Cell Research (2012) 22:1502-1504.

www.nature.com/cr

LETTER TO THE EDITOR

Recognition of methylated DNA by TAL effectors

Cell Research (2012) **22**:1502-1504. doi:10.1038/cr.2012.127; published online 4 September 2012

Dear Editor,

TALE (transcription activator-like effector) DNA-binding repeats, which represent a modular assembly for specific target DNA of almost any sequences, provide a powerful tool for genetic editing. All the codes are for the four fundamental bases. There was no report on the recognition of modified DNA by TALE up to date. In this study, we report two crystal structures of engineered TALE repeats in complex with methylated DNA elements at 1.85 Å and 1.95 Å resolutions, respectively. Biochemical analysis shows that the TALE code NG, but not HD, binds to 5-methylcytosine (mC). Our findings will extend the application of TALE in epigenetic modification and cancer research, but also reveal the previously unconsidered limits for the applications of TALEs.

DNA methylation is a major epigenetic mark, and plays a pivotal role in diverse biological processes in a wide range of organisms. In mammals, DNA methylation usually occurs to the C5 position of cytosine in the CG context. Hypermethylation of the CpG islands may lead to gene silencing [1, 2].

The TALEs are a family of DNA-binding proteins [3-5]. A TALE contains a number of sequence repeats, each recognizing one DNA base. Each TALE repeat consists of 33-35 highly conserved amino acids except for those at positions 12 and 13, which are named RVD (repeat variable diresidue) and determine DNA-binding specificity. The recognition codes between the RVDs and the DNA bases have been established through experimental and computational approaches [6, 7]. For example, the bases A, G, C and T can be recognized by the RVDs NI (Asn and Ile), NN (Asn and Asn), HD (His and Asp) and NG (Asn and Gly), respectively [4]. The modular nature of TALE repeats provides an important tool for genetic manipulation [8-10].

We recently determined the high-resolution structure of DNA-bound TALE dHax3, which provided the molecular basis for base-specific DNA recognition [11]. The 34-residue TALE repeat comprises two α-helices connected by a short loop where RVD resides. The RVD loop tracks along the major groove of DNA (Figure 1A). Only the second residue of RVD, namely the one at position 13, is in direct contact with the base in the sense DNA strand, whereas the first residue helps maintain the RVD loop conformation through hydrogen bond.

Notably, the DNA base T is recognized by Gly^{13} in most cases. The lack of side chain in Gly not only provides sufficient space to accommodate the 5-methyl group of thymine but also allows optimal van der Waals interactions between the Cα atom of Gly^{13} and the 5-methyl group [11] (Figure 1B). This observation immediately suggests the possibility that mC might be recognized by Gly^{13} in RVD, because the only difference between the bases T and mC is at position 4, which is not involved in binding to TALE repeats. To examine this possibility, we replaced three T bases in the sense DNA strand by three mC bases and performed DNA-binding studies using the electrophoretic mobility shift assay (EMSA).

Confirming our prediction, the dHax3 protein binds to the triply modified DNA, with the forward strand 5′-TCCCT(mC)TA(mC)CTC(mC)-3′ (Figure 1C). This binding is very similar to that for the unmodified dsDNA, with the forward strand 5′-TCCCTTTATCTCT-3′ (Figure 1C). This result is rather striking, considering the fact that three T-A base pairs have been replaced by three mC-G base pairs in the dsDNA.

Next, we crystallized the binary complex between dHax3 and the triply modified DNA-binding sequence, and determined its structure at 1.85 Å (Figure 1D and Supplementary information, Figures S1A, S2A and Table S1). As anticipated, the 5-methyl group of mC points to the Cα of Gly^{13} with a distance of 3.4-4.0 Å for the three mC bases (Figure 1D and Supplementary information, Figure S2B). As DNA methylation mostly occurs to cytosine in the CG context, we constructed a dHax3 variant that is expected to recognize the DNA elements 5′-TCCCTT(mC)G(mC)GTCT-3′, where the RVDs NG and NN are designed for bases mC and G, respectively (Supplementary information, Figure S1A). The crystal structure of this dHax3 variant, which we name dHax3-mCG, in complex with its target dsDNA was also obtained and refined at 1.95 Å resolution (Figure 1E and Supplementary information, Figure S2C and Table S1). The coordination of mC bases by Gly^{13} residues is identi-

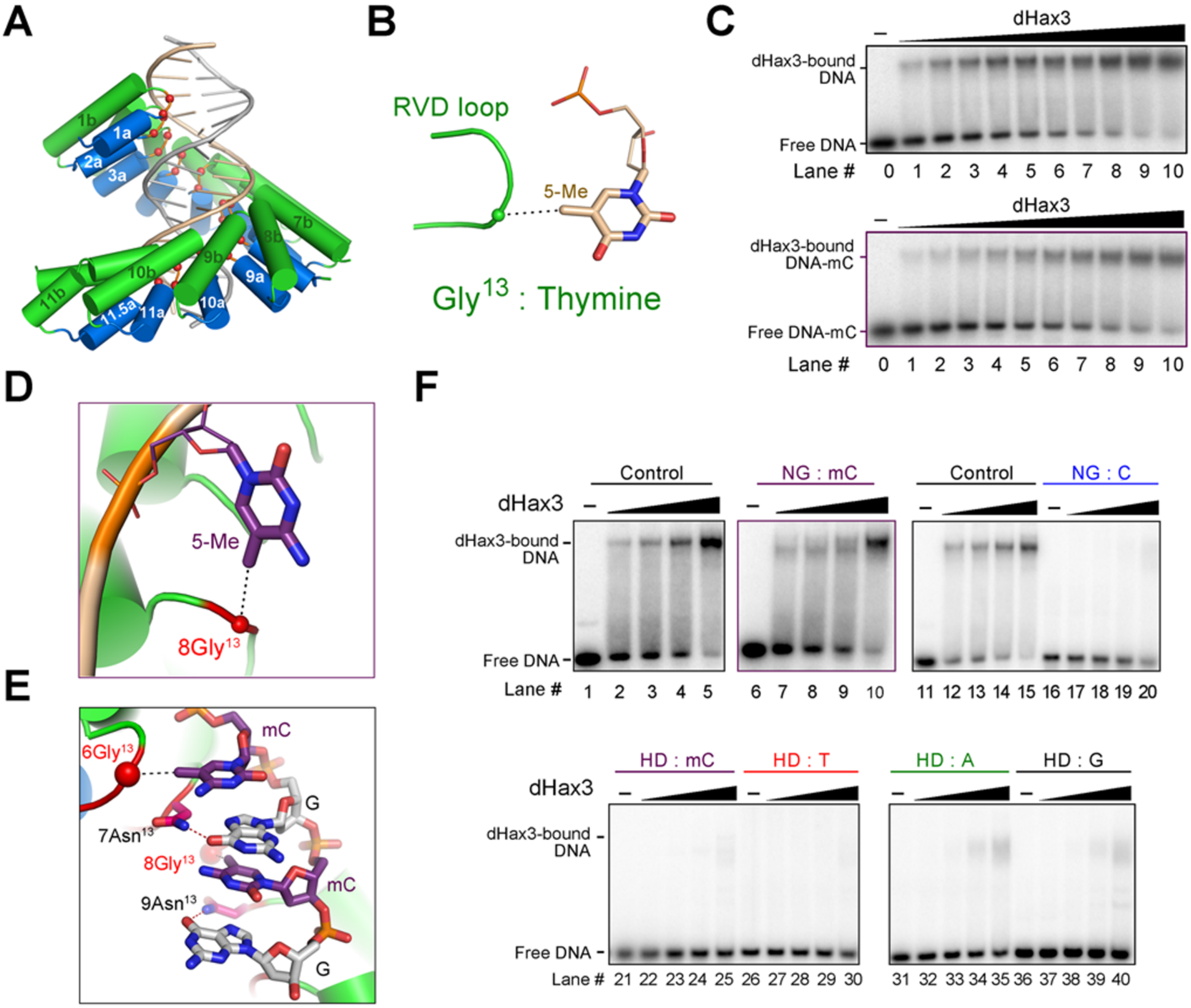

Figure 1 The TALE repeats containing code NG, but not HD, recognize 5-methylcytosine. **(A)** The crystal structure of DNA-bound dHax3 TALE repeats (PDB accession code 3V6T). RVDs in each repeat are shown as red spheres. The upper and lower DNA strands are colored gold and silver. **(B)** Structural analysis of the code NG→T suggested that NG may recognize 5-methycytosine (mC). **(C)** The DNA fragments with three bases T substituted with mC retained binding to TALE repeats. The protocol of EMSA is described in detail in Supplementary information, Data S1. **(D)** Crystal structure of dHax3 in complex with methylated DNA dHax3-5mC at 1.85 Å. The overall structure is identical to Figure 1A; see also Supplementary information, Figure S2A. Only one mC and the Gly^{13} in the corresponding repeat is shown to highlight the van der Waals interaction, which is indicated by the black dashed line, between the 5-methyl group of mC and the Cα atom of Gly. **(E)** Crystal structure of a dHax3 variant in complex with methylated DNA containing (mC)G(mC)G at 1.95 Å resolution. The hydrogen bond between Asn^{13} and base G is shown as red dashed line. The Cα atoms of Gly^{13} residues are shown as red spheres. **(F)** NG, but not HD, recognizes 5-methylcytosine. Lanes 1-20: The DNA of dHax3 box with six bases of T replaced by mC retained binding to dHax3 repeats, whereas replacement by C led to complete loss of binding. Lanes 21-40: HD specifically recognizes C. Substitution of the five cytosines in dHax3 box with any other nucleotides, including mC, leads to almost complete loss of binding with dHax3 repeats. All the structure figures were prepared with PyMOL [12].

cal to that in the first structure (Figure 1D). In fact, the two structures of dHax3 variants in complex with triply and doubly methylated DNA elements are nearly identical to that of dHax3 bound to the unmodified DNA [11], with root-mean-squared deviation values of less than 0.3 Å over more than 900 Cα atoms (Supplementary information, Figure S2D).

Encouraged by the structural findings, we replaced all six T bases by mC in the sense DNA strand. Subsequent EMSA study revealed that dHax3 retained similar binding to this DNA element as to the unmodified DNA (Figure 1F, lanes 1-10; Supplementary information, Figure S1B). In contrast, there was no detectable binding between dHax3 and the DNA element in which the six T bases were replaced by the base C (Figure 1F, lanes 11-20). This is rather striking, because this result suggests a qualitative and reliable method for differentiating the bases mC and C. We next examined whether the RVD code HD, which favors the base C, may also recognize mC. Substitution of the five C bases with mC or T in the sense DNA strand led to complete abrogation of DNA binding by dHax3 (Figure 1F, lanes 21-30). Substitution of the five C bases with A or G resulted in significant impairment of DNA binding (Figure 1F, lanes 31-40). These results illustrate the specific nature of mC recognition by TALE repeats involving the RVD NG, but not HD.

Our experimental characterization provides a molecular basis for distinguishing methylated and unmethylated cytosine. Binding of mC by TALE repeat through the RVD NG extends the DNA recognition code and has potential application in epigenetics and cancer research. For example, specific TALE repeats may be designed to recognize the hypermethylated DNA region; detection can be facilitated by fusing TALEs with fluorescence proteins.

Our study also strongly argues that the *in vivo* methylation status of the target DNA sequence must be considered for the design of specific DNA-binding TALEs. Methylation of the base C *in vivo* might render the DNA sequence unfit for binding by the designed TALEs. Because the methylation status of DNA sequences is frequently under dynamic control, one would have to design at least two TALEs for one DNA sequence (i.e., one for methylated and one for unmethylated). In fact, assessment of methylation status of specific DNA sequences *in vivo* can be greatly facilitated through quantification of fluorescence signal of designed GFP-TALEs. Alternatively, the CpG sequences may be avoided for the application of TALEs, although this practice will somehow limit the potential application. Despite these complexities, the discovery of mC binding by TALEs with RVD NG opens a number of exciting opportunities.

Acknowledgments

This work was supported by funds from the Ministry of Science and Technology of China (2011CB910501, 2009CB918801, and 2009CB918802), the National Natural Science Foundation of China (30888001, 91017011 and 31070644) and Tsinghua University. The atomic coordinates have been deposited in the Protein Data Bank (PDB; accession code: 4GJP and 4GJR).

Dong Deng[1, 2, *], Ping Yin[1, 2, *], Chuangye Yan[2], Xiaojing Pan[1, 2], Xinqi Gong[1, 2], Shiqian Qi[2], Tian Xie[1, 2], Magdy Mahfouz[3], Jian-Kang Zhu[4], Nieng Yan[1, 2], Yigong Shi[2]

[1]*State Key Laboratory of Bio-membrane and Membrane Biotechnology,* [2]*Tsinghua-Peking Center for Life Sciences, Center for Structural Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China;* [3]*Center for Plant Stress Genomics and Technology, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia;* [4]*Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN 47907, USA*

*These two authors contributed equally to this work.
Correspondence: Nieng Yan[a], Yigong Shi[b]
[a]E-mail: nyan@tsinghua.edu.cn
[b]E-mail: shi-lab@tsinghua.edu.cn

References

1 Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 2010; **11**:204-220.

2 He XJ, Chen T, Zhu JK. Regulation and function of DNA methylation in plants and animals. *Cell Res* 2011; **21**:442-465.

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

4 Boch J, Bonas U. *Xanthomonas* AvrBs3 family-type III effectors: discovery and function. *Annu Rev Phytopathol* 2010; **48**:419-436.

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

6 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.

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

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

9 McMahon MA, Rahdar M, Porteus M. Gene editing: not just for translation anymore. *Nat Methods* 2012; **9**:28-31.

10 Li L, Piatek MJ, Atef A, *et al.* Rapid and highly efficient construction of TALE-based transcriptional regulators and nucleases for genome modification. *Plant Mol Biol* 2012; **78**:407-416

11 Deng D, Yan C, Pan X, *et al.* Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* 2012; **335**:720-723.

12 DeLano WL. The PyMOL Molecular Graphics System. 2002; http://www.pymol.org

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