Cell Research (2012) 22:1716-1720.

www.nature.com/cr

LETTER TO THE EDITOR

Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region

Cell Research (2012) **22**:1716-1720. doi:10.1038/cr.2012.156; published online 13 November 2012

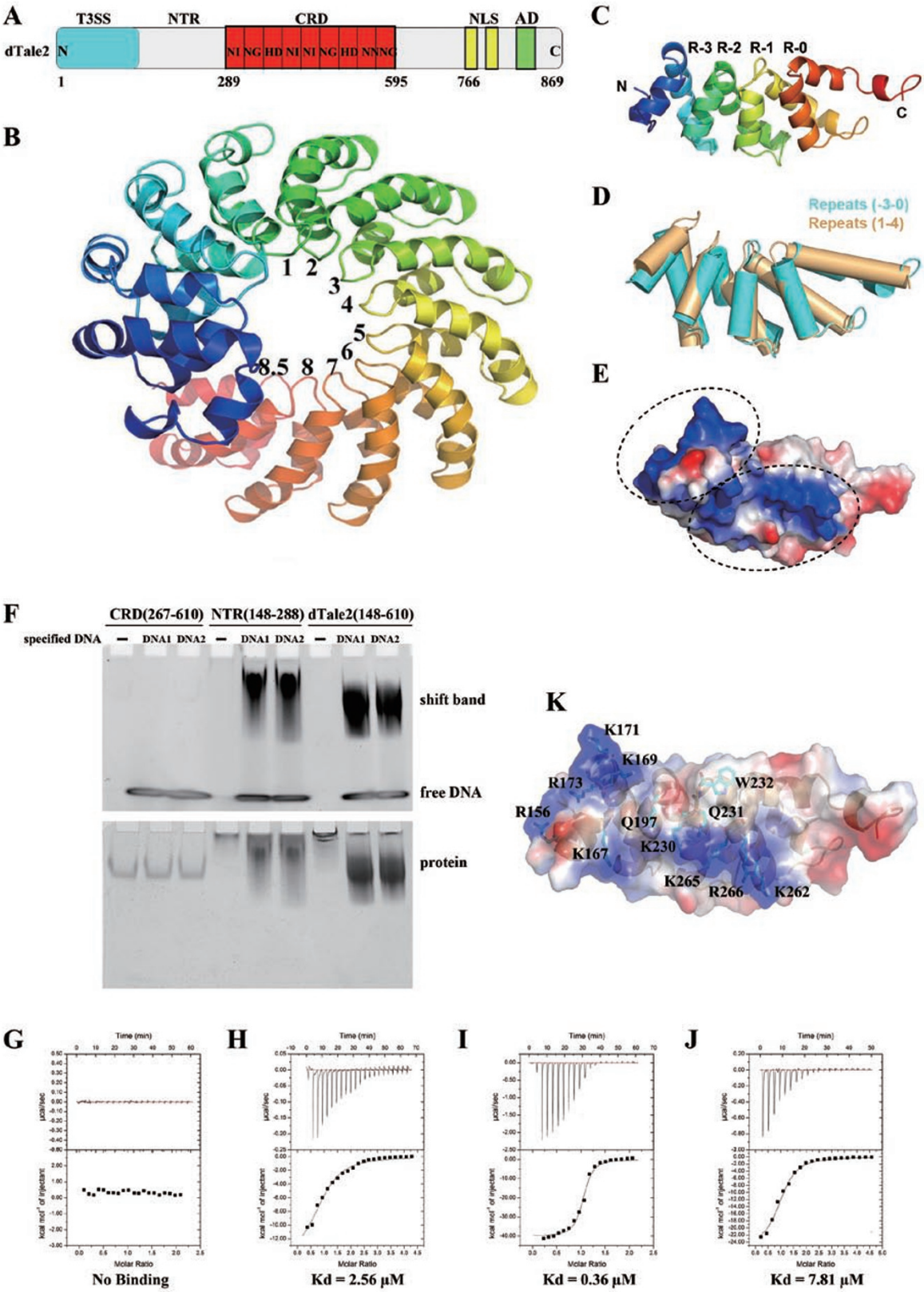
Dear Editor,

Xanthomonas TALEs (transcription activator-like effectors) are modular proteins characterized by an N-terminal T3S signal (T3SS), a central tandem repeat domain, C-terminal nuclear localization signals (NLSs) and an acidic transcriptional activation domain (AD) [1-3]. The central tandem repeats are nearly identical and are typically composed of 34 highly conserved amino acids, containing repeat-variable diresidues (RVDs) at the 12th and 13th positions that mediate their DNA binding specificity [1-5]. More than 20 types of RVDs have been identified thus far, among which HD, NG, NI and NN are the four most common ones with a specificity for the nucleotides C, T, A and G/A, respectively [4, 5]. The codes of RVD-DNA associations have been broadly and successfully employed in a variety of genome editing applications [6-14]. The recent crystal structures of the PthXo1-DNA and dHax3-DNA complexes show that the TALEs form a right-handed superhelix and wrap around the DNA major groove [15, 16]. The specific recognition of target DNAs by TALEs is achieved through the direct interaction of the 13th residue in the repeat with the base group of nucleotides [15, 16].

Most of the customized TALE fusion proteins have been constructed with the full-length TALEs, which might interpose large unnecessary regions and could impair activity [11]. Although the TALE-DNA specificity depends solely on the central repeat domain (CRD) [1-5, 8], the CRD alone is insufficient to bind to the target sites and ensure the activity of the TALE fusion proteins *in vivo* [6, 11, 12, 14]. Consistently, the N-terminal region (NTR) immediately preceding the CRD has been shown to be crucial for the full activity of TALE fusion proteins [11-14], although the underlying mechanism remains less well understood. The structural study of the PthXo1-DNA complex has illustrated the presence and potential roles in DNA binding of two additional degenerate repeats from the NTR [15]; however, the exact boundaries of the NTR involved in DNA binding remain undefined. The TALE minimal DNA binding domain, which is essential for the rational design of customized TALE fusion proteins, needs to be precisely refined and optimized [6, 8, 10, 15].

We first resorted to limited proteolysis to obtain more domain information of TALE proteins. The limited trypsin proteolysis of two TALE proteins, the full-length AvrBs4 and a designed TALE protein 1 (dTale1 (148-977) (with 17.5 and 15.5 Repeats, respectively)), showed similar proteolytic patterns (Supplementary information, Figure S1). Such a pattern was also observed for another TALE protein (dTale2 (148-766), 8.5 Repeats) with the NLSs and AD deleted (Figure 1A), either alone or complexed with its target DNA (Supplementary information, Figure S1). The treatment with trypsin did not disrupt the dTale2-DNA complex, as indicated by a gel shift assay (Supplementary information, Figure S2). Analysis via SDS-PAGE and mass spectrometry identified the NTR (residues 148-288) and the CRD in the protease-resistant fragments (data not shown). Together, these data suggest that the NTR and CRD fold into a continuously intact domain responsible for DNA binding.

We then solved the crystal structure of the intact domain of dTale2 (148-610) containing both the NTR and CRD with molecular replacement (Supplementary information, Table S1). The structurally well-defined region of dTale2 encompasses residues 154-585 and forms a continuously right-handed super-helix, with the nine RVD loops arraying at the inner face of the spiral axis (Figure 1B). The tight packing of the NTR and CRD results in the formation of a highly curved solenoid structure. Each of the 8 TAL repeats is formed by two α-helices with an intervening RVD and reveals an almost identical conformation (Supplementary information, Figure S3). Structural comparison of the dTale2 CRD (303-580) with the DNA-free form of dHax3 shows a highly conserved CRD structure with an RMSD (root mean square deviation) of 1.351 Å for the Cα-atoms over 272 residues. In contrast, the structure of dTale2 (231-580) is significantly different from that of the DNA-bound dHax3, with an RMSD of 2.843 Å over 310 aligned Cα-atoms, further supporting the conformational plasticity of


A
T3SS
NTR
CRD
NLS
AD
dTale2
N
NI NG HD NI NI NG HD NNNG
C
1
289
595
766
869
B
1 2 3 4 5 6 8.5 8 7
C
R-3 R-2 R-1 R-0
N
C
D
Repeats (-3-0)
Repeats (1-4)
E
F
CRD(267-610)
NTR(148-288)
dTale2(148-610)
specified DNA
—
DNA1 DNA2
—
DNA1 DNA2
—
DNA1 DNA2
shift band
free DNA
protein
K
K171
K169
R173
W232
R156
Q197
Q231
K230
K167
K265
R266
K262
G
Time (min)
No Binding
H
Time (min)
Kd = 2.56 μM
I
Time (min)
Kd = 0.36 μM
J
Time (min)
Kd = 7.81 μM

the TALE proteins (Supplementary information, Figure S4A and S4B).

Highly conserved among the TALE proteins, the dTale2 NTR (Supplementary information, Figure S6A) also displays a right-handed superhelical structure, exclusively with α-helices and intervening loops (Figure 1B and 1C), capping the CRD primarily *via* van der Waals contacts (Supplementary information, Figure S5). Unexpectedly, the NTR (residues 162-288) features four continuous repeats, which were not well resolved in previous studies [15, 16]. Each of the four adjacent repeats contains two α-helices and an intervening loop (Figure 1C), a structural feature strikingly similar to that of the TALE CRD. According to the previous nomenclature of PthXo1 [15], we designated the four adjacent repeats as Repeat-3 (residues 162-186), Repeat-2 (residues 187-220), Repeat-1 (residues 221-254) and Repeat-0 (residues 255-288) (Figure 1C). Repeat-0 packs tightly against the first canonical repeat of the CRD, continuing the superhelical structure of the CRD (Supplementary information, Figure S5). Among the four noncanonical repeats, Repeat-1 (221-254) possesses a similar RVD loop to the canonical repeats. Although the intervening loops of R-3, R-2 and R-0 are fairly degenerate to that of the CRD, the four adjacent atypical repeats can be well superimposed with four consecutive TAL repeats in the CRD (Figure 1D), with an RMSD of 1.182 Å over the 58 aligned Cα-atoms, despite their non-conserved amino acids (Supplementary information, Figure S6B). The surface electrostatic potential analysis of dTale2 (154-585) shows both the positively charged patches along the inner ridge of the CRD (Supplementary information, Figure S7), and also two unusual positive patches in the NTR (Figure 1E).

The similar fold to the canonical TAL repeats and the presence of positively charged patches suggest the nucleic acid binding potential of the NTR. Indeed, both the NTR (148-288) and the intact dTale2 (148-610) protein, but not the CRD, exhibited obvious shifts in mobility after incubation with dTale2-specified dsDNA (Supplementary information, Table S2) in gel shift assay (Figure 1F). To verify these results further, we quantified the binding affinities of these TALE proteins with different dsDNA using Isothermal Titration Calorimetry (ITC). Notably, the dTale2 CRD itself did not bind to the dTale2-specified dsDNA (Figure 1G), whereas the NTR did bind to the specified dsDNA, either blunt (DNA1, 16 bp) or sticky dsDNA (DNA2, 17 bp) with similar affinities (dissociation constants Kd of 2.56 μM and 2.51 μM, respectively) (Figure 1H and Supplementary information, Figure S8A). A similar affinity was also observed between the NTR and a random dsDNA (DNA3, 14 bp) (Supplementary information, Figure S8B), suggesting that the NTR binds to DNA in a sequence-independent/nonspecific manner. The NTR immediately preceding the CRD was previously thought to interact specifically with the thymine (T) nucleotide preceding the RVD-specified DNA sequence [2, 4, 8, 15]. However, the NTR bound to a T-rich dsDNA and a T-deficient dsDNA with similar affinities (2.60 μM and 2.05 μM, respectively) (Supplementary information, Figure S9A and S9B), further supporting that the NTR nonspecifically binds to dsDNA.

In contrast to the CRD, dTale2 (148-610), which contains both the NTR and the CRD, did exhibit higher affinities for specified DNA1 and DNA2 (0.36 μM and 0.47 μM, respectively) than that of the NTR (Figure 1I and Supplementary information, Figure S10A). A truncated dTale2 protein (230-610), which is similar to the DNA-bound dHax3 with a premature NTR [16], bound to specified DNA2 with a lower affinity (7.81 μM) (Figure 1J), indicating that the intact NTR is crucial for the DNA binding ability of dTale2 (148-610), consistent with previous reports that suggest that a fragment encompassing more than 100 residues of the NTR is essential for the full activities of customized TALE fusion proteins [11-14]. The dTale2 variant protein (148-766) with an extended C-terminal region, bound to specified DNA2 with a comparable affinity (0.70 μM) to dTale2 (148-610)

Figure 1 Crystal structure of the dTale2 (148-610). **(A)** Domain organization and repeat compositions of dTale2 protein. N, N terminus; T3SS, type III secretion signal (colored in light blue); NLS, colored in yellow; AD, colored in light green; C, C terminus. The CRD is colored in red. Numerals indicate the residue numbers at the boundaries of different subdivisions. **(B)** Overall structure of dTale2 (148-610) (shown as cartoon representation, colored in rainbow). The numbers indicate the repeats along the super-helix. **(C)** Overall structure of the dTale2 NTR (148-288) (shown as cartoon, colored in rainbow). **(D)** Structural comparison of the NTR (162-288, cyan) to the first four canonical repeats (residues 289-424) in the CRD (light yellow). **(E)** Electrostatic potential surface of the NTR (positive potential, blue; negative potential, red). **(F)** Gel shift assay analysis of the dTale2 protein. Both the NTR (148-288) and the intact dTale2 (148-610) protein, but not the CRD (267-610), exhibited obvious shifts in mobility after incubation with dTale2-specified dsDNA. **(G-J)** ITC analysis of dTale2 proteins binding to dsDNA. The curves in the lower panels are the best fit to a one-set-of-sites binding model. The derived dissociation constants Kd are indicated. **(G)** The CRD could not well bind to specified DNA1. **(H)** The NTR (148-288) binds to specified DNA1. **(I)** The dTale2 (148-610) binds to specific DNA1. **(J)** The dTale2 (230-610) binds to specific DNA2. **(K)** The putative amino acids (sticks) in the NTR involved in DNA binding. The NTR is shown as surface representation (transparency 40%).

(Supplementary information, Figure S10A and S10B), suggesting that dTale2 (148-610) was sufficient to bind to the target DNA. Combining the structure of dTale2 (148-610), gel shift assays and ITC analysis, we present the precisely redefined dTale2 minimal and efficient DNA binding domain, which not only includes the CRD, but also encompasses the four atypical repeats in the NTR (Supplementary information, Figure S11). Because of the highly conserved sequences among TALEs, this optimized scaffold is probably applicable for most TALE fusion proteins.

Because the PthXo1-DNA complex indicated that some amino acids (particularly Trp 232, Lys 262, Lys 265 and Arg 266) in the NTR make direct contacts with the dsDNA [15], we speculated that several other amino acids with side chains extending out to the surface of the protein might be also involved in DNA binding (Figure 1E and 1K). As shown by the ITC assay, the single mutation (K169A, W232A or R236A) had little effect on the DNA binding ability of dTale2 (148-610), whereas the multiple-mutations of the positively charged amino acids to alanines (K262A/K265A/R266A), (K171A/K262A/K265A/R266A), (R173A/K262A/K265A/R266A) and (K230A/Q231A/K262A/K265A/R266A) greatly impaired the DNA binding activity of dTale2 (148-610) (Supplementary information, Table S3), implying that these basic amino acids are vital for DNA binding and play synergetic roles in interacting with DNA. Coupled with the previous data from TALEs' applications *in vivo* [11-14], it is suggested that the NTR serves as the indispensable "nucleation site" for the TALE-DNA binding both *in vitro* and *in vivo*.

Remarkably, dTale2 (148-610) also bound to nonspecific dsDNA4 and T-deficient dsDNA with lower binding affinities (1.70 μM and 2.26 μM, respectively) than its binding to specified dsDNA (Supplementary information, Figure S12A and S12B). This nonspecific binding is probably due to the presence of the positively charged patches on the surface of the dTale2 (148-610) NTR and CRD (Supplementary information, Figure S7). As the nucleation site for DNA binding, the NTR might work cooperatively with the CRD both in specific and in nonspecific DNA binding. Although displaying no detectable dsDNA binding activity by the ITC assay, the CRD in the intact TALE markedly enhanced the affinity of the NTR for the specified dsDNA. The lower binding affinity to nonspecific dsDNA might make it feasible for TALEs to rapidly slide along the large host genome to search for target sites, whereas the higher affinity to the target DNA probably ensures the TALEs-DNA specificity and full activity of the TALEs. Although the binding to nonspecific DNA does not always lead to the full activity of TALEs, it might cause the off-target effect of TALE fusion proteins *in vivo*.

Collectively, our results regarding the four atypical repeats in the extended NTR reveal high structural similarities with the canonical TAL repeats. They not only exhibit direct and nonspecific interaction with dsDNA, but also serve as the essential "nucleation site" for TALEs in DNA binding. The refined architecture of the dTale2 minimal DNA binding domain presented here provides new insights into the TALE-DNA interaction and might help in the rational and efficient design of customized TALE fusion proteins for biotechnological applications.

Acknowledgments

We are grateful to the staff at Beamline BL17U at Shanghai Synchrotron Radiation Facility (SSRF, China) for help with data collection. This work was supported by the State Key Program of National Natural Science of China (31130063), Chinese Ministry of Science and Technology (2010CB835300) and the National Outstanding Young Scholar Science Foundation of National Natural Science Foundation of China (20101331722). The atomic coordinates have been deposited in the Protein Data Bank (accession code 4HPZ).

Haishan Gao[1, 2, 3], Xiaojing Wu[2], Jijie Chai[2], Zhifu Han[2]

[1]School of Life Sciences, Peking University, Beijing 100871, China; [2]School of Life Sciences, Tsinghua University, Beijing 100084, China; [3]National Institute of Biological Sciences, Beijing 102206, China

Correspondence: Zhifu Han
Tel: +86-10-62789619
E-mail: hanzhifu0807@163.com

References

1 Boch J, Bonas U. *Xanthomonas* AvrBs3 family-type III effectors: discovery and function. *Annu Rev Phytopathol* 2010; **48**:419-436.
2 Bogdanove AJ, Schornack S, Lahaye T. TAL effectors: finding plant genes for disease and defense. *Curr Opin Plant Biol* 2010; **13**:394-401.
3 Scholze H, Boch J. TAL effectors are remote controls for gene activation. *Curr Opin Microbiol* 2011; **14**:47-53.
4 Boch J, Scholze H, Schornack S, *et al*. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* 2009; **326**:1509-1512.
5 Moscou MJ, Bogdanove AJ. A simple cipher governs DNA recognition by TAL effectors. *Science* 2009; **326**:1501.
6 Christian M, Cermak T, Doyle EL, *et al*. Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* 2010; **186**:757-761.
7 Boch J. TALEs of genome targeting. *Nature Biotechnol* 2011; **29**:135-136.

8 Bogdanove AJ, Voytas DF. TAL effectors: customizable proteins for DNA targeting. *Science* 2011; **333**:1843-1846.
9 Mahfouz MM, Li L. TALE nucleases and next generation GM crops. *GM Crops* 2011; **2**:99-103.
10 Mahfouz MM, Li L, Shamimuzzaman M, Wibowo A, Fang X, Zhu JK. *De novo*-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. *Proc Natl Acad Sci USA* 2011; **108**:2623-2628.
11 Miller JC, Tan S, Qiao G, *et al*. A TALE nuclease architecture for efficient genome editing. *Nature Biotechnol* 2011; **29**:143-148.
12 Mussolino C, Morbitzer R, Lutge F, Dannemann N, Lahaye T, Cathomen T. A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res* 2011; **39**:9283-9293.
13 Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature Biotechnol* 2011; **29**:149-153.
14 Sun N, Liang J, Abil Z, Zhao H. Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. *Mol BioSyst* 2012; **8**:1255-1263.
15 Mak AN, Bradley P, Cernadas RA, Bogdanove AJ, Stoddard BL. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* 2012; **335**:716-719.
16 Deng D, Yan C, Pan X, *et al*. Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* 2012; **335**:720-723.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)