Open
LETTER TO THE EDITOR

Cell Research (2014) 24:628-631.

www.nature.com/cr

Complete decoding of TAL effectors for DNA recognition

Cell Research (2014) **24**:628-631. doi:10.1038/cr.2014.19; published online 11 February 2014

Dear Editor,

The most striking feature of a transcription activator-like effector (TALE) is the presence of a central DNA-binding region composed of tandem repeats of about 34 amino acids [1]. Two hypervariable residues at positions 12 and 13 (repeat-variable diresidues or RVDs) in each repeat bind to DNA, and this modular DNA-binding feature of TALE repeats has inspired the development of custom-designed TALE repeats for gene editing [2, 3, 4, 5]. The nucleotide recognition preference of the commonly used RVDs has been experimentally or computationally determined [2, 5]. For instance, RVD NN has a high preference for both G and A. The rare RVDs, NK and NH, have better specificity for guanine than NN, but their affinity is relatively lower [3, 6, 7]. We thus decided to conduct a thorough investigation of potential RVDs, which cover all possible combinations of amino acid diresidues, for their DNA recognition capabilities.

We set up a screening platform composed of an artificial TALE-VP64-mCherry construct, which expresses RVD (XX') in 3-tandem repeat format (from 7^{th} to 9^{th}, TALE-$(XX')_3$), and 4 corresponding EGFP reporter constructs, in which potential TALE-$(XX')_3$-binding sites composed of 3 consecutive nucleotides (A, T, C or G) are located in front of a minCMV promoter and its downstream *EGFP* gene (Figure 1A, Supplementary information, Figure S1A and Data S1). To test this system, we made a control TALE (TALE-Ctrl) that is identical to TALE-$(XX')_3$ except for the repeat domain (Supplementary information, Figure S1A), and confirmed that it could not activate any of the 4 EGFP reporters (Supplementary information, Figure S1B), thus serving as the control for basal activity. We then constructed 4 TALE-$(XX')_3$ expression plasmids by placing the common RVDs (NI, NG, HD and NN) in the middle to target the 3A, 3T, 3C and 3G EGFP reporters, respectively (Supplementary information, Figure S1C). These TALE-$(XX')_3$ constructs were individually introduced into HEK293T cells together with 1 of the 4 EGFP reporter plasmids to examine their specificities, which were determined by the fold induction of EGFP expression compared with the basal level (Supplementary information, Data S1). The identity of XX' determined TALE-$(XX')_3$ specificity on different EGFP reporters: NI, NG, HD and NN predominantly recognized A, T, C and G or A, respectively. This result is consistent with the current knowledge regarding the base preference of these 4 common RVDs, demonstrating that this artificial system is suitable for testing the DNA recognition ability of RVDs (Supplementary information, Figure S1D-S1E).

To quantitatively measure the base preference of all theoretical RVDs, we created a library of TALE-$(XX')_3$ constructs, which covers a total of 400 types of RVDs, following a special protocol combined with the ULtiMATE assembly method [8] (Supplementary information, Figure S2 and Data S1). X and X' correspond to the 12^{th} and 13^{th} amino acids in a classical TALE module, respectively. We introduced each of the 400 TALE-$(XX')_3$ constructs (Supplementary information, Tables S1 and S2) individually into HEK293T cells together with 1 of the 4 EGFP reporter plasmids and measured both the EGFP and mCherry levels by FACS analysis. We then determined the base-recognition efficiencies of the 400 diresidues. A total of 1 600 data points were summarized in 3 formats: heat map (Figure 1B), histograms categorized by the 13^{th} residue (X') (Supplementary information, Figure S3) and the 12^{th} residue (X) (Supplementary information, Figure S4). The results obtained from this screening provide substantial information regarding the base preference of all theoretical RVDs. In addition to NI, NG, HD and NN, all the natural RVDs and a few artificial RVDs showed base-recognition preferences that were similar to those reported previously [2, 5, 6, 7] (Supplementary information, Table S3). Besides these 25 RVDs, we determined the DNA base-recognition preference of the remaining 375 RVDs that did not evolve naturally and have not been previously examined.

Notably, many of these artificial RVDs showed a distinct preference for DNA bases compared with the 25 reported RVDs, and only a few of them start with 1 of the 2 frequently occurring amino acids, Asn and

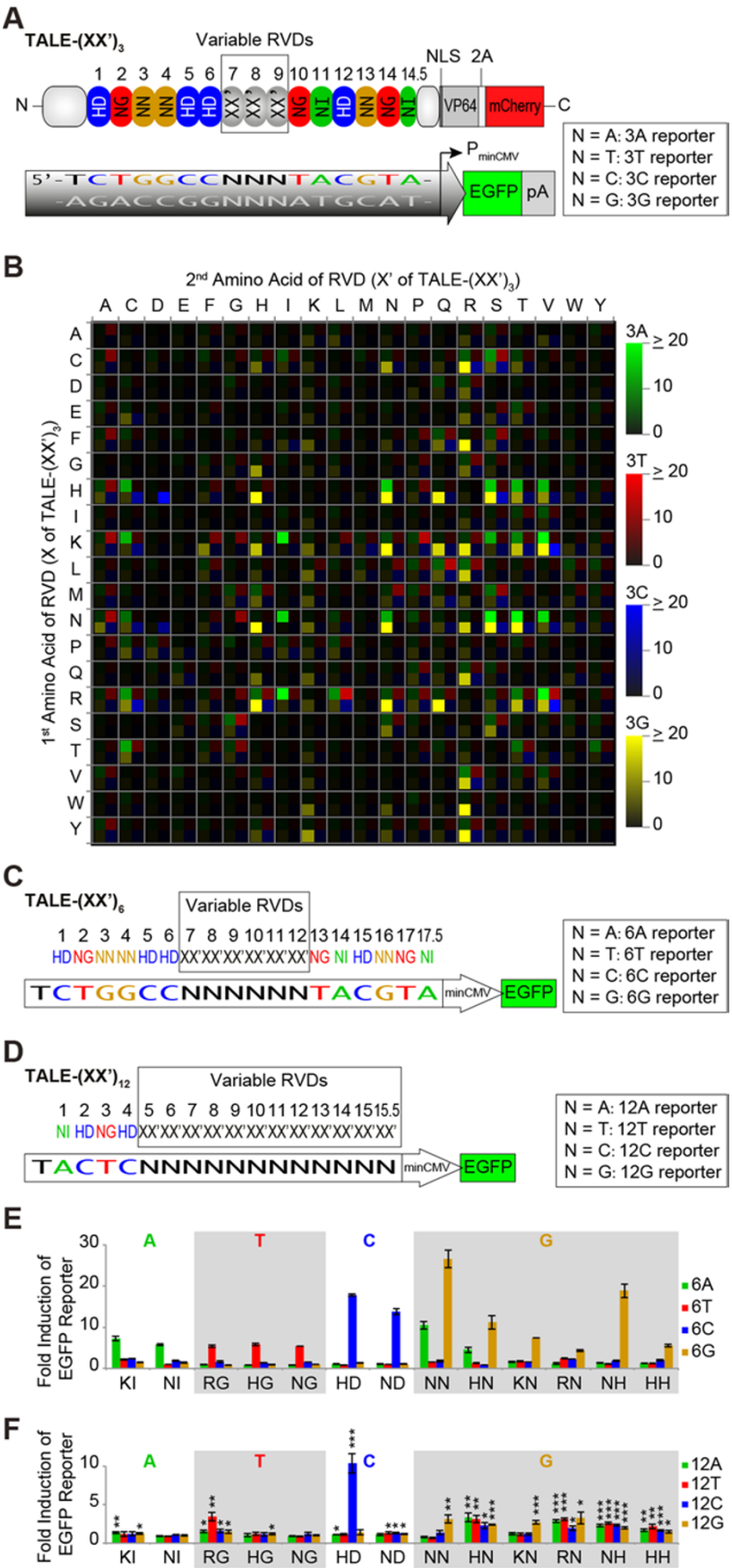
A
TALE-(XX')3
Variable RVDs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 14.5
HD NG NN NN HD HD XX' XX' XX' NG NI HD NN NG NI
N
VP64
mCherry
C
NLS
2A
PminCMV
5'-TCTGGCCNNNTACGTA-
-AGACCGGNNNATGCAT-
EGFP
pA
N = A: 3A reporter
N = T: 3T reporter
N = C: 3C reporter
N = G: 3G reporter
B
2nd Amino Acid of RVD (X' of TALE-(XX')3)
A C D E F G H I K L M N P Q R S T V W Y
1st Amino Acid of RVD (X of TALE-(XX')3)
3A
3T
3C
3G
≥ 20
10
0
C
TALE-(XX')6
Variable RVDs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 17.5
HD NG NN NN HD HD XX' XX' XX' XX' XX' XX' NG NI HD NN NG NI
TCTGGCCNNNNNNTACGTA
minCMV
EGFP
N = A: 6A reporter
N = T: 6T reporter
N = C: 6C reporter
N = G: 6G reporter
D
TALE-(XX')12
Variable RVDs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 15.5
NI HD NG HD XX' XX' XX' XX' XX' XX' XX' XX' XX' XX' XX' XX'
TACTCNNNNNNNNNNNN
minCMV
EGFP
N = A: 12A reporter
N = T: 12T reporter
N = C: 12C reporter
N = G: 12G reporter
E
Fold Induction of EGFP Reporter
30
20
10
0
A
T
C
G
KI NI RG HG NG HD ND NN HN KN RN NH HH
6A
6T
6C
6G
F
Fold Induction of EGFP Reporter
10
5
0
A
T
C
G
KI NI RG HG NG HD ND NN HN KN RN NH HH
12A
12T
12C
12G

Figure 1 A Complete assessment of TALE RVD efficiencies and specificities. **(A)** Design of the screening system for novel TALE RVDs. **(B)** A heat map generated from library screening of TALE-$(XX')_3$ with four reporters (3A, 3T, 3C, and 3G) reflecting the base preference of 400 RVDs. EGFP activities from different reporters were coded by different colors representing the reporter identities (3A, green; 3T, red; 3C, blue; 3G, yellow), and the brightness of the colors indicates the fold induction of reporters by TALE-$(XX')_3$ compared to the basal levels. The single-letter abbreviations for the amino acids are used. **(C)** Design of TALE-$(XX')_6$ and its corresponding reporters. **(D)** Design of TALE-$(XX')_{12}$ and its corresponding reporters. **(E-F)** Base preference of RVDs in TALE-$(XX')_6$ **(E)** and TALE-$(XX')_{12}$ **(F)**. RVDs were clustered by base preference. The x-axis labels indicate the variable RVDs tested in TALE-$(XX')_6$ or TALE-$(XX')_{12}$. Data are means ± SD, $n = 3$; $^*P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.005$.

His (Figure 1B, Supplementary information, Figures S3 and S4). From these artificial RVDs, we selected those that showed potential base-recognition preference based on the criteria shown in Supplementary information, Data S1 for further intensive analyses. We found that the adenine recognition ability of KI and RI was similar to that of NI (Supplementary information, Figure S5A). For thymine recognition, we identified 3 additional RVDs aside from NG, which all end with Gly (RG, KG and HG), and seven RVDs that all end with Ala (KA, CA, FA, YA, RA, PA, and AA), but appeared to have higher background, especially for C recognition (Supplementary information, Figure S5B). HD and ND, as reported previously [2, 5, 7], were optimal RVDs for C recognition, with almost no non-specific recognition of other bases (Supplementary information, Figure S5C). Five groups of RVDs were identified to recognize guanine, with each group sharing the same 13th residue: Asn (N), His (H), Arg (R), Gln (Q), or Lys (K). Most of these RVDs predominantly recognized guanine except for HN and NN (Supplementary information, Figure S5D). These data support the prediction from previous TALE structural studies suggesting that the 13th residues of TALE repeats make the base-specific contact [9, 10]. Nevertheless, our data indicate that the 12th residue also affects RVD specificity. For example, with the same N_{13}, KN and RN only recognized G, whereas HN and NN recognized both A and G, and LN and MN preferred T and C. Similarly, HQ, KQ and RQ preferentially recognized G, whereas LQ preferred T (Supplementary information, Figures S3-S6).

To further examine the base-recognition preference of RVDs, we created two additional artificial platforms with increased stringency, in which multiple TALE repeats carrying the same RVDs were aligned in tandem: TALE-$(XX')_6$ and its corresponding EGFP reporter constructs (6A, 6T, 6C, and 6G) (Figure 1C) were used to test RVDs in 6-tandem repeat format, and TALE-$(XX')_{12}$ and its corresponding EGFP reporter constructs (12A, 12T, 12C, and 12G) (Figure 1D) were used to test RVDs in 12-tandem repeat format (Supplementary information, Table S1 and Figure S2). In addition to the 4 most common RVDs, which were used as controls, we mainly chose those that demonstrated outstanding base-recognition specificities from the initial screening. We found that KI and NI functioned similarly with respect to A recognition in the 6-repeat format (Figure 1E). The activities of TALE-$(RG)_6$ and TALE-$(HG)_6$ were similar to TALE-$(NG)_6$ for 6T recognition (Figure 1E), whereas TALE-$(KG)_6$ showed reduced specificity for 6T (Supplementary information, Figure S7). HD and ND again demonstrated strong C preference (Figure 1E). KN, RN, NH and HH showed specific G recognition with variable efficiencies in 6-tandem repeats, whereas NN and HN recognized both G and A as in the 3-repeat format (Figure 1E), and the 6G preference of TALE-$(XX')_6$ containing either NR, FR, KH, NK, FK or RQ was significantly reduced (Supplementary information, Figure S7). Interestingly, only TALE-$(XX')_{12}$ with RG (for T), HD (for C), NN (for G) and KN (for G) in 12-tandem repeats maintained recognition efficiency and specificity (Figure 1F). This result is somewhat surprising for RG as it is assumed that RVDs ending with Gly cannot form hydrogen bonds with thymine [10]. Consistent with previous reports [6, 7], neither TALE-$(NH)_{12}$ nor TALE-$(HH)_{12}$ could support 12G reporter activation. Considering the strong preference of NH for G in the 6-repeat format, it is unclear why TALE-$(KN)_{12}$ but not TALE-$(NH)_{12}$ retained activity for the 12G reporter. By the same token, it is also unclear why TALE-$(ND)_{12}$ completely lost its preference for the 12C reporter (Figure 1F). Although the combination of the 12th and 13th amino acids determines the ultimate binding activity of TALE, the increase of repeat number also leads to the decrease or even complete loss of DNA-recognition activity of TALE, which is likely due to either steric or static repulsion between consecutive TALE repeat units.

To further evaluate these novel RVDs, we applied KN and RG in TALEN assembly in place of NN and NG, respectively, and compared them with conventional RVDs in TALENs-mediated DNA cleavage by measuring indel rates. $TALENs_{KN}$ for G-targeting showed similar efficiency in creating indels

as $TALENs_{NN}$ in 2 independent tests, and both of them performed better than $TALENs_{NH}$. On the contrary, $TALENs_{RG}$, although functional, were less effective than $TALENs_{NG}$ (Supplementary information, Table S4). It is possible that other diresidues newly revealed in this study could function as valid RVDs in recognizing DNA bases with high specificity. However, rigorous tests are needed in order to more accurately determine their DNA recognition capabilities.

In addition, we identified a significant number of RVDs that target multiple DNA bases (Supplementary information, Table S5). The availability of RVDs that target different combinations of bases in a degenerate manner may provide certain flexibility in future application such as engineering of sophisticated genetic circuitry [11].

By further deciphering the DNA base preference of all RVDs, natural or artificial, we can achieve a clear understanding of the mechanism that guides the base preference of TALE RVDs. Comprehensive information regarding the specific DNA associations of all RVDs may improve the application of TAL effectors in bioengineering and precision therapy.

Acknowledgments

We thank Xiaoyu Li (PKU) for technical advice regarding oligo synthesis and Yanyi Huang (PKU) for critical comments on the manuscript. This work was supported by the National Basic Research Program of China (2010CB911800), the National Natural Science Foundation of China (NSFC31070115, NSFC31170126), and the Peking-Tsinghua Center for Life Sciences.

Junjiao Yang[1, *], Yuan Zhang[1, *], Pengfei Yuan[1], Yuexin Zhou[1], Changzu Cai[1], Qingpeng Ren[1], Dingqiao Wen[1, 3], Coco Chu[3], Hai Qi[2], Wensheng Wei[1]

[1]State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing 100871, China; [2]Tsinghua-Peking Center for Life Sciences, Laboratory of Dynamic Immunobiology, School of Medicine, Tsinghua University, Beijing 100084, China; [3]Current address: Department of Computer Science, Rice University, Houston, TX 77005, USA

*These two authors contributed equally to this work.
Correspondence: Wensheng Wei
Tel: +86-10-62757227
E-mail: wswei@pku.edu.cn

References

1 Boch J, Bonas U. *Annu Rev Phytopathol* 2010; **48**:419-436.
2 Boch J, Scholze H, Schornack S, *et al*. *Science* 2009, **326**:1509-1512.
3 Miller JC, Tan S, Qiao G, *et al*. *Nat Biotech* 2011; **29**:143-148.
4 Bogdanove AJ, Voytas DF. *Science* 2011; **333**:1843-1846.
5 Moscou MJ, Bogdanove AJ. *Science* 2009; **326**:1501.
6 Streubel J, Blucher C, Landgraf A, *et al*. *Nat Biotech* 2012; **30**:593-595.
7 Cong L, Zhou R, Kuo YC, *et al*. *Nat Commun* 2012; **3**:968.
8 Yang J, Yuan P, Wen D, *et al*. *PLoS One* 2013; **8**:e75649.
9 Deng D, Yan C, Pan X, *et al*. *Science* 2012; **335**:720-723.
10 Mak AN, Bradley P, Cernadas RA, *et al*. *Science* 2012; **335**:716-719.
11 Aouida M, Piatek MJ, Bangarusamy DK, *et al*. *Curr Genet* 2013 Oct 1. doi: 10.1007/s00294-013-0412-z

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