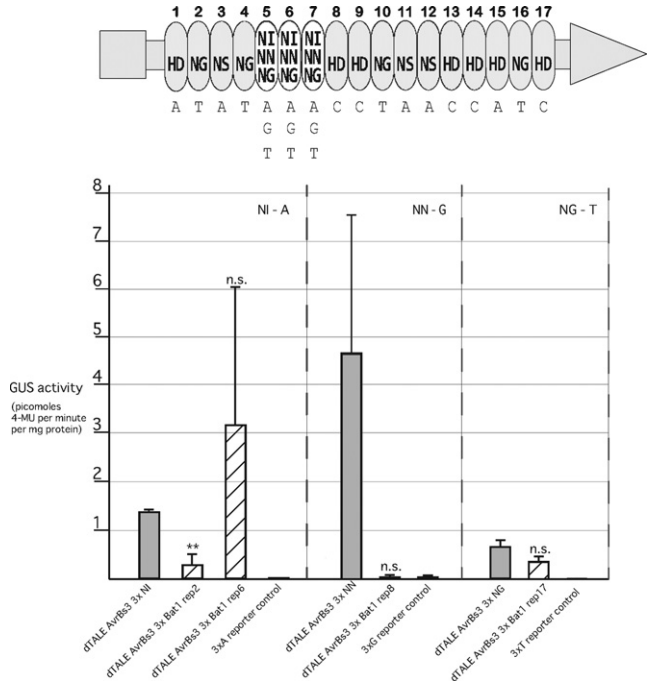


Figure 8. Functional analysis of Bat1 repeats within the context of a TALE repeat array. Trimers of identical Bat1 repeats or TALE repeats with the same RVDs as the Bat1 repeats were embedded into the repeat domain of the 17-repeat TALE AvrBs3 that targets the pepper *Bs3* promoter (*Bs3p*). Repeats 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.

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 $dTALE_{Bat1mimic}$-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 $dTALE_{Bat1mimic}$-FokI had a maximum efficacy of 86% cleavage (19 bp spacer; Figure 4). That $dTALE_{Bat1mimic}$-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.

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–

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-

tivate the reporter to similar levels (Figure 7) suggesting that both the RVD and repeat switch strategies can yield successful constructs. $dBat_{SOX2}$-repeat switch mediated 4.8x reporter activation and thus was slightly more active than the $dBat_{SOX2}$-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_{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 ($dTALE_{SOX2}$) 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).

$dTALE_{AvrBs3\_3xBat1\_rep2}$, 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 ($dTALE_{AvrBs3\_3xNI}$). $dTALE_{AvrBs3\_3xBat1\_rep8}$ (RVD NN) was barely able to elicit any detectable activation, unlike its TALE repeat equivalent ($dTALE_{AvrBs3\_3xNN}$). In contrast, $dTALE_{AvrBs3\_3x\_Bat1\ rep6}$ (RVD NI) and $dTALE_{AvrBs3\_3xBat1\_rep17}$ (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_{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

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-

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

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.

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 $BE_{\text{Bat1 T-0, A-0, C-0 and G-0}}$.

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 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
 2 LTPQQVVAIASNGGGKQALETVQRLLPVLCQAHG
 3 LTPQQVVAIASNSGGKQALETVQRLLPVLCQAHG
 4 LTPEQVVAIASNGGGKQALETVQRLLPVLCQAHG
 5 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
 6 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
 7 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
 8 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
 9 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
10 LTPQQVVAIASNGGGKQALETVQRLLPVLCQAHG
11 LTPEQVVAIASNSGGKQALETVQALLPVLCQAHG
12 LTPEQVVAIASNSGGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
15 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
16 LTPQQVVAIASNGGGRPALETVQRLLPVLCQAHG
17 LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
+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***taa*****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 -
GGGGGGSGGGGSGGGGSDYKDHDGDYKDHDIDYKDDDDKGSSPKKKRKVEASGSGRADALDD
FDLDMLGSDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLINSR – STOP

>Human cell transcriptional activation assay: ΔAD

MGYPYDVPDYASRPKKKRKVGIHAM – Bat1 coding sequence -
GGGGGGSGGGGSGGGGSDYKDHDGDYKDHDIDYKDDDDKGSSPKKKRKVEAS – STOP

>Human cell transcriptional activation assay: ΔNLSs

START – Bat1 coding sequence -
GGGGGGSGGGGSGGGGSDYKDHDGDYKDHDIDYKDDDDKGSGRADALDDFDLDMLGSDALDD
FDLDMLGSDALDDFDLDMLGSDALDDFDLDMLINSR – STOP

>*Planta* transcriptional activation assay: Full construct

START – Bat1 or dTALE coding sequence -
GGGGGGSGGGGSGGGGSDYKDHDGDYKDHDIDYKDDDDKGSSPKKKRKVEASGSGRADALDD
FDLDMLGSDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLINSR – STOP

>*Planta* transcriptional activation assay: ΔAD

START – Bat1 coding sequence -
GGGGGGSGGGGSGGGGSDYKDHDGDYKDHDIDYKDDDDKGSSPKKKRKVEAS – STOP

>*In vitro* nuclease assay

MGLINIFYPYDVPDYAGYPYDVPDYAGSYPYDVPDYAAQCSG – Bat1 coding sequence –
GGQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHL
GGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWW
KVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEE
VRRKFNNGEINF

MGYPYDVPDYASRPKKKRKVGIHAS – TALE coding sequence -
GSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHL
GGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWW
KVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEE
VRRKFNNGEINF

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

BE_{Bat1}/$BE_{Bat1\ T\text{-}0}$
TAGACT**AAGAGAAGCAAAGACGTTAT**ATGC

BE_{Bat2}
TAGACT**TTGTTGAAAAGTTGTAAAAACATTAT**ATGC

BE_{Bat3}
TAGACATAGATTAT**TATATTTG**TAACAAGTAAATGC

$BE_{Bat1\ C\text{-}0}$
TAGACC**AAGAGAAGCAAAGACGTTAT**ATGC

$BE_{Bat1\ G\text{-}0}$
TAGACG**AAGAGAAGCAAAGACGTTAT**ATGC

$BE_{Bat1\ A\text{-}0}$
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.

BE_{Bat1}	AAGAGAAGCAAAGACGTTAT
$BE_{dBatRVDswitch1}$	A**GAGA**AAGCAAAGACGTTAT
$BE_{dBatRVDswitch2}$	AAGAGA**GCA**AAAGACGTTAT
BE_{pSOX2}	TTTATTCCCTGACAGCCCC

CTAGACT**NNNNNNNNNNNNNNNNNNNN**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$_{Bat1}$-*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$_{Bat1}$ 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$_{Bat1mimic}$.

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 BE_{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

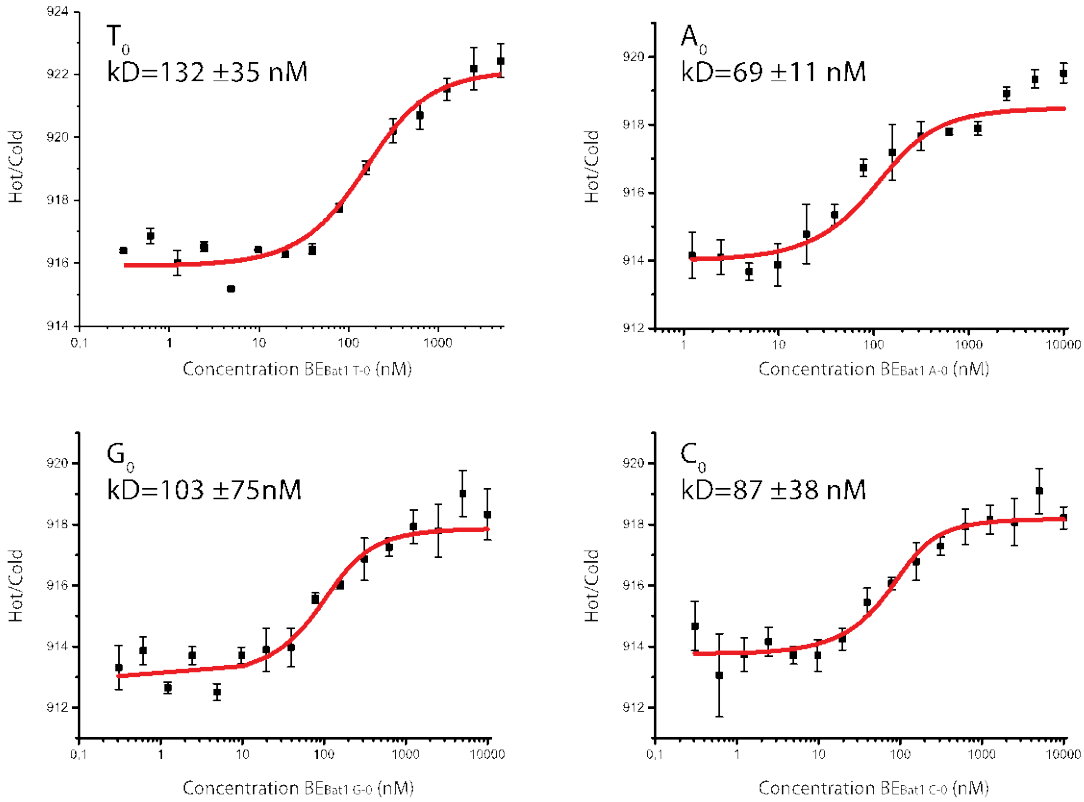
15bp TACGTACGTCTAGAC

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$_{Bat1mimic}$ (for transcriptional activation assays)

```
   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
18 LTPEQVVAIASNGGGKQALETVQRLLPVLCQAHG
19 LTPQQVVAIASNSGGKQALETVQALLPVLCQAHG
+1 LTPQQVVAIASN-GGRPALESIVAQLSRPDPALAA
+2 LTNDHLVALACL-GGRPALDAVKKGLPHAPALIKR
   TNRRIPERTSHRVA
```

>dTALE$_{SOX2}$ (for human cell transcriptional activation assay)

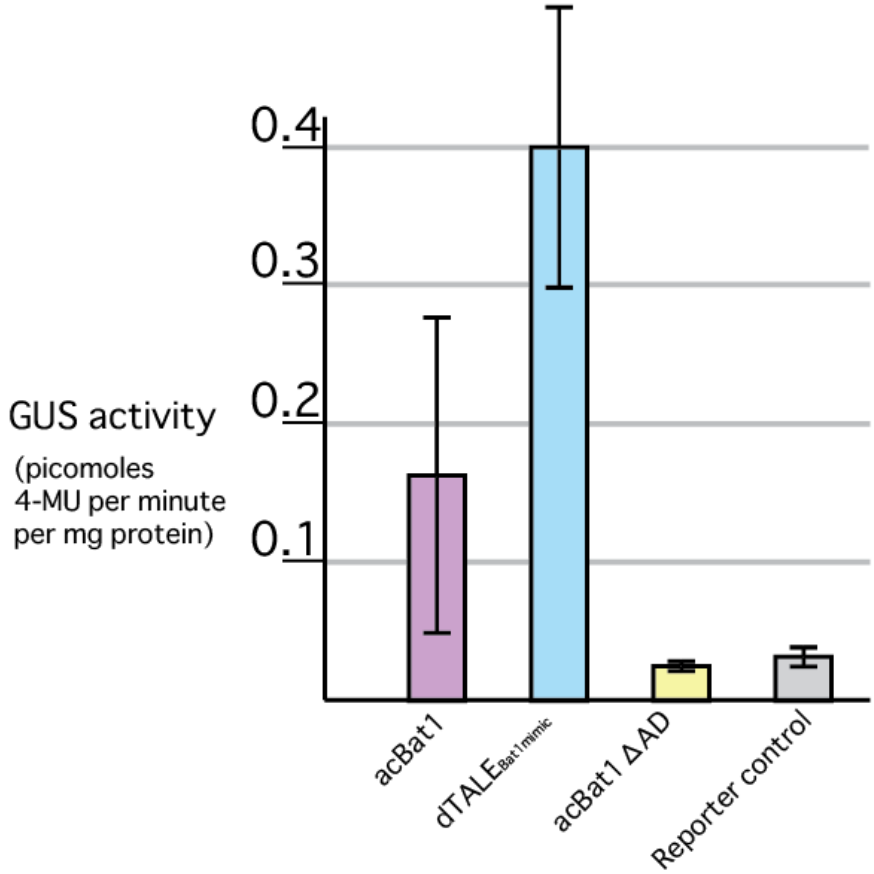
```
MDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGH
GFTHAHIVALSQHPAALGTVAVKYQDMIAALPE
-1 ATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQ
 0 LDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLN
 1 LTPQQVVAIASNIGGKQALETVQRLLPVLCQAHG
 2 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
 3 LTPEQXVAIASNNGGKQALXTVQRLLPVLCQAHG
 4 LTPQQVVAIASNTGGKQALXTVQRLLPVLCQAHG
 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$_{Bat1mimic}$ (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 pGWB442d$TALE_{Bat1mimic}$ 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 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
```

>dBat RVD switch 1

```
NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
0  LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
1  FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
2  FSRDDIAKMAGNNGGAQTLQAVLDLESAFRERG
3  FSQADIVKIAGNIGGAQALYSVLDVEPTLGKRG
4  FSRADIVKIAGNNGGAQALHTVLDLEPALGKRG
5  FSRIDIVKIAANTGGAQALHAVLDLGPTLRECG
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 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$_{SOX2}$-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$_{SOX2}$ 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

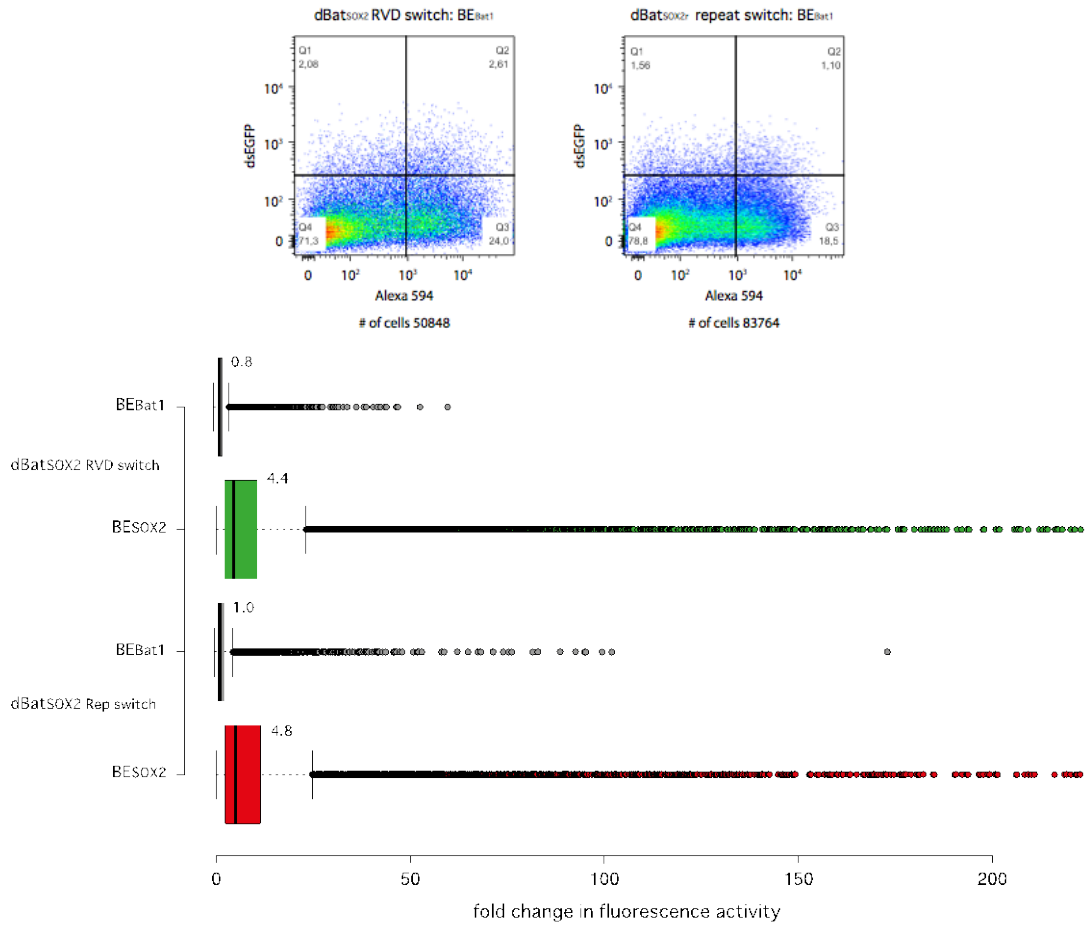
```
NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
 0 LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
 1 FSQSDIVKIAGNGGAQALQAVLDLESMLGKRG
 2 FSRDDIAKMAGNGGAQTLQAVLDLESAFRERG
 3 FSQADIVKIAGNIGGAQALYSVLDVEPTLGKRG
 4 FSRADIVKIAGNGGAQALHTVLDLEPALGKRG
 5 FSRIDIVKIAANGGAQALHAVLDLGPTLRECG
 6 FSQATIAKIAGNDGGAQALQMVLDLGPALGKRG
 7 FSQATIAKIAGNDGGAQALQTVLDLEPALCERG
 8 FSQATIAKMAGNDGGAQALQTVLDLEPALRKRD
 9 FRQADIIKIAGNGGGAQALQAVIEHGPTLRQHG
10 FNLADIVKMAGNNGGAQALQAVLDLKPVLDEHG
11 FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNDGGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGNIGGAQALQAVLALELTLRERG
14 FSQPDIVKIAGNNGGAQALQAVLDLELTFRERG
15 FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16 FSRADIVNVADNNGGAQALKAVLEHEATLNERG
17 FSRADIVKIAGNDGGAQALKAVLEHEATLDESG
18 FSRADIVRIAGNDGGAQALKAVLKYGPVLMQAG
+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 BE_{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 BE_{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

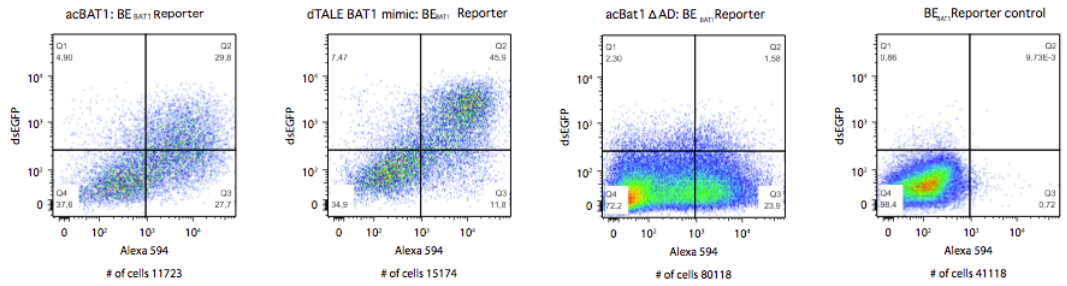


Figure 5

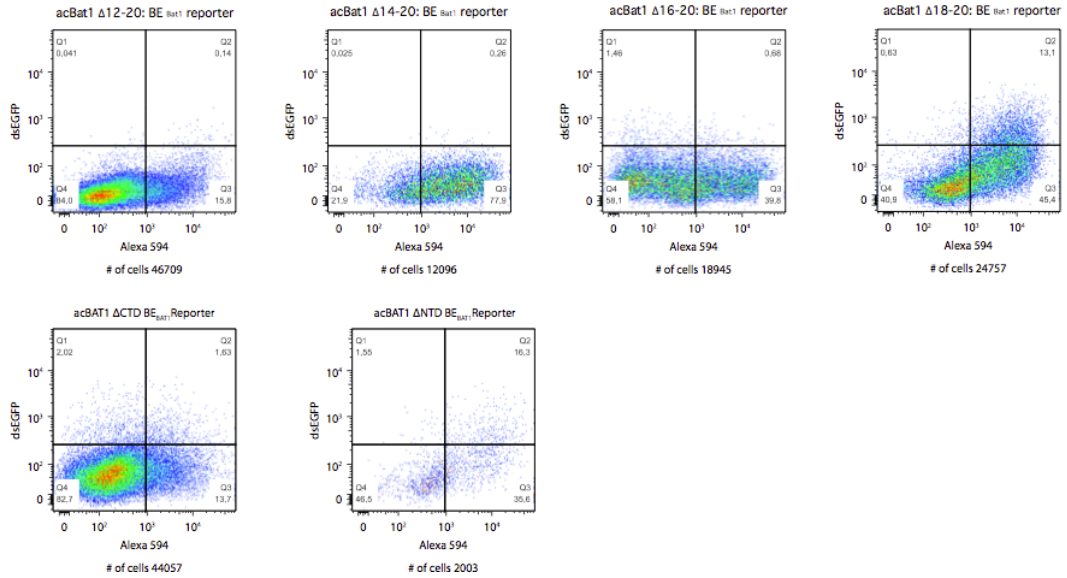


Figure 6

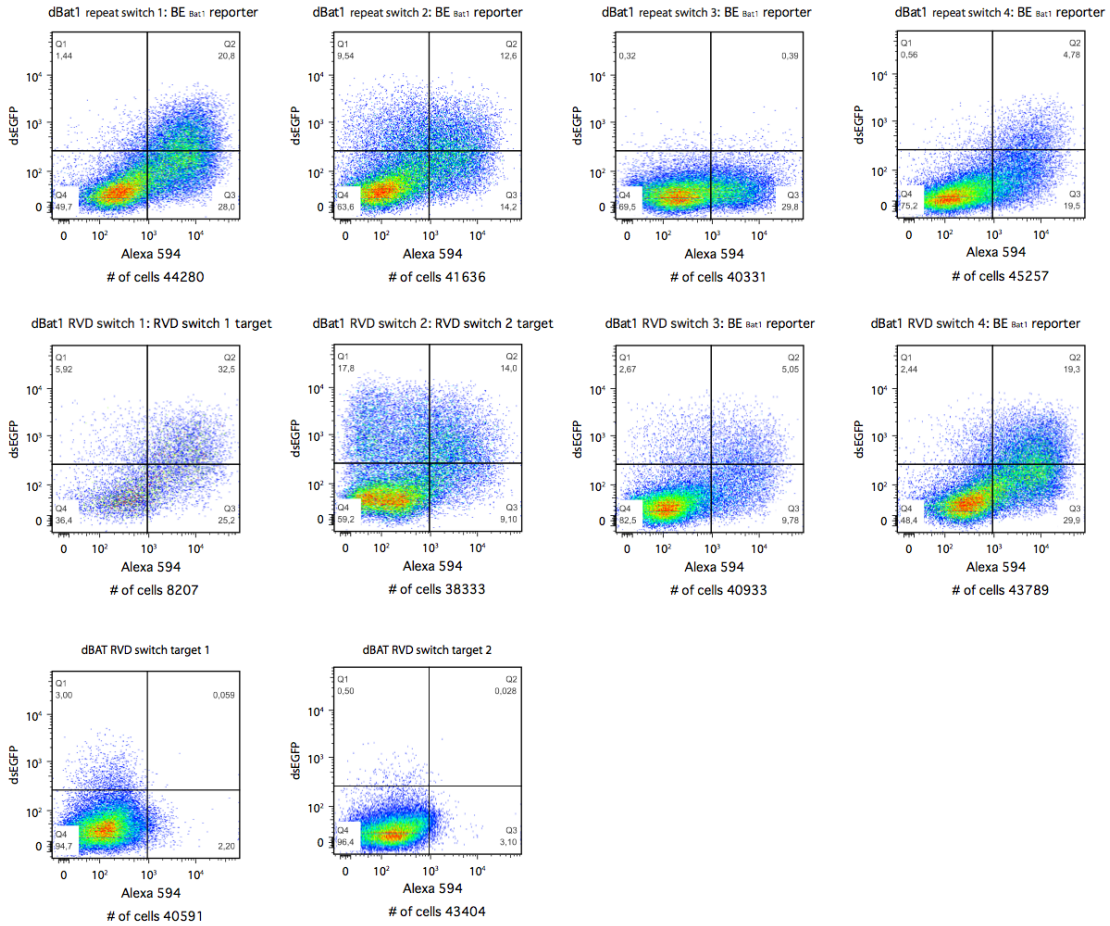


Figure 7

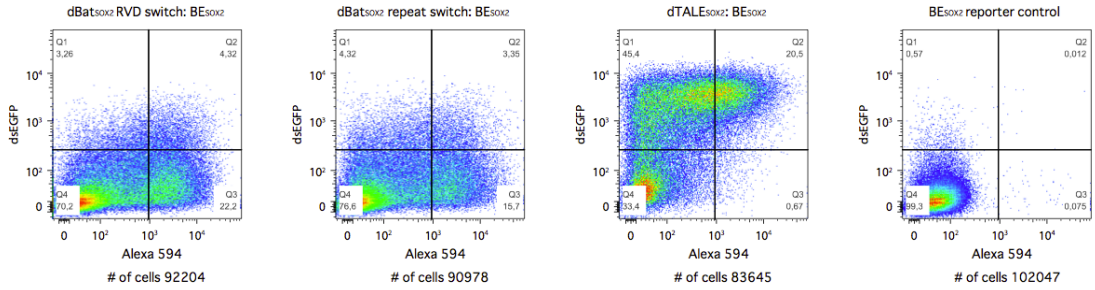


Figure 3

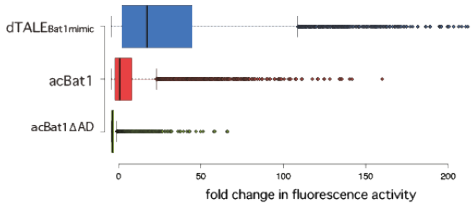
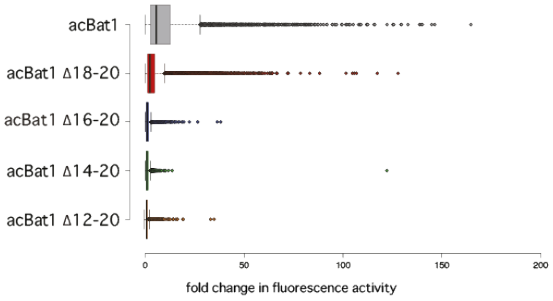


Figure 5



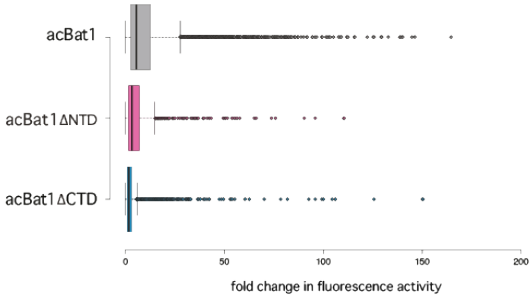
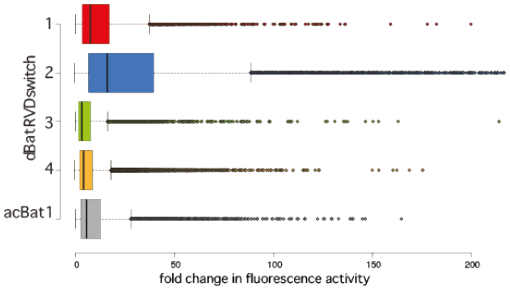


Figure 6



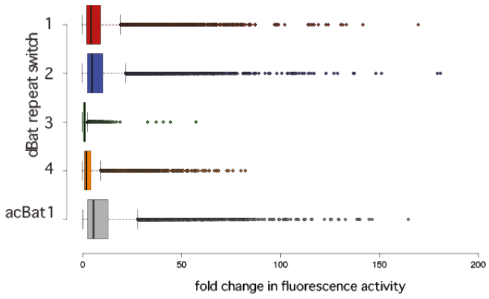
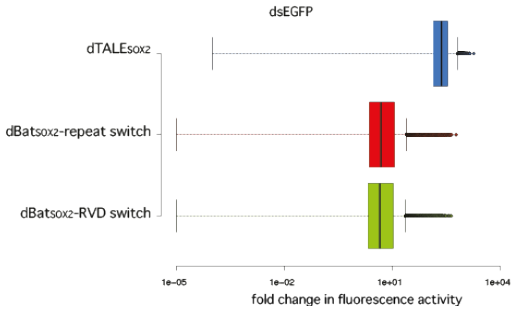


Figure 7



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

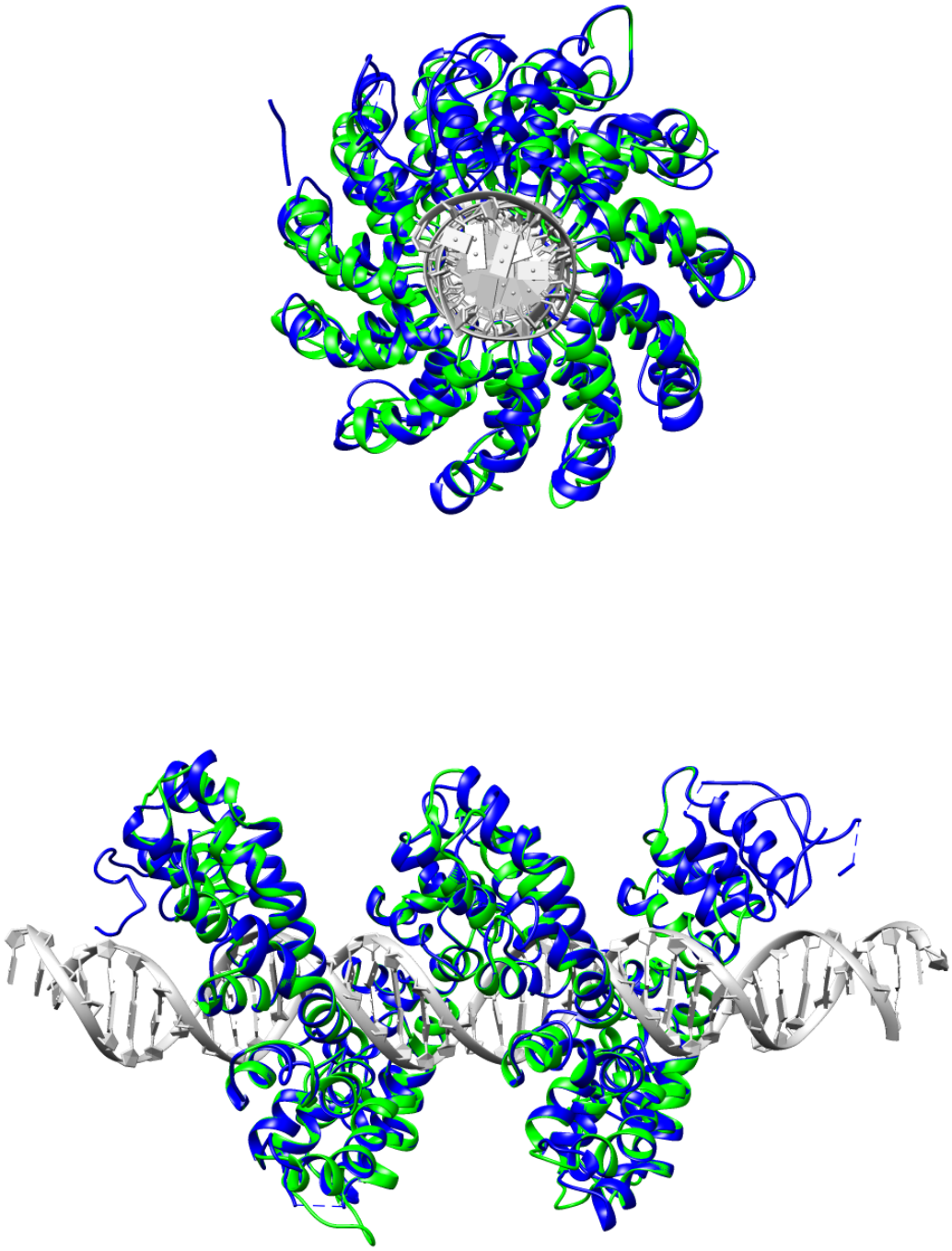
GMQE: 0.70,

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.

Published online 19 October 2015 *Nucleic Acids Research, 2015, Vol. 43, No. 20* ***10065–10080***
doi: 10.1093/nar/gkv1053

DNA-binding proteins from marine bacteria expand the known sequence diversity of TALE-like repeats

Orlando de Lange[1,†], Christina Wolf[1,†], Philipp Thiel[2], Jens Krüger[2], Christian Kleusch[3], Oliver Kohlbacher[2,4] and Thomas Lahaye[1,*]

[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]NanoTemper Technologies, Munich, 81369, Germany and [4]Quantitative Biology Centre and Faculty of Medicine, University of Tuebingen, Sand 14, Tuebingen, Baden-Wuerttemberg, 72076, Germany

Received April 17, 2015; Revised September 11, 2015; Accepted September 14, 2015

ABSTRACT

Transcription Activator-Like Effectors (TALEs) of *Xanthomonas* bacteria are programmable DNA binding proteins with unprecedented target specificity. Comparative studies into TALE repeat structure and function are hindered by the limited sequence variation among TALE repeats. More sequence-diverse TALE-like proteins are known from *Ralstonia solanacearum* (RipTALs) and *Burkholderia rhizoxinica* (Bats), but RipTAL and Bat repeats are conserved with those of TALEs around the DNA-binding residue. We study two novel marine-organism TALE-like proteins (MOrTL1 and MOrTL2), the first to date of non-terrestrial origin. We have assessed their DNA-binding properties and modelled repeat structures. We found that repeats from these proteins mediate sequence specific DNA binding conforming to the TALE code, despite low sequence similarity to TALE repeats, and with novel residues around the BSR. However, MOrTL1 repeats show greater sequence discriminating power than MOrTL2 repeats. Sequence alignments show that there are only three residues conserved between repeats of all TALE-like proteins including the two new additions. 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

Three groups of plant disease associated bacteria have so far been found to encode sequence-related repeat-array proteins known as TALE-likes. The repeat arrays of TALE-likes are DNA-binding domains, with each repeat binding a single DNA base with a common code based on repeat residue 13, the base specifying residue (BSR; use of the term reviewed in (1)). The largest, first discovered and eponymous group are the TALEs, of plant-pathogenic *Xanthomonas* species. Next described and characterised were the RipTALs of *Ralstonia solanacearum* (2,3) and lately the Bats of endofungal bacterium *Burkholderia rhizoxinica* (4–6). Of these groups the TALEs and RipTALs are effector proteins injected into host plants where they mimic eukaryotic transcription factors (7). The repeats bind specific promoter sequences and a domain at the C-terminus of the protein mediates activation of host genes whose products promote bacterial disease. TALEs thus hijack the host's transcriptional machinery and RipTALs are thought to do the same (8,9). The Bats lack the domains necessary to function as eukaryotic transcription factors (6) and their evolutionary relationship to the TALEs and RipTALs remains unclear. The TALE-likes seem to be united only by possession of DNA binding repeats with a conserved code.

TALEs are studied for their applications in biotechnology as much as for their roles in plant disease (10). The reliability of the TALE code allows one to predict the DNA binding element (BE) for any given TALE and to 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 (11), transcriptome (12) and even epigenome (13,14).

One of the potential advantages of the TALE system over the alternative CRISPR/Cas9 system is the diversity of BSR–DNA interactions, contrasting with more restricted Watson–Crick base pairing. BSRs bind their cognate bases with a range of different affinities and specificities, as inferred from studies on arrays with different BSR composi-

*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
†These authors contributed equally to the paper as first authors.

tions (15). In addition, non-BSR polymorphisms might be useful to tune DNA binding properties and further expand the diversity of TALE–DNA interactions. One could then create libraries of dTALEs with a range of binding strengths for the same DNA element, useful for the regulation of synthetic genetic circuits.

One approach to TALE repeat engineering is random mutagenesis and screening, as demonstrated successfully in a recent study by Hubbard *et al.* (16). Alternatively mutations could be introduced in a more targeted fashion, but this requires information on the impact of different types of polymorphisms at different positions in the TALE repeat. Natural variation would provide useful information on what residues can or cannot be tolerated at which positions and with what effect. However, whilst TALES are distributed widely among *Xanthomonas* species, sequence diversity is very low (17). Yet the first characterised RipTAL, Brg11, is only 41% identical to TALE AvrBs3 (18) including numerous repeat sequence polymorphisms. In addition the polymorphism between individual RipTAL repeats is greater than that between TALE repeats. We looked at the DNA recognition properties of each of the repeats of the RipTAL Brg11 and found differences in reporter activation strength even when comparing repeats with identical BSRs, suggesting that non-BSR polymorphisms impact on repeat–DNA interactions (3). Thus RipTAL repeats could be useful as a pool of natural sequence diversity for TALE repeat engineering.

This pool of functionally validated but sequence-diverse TALE-like repeats was further expanded by the molecular characterisation of the Bats of bacterium *B. rhizoxinica* (4–6,19). Repeats of these proteins are below 40% identical to TALE repeats, providing an interesting group for comparison. TALE and Bat repeats mediate DNA binding with broadly the same BSR code and the structures are similar (19,20), but some functional differences were identified (19). This makes the Bats a useful comparison group to inform studies into TALE repeat engineering.

However, residues clustered around the BSR (positions 7–19) are largely invariant across all currently known TALEs, RipTALs and Bats (6). 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 re-engineering attempts aimed at changing DNA binding properties.

We describe here molecular characterisations of two novel repeat proteins predicted from marine bacterial metagenomics sequences (21,22). Repeats of these proteins show 30–40% protein level sequence similarity to TALE repeats. We refer to these predicted proteins as MOrTL1 and MOrTL2 (Marine Organism TALE-Likes) to reflect the 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 support the DNA-binding analysis we build homology models of MOrTL1 and MOrTL2 repeats bound to DNA and carry out molecular dynamics (MD) simulations to test the stability of the modelled interactions. The models show a striking structural similarity to TALE and Bat repeats. Yet MOrTL1 and MOrTL2 repeats bear sequence motifs unknown from TALEs, RipTALs and Bats. Repeats of the two MOrTLs are as distant from one another at the sequence level as they are from any of the other TALE-likes and show functional differences: MOrTL1 repeats exert a greater sequence discriminating power and, unlike MOrTL 2 repeats, they are compatible with both Bat1 and TALE repeats. The sequence diverse MOrTL1 and MOrTL2 repeats could inform future TALE repeat engineering efforts as well as being useful as comparison groups for evolutionary analyses. This makes the MOrTLs a fascinating addition to the growing family of TALE-likes.

MATERIALS AND METHODS

MOrTL construct creation

Genes encoding MOrTL1 (ECG96326) and MOrTL2 (EBN1909), codon optimized for *Escherichia coli* and with additional 5′ and 3′ BsaI recognition sites, were synthesized (GenScript). Sequences are found in Supplementary 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 polymerase chain reaction (PCR) amplified with the primers listed in Supplementary Table S2 bearing BsaI sites corresponding to Block 2 of the previously described Bat1 cloning system (6). The MOrTL blocks, along with Bat1 blocks 1 and 3–5, were assembled into either a Golden Gate compatible pENTR (BsaI overlaps CACC-AAGG) 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 Supplementary Table S2 bearing BsaI sites corresponding to the 5B level 2 repeat blocks of the designer TALE assembly toolkit as previously described (23) 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 BE_{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 (6) 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 (6). 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 (6). Complementary pairs of labelled or corresponding unlabelled oligonucleotides were annealed (list of oligos Supplementary Table S2). 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 min and resolved on a 6% native polyacrylamide gel for 1 h at 100 V, 4°C. Labelled DNA was visualized with a Typhoon FLA 9500 (GE healthcare).

Binding affinity quantifications via MST

Microscale thermophoresis was performed using the Monolith NT.115 (Nanotemper Technologies). Complementary pairs of labelled oligonucleotides (Cy5, Eurofins) were annealed in MST buffer (Tris 20 mM, NaCl 150 mM, 10 mM MgCl2) (18). Affinity measurements were performed by using MST buffer, supplemented with 0.05% Tween as final concentration. Samples were loaded into NT.115 premium capillaries (NanoTemper Technologies. Measurements were performed at 24°C, 30% LED, 20% IR-laser power and constant concentration of 50 nM of labelled oligonucleotides and increasing concentration of purified protein.

Protein melting point analysis

Protein thermal stability was measured in a label-free fluorimetric analysis using the Prometheus NT.48 (NanoTemper Technologies). Briefly, the shift of intrinsic tryptophan fluorescence of proteins upon temperature-induced unfolding was monitored by detecting the emission fluorescence at 330 and 350 nm. Thermal unfolding was performed in nanoDSF grade high-sensitivity glass capillaries (NanoTemper Technologies) at a heating rate of 1°C per minute. Protein melting points (T_m) were calculated from the first derivative of the ratio of tryptophan emission intensities at 330 and 350 nm.

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.* (24), who kindly provided us with plasmids pCherry (mCherry reporter) and TALE expression plasmid $pBT102_{LacO\ dTALE}$ (dTALE targeting *lac* operon, downstream of synthetic constitutive promoter *J23102*).

In order to create reporters for each test protein we inserted novel BEs into the *Trc* promoter of pCherry immediately 3′ of the lac operon (see Supplementary Figures S8 and S9). This was done via PCR amplification of the whole plasmid, using primers listed in Supplementary Table S2 with each bearing one half of the BE as an overhang. The sequences of the novel *Trc* promoter derivatives we created bearing different BEs can be found in Supplementary Figure S9.

We adapted the $pBT102_{LacO\ dTALE}$ plasmid by removing the TALE gene and adding Golden Gate cloning sites in its place. 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; Supplementary Figure S9) 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 Supplementary Figure S8). dTALE blocks with or without a block of MOrTL repeats were then cloned into this vector via BpiI cut-ligation as described above. The resulting genes encode C-terminal GFP fusion proteins. Bat1 repeat blocks alone or together with a MOrTL repeat block were cloned into the pBT102* CACC-GW-AAGG vector via BpiI cut-ligation. These constructs have no GFP tag.

The assay was carried out by co-transforming approximately 25 ng of each plasmid (pCherry and pBT102*) into chemically competent *E. coli* Top10 cells (Life Technologies) 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 repressor of Top10 cells since the mCherry reporter gene has a lac operator in its promoter. Plates were incubated 36 h at 37°C to achieve stationary phase colonies. This is important since the growth rates of subsequent liquid cultures would otherwise differ based on the growth stage of the colonies from which they were inoculated. Single colonies were 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). Picking was done by hand with 200 μl pipette tips scraping only the edge of the colony to avoid taking too much bacterial mass into the low volume liquid cultures since preliminary tests found that too high an initial inoculum led to very high starting mCherry values, and frustrated OD 600 normalisation. Cultures were shaken 3.5 h at 37°C, 180 rpm, determined in preliminary experiments to correspond to the late log phase giving the best reduction of variation via OD 600 normalisation of any tested time point. OD 600 was measured in a plate reader (TECAN) as well as mCherry fluorescence was measured in a TECAN Safire2 microplate reader with the following parameters: Excitation 587 nm, Emission 610 nm, bandwidth ± 12nm, Gain 90, Z-position 6300 μm, followed by an OD 600 measurement for normalisation. Boxplots were generated in RStudio (v. 0.98.501).

Structure modelling

Homology models of $Bat1_{M1\ (3–7)}$ and $Bat1_{M2\ (2–6)}$ 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) (25) 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 sim-

ulations of both models using the software package GROMACS (version 4.6.7) (26). The protocol that was applied to both models used the CHARMM27 all-atom force field (version 2.0) with CMAP (27,28) 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 × 64 × 72 cells was used with fourth-order B-spline interpolation. A single cut-off of 1 nm was used for van der Waals interactions to limit the local interaction distance. 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 (29) (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 Figure 6A and B. Input files and parameter settings for both simulations are 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 used throughout to refer to two predicted proteins: MOrTL1 and MOrTL2, from marine microbial genomic DNA, sequenced as part of the Global Ocean Sampling (GOS) expedition (21). MOrTL1 is synonymous with GenBank protein ID ECG96325 and MOrTL2 with EBN91409. These sequences have been previously suggested to encode modular DNA binding repeats (5) but no functional analysis has been reported until now. Both proteins are tandem repeat arrays, with each repeat 33 amino acids in length; MOrTL1 is formed of 8 repeats, and MOrTL2 of 10 repeats (Figure 1A and B).

Organisms bearing the MOrTLs sequences were sampled from the Gulf of Mexico/Yucatan Channel and are most likely of bacterial origin based on size filtering of the biological material that was used for recovery of DNA (0.1–0.8 μm) (21,22). The biological samples from which MOrTL1 and MOrTL2 were sequenced came from two different locations. The genes are thus at the very least from two different populations and may be from different organisms. Both of the contigs in question are orphans not matching at either end to anything else in the GOS database. Each contig was sequenced with a read from each end covering roughly 1 kb in each case. Both reads contain repeat sequences and a consensus was built in the centre of the contig from the two reads in the case of the MOrTL2-containing-contig. Because of this, the reference sequence in GenBank (*EN814823.1*) indicates two open reading frames (ORFs) separated only by a frameshift in the middle, while the separate reads suggest incomplete sequencing of a larger repeat

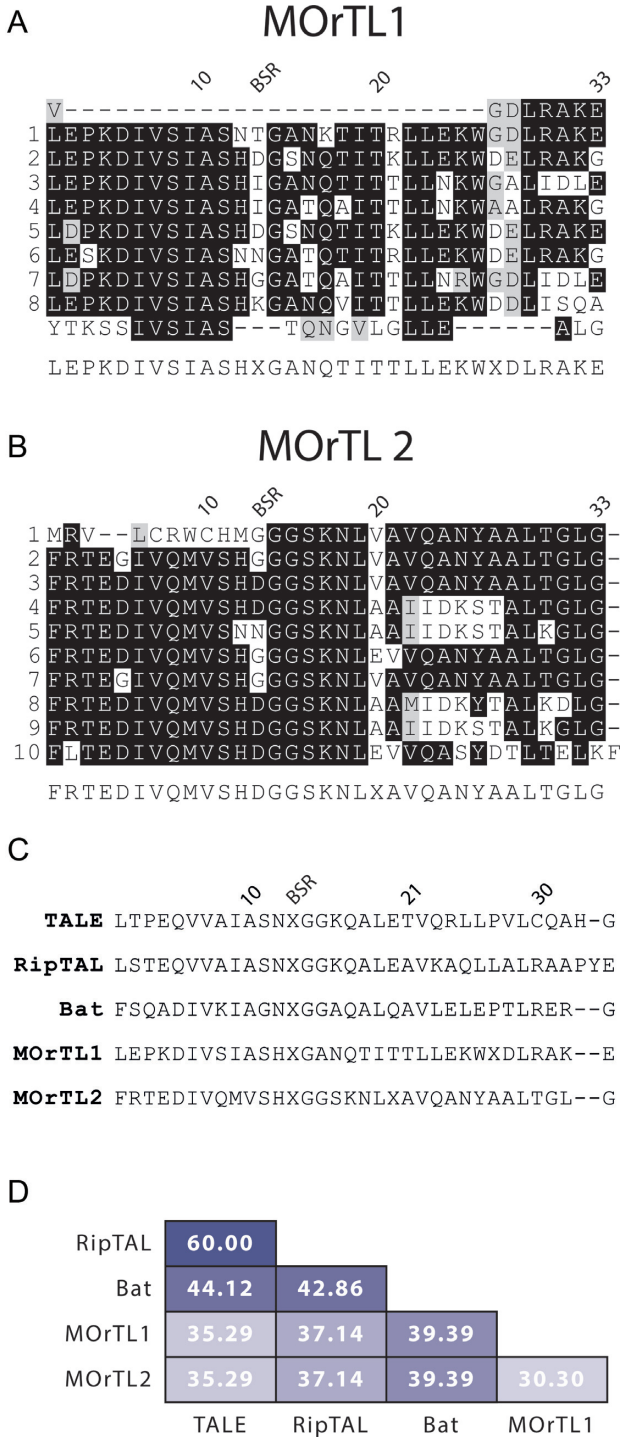


Figure 1. Amino acid sequences of MOrTL1 (ECG96326) and MOrTL2 (EBN19409), and a comparison of consensus TALE-like repeats. **(A, B)** Full amino acid sequences of each protein are displayed as a series of aligned tandem repeats prepared with ClustalW and Boxshade. Identical amino acids are white text on black background, similar amino acids present in 50% of sequences are black on grey background, and dashes indicate gaps. 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. Repeats are numbered down the left-hand side in each case, excluding the degenerate repeat-like sequences framing MOrTL1. **(C)** Consensus core repeats of each TALE-like group. **(D)** Heat map of percentage pairwise sequence similarities of consensus repeats shown in panel (C).

protein. We believe the same has happened for the MOrTL1 contig (*EM567463.*1) although in that case the reference sequence suggests an unresolvable run of N's intervening between two repeat protein ORFs. Sequences of the individual reads from which these contigs were assembled can be found in Supplementary Figure S1.

We compared consensus repeat sequences of MOrTL1 and 2 to consensus repeat sequences of TALEs, RipTALs and Bats (Alignments Supplementary Figure S2, consensus sequences Figure 1C; pairwise identities Figure 1D). Pairwise identities for MOrTL1 and MOrTL2 compared to each other and different TALE-likes are all within 30–40%. MOrTL1 and MOrTL2 consensus repeats share no common sequence features not found in other TALE-like repeats and have the lowest pairwise similarity of any two consensus repeats in the comparison (Figure 1D). MOrTL1 and MOrTL2 repeats differ at more than 60% of positions from each other and from all other TALE-likes. Both for MOrTL1 and MOrTL2 the Bat consensus repeat is the closest relation in terms of sequence identity, though the difference is slight.

Purified MOrTL1 exhibits low affinity DNA binding: database sequences are likely incomplete

It has been shown that TALEs with fewer than 10 repeats are not able to activate reporter genes (30). Thus, there may be too few repeats in the available MOrTLs sequences as they are to achieve high affinity DNA binding. 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 (6,31). Such sequences may also exist in the full-length MOrTL proteins but are not found in the sequences available. Indeed coding sequences (CDSs) of both MOrTLs 1 and 2 begin in what appears to be the middle of a repeat (Supplementary Figure S1D and S1H) supporting this idea. We therefore considered it likely that the reference MOrTL sequences would not yield functional proteins. We nevertheless had genes encoding the reference MOrTL1 and MOrTL2 proteins synthesized. We were able to express and purify MOrTL1 from *Escherichia coli*, while MOrTL2 formed protein aggregates preventing purification (Supplementary 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; BE_{MOrTL1}; Figure 2A) based on the TALE code (Supplementary Table S1). A shift was detectable only with a MOrTL1 concentration of 822 nM or greater (Supplementary Figure S3B). Such weak DNA binding is inconsistent with expectations based on other TALE-likes (6,32). In addition laddering was observed in the gel shift indicating the formation of higher order protein–DNA complexes (Supplementary 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 likely to be fragments of larger, incompletely sequenced genes (Supplementary Figure S1). We considered it worth attempting to fuse together the repeats encoded on both reads of MOrTL2 contig *EN814823.1* even if intervening sequence is lacking. However, the resultant fusion protein (EBN19408-MOrTL2; sequence in Supplementary Figure S4) formed insoluble protein inclusions in *E. coli*, like MOrTL2, preventing functional analysis.

In vitro assays on chimeric Bat1-MOrTL repeat arrays demonstrate DNA binding consistent with the TALE code

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 (3,6). We chose a Bat1 repeat array framework to work with since the Bat consensus repeat was the most similar to MOrTL1 and MOrTL2 repeats at the sequence level (Figure 1D). We tested blocks of five repeats from the central part of each MOrTL embedded within the repeat domain of Bat1 at positions 6–10 ($Bat1_{M1\ (3\text{-}7)}$ and $Bat1_{M2\ (2\text{-}6)}$; Figure 2A). In each case the integrated MOrTL repeats differ in their BSR composition from the Bat1 repeats they replace, which should lead to a modified DNA sequence preference. The design of each chimera is illustrated in Figure 2A. Note that repeats of MOrTL1 and MOrTL2 differ from Bat1 repeats at distinct positions (Figure 2B). To get a first idea of DNA binding properties purified Bat1 and chimera proteins were tested *in vitro* with EMSAs (Figure 2C, Supplementary Figure S6) against cognate BEs, which we predicted with the TALE code (Supplementary Table S1). Clear single shifts, of similar intensity, were observed for Bat1 and Bat1-MOrTL chimeras at 200 nM with their cognate BEs (Figure 2C). We followed this up by using microscale thermophoresis (MST) to quantify the affinity of the binding and calculate a K_D. We found an almost identical affinity in each case: 126 nM for $Bat1_{M1\ (3\text{–}7)}$ with its BE and 128 nM for $Bat1_{M2\ (2\text{–}6)}$ with its BE (Figure 2D, Supplementary Figure S7). We have previously tested the Bat1–BE_{Bat1} interaction in the same system and found a K_D of 132 ± 35 nM (6)). Thus both Bat-MOrTL chimeras were able to bind their cognate TALE-code predicted BE with a strength similar to the wild type Bat1 protein.

Tests with predicted on-target sequences do not alone prove adherence to the TALE code. Specificity also needs to be tested. We designed off-target BEs choosing the worst predicted match for each of the five repeats in positions 6–10 (MOrTL repeat block in the chimeras) of each construct based on the TALE code (Supplementary Table S1): G used for Gly at the BSR and T used for Arg, Asp or Ile at the BSR. Applying this code results in a single off-target oligonucleotide with *GGTTG* at the test position for all three proteins. All other positions in target DNAs were kept constant to isolate the test repeats and test their specificity. We first carried out EMSA competition assays for all three proteins. MST was carried out for the two chimeras to assess affinities for off-target DNAs. In the EMSA competition assays (Figure 2E) the labelled on-target probe is mixed with an excess of either on- or off-target competitor DNA: If the test repeats bind the labelled probe in a specific fashion then an excess of 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

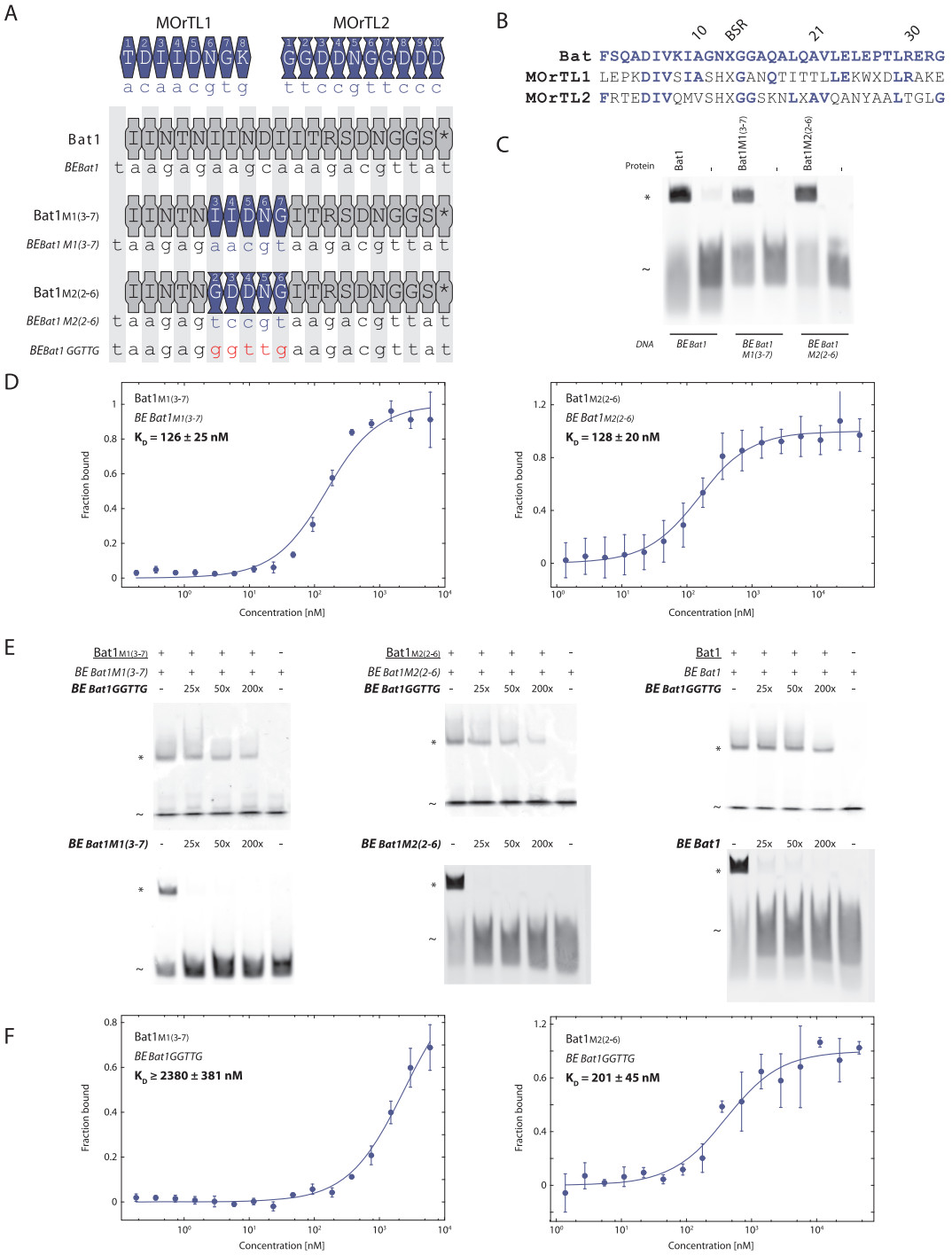


Figure 2. Bat1-MOrTL chimera proteins bind predicted target sequences *in vitro*. **(A)** Schematic display of repeat arrays of Bat1 (grey polygons), MOrTL1 (dark blue hexagons) and MOrTL2 (dark blue vases). Also displayed are the chimeras containing five repeats of MOrTL1 (repeats 3–7) or MOrTL2 (repeats 2–6) in place of repeats 6–10 of Bat1. BSRs of repeats are given in each case, with an asterisk for repeat 20 of Bat1, which lacks an amino acid at position 13 with respect to the consensus sequence. Binding elements (BEs) for each TALE-like chimera were predicted using the TALE code and are given below the cartoon display in each case, with dark blue for bases in the test positions. The off-target sequence, designed to bear mismatch bases for repeats 6–10 of each construct based on the TALE code, for all Bat1 derived proteins (BEBat1$_{GGTTG}$) is shown below with red for bases in the test positions **(B)** Repeat alignments of consensus Bat, MOrTL1 and MOrTL2 repeats as shown in Figure 1C. Amino acids conserved between Bat1 and the MOrTLs are highlighted with blue font letters. Electrophoretic mobility shift assays **(C, E)** 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 native 6% polyacrylamide gel. For the competition assays (E), competitor DNA was added in excess as indicated. In each case the designation of the protein used is underlined, the probe italicized and the competitor bold and italicized. **(D, F)** The interaction between the Bat1-MOrTL chimeras and their predicted on- or off-target DNAs (a) was quantified using microscale thermophoresis. The bound fraction is shown on the *y*-axis against the protein concentration. Standard deviation for three replicates is indicated. Measurements were made with 20% LED and 30% laser power. The dark blue line indicates the K_D fit.

seen in Figure 2E this was indeed observed in every case supporting our hypothesis that repeats of both MOrtTL1 and 2, like Bat1 repeats, have TALE-code-consistent base preferences. However, the discriminating power of the test repeats in the MOrTL2 chimera was lower than that of the other proteins, since a 200x excess of off-target DNA was able to quench on-target binding by 40% relative to the no-competitor lane (quantifications: Supplementary Figure S6). Additionally the protein–DNA interactions with the off-target probe were quantified with MST (Figure 2F, Supplementary Figure S7). The interaction of $Bat1_{M1\ (2–6)}$ with its off-target $BE_{Bat1\ GGTTG}$ was determined to have a K_D of >2300 nM and thus was 19 times lower in affinity than the on-target interaction. In contrast the K_D of the $Bat1_{M2\ (3–7)}$ $BE_{Bat1\ GGTTG}$ interaction was 201 nM, indicating an interaction only about half as strong as the on-target interaction. In each case on-target interactions are stronger than off-target consistent with TALE code base preference but there are differences in discriminating power. The EMSA and MST data together suggest that MOrTL1 and 2 repeats both mediate TALE code base preferences but differ in their discriminating power.

In vivo assays support *in vitro* findings on Bat1-MOrTL chimera DNA binding properties

To study the DNA recognition properties of MOrTL repeats *in vivo* we adapted a TALE-based NOT gate in *E. coli* (24) to serve as a repressor reporter. In this system TALE-like proteins are tested for their ability to bind a constitutive *Trc* promoter and thereby repress expression of a downstream *mCherry* reporter (pCherry). Another plasmid (pBT102) carries either the test TALE-like (Bat1, dTALE or chimera), a *GFP* CDS (negative control) or a positive control dTALE. The positive control TALE is one previously designed and tested for the *lac* operon, which forms part of the *Trc* promoter upstream of the mCherry CDS. The negative-control is simply a constitutively expressed GFP not expected to bind DNA or mediate any repression of the mCherry reporter (this is unconnected to the use of GFP as a secondary reporter in one of the assay set-ups explored by Politz *et al.* (24)). pCherry and pBT102 plasmids are co-transformed into *E. coli* cells and mCherry flourescence is measured in liquid cultures inoculated from the transformants. The reduction of mCherry fluorescence of colonies arising from test and control co-transformations provides a measure of the strength of the interaction of the tested TALE-like and their affinity to the BE in the *Trc* promoter. Lower fluorescence indicates a stronger interaction. The experimental set-up is illustrated in Figure 3A.

Bat1, $Bat1_{M1\ (3–7)}$ and $Bat1_{M2\ (2–6)}$ were tested against cognate reporters and fold repression was calculated relative to the *GFP* negative control. We found that Bat1 mediated 3.4-fold repression while the chimeric $Bat1_{M1\ (3–7)}$ and $Bat1_{M2\ (2–6)}$ mediated 2-fold and 1.8-fold repression, respectively. Thus $Bat1_{M2\ (2–6)}$, which binds its target BE with near identical affinity to $Bat1_{M1\ (3–7)}$ *in vitro* (Figure 2D), shows a slightly weaker repression *in vivo* (Figure 3D). This may reflect the lower discriminatory power of the MOrTL2 repeats observed *in vitro* (Figure 2E, F) leading to more off-target binding across the *E. coli* genome quenching the re-

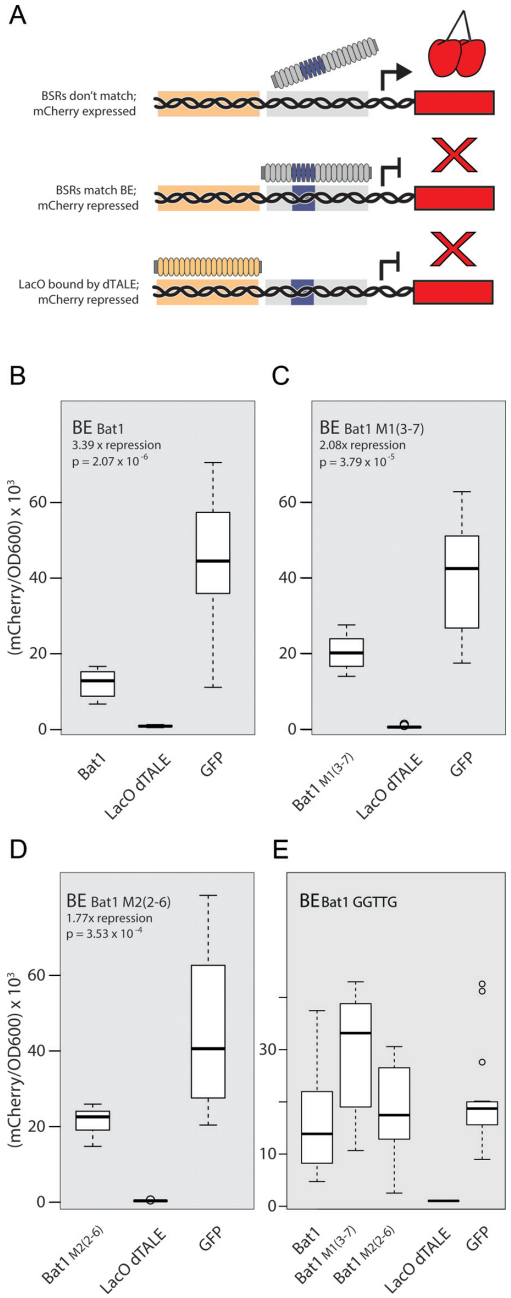


Figure 3. An *in vivo* reporter confirms that MOrTL repeats embedded in a Bat1-repeat array recognize predicted binding elements. **(A)** Schematic display of the repressor assay: mCherry reporter and expression plasmids encoding TALE-likes are co-transformed into *E. coli*. TALE-like chimeras consist of TALE/Bat-repeats (grey ovals) and MOrTL-repeats (dark blue ovals). If the TALE-like binds the given BE (blue rectangle) it should repress the mCherry promoter, observed as a reduction in mCherry fluorescence (red rectangle; cherries). A dTALE that binds the *lac* operon (LacO, orange box) within the mCherry promoter provides a positive control for each reporter. **(B–D)** Box and whisker plots show mCherry fluorescence values for Bat1, $Bat1_{M1\ (3–7)}$ and $Bat1_{M2\ (2–6)}$ tested against reporters bearing corresponding BEs (designation across the top of each plot), normalized to cell density (OD600) and compared to positive (LacO dTALE) and negative (GFP) control expression plasmids. An off-target reporter was created with mismatch bases for repeats 6–10 of each construct based on the TALE code and tested with all test constructs in the same system **(E)**. Fold repression, based on median values, and *P*-values of a two-tailed *t*-test with unequal variances comparing test and GFP samples are given in the top left corner of each plot. $N = 16$ in each case.

pression effect to some extent. We tested the specificity of these *in vivo* interactions assays with the *GGTTG* off-target reporter in each case and showed that the reporter was not repressed by Bat1 or either of the chimeras relative to the GFP control (Figure 3B–E). Overall there is thus clear evidence that DNA binding of MOrTL repeats embedded in a Bat1 repeat array is sequence specific with base preferences consistent with the TALE code.

MOrTL1 and MOrTL2 differ in their compatibility with TALE repeats

In the interests of potential biotechnological applications and to gain further fundamental information on MOrTL1 and 2 repeat properties we created additional chimeras where MOrTLs repeats are embedded in TALE repeat arrays. Specifically we used a dTALE designed to target the same DNA sequence as Bat1 (dTALE-Bat1). The MOrTL repeats chosen for the TALE chimeras were based on ease of primer placement for cloning, this resulted in the same set of five MOrTL1 repeats being taken (dTALE-Bat1$_{M1\,(3–7)}$) but a different set of MOrTL2 repeats (dTALE-Bat1$_{M2\,(4–8)}$). Designs are illustrated in Figure 4A, and construct sequences in Supplementary Figure S10.

As for the Bat1 chimeras we predicted BEs for the TALE-MOrTL chimeras using the TALE code (Supplementary Table S1). We tested purified proteins at 200 nM against these BEs in EMSAs (Figure 4C), revealing a single shift indicative of 1-to-1 DNA binding. This was followed by MST measurements to determine K_D values: 437 nM for the MOrTL1 chimera dTALE-Bat1$_{M1\,(3–7)}$ with its cognate BE and over 5410 nM for the MOrTL2 chimera dTALE-Bat1$_{M2\,(4–8)}$ (Figure 4D, Supplementary Figure S7). While the affinity of the dTALE-MOrTL1 chimera for its target is low compared to what one might expect for a TALE repeat array, this array is rich in Ile BSRs known to mediate low affinity DNA binding (15). However the affinity of the MOrTL2-TALE chimera falls far below the range of measured on-target TALE-like DNA binding interactions. We therefore tested whether these interactions are indeed sequence specific using predicted off-target BEs (note that the off-target used for the MOrTL2 chimera dTALE-Bat1$_{M2\,(4–8)}$ is distinct from the off-target for the other constructs due to the particular BSR composition of the MOrTL2 repeats in this construct (Figure 4A)). EMSA competition assays (Figure 4E) revealed that the on-target binding shift for dTALE-Bat1$_{M2\,(4–8)}$ can be depleted just as easily by the off- as the on-target oligonucleotides. MST with predicted off-target oligonucleotides was also carried out (Figure 4F, Supplementary Figure S7). These tests show that dTALE-Bat1$_{M1\,(3–7)}$ is highly discriminating, with the upper plateau of DNA binding to the off-target BE not even reached at 14 000 nM in the MST measurements (Figure 4F). In contrast the K_D of dTALE-Bat1$_{M2\,(4–8)}$ interacting with BE$_{Bat1\ TTGGT}$ was 6388 nM (Figure 4F), not much weaker than the on-target interaction (Figure 4D). EMSA competition assays and MST thus both support the idea that dTALE-Bat1$_{M2\,(4–8)}$ discriminates poorly between on- and off-target sequences.

In vivo assays support the hypothesis that MOrTL2 repeats are incompatible with TALE repeats

When the same TALE-MOrTL chimeras were tested *in vivo* with the repressor reporter we found similar results. dTALE-Bat1 and the MOrTL1 chimera dTALE-Bat1$_{M1\,(3-7)}$ performed similarly, repressing their reporters 9.5- and 11.6-fold, respectively (Figure 5A,B). In contrast the MOrTL2 chimera dTALE-Bat1$_{M2\,(4–8)}$ mediated only 1.6-fold repression (Figure 5C).

We considered that the poor performance of MOrTL2 repeats 4–8 in a TALE repeat array (dTALE-Bat1$_{M2\,(4–8)}$) may be due to an unfortunate choice of the specific repeats chosen for this construct compared to the Bat1 chimera that contained MOrTL2 repeats 2–6 (Bat1M2 (2–6)). By contrast the same MOrTL1 repeats were used for Bat1 and TALE chimeras. So we created new TALE chimeras for MOrTL1 and MOrTL2. In these new chimeras we took the same MOrTL2 repeats as had previously been used in the Bat1 chimera (repeats 2–6; Figure 5E), and from MOrTL1 we took a different set of repeats (repeats 2–6 versus 3–7 previously; Figure 5D). We tested these in the repressor assay against on- and off-target reporters. These results mirrored the results from the first set of chimeras with the new MOrTL1 chimera mediating 7.1-fold repression of its on-target reporter (Figure 5D), compared to 1.6-fold for the MOrTL2 chimera on its on-target reporter (Figure 5E). In both cases no repression was observed for off-target reporters.

MOrTL1 and 2 repeats both mediated sequence-specific DNA binding interactions of similar strength in the context of a Bat1 repeat array, though the sequence discriminating power of the MOrTL2 chimera was lower (Figures 2C–F and 3C,D). In contrast only the MOrTL1 repeats performed well in a TALE repeat array while MOrTL2 repeats mediated very weak and barely sequence specific DNA binding in a TALE repeat array independent of the particular set of MOrTL2 repeats taken for the chimera. This suggests an incompatibility between MOrTL2 repeats and the surrounding TALE repeat array. Consensus MOrTL1 and 2 repeats are both overall 35% identical to a consensus TALE repeat conforming to that used in our dTALEs. However, conserved residues are at different positions in each case (Figure 4B). Thus differences in compatibility are not surprising. The higher discriminatory power of MOrTL1 repeats (Figure 2E and F) and their compatibility with both Bat1 (Figures 2 and 3) and TALE (Figures 4 and 5) repeats makes them better suited for integration into TALE-like repeat arrays for biotechnological applications.

Functional differences between MOrTL1 and MOrTL2 chimeras are not due to differences in protein stability

We considered that the different functional properties of the MOrTL1 and MOrTL2 chimeras and especially the poor functioning of MOrTL2 repeats in a TALE repeat array might be the consequence of different protein stabilities. To this end we defined melting points for all proteins *in vitro*. The results, shown in Table 1 reveal similar melting points for Bat1 and the two corresponding MOrTL chimera derivatives, and for dTALE-Bat1 and its two corresponding MOrTL chimera derivatives. While MOrTL1 and

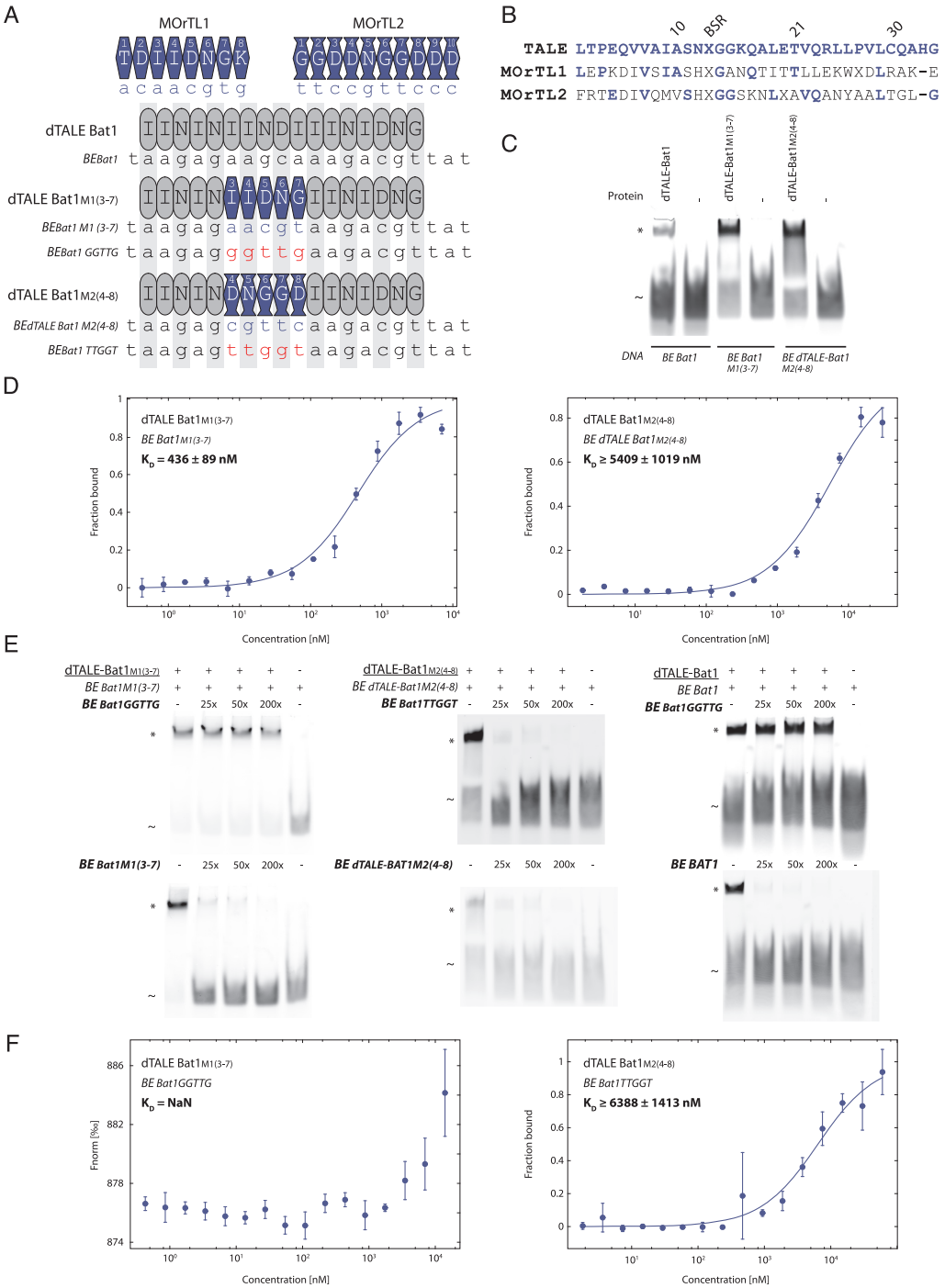


Figure 4. TALE-MOrTL chimeras proteins bind predicted target sequences *in vitro*. **(A)** Schematic display of repeat arrays of dTALE-Bat1 (grey ovals), MOrTL1 (dark blue hexagons) and MOrTL2 (dark blue vases). Also displayed are the chimeras containing five repeats of MOrTL1 (repeats 3–7) or MOrTL2 (repeats 4–8) in place of repeats 6–10 of Bat1 (grey). BSRs of repeats are given in each case. Binding elements (BEs) for each TALE-like chimera were predicted using the TALE code and are given below the cartoon display in each case, with blue bases in the test positions. Off-target sequence BEBat1 $_{GGTTG}$ or BEBat1 $_{TTGGT}$ for the dTALE-Bat1 derived proteins are shown below with red for bases in the test positions **(B)** Alignment of consensus TALE, MOrTL1 and MOrTL2 repeats as shown in Figure 1C. Conserved amino acids are highlighted with dark blue letters. Electrophoretic mobility shift assays **(C, E)** 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. For the competition assays (E) the unlabelled competitor DNA (A) was added in excess as indicated. The off-target sequence was designed to bear mismatch bases for repeats 6–10 of each construct based on the TALE code (BEBat1$_{GGTTG}$ for Bat1$_{M1\,(3–7)}$ and BEBat1$_{TTGGT}$ for dTALE-Bat1$_{M2\,(4–8)}$). In each case the designation of the protein used is underlined, the probe italicized and the competitor bold and italicized. **(D, F)** The interaction between the TALE-MOrTL chimeras and their predicted on- and off target boxes was quantified using microscale thermophoresis. The bound fraction is shown on the *y*-axis against the protein concentration. Standard deviation for three replicates is indicated. Measurements were made with 20% LED and 30% laser power. The dark blue line indicates the K_D fit.

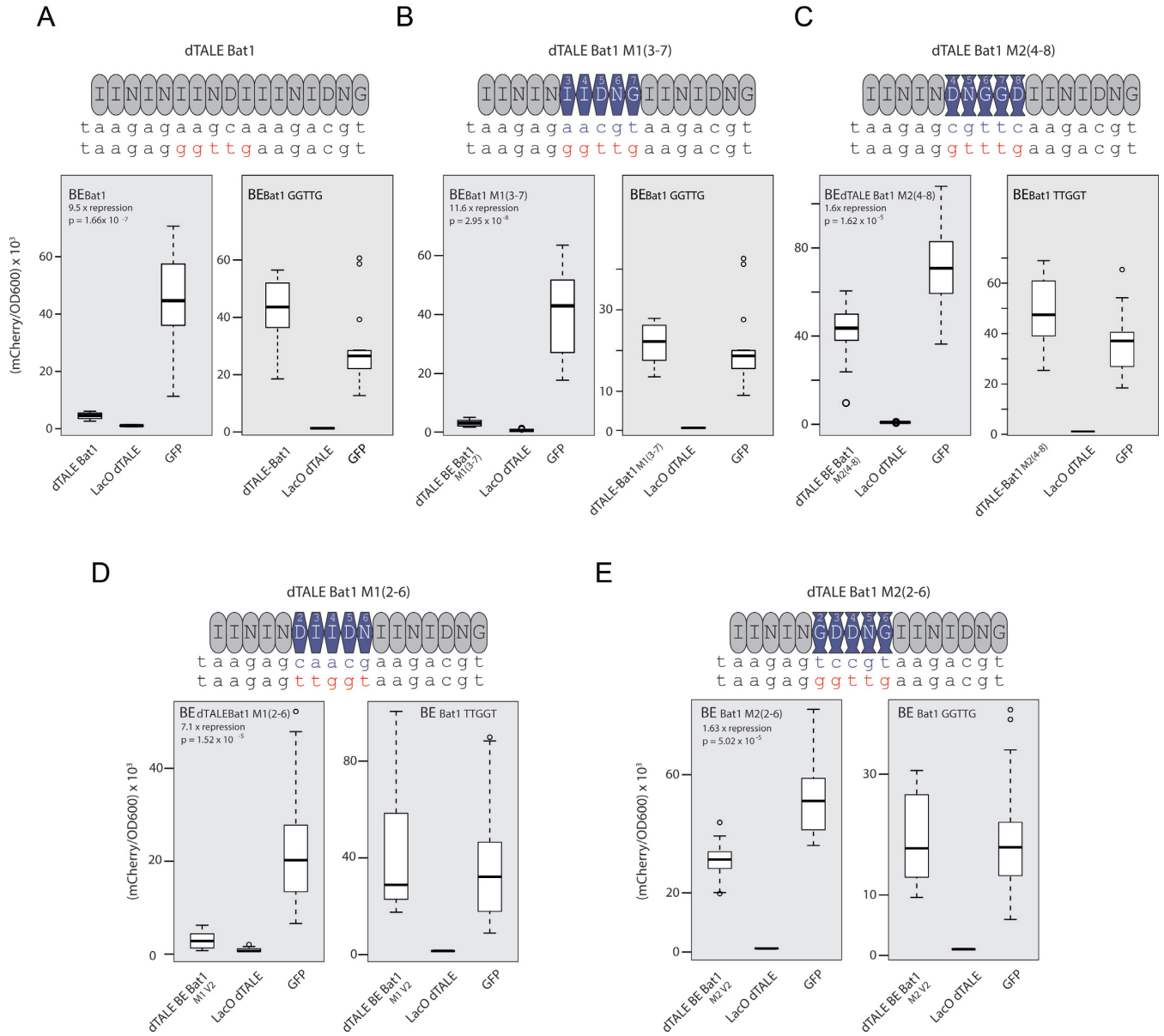


Figure 5. The repressor assay provides evidence for an incompatibility between MOrTL2 and TALE repeats. dTALE-Bat1 **(A)**, MOrTL1 **(B,D)** and MOrTL2 **(C,E)** chimeras were tested against cognate on- and off- target reporters in the repressor assay (Figure 3A). Box and whisker plots show mCherry fluorescence 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. Note that because dTALE-Bat1 and dTALE $Bat1_{M1\ (3–7)}$ were assayed in parallel on their common off-target reporter and the LacO dTALE and GFP control values are thus the same in each plot (A, B off-target reporters).

MOrTL2 do not seem to have a strong impact on the melting points of corresponding chimeras we found a consistent difference between all Bat1 and TALE constructs. All Bat1–derived proteins showed melting points about 15°C higher than all TALE-derived constructs. This might be indicative of a greater thermal stability for Bat proteins compared to TALEs, and consistent with this TALE nucleases have been shown to function poorly at 37°C compared to 30°C (33). This is, however, not relevant to our present characterisation of MOrTL repeats. These data suggest that the introduction of MOrTL repeats does not have a destabilising effect on the Bat1 or TALE proteins and that the incompatibility suggested between MOrTL2 and TALE repeats has a different cause.

Table 1. Comparison of protein melting points of Bat1, dTALE-Bat1 and their MOrTL chimeras

Protein	Melting point
Bat1	44.3 ± 0.1°C
$Bat1_{M1\ (3–7)}$	44.6 ± 0.1°C
$Bat1_{M2\ (2–6)}$	45.6 ± 0.3°C
dTALE-Bat1	31.7 ± 0.1°C
$dTALE\text{-}Bat1_{M1\ (3–7)}$	28.1 ± 0.7°C
$dTALE\text{-}Bat1_{M2\ (4–8)}$	28.4 ± 0.3°C

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 and in the case of MOrTL1 also a TALE repeat array. DNA binding properties seem to be broadly conserved among repeats of TALEs, RipTALs, Bats, MOrTL1 and MOrTL2. By this we mean sequence specific DNA binding with each repeat binding a single base and specificity determined by position 13 with specific BSRs having largely the same base preference in any TALE-like repeat. This functional conservation is suggestive of a structural conservation allowing each repeat to contact a single nucleotide and for position 13 to mediate base specific interactions. A broad functional conservation, together with sequence similarity are suggestive of a conserved structure but further evidence is obviously desirable. There is already evidence in support of a high degree of structural similarity among TALEs and Bats: crystal structures for Bat1 (alternatively termed BuD), with and without its DNA target, have been solved (19) and are similar to analogous structures for TALEs PthXo1, AvrBs3 and dTALE dHax3 (20,34–35), in so far as all proteins form a right-handed super helix that contracts tightly around the B-form DNA helix. The structures are not identical and one of the most noticeable differences is the double-band of electropositive residues allowing the Bat1 repeat array to interact with the phosphate backbone of both DNA strands (19) compared to the single band of TALEs (20,34–35). However, the key structural properties responsible for the 1-to-1 base specific binding behaviour of TALE-likes are similar in TALE and Bat1 structures. The repeats of Bat1 and TALEs are helix-loop-helix structures with BSRs located in the 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 evaluate this hypothesis, we generated models of the functionally validated chimeras $Bat1_{M1\ (3–7)}$ and $Bat1_{M2\ (2–6)}$ using the structure of Bat1 (BuD) bound to DNA as a template. Both models show structural properties similar to those described earlier for TALE-like repeats (Figure 6A and B; 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 interface over time we conducted molecular dynamics (MD). Both independent simulations for predicted structures of MOrTL1 and MOrTL2 repeats embedded in Bat1 revealed highly stable complexes between the proteins and their target DNA, seen in the values for atomic distances between protein and DNA partners (Figure 6A and B). Measuring base–BSR distances during MD simulations showed that under the simulated conditions such interactions were stable and comparable for Bat1 and MOrTL derived repeats (Supplementary Tables S3 and S4). Overlays of identical BSR–base interactions taken from repeats of different origins show that nearly identical interactions were sampled in each case (Figure 6C). Thus the simulated structures and DNA-binding interactions are consistent with our *in vitro* and *in vivo* DNA-binding data. We also wanted to see if the same intra-molecular interactions stabilise MOrTL repeats as have been observed for other TALE-likes. Hydrophobic interactions between specific residues have been predicted to stabilise Bat1 repeat arrays (19). We examined MOrTL1 repeats from our homology model since the DNA binding properties, and thus presumably repeat structures of $Bat1_{M1\ (3–7)}$, more closely resembled Bat1 than did repeats of MOrTL2. Intra- and inter-repeat interactions were indeed present during simulation, and in similar positions on the repeat, but mediated by different residues, than those found in Bat1 repeats (e.g. Val22 of Bat1 repeats versus Leu22 of MOrTL1 repeats (19)) (Figure 6D). Similarly, stabilising interactions are present for TALE repeats at the same or neighbouring positions as those predicted for MOrTL1 repeats but mediated by different residues (36). This would suggest that while TALE-like repeats adopt very similar structures some structural details and particularly the residues involved in stabilising interactions are likely to differ between groups.

Taken together, it seems likely that repeats of TALEs, RipTALs, Bats, MOrTL1 and MOrTL2, adopt similar structures, facilitating a conserved DNA-binding mechanism. We suggest therefore that the designation TALE-like should refer to proteins bearing an array of repeats broadly conserved both functionally and structurally with those of TALEs.

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 in some cases. 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, including but not limited to those used to create the consensus repeats of Figure 1C (see Supplementary Table S5 for list of all 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 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.

It is clear that some positions seem to be more conserved than others both within and between groups. We calculated percentage conservations at each position of each separate alignment (Supplementary Table S6) and all positions at least 75% conserved within all five TALE-like groups are shaded grey (Figure 7, Supplementary Figure S6). Many of these cluster around positions 5–19 (Figure 7B), which is logical considering their proximity to the crucial BSR position (see Figure 6) in addition to the constraint inherent in formation of an alpha-helical structure.

These positions are highly conserved within groups but not necessarily between groups. In fact only three positions are highly-conserved across all five groups (red-lettering; Figure 7B). In contrast other positions are highly conserved

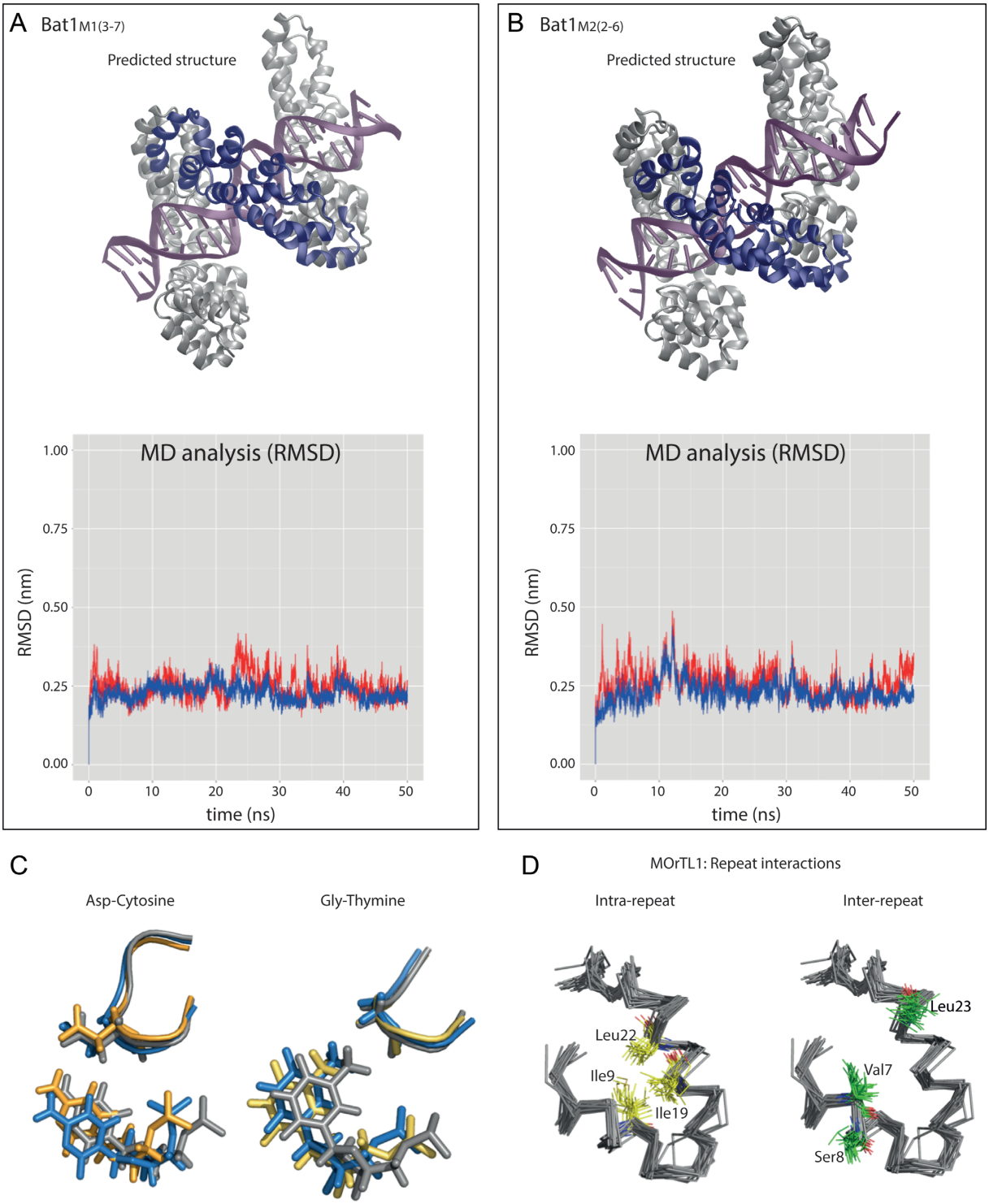


Figure 6. Homology models supported by molecular dynamics (MD) simulations of Bat1-MOrTL chimeras bound to cognate BEs, correspond to known TALE-like structures. Homology models of $Bat1_{M1\ (3–7)}$**(A)** and $Bat1_{M2\ (2–6)}$**(B)** were built using PDB entry 4cja as template structure with template DNA structures mutated *in silico* in order to match the optimal bases for both constructs. The resulting protein–DNA complexes were subjected to 50 ns molecular dynamics simulations. Single snapshots of the models bound to DNA (purple) are shown as well as RMSD read outs from the simulations for DNA (blue traces) and protein C-alpha backbone (red traces). 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. **(C)** Using these models single snapshots of BSR–base interactions were taken from repeats of Bat1 (grey), MOrTL1 (blue) and MOrLT2 (yellow) with Asp or Gly at the BSR position. **(D)** Interactions between MOrTL1 repeats in $Bat1_{M1\ (3–7)}$ were also observed to be mediated by certain residues both within (yellow) and between repeats (green).

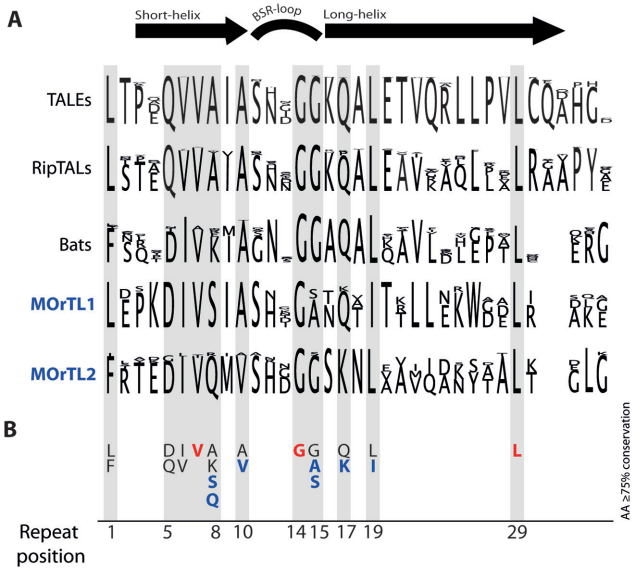


Figure 7. TALE-like repeat alignments show an underlying pattern of sequence conservation around the BSR position. **(A)** Repeat alignments and corresponding sequence logos were derived from representative core repeat arrays from each TALE-like group characterized so far (Supplementary Table S5), using CLC Main Workbench 7. In the sequence logo the total height in each column correlates to conservation at that position. Percentage conservations were calculated for each position (Supplementary Table S6). Positions that are at least 75% conserved in all groups are shaded grey. Predicted secondary structural features are indicated above the alignment (arrows indicate alpha-helices). The most common residues for each TALE-like group at the highly conserved (grey-shaded) positions are indicated underneath the logos **(B)**, and positions within the repeat are numbered. Among these the residues unique to MOrTL1 or MOrTL2 are highlighted with blue lettering, whilst red lettering highlights those positions fully conserved across TALE-likes.

within groups but different residues are found in different TALE-like groups (e.g. positions 8 and 15; Figure 7B). This could be useful as a tool to examine different selective pressures constraining sequence evolution within and between different TALE-like repeat groups.

There is little polymorphism around the BSR of TALE, RipTAL and Bat repeats (Figure 7A). This is limiting for repeat engineering efforts because these residues are especially likely to exert significant influence over DNA binding properties. Previous efforts to exploit natural diversity for TALE-like repeat engineering may have been hindered by the lack of diversity in this key region. Furthermore any effort to create sequence-diverse TALE-likes less prone to repeat recombination (37,38) based on natural diversity will be held back by the lack of sequence diversity in this region, although one approach using codon redundancy to boost the sequence diversity was able to overcome the repeat loss issue in lentiviral delivery vectors (39). Repeats of MOrTL1 and 2, however, have unique residues in otherwise highly conserved positions in this region around the BSR (Figure 7B; dark blue-lettering). At positions, 10, 15, 17 and 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.

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). Blocks of five MOrTL1 repeats, embedded in Bat1 or TALE repeat arrays, were competent to discriminate TALE-code-predicted on-target BEs, from off-target sequences (Figures 2–5). MOrTL2 repeats share no derived sequence features with those of MOrTL1 (Figure 1C, D) and also demonstrated some striking functional differences. MOrTL2 repeats in a Bat1 context mediated strong DNA binding similar to the MOrTL1-Bat1 chimera (Figure 3) and demonstrated a clear base preference (Figure 2D–F). However, there was a difference in specificity in so far as the discriminating power of the repeats is concerned. We see specificity as formed of two components: base-preference and discriminating power. The base-preference of a repeat is a statement of its relative interaction strengths for different bases. The absolute values for each interaction are not important only the ratios. However, the contribution of a particular repeat to the selection of one binding site over another for the whole repeat array is its discriminating power. This comes from the absolute interaction strength for a given repeat binding a given base, in the context of the whole repeat array. If the positive contribution from a best-match interaction or the negative contribution from a mismatch is strong enough, it can make a decisive contribution to target site discrimination. This difference between base preference and discriminating power can be understood for TALE-likes by referring to previous studies on TALE repeat specificity. The SELEX method which uses repeated rounds of selection to identify the preferred target site of an array has consistently shown that every repeat in a TALE array exerts a preference corresponding to the TALE code (11,33). Base preference is constant across all positions in the array (though there are minor qualifications to this (40)). In contrast several lines of evidence have shown that the discriminating power of TALE repeats reduces past repeat 10 (15,41). To us the behaviour of $Bat1_{M2\,(2–6)}$ is suggestive of MOrTL2 repeats having a base preference consistent with the TALE code but low discriminating power. In addition to this possible difference in discriminating power between MOrTL1 and 2 repeats there is also the clear compatibility difference with TALE repeats (Figures 4 and 5). dTALE-MOrTL1 chimeras mediated strong DNA binding (Figure 4) and reporter repression (Figure 5), clearly discriminating on- from off-targets (Figure 4 D–F). dTALE-MOrTL2 chimeras mediated weak and barely sequence-specific DNA binding (Figure 4) and weak reporter repression compared to the other dTALE constructs (Figure 5). Since during all these tests on- and off-target BEs were predicted based on the TALE-code we believe the data demonstrate that MOrTL1 and MOrTL2 repeats are able to mediate DNA binding with a base preference adhering to the TALE code but that there a functional differences between MOrTL1 and MOrTL2 repeats.

We can use this information to make a refined description of the TALE-likes, a grouping until now defined only loosely and inconsistently. We would suggest the designation TALE-like refer only to any protein bearing a tandem

array of 33–35 amino acid repeats mediating 1-to-1 DNA binding with position 13 determining DNA binding specificity in accordance with the TALE code. Repeat arrays of such proteins should also structurally resemble those of TALEs insofar as forming a super helix with each repeat formed of paired alpha-helices.

Comparison of TALE-like repeat sequences may improve understanding of TALE-like repeat structure and the connection between structure and DNA binding properties. This improved understanding will in turn benefit TALE repeat engineering efforts. Until now assumptions on the roles of different TALE or TALE-like repeat residues, apart from the RVD, have been based on structural models (19,34–35). Hypotheses about residue roles remain largely untested in a wet lab setting though molecular dynamics simulations have provided some insights (42). 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 were suggested to form an electropositive stripe along the TALE superhelix and to form hydrogen bonds to the phosphate backbone of the DNA (35). In Bat, MOrTL1 and MOrTL2 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 affect 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. Until 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 (19), 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.

There are additional insights to be gained by comparing sequence conservation within groups to conservation between groups. Certain positions are highly conserved in repeats of every TALE-like group (grey shading Figure 7). However at some of these positions different residues are found in several of the different TALE-like groups (Figure 7B). If one assumes that these sequences should encode protein domains with analogous functions then this observation might suggest that some positions are constrained at the level of array function more than at the level of individual repeat function. That is to say that for some reason, such as inter-repeat interactions, these positions must be conserved within any given array. Alternatives may be equally good as long as they are borne by all repeats in the array. If this were the case then those residues indicated in Figure 7B may be particularly likely to play a role in the compatibility or incompatibility of repeats from different TALE-like groups.

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 they share a common ancestor. To determine whether MOrTL1 and MOrTL2 share this common ancestor requires the identification of a plausible TALE repeat progenitor sequence to use as an outgroup for creation of a phylogenetic tree.

What struck us most clearly when comparing TALE-like repeat sequence diversity (Figure 7) was that TALE repeats display by far the lowest sequence diversity. This sequence conservation is even more apparent when examining individual TALE repeat arrays as opposed to the pooled sequence logo presented in Figure 7A. There is almost no non-RVD repeat polymorphism between the repeats of TALE AvrBs3 for example (Supplementary Figure S11). The low repeat polymorphism among TALEs is thus exceptional and evidence of particular selection pressures or mechanisms of sequence evolution relevant to TALEs only.

Considering the full sequence diversity of TALE-like repeats may also assist with the identification of further TALE-likes. While TALE repeats are highly conserved across most positions only three residues are conserved across all groups (Figure 7B, red lettering): 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 consensus definition of TALE-like repeats as conforming to the sequence motif $X_6\mathbf{V}X_6\mathbf{G}X_{13}\mathbf{L}X_{4-6}$. This motif may be useful as a basis for identifying additional TALE-likes from database DNA sequences, especially if combined with secondary structure predictions to identify the necessary two alpha helices with intervening BSR loop.

By demonstrating that MOrTL repeats mediate DNA binding behaviour analogous to that of other TALE-like repeats (Figures 2–5) 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.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

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. C.K. carried out thermo stability experiments. O.D.L. and C.W. prepared the manuscript with input from all other authors.

FUNDING

Deutsche Forschungsgemeinschaft [SFB924, DFG/LA 1338/6-1]; Two Blades Foundation. Funding for open access charge: Deutsche Forschungsgemeinschaft [SFB924].

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

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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, Struct. Biol. Commun.*, **70**, 87–91.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. Deng,D., Yin,P., Yan,C., Pan,X., Gong,X., Qi,S., Xie,T., Mahfouz,M., Zhu,J.-K., Yan,N. *et al.* (2012) Recognition of methylated DNA by TAL effectors. *Cell Res.*, **22**, 1502–1504.
15. 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.
16. Hubbard,B.P., Badran,A.H., Zuris,J. a, Guilinger,J.P., Davis,K.M., Chen,L., Tsai,S.Q., Sander,J.D., Joung,J.K. and Liu,D.R. (2015) Continuous directed evolution of DNA-binding proteins to improve TALEN specificity. *Nat. Methods*, **12**, 939–942.
17. 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.
18. 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.
19. 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.
20. 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**, 1707–1716.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. Hess,B., Kutzner,C., Van Der Spoel,D. and Lindahl,E. (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435–447.
27. 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.
28. 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.
29. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.
30. 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.
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**, 1697–712.
33. 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*, **12**, 465–471.
34. 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.
35. 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.
36. Deng,D., Yan,C., Wu,J., Pan,X. and Yan,N. (2014) Revisiting the TALE repeat. *Protein Cell*, **5**, 297–306.
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. 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.
40. Rogers,J.M., Barrera,L.a., Reyon,D., Sander,J.D., Kellis,M., Keith Joung,J. and Bulyk,M.L. (2015) Context influences on TALE–DNA binding revealed by quantitative profiling. *Nat. Commun.*, **6**, 7440.
41. 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 *Xanthomonasoryzae* strains. *PLoS One*, **8**, e68464.
42. 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.

Supplementary Material – de Lange, Wolf *et al*., 2015

Supplementary Figures

Supplementary Tables

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.

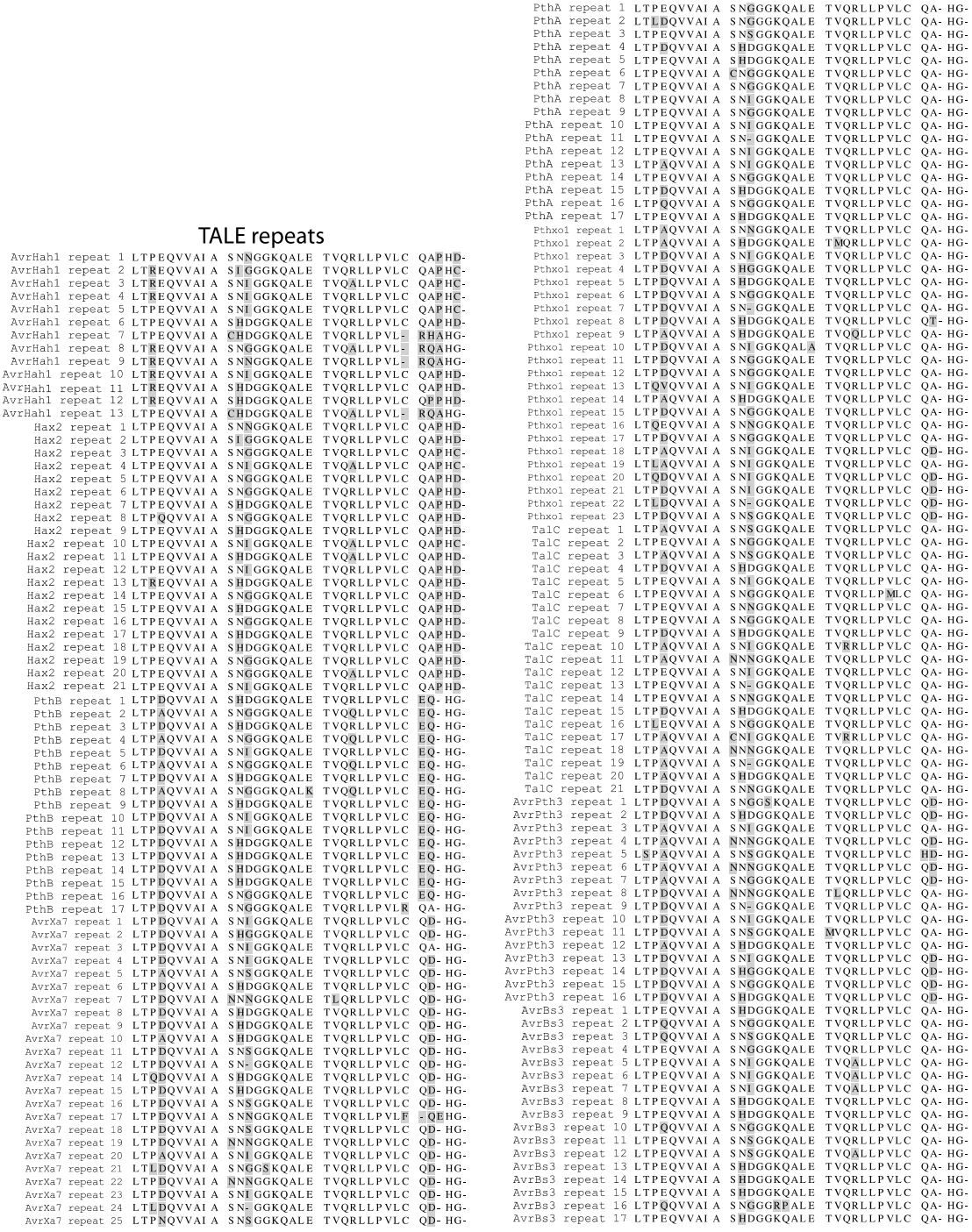
(i) MOrTL 1/ECG96326 – synthesized CDS (GenScript). BsaI restriction enzyme binding sites underlined and overlaps italicized. Start and stop codons bold.

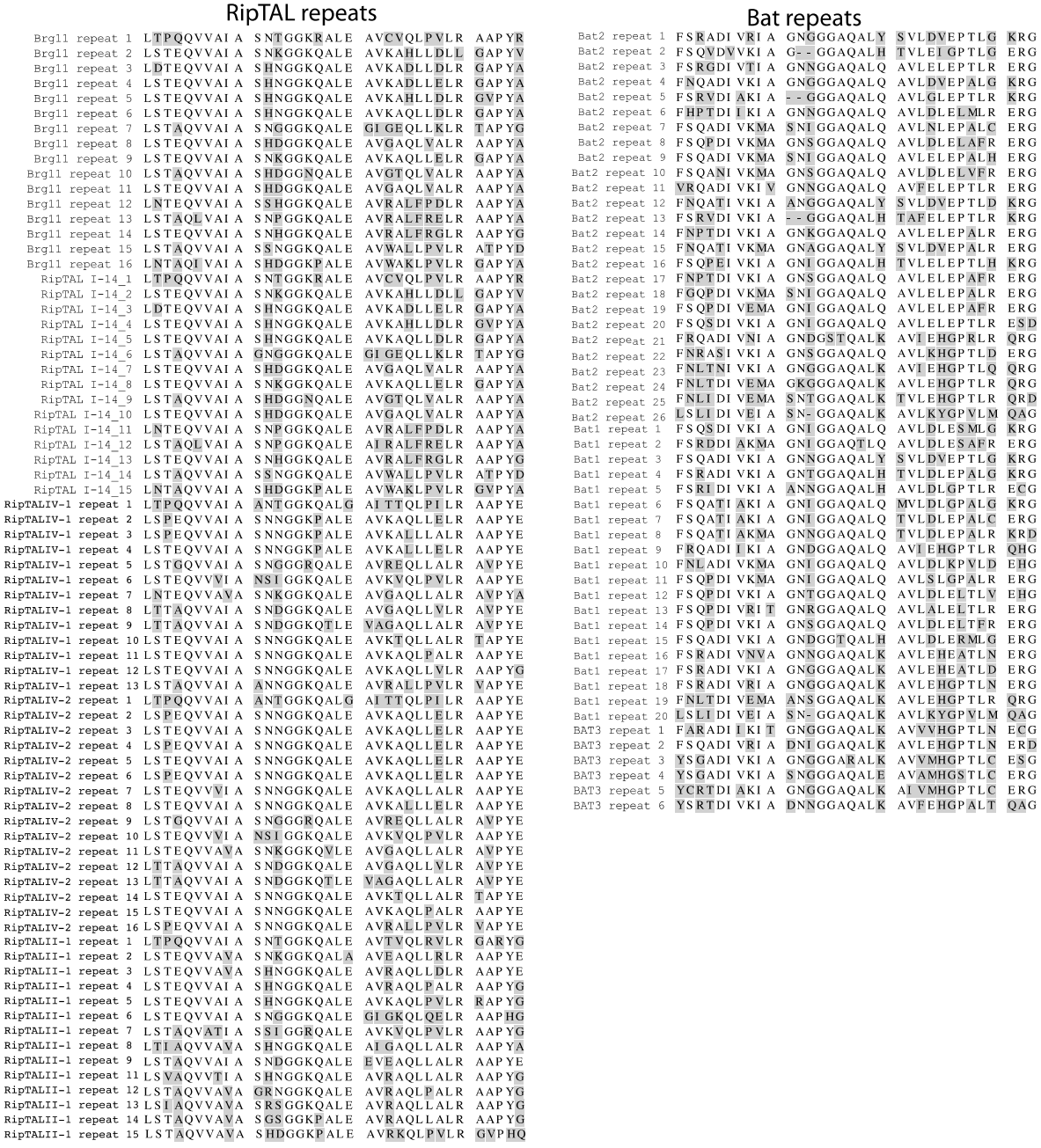
>MOrTL1_CDS_Genscript
GGTCTCA*T**ATG*GTTGGCGATCTGCGTGCGAAAGAACTGGAACCGAAAGACATTGTGA
GCATTGCCTCTAACACCGGCGCGAATAAAACGATTACCCGCCTGCTGGAAAAATGGG
GCGATCTGCGTGCCAAGGAGCTGGAACCGAAAGATATTGTCAGCATCGCCTCTCATG
ACGGCAGTAACCAGACCATTACGAAACTGCTGGAAAAATGGGATGAACTGCGCGCAA
AAGGTCTGGAACCGAAAGATATCGTGAGTATCGCATCCCACATTGGCGCTAACCAAA
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

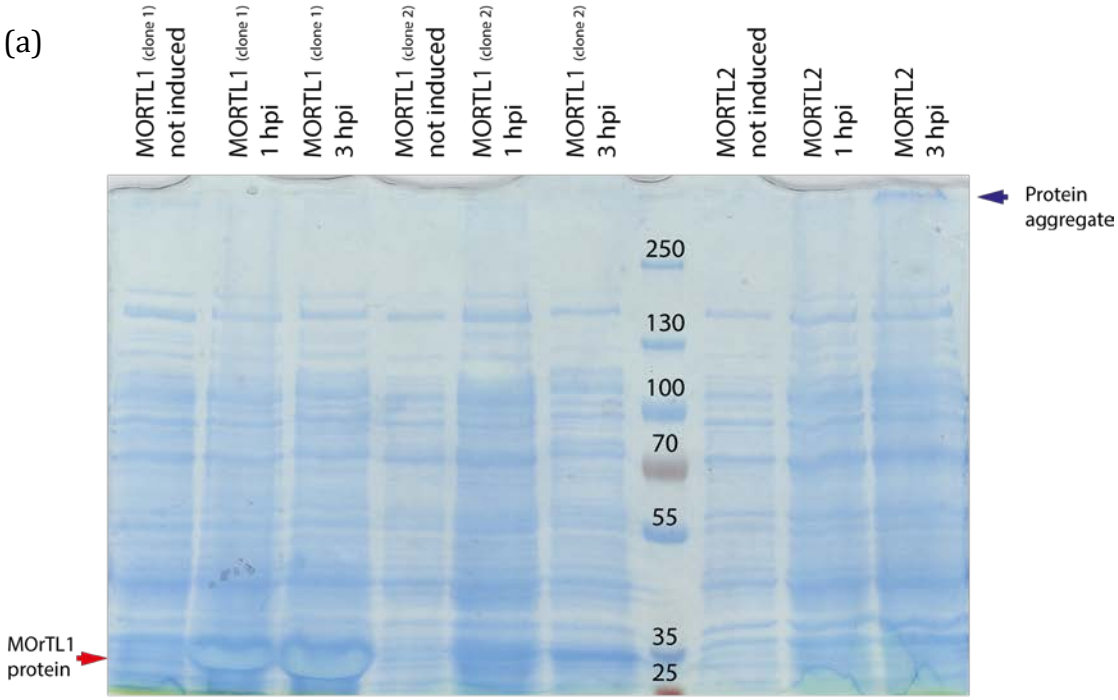
Supplementary Figure 2: Sequences of TALE, RipTAL and Bat repeats used to create the consensus repeats in Figure 1c and for the calculation of percentage identities in Figure 1d. Grey shading indicates conflicting residues.



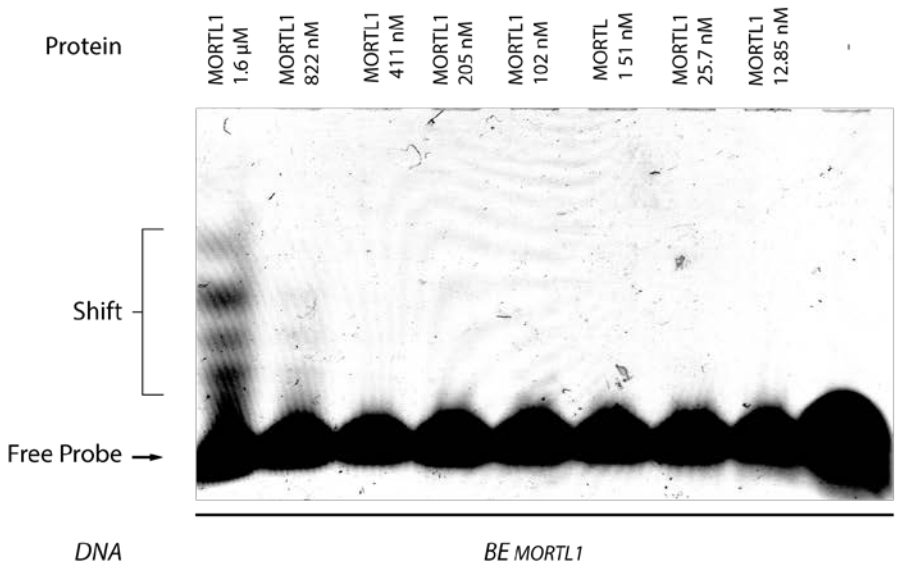

RipTAL repeats
Brgll repeat 1 LTPQQVVAI A SNTGGKRALE AVCVQLPVLR AAPYR
Brgll repeat 2 LSTEQVVAI A SNKGGKQALE AVKAHLLDLL GAPYV
Brgll repeat 3 LDTEQVVAI A SHNGGKQALE AVKADLLDLR GAPYA
Brgll repeat 4 LSTEQVVAI A SHNGGKQALE AVKADLLELR GAPYA
Brgll repeat 5 LSTEQVVAI A SHNGGKQALE AVKAHLLDLR GVPYA
Brgll repeat 6 LSTEQVVAI A SHNGGKQALE AVKAQLLDLR GAPYA
Brgll repeat 7 LSTAQVVAI A SNGGGKQALE GIGEQLLKLR TAPYG
Brgll repeat 8 LSTEQVVAI A SHDGGKQALE AVGAQLVALR AAPYA
Brgll repeat 9 LSTEQVVAI A SNKGGKQALE AVKAQLLELR GAPYA
Brgll repeat 10 LSTAQVVAI A SHDGGNQALE AVGTQLVALR AAPYA
Brgll repeat 11 LSTEQVVAI A SHDGGKQALE AVGAQLVALR AAPYA
Brgll repeat 12 LNTEQVVAI A SSHGGKQALE AVRALFPDLR AAPYA
Brgll repeat 13 LSTAQLVAI A SNPGGKQALE AVRALFRELR AAPYA
Brgll repeat 14 LSTEQVVAI A SNHGGKQALE AVRALFRGLR AAPYG
Brgll repeat 15 LSTAQVVAI A SSNGGKQALE AVWALLPVLR ATPYD
Brgll repeat 16 LNTAQIVAI A SHDGGKPALE AVWAKLPVLR GAPYA
RipTAL I-14_1 LTPQQVVAI A SNTGGKRALE AVCVQLPVLR AAPYR
RipTAL I-14_2 LSTEQVVAI A SNKGGKQALE AVKAHLLDLL GAPYV
RipTAL I-14_3 LDTEQVVAI A SHNGGKQALE AVKADLLELR GAPYA
RipTAL I-14_4 LSTEQVVAI A SHNGGKQALE AVKAHLLDLR GVPYA
RipTAL I-14_5 LSTEQVVAI A SHNGGKQALE AVKAQLLDLR GAPYA
RipTAL I-14_6 LSTAQVVAI A GNGGGKQALE GIGEQLLKLR TAPYG
RipTAL I-14_7 LSTEQVVAI A SHDGGKQALE AVGAQLVALR AAPYA
RipTAL I-14_8 LSTEQVVAI A SNKGGKQALE AVKAQLLELR GAPYA
RipTAL I-14_9 LSTAQVVAI A SHDGGNQALE AVGTQLVALR AAPYA
RipTAL I-14_10 LSTEQVVAI A SHDGGKQALE AVGAQLVALR AAPYA
RipTAL I-14_11 LNTEQVVAI A SNPGGKQALE AVRALFPDLR AAPYA
RipTAL I-14_12 LSTAQLVAI A SNPGGKQALE AIRALFRELR AAPYA
RipTAL I-14_13 LSTEQVVAI A SNHGGKQALE AVRALFRGLR AAPYG
RipTAL I-14_14 LSTAQVVAI A SSNGGKQALE AVWALLPVLR ATPYD
RipTAL I-14_15 LNTAQVVAI A SHDGGKPALE AVWAKLPVLR GVPYA
RipTALIV-1 repeat 1 LTPQQVVAI A ANTGGKQALG AITTQLPILR AAPYE
RipTALIV-1 repeat 2 LSPEQVVAI A SNNGGKPALE AVKAQLLELR AAPYE
RipTALIV-1 repeat 3 LSPEQVVAI A SNNGGKPALE AVKALLLALR AAPYE
RipTALIV-1 repeat 4 LSTEQVVAI A SNNGGKPALE AVKALLLELR AAPYE
RipTALIV-1 repeat 5 LSTGQVVAI A SNGGGRQALE AVREQLLALR AVPYE
RipTALIV-1 repeat 6 LSTEQVVVI A NSIGGKQALE AVKVQLPVLR AAPYE
RipTALIV-1 repeat 7 LNTEQVVAVA SNKGGKQALE AVGAQLLALR AVPYA
RipTALIV-1 repeat 8 LTTAQVVAI A SNDGGKQALE AVGAQLLVLR AVPYE
RipTALIV-1 repeat 9 LTTAQVVAI A SNDGGKQTLE VAGAQLLALR AVPYE
RipTALIV-1 repeat 10 LSTEQVVAI A SNNGGKQALE AVKTQLLALR TAPYE
RipTALIV-1 repeat 11 LSTEQVVAI A SNNGGKQALE AVKAQLPALR AAPYE
RipTALIV-1 repeat 12 LSTEQVVAI A SNNGGKQALE AVKAQLLVLR AAPYG
RipTALIV-1 repeat 13 LSTAQVVAI A ANNGGKQALE AVRALLPVLR VAPYE
RipTALIV-2 repeat 1 LTPQQVVAI A ANTGGKQALG AITTQLPILR AAPYE
RipTALIV-2 repeat 2 LSPEQVVAI A SNNGGKQALE AVKAQLLELR AAPYE
RipTALIV-2 repeat 3 LSTEQVVAI A SNNGGKQALE AVKAQLLELR AAPYE
RipTALIV-2 repeat 4 LSPEQVVAI A SNNGGKQALE AVKAQLLELR AAPYE
RipTALIV-2 repeat 5 LSTEQVVAI A SNNGGKQALE AVKAQLLELR AAPYE
RipTALIV-2 repeat 6 LSPEQVVAI A SNNGGKQALE AVKAQLLELR AAPYE
RipTALIV-2 repeat 7 LSTEQVVVI A SNNGGKQALE AVKAQLLALR AAPYE
RipTALIV-2 repeat 8 LSTEQVVAI A SNNGGKQALE AVKALLLELR AAPYE
RipTALIV-2 repeat 9 LSTGQVVAI A SNGGGRQALE AVREQLLALR AVPYE
RipTALIV-2 repeat 10 LSTEQVVVI A NSIGGKQALE AVKVQLPVLR AAPYE
RipTALIV-2 repeat 11 LSTEQVVAVA SNKGGKQVLE AVGAQLLALR AVPYE
RipTALIV-2 repeat 12 LTTAQVVAI A SNDGGKQALE AVGAQLLVLR AVPYE
RipTALIV-2 repeat 13 LTTAQVVAI A SNDGGKQTLE VAGAQLLALR AVPYE
RipTALIV-2 repeat 14 LSTEQVVAI A SNNGGKQALE AVKTQLLALR TAPYE
RipTALIV-2 repeat 15 LSTEQVVAI A SNNGGKQALE AVKAQLPALR AAPYE
RipTALIV-2 repeat 16 LSPEQVVAI A SNNGGKQALE AVRALLPVLR VAPYE
RipTALII-1 repeat 1 LTPQQVVAI A SNTGGKQALE AVTVQLRVLR GARYG
RipTALII-1 repeat 2 LSTEQVVAVA SNKGGKQALA AVEAQLLRLR AAPYE
RipTALII-1 repeat 3 LSTEQVVAVA SHNGGKQALE AVRAQLLDLR AAPYE
RipTALII-1 repeat 4 LSTEQVVAI A SHNGGKQALE AVRAQLPALR AAPYG
RipTALII-1 repeat 5 LSTEQVVAI A SHNGGKQALE AVKAQLPVLR RAPYG
RipTALII-1 repeat 6 LSTEQVVAI A SNGGGKQALE GIGKQLQELR AAPHG
RipTALII-1 repeat 7 LSTAQVATI A SSIGGRQALE AVKVQLPVLR AAPYG
RipTALII-1 repeat 8 LTIAQVVAVA SHNGGKQALE AIGAQLLALR AAPYA
RipTALII-1 repeat 9 LSTAQVVAI A SNDGGKQALE EVEAQLLALR AAPYE
RipTALII-1 repeat 11 LSVAQVVTI A SHNGGKQALE AVRAQLLALR AAPYG
RipTALII-1 repeat 12 LSTAQVVAVA GRNGGKQALE AVRAQLPALR AAPYG
RipTALII-1 repeat 13 LSIAQVVAVA SRSGGKQALE AVRAQLLALR AAPYG
RipTALII-1 repeat 14 LSTAQVVAVA SGSGGKPALE AVRAQLLALR AAPYG
RipTALII-1 repeat 15 LSTAQVVAVA SHDGGKPALE AVRKQLPVLR GVPHQ
Bat repeats
Bat2 repeat 1 FSRADIVRIA GNGGGAQALY SVLDVEPTLG KRG
Bat2 repeat 2 FSQVDVVKIA G--GGAQALH TVLEIGPTLG ERG
Bat2 repeat 3 FSRGDIVTIA GNNGGAQALQ AVLELEPTLR ERG
Bat2 repeat 4 FNQADIVKIA GNGGGAQALQ AVLDVEPALG KRG
Bat2 repeat 5 FSRVDIAKIA --GGGAQALQ AVLGLEPTLR KRG
Bat2 repeat 6 FHPTDIIKIA GNNGGAQALQ AVLDLELMLR ERG
Bat2 repeat 7 FSQADIVKMA SNIGGAQALQ AVLNLEPALC ERG
Bat2 repeat 8 FSQPDIVKMA GNSGGAQALQ AVLDLELAFR ERG
Bat2 repeat 9 FSQADIVKMA SNIGGAQALQ AVLELEPALH ERG
Bat2 repeat 10 FSQANIVKMA GNSGGAQALQ AVLDLELVFR ERG
Bat2 repeat 11 VRQADIVKIV GNNGGAQALQ AVFELEPTLR ERG
Bat2 repeat 12 FNQATIVKIA ANGGGAQALY SVLDVEPTLD KRG
Bat2 repeat 13 FSRVDIVKIA --GGGAQALH TAFELEPTLR KRG
Bat2 repeat 14 FNPTDIVKIA GNKGGAQALQ AVLELEPALR ERG
Bat2 repeat 15 FNQATIVKMA GNAGGAQALY SVLDVEPALR ERG
Bat2 repeat 16 FSQPEIVKIA GNIGGAQALH TVLELEPTLH KRG
Bat2 repeat 17 FNPTDIVKIA GNSGGAQALQ AVLELEPAFR ERG
Bat2 repeat 18 FGQPDIVKMA SNIGGAQALQ AVLELEPALR ERG
Bat2 repeat 19 FSQPDIVEMA GNIGGAQALQ AVLELEPAFR ERG
Bat2 repeat 20 FSQSDIVKIA GNIGGAQALQ AVLELEPTLR ESD
Bat2 repeat 21 FRQADIVNIA GNDGSTQALK AVIEHGPRLR QRG
Bat2 repeat 22 FNRASIVKIA GNSGGAQALQ AVLKHGPTLD ERG
Bat2 repeat 23 FNLTNIVKIA GNGGGAQALK AVIEHGPTLQ QRG
Bat2 repeat 24 FNLTDIVEMA GKGGGAQALK AVLEHGPTLR QRG
Bat2 repeat 25 FNLIDIVEMA SNTGGAQALK TVLEHGPTLR QRD
Bat2 repeat 26 LSLIDIVEIA SN-GGAQALK AVLKYGPVLM QAG
Bat1 repeat 1 FSQSDIVKIA GNIGGAQALQ AVLDLESMLG KRG
Bat1 repeat 2 FSRDDIAKMA GNIGGAQTLQ AVLDLESAFR ERG
Bat1 repeat 3 FSQADIVKIA GNNGGAQALY SVLDVEPTLG KRG
Bat1 repeat 4 FSRADIVKIA GNTGGAQALH TVLDLEPALG KRG
Bat1 repeat 5 FSRIDIVKIA ANNGGAQALH AVLDLGPTLR ECG
Bat1 repeat 6 FSQATIAKIA GNIGGAQALQ MVLDLGPALG KRG
Bat1 repeat 7 FSQATIAKIA GNIGGAQALQ TVLDLEPALC ERG
Bat1 repeat 8 FSQATIAKMA GNNGGAQALQ TVLDLEPALR KRD
Bat1 repeat 9 FRQADIIKIA GNDGGAQALQ AVIEHGPTLR QHG
Bat1 repeat 10 FNLADIVKMA GNIGGAQALQ AVLDLKPVLD EHG
Bat1 repeat 11 FSQPDIVKMA GNIGGAQALQ AVLSLGPALR ERG
Bat1 repeat 12 FSQPDIVKIA GNTGGAQALQ AVLDLELTLV EHG
Bat1 repeat 13 FSQPDIVRIT GNRGGAQALQ AVLALELTLR ERG
Bat1 repeat 14 FSQPDIVKIA GNSGGAQALQ AVLDLELTFR ERG
Bat1 repeat 15 FSQADIVKIA GNDGGTQALH AVLDLERMLG ERG
Bat1 repeat 16 FSRADIVNVA GNNGGAQALK AVLEHEATLN ERG
Bat1 repeat 17 FSRADIVKIA GNGGGAQALK AVLEHEATLD ERG
Bat1 repeat 18 FSRADIVRIA GNGGGAQALK AVLEHGPTLN ERG
Bat1 repeat 19 FNLTDIVEMA ANSGGAQALK AVLEHGPTLR QRG
Bat1 repeat 20 LSLIDIVEIA SN-GGAQALK AVLKYGPVLM QAG
BAT3 repeat 1 FARADIIKIT GNGGGAQALK AVVVHGPTLN ECG
BAT3 repeat 2 FSQADIVRIA DNIGGAQALK AVLEHGPTLN ERD
BAT3 repeat 3 YSGADIVKIA GNGGGARALK AVVMHGPTLC ESG
BAT3 repeat 4 YSGADIVKIA SNGGGAQALE AVAMHGSTLC ERG
BAT3 repeat 5 YCRTDIAKIA GNGGGAQALK AIVMHGPTLC ERG
BAT3 repeat 6 YSRTDIVKIA DNNGGAQALK AVFEHGPALT QAG

Supplementary Figure 3: (a) Protein expression gel for MOrTL1 and MOrTL2 and (b) EMSA gel for MOrTL1 against BE_{MOrTL1}. hpi= hours post induction with IPTG.

(a)



(b)

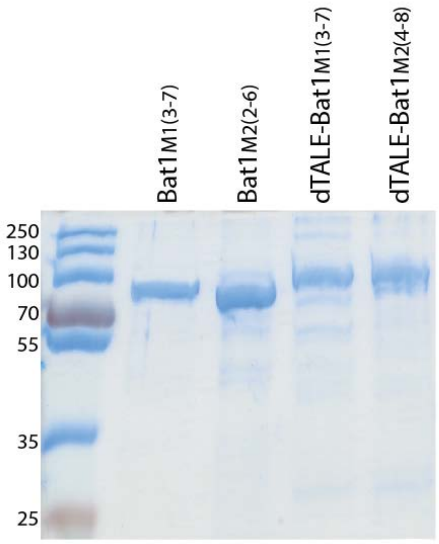


Supplementary Figure 4: Amino acid sequence of fusion protein EBN91408-MOrTL2. The two ORFs of genomic accession *EN814823.1* (see Figure S1) are separated by a frame-shift in the middle of MOrTL1 repeat 1. Removal of a single guanine base allows read through of a longer protein designated EBN91408-MOrTL2. 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.

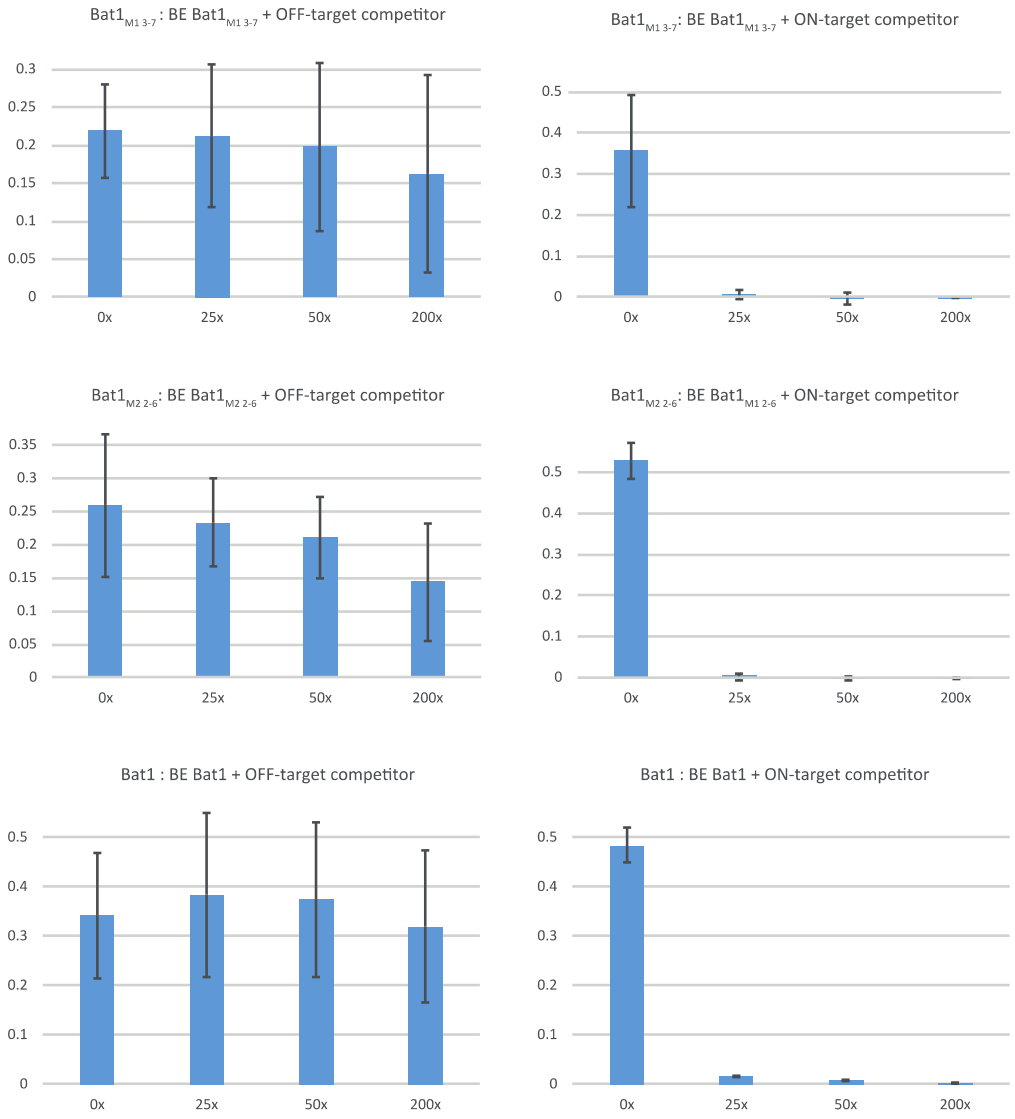
> EBN91408-MOrTL2.

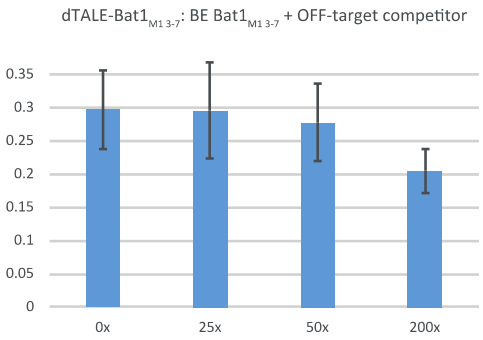
```
    MSNQTEQKILKFKLELRYPTESAQLIRAG
-1  FNRDQADRIILRGSSQRTVAKLLEIHKTLLAHPYR
0   ITYDDLTRIAARNGGSKNLVAVQANYAALTELG
1   FSAKDIVQMVSHGGGSKNLEVVQANYAALTGLG
2   FRTEDIVQMVSHDGGSKNLAAMIDKSTALKDLG
3   FRTEDIVQMVSHDGSSKNLAAMIDKSTALKGLG
4   FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
5   FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
6   FRTEDIVQMVSHDGGSKNLVAVQANYAALTGLG
7   FRTEDIVQMVSHDGGSKNLAAIIDKSTALTGLG
8   FRTEDIVQMVSNNGGSKNLAAIIDKSTALKGLG
9   FRTEDIVQMVSHGGGSKNLEVVQANYAALTGLG
10  FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
11  FRTEDIVQMVSHDGGSKNLAAMIDKYTALKDLG
12  FRTEDIVQMVSHDGGSKNLAAIIDKSTALKGLG
13  FLTEDIVQMVSHDGGSKNLEVVQASYDTLTELKF
```

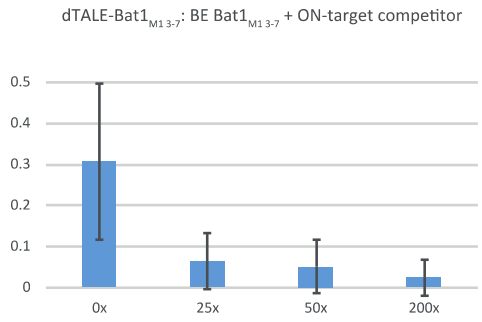
Supplementary Figure 5: Protein expression gel for $Bat1_{M1(3-7)}$, $Bat1_{M2(2-6)}$, dTALE-$Bat1_{M1(3-7)}$ and dTALE-$Bat1_{M2(4-8)}$.

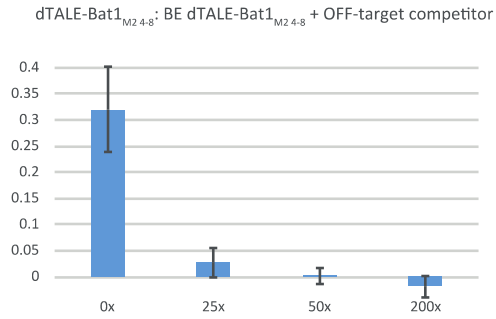


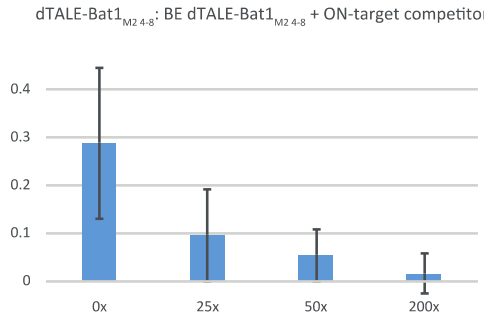
<u>Supplementary Figure 6</u>: Quantifications of protein : DNA relative to free-DNA in EMSAs shown in Figures 2d and 4d. EMSAs were carried out three times and standard deviations are shown. The proportion of shifted probe is expressed as a decimal along the y-axis of each plot. The fold excess of competitor DNA is shown below each bar.

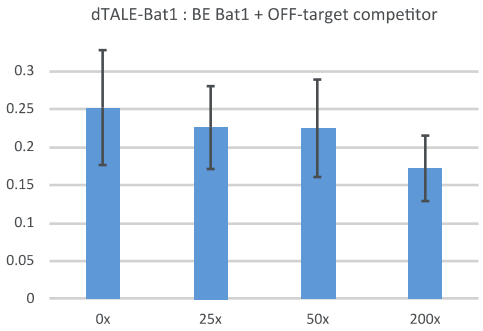


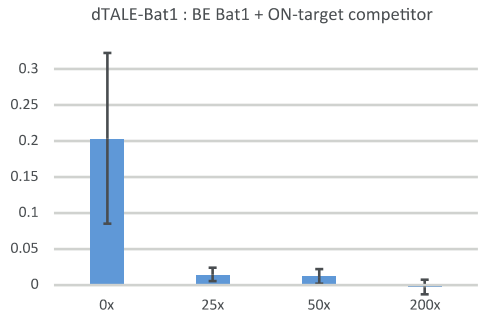

dTALE-Bat1M1 3-7: BE Bat1M1 3-7 + OFF-target competitor
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0x
25x
50x
200x


dTALE-Bat1M1 3-7: BE Bat1M1 3-7 + ON-target competitor
0.5
0.4
0.3
0.2
0.1
0
0x
25x
50x
200x

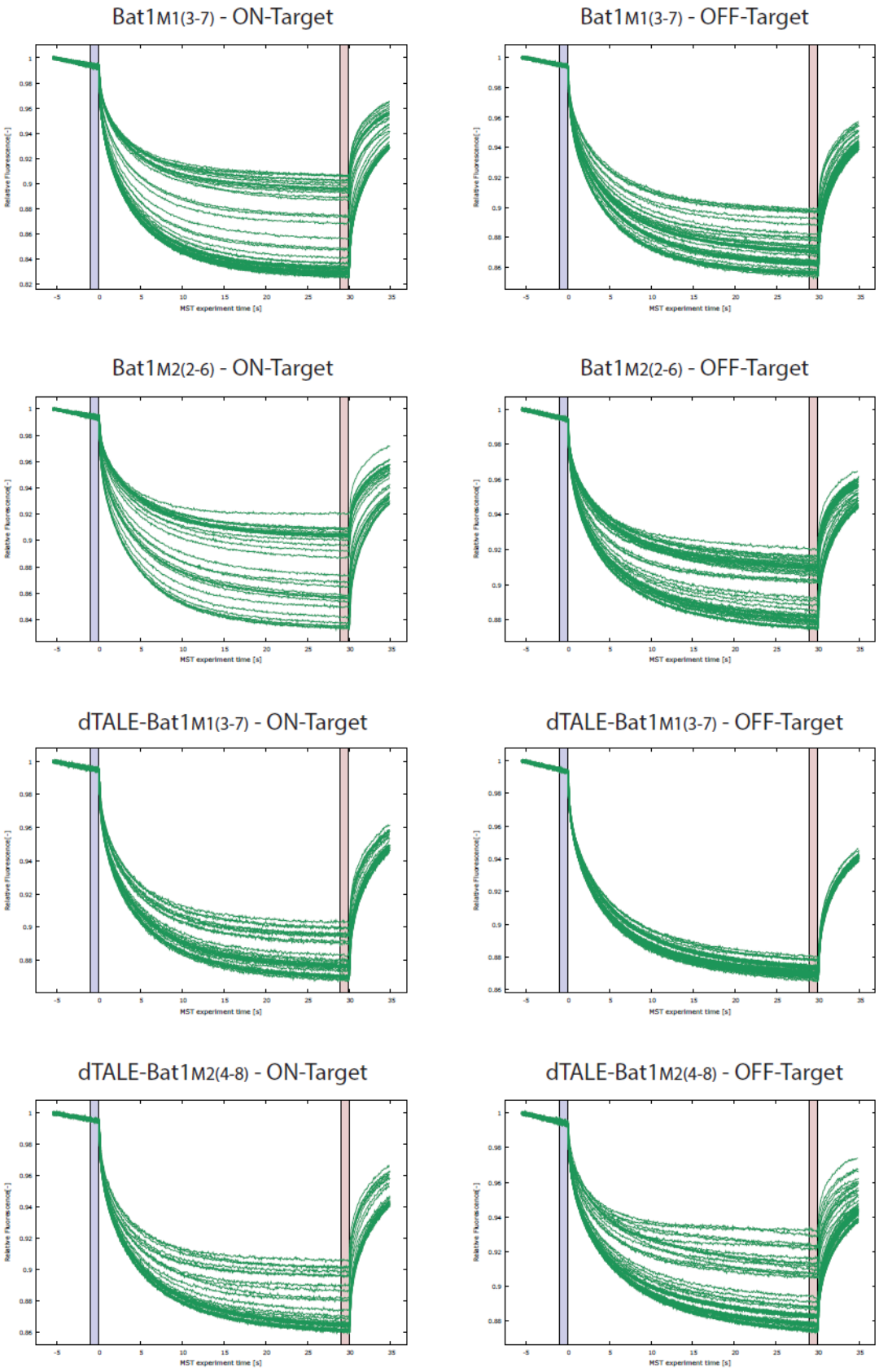

dTALE-Bat1M2 4-8: BE dTALE-Bat1M2 4-8 + OFF-target competitor
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0x
25x
50x
200x


dTALE-Bat1M2 4-8: BE dTALE-Bat1M2 4-8 + ON-target competitor
0.4
0.3
0.2
0.1
0
0x
25x
50x
200x


dTALE-Bat1 : BE Bat1 + OFF-target competitor
0.3
0.25
0.2
0.15
0.1
0.05
0
0x
25x
50x
200x

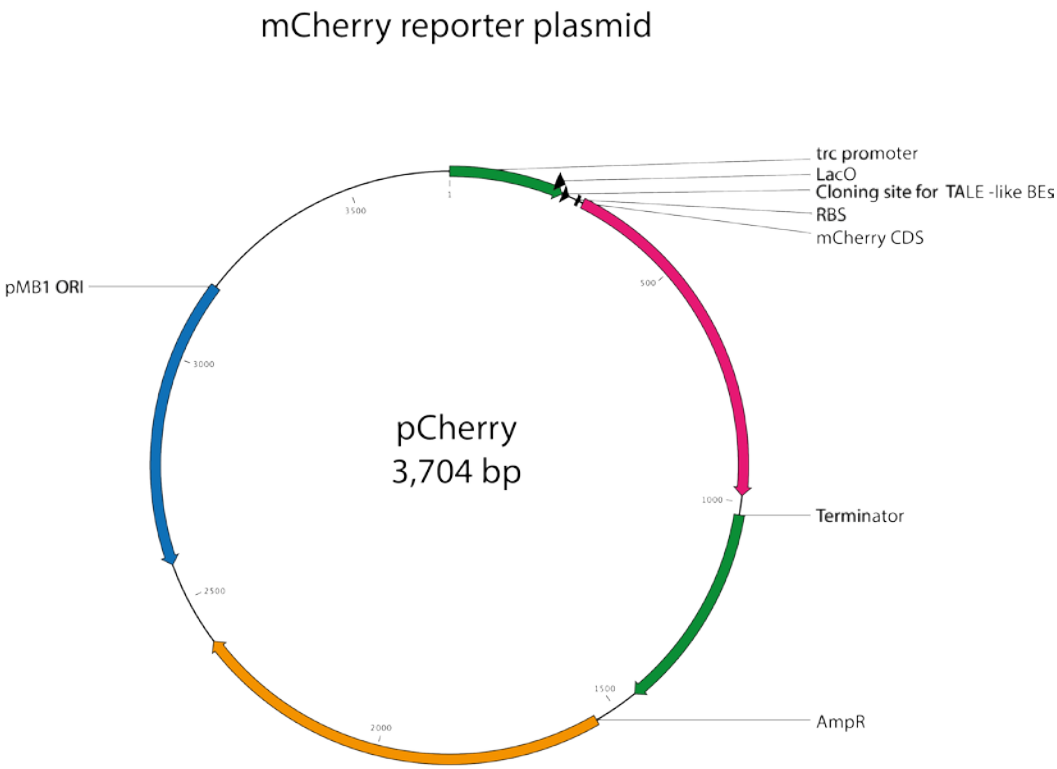

dTALE-Bat1 : BE Bat1 + ON-target competitor
0.3
0.25
0.2
0.15
0.1
0.05
0
0x
25x
50x
200x

Supplementary Figure 7: MST-Traces for figures 4 and 5.



Supplementary Figure 8:

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.



Supplementary Figure 9: 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 (3-7)
aagagAACGTaagacgttat

BE dTALE-Bat1 M1 (2-6)
aagagCAACGaagacgttat

BE Bat1 M2 (2-6)
aagagTCCGTaagacgttat

BE dTALE-Bat1 M2 (4-8)
aagagCGTTCaagacgttat

BE dTALE-Bat1 M2 (2-6)
aagagTCCGTaagacgttat

BE Bat1 GGTTG
aagagGGTTGaagacgttat

BE Bat1 TTGGT
aagagTTGGTaagacgttat

Supplementary Figure 10: 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  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
```

In PBT102 (extra C-terminal residues from cloning vector):
… GGTLIIPDLHSRKSKTSDRRLLT

>Bat1$_{M1\ (3\text{-}7)}$

```
NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
0  LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
1  FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
2  FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
3  FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
4  FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
5  FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
6  LEPKDIVSIASHIGANQTITTLLNKWGALIDLE
7  LEPKDIVSIASHIGATQAITTLLNKWAALRAKG
8  LDPKDIVSIASHDGSNQTITKLLEKWDELRAKE
9  LESKDIVSIASNNGATQTITRLLEKWDELRAKG
10 LDPKDIVSIASHGGATQAITTLLNRWGDLIDLG
```

```
11 FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14 FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15 FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16 FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17 FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18 FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
20 LSLIDIVEIASN-GGAQALKAVLKYGPVLMQAG
+1 RSNEEIVHVAARRGGAGRIRKMVAP---LLERQ
```

In PBT102 (extra C-terminal residues from cloning vector)
… GGTLIIPDLHSRKSKTSDRRLLT

>Bat1$_{M2\ (2\text{-}6)}$

```
NND MSTAFVDQDKQMANRLN
-1 LSPLERSKIEKQYGGATTLAFISNKQNELAQI
0  LSRADILKIASYDCAAHALQAVLDCGPMLGKRG
1  FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRG
2  FSRDDIAKMAGNIGGAQTLQAVLDLESAFRERG
3  FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRG
4  FSRADIVKIAGNTGGAQALHTVLDLEPALGKRG
5  FSRIDIVKIAANNGGAQALHAVLDLGPTLRECG
6  FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
7  FRTEDIVQMVSHDGGSKNLVAVQANYAALTGLG
8  FRTEDIVQMVSHDGGSKNLAAIIDKSTALTGLG
9  FRTEDIVQMVSNNGGSKNLAAIIDKSTALKGLG
10 FRTEDIVQMVSHGGGSKNLEVVQANYAALTGLG
11 FSQPDIVKMAGNIGGAQALQAVLSLGPALRERG
12 FSQPDIVKIAGNTGGAQALQAVLDLELTLVEHG
13 FSQPDIVRITGNRGGAQALQAVLALELTLRERG
14 FSQPDIVKIAGNSGGAQALQAVLDLELTFRERG
15 FSQADIVKIAGNDGGTQALHAVLDLERMLGERG
16 FSRADIVNVAGNNGGAQALKAVLEHEATLNERG
17 FSRADIVKIAGNGGGAQALKAVLEHEATLDERG
18 FSRADIVRIAGNGGGAQALKAVLEHGPTLNERG
19 FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRG
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$_{M1\ (3\text{-}7)}$

…

```
1  LTPQQVVAIASNIGGKQALETVQRLLPVLCQAHG
2  LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
3  LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
4  LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
5  LTPQQVVAFASNNGGKQALTKLLEKWDELRAKG
6  LEPKDIVSIASHIGANQTITTLLNKWGALIDLE
7  LEPKDIVSIASHIGATQAITTLLNKWAALRAKG
8  LDPKDIVSIASHDGSNQTITKLLEKWDELRAKE
9  LESKDIVSIASNNGATQTITRLLEKWDELRAKG
10 LDPKDIVSIASHGGATQAITTLLNRWGDLIDLE
11 LEPKDIVAIASNIGGKQALETVQRLLPVLCQAHG
12 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
15 LTPQQVVAIASHDGGKQALETVQRLLPVLCQAHG
16 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
17 LTPQQVVAIASNGGGKQALETVQALLPVLCQAHG
```

…

>dTALE-Bat1$_{M2\ (4\text{-}8)}$

…

```
1  LTPQQVVAIASNIGGKQALETVQRLLPVLCQAHG
2  LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
3  LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
4  LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
5  LTPQQVVAIASNNGGKQALVAVQANYAALTGLG
6  FRTEDIVQMVSHDGGSKNLAAIIDKSTALTGLG
7  FRTEDIVQMVSNNGGSKNLAAIIDKSTALKGLG
8  FRTEDIVQMVSHGGGSKNLEVVQANYAALTGLG
9  FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
10 FRTEDIVQMVSHDGGSKNLAAMIDKYTALKDLG
11 FRTEDIVAIASNIGGKQALETVQRLLPVLCQAHG
12 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
15 LTPQQVVAIASHDGGKQALETVQRLLPVLCQAHG
16 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
17 LTPQQVVAIASNGGGKQALETVQALLPVLCQAHG
```

…

>dTALE-Bat1$_{M1\ (2\text{-}6)}$

…

```
1  LTPQQVVAIASNIGGKQALETVQRLLPVLCQAHG
2  LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
3  LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
4  LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
5  LTPQQVVAFASNNGGKQALTRLLEKWGDLRAKE
6  LEPKDIVSIASHDGSNQTITKLLEKWDELRAKG
7  LEPKDIVSIASHIGANQTITTLLNKWGALIDLE
8  LEPKDIVSIASHIGATQAITTLLNKWAALRAKG
9  LDPKDIVSIASHDGSNQTITKLLEKWDELRAKE
10 LESKDIVSIASNNGATQTITRLLEKWDELRAKG
11 LDPKDIVAIASNIGGKQALETVQRLLPVLCQAHG
12 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
15 LTPQQVVAIASHDGGKQALETVQRLLPVLCQAHG
16 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
17 LTPQQVVAIASNGGGKQALETVQALLPVLCQAHG
```

…

>dTALE-Bat1$_{M2\ (2\text{-}6)}$

…

```
1  LTPQQVVAIASNIGGKQALETVQRLLPVLCQAHG
2  LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
3  LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
4  LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
5  LTPQQVVAIASNNGGKQALVAVQANYAALTGLG
6  FRTEGIVQMVSHGGGSKNLVAVQANYAALTGLG
7  FRTEDIVQMVSHDGGSKNLVAVQANYAALTGLG
8  FRTEDIVQMVSHDGGSKNLAAIIDKSTALTGLG
9  FRTEDIVQMVSNNGGSKNLAAIIDKSTALKGLG
10 FRTEDIVQMVSHGGGSKNLEVVQANYAALTGLG
11 FRTEGIVAIASNIGGKQALETVQRLLPVLCQAHG
12 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
13 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
14 LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
15 LTPQQVVAIASHDGGKQALETVQRLLPVLCQAHG
16 LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
17 LTPQQVVAIASNGGGKQALETVQALLPVLCQAHG
```

Supplementary Figure 11: 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    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
```

>TALE (AvrBs3)

```
 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
```

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.
GlySL = Glycine Short-loop referring to TALE repeats with a truncated BSR-loop (missing position 13 relative to other repeats). In such repeats the first residue of the glycine di-residue following the BSR position acts as the BSR in these repeats (Mak *et al*., *Science*, 2012)

Supplementary Table 2: A list of oligonucleotides used in this study

Primer name	**Sequence**	**Notes**
EMSA probes		
BE$_{MOrTL1}$ EMSA Fwd	TAGCACAACGTGCTGAC	EMSA Probe for non-chimeric MOrTL1 protein
BE$_{MOrTL1}$ EMSA Rev	GTCAGCACGTTGTGCTA	EMSA Probe for non-chimeric MOrTL1 protein
BE$_{MOrTL2}$ EMSA Fwd	TAGCTTCCGTTCCCTGAC	EMSA Probe for non-chimeric MOrTL2 protein
BE$_{MOrTL2}$ EMSA Rev	GTCAGGGAACGGAAGCTA	EMSA Probe for non-chimeric MOrTL2 protein
BEBat1 M1 6-10 EMSA Fwd	TAGACTAAGAGAACGTAAGACGTTATATGC	EMSA probe for Bat$_{M1\ 6\text{-}10}$ and dTALE-Bat1$_{M1\ 6\text{-}10}$
BEBat1 M1 6-10 EMSA Rev	GCATATAACGTCTTACGTTCTCTTAGTCTA	EMSA probe for Bat$_{M1\ 6\text{-}10}$ and dTALE-Bat1$_{M1\ 6\text{-}10}$
BEBat1 M2 6-10 EMSA Fwd	TAGACTAAGAGTCCGTAAGACGTTATATGC	EMSA probe for Bat$_{M2\ 6\text{-}10}$

BEBat1 M2 6-10 EMSA Rev	GCATATAACGTCTTACG GACTCTTAGTCTA	EMSA probe for $Bat_{M2\ 6\text{-}10}$
BEdTALE-Bat1 M2 6-10 EMSA Fwd	TAGACTAAGAGCGTTCA AGACGTTATATGC	EMSA probe for dTALE- $Bat_{M2\ 6\text{-}10}$
BEdTALE-Bat1 M2 6-10 EMSA Rev	GCATATAACGTCTTGAA CGCTCTTAGTCTA	EMSA probe for dTALE- $Bat_{M2\ 6\text{-}10}$
BEBat1 GGTTG EMSA Fwd	TAGACTAAGAGGGTTGA AGACGTTATATGC	OFF-target EMSA probe for all except dTALE-$Bat1_{M2\ 6\text{-}10}$
BEBat1 GGTTG EMSA REv	GCATATAACGTCTTCAA CCCTCTTAGTCTA	OFF-target EMSA probe for all except dTALE-$Bat1_{M2\ 6\text{-}10}$
BEBat1 TTGGT EMSA Fwd	TAGACTAAGAGTTGGTA AGACGTTATATGC	OFF-target EMSA probe dTALE-$Bat1_{M2\ 6\text{-}10}$
BEBat1 TTGGT EMSA Rev	GCATATAACGTCTTAAC CACTCTTAGTCTA	OFF-target EMSA probe dTALE-$Bat1_{M2\ 6\text{-}10}$
Primers for PCR mutagenesis		
MOrTL1 Bat1 Block2 Mimic Fwd	GGTCTCTTGGGCTGGAA CCGAAAGATATCGTG	To create MOrTL repeat blocks to insert into Bat1
MOrTL1 Bat1 Block2 Mimic Rev	GGTCTCAAACCCAGGTC GATCAGATCGCCCC	To create MOrTL repeat blocks to insert into Bat1
MOrTL2 Bat1 Block2 Mimic Fwd	GGTCTCTTGGGTTTCGT ACCGAAGGCATTGTCCA	To create MOrTL repeat blocks to insert into Bat1
MOrTL2 Bat1 Block2 Mimic Rev	GGTCTCAAACCCAGACC CGTCAGGGCGGCGTAG	To create MOrTL repeat blocks to insert into Bat1
MOrTL1 dTALE 5B mimic Fwd	GAAGACTCTCTGACGAA ACTGCTGGAAAAATG	To create MOrTL repeat blocks to insert into dTALE-Bat1
MOrTL1 dTALE 5B mimic Rev	GAAGACTCCGCTACAAT GTCTTTAGGTTCTAATT C	To create MOrTL repeat blocks to insert into dTALE-Bat1
MOrTL2 dTALE 5B mimic Fwd	GAAGACTCTCTGGTTGC AGTCCAAGCTAATTACG C	To create MOrTL repeat blocks to insert into dTALE-Bat1
MOrTL2dTALE 5B mimic Rev	GAAGACTCCGCTACGAT ATCTTCCGTGCGGAAAC	To create MOrTL repeat blocks to insert into dTALE-Bat1
MOrTL1 dTALE 5B mimic Fwd 2	ATGAAGACTCTCTGACC CGCCTGCTGGAAAAATG GGGCGATC	To create the 2nd set of MOrTL repeats to insert into dTALE-Bat1
MOrTL1 dTALE 5B mimic Rev 2	CAGAAGACTCCGCTACG ATATCTTTCGGATCCAG ACCTTTCG	To create the 2nd set of MOrTL repeats to insert into dTALE-Bat1
MOrTL2 dTALE 5B mimic Fwd 2	ATGAAGACTCTCTGGTT GCTGTTCAAGCTAACTA TGC	To create the 2nd set of MOrTL repeats to insert into dTALE-Bat1
MOrTL2 dTALE 5B mimic Rev 2	CAGAAGACTCCGCTACG ATACCTTCGGTGCGAAA GCC	To create the 2nd set of MOrTL repeats to insert into dTALE-Bat1
GFP-VS-Fwd	ATGGTGTCTAAGGGCGA	To create a GFP only pBT102

	AGAACTC	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	GAACGCTCTTGAAATTGTTATCCGCTC	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	TATCGATAAAAGCACCGCCC	To sequence central section of pDEST17 dTALE-$Bat1M_{2\ 6\text{-}10}$
Sco5B MidSeqR	ACCGTGACTGACCATTTGGA	To sequence central section of pDEST17 dTALE-$Bat1M_{2\ 6\text{-}10}$

Supplementary Table 3: 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 $_{M2\ 6\text{-}10}$**.** 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, RipTALI_14, YP_003750492
Bats	BAT1_BURRH, Bat2 (E5AW45), Bat3(E5AW43)
MOrTL1	EBN91408, EBN91409, ECG96325, ECG96326

Supplementary Table 6: Percentage conservation in each of the sequence logos seen in Figure 7. Grey shading indicates positions with 75% conservation or over in all groups, as displayed also in Figure 7.

Position	TALEs	RipTALs	Bats	MOrTL1	MOrTL2
1	100	100	86.5	100	100
2	99.4	77	65.4	75	83.3
3	88.8	81.1	50	87.5	91.7
4	40	59.5	48.1	100	91.7
5	100	100	82.7	100	83.3
6	100	95.9	98.1	100	100
7	100	98.6	82.7	100	100
8	100	93.2	75	100	100
9	100	87.8	71.2	100	100
10	100	100	94.2	100	100
11	92.9	90.5	73.1	100	100
12	67.6	59.5	92.3	75	91.7
13	29.4	43.2	25	25	58.3
14	100	100	100	100	100
15	98.8	100	98.1	75	91.7
16	99.4	93.2	96.2	62.5	100
17	99.4	87.8	98.1	87.5	100
18	100	95.9	98.1	62.5	100
19	100	100	100	100	100
20	98.8	95.9	50	100	50
21	99.4	91.9	76.9	50	75
22	98.2	87.8	96.2	100	50
23	98.8	40.5	80.8	100	50
24	90.6	75.7	42.3	62.5	50
25	100	70.3	53.8	87.5	50
26	100	91.9	59.6	100	58.3
27	100	55.4	76.9	50	50
28	99.4	33.8	53.8	37.5	91.7
29	100	100	88.5	100	100
30	96.5	97.3	46.2	62.5	58.3
31	87.1	70.3	0	0	0
32	64.1	81.1	0	0	0
33	17.6	98.6	61.5	62.5	75
34	100	97.3	80.8	62.5	100
35	82.4	40.5	92.3	50	91.7