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

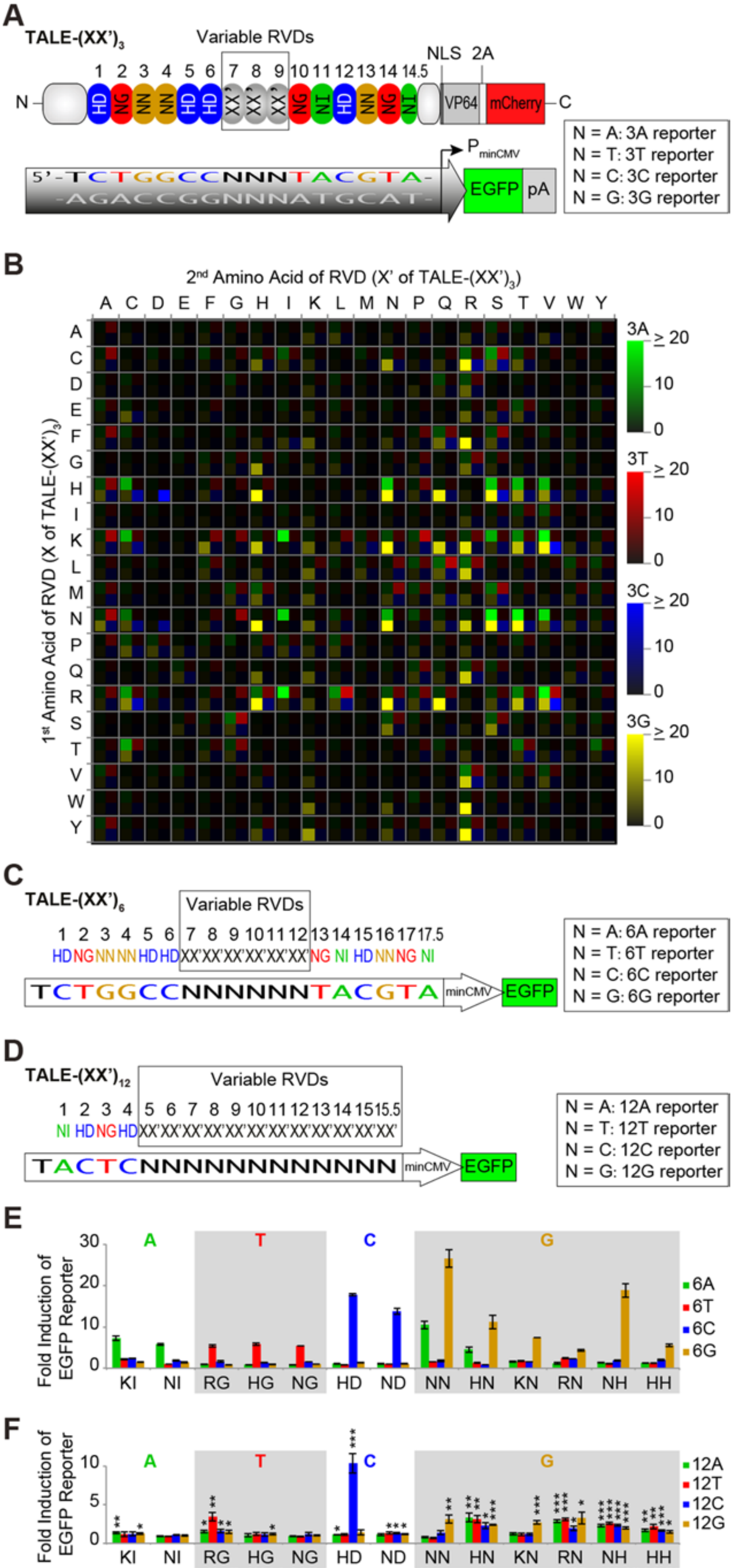

A
TALE-(XX')3
Variable RVDs
NLS 2A
N
HD NG NN NN HD HD XX' XX' XX' NG NI HD NN NG NI
VP64 mCherry
C
P minCMV
5'-TCTGGCCNNNTACGTA-
-AGACCGGNNNATGCAT-
EGFP pA
N = A: 3A reporter
N = T: 3T reporter
N = C: 3C reporter
N = G: 3G reporter
B
2nd Amino Acid of RVD (X' of TALE-(XX')3)
A C D E F G H I K L M N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

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]

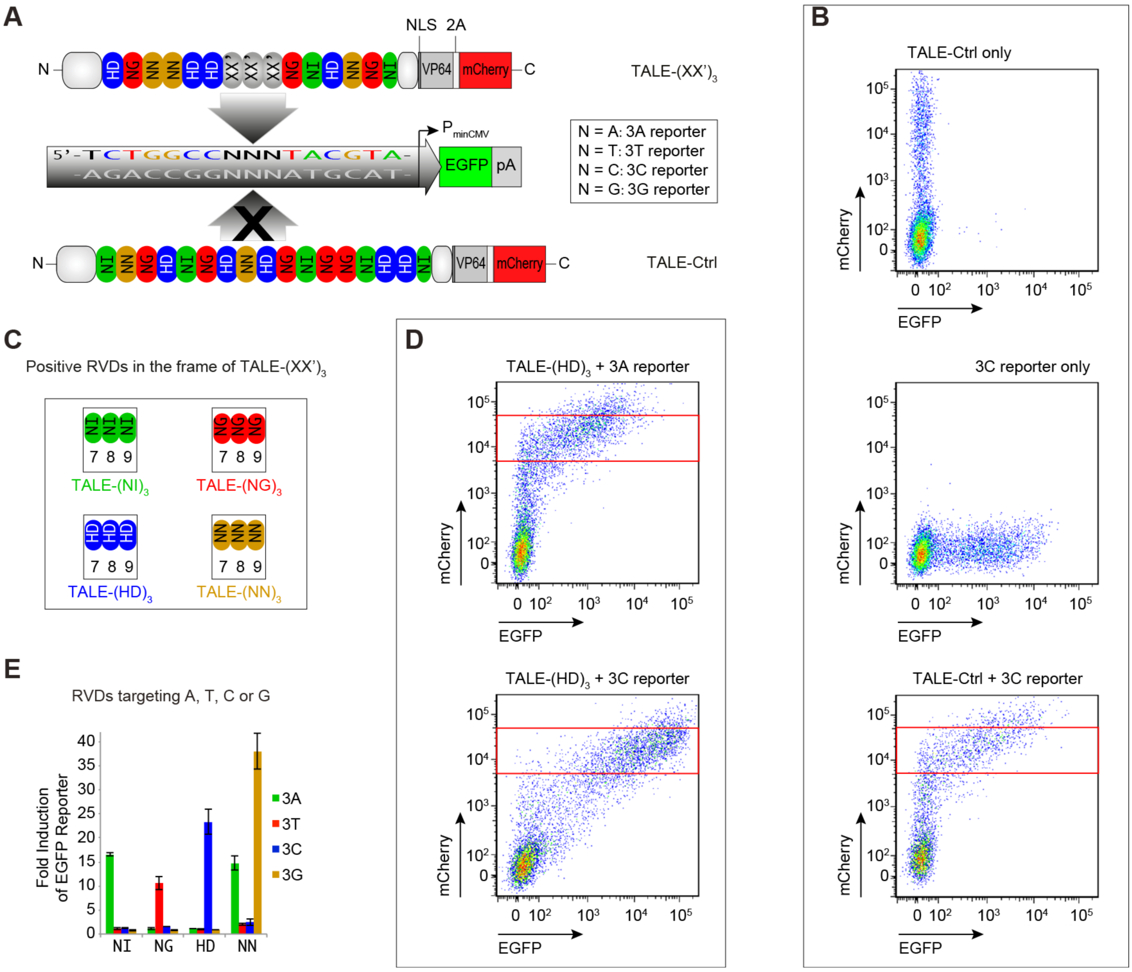
[1]*State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing 100871, China;* [2]*Tsinghua-Peking Center for Life Sciences, Laboratory of Dynamic Immunobiology, School of Medicine, Tsinghua University, Beijing 100084, China;* [3]*Current address: Department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

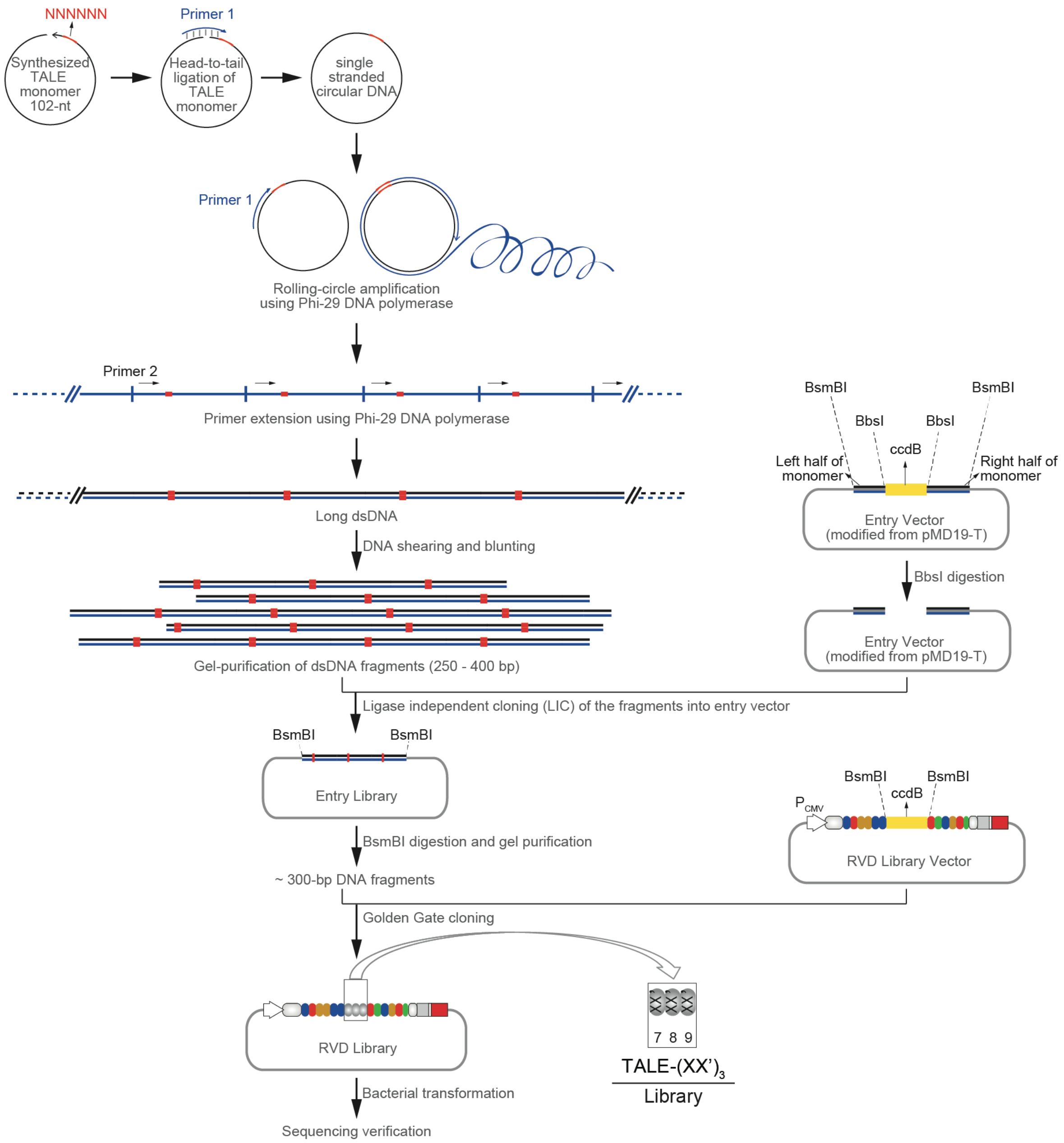
References

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

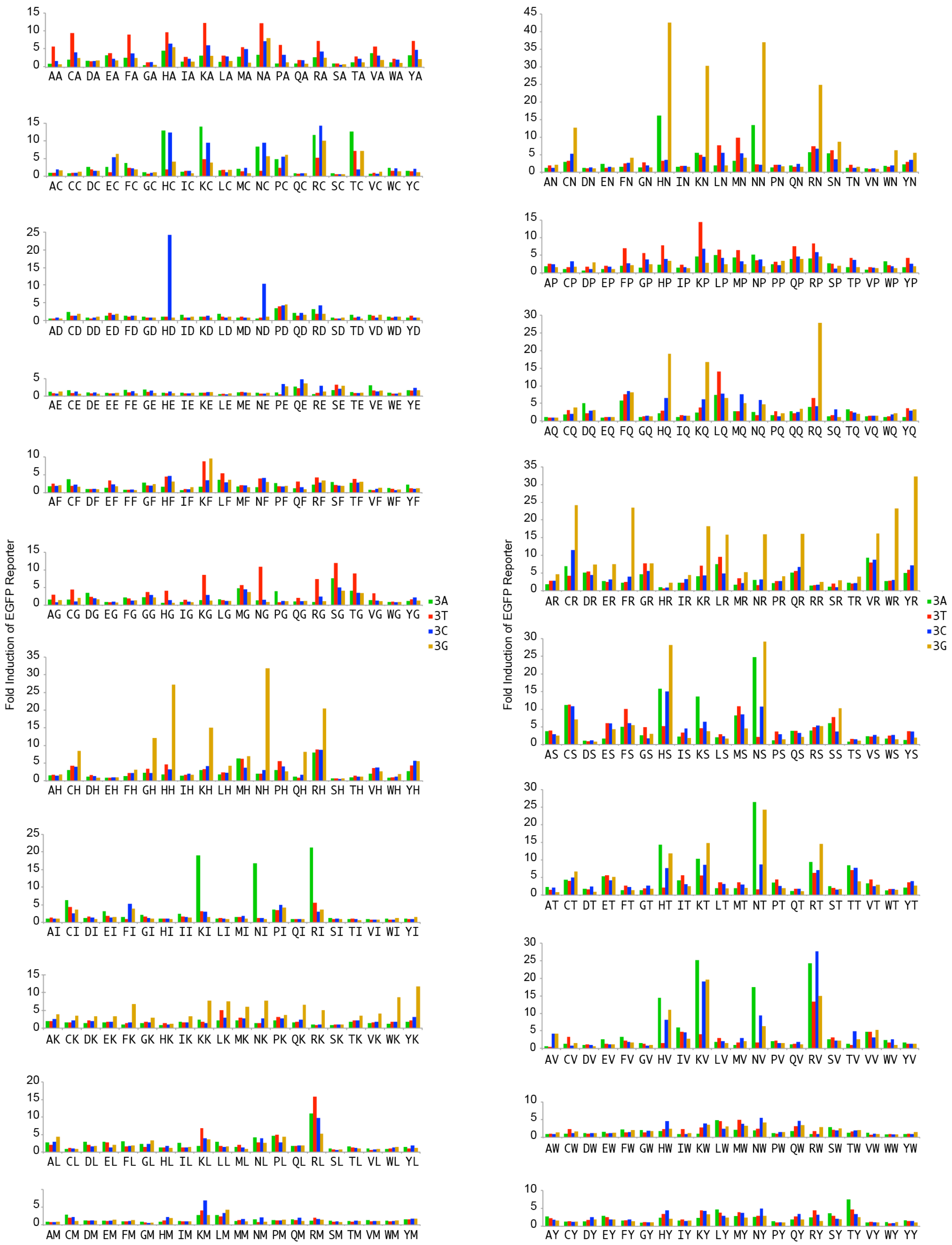
(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)



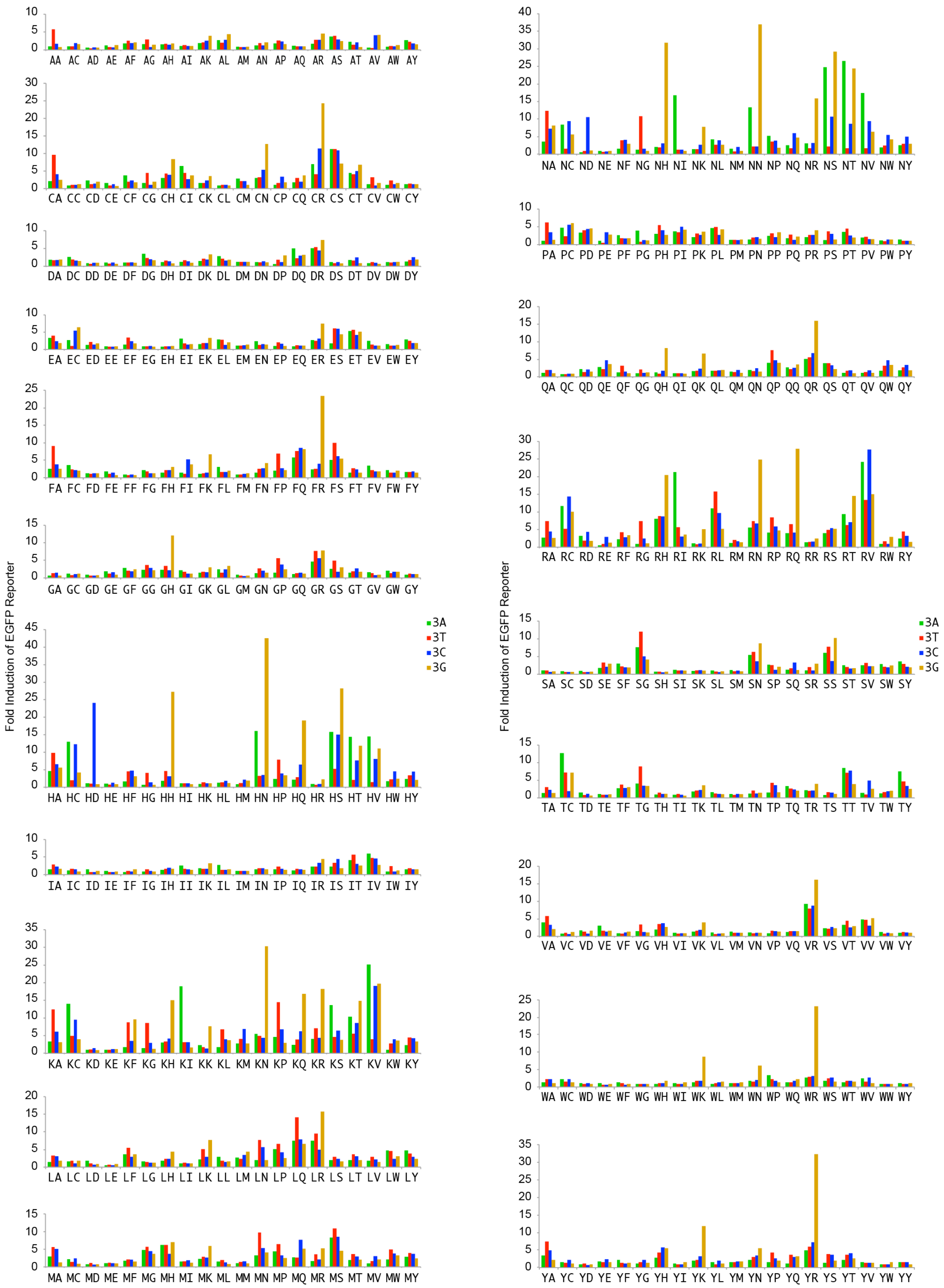
Supplementary information, Figure S1 Screening system for the assessment of TALE RVD efficiencies and specificities. (**A**) Design of the screening system for novel TALE RVDs. The customized TALEs used for RVD screening contained 14.5 repeats fused with the VP64 trans-activation domain and 2A peptide-linked mCherry. The variable diresidues (XX') for testing were placed in the 7^{th} - 9^{th} repeat modules, and the customized TALE was designated as TALE-(XX')$_3$. X and X' represents the 12^{th} and 13^{th} amino acids in the 7^{th} - 9^{th} repeat modules, respectively. To determine the DNA recognition specificity of variable RVDs, four reporters were constructed, consisting of TALE-(XX')$_3$ binding sites with three consecutive nucleotides (A, T, C or G) substituted at positions 7 - 9 in front of a minimal CMV promoter (P_{minCMV}) and its downstream EGFP gene. Construct encoding TALE-Ctrl has the identical backbone as TALE-(XX')$_3$ except that its TALE repeat region is different as indicated, which does not match with any reporters. (**B**) FACS analysis of HEK293T cells transfected with TALE-Ctrl only, 3C reporter only and TALE-Ctrl plus 3C reporter (from top to bottom). Red box indicates the region of data collection. (**C**) Customized TALE-$(XX')_3$ for testing the DNA binding activity of commonly used RVDs (NI, NG, HD and NN). (**D**) Representative FACS analysis of HEK293T cells co-transfected with TALE-$(HD)_3$ and the 3A (top) or 3C reporter (bottom). Red boxes indicate the region of data collection. (**E**) Base binding activity of the common RVDs in (**C**). The horizontal axis labels indicate the variable RVD (XX') for testing in TALE-$(XX')_3$. The color-coded bars represent the fold induction levels of the different reporters (A, green; T, red; C, blue; and G, yellow) in this and the subsequent figures. Data are means ± s.d., n = 3.



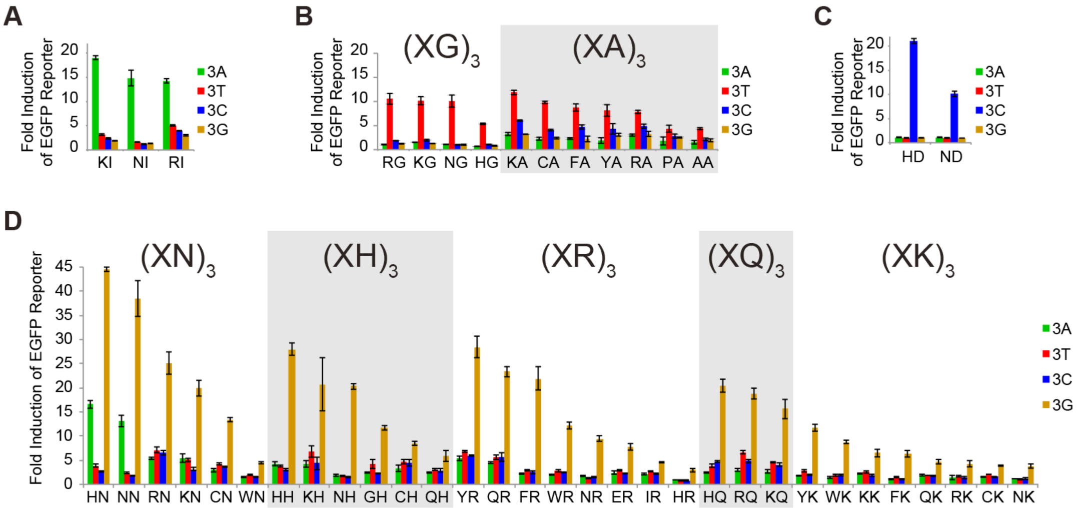
Supplementary information, Figure S2 Schematic of TALE-(XX')$_3$ library construction for novel RVD screening. A 102-nt monomer encoding a standard TALE repeat unit, containing six random nucleotides at the RVD-encoding region, was synthesized and subsequently cyclized. Rolling-circle amplification of these single-strand circular DNA templates was conducted using phi29 DNA polymerase and primer 1, and dsDNA fragments were obtained from primer extension using primer 2. After ultrasonic shearing followed by DNA blunting, 250-400 bp DNA fragments were isolated. After gel purification, these DNA fragments were cloned into a pre-made entry vector through the LIC method. BsmBI digestion of clones in the entry library produced ~300 bp DNA fragments, which were subsequently ligated into a pre-made RVD library vector. After bacterial transformation and sequencing validation, we were able to obtain 400 types of plasmids encoding customized TALEs with three repeat modules in the middle (7^{th} to 9^{th}) carrying the variable RVDs for testing. A detailed protocol is provided in the Supplementary Methods.



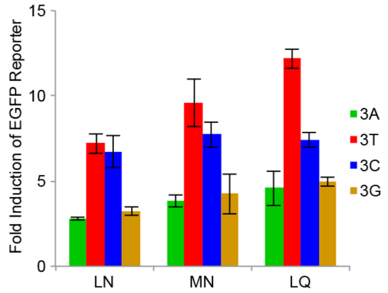
Supplementary information, Figure S3 Base preference of 400 RVDs from the screening of the TALE-(XX')$_3$ library. The binding efficiencies and specificities of the variable RVDs (XX') in each customized TALE were assayed based on the fold induction of EGFP reporters, by comparing with the basal level of EGFP in HEK293T cells transfected with the reporter plasmid and the customized TALE plasmid containing unmatched TALE repeats. The EGFP fluorescence intensity assayed from FACS analysis was normalized to the corresponding mCherry fluorescence intensity. The fold induction of the EGFP reporter of all 400 types of TALE-$(XX')_3$ (corresponding to four reporters) was categorized by the 13^{th} residue (X'), and the data are listed in alphabetical order according to the 12^{th} residue (X) of the variable RVD-carrying TALE. The x-axis labels indicate the variable RVDs (XX') tested in TALE-$(XX')_3$. The color-coded bars represent the fold induction of the different reporters (A, green; T, red; C, blue; and G, yellow) in this and the subsequent figures.


Fold Induction of EGFP Reporter
3A
3T
3C
3G
AA AC AD AE AF AG AH AI AK AL AM AN AP AQ AR AS AT AV AW AY
CA CC CD CE CF CG CH CI CK CL CM CN CP CQ CR CS CT CV CW CY
DA DC DD DE DF DG DH DI DK DL DM DN DP DQ DR DS DT DV DW DY
EA EC ED EE EF EG EH EI EK EL EM EN EP EQ ER ES ET EV EW EY
FA FC FD FE FF FG FH FI FK FL FM FN FP FQ FR FS FT FV FW FY
GA GC GD GE GF GG GH GI GK GL GM GN GP GQ GR GS GT GV GW GY
HA HC HD HE HF HG HH HI HK HL HM HN HP HQ HR HS HT HV HW HY
IA IC ID IE IF IG IH II IK IL IM IN IP IQ IR IS IT IV IW IY
KA KC KD KE KF KG KH KI KK KL KM KN KP KQ KR KS KT KV KW KY
LA LC LD LE LF LG LH LI LK LL LM LN LP LQ LR LS LT LV LW LY
MA MC MD ME MF MG MH MI MK ML MM MN MP MQ MR MS MT MV MW MY
NA NC ND NE NF NG NH NI NK NL NM NN NP NQ NR NS NT NV NW NY
PA PC PD PE PF PG PH PI PK PL PM PN PP PQ PR PS PT PV PW PY
QA QC QD QE QF QG QH QI QK QL QM QN QP QQ QR QS QT QV QW QY
RA RC RD RE RF RG RH RI RK RL RM RN RP RQ RR RS RT RV RW RY
SA SC SD SE SF SG SH SI SK SL SM SN SP SQ SR SS ST SV SW SY
TA TC TD TE TF TG TH TI TK TL TM TN TP TQ TR TS TT TV TW TY
VA VC VD VE VF VG VH VI VK VL VM VN VP VQ VR VS VT VV VW VY
WA WC WD WE WF WG WH WI WK WL WM WN WP WQ WR WS WT WV WW WY
YA YC YD YE YF YG YH YI YK YL YM YN YP YQ YR YS YT YV YW YY

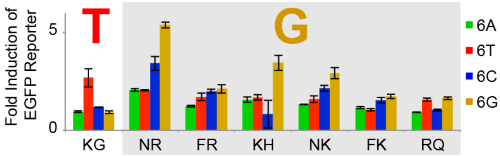
Supplementary information, Figure S4 Base preference of 400 RVDs from the screening of the TALE-(XX')$_3$ library. The data points are the same as those in Figure 1B and Figure S3. The fold induction of the EGFP reporter of all 400 types of TALE-(XX')$_3$ (corresponding to the four reporters) was categorized by the 12th residue (X), and the data are listed in alphabetical order according to the 13th residue (X') of the variable RVD-carrying TALE. The x-axis labels indicate the variable RVDs (XX') tested in TALE-(XX')$_3$.



Supplementary information, Figure S5 Base preferences of RVDs. RVDs were clustered by position 13 (X') of the variable RVDs and ranked by fold induction of EGFP in the corresponding reporter construct. The x-axis labels indicate the variable RVDs tested. Data are means ± s.d., n = 3. (**A**) Adenine-targeting RVDs. (**B**) Thymine-targeting RVDs. (**C**) Cytosine-targeting RVDs. (**D**) Guanine-targeting RVDs.



Supplementary information, Figure S6 Base preference of RVDs in TALE-(XX')$_3$. The x-axis labels indicate the variable RVDs tested in TALE-(XX')$_3$. Data are means ± s.d., n = 3.



Supplementary information, Figure S7 Base preference of RVDs in TALE-(XX')$_6$. The results of the fold induction of TALE-(XX')$_6$ (corresponding to the four reporters) are shown on the y-axis, and the RVDs are clustered by base/reporter preference and ranked by fold induction of EGFP of the corresponding reporter. The x-axis labels indicate the variable RVDs tested in TALE-(XX')$_6$. Data are means ± s.d., n = 3.

Supplementary information, Table S1. Sequences of RVD libraries.

TALE-(XX')$_3$ DNA sequence:

atgaggaccaggctgccatctcccctgccccttcccccgcatttagcgctgggagctttagcgacctgcttaggcagttcgaccccagcttgttcaacaccagcctgtttga
cagcctgcctcccttcggagcgcaccacaccgaagccgccaccggcgagtgggacgaggtgcaaagcggcctgagggcagcggacgctcctccgccaaccatgag
ggtggcagtgacagcagctaggccccctcgggcaaaacctgcacccaggagaagggctgcccaacccagcgacgcgagtccagccgcacaggtggacctcagga
cgctgggctacagccagcaacagcaagagaagatcaagcccaaagtaaggagcaccgtggcccagcaccacgaggccctggtgggtcacggcttcacccacgcg
catatcgttgctctgagccaacatcccgcagctctgggtaccgttgcggtgaagtatcaggacatgatcgcggcactgcctgaagctacacacgaagccatagtgggcgtt
ggcaagcagtggagcggtgccagagcgcttgaggcactgttgacggtggctggcgagctgaggggaccgccactgcaactggacaccggccaactgctgaagatcg
ccaagaggggaggcgtgacggcggtggaggccgtgcatgcctggaggaatgccctgaccggcgcgcccctgaaccttacgcctgagcaagtcgttgcgatcgcctcc
cacgacggcggaaaacaggctttggaaaccgtgcagcggttgctgcccgttttgtgccaagcccacggactcactccggaacaggtggtcgcaatcgcgagcaacggc
ggcggcaagcaagcccttgagacagtccaaagacttttgcctgtcctttgtcaggcgcatggcctcactccggaacaggtggtcgcaatcgcgagcaacaacggcggca
agcaagcccttgagacagtccaaagacttttgcctgtcctttgtcaggcgcatggcttgacccccgaacaggttgtagccatagcttctaacaacggaggtaagcaggcac
tggaaaccgtgcagcgcctgctcccagtactgtgtcaggctcatgggctgacaccagagcaagtagtggctattgcaagtcacgacggtggcaaacaagcgctggaga
ccgtgcagaggctccttccggtgctctgccaagcacacggtcttacgcctgagcaagtcgttgcgatcgcctcccacgacggcggaaaacaggctttggaaaccgtgca
gcggttgctgcccgttttgtgccaagcccacggattgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcc
tgctcccagtactgtgtcaggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcc
cagtactgtgtcaggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcccagta
ctgtgtcaggctcatgggctcactccggaacaggtggtcgcaatcgcgagcaacggcggcggcaagcaagcccttgagacagtccaaagacttttgcctgtcctttgtca
ggcgcatggcctgacaccagagcaagtagtggctattgcaagtaacatcggtggcaaacaagcgctggagaccgtgcagaggctccttccggtgctctgccaagcaca
cggtcttacgcctgagcaagtcgttgcgatcgcctcccacgacggcggaaaacaggctttggaaaccgtgcagcggttgctgcccgttttgtgccaagcccacggactca
ctccggaacaggtggtcgcaatcgcgagcaacaacggcggcaagcaagcccttgagacagtccaaagacttttgcctgtcctttgtcaggcgcatggcctcactccgga
acaggtggtcgcaatcgcgagcaacggcggcggcaagcaagcccttgagacagtccaaagacttttgcctgtcctttgtcaggcgcatggcctgacaccagagcaagt
agtggctattgcaagtaacatcggtggcaaacaagcgctggagagcatcagcatcgtggcccagctgtctcggcccgaccctgccctcgccgctctgaccaacgaccac
ctggtggccctggcttgcctcgggggcaggccagctcttgacgccgtgaagaagggccttcctcacgccccagccctgatcaagcggaccaacagaaggattcccgag
aggacatcacatcgagtggcagatcacgcgcaagtggtccgcgtgctcggattcttccagtgtcactcccaccccgcacaagcgttcgatgacgccatgactcaatttggt
atgtcgagacacggactgctgcagctctttcgtagagtcggtgtcacagaactcgaggcccgctcgggcacactgcctcccgcctcccagcggtgggacaggattctcca
agcgagcggtatgaaacgcgcgaagccttcacctacgtcaactcagacacctgaccaggcgagccttcatgcgttcgcagactcgctggagagggatttggacgcgcc
ctcgcccatgcatgaaggggaccaaactcgcgcgtcagctagccccaagaagaagagaaaggtggaggccagcggttcc**ggacgggctgacgcattggacgatt**
ttgatctggatatgctgggaagtgacgccctcgatgattttgaccttgacatgcttggttcggatgcccttgatgactttgacctcgacatgctcggcagtgacgc
ccttgatgatttcgacctggacatgctgattaactctagaggcagtggagagggcagaggaagtctgctaacatgcggtgacgtcgaggagaatcctggcccagtga
gcaagggcgaggaggataacatggccatcatcaaggagttcatgcgcttcaaggtgcacatggagggctccgtgaacggccacgagttcgagatcgagggcgaggg
cgagggccgcccctacgagggcacccagaccgccaagctgaaggtgaccaagggtggccccctgcccttcgcctgggacatcctgtcccctcagttcatgtacggctcc
aaggcctacgtgaagcaccccgccgacatccccgactacttgaagctgtccttccccgagggcttcaagtgggagcgcgtgatgaacttcgaggacggcggcgtggtga
ccgtgacccaggactcctccctgcaggacggcgagttcatctacaaggtgaagctgcgcggcaccaacttcccctccgacggccccgtaatgcagaagaagaccatgg
gctgggaggcctcctccgagcggatgtaccccgaggacggcgccctgaagggcgagatcaagcagaggctgaagctgaaggacggcggccactacgacgctgag
gtcaagaccacctacaaggccaagaagcccgtgcagctgcccggcgcctacaacgtcaacatcaagttggacatcacctcccacaacgaggactacaccatcgtgga
acagtacgaacgcgccgagggccgccactccaccggcggcatggacgagctgtacaagtaa

TALE-(XX')$_3$ amino acid sequence:

MRTRLPSPPAPSPAFSAGSFSDLLRQFDPSLFNTSLFDSLPPFGAHHTEAATGEWDEVQSGLRAADAPPPTMRVAVTAA
RPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVA
VKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPL
NLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPEQVVAIAS
NNGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETV
QRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX'GGKQALETVQRLLPVLCQAH
GLTPEQVVAIASXX'GGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX'GGKQALETVQRLLPVLCQAHGLTPEQVVAIAS
NGGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQ
RLLPVLCQAHGLTPEQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGL
TPEQVVAIASNIGGKQALESISIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHR
VADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGMSRHGLLQLFRRVGVTELEARSGTLPPASQRWDRILQASGMKRAK
PSPTSTQTPDQASLHAFADSLERDLDAPSPMHEGDQTRASASPKKKRKVEASGS**GRADALDDFDLDMLGSDALDDFDL**
DMLGSDALDDFDLDMLGSDALDDFDLDMLINSRGSGEGRGSLLTCGDVEENPGPVSKGEEDNMAIIKEFMRFKVHMEG
SVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDILSPQFMYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNF
EDGGVVTVTQDSSLQDGEFIYKVKLRGTNFPSDGPVMQKKTMGWEASSERMYPEDGALKGEIKQRLKLKDGGHYDAEV
KTTYKAKKPVQLPGAYNVNIKLDITSHNEDYTIVEQYERAEGRHSTGGMDELYK*

TALE-$(XX')_6$ DNA sequence:

atgaggaccaggctgccatctcccctgccccttcccccgcatttagcgctgggagctttagcgacctgcttaggcagttcgaccccagcttgttcaacaccagcctgtttga
cagcctgcctcccttcggagcgcaccacaccgaagccgccaccggcgagtgggacgaggtgcaaagcggcctgagggcagcggacgctcctccgccaaccatgag
ggtggcagtgacagcagctaggccccctcgggcaaaacctgcacccaggagaagggctgcccaacccagcgacgcgagtccagccgcacaggtggacctcagga
cgctgggctacagccagcaacagcaagagaagatcaagcccaaagtaaggagcaccgtggcccagcaccacgaggccctggtgggtcacggcttcacccacgcg
catatcgttgctctgagccaacatcccgcagctctgggtaccgttgcggtgaagtatcaggacatgatcgcggcactgcctgaagctacacacgaagccatagtgggcgtt
ggcaagcagtggagcggtgccagagcgcttgaggcactgttgacggtggctggcgagctgaggggaccgccactgcaactggacaccggccaactgctgaagatcg
ccaagaggggaggcgtgacggcggtggaggccgtgcatgcctggaggaatgccctgaccggcgcgcccctgaaccttacgcctgagcaagtcgttgcgatcgcctcc
cacgacggcggaaaacaggctttggaaaccgtgcagcggttgctgcccgttttgtgccaagcccacggactcactccggaacaggtggtcgcaatcgcgagcaacggc
ggcggcaagcaagcccttgagacagtccaaagacttttgcctgtcctttgtcaggcgcatggcctcactccggaacaggtggtcgcaatcgcgagcaacaacggcggca
agcaagcccttgagacagtccaaagacttttgcctgtcctttgtcaggcgcatggcttgacccccgaacaggttgtagccatagcttctaacaacggaggtaagcaggcac
tggaaaccgtgcagcgcctgctcccagtactgtgtcaggctcatgggctgacaccagagcaagtagtggctattgcaagtcacgacggtggcaaacaagcgctggaga
ccgtgcagaggctccttccggtgctctgccaagcacacggtcttacgcctgagcaagtcgttgcgatcgcctcccacgacggcggaaaacaggctttggaaaccgtgca
gcggttgctgcccgttttgtgccaagcccacggattgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcc
tgctcccagtactgtgtcaggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcc
cagtactgtgtcaggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcccagta
ctgtgtcaggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcccagtactgtgt
caggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcccagtactgtgtcaggc
tcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcccagtactgtgtcaggctcatgg
gctcactccggaacaggtggtcgcaatcgcgagcaacggcggcggcaagcaagcccttgagacagtccaaagacttttgcctgtcctttgtcaggcgcatggcctgaca
ccagagcaagtagtggctattgcaagtaacatcggtggcaaacaagcgctggagaccgtgcagaggctccttccggtgctctgccaagcacacggtcttacgcctgagc
aagtcgttgcgatcgcctcccacgacggcggaaaacaggctttggaaaccgtgcagcggttgctgcccgttttgtgccaagcccacggactcactccggaacaggtggtc
gcaatcgcgagcaacaacggcggcaagcaagcccttgagacagtccaaagacttttgcctgtcctttgtcaggcgcatggcctcactccggaacaggtggtcgcaatcg
cgagcaacggcggcggcaagcaagcccttgagacagtccaaagacttttgcctgtcctttgtcaggcgcatggcctgacaccagagcaagtagtggctattgcaagtaa
catcggtggcaaacaagcgctggagagcatcagcatcgtggcccagctgtctcggcccgaccctgccctcgccgctctgaccaacgaccacctggtggccctggcttgc
ctcgggggcaggccagctcttgacgccgtgaagaagggccttcctcacgccccagccctgatcaagcggaccaacagaaggattcccgagaggacatcacatcgagt
ggcagatcacgcgcaagtggtccgcgtgctcggattcttccagtgtcactcccaccccgcacaagcgttcgatgacgccatgactcaatttggtatgtcgagacacggact
gctgcagctctttcgtagagtcggtgtcacagaactcgaggcccgctcgggcacactgcctcccgcctcccagcggtgggacaggattctccaagcgagcggtatgaaa
cgcgcgaagccttcacctacgtcaactcagacacctgaccaggcgagccttcatgcgttcgcagactcgctggagagggatttggacgcgccctcgcccatgcatgaag
gggaccaaactcgcgcgtcagctagccccaagaagaagagaaaggtggaggccagcggttcc**ggacgggctgacgcattggacgattttgatctggatatgctg**
ggaagtgacgccctcgatgattttgaccttgacatgcttggttcggatgcccttgatgactttgacctcgacatgctcggcagtgacgcccttgatgatttcgac
ctggacatgctgattaactctagaggcagtggagagggcagaggaagtctgctaacatgcggtgacgtcgaggagaatcctggcccagtgagcaagggcgaggag
gataacatggccatcatcaaggagttcatgcgcttcaaggtgcacatggagggctccgtgaacggccacgagttcgagatcgagggcgagggcgagggccgcccctа
cgagggcacccagaccgccaagctgaaggtgaccaagggtggccccctgcccttcgcctgggacatcctgtcccctcagttcatgtacggctccaaggcctacgtgaag
caccccgccgacatccccgactacttgaagctgtccttccccgagggcttcaagtgggagcgcgtgatgaacttcgaggacggcggcgtggtgaccgtgacccaggact
cctccctgcaggacggcgagttcatctacaaggtgaagctgcgcggcaccaacttcccctccgacggccccgtaatgcagaagaagaccatgggctgggaggcctcct
ccgagcggatgtaccccgaggacggcgccctgaagggcgagatcaagcagaggctgaagctgaaggacggcggccactacgacgctgaggtcaagaccacctac
aaggccaagaagcccgtgcagctgcccggcgcctacaacgtcaacatcaagttggacatcacctcccacaacgaggactacaccatcgtggaacagtacgaacgcg
ccgagggccgccactccaccggcggcatggacgagctgtacaagtaa

TALE-$(XX')_6$ amino acid sequence:

MRTRLPSPPAPSPAFSAGSFSDLLRQFDPSLFNTSLFDSLPPFGAHHTEAATGEWDEVQSGLRAADAPPPTMRVAVTAA
RPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVA
VKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPL
NLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPEQVVAIAS
NNGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETV
QRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX’GGKQALETVQRLLPVLCQAH
GLTPEQVVAIASXX’GGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX’GGKQALETVQRLLPVLCQAHGLTPEQVVAIAS
XX’GGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX’GGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX’GGKQALETV
QRLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPEQVVAIASN
GGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGGKQALESISIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVK
KGLPHAPALIKRTNRRIPERTSHRVADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGMSRHGLLQLFRRVGVTELEARSG
TLPPASQRWDRILQASGMKRAKPSPTSTQTPDQASLHAFADSLERDLDAPSPMHEGDQTRASASPKKKRKVEASGS**GR**
ADALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLINSRGSGEGRGSLLTCGDVEENPGPV
SKGEEDNMAIIKEFMRFKVHMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDILSPQFMYGSKAYVKHP
ADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRGTNFPSDGPVMQKKTMGWEASSERMYPED
GALKGEIKQRLKLKDGGHYDAEVKTTYKAKKPVQLPGAYNVNIKLDITSHNEDYTIVEQYERAEGRHSTGGMDELYK*

TALE-(XX')$_{12}$ DNA sequence:

atgaggaccaggctgccatctcccctgccccttcccccgcatttagcgctgggagctttagcgacctgcttaggcagttcgaccccagcttgttcaacaccagcctgtttga
cagcctgcctcccttcggagcgcaccacaccgaagccgccaccggcgagtgggacgaggtgcaaagcggcctgagggcagcggacgctcctccgccaaccatgag
ggtggcagtgacagcagctaggccccctcgggcaaaacctgcacccaggagaagggctgcccaacccagcgacgcgagtccagccgcacaggtggacctcagga
cgctgggctacagccagcaacagcaagagaagatcaagcccaaagtaaggagcaccgtggcccagcaccacgaggccctggtgggtcacggcttcacccacgcg
catatcgttgctctgagccaacatcccgcagctctgggtaccgttgcggtgaagtatcaggacatgatcgcggcactgcctgaagctacacacgaagccatagtgggcgtt
ggcaagcagtggagcggtgccagagcgcttgaggcactgttgacggtggctggcgagctgaggggaccgccactgcaactggacaccggccaactgctgaagatcg
ccaagaggggaggcgtgacggcggtggaggccgtgcatgcctggaggaatgccctgaccggcgcgcccctgaacctcactccggaacaggtggtcgcaatcgcga
gcaacatcggcggcaagcaagcccttgagacagtccaaagacttttgcctgtcctttgtcaggcgcatggccttacgcctgagcaagtcgttgcgatcgcctcccacgacg
gcggaaaacaggctttggaaaccgtgcagcggttgctgcccgttttgtgccaagcccacggattgacccccgaacaggttgtagccatagcttctaacggcggaggtaag
caggcactggaaaccgtgcagcgcctgctcccagtactgtgtcaggctcatgggctgacaccagagcaagtagtggctattgcaagtcacgacggtggcaaacaagcg
ctggagaccgtgcagaggctccttccggtgctctgccaagcacacggtttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactgga
aaccgtgcagcgcctgctcccagtactgtgtcaggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccg
tgcagcgcctgctcccagtactgtgtcaggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcag
cgcctgctcccagtactgtgtcaggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcct
gctcccagtactgtgtcaggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctccc
agtactgtgtcaggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcccagtact
gtgtcaggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcccagtactgtgtca
ggctcatgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcccagtactgtgtcaggctc
atgggttgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcccagtactgtgtcaggctcatgggt
tgacccccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcccagtactgtgtcaggctcatgggttgacc
cccgaacaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaaccgtgcagcgcctgctcccagtactgtgtcaggctcatgggttgacccccga
acaggttgtagccatagcttctNNNNNNggaggtaagcaggcactggaaagcatcgtggcccagctgtctcggcccgaccctgccctcgccgctctgaccaacgac
cacctggtggccctggcttgcctcgggggcaggccagctcttgacgccgtgaagaagggccttcctcacgccccagccctgatcaagcggaccaacagaaggattccc
gagaggacatcacatcgagtggcagatcacgcgcaagtggtccgcgtgctcggattcttccagtgtcactcccaccccgcacaagcgttcgatgacgccatgactcaattt
ggtatgtcgagacacggactgctgcagctctttcgtagagtcggtgtcacagaactcgaggcccgctcgggcacactgcctcccgcctcccagcggtgggacaggattct
ccaagcgagcggtatgaaacgcgcgaagccttcacctacgtcaactcagacacctgaccaggcgagccttcatgcgttcgcagactcgctggagagggatttggacgc
gccctcgcccatgcatgaaggggaccaaactcgcgcgtcagctagccccaagaagaagagaaaggtggaggccagcggttcc**ggacgggctgacgcattggac
gattttgatctggatatgctgggaagtgacgccctcgatgattttgaccttgacatgcttggttcggatgcccttgatgactttgacctcgacatgctcggcagtg
acgcccttgatgatttcgacctggacatgctgattaa**ctctagaggcagtggagagggcagaggaagtctgctaacatgcggtgacgtcgaggagaatcctggccc
agtgagcaagggcgaggaggataacatggccatcatcaaggagttcatgcgcttcaaggtgcacatggagggctccgtgaacggccacgagttcgagatcgagggcg
agggcgagggccgcccctacgagggcacccagaccgccaagctgaaggtgaccaagggtggccccctgcccttcgcctgggacatcctgtcccctcagttcatgtacg
gctccaaggcctacgtgaagcaccccgccgacatccccgactacttgaagctgtccttccccgagggcttcaagtgggagcgcgtgatgaacttcgaggacggcggcgt
ggtgaccgtgacccaggactcctccctgcaggacggcgagttcatctacaaggtgaagctgcgcggcaccaacttcccctccgacggccccgtaatgcagaagaaga
ccatgggctgggaggcctcctccgagcggatgtaccccgaggacggcgccctgaagggcgagatcaagcagaggctgaagctgaaggacggcggccactacgac
gctgaggtcaagaccacctacaaggccaagaagcccgtgcagctgcccggcgcctacaacgtcaacatcaagttggacatcacctcccacaacgaggactacaccat
cgtggaacagtacgaacgcgccgagggccgccactccaccggcggcatggacgagctgtacaagtaa

TALE-(XX')$_{12}$ amino acid sequence:

MRTRLPSPPAPSPAFSAGSFSDLLRQFDPSLFNTSLFDSLPPFGAHHTEAATGEWDEVQSGLRAADAPPPTMRVAVTAA
RPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVA
VKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPL
NLTPEQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASN
GGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX'GGKQALETVQ
RLLPVLCQAHGLTPEQVVAIASXX'GGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX'GGKQALETVQRLLPVLCQAHG
LTPEQVVAIASXX'GGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX'GGKQALETVQRLLPVLCQAHGLTPEQVVAIASX
X'GGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX'GGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX'GGKQALETVQ
RLLPVLCQAHGLTPEQVVAIASXX'GGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX'GGKQALETVQRLLPVLCQAHG
LTPEQVVAIASXX'GGKQALETVQRLLPVLCQAHGLTPEQVVAIASXX'GGKQALESIVAQLSRPDPALAALTNDHLVALACL
GGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGMSRHGLLQLFRRV
GVTELEARSGTLPPASQRWDRILQASGMKRAKPSPTSTQTPDQASLHAFADSLERDLDAPSPMHEGDQTRASASPKKK
RKVEASGS**GRADALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLIN**SRGSGEGRGSLLTC
GDVEENPGPVSKGEEDNMAIIKEFMRFKVHMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDILSPQFM
YGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRGTNFPSDGPVMQKKTMGWE
ASSERMYPEDGALKGEIKQRLKLKDGGHYDAEVKTTYKAKKPVQLPGAYNVNIKLDITSHNEDYTIVEQYERAEGRHSTG
GMDELYK*

Notes:

RVD-coding sequences and RVDs (XX') are highlighted. Underlined NNNNNN and XX' indicate the DNA and amino acid sequences of the variable RVDs, respectively. The sequences in bold and in red indicate VP64 and mCherry, respectively.

Supplementary information, Table S2. Sequences of amino acid diresidues in RVD libraries (TALE-(XX’)$_3$, TALE-(XX’)$_6$ and TALE-(XX’)$_{12}$).

Amino Acid (XX’)	DNA (NNNNNN)	Amino Acid (XX’)	DNA (NNNNNN)	Amino Acid (XX’)	DNA (NNNNNN)	Amino Acid (XX’)	DNA (NNNNNN)	Amino Acid (XX’)	DNA (NNNNNN)
AA	gctgcg	CA	tgcgcc	DA	gacgct	EA	gaggcg	FA	tttgcc
AC	gcctgc	CC	tgctgc	DC	gattgc	EC	gaatgt	FC	ttctgc
AD	gccgac	CD	tgcgac	DD	gacgac	ED	gaagac	FD	ttcgac
AE	gccgaa	CE	tgcgag	DE	gacgaa	EE	gaggag	FE	ttcgag
AF	gccttc	CF	tgcttt	DF	gatttc	EF	gagttc	FF	ttcttc
AG	gctggt	CG	tgcggt	DG	gatggc	EG	gagggc	FG	ttcgga
AH	gcccac	CH	tgccat	DH	gatcac	EH	gagcac	FH	tttcat
AI	gccatc	CI	tgtatc	DI	gacatt	EI	gaaatc	FI	tttatt
AK	gccaag	CK	tgcaag	DK	gacaag	EK	gaaaag	FK	ttcaag
AL	gcccta	CL	tgcttg	DL	gacctc	EL	gaactc	FL	ttcttg
AM	gccatg	CM	tgtatg	DM	gatatg	EM	gaaatg	FM	tttatg
AN	gccaac	CN	tgcaac	DN	gacaac	EN	gagaac	FN	ttcaat
AP	gccccc	CP	tgccca	DP	gacccc	EP	gagcct	FP	ttcccc
AQ	gcccaa	CQ	tgccag	DQ	gaccaa	EQ	gagcag	FQ	ttccaa
AR	gctcgt	CR	tgccgg	DR	gatcgt	ER	gaacga	FR	ttccgg
AS	gcatcg	CS	tgctcg	DS	gactct	ES	gaatcc	FS	ttctct
AT	gccacc	CT	tgcact	DT	gacact	ET	gaaacc	FT	ttcacc
AV	gctgtc	CV	tgtgtc	DV	gacgtt	EV	gaggtc	FV	ttcgtg
AW	gcgtgg	CW	tgttgg	DW	gactgg	EW	gaatgg	FW	ttctgg
AY	gcgtat	CY	tgctat	DY	gactat	EY	gaatac	FY	ttctat
GA	ggggca	HA	cacgcc	IA	atcgcg	KA	aaggcc	LA	ttggct
GC	ggctgc	HC	cactgc	IC	atctgt	KC	aagtgc	LC	ctatgc
GD	ggcgac	HD	cacgac	ID	atcgac	KD	aaggat	LD	ctcgac
GE	ggcgag	HE	cacgag	IE	atcgaa	KE	aaggaa	LE	ttggag
GF	gggttc	HF	cacttc	IF	atcttt	KF	aagttt	LF	ctgttc
GG	ggtggc	HG	cacggc	IG	atcggc	KG	aagggc	LG	ctcggc
GH	ggccac	HH	caccac	IH	atccat	KH	aaacat	LH	cttcac
GI	ggcatc	HI	cacatc	II	attatc	KI	aagatc	LI	ctcatt
GK	ggcaag	HK	cacaag	IK	attaag	KK	aaaaag	LK	ctcaaa
GL	ggcctc	HL	cacctg	IL	atcttg	KL	aagctc	LL	cttctc
GM	ggtatg	HM	cacatg	IM	atcatg	KM	aagatg	LM	ctcatg
GN	ggcaac	HN	cacaac	IN	atcaac	KN	aagaac	LN	ctcaat
GP	ggtccc	HP	cacccc	IP	atcccc	KP	aagcct	LP	cttccc
GQ	ggccag	HQ	caccag	IQ	atccag	KQ	aagcag	LQ	ctccag
GR	ggtcgc	HR	cacaga	IR	atccgt	KR	aaacgc	LR	ttgcgc
GS	ggctcc	HS	cacagc	IS	atcagc	KS	aagtct	LS	ctctct
GT	ggcacc	HT	cacacc	IT	ataacc	KT	aagacc	LT	ttgact
GV	ggcgtc	HV	cacgtg	IV	atcgtc	KV	aaggtt	LV	ctggtt
GW	ggttgg	HW	cactgg	IW	atctgg	KW	aagtgg	LW	atttgg
GY	ggctac	HY	cactac	IY	atctat	KY	aagtac	LY	ctctac

Amino Acid (XX')	DNA (NNNNNN)	Amino Acid (XX')	DNA (NNNNNN)	Amino Acid (XX')	DNA (NNNNNN)	Amino Acid (XX')	DNA (NNNNNN)	Amino Acid (XX')	DNA (NNNNNN)
MA	atggcg	NA	aacgcc	PA	ccagcc	QA	caggct	RA	agagct
MC	atgtgc	NC	aactgc	PC	ccctgt	QC	cagtgc	RC	cgctgc
MD	atggat	ND	aacgac	PD	cctgac	QD	caggac	RD	agggac
ME	atggaa	NE	aacgag	PE	ccagaa	QE	caggag	RE	agagag
MF	atgttt	NF	aacttc	PF	ccattt	QF	caattt	RF	aggttt
MG	atgggc	NG	aacggc	PG	ccgggg	QG	cagggt	RG	cgcggc
MH	atgcac	NH	aaccac	PH	ccgcat	QH	caacat	RH	agacac
MI	atgata	NI	aacatc	PI	ccaatc	QI	cagatc	RI	agaatc
MK	atgaag	NK	aacaag	PK	cctaag	QK	cagaag	RK	cgcaag
ML	atgtta	NL	aacctg	PL	ccgctc	QL	caactt	RL	cgcctc
MM	atgatg	NM	aacatg	PM	cccatg	QM	cagatg	RM	agaatg
MN	atgaac	NN	aacaac	PN	cctaac	QN	cagaac	RN	cgtaac
MP	atgccc	NP	aacccc	PP	cccccc	QP	caacct	RP	cgcccg
MQ	atgcag	NQ	aaccag	PQ	ccgcag	QQ	cagcaa	RQ	agacag
MR	atgcgg	NR	aacaga	PR	cctcgt	QR	caacgt	RR	cggcga
MS	atgtcc	NS	aacagc	PS	ccttcg	QS	caatcc	RS	cgctcc
MT	atgact	NT	aacacc	PT	ccgacg	QT	cagact	RT	cgcacc
MV	atggtc	NV	aacgtg	PV	cctgtg	QV	caggtg	RV	cgggtt
MW	atgtgg	NW	aactgg	PW	ccgtgg	QW	cagtgg	RW	cgctgg
MY	atgtat	NY	aactac	PY	ccatac	QY	cagtat	RY	agatac
SA	tccgcc	TA	actgcc	VA	gtcgcc	WA	tgggcc	YA	tacgcg
SC	tcctgc	TC	acctgc	VC	gtatgc	WC	tggtgc	YC	tactgt
SD	tcggat	TD	accgac	VD	gtggac	WD	tgggat	YD	tacgat
SE	tccgaa	TE	accgag	VE	gttgag	WE	tgggaa	YE	tatgaa
SF	agcttt	TF	acgttc	VF	gtgttc	WF	tggttt	YF	tacttc
SG	tcgggg	TG	actggc	VG	gtcgga	WG	tggggc	YG	tatggc
SH	tcacac	TH	acccac	VH	gtgcat	WH	tggcat	YH	tatcac
SI	agcatc	TI	accatc	VI	gtcatt	WI	tggatc	YI	tacatt
SK	tccaag	TK	accaag	VK	gtcaag	WK	tggaag	YK	tacaag
SL	tccctc	TL	actctc	VL	gtcttg	WL	tggttg	YL	tactta
SM	agcatg	TM	accatg	VM	gtcatg	WM	tggatg	YM	tatatg
SN	tccaac	TN	accaac	VN	gtgaac	WN	tggaac	YN	tacaac
SP	tccccg	TP	acacca	VP	gtccct	WP	tggccg	YP	tatccg
SQ	tcgcaa	TQ	acccaa	VQ	gtgcag	WQ	tggcag	YQ	taccag
SR	tcgagg	TR	acccgc	VR	gtgcgg	WR	tggcgc	YR	tatcga
SS	tctagc	TS	acatcc	VS	gtgtcc	WS	tggtcg	YS	tattcc
ST	tctacc	TT	actacg	VT	gtgacc	WT	tggact	YT	tatact
SV	tccgtg	TV	acggtc	VV	gttgtt	WV	tgggtt	YV	tacgtt
SW	tcctgg	TW	acttgg	VW	gtgtgg	WW	tggtgg	YW	tactgg
SY	agttac	TY	acttac	VY	gtctac	WY	tggtat	YY	tactat

Supplementary information, Table S3. Base binding specificity of previously reported RVDs obtained from TALE-(XX')$_3$ library screening.

Category	RVDs	A	T	C	G
NX'	NA	-	++	+	+
	NC	+	-	+	-
	ND	-	-	+	-
	NG	-	+	-	-
	NH	-	-	-	++++
	NI	++	-	-	-
	NK	-	-	-	+
	NN	++	-	-	++++
	NP	-	-	-	-
	NQ	-	-	+	-
	NS	++++	-	+	++++
	NT	++++	-	+	++++
	NV	++	-	+	+
HX'	HA	-	+	+	-
	HD	-	-	++++	-
	HG	-	-	-	-
	HH	-	-	-	++++
	HI	-	-	-	-
	HN	++	-	-	++++
IX'	IG	-	-	-	-
	IS	-	-	-	-
SX'	SH	-	-	-	-
	SN	-	+	-	+
	SS	+	+	-	+
YX'	YG	-	-	-	-

Note:

The ranges of fold induction of EGFP reporter for RVDs in TALE-(XX')$_3$ are indicated as follows:

-       < 6
+   6 - 12
++   12 - 18
+++   18 - 24
++++ $\geq$ 24

Supplementary information, Table S4. Efficiencies of TALENs-mediated indels with novel RVDs.

Targeted Gene	Type of TALENs	TALEN[L]	TALEN[R]	Mean Indels (%) ± s.d. (n = 3)
LRP1[a]	$TALENs_{NN}$	NI NI NN NI HD NG NG NN HD NI NN HD HD HD HD	NI HD NI NN NN NG NG NI NG NG NG NN NI NG HD	77.1 ± 4.3
	$TALENs_{NH}$	NI NI NH NI HD NG NG NH HD NI NH HD HD HD HD	NI HD NI NH NH NG NG NI NG NG NG NH NI NG HD	50.7 ± 4.9
	$TALENs_{KN}$	NI NI KN NI HD NG NG KN HD NI KN HD HD HD HD	NI HD NI KN KN NG NG NI NG NG NG KN NI NG HD	78.1 ± 3.4
DKK1[b]	$TALENs_{NN}$	NG HD HD NI NI HD NN HD NG NI NG HD NI NI NN	NN HD HD HD HD NN HD NI NN HD NN HD HD NN HD	54.7 ± 4.8
	$TALENs_{NH}$	NG HD HD NI NI HD NH HD NG NI NG HD NI NI NH	NH HD HD HD HD NH HD NI NH HD NH HD HD NH HD	39.4 ± 7.8
	$TALENs_{KN}$	NG HD HD NI NI HD KN HD NG NI NG HD NI NI KN	KN HD HD HD HD KN HD NI KN HD KN HD HD KN HD	49.9 ± 2.7
LRP1[a]	$TALENs_{NG}$	NI NI NN NI HD NG NG NN HD NI NN HD HD HD HD	NI HD NI NN NN NG NG NI NG NG NG NN NI NG HD	77.1 ± 4.3
	$TALENs_{RG}$	NI NI NN NI HD RG RG NN HD NI NN HD HD HD HD	NI HD NI NN NN RG RG NI RG RG RG NN NI RG HD	40.2 ± 10.3
ATG5[c]	$TALENs_{NG}$	NN NG NG NG HD NI HD NN HD NG NI NG NI NG HD	NI NG NN NN NG NG HD NG NN HD NG NG HD HD HD	11.5 ± 0.6
	$TALENs_{RG}$	NN RG RG RG HD NI HD NN HD RG NI RG NI RG HD	NI RG NN NN RG RG HD RG NN HD RG RG HD HD HD	6.5 ± 0.5

[a] Sequence for TALENs targeting is 5'-TAAGACTTGCAGCCCCAAGCAGTTTGCCTGCAGAGATCAAATAACCTGTA-3', and TALEN[L] and TALEN[R] targeted regions are labeled in grey shades. Sequences of primers used to amplify the TALENs targeting region for the assessment of indels are: 5'-AGCCTAGAGGACTTGGAGGG-3' and 5'-ACCTGGGGCCAATCATCAAG-3'.

[b] Sequence for TALENs targeting is 5'-TTCCAACGCTATCAAGAACCTGCCCCCACCGCTGGGCGGCGCTGCGGGGCA-3', and TALEN[L] and TALEN[R] targeted regions are labeled in grey shades. Sequences of primers used to amplify the TALENs targeting region for the assessment of indels are: 5'-GACCCAGGCTTGCAAAGTGA-3' and 5'-CAGGCGAGACAGATTTGCAC-3'.

[c] Sequence for TALENs targeting is 5'-TGTTTCACGCTATATCAGGATGAGATAACTGAAAGGGAAGCAGAACCATA-3', and TALEN[L] and TALEN[R] targeted regions are labeled in grey shades. Sequences of primers used to amplify the TALENs targeting region for the assessment of indels are: 5'-TAGCAGGTTCATTCATCGTTGCAAGGA-3' and 5'-ATGTAAGGAAAACAAAGTCCAGAACGC-3'.

Supplementary information, Table S5.

RVDs targeting multiple bases.

RVDs targeting two bases

Category	RVDs	A	T	C	G
A/T	SG	+	+	-	-
A/C	HC	++	-	++	-
	KC	++	-	++	-
	NC	+	-	+	-
	KS	++	-	+	-
A/G	HN	++	-	-	++++
	NN	++	-	-	++++
T/C	HA	-	+	+	-
	KA	-	++	+	-
	KP	-	++	+	-
	FS	-	+	+	-
T/G	KF	-	+	-	+
	SN	-	+	-	+
	RQ	-	+	-	++++
	GR	-	+	-	+
	KR	-	+	-	+++
C/G	HQ	-	-	+	+++
	KQ	-	-	+	++
	QR	-	-	+	++
	YR	-	-	+	++++

RVDs targeting three bases

Category	RVDs	A	T	C	G
A/T/C	RL	+	++	+	-
	MS	+	+	+	-
	TT	+	+	+	-
A/T/G	TC	++	+	-	+
	MH	+	+	-	+
	LR	+	+	-	++
	SS	+	+	-	+
A/C/G	RC	+	-	++	+
	CR	+	-	+	+++
	HS	++	-	++	++++
	NS	++++	-	+	++++
	HT	++	-	+	+
	KT	+	-	+	++
	NT	++++	-	+	++++
	HV	++	-	+	+
	KV	++++	-	+++	+++
	NV	++	-	+	+
T/C/G	FQ	-	+	+	+
	NA	-	++	+	+
	RN	-	+	+	++++

RVDs targeting four bases

Category	RVDs	A	T	C	G
A/T/C/G	RH	+	+	+	+++
	LQ	+	++	+	+
	VR	+	+	+	++
	CS	+	+	+	+
	RT	+	+	+	++
	RV	+++	++	+++	++

Note:

The ranges of fold induction of EGFP reporter for RVDs in TALE-(XX')$_3$ are indicated as follows:

-	< 6
+	6 - 12
++	12 - 18
+++	18 - 24
++++	$\geq$ 24

SUPPLEMENTARY REFERENCES

1. Zhang, F. et al. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **29**, 149-153 (2011).
2. Yang, J. et al. ULtiMATE system for rapid assembly of customized TAL effectors. *PLoS One* **8**, e75649 (2013).
3. Boussif, O. et al. A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo: polyethylenimine. *Proc Natl Acad Sci U S A* **92**, 7297-7301 (1995).
4. Mussolino, C. et al. A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic acids research* (2011).

Supplementary information, Data S1 Sequences and material and methods

TALE-Ctrl

```
   1  GTCGACGGATCGGGAGATCTCCCGATCCCCTATGGTGCACTCTCAGTACAATCTGCTCTG
  61  ATGCCGCATAGTTAAGCCAGTATCTGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGT
 121  GCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACCGACAATTGCATGAAGAATC
 181  TGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTTGAC
 241  ATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCAT
 301  ATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACG
 361  ACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTT
 421  TCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAG
 481  TGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGC
 541  ATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAG
 601  TCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGT
 661  TTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGC
 721  ACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGG
 781  GCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGCGCGTTTTGCCTGTACTGGGTCT
 841  CTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTT
 901  AAGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGTGTGAC
 961  TCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGTGGC
1021  GCCCGAACAGGGACTTGAAAGCGAAAGGGAAACCAGAGGAGCTCTCTCGACGCAGGACTC
1081  GGCTTGCTGAAGCGCGCACGGCAAGAGGCGAGGGGCGGCGACTGGTGAGTACGCCAAAAA
1141  TTTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCGAGAGCGTCAGTATTAAGCGGG
1201  GGAGAATTAGATCGCGATGGGAAAAAATTCGGTTAAGGCCAGGGGGAAAGAAAAAATATA
1261  AATTAAAACATATAGTATGGGCAAGCAGGGAGCTAGAACGATTCGCAGTTAATCCTGGCC
1321  TGTTAGAAACATCAGAAGGCTGTAGACAAATACTGGGACAGCTACAACCATCCCTTCAGA
1381  CAGGATCAGAAGAACTTAGATCATTATATAATACAGTAGCAACCCTCTATTGTGTGCATC
1441  AAAGGATAGAGATAAAAGACACCAAGGAAGCTTTAGACAAGATAGAGGAAGAGCAAAACA
1501  AAAGTAAGACCACCGCACAGCAAGCGGCCGGCCGCGCTGATCTTCAGACCTGGAGGAGGA
1561  GATATGAGGGACAATTGGAGAAGTGAATTATATAAATATAAAGTAGTAAAAATTGAACCA
1621  TTAGGAGTAGCACCCACCAAGGCAAAGAGAAGAGTGGTGCAGAGAGAAAAAAGAGCAGTG
1681  GGAATAGGAGCTTTGTTCCTTGGGTTCTTGGGAGCAGCAGGAAGCACTATGGGCGCAGCG
1741  TCAATGACGCTGACGGTACAGGCCAGACAATTATTGTCTGGTATAGTGCAGCAGCAGAAC
1801  AATTTGCTGAGGGCTATTGAGGCGCAACAGCATCTGTTGCAACTCACAGTCTGGGGCATC
1861  AAGCAGCTCCAGGCAAGAATCCTGGCTGTGGAAAGATACCTAAAGGATCAACAGCTCCTG
1921  GGGATTTGGGGTTGCTCTGGAAAACTCATTTGCACCACTGCTGTGCCTTGGAATGCTAGT
1981  TGGAGTAATAAATCTCTGGAACAGATTTGGAATCACACGACCTGGATGGAGTGGGACAGA
2041  GAAATTAACAATTACACAAGCTTAATACACTCCTTAATTGAAGAATCGCAAAACCAGCAA
2101  GAAAAGAATGAACAAGAATTATTGGAATTAGATAAATGGGCAAGTTTGTGGAATTGGTTT
2161  AACATAACAAATTGGCTGTGGTATATAAAATTATTCATAATGATAGTAGGAGGCTTGGTA
2221  GGTTTAAGAATAGTTTTTGCTGTACTTTCTATAGTGAATAGAGTTAGGCAGGGATATTCA
2281  CCATTATCGTTTCAGACCCACCTCCCAACCCCGAGGGGACCCGACAGGCCCGAAGGAATA
2341  GAAGAAGAAGGTGGAGAGAGAGACAGAGACAGATCCATTCGATTAGTGAACGGATCGGCA
2401  CTGCGTGCGCCAATTCTGCAGACAAATGGCAGTATTCATCCACAATTTTAAAAGAAAAGG
2461  GGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAGACATACA
2521  AACTAAAGAATTACAAAAACAAATTACAAAAATTCAAAATTTTCGGGTTTATTACAGGGA
2581  CAGCAGAGATCCAGTTTGGTTAGTACCGGGCCCGCTCTAGACGATGTACGGGCCAGATAT
2641  ACGCGTTGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTT
2701  CATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGA
2761  CCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCA
2821  ATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCA
2881  GTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGG
2941  CCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATC
3001  TACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGT
3061  GGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGT
3121  TTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTG
3181  ACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTCTCTGGCTA
3241  ACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACC
3301  CAAGCTGGCTAGCGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCATGAGGACCAGGC
3361  TGCCATCTCCCCCTGCCCCTTCCCCCGCATTTAGCGCTGGAGCTTTAGCGACCTGCTTA
3421  GGCAGTTCGACCCCAGCTTGTTCAACACCAGCCTGTTTGACAGCCTGCCTCCCTTCGGAG
```

3481 CGCACCACACCGAAGCCGCCACCGGCGAGTGGGACGAGGTGCAAAGCGGCCTGAGGGCAG
3541 CGGACGCTCCTCCGCCAACCATGAGGGTGGCAGTGACAGCAGCTAGGCCCCCTCGGGCAA
3601 AACCTGCACCCAGGAGAAGGGCTGCCCAACCCAGCGACGCGAGTCCAGCCGCACAGGTGG
3661 ACCTCAGGACGCTGGGCTACAGCCAGCAACAGCAAGAGAAGATCAAGCCCAAAGTAAGGA
3721 GCACCGTGGCCCAGCACCACGAGGCCCTGGTGGGTCACGGCTTCACCCACGCGCATATCG
3781 TTGCTCTGAGCCAACATCCCGCAGCTCTGGGTACCGTTGCGGTGAAGTATCAGGACATGA
3841 TCGCGGCACTGCCTGAAGCTACACACGAAGCCATAGTGGGCGTTGGCAAGCAGTGGAGCG
3901 GTGCCAGAGCGCTTGAGGCACTGTTGACGGTGGCTGGCGAGCTGAGGGGACCGCCACTGC
3961 AACTGGACACCGGCCAACTGCTGAAGATCGCCAAGAGGGGAGGCGTGACGGCGGTGGAGG
4021 CCGTGCATGCCTGGAGGAATGCCCTGACCGGCGCGCCCCTGAACCTGACACCAGAGCAAG
4081 TAGTGGCTATTGCAAGTAACATCGGTGGCAAACAAGCGCTGGAGACCGTGCAGAGGCTCC
4141 TTCCGGTGCTCTGCCAAGCACACGGTTTGACCCCCGAACAGGTTGTAGCCATAGCTTCTA
4201 ACAACGGAGGTAAGCAGGCACTGGAAACCGTGCAGCGCCTGCTCCCAGTACTGTGTCAGG
4261 CTCATGGGCTCACTCCGGAACAGGTGGTCGCAATCGCGAGCAACGGCGGCGGCAAGCAAG
4321 CCCTTGAGACAGTCCAAAGACTTTTGCCTGTCCTTTGTCAGGCGCATGGCCTTACGCCTG
4381 AGCAAGTCGTTGCGATCGCCTCCCACGACGGCGGAAAACAGGCTTTGGAAACCGTGCAGC
4441 GGTTGCTGCCCGTTTTGTGCCAAGCCCACGGATTGACCCCCGAACAGGTTGTAGCCATAG
4501 CTTCTAACATCGGAGGTAAGCAGGCACTGGAAACCGTGCAGCGCCTGCTCCCAGTACTGT
4561 GTCAGGCTCATGGGCTCACTCCGGAACAGGTGGTCGCAATCGCGAGCAACGGCGGCGGCA
4621 AGCAAGCCCTTGAGACAGTCCAAAGACTTTTGCCTGTCCTTTGTCAGGCGCATGGCCTTA
4681 CGCCTGAGCAAGTCGTTGCGATCGCCTCCCACGACGGCGGAAAACAGGCTTTGGAAACCG
4741 TGCAGCGGTTGCTGCCCGTTTTGTGCCAAGCCCACGGATTGACCCCCGAACAGGTTGTAG
4801 CCATAGCTTCTAACAACGGAGGTAAGCAGGCACTGGAAACCGTGCAGCGCCTGCTCCCAG
4861 TACTGTGTCAGGCTCATGGGCTTACGCCTGAGCAAGTCGTTGCGATCGCCTCCCACGACG
4921 GCGGAAAACAGGCTTTGGAAACCGTGCAGCGGTTGCTGCCCGTTTTGTGCCAAGCCCACG
4981 GACTCACTCCGGAACAGGTGGTCGCAATCGCGAGCAACGGCGGCGGCAAGCAAGCCCTTG
5041 AGACAGTCCAAAGACTTTTGCCTGTCCTTTGTCAGGCGCATGGCCTGACACCAGAGCAAG
5101 TAGTGGCTATTGCAAGTAACATCGGTGGCAAACAAGCGCTGGAGACCGTGCAGAGGCTCC
5161 TTCCGGTGCTCTGCCAAGCACACGGTTTGACCCCCGAACAGGTTGTAGCCATAGCTTCTA
5221 ACGGCGGAGGTAAGCAGGCACTGGAAACCGTGCAGCGCCTGCTCCCAGTACTGTGTCAGG
5281 CTCATGGGCTCACTCCGGAACAGGTGGTCGCAATCGCGAGCAACGGCGGCGGCAAGCAAG
5341 CCCTTGAGACAGTCCAAAGACTTTTGCCTGTCCTTTGTCAGGCGCATGGCCTGACACCAG
5401 AGCAAGTAGTGGCTATTGCAAGTAACATCGGTGGCAAACAAGCGCTGGAGACCGTGCAGA
5461 GGCTCCTTCCGGTGCTCTGCCAAGCACACGGTCTCACTCCGGAACAGGTGGTCGCAATCG
5521 CGAGCCACGACGGCGGCAAGCAAGCCCTTGAGACAGTCCAAAGACTTTTGCCTGTCCTTT
5581 GTCAGGCGCATGGCCTTACGCCTGAGCAAGTCGTTGCGATCGCCTCCCACGACGGCGGAA
5641 AACAGGCTTTGGAAACCGTGCAGCGGTTGCTGCCCGTTTTGTGCCAAGCCCACGGACTGA
5701 CACCAGAGCAAGTAGTGGCTATTGCAAGTAACATCGGTGGCAAACAAGCGCTGGAGAGCA
5761 TCGTGGCCCAGCTGTCTCGGCCCGACCCTGCCCTCGCCGCTCTGACCAACGACCACCTGG
5821 TGGCCCTGGCTTGCCTCGGGGGCAGGCCAGCTCTTGACGCCGTGAAGAAGGGCCTTCCTC
5881 ACGCCCCAGCCCTGATCAAGCGGACCAACAGAAGGATTCCCGAGAGGACATCACATCGAG
5941 TGGCAGATCACGCGCAAGTGGTCCGCGTGCTCGGATTCTTCCAGTGTCACTCCCACCCCG
6001 CACAAGCGTTCGATGACGCCATGACTCAATTTGGTATGTCGAGACACGGACTGCTGCAGC
6061 TCTTTCGTAGAGTCGGTGTCACAGAACTCGAGGCCCGCTCGGGCACACTGCCTCCCGCCT
6121 CCCAGCGGTGGGACAGGATTCTCCAAGCGAGCGGTATGAAACGCGCGAAGCCTTCACCTA
6181 CGTCAACTCAGACACCTGACCAGGCGAGCCTTCATGCGTTCGCAGACTCGCTGGAGAGGG
6241 ATTTGGACGCGCCCTCGCCCATGCATGAAGGGGACCAAACTCGCGCGTCAGCTAGCCCCA
6301 AGAAGAAGAGAAAGGTGGAGGCCAGCGGTTCCGGACGGGCTGACGCATTGGACGATTTTG
6361 ATCTGGATATGCTGGGAAGTGACGCCCTCGATGATTTTGACCTTGACATGCTTGGTTCGG
6421 ATGCCCTTGATGACTTTGACCTCGACATGCTCGGCAGTGACGCCCTTGATGATTTCGACC
6481 TGGACATGCTGATTAACTCTAGAGGCAGTGGAGAGGGCAGAGGAAGTCTGCTAACATGCG
6541 GTGACGTCGAGGAGAATCCTGGCCCAGTGAGCAAGGGCGAGGAGGATAACATGGCCATCA
6601 TCAAGGAGTTCATGCGCTTCAAGGTGCACATGGAGGGCTCCGTGAACGGCCACGAGTTCG
6661 AGATCGAGGGCGAGGGCGAGGGCCGCCCCTACGAGGGCACCCAGACCGCCAAGCTGAAGG
6721 TGACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTCCCCTCAGTTCATGTACG
6781 GCTCCAAGGCCTACGTGAAGCACCCCGCCGACATCCCCGACTACTTGAAGCTGTCCTTCC
6841 CCGAGGGCTTCAAGTGGGAGCGCGTGATGAACTTCGAGGACGGCGGCGTGGTGACCGTGA
6901 CCCAGGACTCCTCCCTGCAGGACGGCGAGTTCATCTACAAGGTGAAGCTGCGCGGCACCA
6961 ACTTCCCCTCCGACGGCCCCGTAATGCAGAAGAAGACCATGGGCTGGGAGGCCTCCTCCG
7021 AGCGGATGTACCCCGAGGACGGCGCCCTGAAGGGCGAGATCAAGCAGAGGCTGAAGCTGA
7081 AGGACGGCGGCCACTACGACGCTGAGGTCAAGACCACCTACAAGGCCAAGAAGCCCGTGC
7141 AGCTGCCCGGCGCCTACAACGTCAACATCAAGTTGGACATCACCTCCCACAACGAGGACT
7201 ACACCATCGTGGAACAGTACGAACGCGCCGAGGGCCGCCACTCCACCGGCGGCATGGACG

7261 AGCTGTACAAGTAACATGTTTAAGGGTTCCGGTTCCACTAGGTACAATTCGATATCAAGC
7321 TTATCGATAATCAACCTCTGGATTACAAAATTTGTGAAAGATTGACTGGTATTCTTAACT
7381 ATGTTGCTCCTTTTACGCTATGTGGATACGCTGCTTTAATGCCTTTGTATCATGCTATTG
7441 CTTCCCGTATGGCTTTCATTTTCTCCTCCTTGTATAAATCCTGGTTGCTGTCTCTTTATG
7501 AGGAGTTGTGGCCCGTTGTCAGGCAACGTGGCGTGGTGTGCACTGTGTTTGCTGACGCAA
7561 CCCCCACTGGTTGGGGCATTGCCACCACCTGTCAGCTCCTTTCCGGGACTTTCGCTTTCC
7621 CCCTCCCTATTGCCACGGCGGAACTCATCGCCGCCTGCCTTGCCCGCTGCTGGACAGGGG
7681 CTCGGCTGTTGGGCACTGACAATTCCGTGGTGTTGTCGGGGAAATCATCGTCCTTTCCTT
7741 GGCTGCTCGCCTGTGTTGCCACCTGGATTCTGCGCGGGACGTCCTTCTGCTACGTCCCTT
7801 CGGCCCTCAATCCAGCGGACCTTCCTTCCCGCGGCCTGCTGCCGGCTCTGCGGCCTCTTC
7861 CGCGTCTTCGCCTTCGCCCTCAGACGAGTCGGATCTCCCTTTGGGCCGCCTCCCCGCATC
7921 GATACCGTCGACCTCGATCGAGACCTAGAAAAACATGGAGCAATCACAAGTAGCAATACA
7981 GCAGCTACCAATGCTGATTGTGCCTGGCTAGAAGCACAAGAGGAGGAGGAGGTGGGTTTT
8041 CCAGTCACACCTCAGGTACCTTTAAGACCAATGACTTACAAGGCAGCTGTAGATCTTAGC
8101 CACTTTTTAAAAGAAAAGGGGGGACTGGAAGGGCTAATTCACTCCCAACGAAGACAAGAT
8161 ATCCTTGATCTGTGGATCTACCACACACAAGGCTACTTCCCTGATTGGCAGAACTACACA
8221 CCAGGGCCAGGGATCAGATATCCACTGACCTTTGGATGGTGCTACAAGCTAGTACCAGTT
8281 GAGCAAGAGAAGGTAGAAGAAGCCAATGAAGGAGAGAACACCCGCTTGTTACACCCTGTG
8341 AGCCTGCATGGGATGGATGACCCGGAGAGAGAAGTATTAGAGTGGAGGTTTGACAGCCGC
8401 CTAGCATTTCATCACATGGCCCGAGAGCTGCATCCGGACTGTACTGGGTCTCTCTGGTTA
8461 GACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTTAAGCCTCAA
8521 TAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGTGTGACTCTGGTAAC
8581 TAGAGATCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGCATGTGAGCAAAA
8641 GGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTC
8701 CGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACA
8761 GGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCG
8821 ACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCT
8881 CATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGT
8941 GTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAG
9001 TCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGC
9061 AGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTAC
9121 ACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGA
9181 GTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGC
9241 AAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACG
9301 GGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCA
9361 AAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGT
9421 ATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCA
9481 GCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACG
9541 ATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCA
9601 CCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGT
9661 CCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGT
9721 AGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCA
9781 CGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACA
9841 TGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGA
9901 AGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACT
9961 GTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGA
10021 GAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCG
10081 CCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTC
10141 TCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGA
10201 TCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAAT
10261 GCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTT
10321 CAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGT
10381 ATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGAC

2622-3276 bp: CMV promoter

3348-4064 bp: TALE N-terminus

4065-5756 bp: TALE repeat array (NI NN NG HD NI NG HD NN HD NG NI NG NG NI HD HD NI)

5757-6290 bp: TALE C-terminus

6291-6332 bp: NLS

6333-6497 bp: VP64

6504-6566 bp: 2A peptide
6567-7274 bp: mCherry

RVD Library Entry Vector

The following sequence was inserted into the TA-cloning site of pMD19-T vector (Takara, Inc.).

```
   1   TAGCTATACGTCTCATTGACCCCCGAACAGGTTGTAGCCATAGCTTCTAAGTCTTCAGAG
  61   ACGCTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAGCTGGCTAGTTAAG
 121   CTATCAACAAGTTTGTACAAAAAAGCTGAACGAGAAACGTAAAATGATATAAATATCAAT
 181   ATATTAAATTAGATTTTGCATAAAAAACAGACTACATAATACTGTAAAACACAACATATC
 241   CAGTCACTATGAATCAACTACTTAGATGGTATTAGTGACCTGTAGTCGACCGACAGCCTT
 301   CCAAATGTTCTTCGGGTGATGCTGCCAACTTAGTCGACCGACAGCCTTCCAAATGTTCTT
 361   CTCAAACGGAATCGTCGTATCCAGCCTACTCGCTATTGTCCTCAATGCCGTATTAAATCA
 421   TAAAAAGAAATAAGAAAAAGAGGTGCGAGCCTCTTTTTTGTGTGACAAAATAAAAACATC
 481   TACCTATTCATATACGCTAGTGTCATAGTCCTGAAAATCATCTGCATCAAGAACAATTTC
 541   ACAACTCTTATACTTTTCTCTTACAAGTCGTTCGGCTTCATCTGGATTTTCAGCCTCTAT
 601   ACTTACTAAACGTGATAAAGTTTCTGTAATTTCTACTGTATCGACCTGCAGACTGGCTGT
 661   GTATAAGGGAGCCTGACATTTATATTCCCCAGAACATCAGGTTAATGGCGTTTTTGATGT
 721   CATTTTCGCGGTGGCTGAGATCAGCCACTTCTTCCCCGATAACGGAGACCGGCACACTGG
 781   CCATATCGGTGGTCATCATGCGCCAGCTTTCATCCCCGATATGCACCACCGGGTAAAGTT
 841   CACGGGAGACTTTATCTGACAGCAGACGTGCACTGGCCAGGGGGATCACCATCCGTCGCC
 901   CGGGCGTGTCAATAATATCACTCTGTACATCCACAAACAGACGATAACGGCTCTCTCTTT
 961   TATAGGTGTAAACCTTAAACTGCATTTCACCAGCCCCTGTTCTCGTCAGCAAAAGAGCCG
1021   TTCATTTCAATAAACCGGGCGACCTCAGCCATCCCTTCCTGATTTTCCGCTTTCCAGCGT
1081   TCGGCACGCAGACGACGGGCTTCATTCTGCATGGTTGTGCTTACCAGACCGGAGATATTG
1141   ACATCATATATGCCTTGAGCAACTGATAGCTGTCGCTGTCAACTGTCACTGTAATACGCT
1201   GCTTCATAGCATACCTCTTTTTGACATACTTCGGGTATACATATCAGTATATATTCTTAT
1261   ACCGCAAAAATCAGCGCGCAAATACGCATACTGTTATCTGGCTTTTAGTAAGCCGGATCC
1321   ACGCGGCGTTTACGCCCCCCCTGCCACTCATCGCAGTACTGTTGTAATTCATTAAGCATT
1381   CTGCCGACATGGAAGCCATCACAAACGGCATGATGAACCTGAATCGCCAGCGGCATCAGC
1441   ACCTTGTCGCCTTGCGTATAATATTTGCCCATGGTGAAAACGGGGGCGAAGAAGTTGTCC
1501   ATATTGGCCACGTTTAAATCAAAACTGGTGAAACTCACCCAGGGATTGGCTGAGACGAAA
1561   AACATATTCTCAATAAACCCTTTAGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACA
1621   TCTTGCGAATATATGTGTAGAAACTGCCGGAAATCGTCGTGGTATTCACTCCAGAGCGAT
1681   GAAAACGTTTCAGTTTGCTCATGGAAAACGGTGTAACAAGGGTGAACACTATCCCATATC
1741   ACCAGCTCACCGTCTTTCATTGCCATACGGAATTCCGGATGAGCATTCATCAGGCGGGCA
1801   AGAATGTGAATAAAGGCCGGATAAAACTTGTGCTTATTTTTCTTTACGGTCTTTAAAAAG
1861   GCCGTAATATCCAGCTGAACGGTCTGGTTATAGGTACATTGAGCAACTGACTGAAATGCC
1921   TCAAAATGTTCTTTACGATGCCATTGGGATATATCAACGGTGGTATATCCAGTGATTTTT
1981   TTCTCCATTTTAGCTTCCTTAGCTCCTGAAAATCTCGATAACTCAAAAAATACGCCCGGT
2041   AGTGATCTTATTTCATTATGGTGAAAGTTGGAACCTCTTACGTGCCGATCAACGTCTCAT
2101   TTTCGCCAAAAGTTGGCCCAGGGCTTCCCGGTATCAACAGGGACACCAGGATTTATTTAT
2161   TCTGCGAAGTGATCTTCCGTCACAGGTATTTATTCGGCGCAAAGTGCGTCGGGTGATGCT
2221   GCCAACTTAGTCGACTACAGGTCACTAATACCATCTAAGTAGTTGATTCATAGTGACTGG
2281   ATATGTTGTGTTTTACAGTATTATGTAGTCTGTTTTTTATGCAAAATCTAATTTAATATA
2341   TTGATATTTATATCATTTTACGTTTCTCGTTCAGCTTTCTTGTACAAAGTGGTTGATCTA
2401   GAGGGCCCGCGGTTCGAACGTCTCTGAAGACAAGGAGGTAAGCAGGCACTGGAAACCGTG
2461   CAGCGCCTGCTCCCAGTACTGTGTCAGGCTCATGGGTGAGACGTATAGCTA
```

16-48 bp: N-terminus of the TALE repeats
57-2425 bp: ccdB cassette
2434-2496 bp: C-terminus of the TALE repeats

RVD Library Vector

```
   1   GTCGACGGATCGGGAGATCTCCCGATCCCCTATGGTGCACTCTCAGTACAATCTGCTCTG
  61   ATGCCGCATAGTTAAGCCAGTATCTGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGT
 121   GCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACCGACAATTGCATGAAGAATC
 181   TGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTTGAC
 241   ATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCAT
 301   ATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACG
 361   ACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTT
 421   TCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAG
 481   TGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGC
 541   ATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAG
 601   TCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGT
 661   TTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGC
 721   ACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGG
 781   GCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGCGCGTTTTGCCTGTACTGGGTCT
 841   CTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTT
 901   AAGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGTGTGAC
 961   TCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGTGGC
1021   GCCCGAACAGGGACTTGAAAGCGAAAGGGAAACCAGAGGAGCTCTCTCGACGCAGGACTC
1081   GGCTTGCTGAAGCGCGCACGGCAAGAGGCGAGGGGCGGCGACTGGTGAGTACGCCAAAAA
1141   TTTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCGAGAGCGTCAGTATTAAGCGGG
1201   GGAGAATTAGATCGCGATGGGAAAAAATTCGGTTAAGGCCAGGGGGAAAGAAAAAATATA
1261   AATTAAAACATATAGTATGGGCAAGCAGGGAGCTAGAACGATTCGCAGTTAATCCTGGCC
1321   TGTTAGAAACATCAGAAGGCTGTAGACAAATACTGGGACAGCTACAACCATCCCTTCAGA
1381   CAGGATCAGAAGAACTTAGATCATTATATAATACAGTAGCAACCCTCTATTGTGTGCATC
1441   AAAGGATAGAGATAAAAGACACCAAGGAAGCTTTAGACAAGATAGAGGAAGAGCAAAACA
1501   AAAGTAAGACCACCGCACAGCAAGCGGCCGGCCGCGCTGATCTTCAGACCTGGAGGAGGA
1561   GATATGAGGGACAATTGGAGAAGTGAATTATATAAATATAAAGTAGTAAAAATTGAACCA
1621   TTAGGAGTAGCACCCACCAAGGCAAAGAGAAGAGTGGTGCAGAGAGAAAAAAGAGCAGTG
1681   GGAATAGGAGCTTTGTTCCTTGGGTTCTTGGGAGCAGCAGGAAGCACTATGGGCGCAGCG
1741   TCAATGACGCTGACGGTACAGGCCAGACAATTATTGTCTGGTATAGTGCAGCAGCAGAAC
1801   AATTTGCTGAGGGCTATTGAGGCGCAACAGCATCTGTTGCAACTCACAGTCTGGGGCATC
1861   AAGCAGCTCCAGGCAAGAATCCTGGCTGTGGAAAGATACCTAAAGGATCAACAGCTCCTG
1921   GGGATTTGGGGTTGCTCTGGAAAACTCATTTGCACCACTGCTGTGCCTTGGAATGCTAGT
1981   TGGAGTAATAAATCTCTGGAACAGATTTGGAATCACACGACCTGGATGGAGTGGGACAGA
2041   GAAATTAACAATTACACAAGCTTAATACACTCCTTAATTGAAGAATCGCAAAACCAGCAA
2101   GAAAAGAATGAACAAGAATTATTGGAATTAGATAAATGGGCAAGTTTGTGGAATTGGTTT
2161   AACATAACAAATTGGCTGTGGTATATAAAATTATTCATAATGATAGTAGGAGGCTTGGTA
2221   GGTTTAAGAATAGTTTTTGCTGTACTTTCTATAGTGAATAGAGTTAGGCAGGGATATTCA
2281   CCATTATCGTTTCAGACCCACCTCCCAACCCCGAGGGGACCCGACAGGCCCGAAGGAATA
2341   GAAGAAGAAGGTGGAGAGAGAGACAGAGACAGATCCATTCGATTAGTGAACGGATCGGCA
2401   CTGCGTGCGCCAATTCTGCAGACAAATGGCAGTATTCATCCACAATTTTAAAAGAAAAGG
2461   GGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAGACATACA
2521   AACTAAAGAATTACAAAAACAAATTACAAAAATTCAAAATTTTCGGGTTTATTACAGGGA
2581   CAGCAGAGATCCAGTTTGGTTAGTACCGGGCCCGCTCTAGACGATGTACGGGCCAGATAT
2641   ACGCGTTGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTT
2701   CATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGA
2761   CCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCA
2821   ATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCA
2881   GTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGG
2941   CCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATC
3001   TACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGT
3061   GGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGT
3121   TTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTG
3181   ACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTCTCTGGCTA
3241   ACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACC
3301   CAAGCTGGCTAGCGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCATGAGGACCAGGC
3361   TGCCATCTCCCCCTGCCCCTTCCCCCGCATTTAGCGCTGGGAGCTTTAGCGACCTGCTTA
3421   GGCAGTTCGACCCCAGCTTGTTCAACACCAGCCTGTTTGACAGCCTGCCTCCCTTCGGAG
3481   CGCACCACACCGAAGCCGCCACCGGCGAGTGGGACGAGGTGCAAAGCGGCCTGAGGGCAG
3541   CGGACGCTCCTCCGCCAACCATGAGGGTGGCAGTGACAGCAGCTAGGCCCCCTCGGGCAA
```

```
3601  AACCTGCACCCAGGAGAAGGGCTGCCCAACCCAGCGACGCGAGTCCAGCCGCACAGGTGG
3661  ACCTCAGGACGCTGGGCTACAGCCAGCAACAGCAAGAGAAGATCAAGCCCAAAGTAAGGA
3721  GCACCGTGGCCCAGCACCACGAGGCCCTGGTGGGTCACGGCTTCACCCACGCGCATATCG
3781  TTGCTCTGAGCCAACATCCCGCAGCTCTGGGTACCGTTGCGGTGAAGTATCAGGACATGA
3841  TCGCGGCACTGCCTGAAGCTACACACGAAGCCATAGTGGGCGTTGGCAAGCAGTGGAGCG
3901  GTGCCAGAGCGCTTGAGGCACTGTTGACGGTGGCTGGCGAGCTGAGGGGACCGCCACTGC
3961  AACTGGACACCGGCCAACTGCTGAAGATCGCCAAGAGGGGAGGCGTGACGGCGGTGGAGG
4021  CCGTGCATGCCTGGAGGAATGCCCTGACCGGCGCGCCCCTGAACAGAGACGCTGGCTTAT
4081  CGAAATTAATACGACTCACTATAGGGAGACCCAAGCTGGCTAGTTAAGCTATCAACAAGT
4141  TTGTACAAAAAAGCTGAACGAGAAACGTAAAATGATATAAATATCAATATATTAAATTAG
4201  ATTTTGCATAAAAAACAGACTACATAATACTGTAAAACACAACATATCCAGTCACTATGA
4261  ATCAACTACTTAGATGGTATTAGTGACCTGTAGTCGACCGACAGCCTTCCAAATGTTCTT
4321  CGGGTGATGCTGCCAACTTAGTCGACCGACAGCCTTCCAAATGTTCTTCTCAAACGGAAT
4381  CGTCGTATCCAGCCTACTCGCTATTGTCCTCAATGCCGTATTAAATCATAAAAAGAAATA
4441  AGAAAAAGAGGTGCGAGCCTCTTTTTTGTGTGACAAAATAAAAACATCTACCTATTCATA
4501  TACGCTAGTGTCATAGTCCTGAAAATCATCTGCATCAAGAACAATTTCACAACTCTTATA
4561  CTTTTCTCTTACAAGTCGTTCGGCTTCATCTGGATTTTCAGCCTCTATACTTACTAAACG
4621  TGATAAAGTTTCTGTAATTTCTACTGTATCGACCTGCAGACTGGCTGTGTATAAGGGAGC
4681  CTGACATTTATATTCCCCAGAACATCAGGTTAATGGCGTTTTTGATGTCATTTTCGCGGT
4741  GGCTGAGATCAGCCACTTCTTCCCCGATAACGGAGACCGGCACACTGGCCATATCGGTGG
4801  TCATCATGCGCCAGCTTTCATCCCCGATATGCACCACCGGGTAAAGTTCACGGGAGACTT
4861  TATCTGACAGCAGACGTGCACTGGCCAGGGGGATCACCATCCGTCGCCCGGGCGTGTCAA
4921  TAATATCACTCTGTACATCCACAAACAGACGATAACGGCTCTCTCTTTTATAGGTGTAAA
4981  CCTTAAACTGCATTTCACCAGCCCCTGTTCTCGTCAGCAAAAGAGCCGTTCATTTCAATA
5041  AACCGGGCGACCTCAGCCATCCCTTCCTGATTTTCCGCTTTCCAGCGTTCGGCACGCAGA
5101  CGACGGGCTTCATTCTGCATGGTTGTGCTTACCAGACCGGAGATATTGACATCATATATG
5161  CCTTGAGCAACTGATAGCTGTCGCTGTCAACTGTCACTGTAATACGCTGCTTCATAGCAT
5221  ACCTCTTTTTGACATACTTCGGGTATACATATCAGTATATATTCTTATACCGCAAAAATC
5281  AGCGCGCAAATACGCATACTGTTATCTGGCTTTTAGTAAGCCGGATCCACGCGGCGTTTA
5341  CGCCCCCCCTGCCACTCATCGCAGTACTGTTGTAATTCATTAAGCATTCTGCCGACATGG
5401  AAGCCATCACAAACGGCATGATGAACCTGAATCGCCAGCGGCATCAGCACCTTGTCGCCT
5461  TGCGTATAATATTTGCCCATGGTGAAAACGGGGGCGAAGAAGTTGTCCATATTGGCCACG
5521  TTTAAATCAAAACTGGTGAAACTCACCCAGGGATTGGCTGAGACGAAAAACATATTCTCA
5581  ATAAACCCTTTAGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACATCTTGCGAATAT
5641  ATGTGTAGAAACTGCCGGAAATCGTCGTGGTATTCACTCCAGAGCGATGAAAACGTTTCA
5701  GTTTGCTCATGGAAAACGGTGTAACAAGGGTGAACACTATCCCATATCACCAGCTCACCG
5761  TCTTTCATTGCCATACGGAATTCCGGATGAGCATTCATCAGGCGGGCAAGAATGTGAATA
5821  AAGGCCGGATAAAACTTGTGCTTATTTTTCTTTACGGTCTTTAAAAAGGCCGTAATATCC
5881  AGCTGAACGGTCTGGTTATAGGTACATTGAGCAACTGACTGAAATGCCTCAAAATGTTCT
5941  TTACGATGCCATTGGGATATATCAACGGTGGTATATCCAGTGATTTTTTTCTCCATTTTA
6001  GCTTCCTTAGCTCCTGAAAATCTCGATAACTCAAAAAATACGCCCGGTAGTGATCTTATT
6061  TCATTATGGTGAAAGTTGGAACCTCTTACGTGCCGATCAACGTCTCATTTTCGCCAAAAG
6121  TTGGCCCAGGGCTTCCCGGTATCAACAGGGACACCAGGATTTATTTATTCTGCGAAGTGA
6181  TCTTCCGTCACAGGTATTTATTCGGCGCAAAGTGCGTCGGGTGATGCTGCCAACTTAGTC
6241  GACTACAGGTCACTAATACCATCTAAGTAGTTGATTCATAGTGACTGGATATGTTGTGTT
6301  TTACAGTATTATGTAGTCTGTTTTTTATGCAAAATCTAATTTAATATATTGATATTTATA
6361  TCATTTTACGTTTCTCGTTCAGCTTTCTTGTACAAAGTGGTTGATCTAGAGGGCCCGCGG
6421  TTCGAACGTCTCTAGCATCGTGGCCCAGCTGTCTCGGCCCGACCCTGCCCTCGCCGCTCT
6481  GACCAACGACCACCTGGTGGCCCTGGCTTGCCTCGGGGGCAGGCCAGCTCTTGACGCCGT
6541  GAAGAAGGGCCTTCCTCACGCCCCAGCCCTGATCAAGCGGACCAACAGAAGGATTCCCGA
6601  GAGGACATCACATCGAGTGGCAGATCACGCGCAAGTGGTCCGCGTGCTCGGATTCTTCCA
6661  GTGTCACTCCCACCCCGCACAAGCGTTCGATGACGCCATGACTCAATTTGGTATGTCGAG
6721  ACACGGACTGCTGCAGCTCTTTCGTAGAGTCGGTGTCACAGAACTCGAGGCCCGCTCGGG
6781  CACACTGCCTCCCGCCTCCCAGCGGTGGGACAGGATTCTCCAAGCGAGCGGTATGAAACG
6841  CGCGAAGCCTTCACCTACGTCAACTCAGACACCTGACCAGGCGAGCCTTCATGCGTTCGC
6901  AGACTCGCTGGAGAGGGATTTGGACGCGCCCTCGCCCATGCATGAAGGGGACCAAACTCG
6961  CGCGTCAGCTAGCCCCAAGAAGAAGAGAAAGGTGGAGGCCAGCGGTTCCGGACGGGCTGA
7021  CGCATTGGACGATTTTGATCTGGATATGCTGGGAAGTGACGCCCTCGATGATTTTGACCT
7081  TGACATGCTTGGTTCGGATGCCCTTGATGACTTTGACCTCGACATGCTCGGCAGTGACGC
7141  CCTTGATGATTTCGACCTGGACATGCTGATTAACTCTAGAGGCAGTGGAGAGGGCAGAGG
7201  AAGTCTGCTAACATGCGGTGACGTCGAGGAGAATCCTGGCCCAGTGAGCAAGGGCGAGGA
7261  GGATAACATGGCCATCATCAAGGAGTTCATGCGCTTCAAGGTGCACATGGAGGGCTCCGT
7321  GAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGCCCCTACGAGGGCACCCA
```

```
7381   GACCGCCAAGCTGAAGGTGACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTC
7441   CCCTCAGTTCATGTACGGCTCCAAGGCCTACGTGAAGCACCCCGCCGACATCCCCGACTA
7501   CTTGAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAACTTCGAGGACGG
7561   CGGCGTGGTGACCGTGACCCAGGACTCCTCCCTGCAGGACGGCGAGTTCATCTACAAGGT
7621   GAAGCTGCGCGGCACCAACTTCCCCTCCGACGGCCCCGTAATGCAGAAGAAGACCATGGG
7681   CTGGGAGGCCTCCTCCGAGCGGATGTACCCCGAGGACGGCGCCCTGAAGGGCGAGATCAA
7741   GCAGAGGCTGAAGCTGAAGGACGGCGGCCACTACGACGCTGAGGTCAAGACCACCTACAA
7801   GGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTCAACATCAAGTTGGACATCAC
7861   CTCCCACAACGAGGACTACACCATCGTGGAACAGTACGAACGCGCCGAGGGCCGCCACTC
7921   CACCGGCGGCATGGACGAGCTGTACAAGTAACATGTTTAAGGGTTCCGGTTCCACTAGGT
7981   ACAATTCGATATCAAGCTTATCGATAATCAACCTCTGGATTACAAAATTTGTGAAAGATT
8041   GACTGGTATTCTTAACTATGTTGCTCCTTTTACGCTATGTGGATACGCTGCTTTAATGCC
8101   TTTGTATCATGCTATTGCTTCCCGTATGGCTTTCATTTTCTCCTCCTTGTATAAATCCTG
8161   GTTGCTGTCTCTTTATGAGGAGTTGTGGCCCGTTGTCAGGCAACGTGGCGTGGTGTGCAC
8221   TGTGTTTGCTGACGCAACCCCCACTGGTTGGGGCATTGCCACCACCTGTCAGCTCCTTTC
8281   CGGGACTTTCGCTTTCCCCCTCCCTATTGCCACGGCGGAACTCATCGCCGCCTGCCTTGC
8341   CCGCTGCTGGACAGGGGCTCGGCTGTTGGGCACTGACAATTCCGTGGTGTTGTCGGGGAA
8401   ATCATCGTCCTTTCCTTGGCTGCTCGCCTGTGTTGCCACCTGGATTCTGCGCGGGACGTC
8461   CTTCTGCTACGTCCCTTCGGCCCTCAATCCAGCGGACCTTCCTTCCCGCGGCCTGCTGCC
8521   GGCTCTGCGGCCTCTTCCGCGTCTTCGCCTTCGCCCTCAGACGAGTCGGATCTCCCTTTG
8581   GGCCGCCTCCCCGCATCGATACCGTCGACCTCGATCGAGACCTAGAAAAACATGGAGCAA
8641   TCACAAGTAGCAATACAGCAGCTACCAATGCTGATTGTGCCTGGCTAGAAGCACAAGAGG
8701   AGGAGGAGGTGGGTTTTCCAGTCACACCTCAGGTACCTTTAAGACCAATGACTTACAAGG
8761   CAGCTGTAGATCTTAGCCACTTTTTAAAAGAAAAGGGGGGACTGGAAGGGCTAATTCACT
8821   CCCAACGAAGACAAGATATCCTTGATCTGTGGATCTACCACACACAAGGCTACTTCCCTG
8881   ATTGGCAGAACTACACACCAGGGCCAGGGATCAGATATCCACTGACCTTTGGATGGTGCT
8941   ACAAGCTAGTACCAGTTGAGCAAGAGAAGGTAGAAGAAGCCAATGAAGGAGAGAACACCC
9001   GCTTGTTACACCCTGTGAGCCTGCATGGGATGGATGACCCGGAGAGAGAAGTATTAGAGT
9061   GGAGGTTTGACAGCCGCCTAGCATTTCATCACATGGCCCGAGAGCTGCATCCGGACTGTA
9121   CTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACC
9181   CACTGCTTAAGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGT
9241   TGTGTGACTCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTA
9301   GCAGCATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGG
9361   CGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGA
9421   GGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCG
9481   TGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGG
9541   GAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTC
9601   GCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCG
9661   GTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCA
9721   CTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGT
9781   GGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAG
9841   TTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCG
9901   GTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATC
9961   CTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTT
10021  TGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTT
10081  TTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCA
10141  GTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCG
10201  TCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATAC
10261  CGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGG
10321  CCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCC
10381  GGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTA
10441  CAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAAC
10501  GATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTC
10561  CTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCAC
10621  TGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACT
10681  CAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAA
10741  TACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTT
10801  CTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCA
10861  CTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAA
10921  AAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATAC
10981  TCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCG
11041  GATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCC
11101  GAAAAGTGCCACCTGAC
```

2622-3276 bp: CMV promoter
3348-4064 bp: TALE N-terminus
4065-6433 bp: ccdB cassette
6434-6967 bp: TALE C-terminus
6968-7009 bp: NLS
7010-7174 bp: VP64
7181-7243 bp: 2A peptide
7244-7951 bp: mCherry

SUPPLEMENTARY METHODS

Artificial system for RVD screening

The artificial screening system was composed of four reporters and a TALE-VP64 expression library in which the RVDs of three consecutive monomers in the middle of an artificial TALE array were encoded by the same 6 randomly synthesized nucleotides (TALE-$(XX')_3$). TALE-$(XX')_3$ contained 14.5 repeats fused with the VP64 trans-activation domain and 2A peptide-linked mCherry. The variable diresidues (XX') for testing were placed in the 7^{th} - 9^{th} repeat modules, and the variable RVD-carrying TALE modules were purposely designed as triplets to augment the DNA binding capability. X and X' represents the 12^{th} and 13^{th} amino acids in the 7^{th} - 9^{th} repeat modules, respectively. Four reporters consist of TALE-$(XX')_3$ binding sites CTGGCC<u>NNN</u>TACGTA, in which N represents A, T, C or G, was located immediately upstream of a minimal CMV promoter (P_{minCMV}) and its downstream EGFP gene. TALE-Ctrl was constructed to have the identical backbone as TALE-$(XX')_3$ except that its TALE repeats (16.5-mers) are different, not matching with any reporters. The mCherry-normalized EGFP level of TALE-Ctrl co-transfected with reporter served as the corresponding basal level. For each sample, EGFP fluorescence intensity was normalized to mCherry intensity. Fold induction is calculated as the result of normalized sample EGFP intensity divided by normalized basal level shown as follows.

$$\mathbf{Fold\ Induction} = \frac{\left(\prod_{i=1}^{n} \mathbf{Exp}_{\mathbf{EGFP}_i}\right)^{1/n} \Big/ \left(\prod_{i=1}^{n} \mathbf{Exp}_{\mathbf{mCherry}_i}\right)^{1/n}}{\left(\prod_{i=1}^{n} \mathbf{Ctrl}_{\mathbf{EGFP}_i}\right)^{1/n} \Big/ \left(\prod_{i=1}^{n} \mathbf{Ctrl}_{\mathbf{mCherry}_i}\right)^{1/n}}$$

Notes:

$\left(\prod_{i=1}^{n} \mathbf{Exp}_{\mathbf{EGFP}_i}\right)^{1/n}$: Geometric mean of $\mathbf{Exp}_{\mathbf{EGFP}}$ of cells from FACS (n = number of cells)

$\mathbf{Exp}_{\mathbf{EGFP}}$: EGFP intensity of HEK293T cells co-transfected with TALE-(XX')$_m$ plus reporter (m = 3, 6 and 12, corresponding to TALE-(XX')$_3$, TALE-(XX')$_6$ and TALE-(XX')$_{12}$, respectively)

$\mathrm{Exp}_{\mathrm{mCherry}}$: mCherry intensity of HEK293T cells co-transfected with TALE-(XX')$_m$ plus reporter

$\mathbf{Ctrl}_{\mathbf{EGFP}}$: EGFP intensity of HEK293T cells co-transfected with TALE-Ctrl plus reporter

$\mathbf{Ctrl}_{\mathbf{mCherry}}$: mCherry intensity of HEK293T cells co-transfected with TALE-Ctrl plus reporter

Construction of the TALE-(XX')$_3$ library

A 102-nt monomer encoding a standard TALE repeat unit was synthesized with the following sequences: 5'-GCTATGGCTACAACCTGTTCGGGGGTCAACCCATGAGCCT GACACAGTACTGGGAGCAGGCGCTGCACGGTTTCCAGTGCCTGCTTACCTCCNNNN NNAGAA-3'. The six random nucleotides (underlined Ns) corresponding to the RVD-encoding region were purposely placed near the 3' end to ensure unbiased oligonucleotide synthesis. This single-stranded DNA was cyclized using linker primer 1 (5'-GAACAGGTT GTAGCCATAGCTTCT-3') and T4 DNA ligase (NEB). Using the agarose gel-purified single stranded circular DNA as template, rolling-circle amplification was conducted with phi29 DNA polymerase (NEB) and primer 1 (30°C, 90 min) followed by primer extension using primer 2 (5'-AGGTTGTAGCCATAGCT-3') and phi29 (30°C, 90 min) to acquire long dsDNA. After ultrasonic shearing (270 W, work 10 s, pause 10 s, 10 cycles) and T4 DNA polymerase (NEB) treatment to blunt the ends of the DNA fragments (12°C, 15 min, with 400 μM dNTP mix), 250-400 bp DNA fragments were harvested by gel purification. These DNA fragments were then cloned into a pre-made entry vector using the ligase-independent cloning (LIC) method to create an entry library. BsmBI digestion of entry library clones produced ~300 bp DNA fragments that were subsequently cloned, via the Golden Gate approach[1], into a pre-made RVD library vector, which was constructed using the ULtiMATE protocol previously developed by our group[2]. Each plasmid in the final RVD library was verified through sequencing analysis. About 350 kinds of TALE-(XX')$_3$ constructs were obtained from this approach.

Individual construction of TALE-(XX')$_3$

A complementary primer (5'-aaCGTCTCaGTTCGGGGGTCAACCCATGAGCCTGACAC AGTACTGGGAGCAGGCGCTGCACGGTTTCCAGTGCCTGCTT-3') and a specific primer (5'-tCGTCTCaGAACAGGTTGTAGCCATAGCTTCTNNNNNNGGAGGTAAGCAGGCACT GGAA-3'; NNNNNN indicates the RVD codons) were annealed and PCR extended to generate a 102 bp monomer with BsmBI sites at each end. The monomer was ligated via a 6 cycles of the Golden Gate method[1] to generate repeats. The repeat product was PCR amplified using the primers G-lib-F (5'-TAGCTATACGTCTCATTGACCCCCGAACAGGTT GTAGCC-3') and G-lib-R (5'-TAGCTATACGTCTCACCCATGAGCCTGACACAGTACTGG GAGCA-3') and Taq Hifi (Transgen, Inc.). The 3-repeat fragment was gel purified and subsequently cloned into a pre-made RVD library vector via the Golden Gate method[1]. Trans1-T1 competent cells (Transgen, Inc.) were used for bacterial transformation. About 50 kinds of TALE-(XX')$_3$ constructs were obtained from this approach.

Design and construction of TALE-(XX')$_6$ and TALE-(XX')$_{12}$

For TALE-(XX')$_6$, the customized TALEs (17.5-mer) are the same as TALE-(XX')$_3$ except that there are six identical repeats containing the variable RVDs. Accordingly, four reporters were constructed, consisting of TALE-(XX')$_6$ binding sites with six consecutive nucleotides (A, T, C or G) substituted at positions 7 - 12 in front of a minimal CMV promoter and its downstream EGFP gene. The 7th - 9th repeats in TALE-(XX')$_3$ were PCR amplified using the primers G-lib-seq-F (5'-TCTAGGTACCAAGCCCACGGATTGA-3') and G-lib-seq-R (5'-ATCGATCGTCCGGAGTGAGCCCA-3'). The 300 bp PCR product was purified and further PCR amplified using the primers G-lib-F and G-lib-R. The product of the 6-repeat fragment was gel purified and cloned into a pre-made RVD library vector via the Golden Gate method[1] to construct the final TALE-$(XX')_6$ plasmid.

For TALE-$(XX')_{12}$, the customized TALEs (15.5-mer) contained four ACTC-targeting TALE repeat modules followed by 11.5 repeat units with variable RVDs to be tested. Accordingly, four reporters were constructed, consisting of TALE-$(XX')_{12}$ binding sites with 12 consecutive nucleotides (A, T, C or G) substituted at positions 5 - 16 in front of a minimal CMV promoter and its downstream EGFP gene. TALE-$(XX')_{12}$ plasmids were constructed with a similar method using TALE-$(XX')_6$ as a PCR template and applying the ULtiMATE protocol[2].

Backbone plasmid	10 ng
6-repeat fragment	5 ng
10X Tango buffer (Thermo)	1 µl
10 mM ATP	1 µl
BsmBI (Thermo)	0.75 µl
T4 ligase (NEB)	0.25 µl
ddH_2O	up to 10 µl

37°C 5 min \
| 10 cycles
16°C 5 min /

37°C 5 min

Cell culture, transfection and flow cytometric analysis

HEK293T cells (from Stanley Cohen lab at Stanford University) were cultured in DMEM medium with 10% FBS and 1% penicillin-streptomycin at 37°C and 5% CO_2. Cells were seeded 24 h prior to transfection in 24-well plates at a density of 10^5 cells per well. The cells in each well were co-transfected with 0.2 μg TALE-$(XX')_3$ plasmid and 0.4 μg reporter plasmid using polyethylenimine (PEI)[3]. At 48 h post-transfection, the cells were collected and analyzed on a BD LSR II flow cytometer (BD Biosciences). Lasers with wavelengths of 488 nm and 561 nm were used to quantify EGFP and mCherry protein expression, respectively. At least 10,000 events were collected from every sample to obtain sufficient data for analysis. Majority of cells showed mCherry fluorescence intensity of 5×10^4 – 5×10^5, thus were gated for further analysis. The binding efficiencies and specificities of the variable RVDs (XX') in each customized TALE were assayed by comparing the fold induction of EGFP reporters with the basal level of EGFP in HEK293T cells transfected with the reporter plasmid and a customized TALE plasmid containing unmatched TALE repeats (Supplementary Sequences). The EGFP fluorescence intensity assayed from FACS analysis was normalized to the corresponding mCherry fluorescence intensity.

Generation of a heat map illustrating the base-preference of RVDs

The heat map was generated from library screening of TALE-$(XX')_3$ using four reporters (3A, 3T, 3C, and 3G), thereby reflecting the base preference of 400 RVDs. EGFP activities from different reporters are coded by different colors representing the reporter identities (3A, green; 3T, red; 3C, blue; and 3G, yellow). The brightness of the colors indicates the fold induction of the reporters by TALE-$(XX')_3$ compared to their basal levels.

Criteria for selection of novel RVDs from TALE-$(XX')_3$ for intensive study

The standards are as follows: (1) the highest fold induction among four reporters is at least equal to NK for G recognition; and (2) the second highest fold induction is lower than half of the highest one. In addition to these simple rules, we have also included groups of

RVDs that displayed some unique patterns, i.e. seven RVDs ended with Ala (KA, CA, FA, YA, RA, PA, and AA) that showed preference for T-recognition, and RVDs recognizing both A and G (NN and HN).

Statistical analysis

For comparison between the induction level of TALE-$(XX')_{12}$ and the basal level (Fig. 1f), two-sample, one-tailed t-test was performed assuming unequal variance. *P<0.05, **P<0.01, and ***P<0.005.

Construction of TALENs

TALENs with natural RVDs (i.e. NI, NG, HD and NN) were constructed using ULtiMATE system as previously described[2]. For TALEN repeats using novel RVDs KN (in place of NN), NH (in place of NN) or RG (in place of NG), TALE monomers containing new RVDs were individually synthesized. The final assembly of these TALENs constructs was conducted using the same ULtiMATE protocol as above.

Assessment of TALENs-mediated indels

HEK293T cells were seeded in 6-well plates at a density of 3×10^5 cells per well and incubated at 37°C with 5% CO_2. For each well, a pair of TALEN plasmids and pmaxGFP (Lonza Group Ltd.) were co-transfected at a ratio of 9:9:2 (0.9 µg : 0.9 µg : 0.2 µg) using PEI method. The transfected cells were cultured at 37°C for one day followed by 3 days of incubation at 30°C (cold shock) before flow cytometric sorting for GFP positive cells. TALENs-targeting regions were PCR-amplified from the genome DNA of the isolated GFP positive cells. The TALENs-mediated indels were analyzed by mismatch-sensitive T7 endonuclease I (T7E1; New England Biolabs) as described previously[4].