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

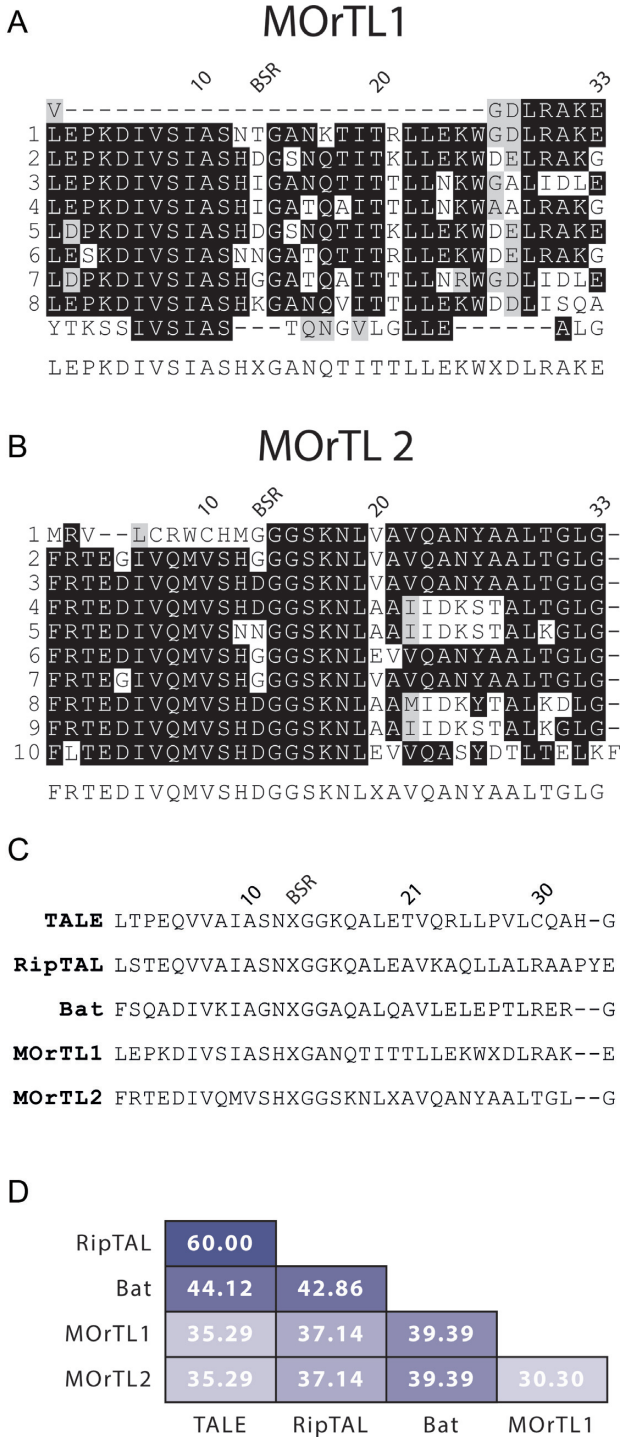

Figure 1. Amino acid sequences of MOrTL1 (ECG96326) and MOrTL2 (EBN19409), and a comparison of consensus TALE-like repeats. **(A, B)** Full amino acid sequences of each protein are displayed as a series of aligned tandem repeats prepared with ClustalW and Boxshade. Identical amino acids are white text 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

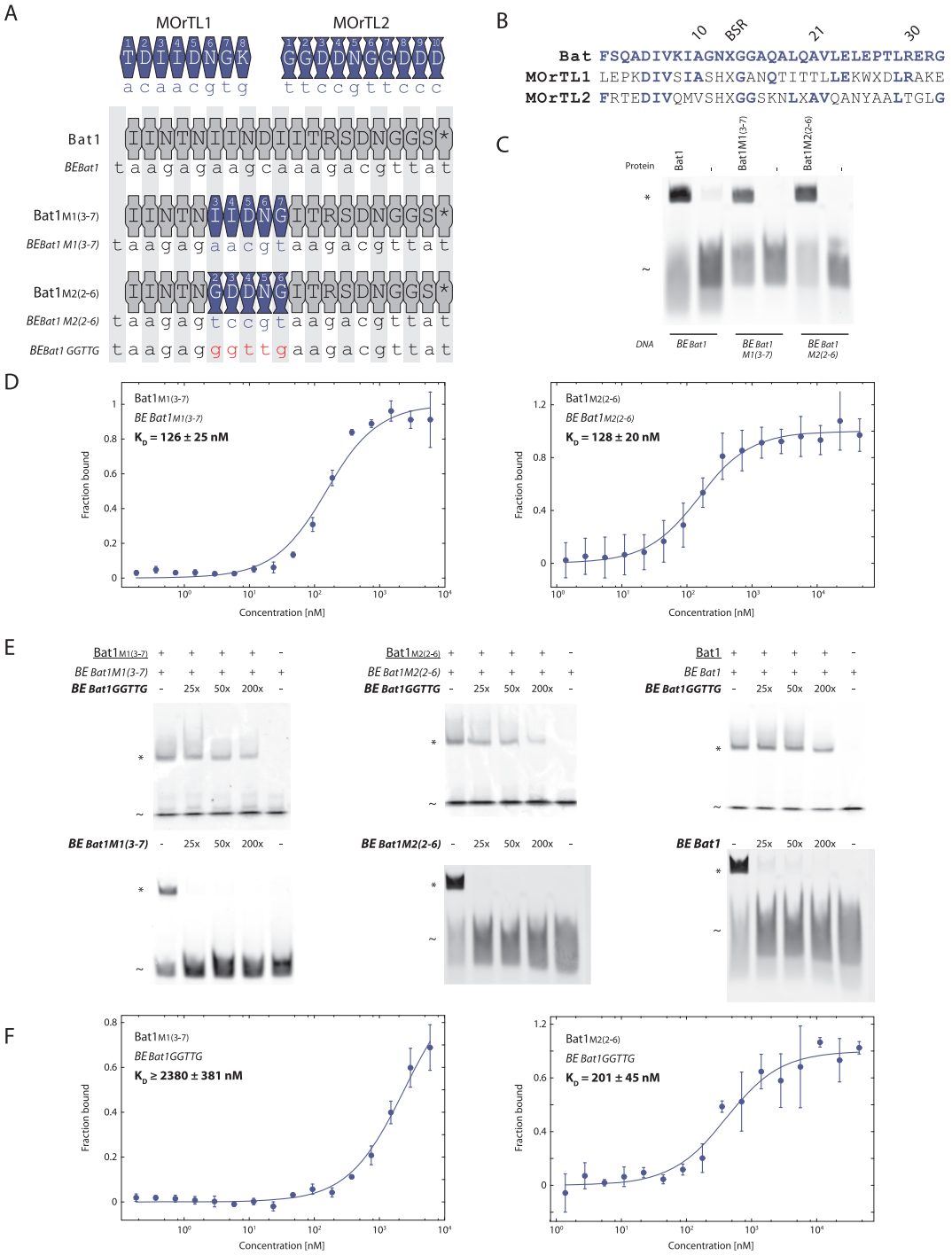

Figure 2. Bat1-MOrTL chimera proteins bind predicted target sequences *in vitro*. **(A)** Schematic display of repeat arrays of Bat1 (grey polygons), MOrTL1 (dark blue hexagons) and MOrTL2 (dark blue vases). Also displayed are the chimeras containing five repeats of MOrTL1 (repeats 3–7) or MOrTL2 (repeats 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-

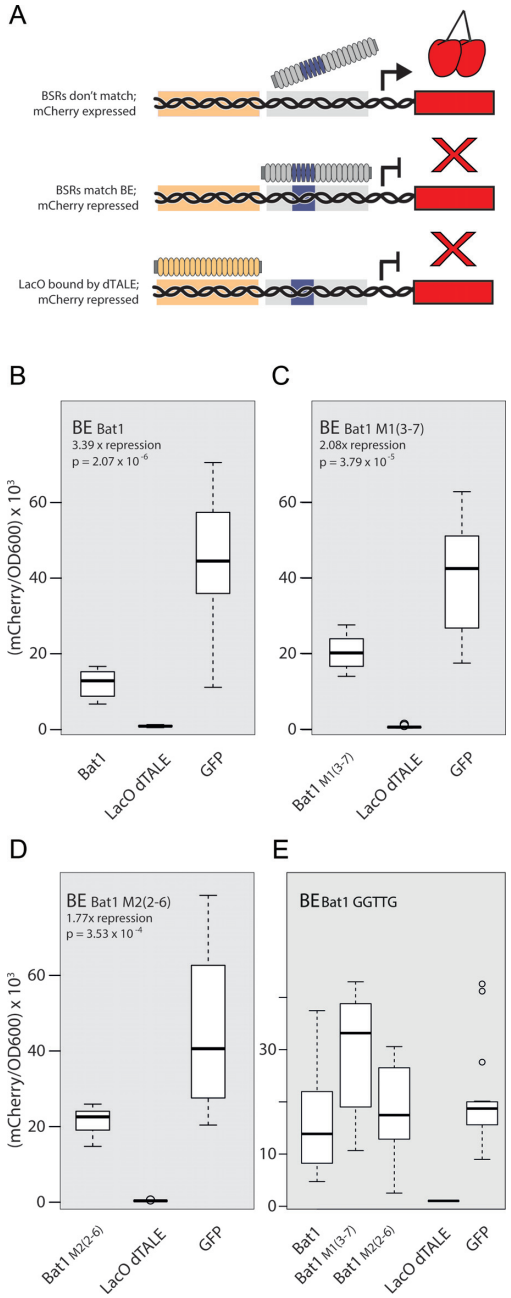

Figure 3. An *in vivo* reporter confirms that MOrTL repeats embedded in a Bat1-repeat array recognize predicted binding elements. **(A)** Schematic display of the repressor assay: mCherry reporter and expression plasmids encoding TALE-likes are co-transformed into *E. coli*. TALE-like chimeras consist 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

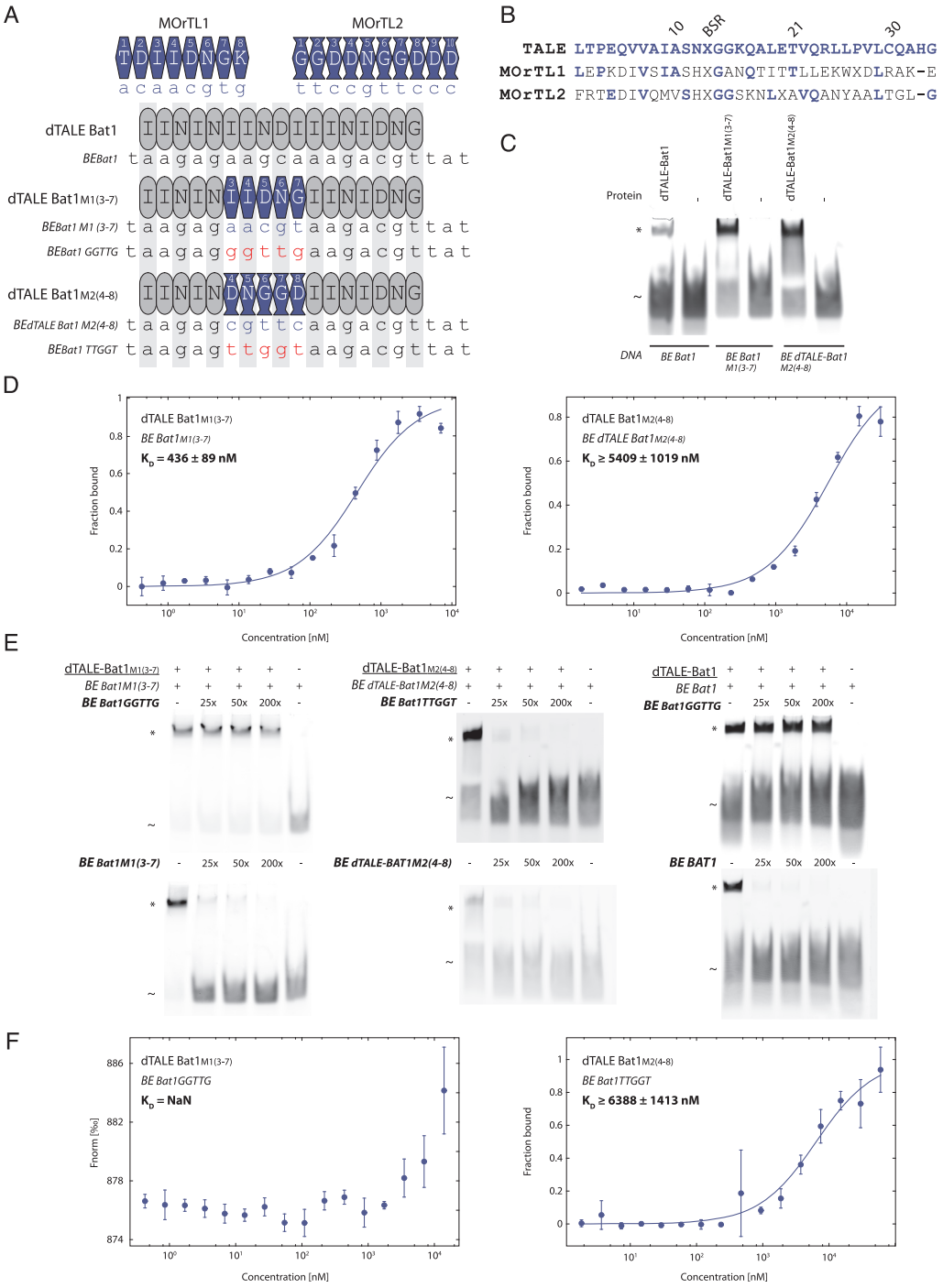

Figure 4. TALE-MOrTL chimeras proteins bind predicted target sequences *in vitro*. **(A)** Schematic display of repeat arrays of dTALE-Bat1 (grey ovals), MOrTL1 (dark blue hexagons) and MOrTL2 (dark blue vases). Also displayed are the chimeras containing five repeats of MOrTL1 (repeats 3–7) or MOrTL2 (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.

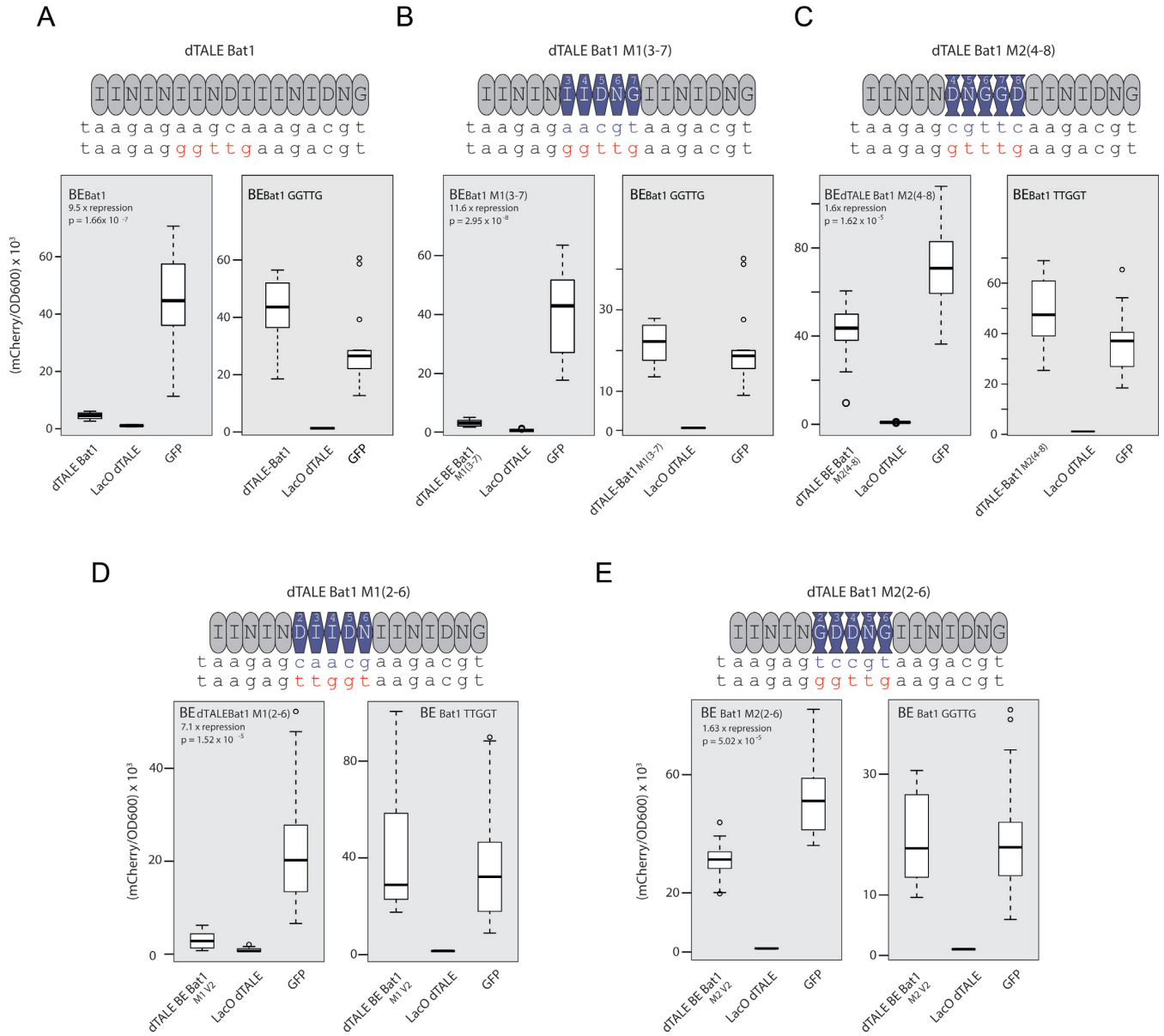

Figure 5. The repressor assay provides evidence for an incompatibility between MOrTL2 and TALE repeats. dTALE-Bat1 **(A)**, MOrTL1 **(B,D)** and MOrTL2 **(C,E)** chimeras were tested against cognate on- and off- target reporters in the repressor assay (Figure 3A). Box and whisker plots show mCherry fluorescence 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

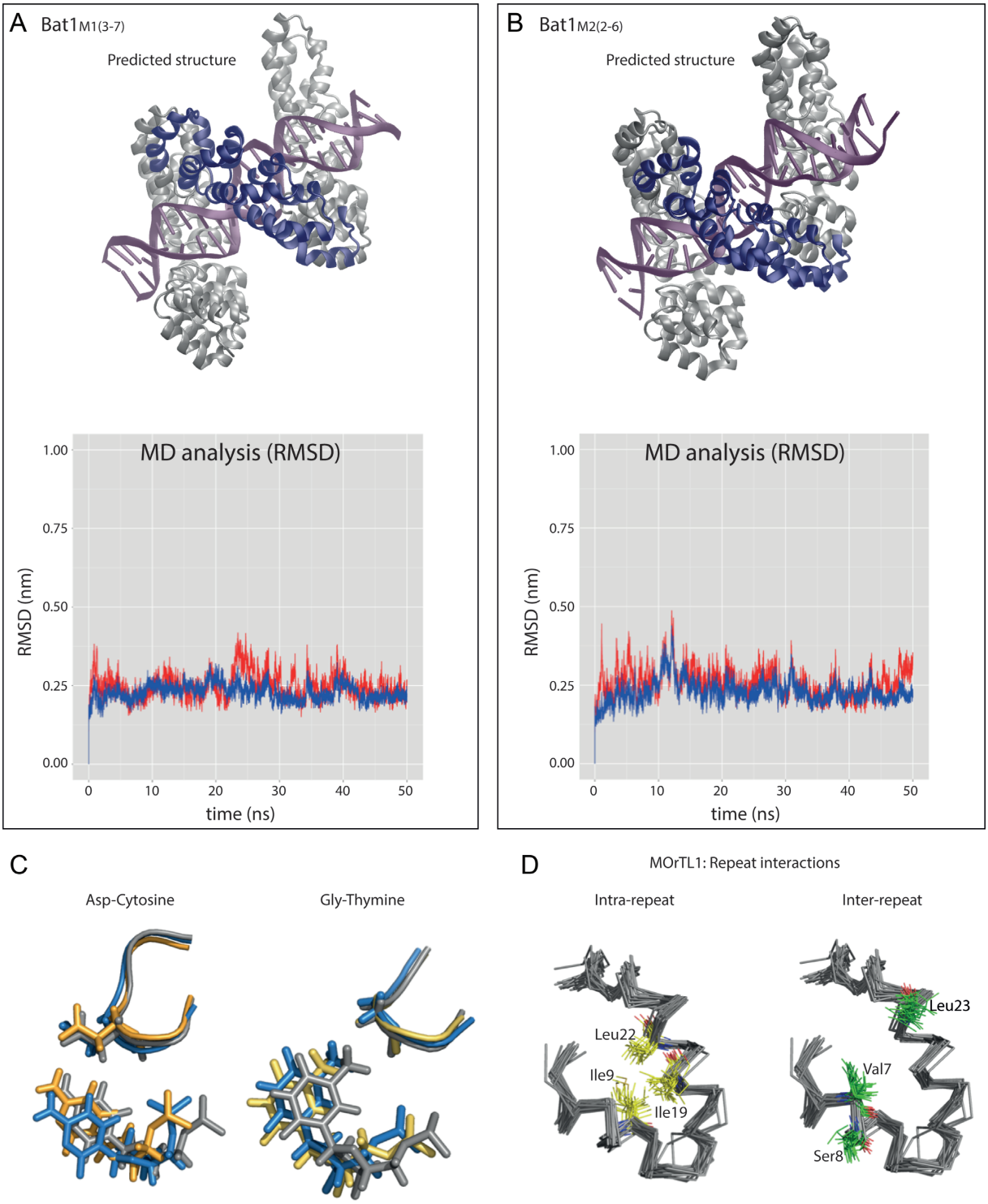

Figure 6. Homology models supported by molecular dynamics (MD) simulations of Bat1-MOrTL chimeras bound to cognate BEs, correspond to known TALE-like structures. Homology models of $Bat1_{M1\ (3–7)}$**(A)** and $Bat1_{M2\ (2–6)}$**(B)** were built using PDB entry 4cja as template structure with template 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).

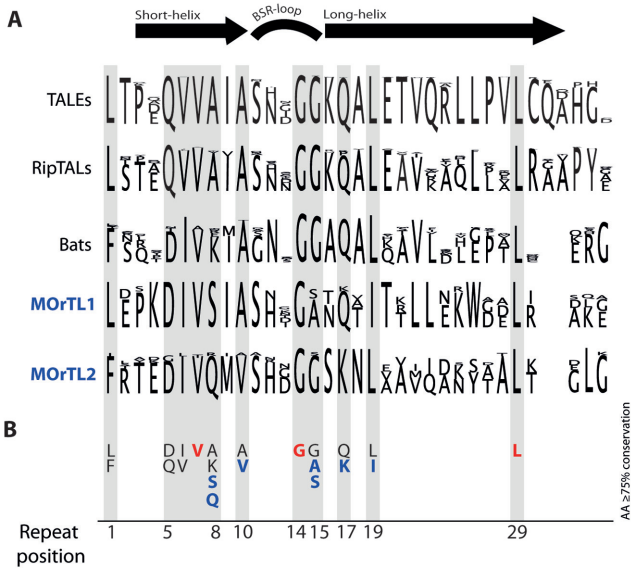

Figure 7. TALE-like repeat alignments show an underlying pattern of sequence conservation around the BSR position. **(A)** Repeat alignments and corresponding sequence logos were derived from representative core repeat arrays from each TALE-like group characterized so far (Supplementary Table S5), using 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.