Sequence-specific recognition of methylated DNA by an engineered transcription activator-like effector protein

*Shogo Tsuji, Shiroh Futaki, Miki Imanishi**

SUPPORTING INFORMATION

Material and methods

Plasmids construction

Mammalian cell expression vectors of TAL_{Dcm}, that targets a DNA sequence "5'-TCCTGGTATATCCCCC-3'", with a nuclear localization signal (NLS) and a synthetic transcription activation domain (VP64)[1] driven by a CMV promoter were constructed as follows. At first, *Mfe*I and *Hin*dIII sites were introduced in RVD regions of pNG2 vector[2] by QuikChange using Phusion Hi-Fidelity DNA polymerase (New England Biolabs, (NEB)), creating pNG2+. Then, according to the method described by Zhang *et al*[2]., 14.5 tandem repeats (using pNG2+ instead of pNG2) were assembled to correspond to the TAL_{Dcm} RVDs and inserted into pCMVt[3], creating TAL_{Dcm}/pCMVt. A luciferase reporter vector, 3×TAL_{Dcm}/pGL3, was constructed by inserting 3×TAL_{Dcm} binding sequence into pGL3 (Promega).

The bacterial expression vectors for B1H screenings were constructed as follows. pB1H2W5-Prd (Addgene plasmid 18039) was cleaved by *Kpn*I/*Xba*I and a DNA fragment containing a His-tag-coding sequence followed by *Sac*I and *Nhe*I sites (5'-GGTACCCACCATCATCACCACCATGAGCTCTGTCGCGCTAGCTCTAGA-3') was inserted[4]. The coding region of TAL_{Dcm} was inserted to the vector, such that the omega-subunit of the RNA polymerase is fused at the N-terminus of TAL_{Dcm}, creating TAL_{Dcm}/pB1H2W5. DNA fragments (5'-CAATTGCCNNSNNSNNSNNSGGCAAGCAAGCTT-3' S = C/G) was inserted into *Mfe*I and *Hin*dIII sites of TAL_{Dcm}/pB1H2W5. The vectors were electroporated into NEB 10β Electrocompetent *E. Coli* (NEB). Plasmids were purified from about 5.0×10^6 transformants, yielding the "XXXX"-library. The "XXAA"-library was constructed in the same way except for using DNA fragments (5'-CAATTGCCNNSNNSGCGGCCGGCAAGCAAGCTT-3'). To create B1H reporter plasmids, a TAL_{Dcm} target sequence was inserted into *Not*I/*Eco*RI sites of pH3U3-mcs (Addgene plasmid 12609) [5]. *Escherichia coli* 11.5TAL_{Dcm} expression vectors were created as previously described[3]. In short, the coding region of 11.5TAL_{Dcm} that targets a DNA sequence "5'-TCCTGGTATATCC-3'", was inserted into pET42b (Novagen), and His-tag coding sequence was inserted into N-terminal of 11.5TAL_{Dcm}, creating N-His 11.5TAL_{Dcm}/pET42b.

The mammalian cell expression vectors of TAL_{RASSF2} were constructed by golden gate assembly.[2] A p300 histone acetyltransferase coding sequence from pcDNA3.1-p300 (Addgene plasmid 23252) was inserted into the C-terminal of TAL_{RASSF2}, creating TAL_{RASSF2}-P300/pCMVt.

The sequences of all these constructed plasmids were confirmed. The amino acid sequences of these proteins are described below.

Luciferase reporter assay

The methylated (5mC) and unmethylated (C) reporter plasmids were prepared by transforming the 3×TAL_{Dcm}/pGL3 into ECOS chemical competent cells (NEB), and dam-/dcm- competent *E. coli* (NEB), respectively. Methylation status of the plasmids was confirmed by digesting the plasmids by *Bam*HI (NEB) and *Psp*GI (NEB) and following gel electrophoresis. One hundred fifty ng of the

expression vector, 200 ng of the reporter vector, and 50 ng of the control vector (pRL-TK; Promega) were transiently co-transfected into HeLa cells using the Lipofectamine LTX (Thermo Fisher Scientific). For mock assay, 150 ng of empty pCMVt vector was used instead of the expression vector. After 24 h, luciferase activity was measured using the dual luciferase reporter system (Promega). The luminescence was obtained by normalization to the transfection control. For the experiments using different percentage of the 5mC reporter vectors, C and 5mC reporters were mixed to be 200 ng in total in the indicated ratios.

Real-time monitoring of luciferase luminescence

One thousand ng expression vector and 200 ng reporter vector were transiently co-transfected into HeLa cells using Lipofectamine LTX (Thermo Fisher Scientific). After 5 h, the medium was replaced with 2 mL of culture medium (10% FBS/DMEM) supplemented with 0.1 mM D-luciferin. Luciferase activity was monitored with an LM-2400 (Hamamatsu Photonics).

Bacterial one-hybrid screening

The TALE library and pH3U3 reporter vectors were co-transformed into the omega knockout bacteria hybrid selection strain, USC hisB- pyrF- rpoZ- (Addgene 18049), and screening was performed as described in Noyes *et al.*[6]. The cells were plated on NM media plates containing kanamycin (25 mg/ml), carbenicillin (100 mg/ml), 3-aminotriazole (5 mM) and isopropyl-β-D-thiogalactopyranoside (IPTG) (10 μM). The cells were cultured for 72 h at 37°C. Individual surviving colonies on 5 mM 3-aminotriazole plates were isolated. Then the sequences of the randomized regions were analyzed and displayed as a sequence logo[7].

Protein purification and Electrophoretic mobility shift assay (EMSA)

The 11.5TAL$_{Dcm}$ proteins with “ASAA” repeat or RVD “NG” at position 2 were expressed in *E. coli* BL21(DE3) cells by induction with 0.2 mM IPTG for 12 h at 18°C. The proteins were purified, and EMSA was performed, as previously described except that 6-FAM-labeled oligo DNA was used. The oligo DNAs used in EMSA are shown below[3].

Bisulfite sequencing

Genomic DNA was extracted using NucleoSpin Tissue (TAKARA) and converted using EZ DNA Methylation-Gold Kit (ZYMO research) according to the manufacturer's instructions. The converted DNA was PCR amplified by ZymoTaq DNA Polymerase (ZYMO research), and cloned into a TOPO TA vector (Thermo Fisher Scientific). Then, sequencing was performed. The sequencing results were analyzed using QUMA software[8].

TAL$_{RASSF2}$ transfection and RT-qPCR

TAL$_{RASSF2}$ expression vectors were transfected into HCT116 and SW480 cells using Lipofectamine

3000 and Lipofectamine LTX, respectively. For mock assay, same amount of empty vector was transfected instead of the TAL_{RASSF2} expression vector. After 24 h, total RNA was isolated using NucleoSpin RNA (TAKARA) from transfected cells. Then, cDNA was synthesized from 500 ng of total RNA using PrimeScript RT reagent Kit with gDNA Eraser (TAKARA). Real-time PCR was carried out using PowerUp SYBR Green Master Mix (Thermo Fisher Scientific), by 7300 Real-Time PCR System (Thermo Fisher Scientific). *Glyceraldehydes-3-phosphate dehydrogenase* (*GAPDH*) was used as an internal control gene. The sequences of the primers are shown below.

Protein sequence of TAL_{Dcm}-based transcription factor

MDPIRPRRPSPARELLPGPQPDRVQPTADRGVSAPAGSPLDGLPARRTVSRTRLPSPPAPSPAFSAGSFSDLLRPFD
PSLLDTSLLDSMPAVGTPHTAAAPAEWDEAQSALRAADDPPPTVRVAVTAARPPRAKPAPRRRAAQPSDASPAAQVD
LRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVTYQHIITALPEATHEDIVGVGKQ
WSGARALEALLTDAGELRGPPLQLDTGQLVKIAKRGGVTAMEAVHASRNALTGAPLN
LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG
LTPDQVVAIAxxxxGKQALETVQRLLPVLCQDHG
LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASNHGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASNHGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG
LTPDQVVAIASHDGGKQALE
SIVAQLSRPDPALAALTNDHLVALACLGGRPAMDAVKKGLPHAPELIRRVNRRIGERTSHRVADYAQVVRVLEFFQC
HSHPAYAFDEAMTQFGMSRNGLVQLFRRVGVTELEARGGTLPPASQRWDRILQASGMKRAKPSPTSAQTPDQASLHA
SPKKKRKVEASGSGRADALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLINSRDYKDDDD
K*

= NG, HD or the sequences of the mutants selected from B1H screening

= RVDs

= NLS

= activation domain(VP64)

= FLAG tag

Protein sequence of omega-TAL$_{Dcm}$

MARVTVQDAVEKIGNRFDLVLVAARRARQMQVGGKDPLVPEENDKTTVIALREIEEGLINNQILDVRERQEQQEQEA
AELQAVTAIAEGRAAADYKDDDDKFRTGSKTPPHGTHHHHHHELATMDPIRPRRPSPARELLPGPQPDRVQPTADRG
VSAPAGSPLDGLPARRTVSRTRLPSPPAPSPAFSAGSFSDLLRPFDPSLLDTSLLDSMPAVGTPHTAAAPAEWDEAQ
SALRAADDPPPTVRVAVTAARPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVG
HGFTHAHIVALSQHPAALGTVAVTYQHIITALPEATHEDIVGVGKQWSGARALEALLTDAGELRGPPLQLDTGQLVK
IAKRGGVTAMEAVHASRNALTGAPLN

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIAXXXXGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNHGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNHGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALE

SIVAQLSRPDPALAALTNDHLVALACLGGRPAMDAVKKGLPHAPELIRRVNRRIGERTSHRVADYAQVVRVLEFFQC
HSHPAYAFDEAMTQFGMSRNGLVQLFRRVGVTELEARGGTLPPASQRWDRILQASGMKRAKPSPTSAQTPDQASLHA
SSRD*

= Randomized region of repeat 2

= RVDs

= FLAG-tag

= His-tag

Protein sequence of N-His 11.5TAL$_{Dcm}$

MHHHHHHELRTRLPSPPAPSPAFSAGSFSDLLRPFDPSLLDTSLLDSMPAVGTPHTAAAPAEWDEAQSALRAADDPP
PTVRVAVTAARPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVA
LSQHPAALGTVAVTYQHIITALPEATHEDIVGVGKQWSGARALEALLTDAGELRGPPLQLDTGQLVKIAKRGGVTAM
EAVHASRNALTGAPLN

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIAxxxxGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNHGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNHGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALE

SIVAQLSRPDPALAALTNDHLVALACLGGRPAMDAVKKGLPHAPELIRRVNRRIGERTSHRVADYAQVVRVLEFFQC
HSHPAYAFDEAMTQFGMSRNGLVQLFRRVGVTELEARGGTLPPASQRWDRILQASGMKRAKPSPTSAQTPDQASLHA
SLHAFADSLERDLDAPSPMHEGDQTRAS*

= ASAA, SNGG, or SHDD

= RVDs

= His-tag

Protein sequence of TAL$_{RASSF2}$-P300

MDPIRPRRPSPARELLPGPQPDRVQPTADRGVSAPAGSPLDGLPARRTVSRTRLPSPPAPSPAFSAGSFSDLLRPFD
PSLLDTSLLDSMPAVGTPHTAAAPAEWDEAQSALRAADDPPPTVRVAVTAARPPRAKPAPRRRAAQPSDASPAAQVD
LRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVTYQHIITALPEATHEDIVGVGKQ
WSGARALEALLTDAGELRGPPLQLDTGQLVKIAKRGGVTAMEAVHASRNALTGAPLN

LTPDQVVAIAxxxxGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNNGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG

LTPDQVVAIASNGGGKQALE

SIVAQLSRPDPALAALTNDHLVALACLGGRPAMDAVKKGLPHAPELIRRVNRRIGE

RTSHRVADYAQVVRVLEFFQCHSHPAYAFDEAMTQFGMSRNGLVQLFRRVGVTELEARGGTLPPASQRWDRILQASG

MKRAKPSPTSAQTPDQASLHASPKKKRKVEASGSG

IFKPEELRQALMPTLEALYRQDPESLPFRQPVDPQLLGIPDY

FDIVKSPMDLSTIKRKLDTGQYQEPWQYVDDIWLMFNNAWLYNRKTSRVYKYCSKLSEVF

EQEIDPVMQSLGYCCGRKLEFSPQTLCCYGKQLCTIPRDATYYSYQNRYHFCEKCFNEIQ

GESVSLGDDPSQPQTTINKEQFSKRKNDTLDPELFVECTECGRKMHQICVLHHEIIWPAG

FVCDGCLKKSARTRKENKFSAKRLPSTRLGTFLENRVNDFLRRQNHPESGEVTVRVVHAS

DKTVEVKPGMKARFVDSGEMAESFPYRTKALFAFEEIDGVDLCFFGMHVQEYGSDCPPPN

QRRVYISYLDSVHFFRPKCLRTAVYHEILIGYLEYVKKLGYTTGHIWACPPSEGDDYIFH

CHPPDQKIPKPKRLQEWYKKMLDKAVSERIVHDYKDIFKQATEDRLTSAKELPYFEGDFW

PNVLEESIKELEQEEEERKREENTSNESTDVTKGDSKNAKKKNNKKTSKNKSSLSRGNKK

KPGMPNVSNDLSQKLYATMEKHKEVFFVIRLIAGPAANSLPPIVDPDPLIPCDLMDGRDA

FLTLARDKHLEFSSLRRAQWSTMCMLVELHTQSQDYPYDVPDYA*

= ASAA or SNGG

= RVDs

= NLS

= p300 histone acetyltransferase

= HA tag

Sequences of oligomers used in this study

Realtime PCR

RASSF2 F 5’-AGAATGGACTACAGCCACCAAAC-3’

RASSF2 R 5’-CACAATGAACTCGTCTTCTTCCTC-3’

GAPDH F 5’-CCTGTTCGACAGTCAGCCG-3’

GAPDH R 5’-CGACCAAATCCGTTGACTCC-3’

EMSA

For TAL_{Dcm}

F 5'-CCGCGGCCGCTCX(C or 5mC)TGGTATATCCCCCGAGGAGTTCG

(6-Carboxyfluorescein was fused to 5'end.)

R 5'- CGAACTCCTCGGGGGATATACCAGGAGCGGCCGCGG

For TAL_{RASSF2}

F 5'-CTCAGCAGTX(C or 5mC)GCCTTTCTTCCTCCCCTTCGTTAGC

(6-Carboxyfluorescein was fused to 5'end.)

R 3'- GCTAACGAAGGGGAGGAAGAAAGGX(C or 5mC)GACTGCTGAG

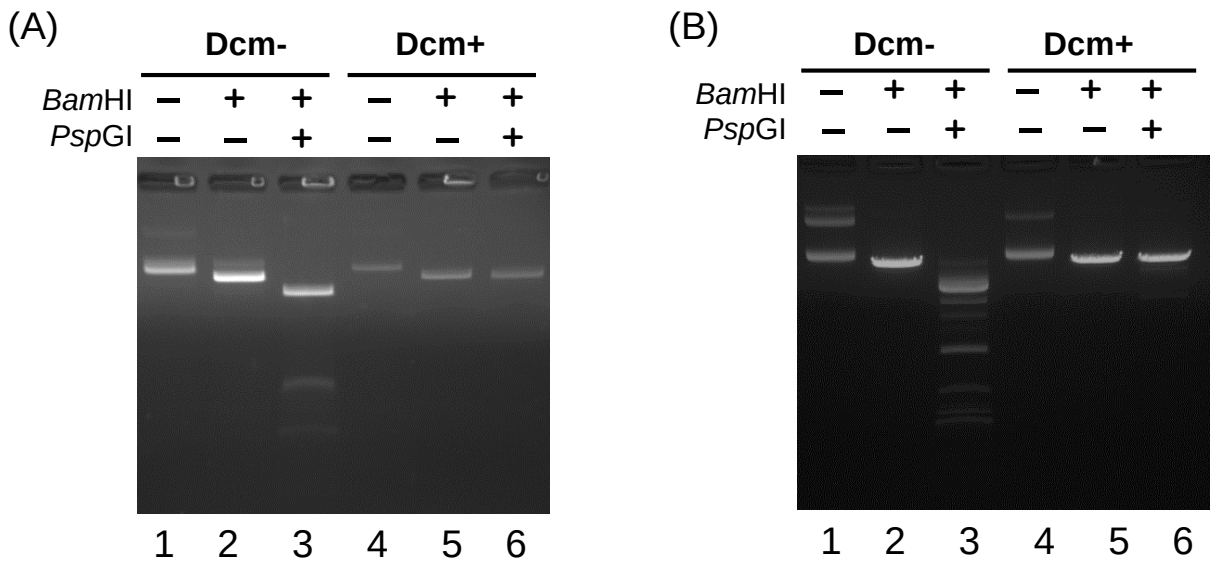


Figure S1. Methylation status confirmation of the reporter vectors. *Bam*HI and *Psp*GI digestion of the pH3U3 reporter vectors (A) or the 3×TAL_{Dcm}/pGL3 reporter vectors (B). Plasmids were digested by *Bam*HI for 1 h at 37 °C for linearization. Then they were digested by *Psp*GI for 1 h at 75 °C. Lanes 1-3; plasmids extracted from the Dcm- strain. Lanes 4-6; plasmids extracted from the Dcm+ strain.

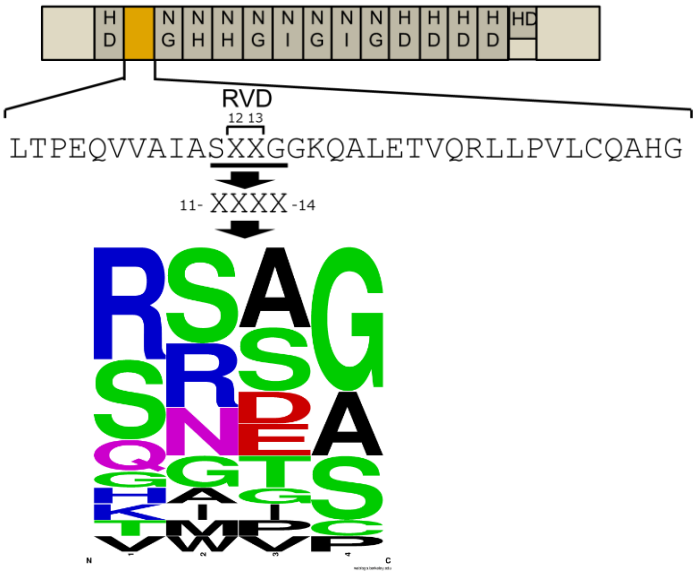
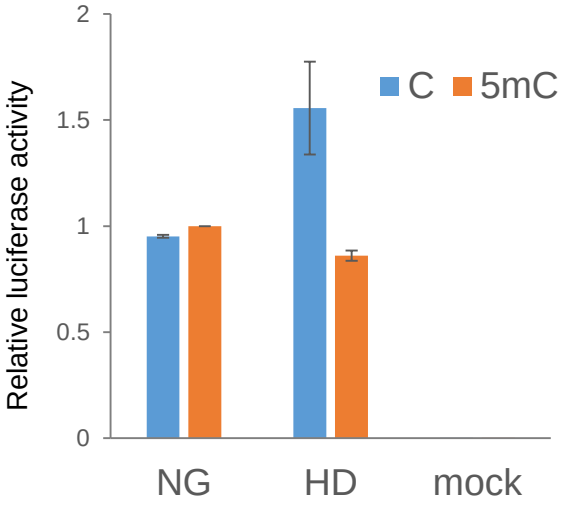


Figure S2. Sequence logo of the RVD and their neighboring amino acid residues of TALE repeats selected from B1H screening. Any pattern was not identified from the logo.

(A)



(B)

Repeat 2	K_d (nM)[a]		Relative K_d (C/5mC)
	C	5mC	
NG	95 ± 23	74 ± 36	1.2
HD	43 ± 11	105 ± 18	0.4

[a] Determined by EMSA

Figure S3. (A) Luciferase reporter activities of TAL_{Dcm} having RVD "NG" or "HD" at repeat 2 position. Luciferase activities were normalized to that of RVD "NG" for the 5mC reporter. (B) K_d values of $11.5TAL_{Dcm}$ having RVD "NG" or "HD" at repeat 2 position to dsDNA containing a TAL_{Dcm} binding site with unmethylated or methylated cytosine. The apparent dissociation constant K_d (nM) determined by EMSA, are shown.

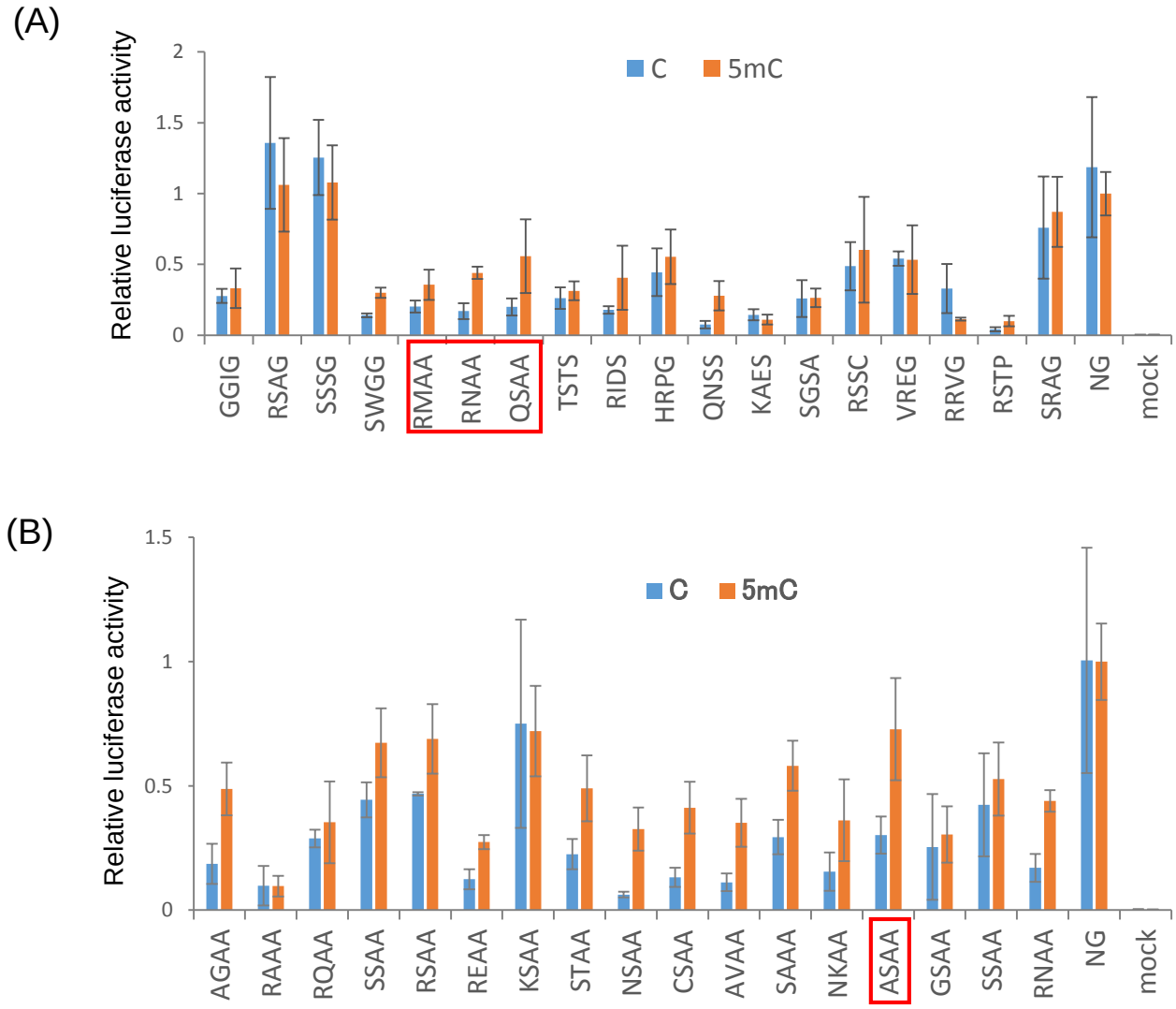


Figure S4. Base specificities of the repeats selected from B1H screening. (A) Luciferase reporter activities of TAL_{Dcm} having the repeats identified from "XXXX" library for reporter vectors with C- or 5mC-binding sites. The red square indicates "XXAA" mutants. (B) Luciferase reporter activities of TAL_{Dcm} having the repeats identified from "XXAA" library for reporter vectors with C- or 5mC-binding sites. The red square indicates the "ASAA" repeat. The activities were normalized to that of RVD "NG" in the 5mC reporter.

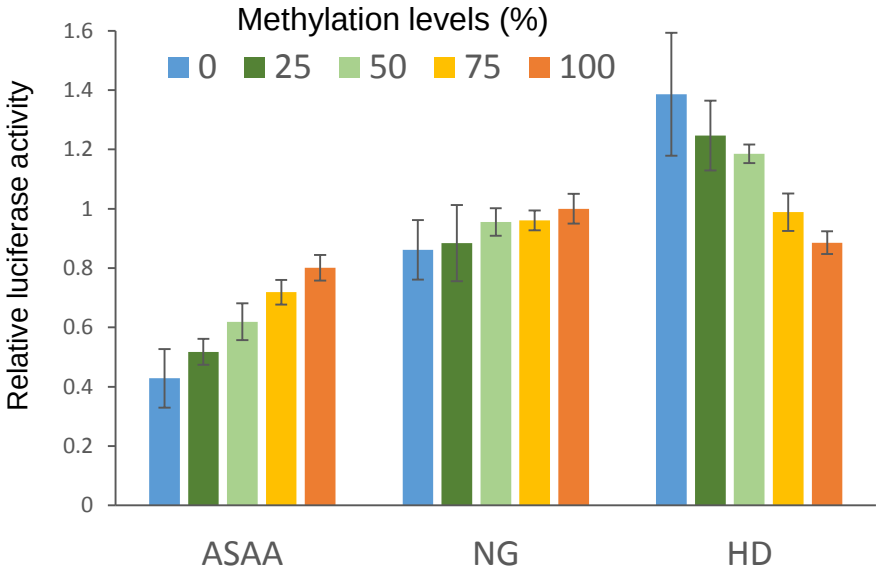
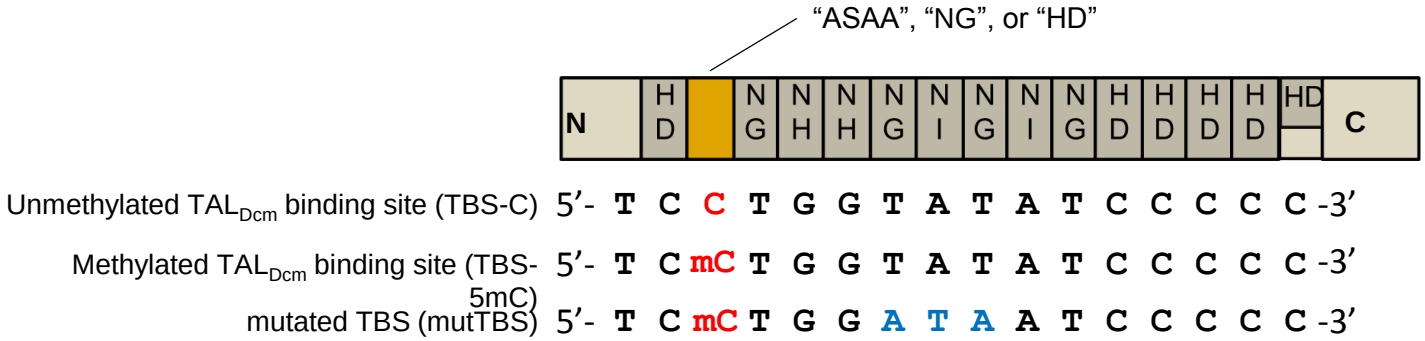


Figure S5. Methylation-dependent transcriptional activation of TAL_{Dcm} with the "ASAA" repeat. Luciferase activities of TAL_{Dcm} for reporter vectors with different methylation percentages were examined. Each luciferase activities were normalized to that of TAL_{Dcm} with RVD "NG" for the 100 % 5mC reporter. Reporter vectors were prepared by mixing C and 5mC reporters in the indicated ratios.

(A)



(B)

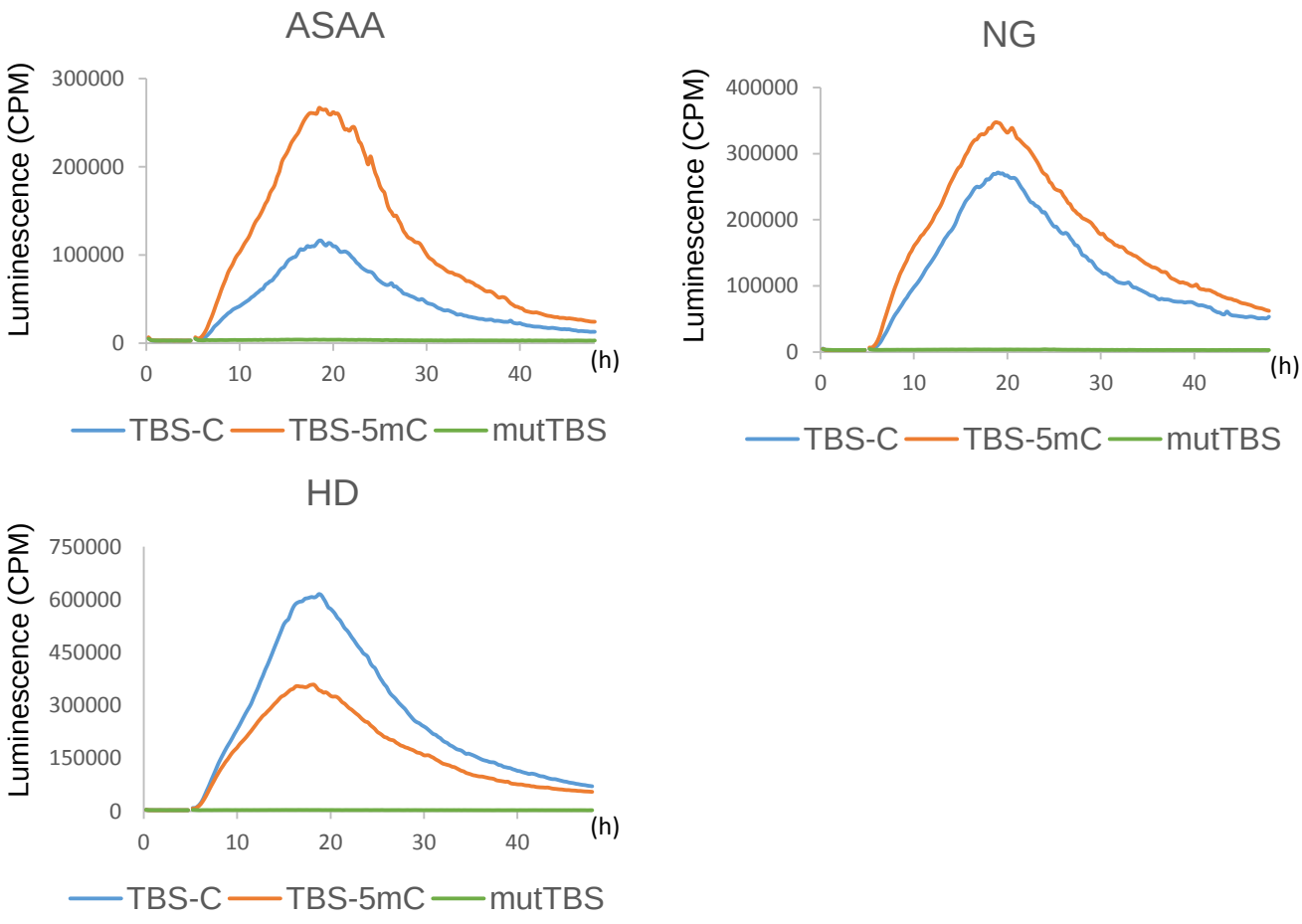
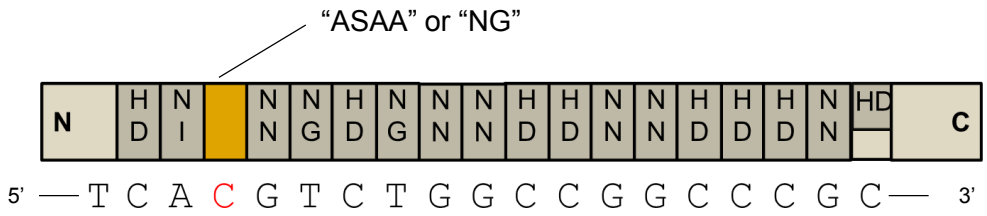
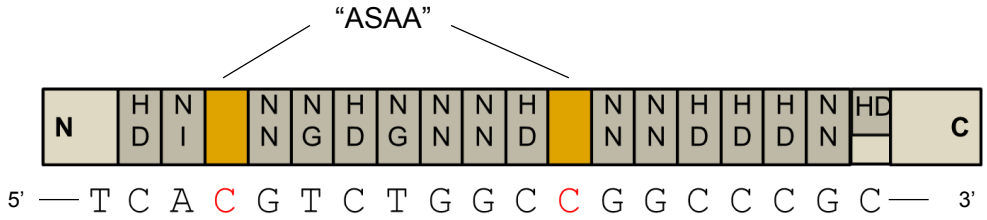


Figure S6. (A) Original and mutated TAL_{Dcm} binding sites (TBS) in the luciferase reporter vectors. Mutated bases are colored in blue. (B) Real-time monitoring of luciferase reporter activities of TAL_{Dcm} having the "ASAA", the RVD "NG", or the RVD "HD" repeats for reporter vectors with TBS-C, TBS-5mC, or mutTBS. Luciferin was added 5 h after transfection in the culture medium.

(A)



(B)



(C)

K_d values of TAL_{RASSF2} having two "ASAA" repeats

	K_d (μM)[a]		Relative K_d (C/5mC)
	C x 2	mC x 2	
ASAA x 2	2.1 ± 0.1	2.0 ± 0.1	1.0

[a] Determined by EMSA

Figure S7. The design of TALE that targets the sequence containing (A) one 5mC or (B) two 5mCs. C or 5mC was used at the cytosine colored in red for EMSA. (C) Apparent dissociation constant K_d (μM) of a TALE containing two "ASAA" repeats to the target sequences containing C or 5mC at the two CpG sites shown in (B).

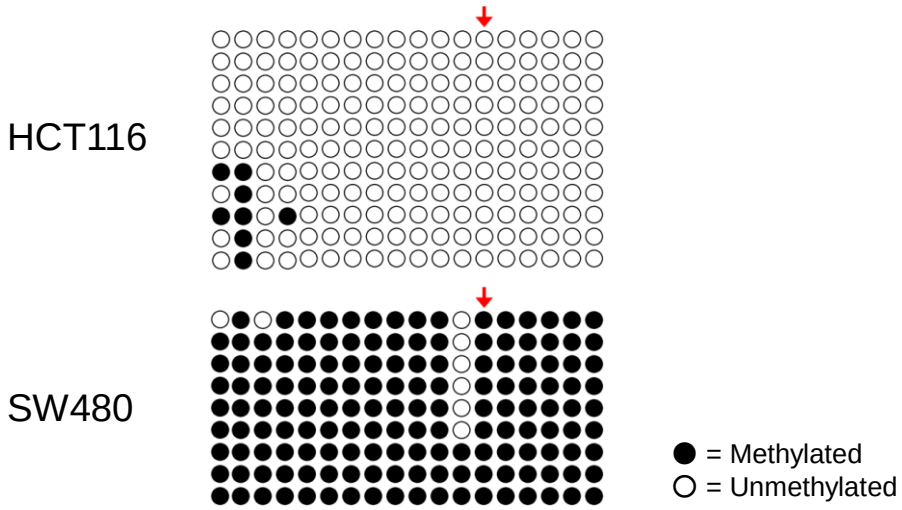


Figure S8. Methylation status of the target region including TAL_{RASSF2} binding site. Data were obtained by bisulfite sequencing of the genome prepared from HCT116 or SW480 cells (HCT116; n=10, SW480; n=9). Red arrows indicate the CpG position targeted by TAL_{RASSF2}.

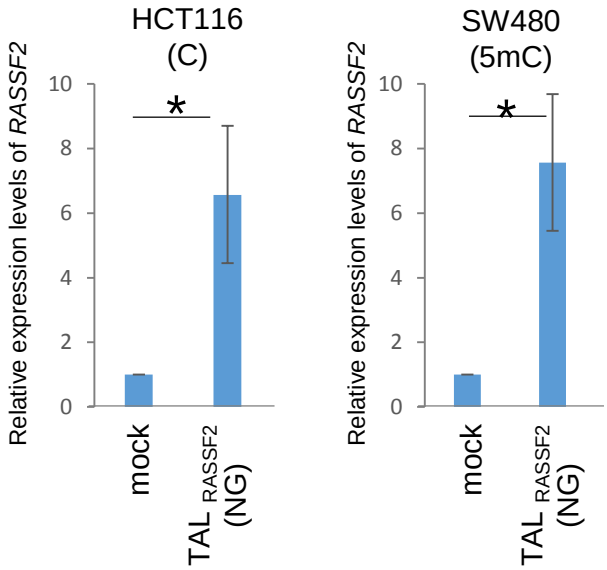


Figure S9. Methylation status-independent activation of *RASSF2* by TAL_{RASSF2}-p300. Twenty four h after TAL_{RASSF2}-p300("NG") transfection, relative expression level of *RASSF2* mRNA, in HCT116 and SW480 cells were examined by RT-qPCR. Expression levels were normalized to those of GAPDH. Data are expressed as means ± SD. n = 3; *P < 0.05.

Table S1 K_d values of TAL_{RASSF2} having "ASAA" or RVD "NG" at repeat1

Repeat 1	K_d (nM)[a]		Relative K_d (C/5mC)
	C	5mC	
ASAA	1128 ± 81	722 ± 118	1.6
NG	510 ± 68	412 ± 14	1.2

[a] Determined by EMSA

References

[1] R. R. Beerli, D. J. Segal, B. Dreier and C. F. Barbas, *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 14628.

[2] F. Zhang, L. Cong, S. Lodato, S. Kosuri, G. M. Church and P. Arlotta, *Nature Biotechnol.*, 2011, **29**, 149.

[3] S. Tsuji, S. Futaki and M. Imanishi, *Biochem. Biophys. Res. Commun.*, 2013, **441**, 262.

[4] M. B. Noyes, X. Meng, A. Wakabayashi, S. Sinha, M. H. Brodsky and S. A. Wolfe, *Nucleic Acids Res.*, 2008, **36**, 2547.

[5] X. D. Meng, M. H. Brodsky and S. A. Wolfe, *Nature Biotechnol.*, 2005, **23**, 988.

[6] M. B. Noyes, *Methods Mol. Biol.*, 2012, **786**, 79.

[7] T. D. Schneider and R. M. Stephens, *Nucleic Acids Res.*, 1990, **18**, 6097.

[8] Y. Kumaki, M. Oda and M. Okano, *Nucleic Acids Res.*, 2008, **36**, W170.