(12) STANDARD PATENT (11) Application No. **AU 2014308713 B2**
(19) AUSTRALIAN PATENT OFFICE

(54) Title
Engineered transcription activator-like effector (TALE) domains and uses thereof

(51) International Patent Classification(s)
C12Q 1/68 (2006.01)

(21) Application No: **2014308713** (22) Date of Filing: **2014.08.22**

(87) WIPO No: **WO15/027134**

(30) Priority Data

(31) Number	(32) Date	(33) Country
14/320,519	**2014.06.30**	**US**
61/868,846	**2013.08.22**	**US**

(43) Publication Date: **2015.02.26**
(44) Accepted Journal Date: **2020.10.08**

(71) Applicant(s)
President and Fellows of Harvard College

(72) Inventor(s)
Liu, David R.;Guilinger, John Paul;Pattanayak, Vikram

(74) Agent / Attorney
Pizzeys Patent and Trade Mark Attorneys Pty Ltd, GPO Box 1374, BRISBANE, QLD, 4001, AU

(56) Related Art
WO 2012138927 A2

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 February 2015 (26.02.2015)

(10) International Publication Number
WO 2015/027134 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/US2014/052231

(22) International Filing Date:
22 August 2014 (22.08.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/868,846 22 August 2013 (22.08.2013) US
14/320,519 30 June 2014 (30.06.2014) US

(71) Applicant: **PRESIDENT AND FELLOWS OF HARVARD COLLEGE** [US/US]; 17 Quincy Street, Cambridge, MA 02138 (US).

(72) Inventors: **LIU, David, R.**; 3 Whitman Circle, Lexington, MA 02420 (US). **GUILINGER, John, Paul**; 340 Bear Cub Drive, Ridgway, CO 81432 (US). **PATTANAYAK, Vikram**; 123 Holden Green, Cambridge, MA 02138 (US).

(74) Agent: **BAKER, C., Hunter**; Wolf, Greenfield & Sacks, P.C., 600 Atlantic Avenue, Boston, MA 02210-2206 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

WO 2015/027134 A1

(54) **Title**: ENGINEERED TRANSCRIPTION ACTIVATOR-LIKE EFFECTOR (TALE) DOMAINS AND USES THEREOF

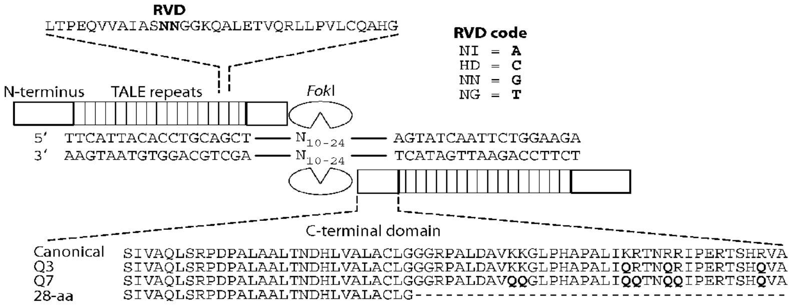

RVD
LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
RVD code
NI = A
HD = C
NN = G
NG = T
N-terminus
TALE repeats
FokI
5' TTCATTACACCTGCAGCT — N_{10-24} — AGTATCAATTCTGGAAGA
3' AAGTAATGTGGACGTCGA — N_{10-24} — TCATAGTTAAGACCTTCT
C-terminal domain
Canonical SIVAQLSRPDPALAALTNDHLVALACLGGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVA
Q3 SIVAQLSRPDPALAALTNDHLVALACLGGGRPALDAVKKGLPHAPALIQRTNQRIPERTSHQVA
Q7 SIVAQLSRPDPALAALTNDHLVALACLGGGRPALDAVQQGLPHAPALIQQTNQQIPERTSHQVA
28-aa SIVAQLSRPDPALAALTNDHLVALACLG----------------------------------

Fig. 1A

(57) **Abstract**: Engineered transcriptional activator-like effectors (TALEs) are versatile tools for genome manipulation with applications in research and clinical contexts. One current drawback of TALEs is their tendency to bind and cleave off-target sequence, which hampers their clinical application and renders applications requiring high-fidelity binding unfeasible. This disclosure provides engineered TALE domains and TALEs comprising such engineered domains, e.g., TALE nucleases (TALENs), TALE transcriptional activators, TALE transcriptional repressors, and TALE epigenetic modification enzymes, with improved specificity and methods for generating and using such TALEs.

ENGINEERED TRANSCRIPTION ACTIVATOR-LIKE EFFECTOR (TALE) DOMAINS AND USES THEREOF

RELATED APPLICATION

[0001] This application claims priority under 35 U.S.C. § 365(c) to U.S. application, U.S.S.N. 14/320,519, filed June 30, 2014, and also claims priority under 35 U.S.C. § 119(e) to U.S. provisional patent application, U.S.S.N. 61/868,846, filed August 22, 2013, each of which is incorporated herein by reference.

GOVERNMENT SUPPORT

[0002] This invention was made with U.S. Government support under grant HR0011-11-2-0003 and N66001-12-C-4207, awarded by the Defense Advanced Research Projects Agency; grant T32GM007753, awarded by the National Institute of General Medical Sciences; and grant DP1 GM105378 awarded by the National Institutes of Health. The U.S. Government has certain rights in this invention.

BACKGROUND OF THE INVENTION

[0003] Transcription activator-like effector nucleases (TALENs) are fusions of the FokI restriction endonuclease cleavage domain with a DNA-binding transcription activator-like effector (TALE) repeat array. TALENs can be engineered to specifically bind and cleave a desired target DNA sequence, which is useful for the manipulation of nucleic acid molecules, genes, and genomes *in vitro* and *in vivo*. Engineered TALENs are useful in the context of many applications, including, but not limited to, basic research and therapeutic applications. For example, engineered TALENs can be employed to manipulate genomes in the context of the generation of gene knockouts or knock-ins via induction of DNA breaks at a target genomic site for targeted gene knockout through non-homologous end joining (NHEJ) or targeted genomic sequence replacement through homology-directed repair (HDR) using an exogenous DNA template, respectively. TALENs are thus useful in the generation of genetically engineered cells, tissues, and organisms.

[0004] TALENs can be designed to cleave any desired target DNA sequence, including naturally occurring and synthetic sequences. However, the ability of TALENs to distinguish target sequences from closely related off-target sequences has not been studied in depth. Understanding this ability and the parameters affecting it is of importance for the design of TALENs having the desired level of specificity and also for choosing unique target sequences to be cleaved, *e.g.*, in order to minimize the chance of undesired off-target cleavage.

SUMMARY OF THE INVENTION

[0005] TALENs are versatile tools for the manipulation of genes and genomes *in vitro* and *in vivo*, as they can be designed to bind and cleave virtually any target sequence within a nucleic acid molecule. For example, TALENs can be used for the targeted deletion of a DNA sequence within a cellular genome via induction of DNA breaks that are then repaired by the cellular DNA repair machinery through non-homologous end joining (NHEJ). TALENs can also be used for targeted sequence replacement in the presence of a nucleic acid comprising a sequence to be inserted into a genomic sequence via homology-directed repair (HDR). As TALENs can be employed to manipulate the genomes of living cells, the resulting genetically modified cells can be used to generate transgenic cell or tissue cultures and organisms.

[0006] In scenarios where a TALEN is employed for the targeted cleavage of a DNA sequence in the context of a complex sample, *e.g.*, in the context of a genome, it is often desirable for the TALEN to bind and cleave the specific target sequence only, with no or only minimal off-target cleavage activity (see, *e.g.*, PCT Application Publication WO2013/066438 A2, the entire contents of which are incorporated herein by reference). In some embodiments, an ideal TALEN would specifically bind only its intended target sequence and have no off-target activity, thus allowing the targeted cleavage of a single sequence, *e.g.*, a single allele of a gene of interest, in the context of a whole genome.

[0007] Some aspects of this disclosure are based on the recognition that the tendency of TALENs to cleave off-target sequences and the parameters affecting the propensity of off-target TALEN activity are poorly understood. The work presented here provides a better understanding of the structural parameters that result in TALEN off-target activity. Methods and systems for the generation of engineered TALENs having no or minimal off-target activity are provided herein, as are engineered TALENs having increased on-target cleavage efficiency and minimal off-target activity. It will be understood by those of skill in the art that the strategies, methods, and reagents provided herein for decreasing non-specific or off-target DNA binding by TALENs are applicable to other DNA-binding proteins as well. In particular, the strategies for modifying the amino acid sequence of DNA-binding proteins for reducing unspecific binding to DNA by substituting cationic amino acid residues with amino acid residues that are not cationic, are uncharged, or are anionic at physiological pH, can be used to decrease the specificity of, for example, other TALE effector proteins, engineered zinc finger proteins (including zinc finger nucleases), and Cas9 proteins.

[0008] Some aspects of this disclosure provide engineered isolated Transcription Activator-Like Effector (TALE) domains. In some embodiments, the isolated TALE domain is an N-terminal TALE domain and the net charge of the isolated N-terminal domain is less than the net charge of the canonical N-terminal domain (SEQ ID NO: 1) at physiological pH. In some embodiments, the isolated TALE domain is a C-terminal TALE domain and the net charge of the C-terminal domain is less than the net charge of the canonical C-terminal domain (SEQ ID NO: 22) at physiological pH. In some embodiments, the isolated TALE domain is an N-terminal TALE domain and the binding energy of the N-terminal domain to a target nucleic acid molecule is smaller than the binding energy of the canonical N-terminal domain (SEQ ID NO: 1). In some embodiments, the isolated TALE domain is a C-terminal TALE domain and the binding energy of the C-terminal domain to a target nucleic acid molecule is smaller than the binding energy of the canonical C-terminal domain (SEQ ID NO: 22). In some embodiments, the net charge of the C-terminal domain is less than or equal to +6, less than or equal to +5, less than or equal to +4, less than or equal to +3, less than or equal to +2, less than or equal to +1, less than or equal to 0, less than or equal to -1, less than or equal to -2, less than or equal to -3, less than or equal to -4, or less than or equal to -5. In some embodiments, the C-terminal domain comprises an amino acid sequence that differs from the canonical C-terminal domain sequence in that at least one cationic amino acid residue of the canonical C-terminal domain sequence is replaced with an amino acid residue that exhibits no charge or a negative charge at physiological pH. In some embodiments, the N-terminal domain comprises an amino acid sequence that differs from the canonical N-terminal domain sequence in that at least one cationic amino acid residue of the canonical N-terminal domain sequence is replaced with an amino acid residue that exhibits no charge or a negative charge at physiological pH. In some embodiments, at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, or at least 15 cationic amino acid(s) in the isolated TALE domain is/are replaced with an amino acid residue that exhibits no charge or a negative charge at physiological pH. In some embodiments, the at least one cationic amino acid residue is arginine (R) or lysine (K). In some embodiments, the amino acid residue that exhibits no charge or a negative charge at physiological pH is glutamine (Q) or glycine (G). In some embodiments, at least one lysine or arginine residue is replaced with a glutamine residue. In some embodiments, the C-terminal domain comprises one or more of the following amino acid replacements:K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q . In some embodiments, the C-terminal domain comprises a Q3 variant sequence (K788Q, R792Q, K801Q). In some embodiments, the C-terminal domain comprises a Q7 variant sequence (K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q). In some embodiments, the N-terminal domain is a truncated version of the canonical N-terminal domain. In some embodiments, wherein the C-terminal domain is a truncated version of the canonical C-terminal domain. In some embodiments, the truncated domain comprises less than 90%, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, or less than 25% of the residues of the canonical domain. In some embodiments, the truncated C-terminal domain comprises less than 60, less than 50, less than 40, less than 30, less than 29, less than 28, less than 27, less than 26, less than 25, less than 24, less than 23, less than 22, less than 21, or less than 20 amino acid residues. In some embodiments, the truncated C-terminal domain comprises 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 residues. In some embodiments, the isolated TALE domain is comprised in a TALE molecule comprising the structure [N-terminal domain]-[TALE repeat array]-[C-terminal domain]-[effector domain]; or [effector domain]-[N-terminal domain]-[TALE repeat array]-[C-terminal domain]. In some embodiments, the effector domain comprises a nuclease domain, a transcriptional activator or repressor domain, a recombinase domain, or an epigenetic modification enzyme domain. In some embodiments, the TALE molecule binds a target sequence within a gene known to be associated with a disease or disorder.

[0009] Some aspects of this disclosure provide Transcription Activator-Like Effector Nucleases (TALENs) having a modified net charge and/or a modified binding energy for binding their target nucleic acid sequence as compared to canonical TALENs. Typically, the inventive TALENs include (a) a nuclease cleavage domain; (b) a C-terminal domain conjugated to the nuclease cleavage domain; (c) a TALE repeat array conjugated to the C-terminal domain; and (d) an N-terminal domain conjugated to the TALE repeat array. In some embodiments, (i) the net charge on the N-terminal domain at physiological pH is less than the net charge on the canonical N-terminal domain (SEQ ID NO: 1) at physiological pH; and/or (ii) the net charge of the C-terminal domain at physiological pH is less than the net charge of the canonical C-terminal domain (SEQ ID NO: 22) at physiological pH. In some embodiments, (i) the binding energy of the N-terminal domain to a target nucleic acid molecule is less than the binding energy of the canonical N-terminal domain (SEQ ID NO:

1); and/or (ii) the binding energy of the C-terminal domain to a target nucleic acid molecule is less than the binding energy of the canonical C-terminal domain (SEQ ID NO: 22). In some embodiments, the net charge on the C-terminal domain at physiological pH is less than or equal to +6, less than or equal to +5, less than or equal to +4, less than or equal to +3, less than or equal to +2, less than or equal to +1, less than or equal to 0, less than or equal to -1, less than or equal to -2, less than or equal to -3, less than or equal to -4, or less than or equal to -5. In some embodiments, the N-terminal domain comprises an amino acid sequence that differs from the canonical N-terminal domain sequence in that at least one cationic amino acid residue of the canonical N-terminal domain sequence is replaced with an amino acid residue that does not have a cationic charge, has no charge, or has an anionic charge. In some embodiments, the C-terminal domain comprises an amino acid sequence that differs from the canonical C-terminal domain sequence in that at least one cationic amino acid residue of the canonical C-terminal domain sequence is replaced with an amino acid residue that does not have a cationic charge, has no charge, or has an anionic charge. In some embodiments, at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, or at least 15 cationic amino acid(s) is/are replaced with an amino acid residue that does not have a cationic charge, has no charge, or has an anionic charge in the N-terminal domain and/or in the C-terminal domain. In some embodiments, the at least one cationic amino acid residue is arginine (R) or lysine (K). In some embodiments, the amino acid residue that replaces the cationic amino acid is glutamine (Q) or glycine (G). Positively charged residues in the C-terminal domain that can be replaced according to aspects of this disclosure include, but are not limited to, arginine (R) residues and lysine (K) residues, *e.g.*, R747, R770, K777, K778, K788, R789, R792, R793, R797, and R801 in the C-terminal domain (see. *e.g.*, SEQ ID NO: 22, the numbering refers to the position of the respective residue in the full-length TALEN protein, the equivalent positions for the C-terminal domain as provide in SEQ ID NO: 22 are R8, R30, K37, K38, K48, R49, R52, R53, R57, R61). Positively charged residues in the N-terminal domain that can be replaced according to aspects of this disclosure include, but are not limited to, arginine (R) residues and lysine (K) residues, *e.g.*, K57, K78, R84, R97, K110, K113, and R114 (see, *e.g.*, SEQ ID NO: 1). In some embodiments, at least one lysine or arginine residue is replaced with a glutamine residue. In some embodiments, the C-terminal domain comprises one or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the C-terminal domain comprises a Q3 variant sequence (K788Q, R792Q, R801Q). In some embodiments, the C-terminal domain comprises a Q7 variant sequence (K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q). In some embodiments, the N-terminal domain is a truncated version of the canonical N-terminal domain. In some embodiments, the C-terminal domain is a truncated version of the canonical C-terminal domain. In some embodiments, the truncated domain comprises less than 90%, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, or less than 25% of the residues of the canonical domain. In some embodiments, the truncated C-terminal domain comprises less than 60, less than 50, less than 40, less than 30, less than 29, less than 28, less than 27, less than 26, less than 25, less than 24, less than 23, less than 22, less than 21, or less than 20 amino acid residues. In some embodiments, the truncated C-terminal domain comprises 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 residues. In some embodiments, the nuclease cleavage domain is a FokI nuclease domain. In some embodiments, the FokI nuclease domain comprises a sequence as provided in SEQ ID NOs: 26-30. In some embodiments, the TALEN is a monomer. In some embodiments, the TALEN monomer dimerizes with another TALEN monomer to form a TALEN dimer. In some embodiments, the dimer is a heterodimer. In some embodiments, the TALEN binds a target sequence within a gene known to be associated with a disease or disorder. In some embodiments, the TALEN cleaves the target sequence upon dimerization. In some embodiments, the disease being treated or prevented is HIV infection or AIDS, or a proliferative disease. In some embodiments, the TALEN binds a CCR5 (C-C chemokine receptor type 5) target sequence in the treatment or prevention of HIV infection or AIDS. In some embodiments, the TALEN binds an ATM (ataxia telangiectasia mutated) target sequence. In some embodiments, the TALEN binds a VEGFA (Vascular endothelial growth factor A) target sequence.

[0010] Some aspects of this disclosure provide compositions comprising a TALEN described herein, *e.g.*, a TALEN monomer. In some embodiments, the composition comprises the inventive TALEN monomer and a different inventive TALEN monomer that form a heterodimer , wherein the dimer exhibits nuclease activity. In some embodiments, the composition is a pharmaceutical composition.

[0011] Some aspects of this disclosure provide a composition comprising a TALEN provided herein. In some embodiments, the composition is formulated to be suitable for contacting with a cell or tissue *in vitro*. In some embodiments, the pharmaceutical composition comprises an effective amount of the TALEN for cleaving a target sequence, *e.g.*, in a cell or in a tissue *in vitro* or *ex vivo*. In some embodiments, the TALEN binds a target sequence within a gene of interest, *e.g.*, a target sequence within a gene known to be associated with a disease or disorder, and the composition comprises an effective amount of the TALEN for alleviating a sign and/or symptom associated with the disease or disorder. Some aspects of this disclosure provide a pharmaceutical composition comprising a TALEN provided herein and a pharmaceutically acceptable excipient. In some embodiments, the pharmaceutical composition is formulated for administration to a subject. In some embodiments, the pharmaceutical composition comprises an effective amount of the TALEN for cleaving a target sequence in a cell in the subject. In some embodiments, the TALEN binds a target sequence within a gene known to be associated with a disease or disorder, and the composition comprises an effective amount of the TALEN for alleviating a sign and/or symptom associated with the disease or disorder.

[0012] Some aspects of this disclosure provide methods of cleaving a target sequence in a nucleic acid molecule using a TALEN provided herein. In some embodiments, the method comprises contacting a nucleic acid molecule comprising the target sequence with an inventive TALEN binding the target sequence under conditions suitable for the TALEN to bind and cleave the target sequence. In some embodiments, the TALEN is provided as a monomer. In some embodiments, the inventive TALEN monomer is provided in a composition comprising a different TALEN monomer that can dimerize with the inventive TALEN monomer to form a heterodimer having nuclease activity. In some embodiments, the inventive TALEN is provided in a pharmaceutical composition. In some embodiments, the target sequence is in the genome of a cell. In some embodiments, the target sequence is in a subject. In some embodiments, the method comprises administering a composition, *e.g.*, a pharmaceutical composition, comprising the TALEN to the subject in an amount sufficient for the TALEN to bind and cleave the target site.

[0013] Some aspects of this disclosure provide methods of preparing engineered TALENs. In some embodiments, the method comprises replacing at least one amino acid in the canonical N-terminal TALEN domain and/or the canonical C-terminal TALEN domain with an amino acid having no charge or a negative charge as compared to the amino acid being replaced at physiological pH; and/or truncating the N-terminal TALEN domain and/or the C-terminal TALEN domain to remove a positively charged fragment; thus generating an engineered TALEN having an N-terminal domain and/or a C-terminal domain of decreased net charge at physiological pH. In some embodiments, the at least one amino acid being replaced comprises a cationic amino acid or an amino acid having a positive charge at physiological pH. Positively charged residues in the C-terminal domain that can be replaced according to aspects of this disclosure include, but are not limited to, arginine (R) residues and lysine (K) residues, *e.g.*, R747, R770, K777, K778, K788, R789, R792, R793, R797, and R801 in the C-terminal domain. Positively charged residues in the N-terminal domain that can be replaced according to aspects of this disclosure include, but are not limited to, arginine (R) residues and lysine (K) residues, *e.g.*, K57, K78, R84, R97, K110, K113, and R114. In some embodiments, the amino acid replacing the at least one amino acid is a cationic amino acid or a neutral amino acid. In some embodiments, the truncated N-terminal TALEN domain and/or the truncated C-terminal TALEN domain comprises less than 90%, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, or less than 25% of the residues of the respective canonical domain. In some embodiments, the truncated C-terminal domain comprises less than 60, less than 50, less than 40, less than 30, less than 29, less than 28, less than 27, less than 26, less than 25, less than 24, less than 23, less than 22, less than 21, or less than 20 amino acid residues. In some embodiments, the truncated C-terminal domain comprises 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 amino acid residues. In some embodiments, the method comprises replacing at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, or at least 15 amino acids in the canonical N-terminal TALEN domain and/or in the canonical C-terminal TALEN domain with an amino acid having no charge or a negative charge at physiological pH. In some embodiments, the amino acid being replaced is arginine (R) or lysine (K). In some embodiments, the amino acid residue having no charge or a negative charge at physiological pH is glutamine (Q) or glycine (G). In some embodiments, the method comprises replacing at least one lysine or arginine residue with a glutamine residue.

[0014] Some aspects of this disclosure provide kits comprising an engineered TALEN as provided herein, or a composition (*e.g.*, a pharmaceutical composition) comprising such a TALEN. In some embodiments, the kit comprises an excipient and instructions for contacting the TALEN with the excipient to generate a composition suitable for contacting a nucleic acid with the TALEN. In some embodiments, the excipient is a pharmaceutically acceptable excipient.

[0015] The summary above is meant to illustrate, in a non-limiting manner, some of the embodiments, advantages, features, and uses of the technology disclosed herein. Other embodiments, advantages, features, and uses of the technology disclosed herein will be apparent from the Detailed Description, the Drawings, the Examples, and the Claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] **Figure 1.** TALEN architecture and selection scheme. (A) Architecture of a TALEN. A TALEN monomer contains an N-terminal domain followed by an array of TALE repeats, a C terminal domain, and a FokI nuclease cleavage domain. The 12th and 13th amino acids (the RVD, red) of each TALE repeat recognize a specific DNA base pair. Two different TALENs bind their corresponding half-sites, allowing FokI dimerization and DNA cleavage. The C-terminal domain variants used in this study are shown. (B) A single-stranded library of DNA oligonucleotides containing partially randomized left half-site (L), spacer (S), right half-site (R) and constant region (thick black line) was circularized, then concatemerized by rolling circle amplification. The resulting DNA libraries were incubated with an *in vitro*-translated TALEN of interest. Cleaved library members were blunted and ligated to adapter #1. The ligation products were amplified by PCR using one primer consisting of adapter #1 and the other primer consisting of adapter #2–constant sequence, which anneals to the constant regions. Amplicons 1½ target-sequence cassettes in length were isolated by gel purification and subjected to high-throughput DNA sequencing and computational analysis.

[0017] **Figure 2.** *In vitro* selection results. The fraction of sequences surviving selection (grey) and before selection (black) are shown for CCR5A TALENs (A) and ATM TALENs (B) as a function of the number of mutations in both half-sites. (C) Specificity scores for the L18+R18 CCR5A TALEN at all positions in the target half-sites plus a single flanking position. The colors range from a maximum specificity score of 1.0 to white (no specificity, score of 0) to a maximum negative score of – 1.0. Boxed bases represent the intended target base. (D) Same as (C) for the L18+R18 ATM TALEN. (E) Enrichment values from the selection of L13+R13 CCR5B TALEN for 16 mutant DNA sequences (mutations in red) relative to on-target DNA (OnB). (F) Correspondence between discrete *in vitro* TALEN cleavage efficiency (cleaved DNA as a fraction of total DNA) for the sequences listed in (E) normalized to on-target cleavage (= 1) versus their enrichment values in the selection normalized to the on-target enrichment value (= 1). (G) Discrete assays of on-target and off-target sequences used in (F) as analyzed by PAGE.

[0018] **Figure 3.** Cellular modification induced by TALENs at on-target and predicted off-target genomic sites. (A) For cells treated with either no TALEN or CCR5A

TALENs containing heterodimeric EL/KK, heterodimeric ELD/KKR, or the homodimeric (Homo) FokI variants, cellular modification rates are shown as the percentage of observed insertions or deletions (indels) consistent with TALEN cleavage relative to the total number of sequences for on-target (On) and predicted off-target sites (Off). (B) Same as (A) for ATM TALENs. (C) Examples of modified sequences at the on-target site and off-target sites for cells treated with CCR5A TALENs containing the ELD/KKR FokI domains. For each example shown, the unmodified genomic site is the first sequence, followed by the top three sequences containing deletions. The numbers in parentheses indicate sequencing counts and the half-sites are underlined and bolded.

[0019] **Figure 4.** Predicted off-target genomic cleavage as a function of TALEN length considering both TALEN specificity and off-target site abundance in the human genome. (A) The enrichment value of on-target (zero mutation) and off-target sequences containing one to six mutations are shown for CCR5B TALENs of varying TALE repeat array lengths. The TALENs targeted DNA sites of 32 bp (L16+R16), 29 bp (L16+R13 or L13+R16), 26 bp (L16+R10 or L13+R13 or L10+R16), 23 bp (L13+R10 or L10+R13) or 20 bp (L10+R10) in length. (B) Number of sites in the human genome related to each of the nine CCR5B on-target sequences (L10, L13, or L16 combined with R10, R13, or R16), allowing for a spacer length from 12 to 25 bps between the two half-sites. (C) For all nine CCR5B TALENs, overall genomic off-target cleavage frequency was predicted by multiplying the number of sites in the human genome containing a certain number of mutations by the enrichment value of off-target sequences containing that same number of mutations shown in (A). Because enrichment values level off at high mutation numbers likely due to the limit of sensitivity of the selection, it was necessary to extrapolate high-mutation enrichment values by fitting enrichment value as function of mutation number (Table 9). The overall predicted genomic cleavage was calculated only for mutation numbers with sites observed to occur more than once in the human genome.

[0020] **Figure 5.** *In vitro* specificity and discrete cleavage efficiencies of TALENs containing canonical or engineered C-terminal domains. (A and B) On-target enrichment values for selections of (A) CCR5A TALENs or (B) ATM TALENs containing canonical, Q3, Q7, or 28-aa C-terminal domains. (C) CCR5A on-target sequence (OnC) and double-mutant sequences with mutations in lower case. (D) ATM on-target sequence (OnA) and single-mutant sequences with mutations in lower case. (E) Discrete in vitro cleavage efficiency of DNA sequences listed in (C) with CCR5A TALENs containing either canonical or engineered Q7 C-terminal domains. (F) Same as (E) for ATM TALENs.

[0021] **Figure 6.** Specificity of engineered TALENs in human cells. The cellular modification efficiency of canonical and engineered TALENs expressed as a percentage of indels consistent with TALEN-induced modification out of total sequences is shown for the on-target CCR5A sequence and for CCR5A off-target site #5, the most highly cleaved off-target substrate tested. Cellular specificity, defined as the ratio of on-target to off-target modification, is shown below each pair of bars.

[0022] **Figure 7.** Target DNA sequences in human CCR5 and ATM genes. The target DNA sequences for the TALENs used in this study are shown in black. The N-terminal TALEN end recognizing the 5′ T for each half-site target is noted (5′) and TALENs are named according to number of base pairs targeted. TALENs targeting the CCR5 L18 and R18 shown are referred to as CCR5A TALENs while TALENs targeting the L10, L13, L16, R10, R13 or R16 half-sites shown are referred to as CCR5B TALENs.

[0023] **Figure 8.** Specificity profiles from all CCR5A TALEN selections as heat maps. Specificity scores for every targeted base pair in selections of CCR5A TALENs are shown. Specificity scores for the L18+R18 CCR5A TALEN at all positions in the target half-sites plus a single flanking position. The colors range from a maximum specificity score of 1.0 to white (score of 0, no specificity) to a maximum negative score of -1.0. Boxed bases represent the intended target base. The titles to the right indicate if the TALEN used in the selection differs from the canonical TALEN architecture, which contains a canonical C-terminal domain, wildtype N-terminal domain, and EL/KK FokI variant. Selections correspond to conditions listed in Table 2. (A) Specificity profiles of canonical, Q3, Q7, 28-aa, 32 nM canonical, 8 nM canonical, 4 nM canonical, 32 nM Q7 and 8 nM Q7 CCR5A TALEN selections. (B) Specificity profiles of 4 nM Q7, N1, N2, N3, canonical ELD/KKR, Q3 ELD/KKR, Q7 ELD/KKR and N2 ELD/KKR CCR5A TALEN selections. When not specified, TALEN concentration was 16 nM.

[0024] **Figure 9.** Specificity profiles from all CCR5A TALEN selections as bar graphs. Specificity scores for every targeted base pair in selections of CCR5A TALENs are shown. Positive specificity scores, up to complete specificity at a specificity score of 1.0, signify enrichment of that base pair over the other possibilities at that position. Negative specificity scores, down to complete antispecificity of –1.0, represents enrichment against that base pair. Specified positions were plotted as stacked bars above the X-axis (multiple specified base pairs at the same position were plotted over each other with the shortest bar in front, and not end-to-end) while anti-specified base pairs were plotted as narrow, grouped bars. The titles to the right indicate if the TALEN used in the selection differs from the canonical TALEN architecture, which contains a canonical C-terminal domain, wild-type N-terminal domain, and EL/KK FokI variant. Selections correspond to conditions listed in Table 2. (A) Specificity profiles of canonical, Q3, Q7, 28-aa, 32 nM canonical, and 8 nM canonical CCR5A TALEN selections. (B) Specificity profiles of 4 nM canonical, 32 nM Q7, 8 nM Q7, 4 nM Q7, N1, and N2 CCR5A TALEN selections. (C) Specificity profiles of N3, canonical ELD/KKR, Q3 ELD/KKR, Q7 ELD/KKR, and N2 ELD/KKR CCR5A TALEN selections. When not specified, TALEN concentration was 16 nM.

[0025] **Figure 10.** Specificity profiles from all ATM TALEN selections as heat maps. Specificity scores for every targeted base pair in selections of ATM TALENs are shown. Specificity scores for the L18+R18 ATM TALEN at all positions in the target half-sites plus a single flanking position. The colors range from a maximum specificity score of 1.0 to white (score of 0, no specificity) to a maximum negative score of -1.0. Boxed bases represent the intended target base. The titles to the right indicate if the TALEN used in the selection differs from the canonical TALEN architecture, which contains a canonical C-terminal domain, wild type N-terminal domain, and EL/KK FokI variant. Selections correspond to conditions listed in Table 2. (A) Specificity profiles of (12 nM) canonical, Q3, (12 nM) Q7, 24 nM canonical, 6 nM canonical, 3 nM canonical, 24 nM Q7, and 6 nM Q7 ATM TALEN selections. (B) Specificity profiles of N1, N2, N3, canonical ELD/KKR, Q3 ELD/KKR, Q7 ELD/KKR, and N2 ELD/KKR ATM TALEN selections. When not specified, TALEN concentration was 12 nM.

[0026] **Figure 11.** Specificity profiles from all ATM TALEN selections as bar graphs. Specificity scores for every targeted base pair in selections of ATM TALENs are shown. Positive specificity scores, up to complete specificity at a specificity score of 1.0, signify enrichment of that base pair over the other possibilities at that position. Negative specificity scores, down to complete antispecificity of –1.0, represents enrichment against that base pair. Specified positions were plotted as stacked bars above the X-axis (multiple specified base pairs at the same position were plotted over each other with the shortest bar in front, and not end-to-end) while anti-specified base pairs were plotted as narrow, grouped bars. The titles to the right indicate if the TALEN used in the selection differs from the canonical TALEN architecture, which contains a canonical C-terminal domain, wild-type N-terminal domain, and EL/KK FokI variant. Selections correspond to conditions listed in Table 2. (A) Specificity profiles of canonical, Q3, Q7, 32 nM canonical, and 8 nM canonical ATM TALEN selections. (B) Specificity profiles of 3 nM canonical, 24 nM Q7, 6 nM Q7, N1, N2, and N3 ATM TALEN selections. (C) Specificity profiles of canonical ELD/KKR, Q3

ELD/KKR, Q7 ELD/KKR, and N2 ELD/KKR ATM TALEN selections. When not specified, TALEN concentration was 12 nM.

[0027] **Figure 12.** Specificity profiles from all CCR5B TALEN selections as heat maps. Specificity scores for every targeted base pair in selections of CCR5B TALENs are shown. Specificity scores for CCR5B TALENs targeting all possible combinations of the left (L10, L13, L16) and right (R10, R13, R16) half-sites at all positions in the target half-sites plus a single flanking position. The colors range from a maximum specificity score of 1.0) to white (score of 0, no specificity) to a maximum negative score of -1.0. Boxed bases represent the intended target base. The titles to the right notes the targeted left (L) and right (R) target half-sites for the CCR5B TALEN used in the selection. Selections correspond to conditions listed in Table 2.

[0028] **Figure 13.** Specificity profiles from all CCR5B TALEN selections as bar graphs. Specificity scores for every targeted base pair in selections of CCR5B TALENs are shown. Positive specificity scores, up to complete specificity at a specificity score of 1.0, signify enrichment of that base pair over the other possibilities at that position. Negative specificity scores, down to complete antispecificity of –1.0, represents enrichment against that base pair. Specified positions were plotted as stacked bars above the X-axis (multiple specified base pairs at the same position were plotted over each other with the shortest bar in front, and not end-to-end) while anti-specified base pairs were plotted as narrow, grouped bars. The titles to the right notes the targeted left (L) and right (R) target half-sites for the CCR5B TALEN used in the selection. Selections correspond to conditions listed in Table 2.

[0029] **Figure 14.** Observed versus predicted double-mutant sequence enrichment values. (A) For the L13+R13 CCR5A TALEN selection, the observed double-mutant enrichment values of individual sequences (post-selection sequence abundance ÷ pre-selection sequence abundance) were normalized to the on-target enrichment value (= 1.0 by definition) and plotted against the corresponding predicted double-mutant enrichment values calculated by multiplying the enrichment value of the component single-mutants normalized to the on-target enrichment. The predicted double mutant enrichment values therefore assume independent contributions from each single mutation to the double-mutant's enrichment value. (B) The observed double-mutant sequence enrichment divided by the predicted double-mutant sequence enrichment plotted as a function of the distance (in base pairs) between the two mutations. Only sequences with two mutations in the same half-site were considered.

[0030] **Figure 15.** Effects of engineered TALEN domains and TALEN concentration on specificity. (A) The specificity score of the targeted base pair at each position of the CCR5A site was calculated for CCR5A TALENs containing the canonical, Q3, Q7, or 28-aa C-terminal domains. The specificity scores of the Q3, Q7, or 28-aa C-terminal domain TALENs subtracted by the specificity scores of the TALEN with the canonical C-terminal domain are shown. (B) Same as (A) but for CCR5A TALENs containing engineered N-terminal domains N1, N2, or N3. (C) Same as (A) but comparing specificity scores differences of the canonical CCR5A TALEN assayed at 16 nM, 8 nM, or 4 nM subtracted by the specificity scores of canonical CCR5A TALENs assayed at 32 nM. (D-F) Same as (A-C) but for ATM TALENs. Selections correspond to conditions listed in Table 2.

[0031] **Figure 16.** Spacer-length preferences of TALENs. (A) For each selection with CCR5A TALENs containing various combinations of the canonical, Q3, Q7, or 28-aa C-terminal domains; N1, N2, or N3 N-terminal mutations; and the EL/KK or ELD/KKR FokI variants and at 4, 8, 16, or 32 nM, the DNA spacer-length enrichment values were calculated by dividing the abundance of DNA spacer lengths in post-selection sequences by the abundance of DNA spacer lengths in the preselection library sequences. (B) Same as (A) but for ATM TALENs.

[0032] **Figure 17.** DNA cleavage-site preferences of TALENs. (A) For each selection with CCR5A TALENs with various combinations of canonical, Q3, Q7, or 28-aa C-terminal domains; N1, N2, or N3 N-terminal mutations; and the EL/KK or ELD/KKR FokI variants and at 4, 8, 16, or 32 nM, histograms of the number of spacer DNA base pairs preceding the right half-site for each possible DNA spacer length, normalized to the total sequence counts of the entire selection, are shown. (B) Same as (A) for ATM TALENs.

[0033] **Figure 18.** DNA cleavage-site preferences of TALENs comprising N-terminal domains with different amino acid substitutions.

[0034] **Figure 19.** Exemplary TALEN plasmid construct.

DEFINITIONS

[0035] As used herein and in the claims, the singular forms "a," "an," and "the" include the singular and the plural reference unless the context clearly indicates otherwise. Thus, for example, a reference to "an agent" includes a single agent and a plurality of agents.

[0036] The term "canonical sequence," as used herein, refers to a sequence of DNA, RNA, or amino acids that reflects the most common choice of base or amino acid at each position amongst known molecules of that type. For example, the canonical amino acid sequence of a protein domain may reflect the most common choice of amino acid resides at each position amongst all known domains of that type, or amongst the majority of known domains of that type. In some embodiments, a canonical sequence is a consensus sequence.

[0037] The terms “consensus sequence” and “consensus site,” as used herein in the context of nucleic acid sequences, refers to a calculated sequence representing the most frequent nucleotide residue found at each position in a plurality of similar sequences. Typically, a consensus sequence is determined by sequence alignment in which similar sequences are compared to each other and similar sequence motifs are calculated. In the context of nuclease target site sequences, a consensus sequence of a nuclease target site may, in some embodiments, be the sequence most frequently bound, bound with the highest affinity, and/or cleaved with the highest efficiency by a given nuclease.

[0038] The terms “conjugating,” “conjugated,” and “conjugation” refer to an association of two entities, for example, of two molecules such as two proteins, two domains (*e.g.*, a binding domain and a cleavage domain), or a protein and an agent(*e.g.*, a protein binding domain and a small molecule). The association can be, for example, via a direct or indirect (*e.g.*, via a linker) covalent linkage or via non–covalent interactions. In some embodiments, the association is covalent. In some embodiments, two molecules are conjugated via a linker connecting both molecules. For example, in some embodiments where two proteins are conjugated to each other, *e.g.*, a binding domain and a cleavage domain of an engineered nuclease, to form a protein fusion, the two proteins may be conjugated via a polypeptide linker, *e.g.*, an amino acid sequence connecting the C-terminus of one protein to the N-terminus of the other protein.

[0039] The term “effective amount,” as used herein, refers to an amount of a biologically active agent that is sufficient to elicit a desired biological response. For example, in some embodiments, an effective amount of a TALE nuclease may refer to the amount of the nuclease that is sufficient to induce cleavage of a target site specifically bound and cleaved by the nuclease, *e.g.*, in a cell-free assay, or in a target cell, tissue, or organism. As will be appreciated by the skilled artisan, the effective amount of an agent, *e.g.*, a nuclease, a hybrid protein, or a polynucleotide, may vary depending on various factors as, for example, on the desired biological response, the specific allele, genome, target site, cell, or tissue being targeted, and the agent being used.

[0040] The term “engineered,” as used herein refers to a molecule, complex, substance, or entity that has been designed, produced, prepared, synthesized, and/or manufactured by a human. Accordingly, an engineered product is a product that does not occur in nature. In some embodiments, an engineered molecule or complex, *e.g.*, an engineered TALEN monomer, dimer, or multimer, is a TALEN that has been designed to meet particular requirements or to have particular desired features *e.g.*, to specifically bind a target sequence of interest with minimal off-target binding, to have a specific minimal or maximal cleavage activity, and/or to have a specific stability.

[0041] As used herein, the term "isolated" refers to a molecule, complex, substance, or entity that has been (1) separated from at least some of the components with which it was associated when initially produced (whether in nature or in an experimental setting), and/or (2) produced, prepared, synthesized, and/or manufactured by a human. Isolated substances and/or entities may be separated from at least about 10%, about 20%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, or more of the other components with which they were initially associated. In some embodiments, isolated agents are more than about 80%, about 85%, about 90%, about 91%, about 92%, about 93%, about 94%, about 95%, about 96%, about 97%, about 98%, about 99%, or more than about 99% pure. As used herein, a substance is "pure" if it is substantially free of other components.

[0042] The term "library," as used herein in the context of nucleic acids or proteins, refers to a population of two or more different nucleic acids or proteins, respectively. For example, a library of nuclease target sites comprises at least two nucleic acid molecules comprising different nuclease target sites. In some embodiments, a library comprises at least 10^1, at least 10^2, at least 10^3, at least 10^4, at least 10^5, at least 10^6, at least 10^7, at least 10^8, at least 10^9, at least 10^{10}, at least 10^{11}, at least 10^{12}, at least 10^{13}, at least 10^{14}, or at least 10^{15} different nucleic acids or proteins. In some embodiments, the members of the library may comprise randomized sequences, for example, fully or partially randomized sequences. In some embodiments, the library comprises nucleic acid molecules that are unrelated to each other, *e.g.*, nucleic acids comprising fully randomized sequences. In other embodiments, at least some members of the library may be related, for example, they may be variants or derivatives of a particular sequence, such as a consensus target site sequence.

[0043] The term "linker," as used herein, refers to a chemical group or a molecule linking two molecules or moieties, *e.g.*, a binding domain and a cleavage domain of a nuclease. Typically, the linker is positioned between, or flanked by, two groups, molecules, or other moieties and connected to each one via a covalent bond, thus connecting the two. In some embodiments, the linker is an amino acid or a plurality of amino acids (*e.g.*, a peptide or protein). In some embodiments, the linker is an organic molecule, group, polymer, or chemical moiety.

[0044] The term “nuclease,” as used herein, refers to an agent, for example a protein or a small molecule, capable of cleaving a phosphodiester bond connecting nucleotide residues in a nucleic acid molecule. In some embodiments, a nuclease is a protein, *e.g.*, an enzyme that can bind a nucleic acid molecule and cleave a phosphodiester bond connecting nucleotide residues within the nucleic acid molecule. A nuclease may be an endonuclease, cleaving a phosphodiester bonds within a polynucleotide chain, or an exonuclease, cleaving a phosphodiester bond at the end of the polynucleotide chain. In some embodiments, a nuclease is a site-specific nuclease, binding and/or cleaving a specific phosphodiester bond within a specific nucleotide sequence, which is also referred to herein as the “recognition sequence,” the “nuclease target site,” or the “target site.” In some embodiments, a nuclease recognizes a single stranded target site, while in other embodiments, a nuclease recognizes a double-stranded target site, for example a double-stranded DNA target site. The target sites of many naturally occurring nucleases, for example, many naturally occurring DNA restriction nucleases, are well known to those of skill in the art. In many cases, a DNA nuclease, such as EcoRI, HindIII, or BamHI, recognize a palindromic, double-stranded DNA target site of 4 to 10 base pairs in length, and cut each of the two DNA strands at a specific position within the target site. Some endonucleases cut a double-stranded nucleic acid target site symmetrically, *i.e.*, cutting both strands at the same position so that the ends comprise base-paired nucleotides, also referred to herein as blunt ends. Other endonucleases cut a double-stranded nucleic acid target sites asymmetrically, *i.e.*, cutting each strand at a different position so that the ends comprise unpaired nucleotides. Unpaired nucleotides at the end of a double-stranded DNA molecule are also referred to as “overhangs,” *e.g.*, as “5′-overhang” or as “3′-overhang,” depending on whether the unpaired nucleotide(s) form(s) the 5′ or the 5′ end of the respective DNA strand. Double-stranded DNA molecule ends ending with unpaired nucleotide(s) are also referred to as sticky ends, as they can “stick to” other double-stranded DNA molecule ends comprising complementary unpaired nucleotide(s). A nuclease protein typically comprises a “binding domain” that mediates the interaction of the protein with the nucleic acid substrate, and a “cleavage domain” that catalyzes the cleavage of the phosphodiester bond within the nucleic acid backbone. In some embodiments, a nuclease protein can bind and cleave a nucleic acid molecule in a monomeric form, while, in other embodiments, a nuclease protein has to dimerize or multimerize in order to cleave a target nucleic acid molecule. Binding domains and cleavage domains of naturally occurring nucleases, as well as modular binding domains and cleavage domains that can be combined to create nucleases that bind specific target sites, are well known to those of skill in the art.

For example, transcriptional activator like elements can be used as binding domains to specifically bind a desired target site, and fused or conjugated to a cleavage domain, for example, the cleavage domain of FokI, to create an engineered nuclease cleaving the desired target site.

[0001] The terms “nucleic acid” and “nucleic acid molecule,” as used herein, refer to a compound comprising a nucleobase and an acidic moiety, *e.g.*, a nucleoside, a nucleotide, or a polymer of nucleotides. Typically, polymeric nucleic acids, *e.g.*, nucleic acid molecules comprising three or more nucleotides are linear molecules, in which adjacent nucleotides are linked to each other via a phosphodiester linkage. In some embodiments, “nucleic acid” refers to individual nucleic acid residues (*e.g.* nucleotides and/or nucleosides). In some embodiments, “nucleic acid” refers to an oligonucleotide chain comprising three or more individual nucleotide residues. As used herein, the terms “oligonucleotide” and “polynucleotide” can be used interchangeably to refer to a polymer of nucleotides (*e.g.*, a string of at least three nucleotides). In some embodiments, “nucleic acid” encompasses RNA as well as single and/or double-stranded DNA. Nucleic acids may be naturally occurring, for example, in the context of a genome, a transcript, an mRNA, tRNA, rRNA, siRNA, snRNA, a plasmid, cosmid, chromosome, chromatid, or other naturally occurring nucleic acid molecule. On the other hand, a nucleic acid molecule may be a non-naturally occurring molecule, *e.g.*, a recombinant DNA or RNA, an artificial chromosome, an engineered genome, or fragment thereof, or a synthetic DNA, RNA, DNA/RNA hybrid, or including non-naturally occurring nucleotides or nucleosides. Furthermore, the terms “nucleic acid,” “DNA,” “RNA,” and/or similar terms include nucleic acid analogs, *i.e.* analogs having other than a phosphodiester backbone. Nucleic acids can be purified from natural sources, produced using recombinant expression systems and optionally purified, chemically synthesized, *etc.* Where appropriate, *e.g.*, in the case of chemically synthesized molecules, nucleic acids can comprise nucleoside analogs such as analogs having chemically modified bases or sugars, and backbone modifications′ A nucleic acid sequence is presented in the 5′ to 3′ direction unless otherwise indicated. In some embodiments, a nucleic acid is or comprises natural nucleosides (*e.g.* adenosine, thymidine, guanosine, cytidine, uridine, deoxyadenosine, deoxythymidine, deoxyguanosine, and deoxycytidine); nucleoside analogs (*e.g.*, 2-aminoadenosine, 2-thiothymidine, inosine, pyrrolo-pyrimidine, 3-methyl adenosine, 5-methylcytidine, 2-aminoadenosine, C5-bromouridine, C5-fluorouridine, C5-iodouridine, C5-propynyl-uridine, C5-propynyl-cytidine, C5-methylcytidine, 2-aminoadenosine, 7-deazaadenosine, 7-deazaguanosine, 8-oxoadenosine, 8-oxoguanosine, O(6)-methylguanine, and 2-thiocytidine); chemically modified bases; biologically modified bases (*e.g.*, methylated bases); intercalated bases; modified sugars (*e.g.*, 2′-fluororibose, ribose, 2′-deoxyribose, arabinose, and hexose); and/or modified phosphate groups (*e.g.*, phosphorothioates and 5′-*N*-phosphoramidite linkages).

[0045] The term “pharmaceutical composition,” as used herein, refers to a composition that can be administrated to a subject in the context of treatment of a disease or disorder. In some embodiments, a pharmaceutical composition comprises an active ingredient, *e.g.*, a nuclease or a nucleic acid encoding a nuclease, and a pharmaceutically acceptable excipient.

[0046] The terms “prevention” or “prevent” refer to the prophylactic treatment of a subject who is at risk of developing a disease, disorder, or condition (*e.g.*, at an elevated risk as compared to a control subject, or a control group of subject, or at an elevated risk as compared to the average risk of an age-matched and/or gender-matched subject), resulting in a decrease in the probability that the subject will develop the disease, disorder, or condition (as compared to the probability without prevention), and/or to the inhibition of further advancement of an already established disorder.

[0047] The term “proliferative disease,” as used herein, refers to any disease in which cell or tissue homeostasis is disturbed in that a cell or cell population exhibits an abnormally elevated proliferation rate. Proliferative diseases include hyperproliferative diseases, such as pre-neoplastic hyperplastic conditions and neoplastic diseases. Neoplastic diseases are characterized by an abnormal proliferation of cells and include both benign and malignant neoplasias. Malignant neoplasms are also referred to as cancers.

[0048] The terms “protein,” “peptide,” and “polypeptide” are used interchangeably herein and refer to a polymer of amino acid residues linked together by peptide (amide) bonds. The terms refer to a protein, peptide, or polypeptide of any size, structure, or function. Typically, a protein, peptide, or polypeptide will be at least three amino acids long. A protein, peptide, or polypeptide may refer to an individual protein or a collection of proteins. One or more of the amino acids in a protein, peptide, or polypeptide may be modified, for example, by the addition of a chemical entity such as a carbohydrate group, a hydroxyl group, a phosphate group, a farnesyl group, an isofarnesyl group, a fatty acid group, a linker for conjugation, functionalization, or other modification, *etc.* A protein, peptide, or polypeptide may also be a single molecule or may be a multi-molecular complex. A protein, peptide, or polypeptide may be just a fragment of a naturally occurring protein or peptide. A protein, peptide, or polypeptide may be naturally occurring, recombinant, or synthetic, or any combination thereof. A protein may comprise different domains, for example, a nucleic acid binding domain and a nucleic acid cleavage domain. In some embodiments, a protein comprises a proteinaceous part, *e.g.*, an amino acid sequence constituting a nucleic acid binding domain, and an organic compound, *e.g.*, a compound that can act as a nucleic acid cleavage agent.

[0049] The term "randomized," as used herein in the context of nucleic acid sequences, refers to a sequence or residue within a sequence that has been synthesized to incorporate a mixture of free nucleotides, for example, a mixture of all four nucleotides A, T, G, and C. Randomized residues are typically represented by the letter N within a nucleotide sequence. In some embodiments, a randomized sequence or residue is fully randomized, in which case the randomized residues are synthesized by adding equal amounts of the nucleotides to be incorporated (*e.g.*, 25% T, 25% A, 25% G, and 25% C) during the synthesis step of the respective sequence residue. In some embodiments, a randomized sequence or residue is partially randomized, in which case the randomized residues are synthesized by adding non-equal amounts of the nucleotides to be incorporated (*e.g.*, 79% T, 7% A, 7% G, and 7% C) during the synthesis step of the respective sequence residue. Partial randomization allows for the generation of sequences that are templated on a given sequence, but have incorporated mutations at a desired frequency. For example, if a known nuclease target site is used as a synthesis template, partial randomization in which at each step the nucleotide represented at the respective residue is added to the synthesis at 79%, and the other three nucleotides are added at 7% each, will result in a mixture of partially randomized target sites being synthesized, which still represent the consensus sequence of the original target site, but which differ from the original target site at each residue with a statistical frequency of 21% for each residue so synthesized (distributed binomially). In some embodiments, a partially randomized sequence differs from the consensus sequence by more than 5%, more than 10%, more than 15%, more than 20%, more than 25%, or more than 30% on average, distributed binomially. In some embodiments, a partially randomized sequence differs from the consensus site by no more than 10%, no more than 15%, no more than 20%, no more than 25%, nor more than 30%, no more than 40%, or no more than 50% on average, distributed binomially.

[0050] The term "subject," as used herein, refers to an individual organism, for example, an individual mammal. In some embodiments, the subject is a human of either sex at any stage of development.. In some embodiments, the subject is a non-human mammal. In some embodiments, the subject is a non-human primate. In some embodiments, the subject is a rodent. In some embodiments, the subject is a sheep, a goat, a cattle, a cat, or a dog. In some embodiments, the subject is a vertebrate, an amphibian, a reptile, a fish, an insect, a fly, or a nematode.

[0051] The terms "target nucleic acid," and "target genome," as used herein in the context of nucleases, refer to a nucleic acid molecule or a genome, respectively, that comprises at least one target site of a given nuclease.

[0052] The term "target site," used herein interchangeably with the term "nuclease target site," refers to a sequence within a nucleic acid molecule that is bound and cleaved by a nuclease. A target site may be single-stranded or double-stranded. In the context of nucleases that dimerize, for example, nucleases comprising a FokI DNA cleavage domain, a target site typically comprises a left-half site (bound by one monomer of the nuclease), a right-half site (bound by the second monomer of the nuclease), and a spacer sequence between the half sites in which the cut is made. This structure ([left-half site]-[spacer sequence]-[right-half site]) is referred to herein as an LSR structure. In some embodiments, the left-half site and/or the right-half site is between 10-18 nucleotides long. In some embodiments, either or both half-sites are shorter or longer. In some embodiments, the left and right half sites comprise different nucleic acid sequences.

[0053] The term "Transcriptional Activator-Like Effector," (TALE) as used herein, refers to proteins comprising a DNA binding domain, which contains a highly conserved 33-34 amino acid sequence comprising a highly variable two-amino acid motif (Repeat Variable Diresidue, RVD). The RVD motif determines binding specificity to a nucleic acid sequence, and can be engineered according to methods well known to those of skill in the art to specifically bind a desired DNA sequence (see, *e.g.*, Miller, Jeffrey; *et.al.* (February 2011). "A TALE nuclease architecture for efficient genome editing". *Nature Biotechnology* **29** (2): 143–8; Zhang, Feng; *et.al.* (February 2011). "Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription". *Nature Biotechnology* **29** (2): 149–53; Geiβler, R.; Scholze, H.; Hahn, S.; Streubel, J.; Bonas, U.; Behrens, S. E.; Boch, J. (2011), Shiu, Shin-Han. ed. "Transcriptional Activators of Human Genes with Programmable DNA-Specificity". *PLoS ONE* **6** (5): e19509; Boch, Jens (February 2011). "TALEs of genome targeting". *Nature Biotechnology* **29** (2): 135–6; Boch, Jens; *et.al.* (December 2009). "Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors". *Science* **326** (5959): 1509–12; and Moscou, Matthew J.; Adam J. Bogdanove (December 2009). "A Simple Cipher Governs DNA Recognition by TAL Effectors". *Science* **326** (5959): 1501; the entire contents of each of which are incorporated herein by reference). The simple relationship between amino acid sequence and DNA recognition has allowed for the engineering of specific DNA binding domains by selecting a combination of repeat segments containing the appropriate RVDs.

[0054] The term "Transcriptional Activator-Like Element Nuclease," (TALEN) as used herein, refers to an artificial nuclease comprising a transcriptional activator like effector DNA binding domain to a DNA cleavage domain, for example, a FokI domain. A number of modular assembly schemes for generating engineered TALE constructs have been reported (Zhang, Feng; *et.al.* (February 2011). "Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription". *Nature Biotechnology* **29** (2): 149–53; Geiβler, R.; Scholze, H.; Hahn, S.; Streubel, J.; Bonas, U.; Behrens, S. E.; Boch, J. (2011), Shiu, Shin-Han. ed. "Transcriptional Activators of Human Genes with Programmable DNA-Specificity". *PLoS ONE* **6** (5): e19509; Cermak, T.; Doyle, E. L.; Christian, M.; Wang, L.; Zhang, Y.; Schmidt, C.; Baller, J. A.; Somia, N. V. *et al.* (2011). "Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting". *Nucleic Acids Research*; Morbitzer, R.; Elsaesser, J.; Hausner, J.; Lahaye, T. (2011). "Assembly of custom TALE-type DNA binding domains by modular cloning". *Nucleic Acids Research*; Li, T.; Huang, S.; Zhao, X.; Wright, D. A.; Carpenter, S.; Spalding, M. H.; Weeks, D. P.; Yang, B. (2011). "Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes". *Nucleic Acids Research.*; Weber, E.; Gruetzner, R.; Werner, S.; Engler, C.; Marillonnet, S. (2011). Bendahmane, Mohammed. ed. "Assembly of Designer TAL Effectors by Golden Gate Cloning". *PLoS ONE* **6** (5): e19722; the entire contents of each of which are incorporated herein by reference).

[0055] The terms "treatment," "treat," and "treating," refer to a clinical intervention aimed to reverse, alleviate, delay the onset of, or inhibit the progress of a disease or disorder, or one or more symptoms thereof, as described herein. As used herein, the terms "treatment," "treat," and "treating" refer to a clinical intervention aimed to reverse, alleviate, delay the onset of, or inhibit the progress of a disease or disorder, or one or more symptoms thereof, as described herein. In some embodiments, treatment may be administered after one or more symptoms have developed and/or after a disease has been diagnosed. In other embodiments, treatment may be administered in the absence of symptoms, *e.g.*, to prevent or delay onset of a symptom or inhibit onset or progression of a disease. For example, treatment may be administered to a susceptible individual prior to the onset of symptoms (*e.g.*, in light of a history of symptoms and/or in light of genetic or other susceptibility factors). Treatment may also be continued after symptoms have resolved, for example to prevent or delay their recurrence.

DETAILED DESCRIPTION OF CERTAIN EMBODIMENTS OF THE INVENTION

[0056] Transcription activator-like effector nucleases (TALENs) are fusions of the FokI restriction endonuclease cleavage domain with a DNA-binding transcription activator-like effector (TALE) repeat array. TALENs can be engineered to reduce off-target cleavage activity and thus to specifically bind a target DNA sequence and can thus be used to cleave a target DNA sequence, *e.g.*, in a genome, *in vitro* or *in vivo*. Such engineered TALENs can be used to manipulate genomes *in vivo* or *in vitro*, *e.g.*, for gene knockouts or knock-ins via induction of DNA breaks at a target genomic site for targeted gene knockout through non-homologous end joining (NHEJ) or targeted genomic sequence replacement through homology-directed repair (HDR) using an exogenous DNA template.

[0057] TALENs can be designed to cleave any desired target DNA sequence, including naturally occurring and synthetic sequences. However, the ability of TALENs to distinguish target sequences from closely related off-target sequences has not been studied in depth. Understanding this ability and the parameters affecting it is of importance for the design of TALENs having the desired level of specificity for their therapeutic use and also for choosing unique target sequences to be cleaved in order to minimize the chance of off-target cleavage.

[0058] Some aspects of this disclosure are based on cleavage specificity data obtained from profiling 41 TALENs on 10^{12} potential off-target sites through *in vitro* selection and high-throughput sequencing. Computational analysis of the selection results predicted off-target substrates in the human genome, thirteen of which were modified by TALENs in human cells. Some aspect of this disclosure are based on the surprising findings that (i) TALEN repeats bind DNA relatively independently; (ii) longer TALENs are more tolerant of mismatches, yet are more specific in a genomic context; and (iii) excessive DNA-binding energy can lead to reduced TALEN specificity. Based on these findings, optimized TALENs were engineered with mutations designed to reduce non-specific DNA binding. Some of these engineered TALENs exhibit improved specificity, *e.g.*, 34- to >116-fold greater specificity, in human cells compared to commonly used TALENs.

[0059] The ability to engineer site-specific changes in genomes represents a powerful research capability with significant therapeutic implications. TALENs are fusions of the FokI restriction endonuclease cleavage domain with a DNA-binding TALE repeat array (Figure

1A). These arrays consist of multiple 34-amino acid TALE repeat sequences, each of which uses a repeat-variable di-residue (RVD), the amino acids at positions 12 and 13, to recognize a single DNA nucleotide.[1,2] Examples of RVDs that enable recognition of each of the four DNA base pairs are known, enabling arrays of TALE repeats to be constructed that can bind virtually any DNA sequence. TALENs can be engineered to be active only as heterodimers through the use of obligate heterodimeric FokI variants.[3,4] In this configuration, two distinct TALEN monomers are each designed to bind one target half-site and to cleave within the DNA spacer sequence between the two half-sites.

[0060] In cells, *e.g.*, in mammalian cells, TALEN-induced double-strand breaks can result in targeted gene knockout through non-homologous end joining (NHEJ)[5] or targeted genomic sequence replacement through homology-directed repair (HDR) using an exogenous DNA template.[6,7] TALENs have been successfully used to manipulate genomes in a variety of organisms[8-11] and cell lines.[7,12,13]

[0061] TALEN-mediated DNA cleavage at off-target sites can result in unintended mutations at genomic loci. While SELEX experiments have characterized the DNA-binding specificities of monomeric TALE proteins,[5,7] the DNA cleavage specificities of active, dimeric nucleases can differ from the specificities of their component monomeric DNA-binding domains.[14] Full-genome sequencing of four TALEN-treated yeast strains[15] and two human cell lines[16] derived from a TALEN-treated cell revealed no evidence of TALE-induced genomic off-target mutations, consistent with other reports that observed no off-target genomic modification in Xenopus[17] and human cell lines.[18] In contrast, TALENs were observed to cleave off-target sites containing two to eleven mutations relative to the on-target sequence *in vivo* in zebrafish,[13,19] rats,[9] human primary fibroblasts,[20] and embryonic stem cells.[7] A systematic and comprehensive profile of TALEN specificity generated from measurements of TALEN cleavage on a large set of related mutant target sites has not been described before. Such a broad specificity profile is fundamental to understand and improve the potential of TALENs as research tools and therapeutic agents.

[0062] Some of the work described herein relates to experiments performed to profile the ability of 41 TALEN pairs to cleave 10^{12} off-target variants of each of their respective target sequences using a modified version of a previously described *in vitro* selection[14] for DNA cleavage specificity. These results from these experiments provide comprehensive profiles of TALEN cleavage specificities. The *in vitro* selection results were used to computationally predict off-target substrates in the human genome, 13 of which were confirmed to be cleaved by TALENs in human cells.

[0063] It was surprisingly found that, despite being less specific per base pair, TALENs designed to cleave longer target sites in general exhibit higher overall specificity than those that target shorter sites when considering the number of potential off-target sites in the human genome. The selection results also suggest a model in which excess non-specific TALEN binding energy gives rise to greater off-target cleavage relative to on-target cleavage. Based on this model, we engineered TALENs with substantially improved DNA cleavage specificity *in vitro*, and 30- to >150-fold greater specificity in human cells, than currently used TALEN constructs.

[0064] Some aspects of this disclosure are based on data obtained from profiling the specificity of 41 heterodimeric TALENs designed to target one of three distinct sequence, as described in more detail elsewhere herein. The profiling was performed using an improved version of an *in vitro* selection method[14] (also described in PCT Application Publication WO2013/066438 A2, the entire contents of which are incorporated herein by reference) with modifications that increase the throughput and sensitivity of the selection (Figure 1B).

[0065] Briefly, TALENs were profiled against libraries of $>10^{12}$ DNA sequences and cleavage products were captured and analyzed to determine the specificity and off-target activity of each TALEN. The selection data accurately predicted the efficiency of off-target TALEN cleavage *in vitro*, and also indicated that TALENs are overall highly specific across the entire target sequence, but that some level of off-target cleavage occurs in conventional TALENs which can be undesirable in some scenarios of TALEN use. As a result of the experiments described herein, it was surprisingly found that that TALE repeats bind their respective DNA base pairs independently beyond a slightly increased tolerance for adjacent mismatches, which informed the recognition that TALEN specificity per base pair is independent of target-site length. It was experimentally validated that shorter TALENs have greater specificity per targeted base pair than longer TALENs, but that longer TALENs are more specific against the set of potential cleavage sites in the context of a whole genome than shorter TALENs for the tested TALEN lengths targeting 20- to 32-bp sites, as described in more detail elsewhere herein.

[0066] Some aspects of this disclosure are based on the surprising discovery that excess binding energy in longer TALENs reduces specificity by enabling the cleavage of off-target sequences without a corresponding increase in the efficiency of on-target cleavage efficiency. Some aspects of this disclosure are based on the surprising discovery that TALENs can be engineered to more specifically cleave their target sequences by reducing off-target binding energy without compromising on-target cleavage efficiency. The recognition that TALEN specificity can be improved by reducing non-specific DNA binding energy beyond what is required to enable efficient on-target cleavage served as the basis for the generation of engineered TALENs with improved target site specificity.

[0067] Typically, a TALEN monomer, *e.g.*, a TALEN monomer as provided herein, comprises or is of the following structure:

[N-terminal domain]-[TALE repeat array]-[C-terminal domain]-[nuclease domain]

wherein each "-" individually indicates conjugation, either covalently or non-covalently, and wherein the conjugation can be direct, *e.g.*, via direct bond, or indirect, *e.g.*, via a linker domain. See also Figure 1.

[0068] Some aspects of this disclosure provide TALENs with enhanced specificity as compared to TALENs that were previously used. In general, the sequence specificity of a TALEN is conferred by the TALE repeat array, which binds to a specific nucleotide sequence. TALE repeat arrays consist of multiple 34-amino acid TALE repeat sequences, each of which uses a repeat-variable di-residue (RVD), the amino acids at positions 12 and 13, to recognize a single DNA nucleotide. Some aspects of this disclosure provide that the specific binding of the TALE repeat array is sufficient for dimerization and nucleic acid cleavage, and that non-specific nucleic acid binding activity is due to the N-terminal and/or C-terminal domains of the TALEN.

[0069] Based on this recognition, improved TALENs have been engineered as provided herein. As it was discovered that non-specific binding via the N-terminal domain can occur through excess binding energy conferred by amino acid residues that are positively charged (cationic) at physiological pH, some of the improved TALENs provided herein have a decreased net charge and/or a decreased binding energy for binding their target nucleic acid sequence as compared to canonical TALENs. This decrease in charge leads to a decrease in off-target binding via the modified N-terminal and C-terminal domains. The portion of target recognition and binding, thus, is more narrowly confined to the specific recognition and binding activity of the TALE repeat array. The resulting TALENs, thus, exhibit an increase in the specificity of binding and, in turn, in the specificity of cleaving the target site by the improved TALEN as compared to a TALEN using non-modified domains.

[0070] In some embodiments, a TALEN is provided in which the net charge of the N-terminal domain is less than the net charge of the canonical N-terminal domain (SEQ ID NO: 1); and/or the net charge of the C-terminal domain is less than the net charge of the canonical C-terminal domain (SEQ ID NO: 22). In some embodiments, a TALEN is provided in which the binding energy of the N-terminal domain to a target nucleic acid molecule is less than the binding energy of the canonical N-terminal domain (SEQ ID NO: 1); and/or the binding energy of the C-terminal domain to a target nucleic acid molecule is less than the binding energy of the canonical C-terminal domain (SEQ ID NO: 22). In some embodiments, a modified TALEN N-terminal domain is provided the binding energy of which to the TALEN target nucleic acid molecule is less than the binding energy of the canonical N-terminal domain (SEQ ID NO: 1). In some embodiments, a modified TALEN C-terminal domain is provided the binding energy of which to the TALEN target nucleic acid molecule is less than the binding energy of the canonical C-terminal domain (SEQ ID NO: 22). In some embodiments, the binding energy of the N-terminal and/or of the C-terminal domain in the TALEN provided is decreased by at least 5%, at least 10%, at least 15%, at least 20%, at least 25%, at least 30%, at least 35%, at least 40%, at least 45%, at least 50%, at least 55%, at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95%, at least 98%, or at least 99%.

[0071] In some embodiments, the canonical N-terminal domain and/or the canonical C-terminal domain is modified to replace an amino acid residue that is positively charged at physiological pH with an amino acid residue that is not charged or is negatively charged. In some embodiments, the modification includes the replacement of a positively charged residue with a negatively charged residue. In some embodiments, the modification includes the replacement of a positively charged residue with a neutral (uncharged) residue. In some embodiments, the modification includes the replacement of a positively charged residue with a residue having no charge or a negative charge. In some embodiments, the net charge of the modified N-terminal domain and/or of the modified C-terminal domain is less than or equal to +10, less than or equal to +9, less than or equal to +8, less than or equal to +7, less than or equal to +6, less than or equal to +5, less than or equal to +4, less than or equal to +3, less than or equal to +2, less than or equal to +1, less than or equal to 0, less than or equal to -1, less than or equal to -2, less than or equal to -3, less than or equal to -4, or less than or equal to -5, or less than or equal to -10. In some embodiments, the net charge of the modified N-terminal domain and/or of the modified C-terminal domain is between +5 and -5, between +2 and -7, between 0 and -5, between 0 and -10, between -1 and -10, or between -2 and -15. In some embodiments, the net charge of the modified N-terminal domain and/or of the modified C-terminal domain is negative. In some embodiments, the net charge of the modified N-terminal domain and of the modified C-terminal domain, together, is negative. In some embodiments, the net charge of the modified N-terminal domain and/or of the modified C-terminal domain is neutral or slightly positive (*e.g.*, less than +2 or less than +1). In some embodiments, the net charge of the modified N-terminal domain and of the modified C-terminal domain, together, is neutral or slightly positive (*e.g.*, less than +2 or less than +1).

[0072] In some embodiments, the modified N-terminal domain and/or the modified C-terminal domain comprise(s) an amino acid sequence that differs from the respective canonical domain sequence in that at least one cationic amino acid residue of the canonical domain sequence is replaced with an amino acid residue that exhibits no charge or a negative charge at physiological pH. In some embodiments, at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, or at least 15 cationic amino acid(s) is/are replaced with an amino acid residue that exhibits no charge or a negative charge at physiological pH in the modified N-terminal domain and/or in the modified C-terminal domain. In some embodiments, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 cationic amino acid(s) is/are replaced with an amino acid residue that exhibits no charge or a negative charge at physiological pH in the modified N-terminal domain and/or in the modified C-terminal domain.

[0073] In some embodiments, the cationic amino acid residue is arginine (R), lysine (K), or histidine (H). In some embodiments, the cationic amino acid residue is R or H. In some embodiments, the amino acid residue that exhibits no charge or a negative charge at physiological pH is glutamine (Q), Glycine (G), Asparagine (N), Threonine (T), Serine (S), Aspartic acid (D), or Glutamic Acid (E). In some embodiments, the amino acid residue that exhibits no charge or a negative charge at physiological pH is Q. In some embodiments, at least one lysine or arginine residue is replaced with a glutamine residue in the modified N-terminal domain and/or in the modified C-terminal domain.

[0074] In some embodiments, the C-terminal domain comprises one or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the C-terminal domain comprises two or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the C-terminal domain comprises three or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the C-terminal domain comprises four or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the C-terminal domain comprises five or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the C-terminal domain comprises six or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the C-terminal domain comprises all seven of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the C-terminal domain comprises a Q3 variant sequence (K788Q, R792Q, R801Q, see SEQ ID NO: 23). In some embodiments, the C-terminal domain comprises a Q7 variant sequence (K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q, see SEQ ID NO: 24).

[0075] In some embodiments, the N-terminal domain is a truncated version of the canonical N-terminal domain. In some embodiments, the C-terminal domain is a truncated version of the canonical C-terminal domain. In some embodiments, the truncated N-terminal domain and/or the truncated C-terminal domain comprises less than 90%, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, or less than 25% of the residues of the canonical domain. In some embodiments, the truncated C-terminal domain comprises less than 60, less than 50, less than 40, less than 30, less than 29, less than 28, less than 27, less than 26, less than 25, less than 24, less than 23, less than 22, less than 21, or less than 20 amino acid residues. In some embodiments, the truncated C-terminal domain comprises 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 residues. In some embodiments, the modified N-terminal domain and/or the modified C-terminal domain is/are truncated and comprise one or more amino acid replacement(s). It will be apparent to those of skill in the art that it is desirable in some embodiments to adjust the DNA spacer length in TALENs using truncated domains, *e.g.*, truncated C-terminal domains, in order to accommodate the truncation.

[0076] In some embodiments, the nuclease domain, also sometimes referred to as a nucleic acid cleavage domain is a non-specific cleavage domain, *e.g.*, a FokI nuclease domain. In some embodiments, the nuclease domain is monomeric and must dimerize or multimerize in order to cleave a nucleic acid. Homo- or heterodimerization or multimerization of TALEN monomers typically occurs via binding of the monomers to binding sequences that are in sufficiently close proximity to allow dimerization, *e.g.*, to sequences that are proximal to each other on the same nucleic acid molecule (*e.g.*, the same double-stranded nucleic acid molecule).

[0077] The most commonly used domains, *e.g.*, the most widely used N-terminal and C-terminal domains, are referred to herein as canonical domains. Exemplary sequences of a canonical N-terminal domain (SEQ ID NO: 1) and a canonical C-terminal domain (SEQ ID NO: 22) are provided herein. Exemplary sequences of FokI nuclease domains are also provided herein. In addition, exemplary sequences of TALE repeats forming a CCR5-binding TALE repeat array are provided. It will be understood that the sequences provided below are exemplary and provided for the purpose of illustrating some embodiments embraced by the present disclosure. They are not meant to be limiting and additional sequences useful according to aspects of this disclosure will be apparent to the skilled artisan based on this disclosure.

[0078] Canonical N-terminal domain:

VDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAV**K**YQDMI
AALPEATHEAIVGVG**K**QWSGA**R**ALEALLTVAGEL**R**GPPLQLDTGQLL**K**IA**KR**GGVTAVEAVH
AWRNALTGAPLN (SEQ ID NO: 1)

[0079] Modified N-terminal domain: N1

VDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMI
AALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLL**Q**IAKRGGVTAVEAVH
AWRNALTGAPLN (SEQ ID NO: 2)

[0080] Modified N-terminal domain: N2

VDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMI
AALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLL**Q**IA**Q**RGGVTAVEAVH
AWRNALTGAPLN (SEQ ID NO: 3)

[0081] Modified N-terminal domain: N3

VDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMI
AALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLL**Q**IA**QQ**GGVTAVEAVH
AWRNALTGAPLN (SEQ ID NO: 4)

[0082] TALE repeat array: L18 CCR5A

MTPDQVVAIASNGGGKQALETVQRLLPVLCQDH (SEQ ID NO: 5)
GLTPEQVVAIASHDGGKQALETVQRLLPVLCQAH (SEQ ID NO: 6)
GLTPDQVVAIASNIGGKQALETVQRLLPVLCQAH (SEQ ID NO: 7)
GLTPAQVVAIASNGGGKQALETVQRLLPVLCQDH (SEQ ID NO: 8)

GLTPDQVVAIASNGGGKQALETVQRLLPVLCQDH (SEQ ID NO: 9)

GLTPEQVVAIASNIGGKQALETVQRLLPVLCQAH (SEQ ID NO: 10)

GLTPDQVVAIASHDGGKQALETVQRLLPVLCQAH (SEQ ID NO: 11)

GLTPAQVVAIASNIGGKQALETVQRLLPVLCQDH (SEQ ID NO: 12)

GLTPDQVVAIASHDGGKQALETVQRLLPVLCQDH (SEQ ID NO: 13)

GLTPEQVVAIASHDGGKQALETVQRLLPVLCQAH (SEQ ID NO: 14)

GLTPDQVVAIASNGGGKQALETVQRLLPVLCQAH (SEQ ID NO: 15)

GLTPAQVVAIANNNGGKQALETVQRLLPVLCQDH (SEQ ID NO: 16)

GLTPDQVVAIASHDGGKQALETVQRLLPVLCQDH (SEQ ID NO: 17)

GLTPEQVVAIASNIGGKQALETVQRLLPVLCQAH (SEQ ID NO: 18)

GLTPDQVVAIANNNGGKQALETVQRLLPVLCQAH (SEQ ID NO: 19)

GLTPAQVVAIASHDGGKQALETVQRLLPVLCQDH (SEQ ID NO: 20)

GLTPEQVVAIASNGGGRPALE (SEQ ID NO: 21)

[0083] Canonical C-terminal domain:

SIVAQLS**R**PDPALAALTNDHLVALACLGG**R**PALDAV**KK**GLPHAPALI**KR**TN**RR**IPE**R**TSH**R**VA (SEQ ID NO: 22)

[0084] Modified C-terminal domain: Q3

SIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALI**Q**RTN**Q**RIPERTSH**Q**VA (SEQ ID NO: 23)

[0085] Modified C-terminal domain: Q7

SIVAQLSRPDPALAALTNDHLVALACLGGRPALDAV**QQ**GLPHAPALI**QQ**TN**QQ**IPERTSH**Q**VA (SEQ ID NO: 24)

[0086] Modified C-terminal domain: 28-aa

SIVAQLSRPDPALAALTNDHLVALACLG (SEQ ID NO: 25)

[0087] FokI: homodimeric

GSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF* (SEQ ID NO: 26)

[0088] FokI: EL

GSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEM**E**RYVEENQTRNKH**L**NPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF* (SEQ ID NO: 27)

[0089] FokI: KK

GSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYV**K**ENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNH**K**TNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF* (SEQ ID NO: 28)

[0090] FokI: ELD

GSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEM**E**RYVEENQTR**D**KH**L**NPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF* (SEQ ID NO: 29)

[0091] FokI: KKR

GSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYV**K**ENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLN**RK**TNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF* (SEQ ID NO: 30)

[0092] In some embodiments, a TALEN is provided herein that comprises a canonical N-terminal domain, a TALE repeat array, a modified C-terminal domain, and a nuclease domain. In some embodiments, a TALEN is provided herein that comprises a modified N-terminal domain, a TALE repeat array, a canonical C-terminal domain, and a nuclease domain. In some embodiments, a TALEN is provided herein that comprises a modified N-terminal domain, a TALE repeat array, a modified C-terminal domain, and a nuclease domain. In some embodiments, the nuclease domain is a FokI nuclease domain. In some embodiments, the FokI nuclease domain is a homodimeric FokI domain, or a FokI-EL, FokI-KK, FokI-ELD, or FokI-KKR domain.

[0093] All possible combinations of the specific sequences of canonical and modified domains provided herein are embraced by this disclosure, including the following:

TALEN	N-terminal domain	TALE repeat array	C-terminal domain	Nuclease domain
1	Canonical	Sequence-specific	Q3	Homodimeric
2	Canonical	Sequence-specific	Q3	EL
3	Canonical	Sequence-specific	Q3	KK
4	Canonical	Sequence-specific	Q3	ELD
5	Canonical	Sequence-specific	Q3	KKR
6	Canonical	Sequence-specific	Q7	Homodimeric
7	Canonical	Sequence-specific	Q7	EL
8	Canonical	Sequence-specific	Q7	KK
9	Canonical	Sequence-specific	Q7	ELD
10	Canonical	Sequence-specific	Q7	KKR
11	Canonical	Sequence-specific	Truncated (28aa)	Homodimeric
12	Canonical	Sequence-specific	Truncated (28aa)	EL
13	Canonical	Sequence-specific	Truncated (28aa)	KK
14	Canonical	Sequence-specific	Truncated (28aa)	ELD
15	Canonical	Sequence-specific	Truncated (28aa)	KKR
16	N1	Sequence-specific	Canonical	Homodimeric
17	N1	Sequence-specific	Canonical	EL
18	N1	Sequence-specific	Canonical	KK
19	N1	Sequence-specific	Canonical	ELD
20	N1	Sequence-specific	Canonical	KKR
21	N1	Sequence-specific	Q3	Homodimeric
22	N1	Sequence-specific	Q3	EL
23	N1	Sequence-specific	Q3	KK
24	N1	Sequence-specific	Q3	ELD
25	N1	Sequence-specific	Q3	KKR
26	N1	Sequence-specific	Q7	Homodimeric
27	N1	Sequence-specific	Q7	EL

TALEN	N-terminal domain	TALE repeat array	C-terminal domain	Nuclease domain
28	N1	Sequence-specific	Q7	KK
29	N1	Sequence-specific	Q7	ELD
30	N1	Sequence-specific	Q7	KKR
31	N1	Sequence-specific	Truncated (28aa)	Homodimeric
32	N1	Sequence-specific	Truncated (28aa)	EL
33	N1	Sequence-specific	Truncated (28aa)	KK
34	N1	Sequence-specific	Truncated (28aa)	ELD
35	N1	Sequence-specific	Truncated (28aa)	KKR
36	N2	Sequence-specific	Canonical	Homodimeric
37	N2	Sequence-specific	Canonical	EL
38	N2	Sequence-specific	Canonical	KK
39	N2	Sequence-specific	Canonical	ELD
40	N2	Sequence-specific	Canonical	KKR
41	N2	Sequence-specific	Q3	Homodimeric
42	N2	Sequence-specific	Q3	EL
43	N2	Sequence-specific	Q3	KK
44	N2	Sequence-specific	Q3	ELD
45	N2	Sequence-specific	Q3	KKR
46	N2	Sequence-specific	Q7	Homodimeric
47	N2	Sequence-specific	Q7	EL
48	N2	Sequence-specific	Q7	KK
49	N2	Sequence-specific	Q7	ELD
50	N2	Sequence-specific	Q7	KKR
51	N2	Sequence-specific	Truncated (28aa)	Homodimeric
52	N2	Sequence-specific	Truncated (28aa)	EL
53	N2	Sequence-specific	Truncated (28aa)	KK
54	N2	Sequence-specific	Truncated (28aa)	ELD
55	N2	Sequence-specific	Truncated (28aa)	KKR
56	N3	Sequence-specific	Canonical	Homodimeric
57	N3	Sequence-specific	Canonical	EL

TALEN	N-terminal domain	TALE repeat array	C-terminal domain	Nuclease domain
58	N3	Sequence-specific	Canonical	KK
59	N3	Sequence-specific	Canonical	ELD
60	N3	Sequence-specific	Canonical	KKR
61	N3	Sequence-specific	Q3	Homodimeric
62	N3	Sequence-specific	Q3	EL
63	N3	Sequence-specific	Q3	KK
64	N3	Sequence-specific	Q3	ELD
65	N3	Sequence-specific	Q3	KKR
66	N3	Sequence-specific	Q7	Homodimeric
67	N3	Sequence-specific	Q7	EL
68	N3	Sequence-specific	Q7	KK
69	N3	Sequence-specific	Q7	ELD
70	N3	Sequence-specific	Q7	KKR
71	N3	Sequence-specific	Truncated (28aa)	Homodimeric
72	N3	Sequence-specific	Truncated (28aa)	EL
73	N3	Sequence-specific	Truncated (28aa)	KK
74	N3	Sequence-specific	Truncated (28aa)	ELD
75	N3	Sequence-specific	Truncated (28aa)	KKR
76	Canonical	Sequence-specific	Canonical	EL
77	Canonical	Sequence-specific	Canonical	KK
78	Canonical	Sequence-specific	Canonical	ELD
79	Canonical	Sequence-specific	Canonical	KKR
80	Canonical	Sequence-specific	Truncated (28aa)	Homodimeric
81	Canonical	Sequence-specific	Truncated (28aa)	EL
82	Canonical	Sequence-specific	Truncated (28aa)	KK
83	Canonical	Sequence-specific	Truncated (28aa)	ELD
84	Canonical	Sequence-specific	Truncated (28aa)	KKR

Table 1: Exemplary TALENs embraced by the present disclosure. The respective TALE repeat array employed will depend on the specific target sequence. Those of skill in the art will be able to design such sequence-specific TALE repeat arrays based on the instant

disclosure and the knowledge in the art. Sequences for the different N-terminal, C-terminal, and Nuclease domains are provided above (See, SEQ ID NOs 1-4 and 22-30).

[0094] It will be understood by those of skill in the art that the exemplary sequences provided herein are for illustration purposes only and are not intended to limit the scope of the present disclosure. The disclosure also embraces the use of each of the inventive TALEN domains, *e.g.*, the modified N-terminal domains, C-terminal domains, and nuclease domains described herein, in the context of other TALEN sequences, *e.g.*, other modified or unmodified TALEN structures. Additional sequences satisfying the described principles and parameters that are useful in accordance to aspects of this disclosure will be apparent to the skilled artisan.

[0095] In some embodiments, the TALEN provided is a monomer. In some embodiments, the TALEN monomer can dimerize with another TALEN monomer to form a TALEN dimer. In some embodiments the formed dimer is a homodimer. In some embodiments, the dimer is a heterodimer.

[0096] In some embodiments, TALENs provided herein cleave their target sites with high specificity. For example, in some embodiments an improved TALEN is provided that has been engineered to cleave a desired target site within a genome while binding and/or cleaving less than 1, less than 2, less than 3, less than 4, less than 5, less than 6, less than 7, less than 8, less than 9 or less than 10 off-target sites at a concentration effective for the nuclease to cut its intended target site. In some embodiments, a TALEN is provided that has been engineered to cleave a desired unique target site that has been selected to differ from any other site within a genome by at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, or at least 10 nucleotide residues.

[0097] Some aspects of this disclosure provide nucleic acids encoding the TALENs provided herein. For example, nucleic acids are provided herein that encode the TALENs described in Table 1. In some embodiments, the nucleic acids encoding the TALEN are under the control of a heterologous promoter. In some embodiments, the encoding nucleic acids are included in an expression construct, *e.g.*, a plasmid, a viral vector, or a linear expression construct. In some embodiments, the nucleic acid or expression construct is in a cell, tissue, or organism.

[0098] The map of an exemplary nucleic acid encoding a TALEN provided herein is illustrated in Figure 19. An exemplary sequence of such a nucleic acid is provided below. It will be understood by those of skill in the art that the maps and sequences provided herein are exemplary and do not limit the scope of this disclosure.

[0099] As described elsewhere herein, TALENs, including the improved TALENs provided by this disclosure, can be engineered to bind (and cleave) virtually any nucleic acid sequence based on the sequence-specific TALE repeat array employed. In some embodiments, an improved TALEN provided herein binds a target sequence within a gene known to be associated with a disease or disorder. In some embodiments, TALENs provided herein may be used for therapeutic purposes. For example, in some embodiments, TALENs provided herein may be used for treatment of any of a variety of diseases, disorders, and/or conditions, including but not limited to one or more of the following: autoimmune disorders (*e.g.* diabetes, lupus, multiple sclerosis, psoriasis, rheumatoid arthritis); inflammatory disorders (*e.g.* arthritis, pelvic inflammatory disease); infectious diseases (*e.g.* viral infections (*e.g.*, HIV, HCV, RSV), bacterial infections, fungal infections, sepsis); neurological disorders (*e.g.* Alzheimer's disease, Huntington's disease; autism; Duchenne muscular dystrophy); cardiovascular disorders (*e.g.* atherosclerosis, hypercholesterolemia, thrombosis, clotting disorders, angiogenic disorders such as macular degeneration); proliferative disorders (*e.g.* cancer, benign neoplasms); respiratory disorders (*e.g.* chronic obstructive pulmonary disease); digestive disorders (*e.g.* inflammatory bowel disease, ulcers); musculoskeletal disorders (*e.g.* fibromyalgia, arthritis); endocrine, metabolic, and nutritional disorders (*e.g.* diabetes, osteoporosis); urological disorders (*e.g.* renal disease); psychological disorders (*e.g.* depression, schizophrenia); skin disorders (*e.g.* wounds, eczema); blood and lymphatic disorders (*e.g.* anemia, hemophilia); *etc.* In some embodiments, the TALEN cleaves the target sequence upon dimerization. In some embodiments, a TALEN provided herein cleaves a target site within an allele that is associated with a disease or disorder. In some embodiments, the TALEN cleaves a target site the cleavage of which results in the treatment or prevention of a disease or disorder. In some embodiments, the disease is HIV/AIDS. In some embodiments, the disease is a proliferative disease. In some embodiments, the TALEN binds a CCR5 target sequence(*e.g.*, a CCR5 sequence associated with HIV). In some embodiments, the TALEN binds an ATM target sequence (*e.g.*, an ATM target sequence associated with ataxia telangiectasia). In some embodiments, the TALEN binds a VEGFA target sequence (*e.g.*, a VEGFA sequence associated with a proliferative disease). In some embodiments, the TALEN binds a CFTR target sequence (*e.g.*, a CFTR sequence associated with cystic fibrosis). In some embodiments, the TALEN binds a dystrophin target sequence (*e.g.*, a dystrophin gene sequence associated with Duchenne muscular dystrophy). In some embodiments, the TALEN binds a target sequence associated with haemochromatosis, haemophilia, Charcot–Marie–Tooth disease, neurofibromatosis, phenylketonuria, polycystic kidney disease, sickle-cell disease, or Tay–Sachs disease. Suitable target genes, *e.g.*, genes causing the listed diseases, are known to those of skill in the art. Additional genes and gene sequences associated with a disease or disorder will be apparent to those of skill in the art.

[00100] Some aspects of this disclosure provide isolated TALE effector domains, *e.g.*, N- and C-terminal TALE effector domains, with decreased non-specific nucleic acid binding activity as compared to previously used TALE effector domains. The isolated TALE effector domains provided herein can be used in the context of suitable TALE effector molecules, *e.g.*, TALE nucleases, TALE transcriptional activators, TALE transcriptional repressors, TALE recombinases, and TALE epigenome modification enzymes. Additional suitable TALE effectors in the context of which the isolated TALE domains can be used will be apparent to those of skill in the art based on this disclosure. In general, the isolated N- and C-terminal domains provided herein are engineered to optimize, *e.g.*, minimize, excess binding energy conferred by amino acid residues that are positively charged (cationic) at physiological pH. Some of the improved N-terminal or C-terminal TALE domains provided herein have a decreased net charge and/or a decreased binding energy for binding a target nucleic acid sequence as compared to the respective canonical TALE domains. When used as part of a TALE effector molecule, *e.g.*, a TALE nuclease, TALE transcriptional activator, TALE transcriptional repressor, TALE recombinase, or TALE epigenome modification enzyme, this decrease in charge leads to a decrease in off-target binding via the modified N-terminal and C-terminal domain(s). The portion of target recognition and binding, thus, is more narrowly confined to the specific recognition and binding activity of the TALE repeat array, as explained in more detail elsewhere herein. The resulting TALE effector molecule, thus, exhibits an increase in the specificity of binding and, in turn, in the specificity of the respective effect of the TALE effector (*e.g.*, cleaving the target site by a TALE nuclease, activation of a target gene by a TALE transcriptional activator, repression of expression of a target gene by a TALE transcriptional repressor, recombination of a target sequence by a TALE recombinase, or epigenetic modification of a target sequence by a TALE epigenome modification enzyme) as compared to TALE effector molecules using unmodified domains.

[00101] In some embodiments, an isolated N-terminal TALE domain is provided in which the net charge is less than the net charge of the canonical N-terminal domain (SEQ ID NO: 1). In some embodiments, an isolated C-terminal TALE domain is provided in which the net charge is less than the net charge of the canonical C -terminal domain (SEQ ID NO:

22). In some embodiments, an isolated N-terminal TALE domain is provided in which the binding energy to a target nucleic acid molecule is less than the binding energy of the canonical N-terminal domain (SEQ ID NO: 1). In some embodiments, an isolated C-terminal TALE domain is provided in which the binding energy to a target nucleic acid molecule is less than the binding energy of the canonical C-terminal domain (SEQ ID NO: 22). In some embodiments, the binding energy of the isolated N-terminal and/or of the isolated C-terminal TALE domain provided herein is decreased by at least 5%, at least 10%, at least 15%, at least 20%, at least 25%, at least 30%, at least 35%, at least 40%, at least 45%, at least 50%, at least 55%, at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95%, at least 98%, or at least 99%.

[00102] In some embodiments, the canonical N-terminal domain and/or the canonical C-terminal domain is modified to replace an amino acid residue that is positively charged at physiological pH with an amino acid residue that is not charged or is negatively charged to arrive at the isolated N-terminal and/or C-terminal domain provided herein. In some embodiments, the modification includes the replacement of a positively charged residue with a negatively charged residue. In some embodiments, the modification includes the replacement of a positively charged residue with a neutral (uncharged) residue. In some embodiments, the modification includes the replacement of a positively charged residue with a residue having no charge or a negative charge. In some embodiments, the net charge of the isolated N-terminal domain and/or of the isolated C-terminal domain provided herein is less than or equal to +10, less than or equal to +9, less than or equal to +8, less than or equal to +7, less than or equal to +6, less than or equal to +5, less than or equal to +4, less than or equal to +3, less than or equal to +2, less than or equal to +1, less than or equal to 0, less than or equal to -1, less than or equal to -2, less than or equal to -3, less than or equal to -4, or less than or equal to -5, or less than or equal to -10 at physiological pH. In some embodiments, the net charge of the isolated N-terminal domain and/or of the isolated C-terminal domain is between +5 and -5, between +2 and -7, between 0 and -5, between 0 and -10, between -1 and -10, or between -2 and -15 at physiological pH. In some embodiments, the net charge of the isolated N-terminal TALE domain and/or of the isolated C-terminal TALE domain is negative. In some embodiments, an isolated N-terminal TALE domain and an isolated C-terminal TALE domain are provided and the net charge of the isolated N-terminal TALE domain and of the isolated C-terminal TALE domain, together, is negative. In some embodiments, the net charge of the isolated N-terminal TALE domain and/or of the isolated C-terminal TALE domain is neutral or slightly positive (*e.g.*, less than +2 or less than +1 at physiological pH). In some embodiments, an isolated N-terminal TALE domain and an isolated C-terminal TALE domain are provided, and the net charge of the isolated N-terminal TALE domain and of the isolated C-terminal TALE domain, together, is neutral or slightly positive (*e.g.*, less than +2 or less than +1 at physiological pH).

[00103] In some embodiments, the isolated N-terminal domain and/or the isolated C-terminal domain provided herein comprise(s) an amino acid sequence that differs from the respective canonical domain sequence in that at least one cationic amino acid residue of the canonical domain sequence is replaced with an amino acid residue that exhibits no charge or a negative charge at physiological pH. In some embodiments, at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, or at least 15 cationic amino acid(s) is/are replaced with an amino acid residue that exhibits no charge or a negative charge at physiological pH in the isolated N-terminal domain and/or in the isolated C-terminal domain provided. In some embodiments, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 cationic amino acid(s) is/are replaced with an amino acid residue that exhibits no charge or a negative charge at physiological pH in the isolated N-terminal domain and/or in the isolated C-terminal domain.

[00104] In some embodiments, the cationic amino acid residue is arginine (R), lysine (K), or histidine (H). In some embodiments, the cationic amino acid residue is R or H. In some embodiments, the amino acid residue that exhibits no charge or a negative charge at physiological pH is glutamine (Q), glycine (G), asparagine (N), threonine (T), serine (S), aspartic acid (D), or glutamic acid (E). In some embodiments, the amino acid residue that exhibits no charge or a negative charge at physiological pH is Q. In some embodiments, at least one lysine or arginine residue is replaced with a glutamine residue in the isolated N-terminal domain and/or in the isolated C-terminal domain.

[00105] In some embodiments, an isolated C-terminal TALE domain is provided herein that comprises one or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the isolated C-terminal domain comprises two or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the isolated C-terminal domain comprises three or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the isolated C-terminal domain comprises four or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the isolated C-terminal domain comprises five or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the isolated C-terminal domain comprises six or more of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the isolated C-terminal domain comprises all seven of the following amino acid replacements: K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q. In some embodiments, the isolated C-terminal domain comprises a Q3 variant sequence (K788Q, R792Q, R801Q, see SEQ ID NO: 23). In some embodiments, the isolated C-terminal domain comprises a Q7 variant sequence (K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q, see SEQ ID NO: 24).

[00106] In some embodiments, an isolated N-terminal TALE domain is provided that is a truncated version of the canonical N-terminal domain. In some embodiments, an isolated C-terminal TALE domain is provided that is a truncated version of the canonical C-terminal domain. In some embodiments, the truncated N-terminal domain and/or the truncated C-terminal domain comprises less than 90%, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, or less than 25% of the residues of the canonical domain. In some embodiments, the truncated C-terminal domain comprises less than 60, less than 50, less than 40, less than 30, less than 29, less than 28, less than 27, less than 26, less than 25, less than 24, less than 23, less than 22, less than 21, or less than 20 amino acid residues. In some embodiments, the truncated C-terminal domain comprises 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 residues. In some embodiments, an isolated N-terminal TALE domain and/or an isolated C-terminal domain is provided herein that is/are truncated and comprise(s) one or more amino acid replacement(s). In some embodiments, the isolated N-terminal TALE domains comprise an amino acid sequence as provided in any of SEQ ID NOs 2-5. In some embodiments, the isolated C-terminal TALE domains comprise an amino acid sequence as provided in any of SEQ ID NOs 23-25.

[00107] It will be apparent to those of skill in the art that the isolated C- and N-terminal TALE domains provided herein may be used in the context of any TALE effector molecule, *e.g.*, as part of a TALE nuclease, a TALE transcriptional activator, a TALE transcriptional repressor, a TALE recombinase, a TALE epigenome modification enzyme, or any other suitable TALE effector molecule. In some embodiments, a TALE domain provided herein is used in the context of a TALE molecule comprising or consisting essentially of the following structure

[N-terminal domain]-[TALE repeat array]-[C-terminal domain]-[effector domain] or

[effector domain]-[N-terminal domain]-[TALE repeat array]-[C-terminal domain], wherein the effector domain may, in some embodiments, be a nuclease domain, a transcriptional activator or repressor domain, a recombinase domain, or an epigenetic modification enzyme domain.

[00108] It will also be apparent to those of skill in the art that it is desirable, in some embodiments, to adjust the DNA spacer length in TALE effector molecules comprising such a spacer, when using a truncated domain, *e.g.*, truncated C-terminal domain as provided herein, in order to accommodate the truncation.

[00109] Some aspects of this disclosure provide compositions comprising a TALEN provided herein, *e.g.*, a TALEN monomer. In some embodiments, the composition comprises the TALEN monomer and a different TALEN monomer that can form a heterodimer with the TALEN, wherein the dimer exhibits nuclease activity.

[00110] In some embodiments, the TALEN is provided in a composition formulated for administration to a subject, *e.g.*, to a human subject. For example, in some embodiments, a pharmaceutical composition is provided that comprises the TALEN and a pharmaceutically acceptable excipient. In some embodiments, the pharmaceutical composition is formulated for administration to a subject. In some embodiments, the pharmaceutical composition comprises an effective amount of the TALEN for cleaving a target sequence in a cell in the subject. In some embodiments, the TALEN binds a target sequence within a gene known to be associated with a disease or disorder and wherein the composition comprises an effective amount of the TALEN for alleviating a symptom associated with the disease or disorder.

[00111] For example, some embodiments provide pharmaceutical compositions comprising a TALEN as provided herein, or a nucleic acid encoding such a nuclease, and a pharmaceutically acceptable excipient. Pharmaceutical compositions may optionally comprise one or more additional therapeutically active substances.

[00112] Formulations of the pharmaceutical compositions described herein may be prepared by any method known or hereafter developed in the art of pharmacology. In general, such preparatory methods include the step of bringing the active ingredient into association with an excipient and/or one or more other accessory ingredients, and then, if necessary and/or desirable, shaping and/or packaging the product into a desired single- or multi-dose unit.

[00113] Pharmaceutical formulations may additionally comprise a pharmaceutically acceptable excipient, which, as used herein, includes any and all solvents, dispersion media, diluents, or other liquid vehicles, dispersion or suspension aids, surface active agents, isotonic agents, thickening or emulsifying agents, preservatives, solid binders, lubricants and the like, as suited to the particular dosage form desired. Remington's *The Science and Practice of Pharmacy*, 21st Edition, A. R. Gennaro (Lippincott, Williams & Wilkins, Baltimore, MD, 2006; incorporated herein by reference) discloses various excipients used in formulating pharmaceutical compositions and known techniques for the preparation thereof. Except insofar as any conventional excipient medium is incompatible with a substance or its derivatives, such as by producing any undesirable biological effect or otherwise interacting in a deleterious manner with any other component(s) of the pharmaceutical composition, its use is contemplated to be within the scope of this invention.

[00114] In some embodiments, a composition provided herein is administered to a subject, for example, to a human subject, in order to effect a targeted genomic modification within the subject. In some embodiments, cells are obtained from the subject and contacted with a nuclease or a nuclease-encoding nucleic acid ex vivo, and re-administered to the subject after the desired genomic modification has been effected or detected in the cells. Although the descriptions of pharmaceutical compositions provided herein are principally directed to pharmaceutical compositions which are suitable for administration to humans, it will be understood by the skilled artisan that such compositions are generally suitable for administration to animals of all sorts. Modification of pharmaceutical compositions suitable for administration to humans in order to render the compositions suitable for administration to various animals is well understood, and the ordinarily skilled veterinary pharmacologist can design and/or perform such modification with no more than routine experimentation. Subjects to which administration of the pharmaceutical compositions is contemplated include, but are not limited to, humans and/or other primates; mammals, including, but not limited to, cattle, pigs, horses, sheep, cats, dogs, mice, and/or rats; and/or birds, including commercially relevant birds such as chickens, ducks, geese, and/or turkeys.

[00115] The scope of this disclosure embraces methods of using the TALENs provided herein. It will be apparent to those of skill in the art that the TALENs provided herein can be used in any method suitable for the application of TALENs, including, but not limited to, those methods and applications known in the art. Such methods may include TALEN-mediated cleavage of DNA, *e.g.*, in the context of genome manipulations such as, for example, targeted gene knockout through non-homologous end joining (NHEJ) or targeted genomic sequence replacement through homology-directed repair (HDR) using an exogenous DNA template, respectively. The improved features of the TALENs provided herein, *e.g.*, the improved specificity of some of the TALENs provided herein, will typically allow for such methods and applications to be carried out with greater efficiency. All methods and applications suitable for the use of TALENs, and performed with the TALENs provided herein, are contemplated and are within the scope of this disclosure. For example, the instant disclosure provides the use of the TALENs provided herein in any method suitable for the use of TALENs as described in Boch, Jens (February 2011). "TALEs of genome targeting". Nature Biotechnology 29 (2): 135–6. doi:10.1038/nbt.1767. PMID 21301438; Boch, Jens; *et.al.* (December 2009). "Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors". Science 326 (5959): 1509–12. Bibcode:2009Sci...326.1509B. doi:10.1126/science.1178811. PMID 19933107; Moscou, Matthew J.; Adam J. Bogdanove (December 2009). "A Simple Cipher Governs DNA Recognition by TAL Effectors". Science 326 (5959): 1501. Bibcode:2009Sci...326.1501M. doi:10.1126/science.1178817. PMID 19933106; Christian, Michelle; *et.al.* (October 2010). "Targeting DNA Double-Strand Breaks with TAL Effector Nucleases". Genetics 186 (2): 757–61. doi:10.1534/genetics.110.120717. PMC 2942870. PMID 20660643; Li, Ting; *et.al.* (August 2010). "TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain". Nucleic Acids Research 39: 1–14. doi:10.1093/nar/gkq704. PMC 3017587. PMID 20699274; Mahfouz, Magdy M.; et.al. (February 2010). "De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks". PNAS 108 (6): 2623–8. Bibcode:2011PNAS..108.2623M. doi:10.1073/pnas.1019533108. PMC 3038751. PMID 21262818; Cermak, T.; Doyle, E. L.; Christian, M.; Wang, L.; Zhang, Y.; Schmidt, C.; Baller, J. A.; Somia, N. V. *et al.* (2011). "Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting". Nucleic Acids Research. doi:10.1093/nar/gkr218; Miller, Jeffrey; et.al. (February 2011). "A TALE nuclease architecture for efficient genome editing". Nature Biotechnology 29 (2): 143–8. doi:10.1038/nbt.1755. PMID 21179091; Hockemeyer, D.; Wang, H.; Kiani, S.; Lai, C. S.; Gao, Q.; Cassady, J. P.; Cost, G. J.; Zhang, L. *et al.* (2011). "Genetic engineering of human pluripotent cells using TALE nucleases". Nature Biotechnology 29 (8). doi:10.1038/nbt.1927; Wood, A. J.; Lo, T. -W.; Zeitler, B.; Pickle, C. S.; Ralston, E. J.; Lee, A. H.; Amora, R.; Miller, J. C. *et al.* (2011). "Targeted Genome Editing Across Species Using ZFNs and TALENs". Science 333 (6040): 307. doi:10.1126/science.1207773. PMC 3489282. PMID 21700836; Tesson, L.; Usal, C.; Ménoret, S. V.; Leung, E.; Niles, B. J.; Remy, S. V.; Santiago, Y.; Vincent, A. I. *et al.* (2011). "Knockout rats generated by embryo microinjection of TALENs". Nature Biotechnology 29 (8): 695. doi:10.1038/nbt.1940; Huang, P.; Xiao, A.; Zhou, M.; Zhu, Z.; Lin, S.; Zhang, B. (2011). "Heritable gene targeting in zebrafish using customized TALENs". Nature Biotechnology 29 (8): 699. doi:10.1038/nbt.1939; Doyon, Y.; Vo, T. D.; Mendel, M. C.; Greenberg, S. G.; Wang, J.; Xia, D. F.; Miller, J. C.; Urnov, F. D. *et al.* (2010). "Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures". Nature Methods 8 (1): 74–79. doi:10.1038/nmeth.1539. PMID 21131970; Szczepek, M.; Brondani, V.; Büchel, J.; Serrano, L.; Segal, D. J.; Cathomen, T. (2007). "Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases". Nature Biotechnology 25 (7): 786. doi:10.1038/nbt1317. PMID 17603476; Guo, J.; Gaj, T.; Barbas Iii, C. F. (2010). "Directed Evolution of an Enhanced and Highly Efficient FokI Cleavage Domain for Zinc Finger Nucleases". Journal of Molecular Biology 400 (1): 96. doi:10.1016/j.jmb.2010.04.060. PMC 2885538. PMID 20447404; Mussolino, C.; Morbitzer, R.; Lutge, F.; Dannemann, N.; Lahaye, T.; Cathomen, T. (2011). "A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity". Nucleic Acids Research. doi:10.1093/nar/gkr597; Zhang, Feng; et.al. (February 2011). "Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription". Nature Biotechnology 29 (2): 149–53. doi:10.1038/nbt.1775. PMC 3084533. PMID 21248753; Morbitzer, R.; Elsaesser, J.; Hausner, J.; Lahaye, T. (2011). "Assembly of custom TALE-type DNA binding domains by modular cloning". Nucleic Acids Research. doi:10.1093/nar/gkr151; Li, T.; Huang, S.; Zhao, X.; Wright, D. A.; Carpenter, S.; Spalding, M. H.; Weeks, D. P.; Yang, B. (2011). "Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes". Nucleic Acids Research. doi:10.1093/nar/gkr188; Geiβler, R.; Scholze, H.; Hahn, S.; Streubel, J.; Bonas, U.; Behrens, S. E.; Boch, J. (2011). "Transcriptional Activators of Human Genes with Programmable DNA-Specificity". In Shiu, Shin-Han. PLoS ONE 6 (5): e19509. doi:10.1371/journal.pone.0019509; Weber, E.; Gruetzner, R.; Werner, S.; Engler, C.; Marillonnet, S. (2011). "Assembly of Designer TAL Effectors by Golden Gate Cloning". In Bendahmane, Mohammed. PLoS ONE 6 (5): e19722. doi:10.1371/journal.pone.0019722; Sander *et al.* "Targeted gene disruption in somatic zebrafish cells using engineered TALENs". Nature Biotechnology Vol 29:697-98 (5 August 2011) Sander, J. D.; Cade, L.; Khayter, C.; Reyon, D.; Peterson, R. T.; Joung, J. K.; Yeh, J. R. J. (2011). "Targeted gene disruption in somatic zebrafish cells using engineered TALENs". Nature Biotechnology 29 (8): 697. doi:10.1038/nbt.1934; the entire contents of each of which are incorporated herein by reference.

[00116] In some embodiments, the TALENs, TALEN domains, TALEN-encoding or TALEN domain-encoding nucleic acids, compositions, and reagents described herein are isolated. In some embodiments, the TALENs, TALEN domains, TALEN-encoding or TALEN domain-encoding nucleic acids, compositions, and reagents described herein are purified, *e.g.*, at least 60%, at least 70%, at least 80%, at least 90%, or at least 95% pure.

[00117] Some aspects of this disclosure provide methods of cleaving a target sequence in a nucleic acid molecule using an inventive TALEN as described herein. In some embodiments, the method comprises contacting a nucleic acid molecule comprising the target sequence with a TALEN binding the target sequence under conditions suitable for the TALEN to bind and cleave the target sequence. In some embodiments, the TALEN is provided as a monomer. In some embodiments, the inventive TALEN monomer is provided in a composition comprising a different TALEN monomer that can dimerize with the first inventive TALEN monomer to form a heterodimer having nuclease activity. In some embodiments, the inventive TALEN is provided in a pharmaceutical composition. In some embodiments, the target sequence is in a cell. In some embodiments, the target sequence is in the genome of a cell. In some embodiments, the target sequence is in a subject. In some embodiments, the method comprises administering a composition, *e.g.*, a pharmaceutical composition, comprising the TALEN to the subject in an amount sufficient for the TALEN to bind and cleave the target site.

[00118] Some aspects of this disclosure provide methods of preparing engineered TALENs. In some embodiments, the method comprises replacing at least one amino acid in the canonical N-terminal TALEN domain and/or the canonical C-terminal TALEN domain with an amino acid having no charge or a negative charge at physiological pH; and/or truncating the N-terminal TALEN domain and/or the C-terminal TALEN domain to remove a positively charged fragment; thus generating an engineered TALEN having an N-terminal domain and/or a C-terminal domain of decreased net charge. In some embodiments, the at least one amino acid being replaced comprises a cationic amino acid or an amino acid having a positive charge at physiological pH. In some embodiments, the amino acid replacing the at least one amino acid is a cationic amino acid or a neutral amino acid. In some embodiments, the truncated N-terminal TALEN domain and/or the truncated C-terminal TALEN domain comprises less than 90%, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, or less than 25% of the residues of the respective canonical domain. In some embodiments, the truncated C-terminal domain comprises less than 60, less than 50,

less than 40, less than 30, less than 29, less than 28, less than 27, less than 26, less than 25, less than 24, less than 23, less than 22, less than 21, or less than 20 amino acid residues.

[00119] In some embodiments, the truncated C-terminal domain comprises 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 amino acid residues. In some embodiments, the method comprises replacing at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, or at least 15 amino acids in the canonical N-terminal TALEN domain and/or in the canonical C-terminal TALEN domain with an amino acid having no charge or a negative charge at physiological pH. In some embodiments, the amino acid being replaced is arginine (R) or lysine (K). In some embodiments, the amino acid residue having no charge or a negative charge at physiological pH is glutamine (Q) or glycine (G). In some embodiments, the method comprises replacing at least one lysine or arginine residue with a glutamine residue.

[00120] In some embodiments, the improved TALENs provided herein are designed and/or generated by recombinant technology. In some embodiments, designing and/or generating comprises designing a TALE repeat array that specifically binds a desired target sequence, or a half-site thereof.

[00121] Some aspects of this disclosure provide kits comprising an engineered TALEN as provided herein, or a composition (*e.g.*, a pharmaceutical composition) comprising such a TALEN. In some embodiments, the kit comprises an excipient and instructions for contacting the TALEN with the excipient to generate a composition suitable for contacting a nucleic acid with the TALEN. In some embodiments, the excipient is a pharmaceutically acceptable excipient.

[00122] Typically, the kit will comprise a container housing the components of the kit, as well as written instructions stating how the components of the kit should be stored and used.

[00123] The function and advantage of these and other embodiments of the present invention will be more fully understood from the Examples below. The following Examples are intended to illustrate the benefits of the present invention and to describe particular embodiments, but are not intended to exemplify the full scope of the invention. Accordingly, it will be understood that the Examples are not meant to limit the scope of the invention.

EXAMPLES

EXAMPLE 1

Materials and Methods

Oligonucleotides, PCR and DNA Purification

[00124] All oligonucleotides were purchased from Integrated DNA Technologies (IDT). Oligonucleotide sequences are listed in Table 10. PCR was performed with 0.4 μL of 2 U/μL Phusion Hot Start II DNA polymerase (Thermo-Fisher) in 50 μL with 1x HF Buffer, 0.2 mM dNTP mix (0.2 mM dATP, 0.2 mM dCTP, 0.2 mM dGTP, 0.2 mM dTTP) (NEB), 0.5 μM to 1 μM of each primer and a program of: 98 °C, 1 min; 35 cycles of [98 °C, 15 s; 62 °C, 15 s; 72 °C, 1 min] unless otherwise noted. Many DNA reactions were purified with a QIAquick PCR Purification Kit (Qiagen) referred to below as Q-column purification or MinElute PCR Purification Kit (Qiagen) referred to below as M-column purification.

TALEN Construction

[00125] The canonical TALEN plasmids were constructed by the FLASH method[12] with each TALEN targeting 10-18 base pairs. N-terminal mutations were cloned by PCR with Q5 Hot Start Master Mix (NEB) [98 °C, 22 s; 62 °C, 15 s; 72 °C, 7 min]) using phosphorylated TAL-N1fwd (for N1), phosphorylated TAL-N2fwd (for N2), or phosphorylated TAL-N3fwd (for N3) and phosphorylated TALNrev as primers. 1 μL DpnI (NEB) was added and the reaction was incubated at 37 °C for 30 min then M-column purified. ~25 ng of eluted DNA was blunt-end ligated intramolecularly in 10 μL 2x Quick Ligase Buffer, 1 μL of Quick Ligase (NEB) in a total volume of 20 μL at room temperature (~21 °C) for 15 min. 1 μL of this ligation reaction was transformed into Top10 chemically competent cells (Invitrogen). C-terminal domain mutations were cloned by PCR using TAL-Cifwd and TAL-Cirev primers, then Q-column purified. ~1 ng of this eluted DNA was used as the template for PCR with TALCifwd and either TAL-Q3 (for Q3) or TAL-Q7 (for Q7) for primers, then Q-column purified. ~1 ng of this eluted DNA was used as the template for PCR with TAL-Cifwd and TAL-Ciirev for primers, then Qcolumn purified. ~1 μg of this DNA fragment was digested with HpaI and BamHI in 1x NEBuffer 4 and cloned into ~2 μg of desired TALEN plasmid pre-digested with HpaI and BamHI.

In Vitro TALEN Expression

[00126] TALEN proteins, all containing a 3xFLAG tag, were expressed by in vitro transcription/translation. 800 ng of TALEN-encoding plasmid or no plasmid ("empty lysate" control) was added to an in vitro transcription/translation reaction using the TNT® Quick Coupled Transcription/Translation System, T7 Variant (Promega) in a final volume of 20 μL at 30 °C for 1.5 h. Western blots were used to visualize protein using the anti-FLAG M2 monoclonal antibody (Sigma-Aldrich). TALEN concentrations were calculated by comparison to standard curve of 1 ng to 16 ng N-terminally FLAG-tagged bacterial alkaline phosphatase (Sigma-Aldrich).

In Vitro Selection for DNA Cleavage

[00127] Pre-selection libraries were prepared with 10 pmol of oligo libraries containing partially randomized target half-site sequences (CCR5A, ATM, or CCR5B) and fully randomized 10- to 24-bp spacer sequences (Table 10). Oligonucleotide libraries were separately circularized by incubation with 100 units of CircLigase II ssDNA Ligase (Epicentre) in 1x CircLigase II Reaction Buffer (33 mM Tris-acetate, 66 mM potassium acetate, 0.5 mM dithiothreitol, pH 7.5) supplemented with 2.5 mM MnCl2 in 20 μL total for 16 h at 60 °C then incubated at 80 °C for 10 min. 2.5 μL of each circularization reaction was used as a substrate for rolling-circle amplification at 30 °C for 16 h in a 50-μL reaction using the Illustra TempliPhi 100 Amplification Kit (GE Healthcare). The resulting concatemerized libraries were quantified with Quant-iT™ PicoGreen® dsDNA Kit (Invitrogen) and libraries with different spacer lengths were combined in an equimolar ratio.

[00128] For selections on the CCR5B sequence libraries, 500 ng of pre-selection library was digested for 2 h at 37 °C in 1x NEBuffer 3 with in vitro transcribed/translated TALEN plus empty lysate (30 μL total). For all CCR5B TALENs, in vitro transcribed/translated TALEN concentrations were quantified by Western blot (during the blot, TALENs were stored for 16 h at 4 °C) and then TALEN was added to 40 nM final concentration per monomer. For selections on CCR5A and ATM sequence libraries, the combined pre-selection library was further purified in a 300,000 MWCO spin column (Sartorius) with three 500-μL washes in 1x NEBuffer 3. 125 ng pre-selection library was digested for 30 min at 37 °C in 1x NEBuffer 3 with a total 24 μL of fresh in vitro transcribed/translated TALENs and empty lysate. For all CCR5A and ATM TALENs, 6 μL of in vitro transcription/translation left TALEN and 6 μL of right TALEN were used, corresponding to a final concentration in a cleavage reaction of 16 nM ± 2 nM or 12 nM ± 1.5 nM for CC5A or ATM TALENs, respectively. These TALEN concentrations were quantified by Western blot performed in parallel with digestion.

[00129] For all selections, the TALEN-digested library was incubated with 1 µL of 100 µg/µL RNase A (Qiagen) for 2 min and then Q-column purified. 50 µL of purified DNA was incubated with 3 µL of 10 mM dNTP mix (10 mM dATP, 10 mM dCTP, 10 mM dGTP, 10 mM dTTP) (NEB), 6 µL of 10x NEBuffer 2, and 1 µL of 5 U/µL Klenow Fragment DNA Polymerase (NEB) for 30 min at room temperature and Q-column purified. 50 µL of the eluted DNA was ligated with 2 pmol of heated and cooled #1 adapters containing barcodes corresponding to each sample (selections with different TALEN concentrations or constructs) (Table 10A). Ligation was performed in 1x T4 DNA Ligase Buffer (50 mM Tris-HCl, 10 mM MgCl2 , 1 mM ATP, 10 mM DTT, pH 7.5) with 1 µL of 400 U/µL T4 DNA ligase (NEB) in 60 µL total volume for 16 h at room temperature, then Q-column purified.

[00130] 6 µL of the eluted DNA was amplified by PCR in 150 µL total reaction volume (divided into 3x 50 µL reactions) for 14 to 22 cycles using the #2A adapter primers in Table 10A. The PCR products were purified by Q-column. Each DNA sample was quantified with Quant-iT™ PicoGreen® dsDNA Kit (Invitrogen) and then pooled into an equimolar mixture. 500 ng of pooled DNA was run a 5% TBE 18-well Criterion PAGE gel (BioRad) for 30 min at 200 V and DNAs of length ~230 bp (corresponding to 1.5 target site repeats plus adapter sequences) were isolated and purified by Qcolumn. ~2 ng of eluted DNA was amplified by PCR for 5 to 8 cycles with #2B adapter primers (Table 10A) and purified by M-column.

[00131] 10 µL of eluted DNA was purified using 12 µL of AMPure XP beads (Agencourt) and quantified with an Illumina/Universal Library Quantification Kit (Kapa Biosystems). DNA was prepared for high-throughput DNA sequencing according to Illumina instructions and sequenced using a MiSeq DNA Sequencer (Illumina) using a 12 pM final solution and 156-bp paired-end reads. To prepare the preselection library for sequencing, the pre-selection library was digested with 1 µL to 4 µL of appropriate restriction enzyme (CCR5A = Tsp45I, ATM = Acc65I, CCR5B = AvaI (NEB)) for 1 h at 37 °C then ligated as described above with 2 pmol of heated and cooled #1 library adapters (Table 10A). Pre-selection library DNA was prepared as described above using #2A library adapter primers and #2B library adapter primers in place of #2A adapter primers and #2B adapter primers, respectively (Table 10A). The resulting pre-selection library DNA was sequenced together with the TALEN-digested samples.

Discrete In Vitro TALEN Cleavage Assays

[00132] Discrete DNA substrates for TALEN digestion were constructed by combining pairs of oligonucleotides as specified in Table 9B with restriction cloning14 into pUC19 (NEB). Corresponding cloned plasmids were amplified by PCR (59 °C annealing for 15 s) for 24 cycles with pUC19Ofwd and pUC19Orev primers (Table 10B) and Q-column purified. 50 ng of amplified DNAs were digested in 1x NEBuffer 3 with 3 μL each of in vitro transcribed/translated TALEN left and right monomers (corresponding to a ~16 nM to ~12 nM final TALEN concentration), and 6 μL of empty lysate in a total reaction volume of 120 μL. The digestion reaction was incubated for 30 min at 37 °C, then incubated with 1 μL of 100 μg/μL RNase A (Qiagen) for 2 min and purified by M-column. The entire 10 μL of eluted DNA with glycerol added to 15% was analyzed on a 5% TBE 18-well Criterion PAGE gel (Bio-Rad) for 45 min at 200 V, then stained with 1x SYBR Gold (Invitrogen) for 10 min. Bands were visualized and quantified on an AlphaImager HP (Alpha Innotech).

Cellular TALEN Cleavage Assays

[00133] TALENs were cloned into mammalian expression vectors12 and the resulting TALEN vectors transfected into U2OS-EGFP cells as previously described.[12] Genomic DNA was isolated after 2 days as previously described.[12] For each assay, 50 ng of isolated genomic DNA was amplified by PCR [98 °C, 15s 67.5 °C, 15 s; 72 °C, 22s] for 35 cycles with pairs of primers with or without 4% DMSO as specified in Table 10C. The relative DNA content of the PCR reaction for each genomic site was quantified with Quant-iT™ PicoGreen ® dsDNA Kit (Invitrogen) and then pooled into an equimolar mixture, keeping no-TALEN and all TALEN-treated samples separate. DNA corresponding to 150 to 350 bp was purified by PAGE as described above.

[00134] 44 μL of eluted DNA was incubated with 5 μL of 1x T4 DNA Ligase Buffer and 1 μL of 10 U/μL Polynucleotide kinase (NEB) for 30 min at 37 °C and Q-column purified. 43 μL of eluted DNA was incubated with 1 μL of 10 mM dATP (NEB), 5 μL of 10x NEBuffer 2, and 1 μL of 5 U/μL DNA Klenow Fragment (3′→ 5′ exo–) (NEB) for 30 min at 37 °C and purified by M-column. 10μL of eluted DNA was ligated as above with 10 pmol of heated and cooled G (genomic) adapters (Table 10A). 8 μL of eluted DNA was amplified by PCR for 6 to 8 cycles with G-B primers containing barcodes corresponding to each sample. Each sample DNA was quantified with Quant-iT™ PicoGreen ® dsDNA Kit (Invitrogen) and then pooled into an equimolar mixture. The combined DNA was subjected to high throughput sequencing using a MiSeq as described above.

Data Analysis

[00135] Illumina sequencing reads were filtered and parsed with scripts written in Unix Bash as outlined in the Algorithms section. The source code is available upon request. Specificity scores were calculated as previously described.[14] Statistical analysis on the distribution of number of mutations in various TALEN selections in Table 3 was performed as previously described.[14] Statistical analysis of modified sites in Table 7 was performed as previously described.[14]

Algorithms

[00136] All scripts were written in bash or MATLAB.

Computational Filtering of Pre-selection Sequences and Selected Sequences

[00137] For Pre-selection Sequences

1) Search for 16 bp constant sequence (CCR5A = CGTCACGCTCACCACT, CCR5B = CCTCGGGACTCCACGCT, ATM = GGTACCCCACTCCGCGT) immediately after first 4 bases read (random bases), accepting only sequences with the 16bp constant sequence allowing for one mutation.
2) Search for 9 bp final sequence at a position at least the minimum possible full site length away and up to the max full site length away from constant sequence to confirm the presence of a full site, accept only sequences with this 9 bp final sequence. (Final sequence: CCR5A = CGTCACGCT, CCR5B = CCTCGGGAC, ATM = GGTACGTGC)
3) Search for best instances of each half site in the full site, accept any sequences with proper left and right half-site order of left then right.
4) Determine DNA spacer sequence between the two half sites, the single flanking nucleotide to left of the left half-site and single flanking nucleotide to right of the right half-site (sequence between half sites and constant sequences).
5) Filter by sequencing read quality scores, accepting sequences with quality scores of A or better across three fourths of the half site positions.

[00138] For Selected Sequences

1) Output to separate files all sequence reads and position quality scores of all sequences starting with correct 5 bp barcodes corresponding to different selection conditions.
2) Search for the initial 16 bp sequence immediately after the 5 bp barcode repeated at a position at least the minimum possible full site length away and up to the max full site

length away from initial sequence to confirm the presence of a full site with repeated sequence, accept only sequences with a 16bp repeat allowing for 1 mutation.

3) Search for 16 bp constant sequence within the full site, accept only sequences with a constant sequence allowing for one mutation. Parse sequence to start with constant sequence plus 5' sequence to second instance of repeated sequence then initial sequence after barcode to constant sequence resulting in constant sequences sandwiching the equivalent of one full site:

CONSTANT – LFLANK – LHS – SPACER – RHS – RFLANK – CONSTANT
LFLANK = Left Flank Sequence (designed as a single random base)
LHS = Left Half Site Sequence
RHS = Right Half Site Sequence
RFLANK = Right Flank Sequence (designed as a single random base)
CONSTANT = Constant Sequence (CCR5A = CGTCACGCTCACCACT, CCR5B = CCTCGGGACTCCACGCT, ATM = GGTACCCCACTCCGCGT)

4) Search for best instances of each half site in the full site, accept any sequences with proper left and right half-site order of left then right.

5) With half site positions determine corresponding spacer (sequence between the two half sites), left flank and right flank sequences (sequence between half sites and constant sequences).

6) Determine sequence end by taking sequence from the start of read after the 5 bp barcode sequence to the beginning of the constant sequence.

SEQUENCESTART – RHS – RFLANK – CONSTANT

7) Filter by sequencing read quality scores, accepting sequences with quality scores of A or better across three fourths of the half site positions.

8) Selected sequences were filtered by sequence end, by accepting only sequences with sequence ends in the spacer that were 2.5-fold more abundant than the amount of sequence end background calculated as the mean of the number of sequences with ends zero to five base pairs into each half-site from the spacer side (sequence end background number was calculated for both half sites with the closest half site to the sequence end utilized as sequence end background for comparison).

[00139] Computational Search for Genomic Off-Target Sites Related to the CCR5B Target Site

1) The Patmatch program[39] was used to search the human genome (GRCh37/hg19 build) for pattern sequences as follows: CCR5B left half-site sequence (L16, L13 or L10)

NNNNNNNNN... CCR5B right half-site sequence (R16, R13 or R10)[M,0,0] where number of Ns varied from 12 to 25

and M (indicating mutations allowed) varied from 0 to 14.

2) The number of output off-target sites were de-cumulated since the program outputs all sequences with X or fewer mutations, resulting in the number of off-target sites in the human genome that are a specific number of mutations away from the target site.

[00140] Identification of Indels in Sequences of Genomic Sites

1) For each sequence the primer sequence was used to identify the genomic site.

2) Sequences containing the reference genomic sequence corresponding to 8 bp to the left of the target site and reference genomic sequence 8 bp (or 6 bp for genomic sites at the very end of sequencing reads) to the right of the full target site were considered target site sequences.

3) Any target site sequences corresponding to the same size as the reference genomic site were considered unmodified and any sequences not the reference size were aligned with ClustalW[40] to the reference genomic site.

4) Aligned sequences with more than two insertions or two deletions in the DNA spacer sequence between the two half-site sequences were considered indels.

Results

Specificity Profiling of TALENs targeting CCR5 and ATM

[00141] We profiled the specificity of 41 heterodimeric TALEN pairs (hereafter referred to as TALENs) in total, comprising TALENs targeting left and right half-sites of various lengths and TALENs with different domain variants. Each of the 41 TALENs was designed to target one of three distinct sequences, which we refer to as CCR5A, CCR5B, or ATM, in two different human genes, CCR5 and ATM (Figure 7). We used an improved version of a previously described in vitro selection method[14] with modifications that increase the throughput and sensitivity of the selection (Figure 1B).

[00142] Briefly, preselection libraries of > 10^{12} DNA sequences each were digested with 3 nM to 40 nM of an in vitro translated TALEN. These concentrations correspond to ~20 to ~200 dimeric TALEN molecules per human cell nucleus,[21] a relatively low level of cellular protein expression.[22,23] Cleaved library members contained a free 5′ monophosphate that was captured by adapter ligation and isolated by gel purification (Figure 1B). In the control sample, all members of the pre-selection library were cleaved by a restriction endonuclease at a constant sequence to enable them to be captured by adapter ligation and isolated by gel purification. High-throughput sequencing of TALEN-treated or control samples surviving this selection process and computational analysis revealed the abundance of all TALEN-cleaved sequences as well as the abundance of the corresponding sequences before selection. The enrichment value for each library member surviving selection was calculated by dividing its post-selection sequence abundance by its preselection abundance. The pre-selection DNA libraries were sufficiently large that they each contain, in theory, at least ten copies of all possible DNA sequences with six or fewer mutations relative to the on-target sequence.

[00143] For all 41 TALENs tested, the DNA that survived the selection contained significantly fewer mean mutations in the targeted half-sites than were present in the pre-selection libraries (Table 3 and 4). For example, the mean number of mutations in DNA sequences surviving selection after treatment with TALENs targeting 18-bp left and right half-sites was 4.06 for CCR5A and 3.18 for ATM sequences, respectively, compared to 7.54 and 6.82 mutations in the corresponding pre-selection libraries (Figure 2A and 2B). For all selections, the on-target sequences were enriched by 8- to 640-fold (Table 5). To validate our selection results *in vitro*, we assayed the ability of the CCR5B TALENs targeting 13-bp left and right half-sites (L13+R13) to cleave each of 16 diverse off-target substrates (Figure 2E and 2F). The resulting discrete in vitro cleavage efficiencies correlated well with the observed enrichment values (Figure 2G).

[00144] To determine the specificity at each position in the TALEN target site for all four possible base pairs, a specificity score was calculated as the difference between pre-selection and post-selection base pair frequencies, normalized to the maximum possible change of the pre-selection frequency from complete specificity (defined as 1.0) to complete anti-specificity (defined as –1.0). For all TALENs tested, the targeted base pair at every position in both half-sites is preferred, with the sole exception of the base pair closest to the spacer for some ATM TALENs at the right-half site (Figure 2C, 2D and Figures 8 through 13). The 5′ T nucleotide recognized by the N-terminal domain is highly specified, and the 5′ DNA end (the N-terminal TALEN end) generally exhibits higher specificity than the 3' DNA end; both observations are consistent with previous reports.[24,25] Taken together, these results show that the selection data accurately predicts the efficiency of off-target TALEN cleavage in vitro, and that TALENs are overall highly specific across the entire target sequence.

TALEN Off-Target Cleavage in Cells

[00145] To test if off-target cleavage activities reported by the selection are relevant to off-target cleavage in cells, we used the in vitro selection results to train a machine-learning algorithm to generate potential TALEN off-target sites in the human genome.[26] This computational step was necessary because the preselection libraries cover all sequences with six or fewer mutations, while almost all potential off-target sites in the human genome for CCR5 and ATM sequences differ at more than six positions relative to the target sequence. The algorithm calculates the posterior probability of each nucleotide in each position of a target to occur in a sequence that was cleaved by the TALENs in opposition to sequences from the target library that were not observed to be cleaved.[27] These posterior probabilities were then used to score the likelihood that the TALEN used to train the algorithm would cleave every possible target sequence in the human genome with monomer spacing of 10 to 30 bps. Using the machine-learning algorithm, we identified 36 CCR5A and 36 ATM TALEN off-target sites that differ from the on-target sequence at seven to fourteen positions (Table 6).

[00146] The 72 best-scoring genomic off-target sites for CCR5A and ATM TALENs were amplified from genomic DNA purified from human U2OS-EGFP cells12 expressing either CCR5A or ATM TALENs.[3] Sequences containing insertions or deletions of three or more base pairs in the DNA spacer of the potential genomic off-target sites and present in significantly greater numbers in the TALEN-treated samples versus the untreated control sample were considered TALEN-induced modifications. Of the 35 CCR5A off-target sites that we successfully amplified, we identified six off-target sites with TALEN-induced modifications; likewise, of the 31 ATM off-target sites that we successfully amplified, we observed seven off-target sites with TALEN-induced modifications (Figure 3 and Table 7). The inspection of modified on-target and off-target sites yielded a prevalence of deletions ranging from three to dozens of base pairs (Figure 3), consistent with previously described characteristics of TALEN-induced genomic modification.[28]

[00147] These results collectively indicate that the in vitro selection data, processed through a machine-learning algorithm, can predict bona fide off-target substrates that undergo TALEN-induced modification in human cells. TALE Repeats Productively Bind Base Pairs with Relative Independence The extensive number of quantitatively characterized off-target substrates in the selection data enabled us to assess whether mutations at one position in the target sequence affect the ability of TALEN repeats to productively bind other positions. We generated an expected enrichment value for every possible double-mutant sequence for the

L13+R13 CCR5B TALENs assuming independent contributions from the two corresponding single-mutation enrichments. In general, the predicted enrichment values closely resembled the actual observed enrichment values for each double-mutant sequence (Figure 14A), suggesting that component single mutations independently contributed to the overall cleavability of double-mutant sequences. The difference between the observed and predicted double-mutant enrichment values was relatively independent of the distance between the two mutations, except that two neighboring mismatches were slightly better tolerated than would be expected (Figure 14B).

[00148] To determine the potential interdependence of more than two mutations, we evaluated the relationship between selection enrichment values and the number of mutations in the post-selection target for the L13+R13 CCR5B TALEN (Figure 4A, black line). For 0 to 5 mutations, enrichment values closely followed a simple exponential function of the mean number of mutations (m) (Table 8). This relationship is consistent with a model in which each successive mutation reduces the binding energy by a constant amount (ΔG), resulting in an exponential decrease in TALEN binding (Keq(m)) such that Keq(m) ~ eΔG*m. The observed exponential relationship therefore suggests that the mean reduction in binding energy from a typical mismatch is independent of the number of mismatches already present in the TALEN:DNA interaction. Collectively, these results indicate that TALE repeats bind their respective DNA base pairs independently beyond a slightly increased tolerance for adjacent mismatches.

Longer TALENs are Less Specific Per Recognized Base Pair

[00149] The independent binding of TALE repeats simplistically predicts that TALEN specificity per base pair is independent of target-site length. To experimentally characterize the relationship between TALE array length and off-target cleavage, we constructed TALENs targeting 10, 13, or 16 bps (including the 5′ T) for both the left (L10, L13, L16) and right (R10, R13, R16) half-sites. TALENs representing all nine possible combinations of left and right CCR5B TALENs were subjected to in vitro selection. The results revealed that shorter TALENs have greater specificity per targeted base pair than longer TALENs (Table 3). For example, sequences cleaved by the L10+R10 TALEN contained a mean of 0.032 mutations per recognized base pair, while those cleaved by the L16+R16 TALEN contained a mean of 0.067 mutations per recognized base pair. For selections with the longest CCR5B TALENs targeting 16+16 base pairs or CCR5A and ATM TALENs targeting 18+18 bp, the mean selection enrichment values do not follow a simple exponential decrease as function of mutation number (Figure 4A and Table 8).

[00150] We hypothesized that excess binding energy from the larger number of TALE repeats in longer TALENs reduces specificity by enabling the cleavage of sequences with more mutations, without a corresponding increase in the cleavage of sequences with fewer mutations, because the latter are already nearly completely cleaved. Indeed, the in vitro cleavage efficiencies of discrete DNA sequences for these longer TALENs are independent of the presence of a small number of mutations in the target site (Figures 5C-5F), suggesting there is nearly complete binding and cleavage of sequences containing few mutations. Likewise, higher TALEN concentrations also result in decreased enrichment values of sequences with few mutations while increasing the enrichment values of sequences with many mutations (Table 5). These results together support a model in which excessive TALEN binding arising from either long TALE arrays or high TALEN concentrations decreases observed TALEN DNA cleavage specificity of each recognized base pair.

Longer TALENs Induce Less Off-Target Cleavage in a Genomic Context

[00151] Although longer TALENs are more tolerant of mismatched sequences (Figure 4A) than shorter TALENs, in the human genome there are far fewer closely related off-target sites for a longer target site than for a shorter target site (Figure 4B). Since off-target site abundance and cleavage efficiency both contribute to the number of off-target cleavage events in a genomic context, we calculated overall genome cleavage specificity as a function of TALEN length by multiplying the extrapolated mean enrichment value of mutant sequences of a given length with the number of corresponding mutant sequences in the human genome. The decrease in potential off-target site abundance resulting from the longer target site length is large enough to outweigh the decrease in specificity per recognized base pair observed for longer TALENs (Figure 4C). As a result, longer TALENs are predicted to be more specific against the set of potential cleavage sites in the human genome than shorter TALENs for the tested TALEN lengths targeting 20- to 32-bp sites.

Engineering TALENs with Improved Specificity

[00152] The findings above suggest that TALEN specificity can be improved by reducing non-specific DNA binding energy beyond what is needed to enable efficient on-target cleavage. The most widely used 63-aa C-terminal domain between the TALE repeat array and the FokI nuclease domain contains ten cationic residues. We speculated that reducing the cationic charge of the canonical TALE C-terminal domain would decrease non-specific DNA binding[29] and improve TALEN specificity.

[00153] We constructed two C-terminal domain variants in which three ("Q3", K788Q, R792Q, R801Q) or seven ("Q7", K777Q, K778Q, K788Q, R789Q, R792Q, R793Q, R801Q) cationic Arg or Lys residues in the canonical 63-aa C-terminal domain were mutated to Gln. We performed in vitro selections on CCR5A and ATM TALENs containing the canonical, engineered Q3, and engineered Q7 C-terminal domains, as well as a previously reported 28-aa truncated C-terminal domain[5] with a theoretical net charge identical to that of the Q7 C-terminal domain (–1).

[00154] The on-target sequence enrichment values for the CCR5A and ATM selections increased substantially as the net charge of the C-terminal domain decreased (Figure 5A and 5B). For example, the ATM selections resulted in on-target enrichment values of 510, 50, and 20 for the Q7, Q3, and canonical 63-aa C-terminal variants, respectively. These results suggest that the TALEN variants in which cationic residues in the C-terminal domain have been partially replaced by neutral residues or completely removed are substantially more specific in vitro than the TALENs that containing the canonical 63-aa C-terminal domain. Similarly, mutating one, two, or three cationic residues in the TALEN N-terminus to Gln also increased cleavage specificity (Table 5, and Figures 8-11).

[00155] In order to confirm the greater DNA cleavage specificity of Q7 over canonical 63-aa C-terminal domains in vitro, a representative collection of 16 off-target DNA substrates were digested in vitro with TALENs containing either canonical or engineered Q7 C-terminal domains. ATM and CCR5A TALENs with the canonical 63-aa C-terminal domain TALEN demonstrate comparable in vitro cleavage activity on target sites with zero, one, or two mutations (Figures 5C-5F). In contrast, for 11 of the 16 off-target substrates tested, the engineered Q7 TALEN variants showed substantially higher (~4-fold or greater) discrimination against off-target DNA substrates with one or two mutations than the canonical 63-aa C-terminal domain TALENs, even though the Q7 TALENs cleaved their respective on-target sequences with comparable or greater efficiency than TALENs with the canonical 63-aa C-terminal domains (Figures 5C-5F). Overall, the discrete cleavage assays are consistent with the selection results and indicate that TALENs with engineered Q7 C-terminal domains are substantially more specific than TALENs with canonical 63-aa C-terminal domains in vitro.

Improved Specificity of Engineered TALENs in Human Cells

[00156] To determine if the increased specificity of the engineered TALENs observed in vitro also applies in human cells, TALEN-induced modification rates of the on-target and top 36 predicted off-target sites were measured for CCR5A and ATM TALENs containing all six possible combinations of the canonical 63-aa, Q3, or Q7 C-terminal domains and the EL/KK or ELD/KKR FokI domains (12 TALENs total).

[00157] For both FokI variants, the TALENs with Q3 C-terminal domains demonstrate significant on-target activities ranging from 8% to 24% modification, comparable to the activity of TALENs with the canonical 63-aa C-terminal domains. TALENs with canonical 63-aa or Q3 C-terminal domains and the ELD/KKR FokI domain are both more active in modifying the CCR5A and ATM on-target site in cells than the corresponding TALENs with the Q7 C-terminal domain by ~5-fold (Figure 6 and Table 7).

[00158] Consistent with the improved specificity observed in vitro, the engineered Q7 TALENs are more specific than the Q3 variants, which in turn are more specific than the canonical 63-aa C-terminal domain TALENs. Compared to the canonical 63-aa C-terminal domains, TALENs with Q3 C-terminal domains demonstrate a mean increase in cellular specificity (defined as the ratio of the cellular modification percentage for on-target to off-target sites) of more than 13-fold and more than 9-fold for CCR5A and ATM sites, respectively, with the ELD/KKR FokI domain (Table 7). These mean improvements can only be expressed as lower limits due to the absence or near-absence of observed cleavage events by the engineered TALENs for many off-target sequences. For the most abundantly cleaved off-target site (CCR5A off-target site #5), the Q3 C-terminal domain is 34-fold more specific (Figure 6), and the Q7 C-terminal domain is > 116-fold more specific, than the canonical 63-aa C-terminal domain.

[00159] Together, these results reveal that for targeting the CCR5 and ATM sequences, replacing the canonical 63-aa C-terminal domain with the engineered Q3 C-terminal domain results in comparable activity for the on-target site in cells, a 34-fold improvement in specificity in cells for the most readily cleaved off-target site, and a consistent increase in specificity for other off-target sites. When less activity is required, the engineered Q7 C-terminal domain offers additional gains in specificity.

Engineering N-Terminal Domains for Improved TALEN DNA Cleavage Specificity

[00160] The model of TALEN binding and specificity described herein predicts that reducing excess TALEN binding energy will increase TALEN DNA cleavage specificity. To

further test this prediction and potentially further augment TALEN specificity, we mutated one ("N1", K150Q), two ("N2", K150Q and K153Q), or three ("N3", K150Q, K153Q, and R154Q) Lys or Arg residues to Gln in the N-terminal domain of TALENs targeting CCR5A and ATM. These N-terminal residues have been shown in previous studies to bind non-specifically to DNA, and mutations at these specific residues to neutralize the cationic charge decrease non-specific DNA binding energy.[33] We hypothesized the reduction in non-specific binding energy from these N-terminal mutations would decrease excess TALEN binding energy resulting in increased specificity. In vitro selections on these three TALEN variants revealed that the less cationic N-terminal TALENs indeed exhibit greater enrichment values of on-target cleavage (Table 5).

Effects of N-Terminal and C-Terminal Domains and TALEN Concentration on Specificity

[00161] All TALEN constructs tested specifically recognize the intended base pair across both half-sites (Figures 8 to 13), except that some of the ATM TALENs do not specifically interact with the base pair adjacent to the spacer (targeted by the most C-terminal TALE repeat) (Figures 10 and 11). To compare the broad specificity profiles of canonical TALENs with those containing engineered C-terminal or N-terminal domains, the specificity scores of each target base pair from selections using CCR5A and ATM TALENs with the canonical, Q3, or Q7 C-terminal domains and N1, N2, or N3 N-terminal domains were subtracted by the corresponding specificity scores from selections on the canonical TALEN (canonical 63-aa C-terminal domain, wild-type N-terminal domain).

[00162] The results are shown in Figure 15. Mutations in the C-terminal domain that increase specificity did so most strongly in the middle and at the C-terminal end of each half-site. Likewise, the specificity-increasing mutations in the N-terminus tended to increase specificity most strongly at positions near the TALEN N-terminus (5' DNA end) although mutations in the N-terminus of ATM TALEN targeting the right half-site did not significantly alter specificity. These results are consistent with a local binding compensation model in which weaker binding at either terminus demands increased specificity in the TALE repeats near this terminus. To characterize the effects of TALEN concentration on specificity, the specificity scores from selections of ATM and CCR5A TALENs performed at three different concentrations ranging from 3 nM to 16 nM were each subtracted by the specificity scores of corresponding selections performed at the highest TALEN concentration assayed, 24 nM for ATM, or 32 nM for CCR5A. The results (Figure 15) indicate that specificity scores increase fairly uniformly across the half-sites as the concentration of TALEN is decreased.

DNA Spacer-Length and Cut-Site Preferences

[00163] To assess the spacer-length preference of various TALEN architectures (C-terminal mutations, N-terminal mutations, and FokI variants) and various TALEN concentrations, the enrichment values of library members with 10- to 24- base pair spacer lengths in each of the selections with CCR5A and ATM TALEN with various combinations of the canonical, Q3, Q7, or 28-aa C-terminal domains; N1, N2, or N3 N-terminal mutations; and the EL/KK or ELD/KKR FokI variants at 4 nM to 32 nM CCR5A and ATM TALEN were calculated (Figure 16). All of the tested concentrations, N-terminal variants, C-terminal variants, and FokI variants demonstrated a broad DNA spacer-length preference ranging from 14- to 24- base pairs with three notable exceptions. First, the CCR5A 28-aa C-terminal domain exhibited a much narrower DNA spacer-length preference than the broader DNA spacer-length preference of the canonical C-terminal domain, consistent with previous reports.[34-36] Second, the CCR5A TALENs containing Q7 C-terminal domains showed an increased tolerance for 12-base spacers compared to the canonical C-terminal domain variant (Figure 16). This slightly broadened spacer-length preference may reflect greater conformational flexibility in the Q7 C-terminal domain, perhaps resulting from a smaller number of non-specific protein:DNA interactions along the TALEN:DNA interface. Third, the ATM TALENs with Q7 C-terminal domains and the ATM TALENs with N3 mutant N-terminal domains showed a narrowed spacer preference.

[00164] These more specific TALENs (Table 5) with lower DNA-binding affinity may have faster off-rates that are competitive with the rate of cleavage of non-optimal DNA spacer lengths, altering the observed spacer-length preference. While previous reports have focused on the length of the TALEN C-terminal domain as a primary determinant of DNA spacer-length preference, these results suggest the net charge of the C-terminal domain as well as overall DNA-binding affinity can also affect TALEN spacer-length preference.

[00165] We also characterized the location of TALEN DNA cleavage within the spacer. We created histograms reporting the number of spacer DNA bases observed preceding the right half-site in each of the sequences from the selections with CCR5A and ATM TALEN with various combinations of the canonical, Q3, Q7, or 28-aa C-terminal domains; N1, N2, or N3 N-terminal mutations; and the EL/KK or ELD/KKR FokI variants (Figure 17). The peaks in the histogram were interpreted to represent the most likely locations of DNA cleavage within the spacer. The cleavage positions are dependent on the length of the DNA spacer between the TALEN binding half-sites, as might be expected from conformational constraints imposed by the TALEN C-terminal domain and DNA spacer lengths.

Discussion

[00166] The in vitro selection of 41 TALENs challenged with 10^{12} closed related off-target sequences and subsequent analysis inform our understanding of TALEN specificity through four key findings: (i) TALENs are highly specific for their intended target base pair at all positions with specificity increasing near the N-terminal TALEN end of each TALE repeat array (corresponding to the 5′ end of the bound DNA); (ii) longer TALENs are more specific in a genomic context while shorter TALENs have higher specificity per nucleotide; (iii) TALE repeats each bind their respective base pair relatively independently; and (iv) excess DNA-binding affinity leads to increased TALEN activity against off-target sites and therefore decreased specificity.

[00167] The observed decrease in specificity for TALENs with more TALE repeats or more cationic residues in the C-terminal domain or N-terminus are consistent with a model in which excess TALEN binding affinity leads to increased promiscuity. Excess binding energy could also explain the previously reported promiscuity at the 5′ terminal T of TALENs with longer C-terminal domains[30] and is also consistent with a report of higher TALEN protein concentrations resulting in more off-target site cleavage in vivo.[9] While decreasing TALEN protein expression in cells in theory could reduce off-target cleavage, the Kd values of some TALEN constructs for their target DNA sequences are likely already comparable to, or below, the theoretical minimum protein concentration in a human cell nucleus, ~0.2 nM.[21]

[00168] The difficulty of improving the specificity of such TALENs by lowering their expression levels, coupled with the need to maintain sufficient TALEN concentrations to effect desired levels of on-target cleavage, highlight the value of engineering TALENs with higher intrinsic specificity such as those described in this work. Our findings suggest that mutant C-terminal domains with reduced non-specific DNA binding may be used to fine-tune the DNA-binding affinity of TALENs such that on-target sequences are cleaved efficiently but with minimal excess binding energy, resulting in better discrimination between on-target and off-target sites. Since TALENs targeting up to 46 total base pairs have been shown to be active in cells,[15] the results presented here are consistent with the notion that specificity may be even further improved by engineering TALENs with a combination of mutant N-terminal and C-terminal domains that impart reduced non-specific DNA binding, a greater number of

TALE repeats to contribute additional on-target DNA binding, and the more specific (but lower-affinity) NK RVD to recognize G.[25,31]

[00169] Our study has identified more bona fide TALEN genomic off-target sites than other studies using methods such as SELEX or integrase-deficient lentiviral vectors (IDLVs).[32] Our model and the resulting improved TALENs would have been difficult to derive from cellular off-target cleavage methods, which are intrinsically limited by the small number of sequences closely related to a target sequence of interest that are present in a genome, or from SELEX experiments with monomeric TALE repeat arrays,[5] which do not measure DNA cleavage activity and therefore does not characterize active, dimeric TALENs. In contrast, each TALEN in this study was evaluated for its ability to cleave any of 10^{12} close variants of its on-target sequence, a library size several orders of magnitude greater than the number of different sequences in a mammalian genome. This dense coverage of off-target sequence space enabled the elucidation of detailed relationships between DNA-cleavage specificity and target base pair position, TALE repeat length, TALEN concentration, mismatch location, and engineered TALEN domain composition.

EXAMPLE 2

[00170] A number of TALENs were generated in which at least one cationic amino acid residue of the canonical N-terminal domain sequence was replaced with an amino acid residue that exhibits no charge or a negative charge at physiological pH. The TALENs comprised substitutions of glycine (G) and/or glutamine (Q) in their N-terminal domains (see Figure 18). An evaluation of the cutting preferences of the engineered TALENs demonstrated that mutations to glycine (G) are equivalent to glutamine (Q). Mutating the positively charged amino acids in the TALEN N-terminal domain (K150Q, K153Q, and R154Q) result in similar decreases in binding affinity and off-target cleavage for mutations to either Q or G. For example, TALENs comprising the M3 and M4 N-terminus, which comprises the same amino acid (R154) mutated to either Q or G, respectively, demonstrated roughly equivalent amounts of cleavage. Similarly TALENs comprising the M6 and M8 N-terminus, varying only in whether Q or G substitutions were introduced at positions K150 and R154, and TALENs comprising the M9 and M10 N-terminus, varying only in whether Q or G substitutions were introduced at positions K150, K153, and R154, showed similar cleavage activity.

EXAMPLE 3

[00171] A plasmid was generated for cloning and expression of engineered TALENs as provided herein. A map of the plasmid is shown in Figure 19. The plasmid allows for the modular cloning of N-terminal and C-terminal domains, *e.g.*, engineered domains as provided herein, and for TALE repeats, thus generating a recombinant nucleic acid encoding the desired engineered TALEN. The plasmid also encodes amino acid tags, e.g., an N-terminal FLAG tag and a C-terminal V5 tag, which can, optionally be utilized for purification or detection of the encoded TALEN. Use of these tags is optional and one of skill in the art will understand that the TALEN-encoding sequences will have to be cloned in-frame with the tag-encoding sequences in order to result in a tagged TALEN protein being encoded.

[00172] An exemplary sequence of a cloning vector as illustrated in Figure 19 is provided below. Those of skill in the art will understand that the sequence below is illustrative of an exemplary embodiment and does not limit this disclosure.

```
>pExpCCR5A-L18_(63aa)
GACGGATCGGGAGATCTCCCGATCCCCTATGGTCGACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCC
AGTATCTGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAG
GCTTGACCGACAATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGAT
ATACGCGTTGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATA
TGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGT
CAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGACTATTTACGGT
AAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAAT
GGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCA
TCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTC
CAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTA
ACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTCTCTGGC
TAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAGCTGGCTAGC
ACCATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAGATG
GCCCCCAAGAAGAAGAGGAAGGTGGGCATTCACCGCGGGGTACCTATGGTGGACTTGAGGACACTCGGTTATTCG
CAACAGCAACAGGAGAAAATCAAGCCTAAGGTCAGGAGCACCGTCGCGCAACACCACGAGGCGCTTGTGGGGCAT
GGCTTCACTCATGCGCATATTGTCGCGCTTTCACAGCACCCTGCGGCGCTTGGGACGGTGGCTGTCAAATACCAA
GATATGATTGCGGCCCTGCCCGAAGCCACGCACGAGGCAATTGTAGGGGTCGGTAAACAGTGGTCGGGAGCGCGA
GCACTTGAGGCGCTGCTGACTGTGGCGGGTGAGCTTAGGGGGCCTCCGCTCCAGCTCGACACCGGGCAGCTGCTG
AAGATCGCGAAGAGAGGGGGAGTAACAGCGGTAGAGGCAGTGCACGCCTGGCGCAATGCGCTCACCGGGGCCCCC
TTGAACCTGACCCCAGACCAGGTAGTCGCAATCGCGTCAAACGGAGGGGGAAAGCAAGCCCTGGAAACCGTGCAA
AGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTACACCGGAGCAAGTCGTGGCCATTGCATCCCACGACGGT
GGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTCTGTCAAGCCCACGGGCTGACTCCCGATCAA
GTTGTAGCGATTGCGTCGAACATTGGAGGGAAACAAGCATTGGAGACTGTCCAACGGCTCCTTCCCGTGTTGTGT
```

```
CAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCGAATGGCGGCGGTAAGCAGGCGCTGGAAACA
GTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAATCGCGTCAAAC
GGAGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTACACCG
GAGCAAGTCGTGGCCATTGCAAGCAACATCGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTT
CTCTGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGCATGACGGAGGGAAACAAGCATTG
GAGACTGTCCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCC
TCCAATATTGGCGGTAAGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTG
ACCCCAGACCAGGTAGTCGCAATCGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTG
CCGGTCCTTTGTCAAGACCACGGCCTTACACCGGAGCAAGTCGTGGCCATTGCATCCCACGACGGTGGCAAACAG
GCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTCTGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCG
ATTGCGTCCAACGGTGGAGGGAAACAAGCATTGGAGACTGTCCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCAC
GGTTTGACGCCTGCACAAGTGGTCGCCATCGCCAACAACAACGGCGGTAAGCAGGCGCTGGAAACAGTACAGCGC
CTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAATCGCGTCACATGACGGGGGA
AAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTACACCGGAGCAAGTC
GTGGCCATTGCAAGCAACATCGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTCTGTCAA
GCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGAATAACAATGGAGGGAAACAAGCATTGGAGACTGTC
CAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCAGCCATGAT
GGCGGTAAGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACACCCGAA
CAGGTGGTCGCCATTGCTTCTAATGGGGGAGGACGGCCAGCCTTGGAGTCCATCGTAGCCCAATTGTCCAGGCCC
GATCCCGCGTTGGCTGCGTTAACGAATGACCATCTGGTGGCGTTGGCATGTCTTGGTGGACGACCCGCGCTCGAT
GCAGTCAAAAAGGGTCTGCCTCATGCTCCCGCATTGATCAAAAGAACCAACCGGCGGATTCCCGAGAGAACTTCC
CATCGAGTCGCGGGATCCCAACTAGTCAAAAGTGAACTGGAGGAGAAGAAATCTGAACTTCGTCATAAATTGAAA
TATGTGCCTCATGAATATATTGAATTAATTGAAATTGCCAGAAATTCCACTCAGGATAGAATTCTTGAAATGAAG
GTAATGGAATTTTTTATGAAAGTTTATGGATATAGAGGTAAACATTTGGGTGGATCAAGGAAACCGGACGGAGCA
ATTTATACTGTCGGATCTCCTATTGATTACGGTGTGATCGTGGATACTAAAGCTTATAGCGGAGGTTATAATCTG
CCAATTGGCCAAGCAGATGAAATGGAGCGATATGTCGAAGAAAATCAAACACGAAACAAACATATCAACCCTAAT
GAATGGTGGAAAGTCTATCCATCTTCTGTAACGGAATTTAAGTTTTTATTTGTGAGTGGTCACTTTAAAGGAAAC
TACAAAGCTCAGCTTACACGATTAAATCATATCACTAATTGTAATGGAGCTGTTCTTAGTGTAGAAGAGCTTTTA
ATTGGTGGAGAAATGATTAAAGCCGGCACATTAACCTTAGAGGAAGTGAGACGGAAATTTAATAACGGCGAGATA
AACTTTTAAGGGCCCTTCGAAGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACGCGTACCGGTCAT
CATCACCATCACCATTGAGTTTAAACCCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTT
TGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATT
GCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGG
GAAGACAATAGCAGGCATGCTGGGGATGCGGTGGGCTCTATGGCTTCTGAGGCGGAAAGAACCAGCTGGGGCTCT
AGGGGGTATCCCCACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCT
ACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCC
CGTCAAGCTCTAAATCGGGGCATCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTT
GATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACG
TTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAA
GGGATTTTGGGGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTAATTCTGT
GGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGGCTCCCCAGGCAGGCAGAAGTATGCAAAGCATGCATCTC
```

AATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAAT

TAGTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCG

CCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCTGCCTCTGAGCTATTCCAGAAGTAG

TGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTCCCGGGAGCTTGTATATCCATTTTCGGATCTGAT

CAGCACGTGTTGACAATTAATCATCGGCATAGTATATCGGCATAGTATAATACGACAAGGTGAGGAACTAAACCA

TGGCCAAGCCTTTGTCTCAAGAAGAATCCACCCTCATTGAAAGAGCAACGGCTACAATCAACAGCATCCCCATCT

CTGAAGACTACAGCGTCGCCAGCGCAGCTCTCTCTAGCGACGGCCGCATCTTCACTGGTGTCAATGTATATCATT

TTACTGGGGGACCTTGTGCAGAACTCGTGGTGCTGGGCACTGCTGCTGCTGCGGCAGCTGGCAACCTGACTTGTA

TCGTCGCGATCGGAAATGAGAACAGGGGCATCTTGAGCCCCTGCGGACGGTGTCGACAGGTGCTTCTCGATCTGC

ATCCTGGGATCAAAGCGATAGTGAAGGACAGTGATGGACAGCCGACGGCAGTTGGGATTCGTGAATTGCTGCCCT

CTGGTTATGTGTGGGAGGGCTAAGCACTTCGTGGCCGAGGAGCAGGACTGACACGTGCTACGAGATTTCGATTCC

ACCGCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAGCGCGGG

GATCTCATGCTGGAGTTCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGC

ATCACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCT

TATCATGTCTGTATACCGTCGACCTCTAGCTAGAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAAT

TGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTG

AGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAA

TGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTG

CGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGG

GATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCG

TTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACA

GGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACC

GGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCAATGCTCACGCTGTAGGTATCTCAGTTCG

GTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGT

AACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGC

AGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTA

TTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACC

ACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCT

TTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCA

AAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACT

TGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTT

GCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCG

CGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGT

CCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAAT

AGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGC

TCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCT

CCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACT

GTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGG

CGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATC

ATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACT

CGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAAT

```
GCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGC
ATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCG
CGCACATTTCCCCGAAAAGTGCCACCTGACGTC (SEQ ID NO: 42)
```

REFERENCES

1. Moscou, M.J. & Bogdanove, A.J. A simple cipher governs DNA recognition by TAL effectors. Science 326, 1501 (2009).
2. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. Science 326, 1509-1512 (2009).
3. Doyon, Y. *et al.* Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. Nat Methods 8, 74-79 (2011).
4. Cade, L. *et al.* Highly efficient generation of heritable zebrafish gene mutations using homo- and heterodimeric TALENs. Nucleic Acids Res 40, 8001-8010 (2012).
5. Miller, J.C. *et al.* A TALE nuclease architecture for efficient genome editing. Nat Biotechnol 29, 143-148 (2011).
6. Bedell, V.M. *et al.* In vivo genome editing using a high-efficiency TALEN system. Nature 491, 114-118 (2012).
7. Hockemeyer, D. *et al.* Genetic engineering of human pluripotent cells using TALE nucleases. Nat Biotechnol 29, 731-734 (2011).
8. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res 39, e82 (2011).
9. Tesson, L. *et al.* Knockout rats generated by embryo microinjection of TALENs. Nat Biotechnol 29, 695-696 (2011).
10. Moore, F.E. *et al.* Improved somatic mutagenesis in zebrafish using transcription activator-like effector nucleases (TALENs). PLoS One 7, e37877 (2012).
11. Wood, A.J. *et al.* Targeted genome editing across species using ZFNs and TALENs. Science 333, 307 (2011).
12. Reyon, D. *et al.* FLASH assembly of TALENs for high-throughput genome editing. Nat Biotechnol 30, 460-465 (2012).
13. Mussolino, C. *et al.* A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. Nucleic Acids Res 39, 9283-9293 (2011).
14. Pattanayak, V., Ramirez, C.L., Joung, J.K. & Liu, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. Nat Methods 8, 765-770 (2011).

15. Li, T. *et al.* Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. Nucleic Acids Res 39, 6315-6325 (2011).
16. Ding, Q. *et al.* A TALEN Genome-Editing System for Generating Human Stem Cell-Based Disease Models. Cell Stem Cell (2012).
17. Lei, Y. *et al.* Efficient targeted gene disruption in Xenopus embryos using engineered transcription activator-like effector nucleases (TALENs). Proc Natl Acad Sci U S A 109, 17484-17489 (2012).
18. Kim, Y. *et al.* A library of TAL effector nucleases spanning the human genome. Nat Biotechnol 31, 251-258 (2013).
19. Dahlem, T.J. *et al.* Simple methods for generating and detecting locus-specific mutations induced with TALENs in the zebrafish genome. PLoS Genet 8, e1002861 (2012).
20. Osborn, M.J. *et al.* TALEN-based Gene Correction for Epidermolysis Bullosa. Molecular Therapy (2013).
21. Maul, G.G. & Deaven, L. Quantitative determination of nuclear pore complexes in cycling cells with differing DNA content. J Cell Biol 73, 748-760 (1977).
22. Huang, B. *et al.* Counting low-copy number proteins in a single cell. Science 315, 81-84 (2007).
23. Beck, M. *et al.* The quantitative proteome of a human cell line. Mol Syst Biol 7, 549 (2011).
24. Meckler, J.F. *et al.* Quantitative analysis of TALE-DNA interactions suggests polarity effects. Nucleic Acids Res (2013).
25. Christian, M.L. *et al.* Targeting G with TAL effectors: a comparison of activities of TALENs constructed with NN and NK repeat variable di-residues. PLoS One 7, e45383 (2012).
26. Sander, J.D. *et al.* Abstraction of zinc finger nuclease cleavage profiles reveals an expanded landscape of off-target mutations. Submitted (2013).
27. Witten, I.H. & Frank, E. Data mining: practical machine learning tools and techniques, Edn. 2nd. (Morgan Kaufman, San Francisco; 2005).
28. Kim, Y., Kweon, J. & Kim, J.S. TALENs and ZFNs are associated with different mutation signatures. Nat Methods 10, 185 (2013).
29. McNaughton, B.R., Cronican, J.J., Thompson, D.B. & Liu, D.R. Mammalian cell penetration, siRNA transfection, and DNA transfection by supercharged proteins. Proc Natl Acad Sci U S A 106, 6111-6116 (2009).

30. Sun, N., Liang, J., Abil, Z. & Zhao, H. Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. Mol Biosyst 8, 1255-1263 (2012).
31. Cong, L., Zhou, R., Kuo, Y.C., Cunniff, M. & Zhang, F. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. Nat Commun 3, 968 (2012).
32. Gabriel, R. *et al.* An unbiased genome-wide analysis of zinc-finger nuclease specificity. Nat Biotechnol 29, 816-823 (2011).
33. Gao, H., Wu, X., Chai, J. & Han, Z. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. Cell Res 22, 1716-1720 (2012).
34. Li, T. *et al.* Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. Nucleic Acids Res 39, 6315-6325 (2011).
35. Miller, J.C. *et al.* A TALE nuclease architecture for efficient genome editing. Nat Biotechnol 29, 143-148 (2011).
36. Mahfouz, M.M. *et al.* De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. Proc Natl Acad Sci U S A 108, 2623-2628 (2011).
37. Pattanayak, V., Ramirez, C.L., Joung, J.K. & Liu, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. Nat Methods 8, 765-770 (2011).
38. Sander, J.D. *et al.* Abstraction of zinc finger nuclease cleavage profiles reveals an expanded landscape of off-target mutations. Submitted (2013).
39. Yan, T. *et al.* PatMatch: a program for finding patterns in peptide and nucleotide sequences. Nucleic Acids Res 33, W262-266 (2005).
40. Larkin, M.A. *et al.* Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948 (2007).

[00173] All publications, patents, patent applications, publication, and database entries (*e.g.*, sequence database entries) mentioned herein, *e.g.*, in the Background, Summary, Detailed Description, Examples, and/or References sections, are hereby incorporated by reference in their entirety as if each individual publication, patent, patent application, publication, and database entry was specifically and individually incorporated herein by reference. In case of conflict, the present application, including any definitions herein, will control.

EQUIVALENTS AND SCOPE

[00174] Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. The scope of the present invention is not intended to be limited to the above description, but rather is as set forth in the appended claims.

[00175] In the claims articles such as "a," "an," and "the" may mean one or more than one unless indicated to the contrary or otherwise evident from the context. Claims or descriptions that include "or" between one or more members of a group are considered satisfied if one, more than one, or all of the group members are present in, employed in, or otherwise relevant to a given product or process unless indicated to the contrary or otherwise evident from the context. The invention includes embodiments in which exactly one member of the group is present in, employed in, or otherwise relevant to a given product or process. The invention also includes embodiments in which more than one, or all of the group members are present in, employed in, or otherwise relevant to a given product or process.

[00176] Furthermore, it is to be understood that the invention encompasses all variations, combinations, and permutations in which one or more limitations, elements, clauses, descriptive terms, *etc.*, from one or more of the claims or from relevant portions of the description is introduced into another claim. For example, any claim that is dependent on another claim can be modified to include one or more limitations found in any other claim that is dependent on the same base claim. Furthermore, where the claims recite a composition, it is to be understood that methods of using the composition for any of the purposes disclosed herein are included, and methods of making the composition according to any of the methods of making disclosed herein or other methods known in the art are included, unless otherwise indicated or unless it would be evident to one of ordinary skill in the art that a contradiction or inconsistency would arise.

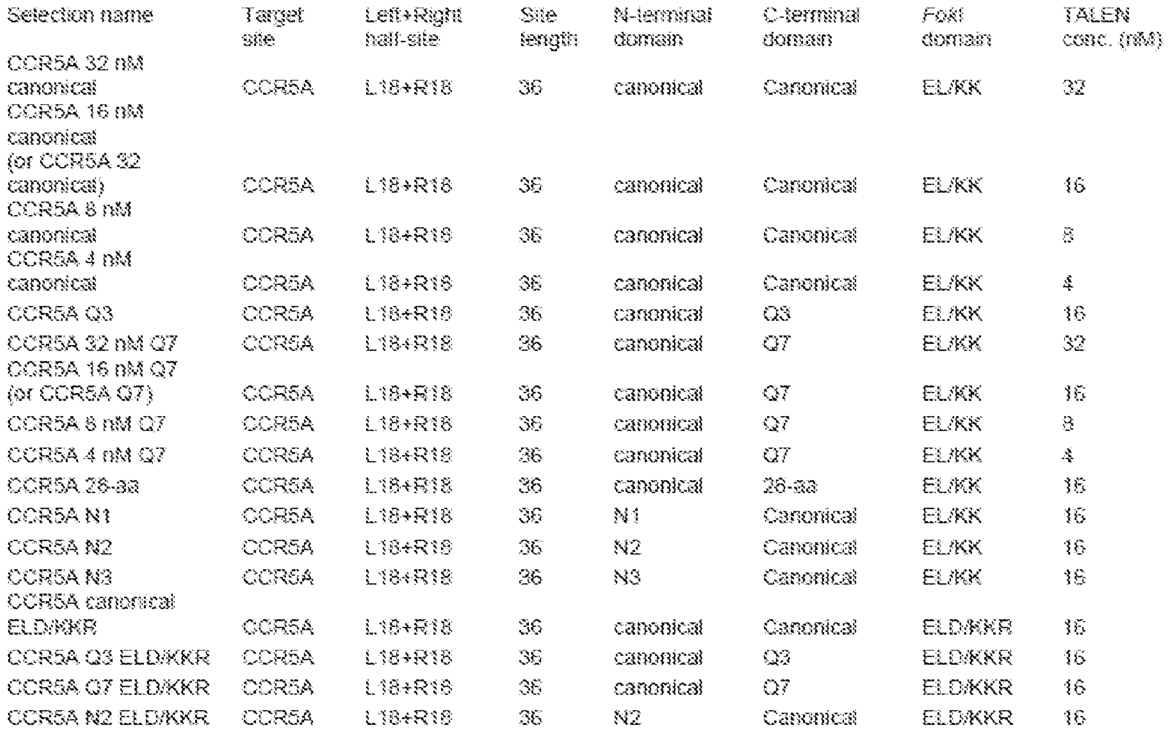
[00177] Where elements are presented as lists, *e.g.*, in Markush group format, it is to be understood that each subgroup of the elements is also disclosed, and any element(s) can be removed from the group. It is also noted that the term "comprising" is intended to be open and permits the inclusion of additional elements or steps. It should be understood that, in general, where the invention, or aspects of the invention, is/are referred to as comprising particular elements, features, steps, *etc.*, certain embodiments of the invention or aspects of the invention consist, or consist essentially of, such elements, features, steps, *etc.* For purposes of simplicity those embodiments have not been specifically set forth *in haec verba* herein. Thus for each embodiment of the invention that comprises one or more elements, features, steps, *etc*., the invention also provides embodiments that consist or consist essentially of those elements, features, steps, *etc*.

[00178] Where ranges are given, endpoints are included. Furthermore, it is to be understood that unless otherwise indicated or otherwise evident from the context and/or the understanding of one of ordinary skill in the art, values that are expressed as ranges can assume any specific value within the stated ranges in different embodiments of the invention, to the tenth of the unit of the lower limit of the range, unless the context clearly dictates otherwise. It is also to be understood that unless otherwise indicated or otherwise evident from the context and/or the understanding of one of ordinary skill in the art, values expressed as ranges can assume any subrange within the given range, wherein the endpoints of the subrange are expressed to the same degree of accuracy as the tenth of the unit of the lower limit of the range.

[00179] In addition, it is to be understood that any particular embodiment of the present invention may be explicitly excluded from any one or more of the claims. Where ranges are given, any value within the range may explicitly be excluded from any one or more of the claims. Any embodiment, element, feature, application, or aspect of the compositions and/or methods of the invention, can be excluded from any one or more claims. For purposes of brevity, all of the embodiments in which one or more elements, features, purposes, or aspects is excluded are not set forth explicitly herein.

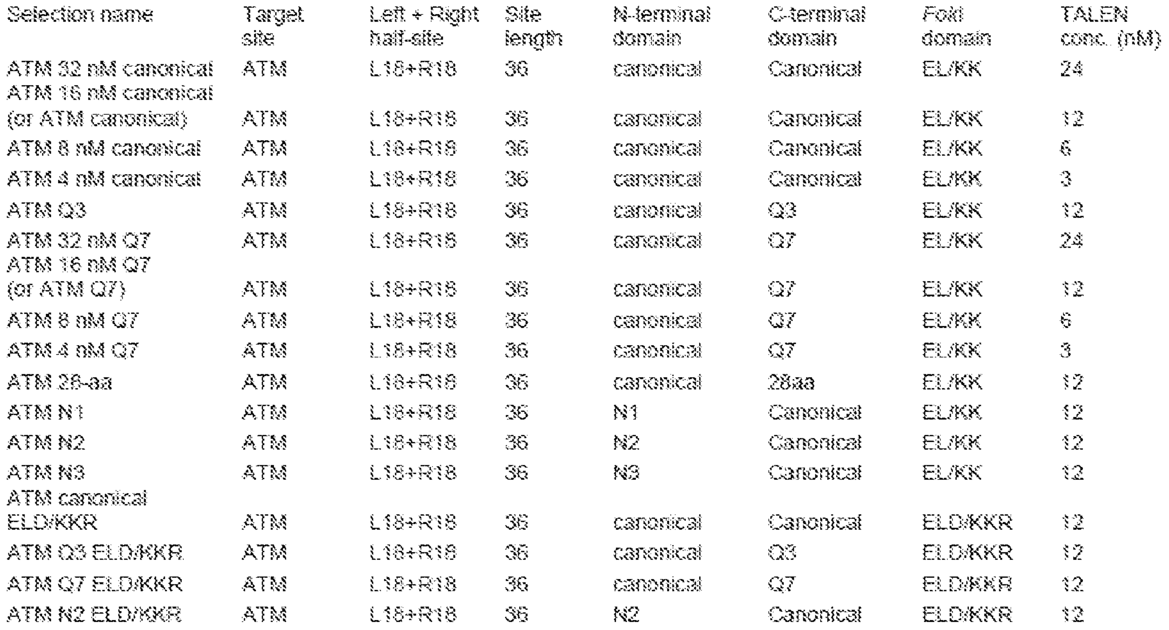
TABLES

A



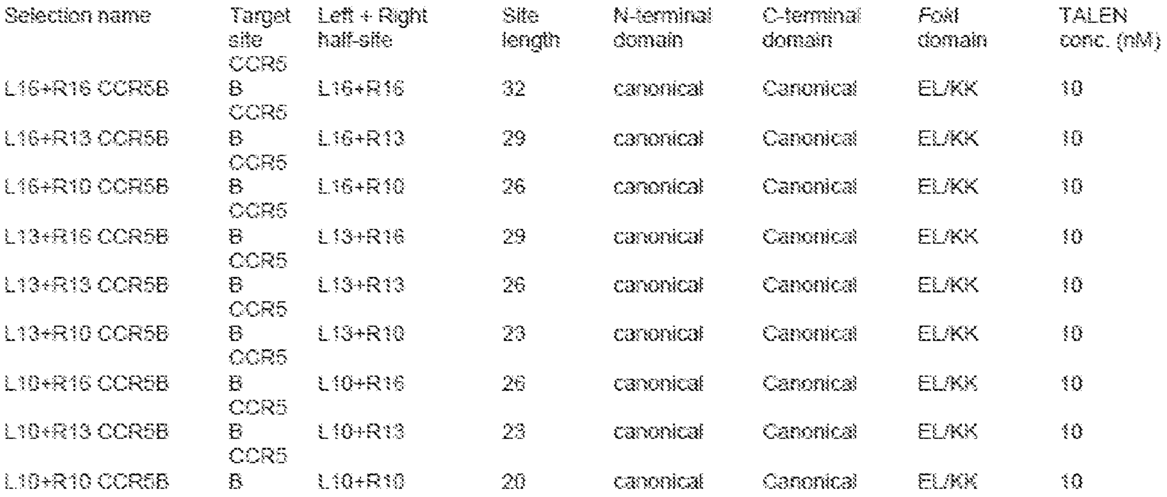
Selection name	Target site	Left+Right half-site	Site length	N-terminal domain	C-terminal domain	*FokI* domain	TALEN conc. (nM)
CCR5A 32 nM canonical	CCR5A	L18+R18	36	canonical	Canonical	EL/KK	32
CCR5A 16 nM canonical (or CCR5A 32 canonical)	CCR5A	L18+R18	36	canonical	Canonical	EL/KK	16
CCR5A 8 nM canonical	CCR5A	L18+R18	36	canonical	Canonical	EL/KK	8
CCR5A 4 nM canonical	CCR5A	L18+R18	36	canonical	Canonical	EL/KK	4
CCR5A Q3	CCR5A	L18+R18	36	canonical	Q3	EL/KK	16
CCR5A 32 nM Q7	CCR5A	L18+R18	36	canonical	Q7	EL/KK	32
CCR5A 16 nM Q7 (or CCR5A Q7)	CCR5A	L18+R18	36	canonical	Q7	EL/KK	16
CCR5A 8 nM Q7	CCR5A	L18+R18	36	canonical	Q7	EL/KK	8
CCR5A 4 nM Q7	CCR5A	L18+R18	36	canonical	Q7	EL/KK	4
CCR5A 28-aa	CCR5A	L18+R18	36	canonical	28-aa	EL/KK	16
CCR5A N1	CCR5A	L18+R18	36	N1	Canonical	EL/KK	16
CCR5A N2	CCR5A	L18+R18	36	N2	Canonical	EL/KK	16
CCR5A N3	CCR5A	L18+R18	36	N3	Canonical	EL/KK	16
CCR5A canonical ELD/KKR	CCR5A	L18+R18	36	canonical	Canonical	ELD/KKR	16
CCR5A Q3 ELD/KKR	CCR5A	L18+R18	36	canonical	Q3	ELD/KKR	16
CCR5A Q7 ELD/KKR	CCR5A	L18+R18	36	canonical	Q7	ELD/KKR	16
CCR5A N2 ELD/KKR	CCR5A	L18+R18	36	N2	Canonical	ELD/KKR	16

B



Selection name	Target site	Left + Right half-site	Site length	N-terminal domain	C-terminal domain	*FokI* domain	TALEN conc. (nM)
ATM 32 nM canonical	ATM	L18+R18	36	canonical	Canonical	EL/KK	24
ATM 16 nM canonical (or ATM canonical)	ATM	L18+R18	36	canonical	Canonical	EL/KK	12
ATM 8 nM canonical	ATM	L18+R18	36	canonical	Canonical	EL/KK	6
ATM 4 nM canonical	ATM	L18+R18	36	canonical	Canonical	EL/KK	3
ATM Q3	ATM	L18+R18	36	canonical	Q3	EL/KK	12
ATM 32 nM Q7	ATM	L18+R18	36	canonical	Q7	EL/KK	24
ATM 16 nM Q7 (or ATM Q7)	ATM	L18+R18	36	canonical	Q7	EL/KK	12
ATM 8 nM Q7	ATM	L18+R18	36	canonical	Q7	EL/KK	6
ATM 4 nM Q7	ATM	L18+R18	36	canonical	Q7	EL/KK	3
ATM 28-aa	ATM	L18+R18	36	canonical	28aa	EL/KK	12
ATM N1	ATM	L18+R18	36	N1	Canonical	EL/KK	12
ATM N2	ATM	L18+R18	36	N2	Canonical	EL/KK	12
ATM N3	ATM	L18+R18	36	N3	Canonical	EL/KK	12
ATM canonical ELD/KKR	ATM	L18+R18	36	canonical	Canonical	ELD/KKR	12
ATM Q3 ELD/KKR	ATM	L18+R18	36	canonical	Q3	ELD/KKR	12
ATM Q7 ELD/KKR	ATM	L18+R18	36	canonical	Q7	ELD/KKR	12
ATM N2 ELD/KKR	ATM	L18+R18	36	N2	Canonical	ELD/KKR	12

C



Selection name	Target site	Left + Right half-site	Site length	N-terminal domain	C-terminal domain	FokI domain	TALEN conc. (nM)
L16+R16 CCR5B	CCR5 B	L16+R16	32	canonical	Canonical	EL/KK	10
L16+R13 CCR5B	CCR5 B	L16+R13	29	canonical	Canonical	EL/KK	10
L16+R10 CCR5B	CCR5 B	L16+R10	26	canonical	Canonical	EL/KK	10
L13+R16 CCR5B	CCR5 B	L13+R16	29	canonical	Canonical	EL/KK	10
L13+R13 CCR5B	CCR5 B	L13+R13	26	canonical	Canonical	EL/KK	10
L13+R10 CCR5B	CCR5 B	L13+R10	23	canonical	Canonical	EL/KK	10
L10+R16 CCR5B	CCR5 B	L10+R16	26	canonical	Canonical	EL/KK	10
L10+R13 CCR5B	CCR5 B	L10+R13	23	canonical	Canonical	EL/KK	10
L10+R10 CCR5B	CCR5 B	L10+R10	20	canonical	Canonical	EL/KK	10

Table 2. TALEN constructs and concentrations used in the selections. For each selection using TALENs targeting the CCR5A target sequence (A), ATM target sequence (B) and CCR5B target sequence (C), the selection name, the target DNA site, the TALEN N-terminal domain, the TALEN C-terminal domain, the TALEN FokI domain, and the TALEN concentration (conc.) are shown.

A

Selection name	Seq. count	Mean mut.	Stdev mut.	Mut./bp	P-value vs. library	P-value vs. other TALENs
CCR5A 32 nM canonical	53883	4.327	1.483	0.120	3.3E-10	vs. CCR5A canonical ELD/KKR = 0.260
CCR5A 16 nM canonical	28940	4.061	1.436	0.113	5.4E-10	vs. CCR5A Q3 ELD/KKR = 0.028
CCR5A 8 nM canonical	29568	3.751	1.394	0.104	3.3E-10	
CCR5A 4 nM canonical	34355	3.347	1.355	0.093	1.5E-10	
CCR5A Q3	51694	3.841	1.380	0.107	1.7E-10	
CCR5A 32 nM Q7	48473	2.718	1.197	0.076	4.4E-11	
CCR5A 16 nM Q7	56593	2.559	1.154	0.071	3.1E-11	
CCR5A 8 nM Q7	43895	2.303	1.157	0.064	3.0E-11	
CCR5A 4 nM Q7	43737	2.018	1.234	0.056	2.1E-11	
CCR5A 28-aa	47395	2.614	1.203	0.073	4.0E-11	
CCR5A N1	64257	3.721	1.379	0.103	1.1E-10	vs. CCR5A 8 nM canonical =0.039
CCR5A N2	45467	3.148	1.306	0.087	8.2E-11	
CCR5A N3	24064	2.474	1.493	0.069	8.1E-11	
CCR5A canonical ELD/KKR	46998	4.336	1.491	0.120	4.0E-10	
CCR5A Q3 ELD/KKR	56978	4.098	1.415	0.114	2.2E-10	
CCR5A Q7 ELD/KKR	54903	3.234	1.330	0.090	7.3E-11	
CCR5A N2 ELD/KKR	79632	3.286	1.341	0.091	5.2E-11	

B

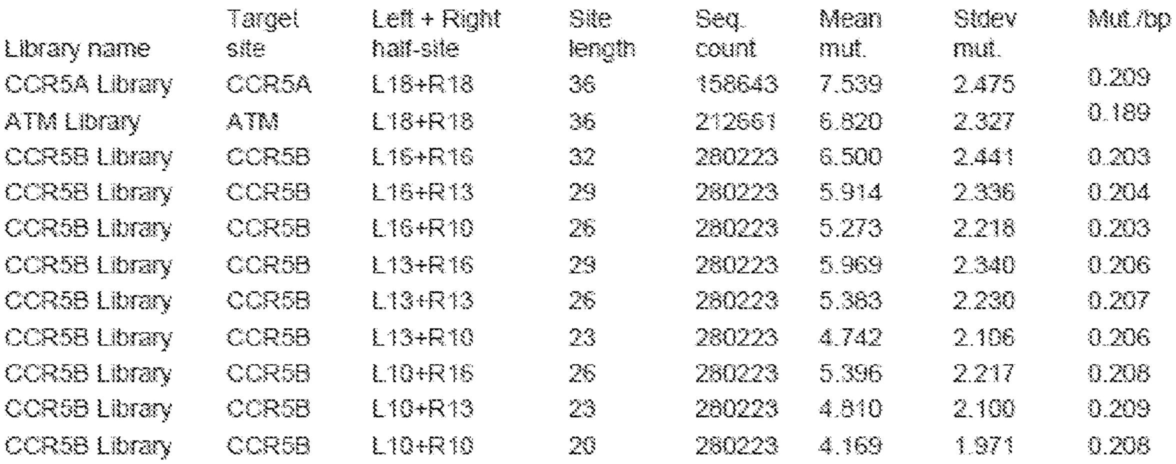
Selection name	Seq. count	Mean mut.	Stdev mut.	Mut./bp	P-value vs. library	P-value vs. other TALENs
ATM 24 nM canonical	89571	3.262	1.360	0.091	6.54E-11	vs. ATM canonical ELD/KKR =0.012
ATM 12 nM canonical (or ATM canonical)	96703	3.181	1.307	0.088	5.36E-11	
ATM 6 nM canonical	78652	2.736	1.259	0.076	3.63E-11	
ATM 3 nM canonical	82527	2.552	1.258	0.071	2.71E-11	
ATM Q3	96582	2.551	1.248	0.071	2.31E-11	vs. ATM 4 nM canonical =0.222
ATM 24 nM Q7	10166	1.885	2.125	0.052	2.06E-10	
ATM 12 nM Q7 (or ATM Q7)	4662	1.626	2.083	0.045	5.31E-10	
ATM 6 nM Q7	1298	1.700	2.376	0.047	7.16E-09	vs. ATM 16 nM Q7 =0.035
ATM N1	84402	2.627	1.318	0.073	2.92E-11	
ATM N2	62470	2.317	1.516	0.064	2.69E-11	
ATM N3	1605	2.720	2.363	0.076	2.69E-08	
ATM canonical ELD/KKR	107970	3.279	1.329	0.091	5.48E-11	
ATM Q3 ELD/KKR	104099	2.846	1.244	0.079	3.15E-11	

ATM Q7 ELD/KKR	21108	1.444	1.56	0.040	3.02E-11
ATM N2 ELD/KKR	70185	2.45	1.444	0.06805	2.82E-11

C

Selection name	Seq. count	Mean mut.	Stdev mut.	Mut./bp	P-value vs. library	P-value vs. other TALENs
L16+R16 CCR5B	34904	2.134	1.168	0.067	4.7E-11	
L16+R13 CCR5B	38229	1.581	1.142	0.055	2.7E-11	
L16+R10 CCR5B	37801	1.187	0.949	0.046	2.2E-11	
L13+R16 CCR5B	46608	1.505	1.090	0.052	1.7E-11	
L13+R13 CCR5B	53973	0.996	1.025	0.038	8.8E-12	
L13+R10 CCR5B	60550	0.737	0.684	0.032	7.4E-12	
L10+R16 CCR5B	36927	1.367	0.971	0.053	3.0E-11	
L10+R13 CCR5B	58170	0.839	0.882	0.036	9.1E-12	
L10+R10 CCR5B	57331	0.646	0.779	0.032	1.0E-11	

Table 3. Statistics of sequences selected by TALEN digestion. Statistics are shown for each TALEN selection on the CCR5A target sequence (A), ATM target sequence (B), and CCR5B target sequences (C). Seq. counts: total counts of high-throughput sequenced and computationally filtered selection sequences. Mean mut.: mean mutations in selected sequences. Stdev. mut.: standard deviation of mutations in selected sequences. Mut./bp: mean mutation normalized to target site length (bp). P-value vs. library: P-values between the TALEN selection sequence distributions to the corresponding pre-selection library sequence distributions (Table 5) were determined as previously reported.5 P-value vs. other TALENs: all pair-wise comparisons between all TALEN digestions were calculated and P-values between 0.01 and 0.5 are shown. Note that for the 3 nM Q7 ATM and the 28-aa ATM selection not enough sequences were obtained to interpret, although these selections were performed.



Library name	Target site	Left + Right half-site	Site length	Seq. count	Mean mut.	Stdev mut.	Mut./bp
CCR5A Library	CCR5A	L18+R18	36	158643	7.539	2.475	0.209
ATM Library	ATM	L18+R18	36	212661	6.820	2.327	0.189
CCR5B Library	CCR5B	L16+R16	32	280223	6.500	2.441	0.203
CCR5B Library	CCR5B	L16+R13	29	280223	5.914	2.336	0.204
CCR5B Library	CCR5B	L16+R10	26	280223	5.273	2.218	0.203
CCR5B Library	CCR5B	L13+R16	29	280223	5.969	2.340	0.206
CCR5B Library	CCR5B	L13+R13	26	280223	5.383	2.230	0.207
CCR5B Library	CCR5B	L13+R10	23	280223	4.742	2.106	0.206
CCR5B Library	CCR5B	L10+R16	26	280223	5.396	2.217	0.208
CCR5B Library	CCR5B	L10+R13	23	280223	4.810	2.100	0.209
CCR5B Library	CCR5B	L10+R10	20	280223	4.169	1.971	0.208

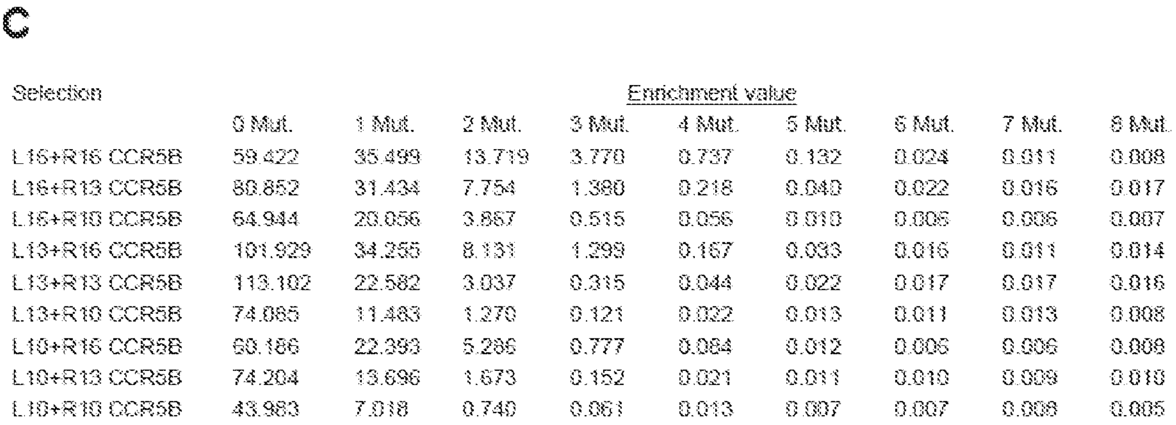
Table 4. Statistics of sequences from pre-selection libraries. For each preselection library containing a distribution of mutant sequences of the CCR5A target sequence, ATM target sequence and CCR5B target sequences. Seq. counts: total counts of high-throughput sequenced and the computationally filtered selection sequences. Mean mut.: mean mutations of sequences. Stdev. mut.: standard deviation of sequences. Mut./bp: mean mutation normalized to target site length (bp).

A

Selection					Enrichment value				
	0 Mut.	1 Mut.	2 Mut.	3 Mut.	4 Mut.	5 Mut.	6 Mut.	7 Mut.	8 Mut.
CCR5A 32 nM canonical	9.879	9.191	8.335	6.149	4.205	2.269	1.005	0.325	0.065
CCR5A 16 nM canonical	12.182	13.200	10.322	7.155	4.442	2.127	0.748	0.216	0.052
CCR5A 8 nM canonical	19.673	17.935	13.731	8.505	4.512	1.756	0.531	0.116	0.028
CCR5A 4 nM canonical	36.737	29.407	19.224	9.958	4.047	1.242	0.302	0.058	0.014
CCR5A Q3	18.550	16.456	12.024	8.070	4.532	1.938	0.572	0.126	0.025
CCR5A 32 nM Q7	60.583	54.117	31.082	11.031	2.640	0.469	0.073	0.013	0.006
CCR5A 16 nM Q7	62.294	64.689	35.035	10.538	2.183	0.322	0.046	0.010	0.005
CCR5A 8 nM Q7	97.020	91.633	38.634	8.974	1.485	0.189	0.029	0.010	0.007
CCR5A 4 nM Q7	197.239	130.497	38.361	6.535	0.896	0.120	0.025	0.019	0.017
CCR5A 28-aa	70.441	62.213	33.481	10.486	2.317	0.402	0.064	0.012	0.006
CCR5A N1	19.038	18.052	13.858	8.788	4.546	1.697	0.499	0.115	0.025
CCR5A N2	41.715	35.752	22.638	10.424	3.777	0.989	0.194	0.038	0.007
CCR5A N3	173.897	86.392	31.503	6.770	1.853	0.350	0.069	0.036	0.027
CCR5A canonical ELD/KKR	8.101	10.012	8.220	6.147	4.119	2.291	1.019	0.330	0.063
CCR5A Q3 ELD/KKR	14.664	12.975	9.409	6.819	4.544	2.235	0.797	0.198	0.041
CCR5A Q7 ELD/KKR	37.435	32.922	21.033	10.397	3.667	1.087	0.238	0.046	0.010
CCR5A N2 ELD/KKR	35.860	31.469	20.135	10.189	3.983	1.155	0.260	0.050	0.013

B

Selection					Enrichment value				
	0 Mut.	1 Mut.	2 Mut.	3 Mut.	4 Mut.	5 Mut.	6 Mut.	7 Mut.	8 Mut.
ATM 24 nM canonical	19.900	16.881	12.162	6.318	2.529	0.884	0.226	0.057	0.015
ATM 12 nM canonical	20.472	17.645	12.724	6.549	2.606	0.803	0.169	0.039	0.007
ATM 6 nM canonical	41.141	29.522	17.153	6.551	1.872	0.431	0.062	0.017	0.006
ATM 3 nM canonical	56.152	37.152	18.530	6.196	1.562	0.308	0.058	0.015	0.008
ATM Q3	50.403	36.687	19.031	6.245	1.513	0.294	0.057	0.016	0.010
ATM 24 nM Q7	353.148	90.350	13.475	1.531	0.186	0.128	0.116	0.118	0.103
ATM 12 nM Q7	513.385	89.952	11.310	0.860	0.190	0.093	0.115	0.092	0.111
ATM 6 nM Q7	644.427	82.074	7.650	0.677	0.170	0.205	0.163	0.164	0.071
ATM N1	57.218	35.388	17.808	6.124	1.644	0.383	0.076	0.023	0.011
ATM N2	119.240	53.618	18.977	4.742	0.992	0.233	0.076	0.044	0.037
ATM N3	201.158	55.468	15.244	3.187	0.764	0.307	0.154	0.173	0.267
ATM canonical ELD/KKR	19.356	15.692	11.855	6.403	2.706	0.899	0.224	0.054	0.011
ATM Q3 ELD/KKR	32.816	25.151	16.172	6.727	2.095	0.506	0.095	0.018	0.004
ATM Q7 ELD/KKR	447.509	93.166	13.505	1.543	0.170	0.053	0.049	0.045	0.045
ATM N2 ELD/KKR	90.625	45.525	18.683	5.369	1.267	0.274	0.076	0.035	0.027



C

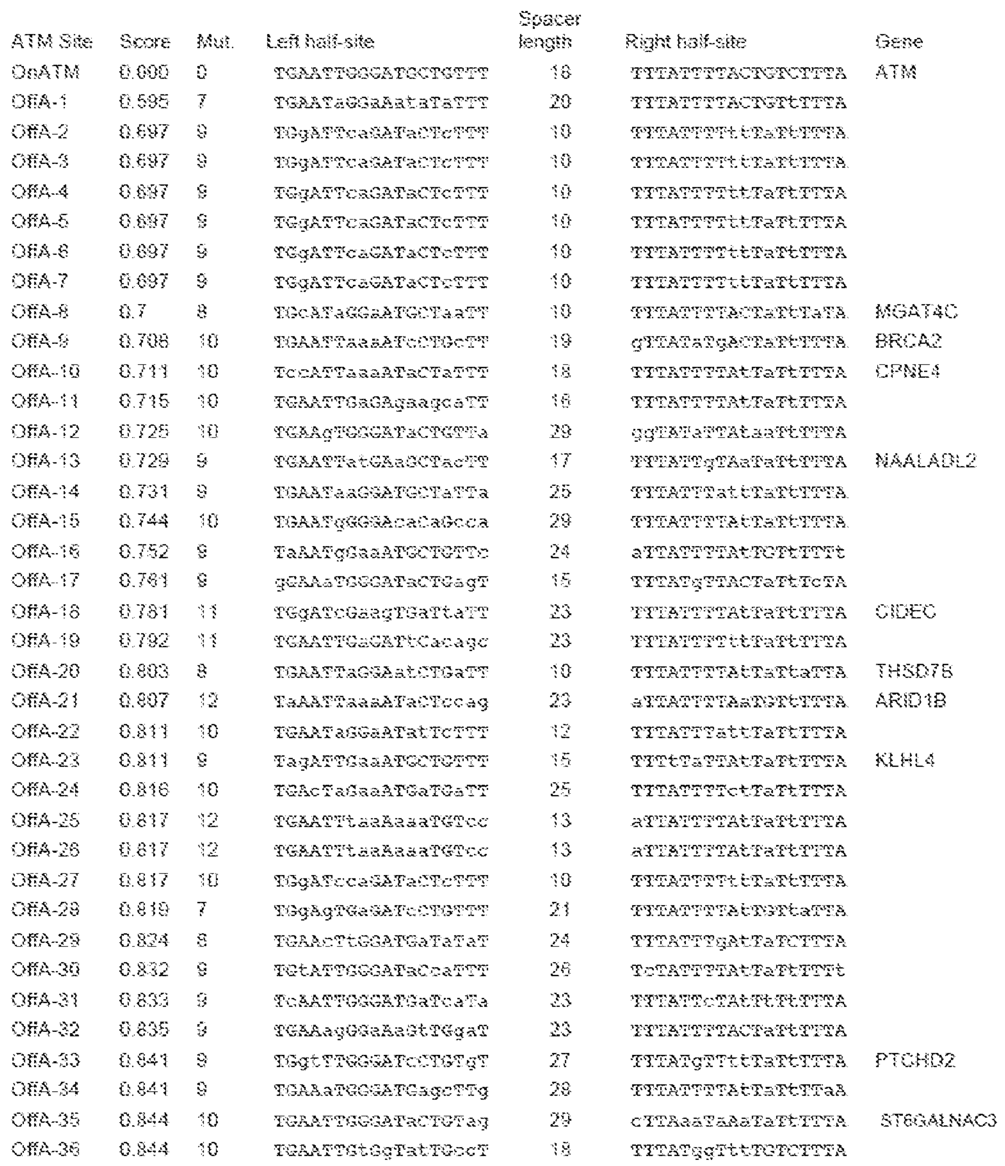
Selection	Enrichment value								
	0 Mut.	1 Mut.	2 Mut.	3 Mut.	4 Mut.	5 Mut.	6 Mut.	7 Mut.	8 Mut.
L16+R16 CCR5B	59.422	35.499	13.719	3.770	0.737	0.132	0.024	0.011	0.008
L16+R13 CCR5B	80.852	31.434	7.754	1.380	0.218	0.040	0.022	0.016	0.017
L16+R10 CCR5B	64.944	20.056	3.867	0.515	0.056	0.010	0.005	0.006	0.007
L13+R16 CCR5B	101.929	34.255	8.131	1.299	0.167	0.033	0.016	0.011	0.014
L13+R13 CCR5B	113.102	22.582	3.037	0.315	0.044	0.022	0.017	0.017	0.016
L13+R10 CCR5B	74.085	11.483	1.270	0.121	0.022	0.013	0.011	0.013	0.008
L10+R16 CCR5B	60.186	22.393	5.286	0.777	0.084	0.012	0.006	0.006	0.009
L10+R13 CCR5B	74.204	13.696	1.673	0.152	0.021	0.011	0.010	0.009	0.010
L10+R10 CCR5B	43.983	7.018	0.740	0.061	0.013	0.007	0.007	0.008	0.005

Table 5. Enrichment values of sequences as a function of number of mutations. For each TALEN selection on the CCR5A target sequence (A), ATM target sequence (B) and CCR5B target sequence (C), enrichment values calculated by dividing the fractional abundance of post-selection sequences from a TALEN digestion by the fractional abundance of pre-selection sequences as a function of total mutations (Mut.) in the half-sites.

A

CCR5A Site	Score	Mut.	Left half-site	Spacer length	Right half-site	Gene
OnCCR5A	0.008	0	TTCATTACACCTGCAGCT	18	AGTATCAATTCTGGAAGA	*CCR5*
OffC-1	0.747	9	TaCATcACAtaTGCAaaT	29	tGTATCAtTTCTGGgAGA	*ARL17A & LRRC37A*
OffC-2	0.747	9	TaCATcACAtaTGCAaaT	29	tGTATCAtTTCTGGgAGA	*ARL17A & LRRC37A*
OffC-3	0.747	9	TaCATcACAtaTGCAaaT	29	tGTATCAtTTCTGGgAGA	*ARL17A & LRRC37A*
OffC-4	0.747	11	TcCATaACACaTctttCT	10	tGcATCAtTcCTGGAAGA	*ZSCAN5A*
OffC-5	0.804	11	TcCAaTACctCTGCcaCa	14	AGgAgCAAcTCTGGgAGA	
OffC-6	0.818	10	TTCAgTcCAtCTGaAaac	16	gGTATCAtTTCTGGAgGA	*KL*
OffC-7	0.834	14	TaCAaaACcCtTGCcaaa	27	taTATCAATTtgGGgAGA	
OffC-8	0.837	12	TcCAagACACCTGCttac	26	tcTATCAATTtgGGgAGA	
OffC-9	0.874	10	TTCATaACAtCTtaAaaT	27	AaTAcCAAcTCTGGAtGA	*ZEB2*
OffC-10	0.89	12	TcCAaaACAtCTGaAaaT	25	tGgATCAAaTtgGGAAGA	
OffC-11	0.896	12	TTCAgaACACaTGactac	21	tGTATCAgTTaTGGAtGA	*GABPA*
OffC-12	0.904	13	TcCATaAtAtCTtCctCT	28	gGgATtAATTtgGGAgGA	
OffC-13	0.905	11	TgCAaTAtACCTGttGaT	16	ctcATCAATTCTGGgtGA	
OffC-14	0.906	12	TTCATaACACtccacctT	16	gGTATCAAaTCTGGggGA	*SYN3*
OffC-15	0.906	12	TcCATgACACaaaagaCT	26	gGTATCtATcCTGGAAtA	*SPOCK3*
OffC-16	0.906	9	TTCcTTcCACCaGtgtCc	28	AGcATCAATcCTGGAAGA	
OffC-17	0.907	10	TTaATaACAtCTcCAaCT	24	gGcAcCAAaTCTGGAtGA	*ATP13A5*
OffC-18	0.909	13	TcCATcACcCCTcCctCc	10	gGTgcCAgcTCTGGAgGA	*TBC1D7*
OffC-19	0.909	8	TTCATTACtCCTcCttCT	30	ctTATCAcTTtTGGAAGA	
OffC-20	0.912	10	TgCATTACACaTtatGtg	17	AGcAgCAcTTCTGGAAGA	
OffC-21	0.913	11	TTCAaaACACaTaCAtCT	28	AacAaCAtTcCTGtAAGA	*PRKAG2*
OffC-22	0.913	10	TcCATTACcaCTGCAGaT	25	gacATCAgTTaTGGAtGA	
OffC-23	0.925	13	TTCcagACcCCTtCctCa	13	gacATCAAaTCTGGgAGA	
OffC-24	0.927	12	TTCcaaACACCcGCttCc	26	taTATCctTTCTGGAAtA	
OffC-25	0.93	12	TgaAaTACACCTGCctaT	13	gGccTCAAggCTGGAtGA	*IL15*
OffC-26	0.93	12	TgCcaaACctCTGtcaCc	22	AGgATCAcTTCTGGAAGA	
OffC-27	0.931	12	TgCcaaACctCTGtcaCc	22	AGgATCAcTTCTGGAAGA	
OffC-28	0.931	8	TTtATTACACtTcCAGaT	19	gaTATCctTTCTGGAAGA	*ADIPOR2*
OffC-29	0.932	13	TaCAaaAaACtTtCtGag	27	tGTATCAATTtgGGgAGA	*FBXL17*
OffC-30	0.932	11	TcCAaaACACCcaCAGac	19	gGTATagATTgTGGAAGA	*ZNF365*
OffC-31	0.934	13	TTCATTcCACaTcCccac	25	gtTATCAAcatgGGAAGA	*MYO18B*
OffC-32	0.934	11	TTCAaTAtgCCaaCAGCT	11	AGctTCAATctgGGAgGA	
OffC-33	0.934	12	TTCAaTACACtTGtctaT	12	tGTgTCAtTTCTGGgttA	
OffC-34	0.935	11	TTCAacACACCTtCAaaa	12	tGTgTCAtTaaTGGAAGA	
OffC-35	0.935	10	TTCAaaACAtCTGacatT	10	AaTAgaAATTCTGGAAGA	
OffC-36	0.935	11	cTCcTaAtACCTGCAaaT	21	gaTATtAtTTCTGGAgGA	

B



ATM Site	Score	Mut.	Left half-site	Spacer length	Right half-site	Gene
OnATM	0.000	0	TGAATTGGGATGCTGTTT	18	TTTATTTTACTGTCTTTA	ATM
OffA-1	0.595	7	TGAATaGGaAataTaTTT	20	TTTATTTTACTGTtTTTA	
OffA-2	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA	
OffA-3	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA	
OffA-4	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA	
OffA-5	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA	
OffA-6	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA	
OffA-7	0.697	9	TGgATTcaGATaCTcTTT	10	TTTATTTTttTaTtTTTA	
OffA-8	0.7	8	TGcATaGGaATGCTaaTT	10	TTTATTTTACTaTtTaTA	MGAT4C
OffA-9	0.708	10	TGAATTaaaATcCTGcTT	19	gTTATaTgACTaTtTTTA	BRCA2
OffA-10	0.711	10	TccATTaaaATaCTaTTT	18	TTTATTTTAtTaTtTTTA	CPNE4
OffA-11	0.715	10	TGAATTGaGAgaagcaTT	16	TTTATTTTAtTaTtTTTA	
OffA-12	0.725	10	TGAAgTGGGATaCTGTTa	29	ggTATaTTAtaaTtTTTA	
OffA-13	0.729	9	TGAATTatGAaGCTacTT	17	TTTATTgTAaTaTtTTTA	NAALADL2
OffA-14	0.731	9	TGAATaaGGATGCTaTTa	25	TTTATTTattTaTtTTTA	
OffA-15	0.744	10	TGAATgGGGAcaCaGcca	29	TTTATTTTAtTaTtTTTA	
OffA-16	0.752	9	TaAATgGaaATGCTGTTc	24	aTTATTTTAtTGTtTTTt	
OffA-17	0.761	9	gGAAaTGGGATaCTGagT	15	TTTATgTTACTaTtTcTA	
OffA-18	0.781	11	TGgATcGaagTGaTtaTT	23	TTTATTTTAtTaTtTTTA	CIDEC
OffA-19	0.792	11	TGAATTGaGATtCacagc	23	TTTATTTTttTaTtTTTA	
OffA-20	0.803	8	TGAATTaGGAatCTGaTT	10	TTTATTTTAtTaTtaTTA	THSD7B
OffA-21	0.807	12	TaAATTaaaATaCTccag	23	aTTATTTTAaTGTtTTTA	ARID1B
OffA-22	0.811	10	TGAATaGGaATatTcTTT	12	TTTATTTattTaTtTTTA	
OffA-23	0.811	9	TagATTGaaATGCTGTTT	15	TTTtTaTTAtTaTtTTTA	KLHL4
OffA-24	0.816	10	TGAcTaGaaATGaTGaTT	25	TTTATTTTctTaTtTTTA	
OffA-25	0.817	12	TGAATTtaaAaaaTGTcc	13	aTTATTTTAtTaTtTTTA	
OffA-26	0.817	12	TGAATTtaaAaaaTGTcc	13	aTTATTTTAtTaTtTTTA	
OffA-27	0.817	10	TGgATccaGATaCTcTTT	10	TTTATTTTttTaTtTTTA	
OffA-28	0.819	7	TGgAgTGaGATcCTGTTT	21	TTTATTTTAtTGTtaTTA	
OffA-29	0.824	8	TGAAcTtGGATGaTaTaT	24	TTTATTTgAtTaTCTTTA	
OffA-30	0.832	9	TGtATTGGGATaCcaTTT	26	TcTATTTTAtTaTtTTTt	
OffA-31	0.833	9	TcAATTGGGATGaTcaTa	23	TTTATTcTAtTtTtTTTA	
OffA-32	0.835	9	TGAAagGGaAaGtTGgaT	23	TTTATTTTACTaTtTTTA	
OffA-33	0.841	9	TGgtTTGGGATcCTGTgT	27	TTTATgTTttTaTtTTTA	PTCHD2
OffA-34	0.841	9	TGAAaTGGGATGagcTTg	28	TTTATTTTAtTaTtTTaA	
OffA-35	0.844	10	TGAATTGGGATaCTGTag	29	cTTAaaTaAaTaTtTTTA	ST6GALNAC3
OffA-36	0.844	10	TGAATTGtGgTatTGccT	18	TTTATggTttTGTCTTTA	

Table 6. Predicted off-target sites in the human genome. (A) Using a machine learning "classifier" algorithm trained on the output of the in vitro CCR5A TALEN selection,6 mutant sequences of the target site allowing for spacer lengths of 10 to 30 base pairs were scored. The resulting 36 predicted off-targets sites with the best scores for the CCR5A TALENs are shown with classifier scores, mutation numbers, left and right half-site sequences (mutations from on-target in lower case), the length of the spacer between half-sites in base pairs, and the gene (including introns) in which the predicted off-target sites occurs, if it lies within a gene. (B) Same as (A) for ATM TALENs. Sequences relate to SEQ ID NOs: 43-XX.

A

C-terminal domain	No TALEN	Q7	Q7	Q3	Q3	Canonical	Canonical	Canonical
FokI domain	No TALEN	EL/KK	ELD/KKR	EL/KK	ELD/KKR	EL/KK	ELD/KKR	Homo
CCR5A Sites								
OnC								
Indels	5	147	705	1430	3731	841	2004	3943
Total	23644	7192	12667	16843	15381	8546	7267	8422
% Modified	0.021%	2.044%	5.566%	8.490%	24.257%	9.841%	27.577%	46.818%
P-value		1.3E-33	2.5E-160	<1.0E-200	<1.0E-200	5.9E-200	<1.0E-200	<1.0E-200
Specificity								
OffC-1								
Indels	0	0	1	1	0	0	1	1
Total	51248	38975	79858	35491	77804	34227	87497	42498
% Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-2								
Indels	0	0	0	0	0	0	10	0
Total	124356	96280	157387	93337	159817	85603	163332	114663
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	0.006%	<0.006%
P-value							1.6E-03	
Specificity		>307	>835	>1274	>3639	>1476	>4137	>7023
OffC-3								
Indels	5	0	4	1	0	0	6	3
Total	93085	75358	130027	72919	131132	67192	136796	90039
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-4								
Indels	0	1	0	0	0	0	0	0
Total	45377	44674	52876	35133	53909	26034	42284	40452
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-5								
Indels	0	0	0	3	22	134	385	395
Total	27009	28172	26036	22432	25800	25273	17045	17077
Modified	<0.006%	<0.006%	<0.006%	0.013%	0.085%	0.527%	2.209%	2.261%
P-value					2.7E-06	4.5E-31	4.9E-87	2.8E-89

Specificity		>576	>1450	635	285	19	12	21
OffC-6								
Indels	0	0	0	0	0	0	0	0
Total	10766	12309	10886	9240	10558	10500	5943	6560
Modified	<0.009%	<0.008%	<0.009%	<0.011%	<0.009%	<0.010%	<0.017%	<0.015%
P-value								
Specificity								
OffC-7								
Total	0	0	0	0	0	0	0	0
Modified	15626	28825	22138	31742	19577	11902	33200	15400
P-value	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
Specificity								
OffC-8								
Indels	0	0	0	1	0	0	0	0
Total	40603	39765	47974	51595	44002	34520	25211	30771
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-10								
Indels	0	0	0	0	0	0	0	0
Total	4142	9591	5187	1413	7975	4378	2215	3779
Modified	<0.024%	<0.010%	<0.019%	<0.071%	<0.013%	<0.023%	<0.045%	<0.026%
P-value								
Specificity								
OffC-11								
Indels	0	0	0	0	0	0	0	0
Total	71160	55455	65015	44847	70907	50967	65257	60191
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-12								
Indels	0	0	0	0	0	0	0	0
Total	3242	1784	30274	14006	4897	19830	9747	12910
Modified	<0.031%	<0.056%	<0.006%	<0.007%	<0.020%	<0.006%	<0.010%	<0.008%
P-value								
Specificity								
OffC-13								
Indels	0	0	0	0	0	0	0	0

Total	65518	52459	53413	38156	61600	47922	57211	78546
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

OffC-14

Indels	0	0	0	0	0	0	2	0
Total	34607	7217	26301	8339	29845	1081	9471	19026
Modified	<0.006%	<0.014%	<0.006%	<0.012%	<0.006%	<0.093%	0.021%	<0.006%
P-value								
Specificity								

OffC-15

Indels	0	0	0	0	0	0	16	2
Total	4989	4880	6026	9370	9156	7371	6967	4662
Modified	<0.020%	<0.020%	<0.017%	<0.011%	<0.011%	<0.014%	0.230%	0.043%
P-value							6.3E-05	
Specificity		>100	>335	>796	>2221	>725	120	1091

OffC-16

Indels	0	1	1	1	14	1	12	0
Total	36228	34728	34403	34866	44362	38384	38536	32636
Modified	<0.006%	<0.006%	<0.006%	<0.006%	0.032%	<0.006%	0.031%	<0.006%
P-value					1.8E-04		5.3E-04	
Specificity		>307	>835	>1274	769	>1476	886	>7023

OffC-17

Indels	0	0	0	0	0	0	0	0
Total	32112	23901	31273	33968	27437	29670	27133	31299
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

OffC-18

Indels	0	0	0	0	0	0	0	0
Total	9437	9661	13505	14900	13848	12720	6624	12804
Modified	<0.011%	<0.010%	<0.007%	<0.007%	<0.007%	<0.008%	<0.015%	<0.008%
P-value								
Specificity								

OffC-19

Indels	1	1	1	2	2	2	1	0
Total	22869	11479	22702	15258	20733	17449	14638	28478
Modified	<0.006%	0.009%	<0.006%	0.013%	0.010%	0.011%	0.007%	<0.006%
P-value								

Specificity								
OffC-20								
Indels	0	0	0	0	0	1	0	0
Total	23335	26164	30782	15261	20231	21184	14144	18972
Modified	<0.006%	<0.006%	<0.006%	<0.007%	<0.006%	<0.006%	<0.007%	<0.006%
P-value								
Specificity								
OffC-21								
Indels	0	0	0	0	0	0	0	0
Total	34302	27573	31694	24451	25826	27192	18110	21161
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-22								
Indels	1	0	0	0	0	0	0	0
Total	81037	86687	74274	79004	93477	92099	75359	104857
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-23								
Indels	0	0	0	0	0	0	0	0
Total	18812	19337	23034	25603	25023	28615	17172	21033
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-24								
Indels	0	0	1	0	0	0	0	1
Total	23538	21673	24594	27687	18343	29113	21709	26610
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-25								
Indels	0	0	0	0	0	0	0	0
Total	28941	25326	25871	10641	21422	20171	18946	18711
Modified	<0.006%	<0.006%	<0.006%	<0.009%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffC-26								

Indels	0	0	1	0	0	0	0	0
Total	71631	48494	62550	45801	60175	65137	28735	64632
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.005%
P-value								
Specificity								

OffC-27

Indels	0	0	0	0	0	0	0	0
Total	12181	2423	11258	7188	5126	4003	2116	4603
% Modified	<0.008%	<0.041%	<0.009%	<0.014%	<0.020%	<0.025%	<0.047%	<0.022%
P-value								
Specificity								

OffC-28

Indels	0	0	0	0	6	1	12	5
Total	10651	6410	16179	13980	13022	7232	7379	8998
% Modified	<0.009%	<0.016%	<0.006%	<0.007%	0.046%	0.014%	0.163%	0.056%
P-value					1.4E-02		6.3E-04	
Specificity		>131	>835	>1187	526	712	170	843

OffC-29

Indels	0	0	0	0	0	0	0	0
Total	4252	3766	4228	6960	3234	1516	2466	1810
% Modified	<0.023%	<0.027%	<0.024%	<0.014%	<0.031%	<0.066%	<0.041%	<0.055%
P-value								
Specificity								

OffC-30

Indels	0	0	0	0	0	0	0	0
Total	11640	12257	9617	34097	20507	5029	22248	6285
% Modified	<0.008%	<0.008%	<0.010%	<0.006%	<0.006%	<0.020%	<0.006%	<0.016%
P-value								
Specificity								

OffC-31

Indels	0	0	0	0	0	0	0	0
Total	64522	67791	50085	50056	56241	48287	72230	100410
% Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.005%
P-value								
Specificity								

OffC-32

Indels	0	0	0	0	0	0	0	0
Total	1944	6888	9330	3207	4591	6699	13807	19115
% Modified	<0.051%	<0.015%	<0.011%	<0.031%	<0.022%	<0.015%	<0.007%	<0.009%

P-value	
Specificity	

OffC-33

Indels	0	0	0	0	0	0	0	0
Total	34475	27039	18547	33457	15745	17075	4	18844
% Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<25.000%	<0.006%
P-value								
Specificity								

OffC-34

Indels	0	0	0	0	0	0	0	0
Total	9052	16858	13647	11796	6945	6114	4979	9072
% Modified	<0.011%	<0.006%	<0.007%	<0.008%	<0.014%	<0.016%	<0.020%	<0.011%
P-value								
Specificity								

OffC-35

Indels	0	0	0	0	0	0	0	0
Total	23839	22290	25133	24190	10	10459	22554	11897
% Modified	<0.006%	<0.006%	<0.006%	<0.006%	<10.000%	<0.010%	<0.006%	<0.008%
P-value								
Specificity								

OffC-36

Indels	1	0	0	1	2	1	19	5
Total	23412	24394	23427	24132	19723	28369	12461	18052
Modified	<0.006%	<0.006%	<0.006%	<0.006%	0.010%	<0.006%	0.152%	0.028%
P-value							2.6E-05	
Specificity		>307	>835	>1274	2392	>1476	181	1690

B

C-term. Domain:	No TALEN	Q7	Q7	Q3	Q3	Canonical	Canonical	Canonical
FokI Domain:	No TALEN	EL/KK	ELD/KKR	EL/KK	ELD/KKR	EL/KK	ELD/KKR	Homo
ATM Sites								
On-A								
Indels	3	0	46	104	309	1289	410	909
Total	6888	1669	2520	1198	1808	19025	2533	5003
Modified	0.03%	0.00%	1.83%	8.68%	17.09%	6.78%	16.19%	18.17%
P-value	0		2.2E-11	3.2E-26	4.9E-81	6.4E-276	4.5E-105	1.5E-228
Specificity								

OffA-1

Indels	0	0	1	0	1	0	13	34

Total	52490	45383	34195	32325	47589	39704	50349	44056
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	0.026%	0.077%
P-value							3.1E-04	5.5E-09
Specificity		>0	>274	>1302	>2564	>1016	627	235

OffA-2

Indels	0	0	0	0	0	0	0	0
Total	8777	11846	11362	12273	20704	3776	5650	5025
Modified	<0.011%	<0.008%	<0.009%	<0.008%	<0.006%	<0.026%	<0.018%	<0.020%
P-value								
Specificity								

OffA-3

Indels	0	0	0	0	1	0	0	0
Total	47338	14352	21253	17777	26512	19483	43728	29469
Modified	<0.006%	<0.007%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

OffA-4

Indels	0	0	0	0	0	0	0	0
Total	12292	532	1383	2597	861	2598	1356	3573
Modified	<0.008%	<0.188%	<0.072%	<0.039%	<0.116%	<0.038%	<0.074%	<0.028%
P-value								
Specificity								

OffA-5

Indels	0	0	0	0	0	0	0	0
Total	60859	22846	25573	19054	25315	31754	66622	60925
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

OffA-6

Indels	0	0	0	0	0	0	0	0
Total	60859	22846	25573	19054	25315	31754	66622	60925
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

OffA-7

Indels	0	0	0	0	0	0	0	0
Total	60859	22846	25573	19054	25315	31754	66622	60925
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								

Specificity								
OffA-8								
Indels	0	0	0	0	0	0	0	0
Total	9170	1614	5934	3215	2450	12750	10120	13003
Modified	<0.011%	<0.062%	<0.017%	<0.031%	<0.041%	<0.008%	<0.010%	<0.008%
P-value								
Specificity								
OffA-9								
Indels	0	0	0	0	0	0	0	3
Total	8753	12765	9504	10114	11086	10676	9013	11110
Modified	<0.011%	<0.008%	<0.011%	<0.010%	<0.009%	<0.009%	<0.011%	0.027%
P-value								
Specificity								
OffA-10								
Indels	1	0	0	2	2	3	5	7
Total	8151	16888	8804	7061	8891	32138	14889	40120
Modified	0.012%	<0.006%	<0.011%	0.028%	0.022%	0.009%	0.034%	0.017%
P-value								
Specificity								
OffA-11								
Indels	0	0	1	0	0	0	9	76
Total	41343	32352	28834	28709	26188	32519	24894	19586
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	0.036%	0.388%
P-value							2.7E-03	2.5E-18
Specificity		>0	>274	>1302	>2564	>1016	446	47
OffA-12								
Indels	0	0	0	0	0	0	0	0
Total	13186	2328	13981	12911	21134	9220	7792	8068
Modified	<0.008%	<0.043%	<0.007%	<0.008%	<0.005%	<0.011%	<0.013%	<0.012%
P-value								
Specificity								
OffA-13								
Indels	0	0	0	0	0	2	9	0
Total	32704	32015	12312	23645	26315	24078	36111	22364
Modified	<0.006%	<0.006%	<0.008%	<0.006%	<0.006%	0.008%	0.025%	<0.006%
P-value							2.7E-03	
Specificity		>0	>225	>1302	>2564	816	649	>2725
OffA-15								

Indels	0	0	0	0	1	0	0	0
Total	14654	15934	12313	6581	13053	18996	10916	21519
Modified	<0.007%	<0.006%	<0.008%	<0.015%	0.008%	<0.006%	<0.009%	<0.005%
P-value								
Specificity								
OffA-16								
Indels	1	0	0	0	0	0	0	12
Total	65190	35639	37252	30378	31489	22590	13594	20922
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.007%	0.057%
P-value								7.9E-04
Specificity		>0	>274	>1302	>2564	>1016	>2200	317
OffA-17								
Indels	0	0	0	0	0	0	0	6
Total	1972	606	1439	2113	2862	729	597	636
Modified	<0.051%	<0.165%	<0.069%	<0.047%	<0.035%	<0.137%	<0.168%	0.943%
P-value								1.4E-02
Specificity		>0	>26	>183	>469	>49	>97	19
OffA-18								
Indels	0	0	0	0	0	0	0	0
Total	5425	995	1453	1831	3132	1934	1534	5816
Modified	<0.018%	<0.101%	<0.069%	<0.055%	<0.032%	<0.052%	<0.065%	<0.017%
P-value								
Specificity								
OffA-19								
Indels	1	2	0	1	1	1	1	3
Total	31094	41252	33213	29518	32337	25904	27575	38711
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	0.008%
P-value								
Specificity								
OffA-21								
Indels	0	0	0	0	0	0	0	0
Total	15297	9710	15719	12119	15483	21692	16558	16418
Modified	<0.007%	<0.010%	<0.006%	<0.008%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								
OffA-22								
Indels	27	41	38	46	32	50	55	57
Total	9405	11150	11516	10269	13814	14057	11685	14291
Modified	0.267%	0.368%	0.330%	0.448%	0.232%	0.356%	0.471%	0.399%

P-value								
Specificity								

OffA-23

Indels	1	0	0	0	0	0	10	20
Total	5671	9353	2203	7011	7078	12066	3484	8619
Modified	0.018%	<0.011%	<0.045%	<0.014%	<0.014%	<0.008%	0.287%	0.232%
P-value							3.5E-03	9.1E-05
Specificity		>0	>40	>609	>1210	>818	56	78

OffA-24

Indels	4	0	0	1	0	1	0	2
Total	17288	7909	14261	29936	6943	6333	14973	19953
Modified	0.023%	<0.013%	<0.007%	<0.006%	<0.014%	0.016%	<0.007%	0.010%
P-value								
Specificity								

OffA-25

Indels	0	0	0	0	0	0	0	0
Total	20089	45320	50758	108581	11574	20948	123827	74151
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.009%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

OffA-27

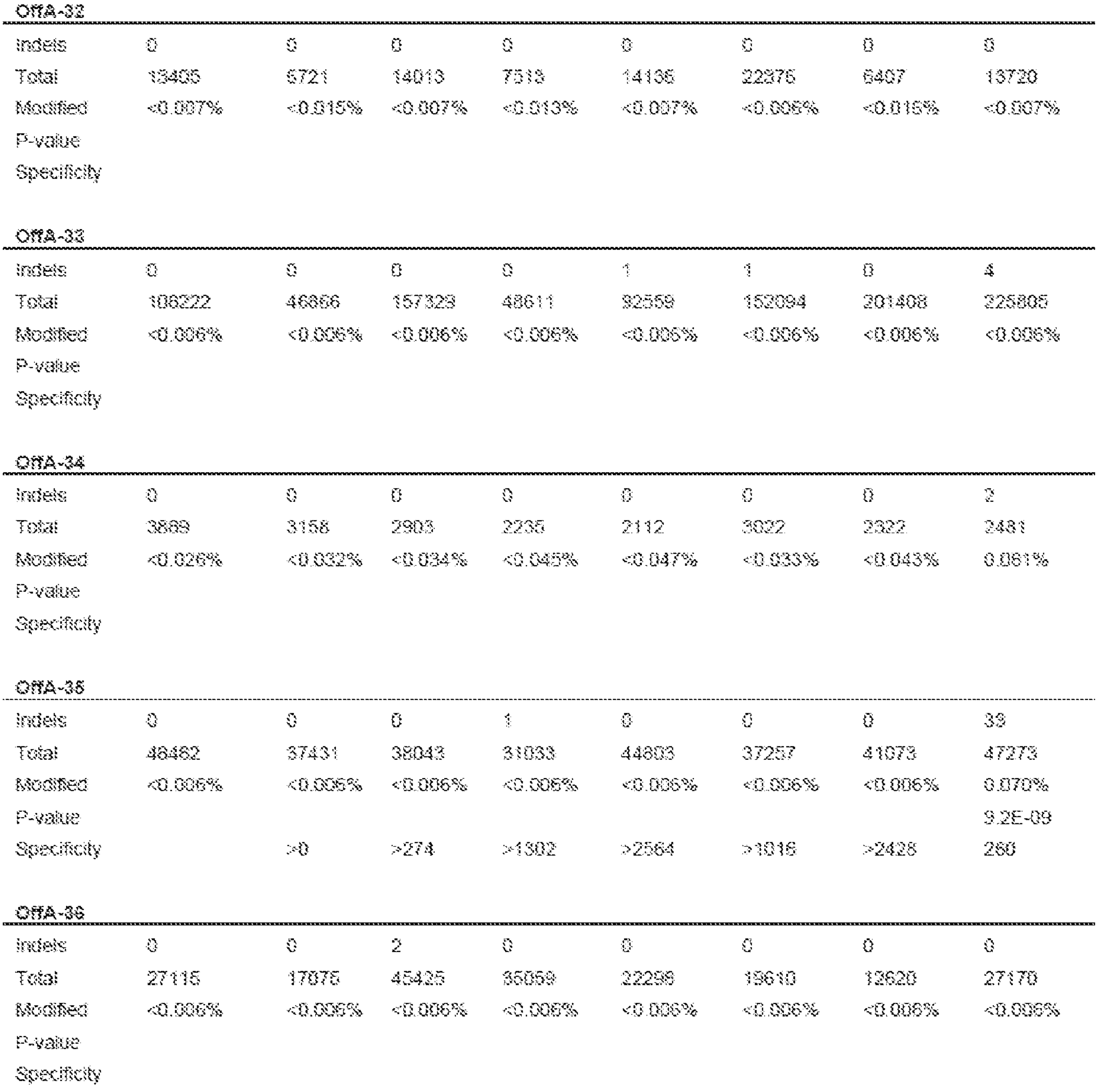
Indels	0	0	0	0	1	0	0	0
Total	47338	14352	21253	17777	26512	19483	43728	29469
Modified	<0.006%	<0.007%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

OffA-29

Indels	0	0	0	0	0	0	0	0
Total	5174	12618	36909	18063	16486	17934	9999	35072
Modified	<0.019%	<0.008%	<0.006%	<0.006%	<0.006%	<0.006%	<0.010%	<0.006%
P-value								
Specificity								

OffA-30

Indels	4	4	0	7	4	4	0	3
Total	45082	56531	36333	68651	69552	20362	29180	21350
Modified	0.009%	0.007%	<0.006%	0.008%	<0.006%	0.020%	<0.006%	0.014%
P-value								
Specificity								



OffA-32

Indels	0	0	0	0	0	0	0	0
Total	13405	6721	14013	7513	14136	22376	6407	13720
Modified	<0.007%	<0.015%	<0.007%	<0.013%	<0.007%	<0.006%	<0.016%	<0.007%
P-value								
Specificity								

OffA-33

Indels	0	0	0	0	1	1	0	4
Total	106222	46866	157323	48611	92559	152094	201408	225805
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%
P-value								
Specificity								

OffA-34

Indels	0	0	0	0	0	0	0	2
Total	3889	3158	2903	2235	2112	3022	2322	2481
Modified	<0.026%	<0.032%	<0.034%	<0.045%	<0.047%	<0.033%	<0.043%	0.081%
P-value								
Specificity								

OffA-35

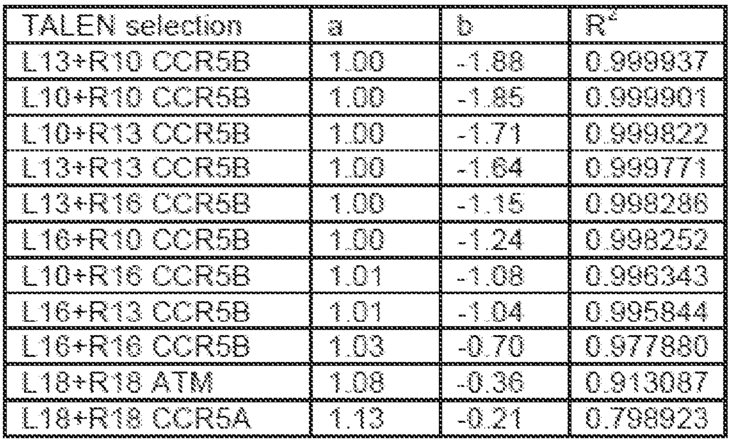
Indels	0	0	0	1	0	0	0	33
Total	46462	37431	38043	31033	44603	37257	41073	47273
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	0.070%
P-value								9.2E-09
Specificity		>0	>274	>1302	>2564	>1016	>2428	260

OffA-36

Indels	0	0	2	0	0	0	0	0
Total	27115	17075	45425	35059	22298	19610	12620	27170
Modified	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.006%	<0.008%	<0.006%
P-value								
Specificity								

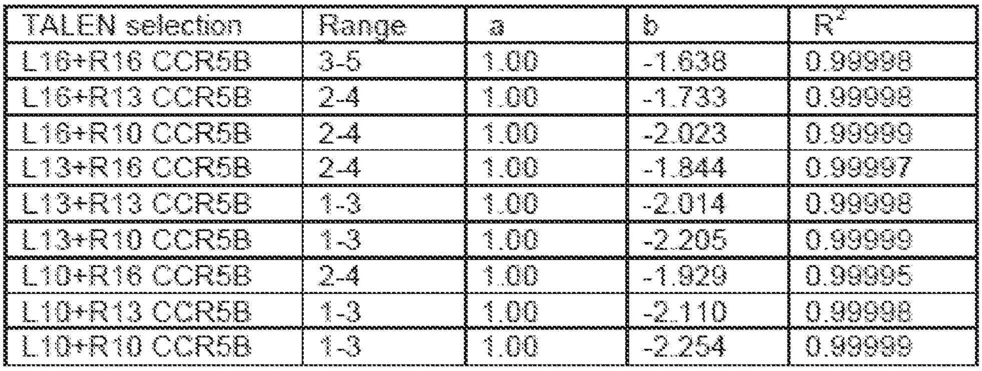
Table 7. Cellular modification induced by TALENs at on-target and predicted off-target genomic sites. (A) Results from sequencing CCR5A on-target and each predicted genomic off-target site that amplified from genomic DNA isolated from human cells treated with either no TALEN or TALENs containing canonical, Q3 or Q7 C-terminal domains, and either EL/KK heterodimeric, ELD/KKR heterodimeric, or homodimeric (Homo) FokI domains. Indels: the number of observed sequences containing insertions or deletions consistent with TALEN-induced cleavage. Total: total number of sequence counts. Modified: number of indels divided by total number of sequences as percentages. Upper limits of potential modification were calculated for sites with no observed indels by assuming there is less than one indel then dividing by the total sequence count to arrive at an upper limit modification

percentage, or taking the theoretical limit of detection (1/16,400), whichever value was more conservative (larger). P-values: calculated as previously reported5 between each TALEN-treated sample and the untreated control sample. P-values less than 0.05 are shown. Specificity: the ratio of ontarget to off-target genomic modification frequency for each site. (B) Same as (A) for the ATM target sites.



TALEN selection	a	b	R^2
L13+R10 CCR5B	1.00	-1.88	0.999937
L10+R10 CCR5B	1.00	-1.85	0.999901
L10+R13 CCR5B	1.00	-1.71	0.999822
L13+R13 CCR5B	1.00	-1.64	0.999771
L13+R16 CCR5B	1.00	-1.15	0.998286
L16+R10 CCR5B	1.00	-1.24	0.998252
L10+R16 CCR5B	1.01	-1.08	0.996343
L16+R13 CCR5B	1.01	-1.04	0.995844
L16+R16 CCR5B	1.03	-0.70	0.977880
L18+R18 ATM	1.08	-0.36	0.913087
L18+R18 CCR5A	1.13	-0.21	0.798923

Table 8. Exponential fitting of enrichment values as function of mutation number. Enrichment values of post-selection sequences as function of mutation were normalized relative to on-target enrichment (= 1.0 by definition). Normalized enrichment values of sequences with zero to four mutations were fit to an exponential function, a*eb, with R2 reported using the non-linear least squares method.



TALEN selection	Range	a	b	R^2
L16+R16 CCR5B	3-5	1.00	-1.638	0.99998
L16+R13 CCR5B	2-4	1.00	-1.733	0.99998
L16+R10 CCR5B	2-4	1.00	-2.023	0.99999
L13+R16 CCR5B	2-4	1.00	-1.844	0.99997
L13+R13 CCR5B	1-3	1.00	-2.014	0.99998
L13+R10 CCR5B	1-3	1.00	-2.205	0.99999
L10+R16 CCR5B	2-4	1.00	-1.929	0.99995
L10+R13 CCR5B	1-3	1.00	-2.110	0.99998
L10+R10 CCR5B	1-3	1.00	-2.254	0.99999

Table 9. Exponential fitting and extrapolation of enrichment values as function of mutation number. Enrichment values of all sequences from all nine of the CCR5B selections as function of mutation number were normalized relative to enrichment values of sequences with the lowest mutation number in the range shown (= 1.0 by definition). Normalized enrichment values of sequences from the range of mutations specified were fit to an exponential function, $a*e^b$, with R^2 reported utilizing the non-linear least squares method. These exponential decrease, b, were used to extrapolate all mean enrichment values beyond five mutations.

A

oligonucleotide name	oligonucleotide sequence (5'->3')
TAL-Nrev	5Phos/CAGCAGCTGCCCGGT
TAL-N1fwd	5Phos/cAGATCGCGAAGAGAGGGGGAGTAACAGCGGTAG
TAL-N2fwd	5Phos/cAGATCGCGcAGAGAGGGGGAGTAACAGCGGTAG
TAL-N3fwd	5Phos/cAGATCGCGcAGcagGGGGGAGTAACAGCGGTAG
TAL-Cifwd	ATC GTA GCC CAA TTG TCC A
TAL-Cirev	GTTGGTTCTTTGGATCAATGCG
TAL-Q3	AAGTTCTCTCGGGAATCCGTTGGTTGGTTCTTTGGATCA
TAL-Q7	GAAGTTCTCTCGGGAATTTGTTGGTTGGTTTGTTGGATCAATGCGGGAGCATGAGGCAGACCTT GTTGGACTGCATC
TAL-Ciirev	CTTTTGACTAGTTGGGATCCCGCGACTTGATGGGAAGTTCTCTCGGGAAT
CCR5A Library10	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNN A%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library12	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNN NNA%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library14	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNN NNNNA%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library16	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNN NNNNNNA%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library18	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNN NNNNNNNNA%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library20	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNN NNNNNNNNNNA%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library22	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNN NNNNNNNNNNNNA%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACGCT
CCR5A Library24	5Phos/CACCACTNT%T%C%A%T%T%A%C%A%C%C%T%G%C%A%G%C%T%NNNNNNNNNN NNNNNNNNNNNNNNA%G%T%A%T%C%A%A%T%T%C%T%G%G%A%A%G%A%NCGTCACG CT
CCR5B Library10	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNC%A% T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library12	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNC% A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library14	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNN C%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library16	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNN NNC%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library18	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNN NNNNC%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library20	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNN NNNNNNC%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library22	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNN NNNNNNNNC%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
CCR5B Library24	5Phos/CCACGCTNT%C%T%T%C%A%T%T%A%C%A%C%C%T%G%C%NNNNNNNNNNNNNN NNNNNNNNNNC%A%T%A%C%A%G%T%C%A%G%T%A%T%C%A%NCCTCGGGACT
ATM Library10	Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNNN T%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library12	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNN NNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library14	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNN NNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library16	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNN NNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library18	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNN NNNNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library20	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNN NNNNNNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library22	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNN NNNNNNNNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
ATM Library24	5Phos/CTCCGCGTNT%G%A%A%T%T%G%G%G%A%T%G%C%T%G%T%T%T%NNNNNNNNN

	NNNNNNNNNNNNNNNNT%T%T%A%T%T%T%T%A%C%T%G%T%C%T%T%T%A%GGTACCCCA
#1 adapter-fwd**1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGT
#1 adapter-rev**1	ACAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGAA
#1 adapter-rev**2	TTCAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCAA
#1 adapter-rev**3	TTGCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGACT
#1 adapter-rev**4	AGTCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCATT
#1 adapter-rev**5	AATGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**6	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCATGA
#1 adapter-rev**6	TCATGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**7	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGCT
#1 adapter-rev**7	AGCATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**8	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTAGT
#1 adapter-rev**8	ACTAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**9	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTAA
#1 adapter-rev**10	TTAGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**10	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGTA
#1 adapter-rev**11	TACTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**11	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTACT
#1 adapter-rev**12	AGTACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**12	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGT
#1 adapter-rev**13	ACAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**13	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTAA
#1 adapter-rev**14	TTAGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**14	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGTA
#1 adapter-rev**14	TACTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**15	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTACT
#1 adapter-rev**15	AGTACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#1 adapter-fwd**16	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGT
#1 adapter-rev**16	ACAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGG
#2A primer-fwd	AATGATACGGCGACCAC
#2A primer-rev*CCR5A	GTTCAGACGTGTGCTCTTCCGATCTNNNNAGTGGTGAGCGTGACG
#2A primer-rev*ATM	GTTCAGACGTGTGCTCTTCCGATCTNNNNACGCGGAGTGGGGTACC
#2A primer-rev*CCR5B	CAGACGTGTGCTCTTCCGATCNNNNAGCGTGGAGTCCCGAGG
#2B primer-fwd	AATGATACGGCGACCAC
#2B primer-rev**1	CAAGCAGAAGACGGCATACGAGATTGTTGACTGTGACTGGAGTTCAGACGTGTGCTCTTC
#2B primer-rev**2	CAAGCAGAAGACGGCATACGAGATACGGAACTGTGACTGGAGTTCAGACGTGTGCTCTTC
#2B primer-rev **3	CAAGCAGAAGACGGCATACGAGATTCTAACATGTGACTGGAGTTCAGACGTGTGCTCTTC
#2B primer-rev **4	CAAGCAGAAGACGGCATACGAGATCGGGACGGGTGACTGGAGTTCAGACGTGTGCTCTTC
	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCG
#1 Lib. adapter - fwd*CCR5A	GTACCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG
#1 Lib. adapter - rev*CCR5A	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTG
#1 Lib. adapter - fwd*ATM	GTACGATGCGATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG
#1 Lib. adapter - rev*ATM	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGCATC
#1 Lib. adapter – fwd*CCR5B	TCGGGAACGTGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGTCTAATCTCGTATGCCGTCTTCTGCTTG
#1 Lib. adapter - rev*CCR5B	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCACGTT
#2A Lib. primer-rev	CAAGCAGAAGACGGCATACGA
#2A Lib. primer-fwd*CCR5A	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCGTCACGCTCACCACT

#2A Lib. primer-fwd*ATM	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGGTACCCCACTCCGCGT
#2A Lib. primer-fwd*CCR5B	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCCTCGGGACTCCACGCT
#2B Lib. primer-rev	CAAGCAGAAGACGGCATACGA
#2B Lib. primer-fwd	AATGATACGGCGACCAC
G adapter-fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
G adapter-rev	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCA
G-B primer-fwd	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC
G-B primer-rev**1	CAAGCAGAAGACGGCATACGAGATGTGCGGACGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**2	CAAGCAGAAGACGGCATACGAGATCGTTTCACGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**3	CAAGCAGAAGACGGCATACGAGATAAGGCCACGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**4	CAAGCAGAAGACGGCATACGAGATTCCGAAACGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**5	CAAGCAGAAGACGGCATACGAGATTACGTACGGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**6	CAAGCAGAAGACGGCATACGAGATATCCACTCGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**7	CAAGCAGAAGACGGCATACGAGATAAAGGAATGTGACTGGAGTTCAGACGTGTGCT
G-B primer-rev**8	CAAGCAGAAGACGGCATACGAGATATATCAGTGTGACTGGAGTTCAGACGTGTGCT
CCR5AonCfwd	CGACGGTCTAGAGTCTTCATTACACCTGCAGCTCTCATTTTCCATACAGT
CCR5Amut1fwd	CGACGGTCTAGAGTCTTCATTACAtCTGCAcCTCTCATTTTCCATACAGT
CCR5Amut2fwd	CGACGGTCTAGAGTCTTCAaTACACCTGtAGCTCTCATTTTCCATACAGT
CCR5Amut3fwd	CGACGGTCTAGAGTCTTCgTTACACCTGCAtCTCTCATTTTCCATACAGT
CCR5Amut4fwd	CGACGGTCTAGAGTCTTaATTgCACCTGCAGCTCTCATTTTCCATACAGT
CCR5AonCrev	CCGACGAAGCTTTTCTTCCAGAATTGATACTGACTGTATGGAAAATGA
CCR5Amut1rev	CCGACGAAGCTTTTCTTaCAGAATTcATACTGACTGTATGGAAAATGA
CCR5Amut2rev	CCGACGAAGCTTTTCcTCCAGAgTTGATACTGACTGTATGGAAAATGA
CCR5Amut3rev	CCGACGAAGCTTTTCTTCCtGAATTGATAaTGACTGTATGGAAAATGA
CCR5Amut4rev	CCGACGAAGCTTTTCTTCCAGcATTGtTACTGACTGTATGGAAAATGA
ATMonAfwd	CGACGGTCTAGATTTGAATTGGGATGCTGTTTTTAGGTATTCTATTCAAATT
ATMmut1fwd	CGACGGTCTAGATTTGAATTGGGtTGCTGTTTTTAGGTATTCTATTCAAATT
ATMmut2fwd	CGACGGTCTAGATTTGAATTGcGATGCTGTTTTTAGGTATTCTATTCAAATT
ATMmut3wd	CGACGGTCTAGATTTGAgTTGGGATGCTGTTTTTAGGTATTCTATTCAAATT
ATMmut4fwd	CGACGGTCTAGATTTGAATTGGGATGCTGaTTTTAGGTATTCTATTCAAATT
ATMonArev	CCGACGAAGCTTAATAAAGACAGTAAAATAAATTTGAATAGAATACCTAAAA
ATMmut1rev	CCGACGAAGCTTAATAAAGACAGTgAAATAAATTTGAATAGAATACCTAAAA
ATMmut2rev	CCGACGAAGCTTAATAAAGAtAGTAAAATAAATTTGAATAGAATACCTAAAA
ATMmut3rev	CCGACGAAGCTTAATAAAGACAGTAAgATAAATTTGAATAGAATACCTAAAA
ATMmut4rev	CCGACGAAGCTTAATAAcGACAGTAAAATAAATTTGAATAGAATACCTAAAA
CCR5BonBfwd	CGACGGTCTAGAAAGGTCTTCATTACACCTGCAGCTCTCATTTTCCATACAGTCA
CCR5Bmut1fwd	CGACGGTCTAGAGTCTTCATTACACCTGtAGCTCTCATTTTC
CCR5Bmut2fwd	CGACGGTCTAGAGTCTTCATaACACCTGCAGCTCTCATTTTC
CCR5Bmut3fwd	CGACGGTCTAGAGTCTTCATTACACCcGCAGCTCTCATTTTC
CCR5Bmut4fwd	CGACGGTCTAGAGTCTTCATaACACCTGtAGCTCTCATTTTC
CCR5Bmut5fwd	CGACGGTCTAGAGTCTTCATTAtACCTaCAGCTCTCATTTTC
CCR5Bmut6fwd	CGACGGTCTAGAGTCTTCATTgCACCcGCAGCTCTCATTTTC
CCR5BonBrev	CCGACGAAGCTTTCTTCCAGAATTGATACTGACTGTATGGAAAATGAGAGCT
CCR5Bmut1rev	CCGACGAAGCTTTCTTCCAGAATTGATACTaACTGTATGGAAAATGAGAGCT
CCR5Bmut2rev	CCGACGAAGCTTTCTTCCAGAATTGATACTGACTGTATcGAAAATGAGAGCT
CCR5Bmut3rev	CCGACGAAGCTTTCTTCCAGAATTGATACTGACTGaATGGAAAATGAGAGCT
CCR5Bmut4rev	CCGACGAAGCTTTCTTCCAGAATTGATACcGACTGTATGGAAAATGAGAGCT
CCR5Bmut5rev	CCGACGAAGCTTTCTTCCAGAATTGATACTaACTGTATcGAAAATGAGAGCT
CCR5Bmut6rev	CCGACGAAGCTTTCTTCCAGAATTGATACTGAaTGTgTGGAAAATGAGAGCT
CCR5Bmut7rev	CCGACGAAGCTTTCTTCCAGAATTGATACTGAtaGTATGGAAAATGAGAGCT
pUC19Ofwd	GCGACACGGAAATGTTGAATACTCAT
pUC19Orev	CAGCGAGTCAGTGAGCGA

B

DNA substrate name	Oligonucleotide Combination
A1	ATMmut1fwd + ATMonArev
A2	ATMmut2fwd + ATMonArev
A3	ATMonAfwd +ATMmut1rev
A4	ATMonAfwd +ATMmut2rev
A5	ATMmut2fwd +ATMmut2rev
A6	ATMmut3fwd +ATMmut3rev
A7	ATMmut1fwd +ATMmut1rev
A8	ATMmut4fwd +ATMmut4rev
C1	CCR5Amut1fwd + CCR5AonCrev
C2	CCR5Amut2fwd + CCR5AonCrev
C3	CCR5Amut3fwd + CCR5AonCrev
C4	CCR5Amut4fwd + CCR5AonCrev
C5	CCR5AonAfwd + CCR5Amut1rev
C6	CCR5AonAfwd + CCR5Amut2rev
C7	CCR5AonAfwd + CCR5Amut3rev
C8	CCR5AonAfwd + CCR5Amut4rev
B1	CCR5Bmut1fwd + CCR5BonBrev
B2	CCR5Bmut2fwd + CCR5BonBrev
B3	CCR5Bmut3fwd + CCR5BonBrev
B4	CCR5BonBfwd + CCR5Bmut1rev
B5	CCR5BonBfwd + CCR5Bmut2rev
B6	CCR5BonBfwd + CCR5Bmut3rev
B7	CCR5BonBfwd + CCR5Bmut4rev
B8	CCR5Bmut4fwd + CCR5BonBrev
B9	CCR5Bmut5fwd + CCR5BonBrev
B10	CCR5Bmu6fwd + CCR5BonBrev
B11	CCR5BonBfwd + CCR5Bmut5rev
B12	CCR5BonBfwd + CCR5Bmut6rev
B13	CCR5BonBfwd + CCR5Bmut7rev
B14	CCR5Bmut1fwd + CCR5Bmut1rev
B15	CCR5Bmut2fwd + CCR5Bmut2rev
B16	CCR5Bmut1fwd + CCR5Bmut3rev

C

Site	Fwd primer	Rev primer	PCR
OnCCR5A	TCACTTGGGTGGTGGCTGTG	GACCATGACAAGCAGCGGCA	
OffC-1	AGTCCAAGACCAGCCTGGGG	AAGAACCTGTTGTCTAATCCAGCA	
OffC-2	GAACCTGTTGTCTAATCCAGCGTC	CTGCAAAGAAGGCCAGGCA	
OffC-3	AGTCCAAGACCAGCCTGGGG	AAGAACCTGTTGTCTAATCCAGCA	
OffC-4	TGACCTGTTTGTTCAGGTCTTCC	CCATATGGTCCCTGTCGCAA	
OffC-5	TCCAGTTGCTGTCCCTTCAGA	ACAGGGAGAGCCACCAATGC	
OffC-6	GCCCGGCCTGTCCTGTATTT	CACCCACACATGCACTTCCC	
OffC-7	TGGCTATTCTAGTTCTTTTGCAT	CCATGCCCTAGGGATTTGTGGA	
OffC-8	CGCTGAAGGCTGTCACCCTAA	TGGACCTAAGAGTCCTGCCCAT	
OffC-9	CCACCACCACACAACTTCACA	CAGCTGGCGAGAACTGCAAA	ND
OffC-10	TTCCAGGTCCTTTGCACAAATA	GCAAGGTCGTTGGATAGAAGTTGA	
OffC-11	CACCGAAAGCAACCCATTCC	TGATCTGCCCACCCCAGACT	
OffC-12	TTCATTCTCACCATCTGGAATTGG	TCTGGCTGGACTGCTCTGGTT	
OffC-13	TGGCATGTGGATCAGTACCCA	TAGAACATGCCCGCGAACAG	
OffC-14	CTGACGTCCATGTCAACGGG	TTTGAATTCCCCCTCCCCAT	
OffC-15	GCTCCTTTCTGAGAAGCACCCAT	GGCAGATGGTGGCAGGTCTT	
OffC-16	ATGAGGGCTTGGATTGGCTG	CCACCTCCCCCACTGCAATA	

OffC-17	GGAGGCCTTCATTGTGTCACG	AACTCCACCTGGGTGCCCTA	
OffC-18	CGTGGTCCCCCAGAAATCAC	GGAGCAGGAGTTGGTGGCAT	
OffC-19	GATTGCATAGGTTAGCATTGCC	GCCCCTGTTGGTTGACTCCC	
OffC-20	TTCCAGCGAATGGAAAGTGCT	AAGCCCAGGAATAAGGGCCA	
OffC-21	AAGCATGCTCACACTGTGGTGTA	TTGCTTGAGGCGGAAGTTGC	
OffC-22	TGACCCTCCAGCAAAGGTGA	CCCCAGGGACTGAGCATGAG	
OffC-23	GCTTTGCTTGCACTGTGCCTT	GGGGACAGACTGTGAGGGCT	
OffC-24	TCAAAAGGATGTGATCTGCCACA	GGCCTCTTTGAGGGCCAGTT	
OffC-25	CCAGGGCTCAATTCTTAGACCG	AAAAGAGCAGGGCTGCCATC	
OffC-26	TGTTCATGCCTGCACAGTGG	TGGATGTGCCCTCTACCACA	
OffC-27	TTTGGCAAGGAATTCACAGTTC	TCATGCCTGCACAGTGGTTG	
OffC-28	GGAGGATGTCTTTGTGGTAGGGG	CGCTGCCAAGCAAACTCAAA	
OffC-29	TCCCCCAACTTCACTGTTTTT	GCAATGAGCATGTGGACACCA	
OffC-30	TTCTCTGTTTCCAGTGATTTCAGA	GTCGCAAAACAGCCAGTTGC	+DMSO
OffC-31	TGGCTTGGTTAATGGACAATGG	CCTGCAAGGAGCAAGGCTTC	+DMSO
OffC-32	TGGGCTTCGTTGACTTAAAGAG	GGACAAGAGGGCCAGGGTTT	
OffC-33	TCTTAAACATGTGGAACCCAGTCAT	TGAAAACCCACAGAGTGGGAGA	
OffC-34	GCAGATTCATTAGCGTTTGTGGC	TGCATGGGTGTAAATGTAGCAGAAA	
OffC-35	CCAAGGATCAATACCTTTGGAGGA	GCCCTCCCTTGAATCAGGCT	
OffC-36	TTCCCCTAACCAGGGGCAGT	GTGGTGAGTGGGTGTGGCAG	+DMSO
OnATM	AGCGCCTGATTCGAGATCCT	AGCGCCTGATTCGAGATCCT	
OffA-1	CCTGCCATTGAATTCCAGCCT	TGTCTGCCTTTCCTGTCCCC	
OffA-2	GACTGCCACTGCACTCCCAC	GGATACCCTTGCCTCCCCAC	
OffA-3	CCTCCCATTTTCCTTCCTCCA	CTGGGAGACACAGGTGGCAG	
OffA-4	TCCTCCAATTTTCCTTCCTCCA	CTGGGAGACACAGGTGGCAG	
OffA-5	CTGGGAGACACAGGTGGCAG	AGGACCAATGGGGCCAATCT	
OffA-6	CTGGGAGACACAGGTGGCAG	AGGACCAATGGGGCCAATCT	
OffA-7	CTGGGAGACACAGGTGGCAG	AGGACCAATGGGGCCAATCT	
OffA-8	GCATGCCAAAGAAATTGTAGGC	TTCCCCCTGTCATGGTCTTCA	
OffA-9	GCATCTCTGCATTCCTCAGAAGTGG	AGAAACTGAGCAAGCCTCAGTCAA	
OffA-10	GGGATACCAAAGAGCTTTTGTTTTGTT	CAGAGGCTGCATGATGCCTAATA	
OffA-11	TGCAGCTACGGATGAAAACCAT	TCAGAATACCTCCCCGCCAG	
OffA-12	GCATAAAGCACAGGATGGGAGA	TCCCTCTTTAACGGTTATGTTGGC	
OffA-13	TGGGTTAAGTAATTTCGAAAGGAGAA	ATGTGCCCCACACATTGCC	+DMSO
OffA-14	GAGTGAGCCACTGCACCCAG	CGTGTGGTGGTGGCACAAG	ND
OffA-15	CCTCCCTCTGGCTCCCTCCC	ACCAGGGCCTGTTGGGGGTT	
OffA-16	TGCTCCCTGACCTTCCTGAGA	CCATTGGAATGAGAACCTTCTGG	
OffA-17	GGTGGAACAATCCACCTGTATTAGC	GAATGTGACACCACCACCGC	
OffA-18	GGCTTTGCAAACATAAACACTCA	CCTTCTGAGCAGCTGGGACAA	
OffA-19	CACTGGAACCCAGGAGGTGG	CCTCCCATTGGAGCCTTGGT	
OffA-20	CAGCCTGCCTGGGTGACAG	CATCTGAGCTCAAAACTGCTGC	+DMSO
OffA-21	GCCACTGCATTGCATTTTCC	TGAGGGCAGGTCTGTTTCCTG	ND
OffA-22	GGGAGGATCTCTCGAGTCCAGG	CCTTGCCTGACTTGCCCTGT	
OffA-23	TGTTTAGTAATTAAGACCCTGGCTTTC	GCGACAGGTACAAAGCAGTCCAT	
OffA-24	GCCCTTTGATTTCATCTGTTTCCC	CATTGCTGCCATTGCACTCC	
OffA-25	AAACTGGCACATGTACTCCT	ACATGATTTGATTTTTCATGTGTTT	
OffA-26	GGGTGGAAGGTGAGAGGAGATT	CGCAGATGGGCATGTTATTG	ND
OffA-27	CCTCCCATTTTCCTTCCTCCA	GACTGCCACTGCACTCCCAC	
OffA-28	AGCCAAGATTGCACCATTGC	GTCCCTGACGGAGGCTGAGA	ND
OffA-29	TGGTTGGATTTTGGCTCTGTCAC	TGTCAATATCAATACCCTGCTTTCCTC	

OffA-30	TGGTTACTTTTAAAGGGTCATGATGGA	AAAAATGGATGCAAAGCCAAA	+DMSO
OffA-31	GGGACACAGAGCCAAACCGT	TGTGCACATGTACCCTAAAACT	ND
OffA-32	CAGTCATTGTTTCTAGGTAGGGGA	TTGGCAATTTGGGTGCAACA	
OffA-33	TGGATAACCTGCAGATTTGTTTCTG	TGAGCCCAGGAGTTTCAGGC	
OffA-34	TCGTGTGTGTGTGTTTGCTTCA	CAGTGGTTCGGGAAACAGCA	
OffA-35	TGGGAATGTAAATCTGACTGGCTG	CTGGAACTCTGGGCATGGCT	
OffA-36	GCTGCAATTGCTTTTTGGCA	TGGACCCCTCCCTTACACC	

Table 10. Oligonucleotides used in this study. (A) All oligonucleotides were purchased from Integrated DNA Technologies. '/5Phos/' indicates 5' phosphorylated oligonucleotides. A % symbol indicates that the preceding nucleotide was incorporated as a mixture of phosphoramidites consisting of 79 mol% of the phosphoramidite corresponding to the preceding nucleotide and 7 mol% of each of the other three canonical phosphoramidites. An (*) indicates that the oligonucleotide primer was specific to a selection sequence (either CCR5A, ATM or CCR5B). An (**) indicates that the oligonucleotide adapter or primer had a unique sequence identifier to distinguish between different samples (selection conditions or cellular TALEN treatment). (B) Combinations of oligonucleotides used to construct discrete DNA substrates used in TALEN digestion assays. (C) Primer pairs for PCR amplifying on-target and off-target genomic sites. +DMSO: DMSO was used in the PCR; ND: no correct DNA product was detected from the PCR reaction. Sequences relate to SEQ ID NOs: XX-XX.

CLAIMS

What is claimed is:

1. A Transcription Activator-Like Effector (TALE) domain, wherein

(i) the TALE domain comprises an N-terminal TALE domain according to SEQ ID NO: 1, wherein at least one of K110, K113, and R114 is replaced with an amino acid residue that does not have a cationic charge, has no charge, or has an anionic charge at physiological pH; or

(ii) the TALE domain comprises a C-terminal TALE domain according to SEQ ID NO: 22, wherein at least one of R8, R30, K37, K38, K48, R49, R52, R53, R57, and R61 is replaced with an amino acid residue that does not have a cationic charge, has no charge, or has an anionic charge at physiological pH.

2. The TALE domain of claim 1, wherein the net charge of the C-terminal domain is less than or equal to +6, less than or equal to +5, less than or equal to +4, less than or equal to +3, less than or equal to +2, less than or equal to +1, less than or equal to 0, less than or equal to -1, less than or equal to -2, less than or equal to -3, less than or equal to -4, or less than or equal to -5.

3. The TALE domain of any one of claim 1 or claim 2, wherein at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, or at least 15 cationic amino acid(s) in the TALE domain is/are replaced with an amino acid residue that exhibits no charge or a negative charge at physiological pH.

4. The TALE domain of claim 3, wherein the at least one cationic amino acid residue is arginine (R) or lysine (K).

5. The TALE domain of any one of claims 1-4, wherein the amino acid residue that does not have a cationic charge, has no charge, or has an anionic charge at physiological pH is glutamine (Q) or glycine (G).

6. The TALE domain of any one of claims 3-5, wherein at least one lysine or arginine residue is replaced with a glutamine residue.

7. The TALE domain of any one of claims 1-6, wherein the C-terminal domain comprises one or more of the following amino acid replacements: K37Q, K38Q, K48Q, R49Q, R52Q, R53Q, or R61Q in SEQ ID NO: 22.

8. The TALE domain of any one of claims 1-7, wherein the C-terminal domain comprises a Q3 variant sequence (K48Q, R52Q, and R61Q in SEQ ID NO: 22).

9. The TALE domain of any one of claims 1-7, wherein the C-terminal domain comprises a Q7 variant sequence (K37Q, K38Q, K48Q, R49Q, R52Q, R53Q, and R61Q in SEQ ID NO: 22).

10. The TALE domain of any one of claims 1-9, wherein the N-terminal domain is a truncated version of the canonical N-terminal domain.

11. The TALE domain of any one of claims 1-7 or 10, wherein the C-terminal domain is a truncated version of the canonical C-terminal domain.

12. The TALE domain of claim 11, wherein the truncated domain comprises less than 90%, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, or less than 25% of the residues of the canonical domain.

13. The TALE domain of claim 11, wherein the truncated C-terminal domain comprises less than 60, less than 50, less than 40, less than 30, less than 29, less than 28, less than 27, less than 26, less than 25, less than 24, less than 23, less than 22, less than 21, or less than 20 amino acid residues.

14. The TALE domain of claim 11, wherein the truncated C-terminal domain comprises 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33,

2014308713 03 Sep 2020

32, 31, 30, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 residues.

15. The TALE domain of any one of claims 1-14, wherein the TALE domain is comprised in a TALE molecule comprising the structure:

[N-terminal domain]-[TALE repeat array]-[C-terminal domain]-[effector domain]

or

[effector domain]-[N-terminal domain]-[TALE repeat array]-[C-terminal domain].

16. The TALE domain of claim 15, wherein the effector domain comprises a nuclease domain, a transcriptional activator or repressor domain, a recombinase domain, or an epigenetic modification enzyme domain.

17. The TALE domain of claim 15 or 16, wherein the TALE molecule binds a target sequence within a gene known to be associated with a disease or disorder.

18. The TALE domain of any preceding claim, wherein the binding energy of the N-terminal domain to a target nucleic acid molecule is less than the binding energy of the canonical N-terminal domain (SEQ ID NO: 1).

19. The TALE domain of any preceding claim, wherein the binding energy of the C-terminal domain to a target nucleic acid molecule is less than the binding energy of the canonical C-terminal domain (SEQ ID NO: 22).

20. The TALE domain of any preceding claim, wherein the N-terminal TALE domain comprises the amino acid sequence of SEQ ID NO: 2, SEQ ID NO: 3, or SEQ ID NO: 4.

21. The TALE domain of any preceding claim, wherein the C-terminal TALE domain comprises the amino acid sequence of SEQ ID NO: 23 or SEQ ID NO: 24.

22. A Transcription Activator-Like Effector Nuclease (TALEN), comprising

(a) a nuclease cleavage domain;

(b) a C-terminal domain as defined in any one of claims 1-14, conjugated to the nuclease cleavage domain;

(c) a TALE repeat array conjugated to the C-terminal domain; and

(d) an N-terminal domain as defined in any one of claims 1-10, conjugated to the TALE repeat array.

23. The TALEN of claim 22, wherein the nuclease cleavage domain is a FokI nuclease domain.

24. The TALEN of claim 23, wherein the FokI nuclease domain comprises a sequence as provided in SEQ ID NOs: 26-30.

25. The TALEN of any one of claims 22, 23, and 24, wherein the TALEN is a monomer.

26. The TALEN of claim 25, wherein the TALEN monomer can dimerize with another TALEN monomer to form a TALEN dimer.

27. The TALEN of claim 26, wherein the dimer is a heterodimer.

28. The TALEN of any one of claims 22-27, wherein the TALEN binds a target sequence within a gene known to be associated with a disease or disorder.

29. The TALEN of claim 28, wherein the TALEN cleaves the target sequence upon dimerization.

30. The TALEN of claim 28 or 29, wherein the disease is HIV/AIDS or a proliferative disease.

31. The TALEN of any one of claims 23-30, wherein the TALEN binds a CCR5 target sequence.

32. The TALEN of any one of claims 21-30, wherein the TALEN binds an ATM target sequence.

33. The TALEN of any one of claims 21-30, wherein the TALEN binds a VEGFA target sequence.

34. A composition comprising the TALEN of any one of claims 21-33 and a different TALEN that can form a heterodimer with the TALEN, wherein the dimer exhibits nuclease activity.

35. A pharmaceutical composition comprising the TALEN of any one of claims 21-33, or the composition of claim 34, and a pharmaceutically acceptable excipient.

36. The pharmaceutical composition of claim 35, wherein the pharmaceutical composition is formulated for administration to a subject.

37. The pharmaceutical composition of claim 35 or 36, wherein the pharmaceutical composition comprises an effective amount of the TALEN for cleaving a target sequence in a cell in the subject.

38. The pharmaceutical composition of claim 35 or 36, wherein the TALEN binds a target sequence within a gene known to be associated with a disease or disorder and wherein the composition comprises an effective amount of the TALEN for alleviating a symptom associated with the disease or disorder.

39. A method of cleaving a target sequence in a nucleic acid molecule, comprising contacting a nucleic acid molecule comprising the target sequence with a TALEN binding the target sequence under conditions suitable for the TALEN to bind and cleave the target sequence, wherein the TALEN is a TALEN of any one of claims 21-33, or wherein the TALEN is comprised in the composition of claim 34 or the pharmaceutical composition of any one of claims 35-38.

40. The method of claim 39, wherein the target sequence is comprised in a cell.

41. The method of claim 39 or 40, wherein the target sequence is comprised in a subject.

42. The method of claim 41, wherein the method comprises administering the composition or the pharmaceutical composition comprising the TALEN to the subject in an amount sufficient for the TALEN to bind and cleave the target site.

43. A method of preparing an engineered TALEN comprising the TALE domain as defined in any one of claims 1-21, the method comprising:

replacing at least one amino acid in the canonical N-terminal TALEN domain and/or the canonical C-terminal TALEN domain with an amino acid residue that does not have a cationic charge, has no charge, or has an anionic charge at physiological pH; and/or

truncating the N-terminal TALEN domain and/or the C-terminal TALEN domain to remove a positively charged fragment;

thus generating an engineered TALEN having an N-terminal domain and/or a C-terminal domain of a decreased net charge.

44. The method of claim 43, wherein the at least one amino acid being replaced comprises a cationic amino acid or an amino acid having a positive charge at physiological pH.

45. The method of claim 43 or 44, wherein the amino acid replacing the at least one amino acid is a cationic amino acid or a neutral amino acid.

46. The method of any one of claims 43-45, wherein the truncated N-terminal TALEN domain and/or the truncated C-terminal TALEN domain comprises less than 90%, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, or less than 25% of the residues of the respective canonical domain.

47. The method of any one of claims 43-46, wherein the truncated C-terminal domain comprises less than 60, less than 50, less than 40, less than 30, less than 29, less than 28, less than 27, less than 26, less than 25, less than 24, less than 23, less than 22, less than 21, or less than 20 amino acid residues.

48. The method of any one of claims 43-46, wherein the truncated C-terminal domain comprises 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 amino acid residues.

49. The method of any one of claims 43-48, wherein the method comprises replacing at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, or at least 15 amino acids in the canonical N-terminal TALEN domain and/or in the canonical C-terminal TALEN domain with an amino acid having no charge or a negative charge at physiological pH.

50. The method of any one of claims 43-49, wherein the amino acid residue having no charge or a negative charge at physiological pH is glutamine (Q) or glycine (G).

51. The method of any one of claims 43-50, wherein the method comprises replacing at least one lysine or arginine residue with a glutamine residue.

52. The method of any one of claims 43-51, wherein the net charge of the C-terminal domain is less than or equal to +5 at physiological pH.

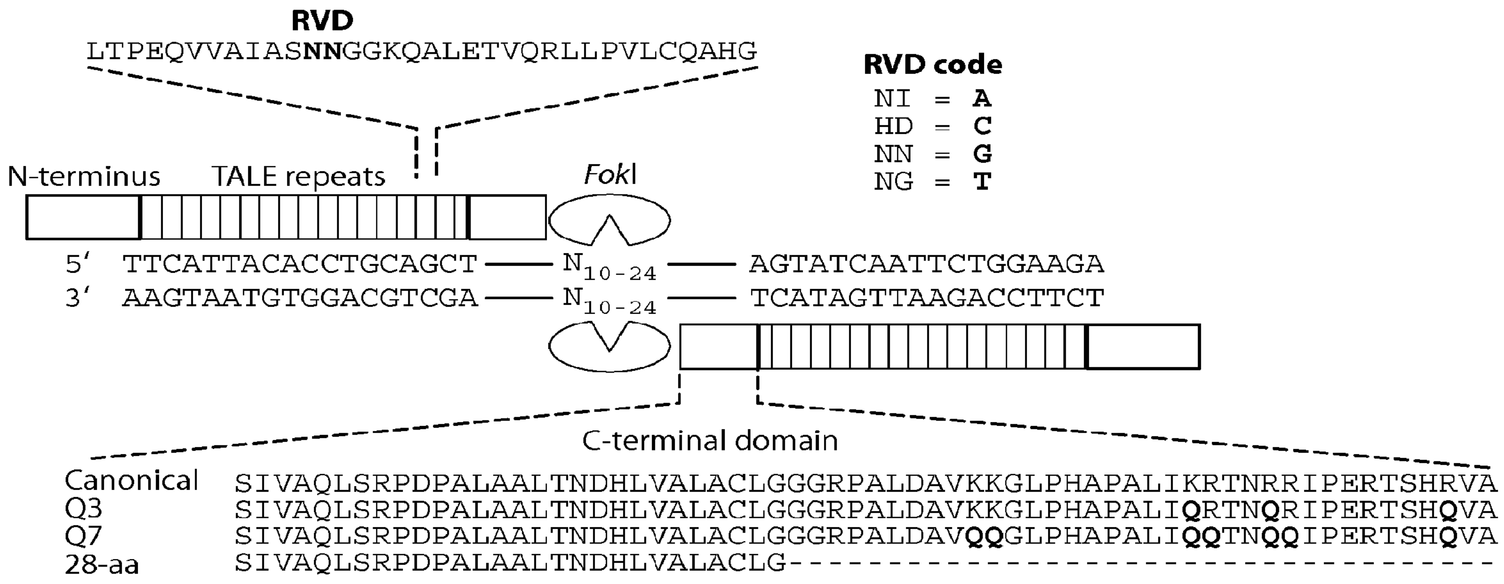

RVD
LTPEQVVAIASNNGGKQALETVQRLLPVLCQAHG
RVD code
NI = A
HD = C
NN = G
NG = T
N-terminus
TALE repeats
FokI
5′ TTCATTACACCTGCAGCT — N_{10-24} — AGTATCAATTCTGGAAGA
3′ AAGTAATGTGGACGTCGA — N_{10-24} — TCATAGTTAAGACCTTCT
C-terminal domain
Canonical SIVAQLSRPDPALAALTNDHLVALACLGGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVA
Q3 SIVAQLSRPDPALAALTNDHLVALACLGGGRPALDAVKKGLPHAPALIQRTNQRIPERTSHQVA
Q7 SIVAQLSRPDPALAALTNDHLVALACLGGGRPALDAVQQGLPHAPALIQQTNQQIPERTSHQVA
28-aa SIVAQLSRPDPALAALTNDHLVALACLG-----------------------------------

Fig. 1A

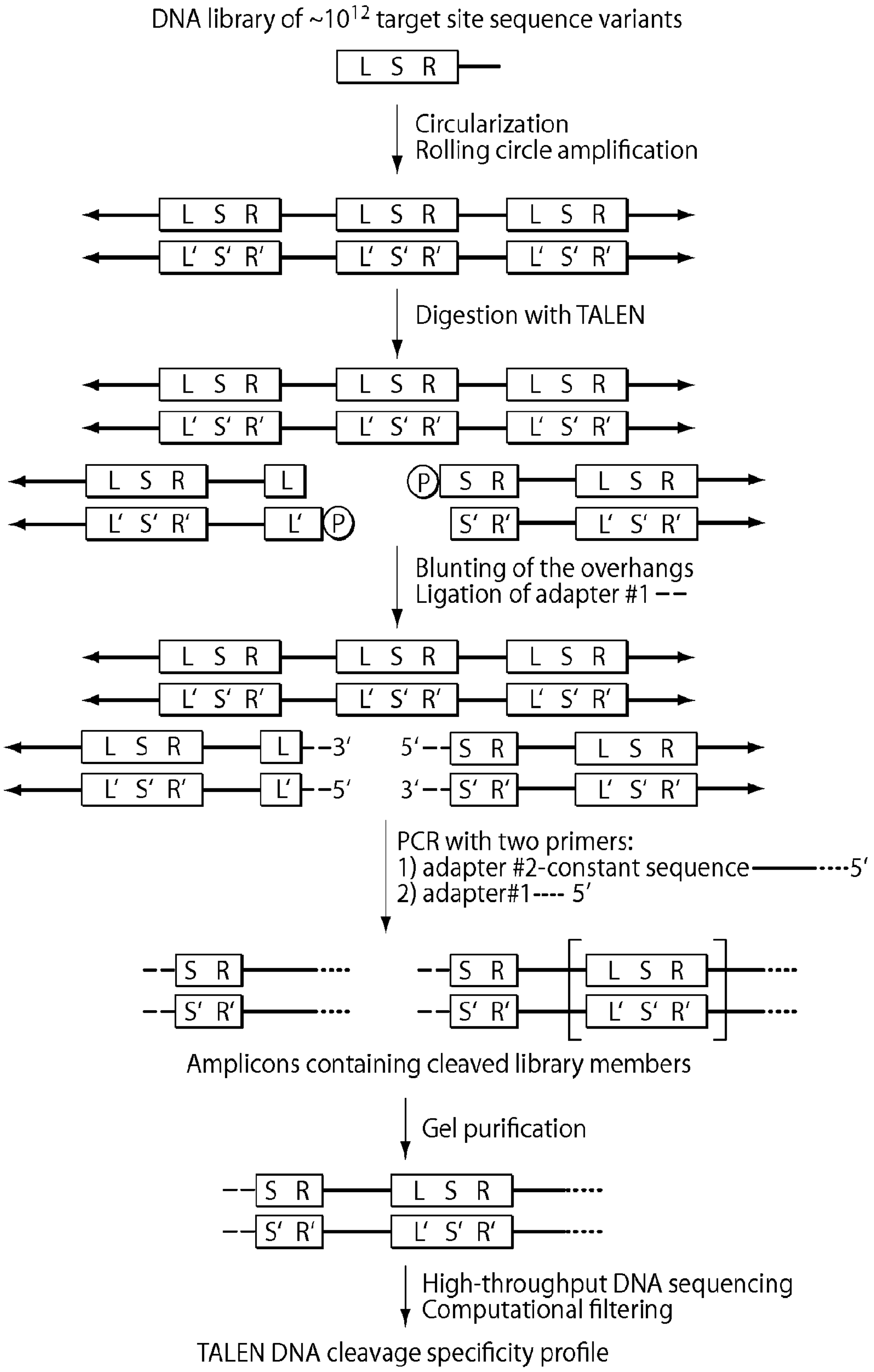

DNA library of ~10¹² target site sequence variants
L S R
Circularization
Rolling circle amplification
L S R
L' S' R'
Digestion with TALEN
P
Blunting of the overhangs
Ligation of adapter #1 --
3'
5'
PCR with two primers:
1) adapter #2-constant sequence ----5'
2) adapter#1---- 5'
Amplicons containing cleaved library members
Gel purification
High-throughput DNA sequencing
Computational filtering
TALEN DNA cleavage specificity profile

Fig. 1B

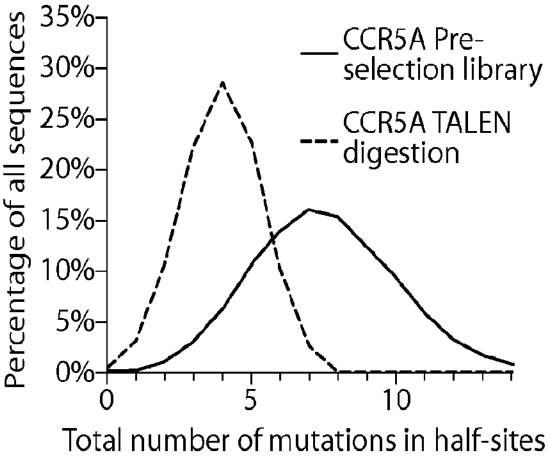

Percentage of all sequences
35%
30%
25%
20%
15%
10%
5%
0%
0
5
10
Total number of mutations in half-sites
CCR5A Pre-selection library
CCR5A TALEN digestion

Fig. 2A

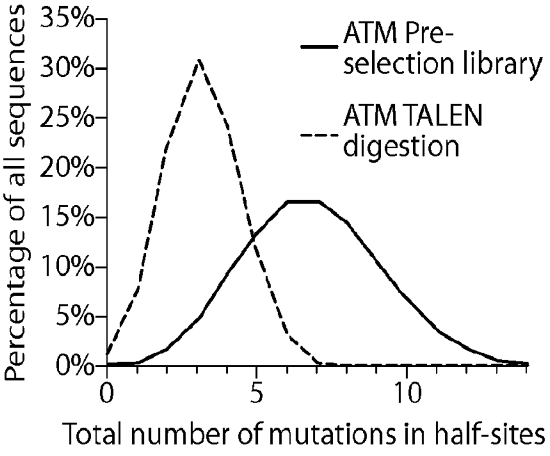

Percentage of all sequences
35%
30%
25%
20%
15%
10%
5%
0%
0
5
10
Total number of mutations in half-sites
ATM Pre-selection library
ATM TALEN digestion

Fig. 2B

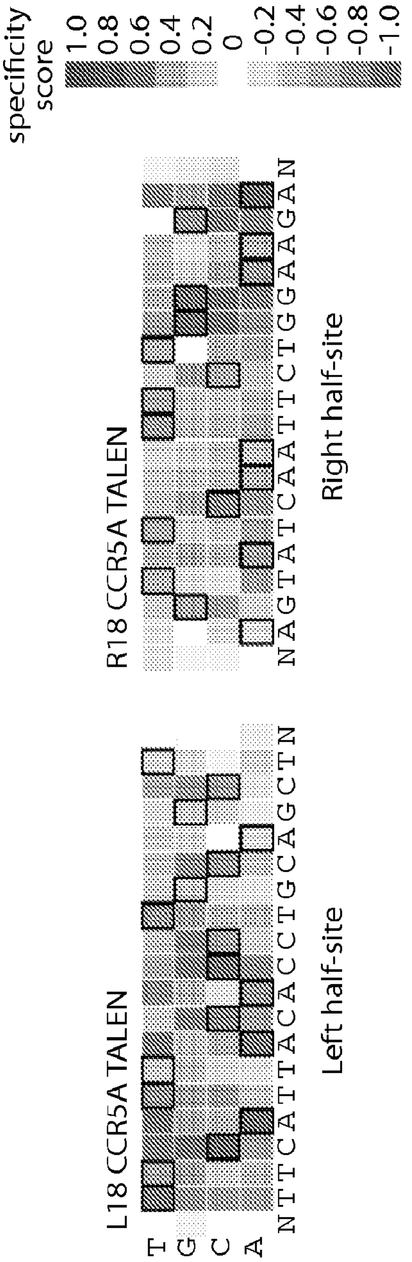

L18 CCR5A TALEN
T
G
C
A
NTTCATTTACACCTGCAGCTN
Left half-site
R18 CCR5A TALEN
NAGTATCAATTCTGGAAGAN
Right half-site
specificity score
1.0
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1.0

Fig. 2C

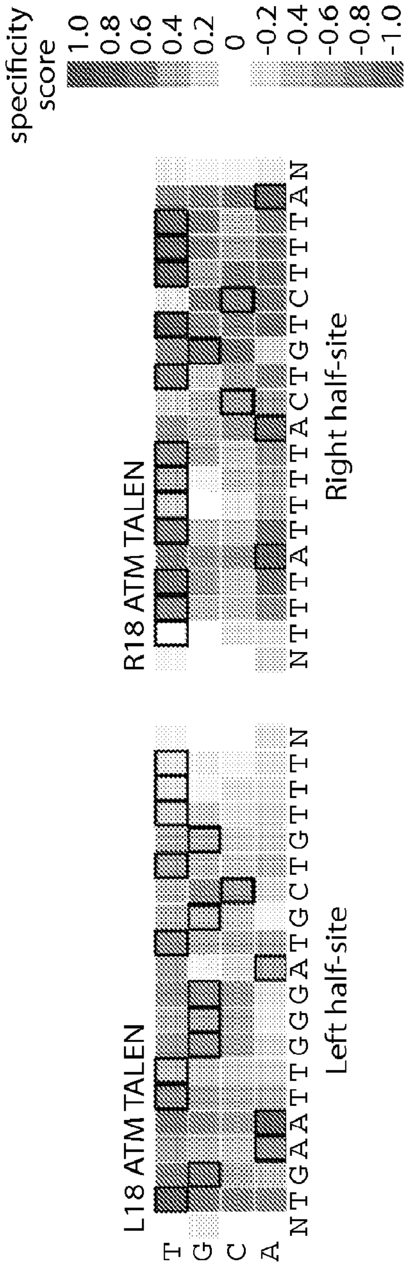

L18 ATM TALEN
T
G
C
A
NTGAATTGGGATGCTGTTTN
Left half-site
R18 ATM TALEN
NTTTATTTTACTGTCTTTAN
Right half-site
specificity score
1.0
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1.0

Fig. 2D

DNA	Enrichment	Left half-site	Right half-site
OnB	1.0	TCATTACACCTGC	CATACAGTCAGTA
B1	0.47	TCATTACACCTG**t**	CATACAGTCAGTA
B2	0.46	TCAT**a**ACACCTGC	CATACAGTCAGTA
B3	0.13	TCATTACACC**c**GC	CATACAGTCAGTA
B4	1.1	TCATTACACCTGC	CATACAGT**a**AGTA
B5	0.84	TCATTACACCTGC	**g**ATACAGTCAGTA
B6	0.21	TCATTACACCTGC	CAT**t**CAGTCAGTA
B7	0.03	TCATTACACCTGC	CATACAGTC**g**GTA
B8	0.25	TCAT**a**ACACCTG**t**	CATACAGTCAGTA
B9	0.25	TCATTA**t**ACCT**a**C	CATACAGTCAGTA
B10	0.00	TCATT**g**CACC**c**GC	CATACAGTCAGTA
B11	0.55	TCATTACACCTGC	**g**ATACAGT**a**AGTA
B12	0.04	TCATTACACCTGC	CA**c**ACA**t**TCAGTA
B13	0.03	TCATTACACCTGC	CATAC**ta**TCAGTA
B14	0.30	TCATTACACCTG**t**	CATACAGT**a**AGTA
B15	0.18	TCAT**a**ACACCTGC	**g**ATACAGTCAGTA
B16	0.04	TCATTACACCTG**t**	CAT**t**CAGTCAGTA

Fig. 2E

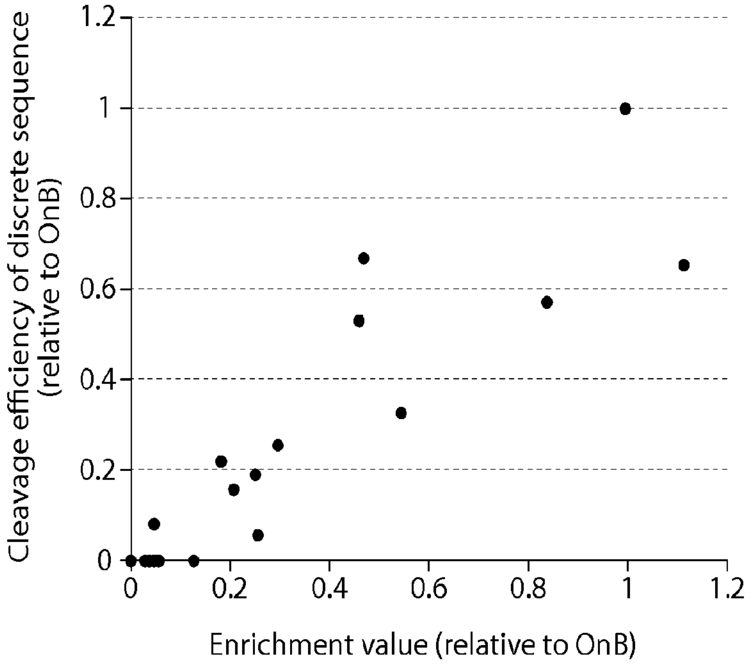

Cleavage efficiency of discrete sequence (relative to OnB)
1.2
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
1.2
Enrichment value (relative to OnB)

Fig. 2F

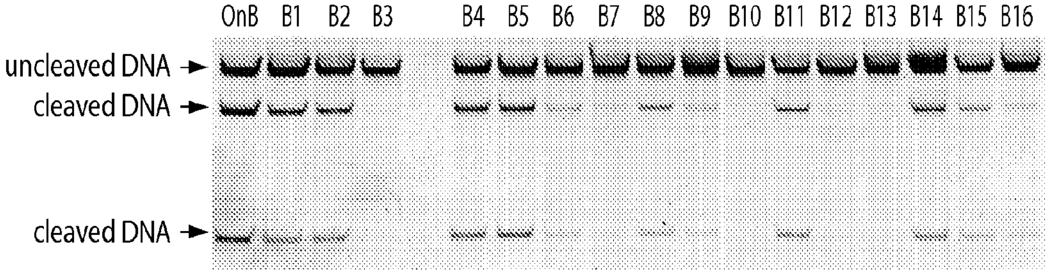

OnB
B1
B2
B3
B4
B5
B6
B7
B8
B9
B10
B11
B12
B13
B14
B15
B16
uncleaved DNA
cleaved DNA
cleaved DNA

Fig. 2G

Site	No TALEN	CCR5A EL/KK *Fok*I	CCR5A ELD/KKR *Fok*I	CCR5A Homo *Fok*I
OnCCR5A	<0.006%	9.84%	27.6%	46.8%
OffC-2	<0.006%	<0.006%	0.006%	<0.006%
OffC-5	<0.006%	0.45%	1.96%	2.78%
OffC-15	<0.020%	<0.014%	0.230%	0.043%
OffC-16	<0.006%	<0.006%	0.031%	<0.006%
OffC-28	<0.009%	0.014%	0.163%	0.056%
OffC-36	<0.006%	<0.006%	0.152%	0.028%

Fig. 3A

Site	No TALEN	ATM EL/KK *Fok*I	ATM ELD/KKR *Fok*I	ATM Homo *Fok*I
OnATM	0.006%	6.78%	16.2%	18.2%
OffA-1	<0.006%	<0.006%	0.03%	0.08%
OffA-11	<0.006%	<0.006%	0.04%	0.39%
OffA-13	<0.006%	0.01%	0.02%	<0.006%
OffA-16	<0.006%	<0.006%	<0.006%	0.06%
OffA-17	<0.05%	<0.14%	<0.17%	0.94%
OffA-23	0.02%	<0.006%	0.29%	0.23%
OffA-35	<0.006%	<0.006%	<0.006%	0.07%

Fig. 3B

OnCCR5A

```
TTCATTACACCTGCAGCTCTCATTTTCCATACAGTCAGTATCAATTCTGGAAGA (7267)
TTCATTACACCTGCAGCTCTCAT-------ACAGTCAGTATCAATTCTGGAAGA (76)
TTCATTACACCTGCAG------------------TCAGTATCAATTCTGGAAGA (63)
TTCATTACACCTG------------------------------------GAAGA (61)
```

OffC-2

```
TACATCACATATGCAAATTGACTCAAAATGGATCATAGACCTAAATGTGTATCATTTCTGGGAGA (163332)
TACATCACATATGCAAATTGACTCAAAATGGATCA---ACCTAAATGTGTATCATTTCTGGGAGA (6)
TACATCACATATGCAAATTGACTCAAAATG-------GACCTAAATGTGTATCATTTCTGGGAGA (4)
```

OffC-5

```
TCCAATACCTCTGCCACACCCAGGCATTGGCCAGGAGCAACTCTGGGAGA (17045)
TCCAATACCTCTGCCACAC-----------CCAGGAGCAACTCTGGGAGA (28)
TCCAATACCTCTG----------GCATTGGCCAGGAGCAACTCTGGGAGA (12)
TCCAATAC--------------------------------CTCTGGGAGA (10)
```

OffC-15

```
TCCATGACACAAAAGACTTCCCTGATTTCTTCTAAGGCATCACTGGTATCTATCCTGGAATA (6967)
TCCATGACACAAAAGACTTCCCTGATTTCTTCTAAGG-----CTGGTATCTATCCTGGAATA (6)
```

OffC-16

```
TTCCTTCCACCAGTGTCCACAGTCTTCACACTGATCACCAAATCCCAGCATCAATCCTGGAAGA (38536)
TTCCTTCCACCAGTGTCCACAGTC-----------CACCAAATCCCAGCATCAATCCTGGAAGA (4)
```

OffC-28

```
TTTATTACACTTCCAGATCTTTTATTTTAAGTTACCAGATATCCTTTCTGGAAGA (7379)
TTTATTACACTT----------------------CCAGATATCCTTTCTGGAAGA (3)
TTTATTACACTTCCAGATCTTTT---------------ATATCCTTTCTGGAAGA (2)
TTTATTACACTTCCAGATCTTT-----------------TATCCTTTCTGGAAGA (2)
```

OffC-36

```
CTCCTAATACCTGCAAATTATAAGGACACTATTTGACTTGATATTATTTCTGGAGGA (12461)
CTCCTAATACCTGCAAATTATAAGGACACT----GACTTGATATTATTTCTGGAGGA (11)
```

Fig. 3C

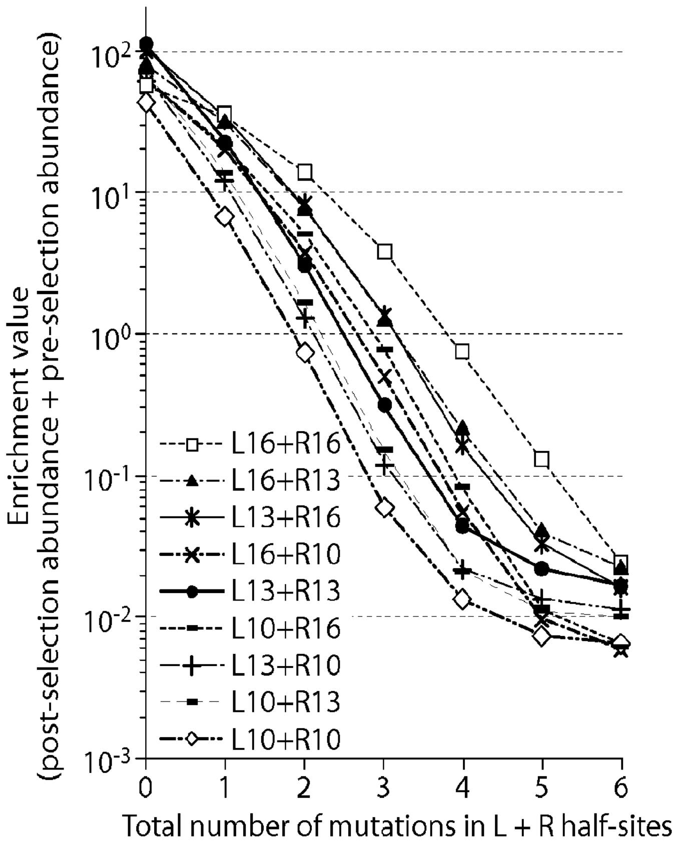

Enrichment value
(post-selection abundance ÷ pre-selection abundance)
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
0
1
2
3
4
5
6
Total number of mutations in L + R half-sites
L16+R16
L16+R13
L13+R16
L16+R10
L13+R13
L10+R16
L13+R10
L10+R13
L10+R10

Fig. 4A

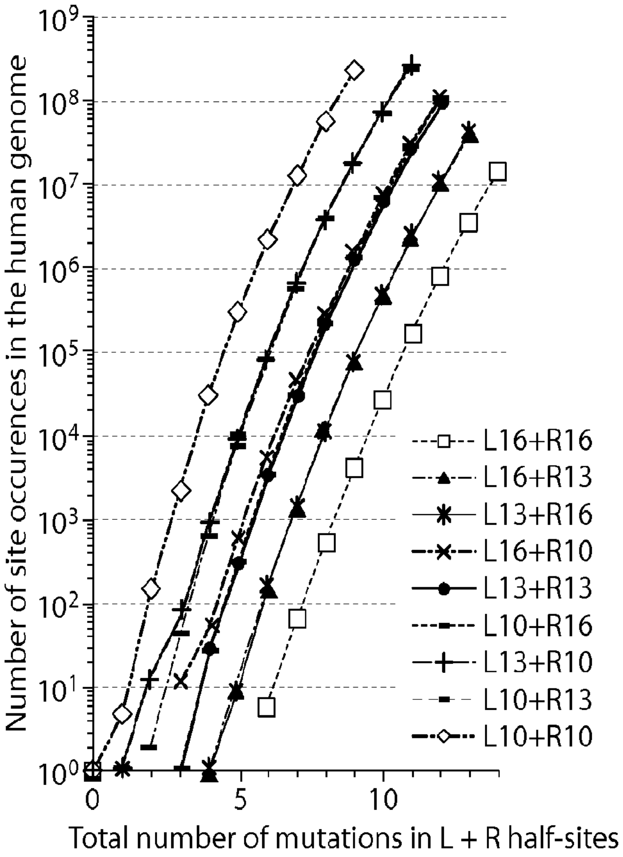

Number of site occurences in the human genome
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
5
10
Total number of mutations in L + R half-sites
L16+R16
L16+R13
L13+R16
L16+R10
L13+R13
L10+R16
L13+R10
L10+R13
L10+R10

Fig. 4B

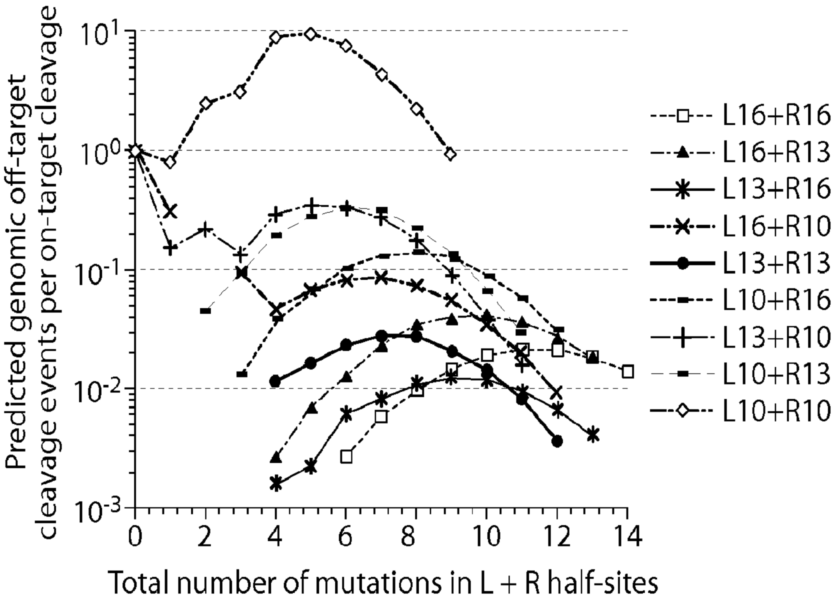

Predicted genomic off-target cleavage events per on-target cleavage
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
0
2
4
6
8
10
12
14
Total number of mutations in L + R half-sites
L16+R16
L16+R13
L13+R16
L16+R10
L13+R13
L10+R16
L13+R10
L10+R13
L10+R10

Fig. 4C

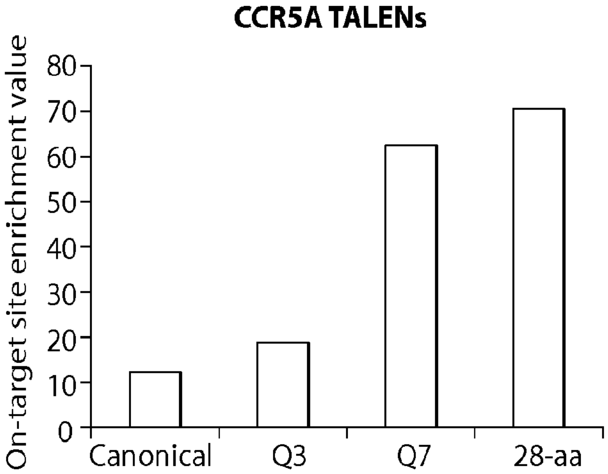

CCR5A TALENs
On-target site enrichment value
0
10
20
30
40
50
60
70
80
Canonical
Q3
Q7
28-aa

Fig. 5A

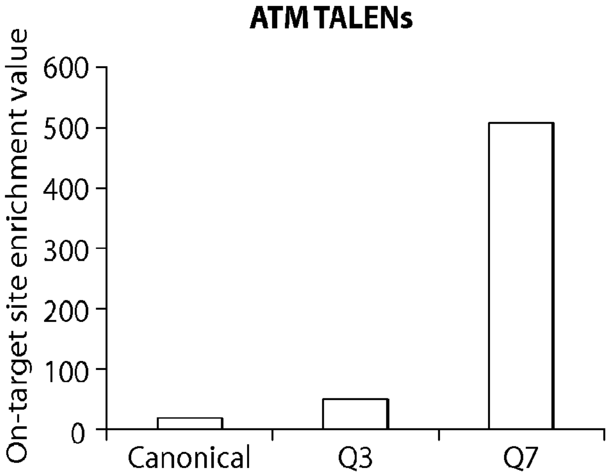

ATM TALENs
On-target site enrichment value
0
100
200
300
400
500
600
Canonical
Q3
Q7

Fig. 5B

DNA	Left half-site	Right half-site
OnC	TTCATTACACCTGCAGCT	AGTATCAATTCTGGAAGA
C1	TTCATTACA**t**CTGCA**c**CT	AGTATCAATTCTGGAAGA
C2	TTCA**a**TACACCTG**t**AGCT	AGTATCAATTCTGGAAGA
C3	TTCATTACACC**c**GCAGC**a**	AGTATCAATTCTGGAAGA
C4	TT**a**ATT**g**CACCTGCAGCT	AGTATCAATTCTGGAAGA
C5	TTCATTACACCTGCAGCT	AGTAT**g**AATTCTG**t**AAGA
C6	TTCATTACACCTGCAGCT	AGTATCAA**c**TCTGGA**g**GA
C7	TTCATTACACCTGCAGCT	A**t**TATCAATTC**a**GGAAGA
C8	TTCATTACACCTGCAGCT	AGTA**a**CAAT**g**CTGGAAGA

Fig. 5C

DNA	Left half-site	Right half-site
OnA	TGAATTGGGATGCTGTTT	TTTATTTTACTGTCTTTA
A1	TGAATTGGG**t**TGCTGTTT	TTTATTTTACTGTCTTTA
A2	TGAATTG**c**GATGCTGTTT	TTTATTTTACTGTCTTTA
A3	TGAATTGGGATGCTGTTT	TTTATTT**c**ACTGTCTTTA
A4	TGAATTGGGATGCTGTTT	TTTATTTTACT**a**TCTTTA
A5	TGAATTG**c**GATGCTGTTT	TTTATTTTACT**a**TCTTTA
A6	TGA**g**TTGGGATGCTGTTT	TTTAT**g**TTACTGTCTTTA
A7	TGAATTGGG**t**TGCTGTTT	TTTATTT**g**ACTGTCTTTA
A8	TGAATTGGGATGCTG**a**TT	TTTATTTTACTGTC**c**TTA

Fig. 5D

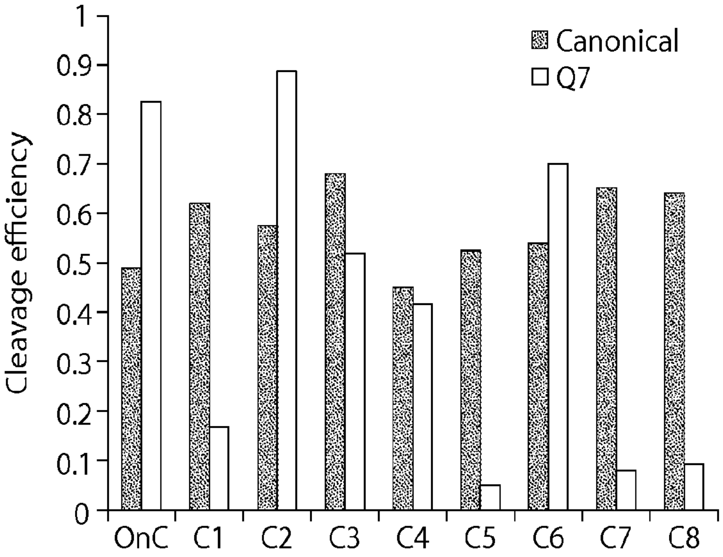

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Cleavage efficiency
▨ Canonical
□ Q7
OnC
C1
C2
C3
C4
C5
C6
C7
C8

Fig. 5E

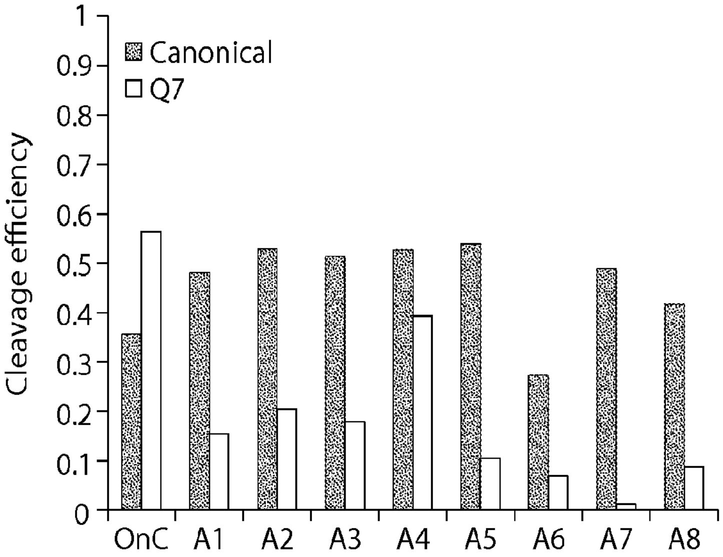

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Cleavage efficiency
▨ Canonical
□ Q7
OnC
A1
A2
A3
A4
A5
A6
A7
A8

Fig. 5F

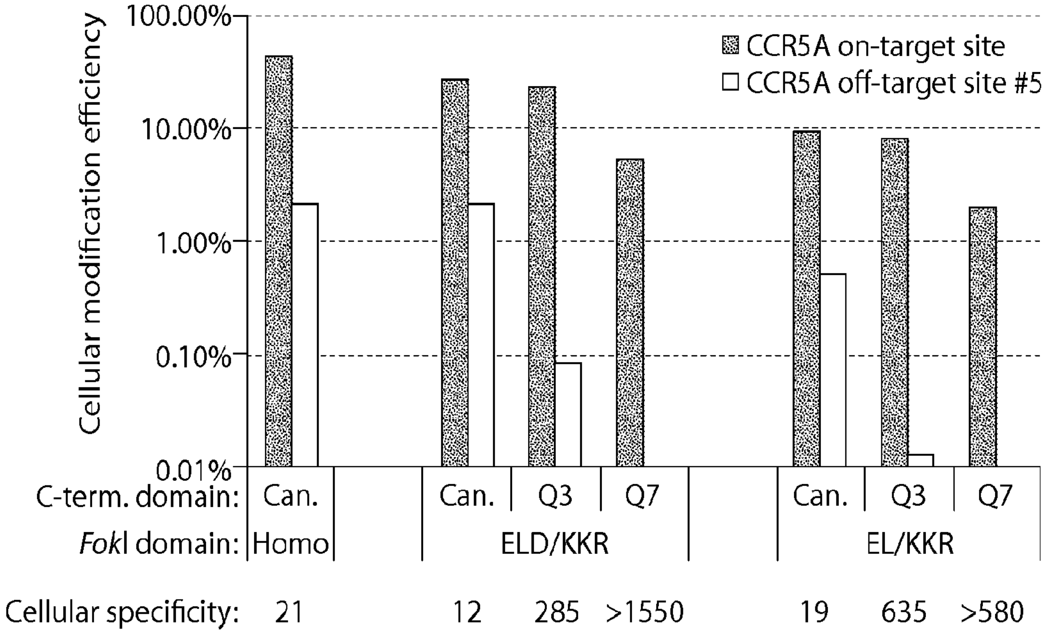

Cellular modification efficiency
100.00%
10.00%
1.00%
0.10%
0.01%
▨ CCR5A on-target site
□ CCR5A off-target site #5
C-term. domain:
Can.
Can.
Q3
Q7
Can.
Q3
Q7
*Fok*I domain:
Homo
ELD/KKR
EL/KKR
Cellular specificity:
21
12
285
>1550
19
635
>580

Fig. 6

CCR5 target sites

TALEN monomer

```
CCR5A L18           5′-TTCATTACACCTGCAGCT
CCR5B L16         5′-TCTTCATTACACCTGC
CCR5B L13            5′-TCATTACACCTGC
CCR5B L10               5′-TTACACCTGC

                     TCTTCATTACACCTGCAGCTCTCATTTTCCATACAGTCAGTATCAATTCTGGAAGA
                     AGAAGTAATGTGGACGTCGAGAGTAAAAGGTATGTCAGTCATAGTTAAGACCTTCT

CCR5A R18                                                  TCATAGTTAAGACCTTCT-5′
CCR5B R16                                         GTATGTCAGTCATAGT-5′
CCR5B R13                                         GTATGTCAGTCAT-5′
CCR5B R10                                         GTATGTCAGT-5′
```

Fig. 7A

ATM target site

TALEN monomer

```
ATM L18     5′-TGAATTGGGATGCTGTTT
               TGAATTGGGATGCTGTTTTTAGGTATTCTATTCAAATTTATTTTACTGTCTTTA
               ACTTAACCCTACGACAAAAATCCATAAGATAAGTTTAAATAAAATGACAGAAAT

ATM R18                                            AAATAAAATGACAGAAAT-5′
```

Fig. 7B

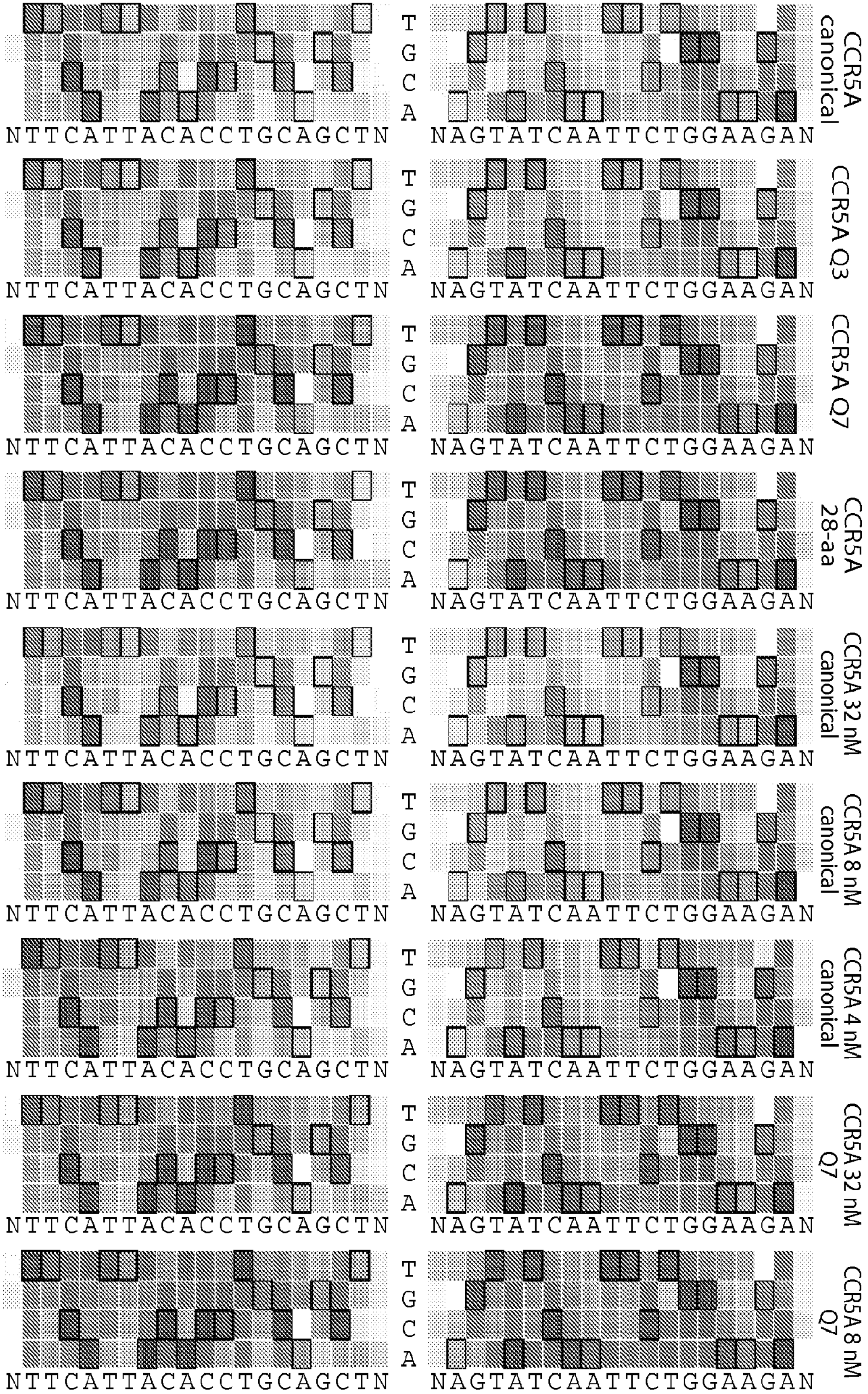

T
G
C
A
NTTCATTACACCTGCAGCTN
NAGTATCAATTCTGGAAGAN
CCR5A canonical
CCR5A Q3
CCR5A Q7
CCR5A 28-aa
CCR5A 32 nM canonical
CCR5A 8 nM canonical
CCR5A 4 nM canonical
CCR5A 32 nM Q7
CCR5A 8 nM Q7

Fig. 8A

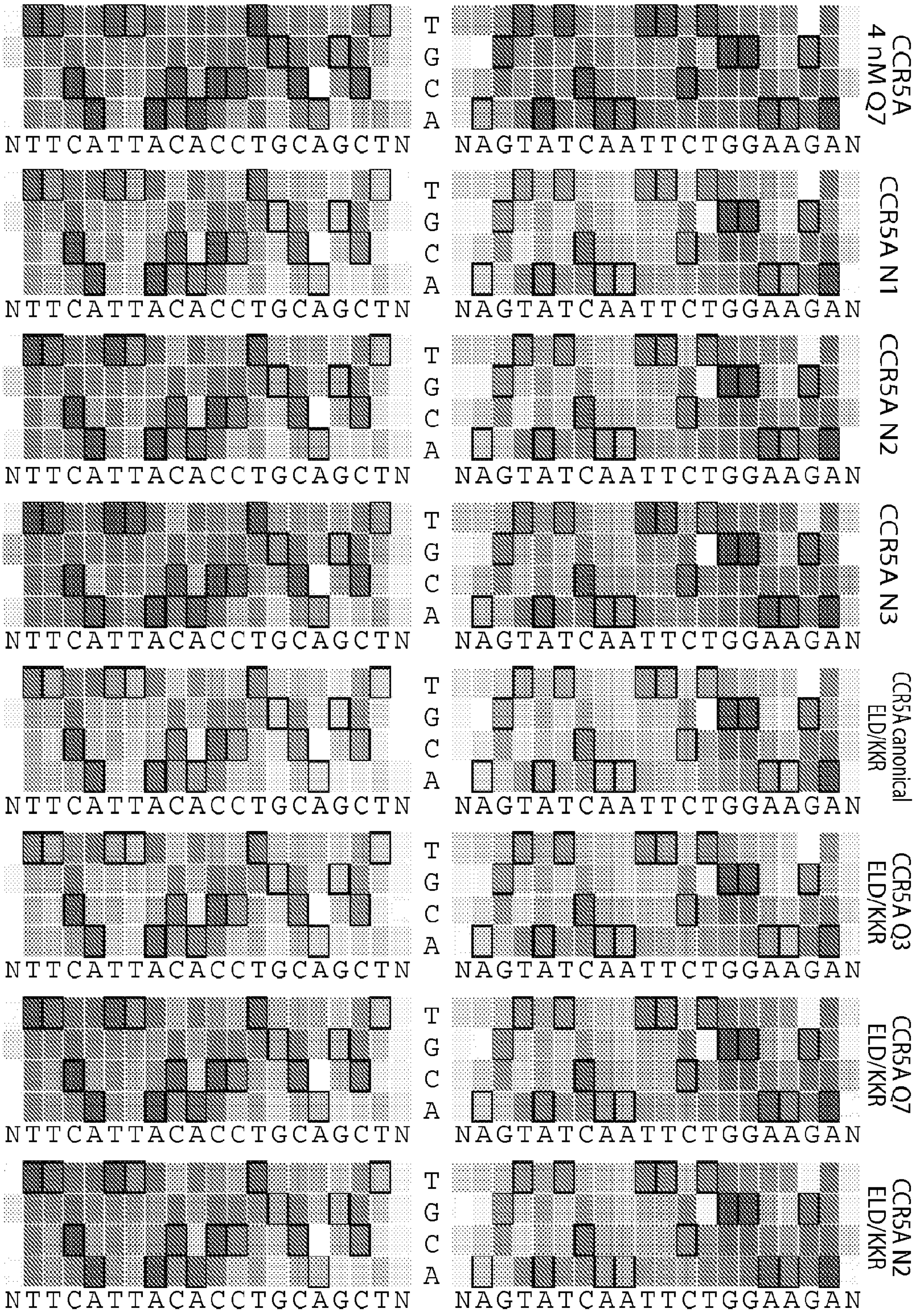

T
G
C
A
NTTCATTACACCTGCAGCTN
NAGTATCAATTCTGGAAGAN
CCR5A 4 nM Q7
CCR5A N1
CCR5A N2
CCR5A N3
CCR5A canonical ELD/KKR
CCR5A Q3 ELD/KKR
CCR5A Q7 ELD/KKR
CCR5A N2 ELD/KKR

Fig. 8B

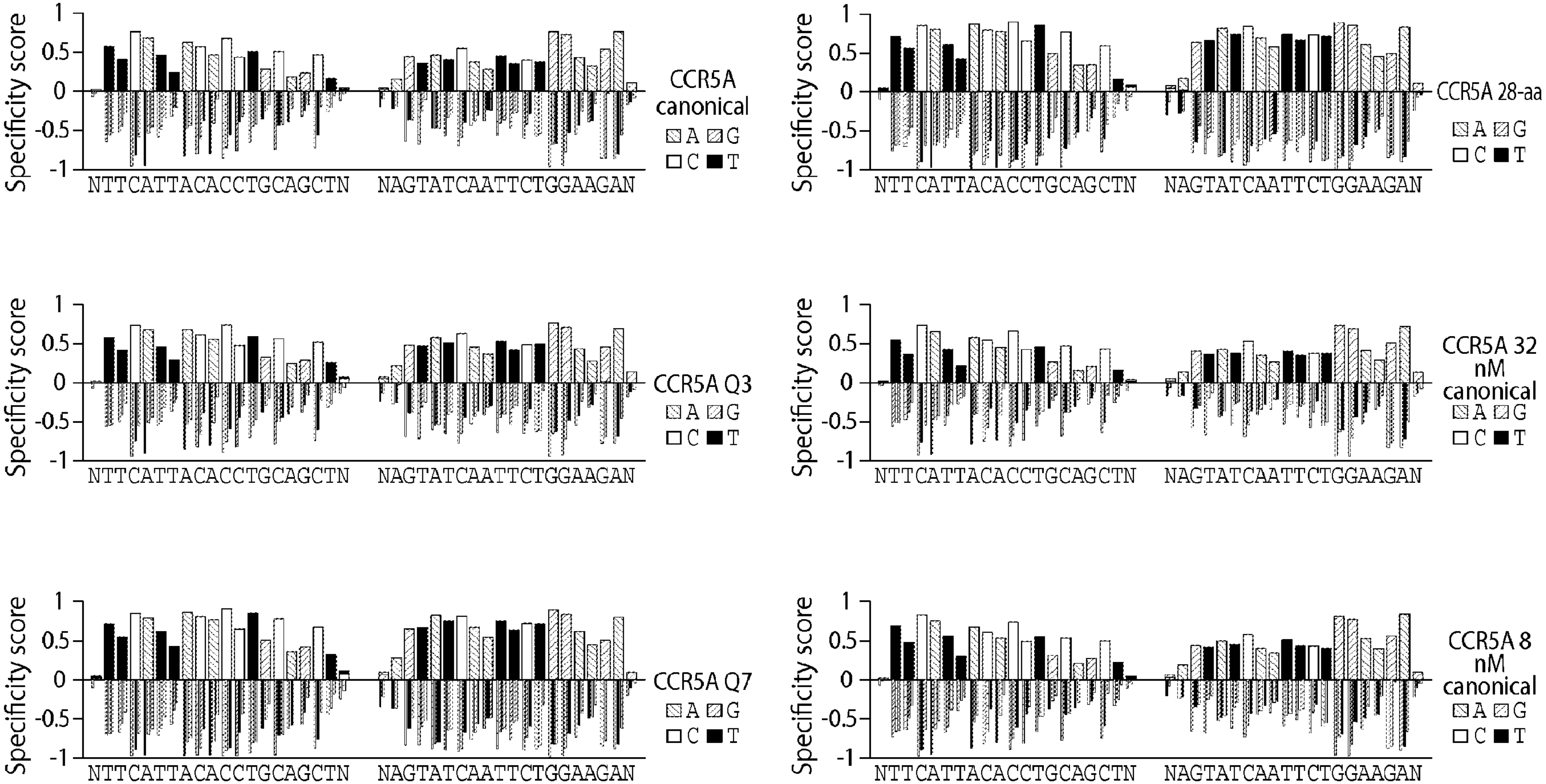

Specificity score
1
0.5
0
-0.5
-1
NTTCATTACACCTGCAGCTN
NAGTATCAATTCTGGAAGAN
CCR5A canonical
▨ A ▨ G
□ C ■ T
CCR5A 28-aa
CCR5A Q3
CCR5A 32 nM canonical
CCR5A Q7
CCR5A 8 nM canonical

Fig. 9A

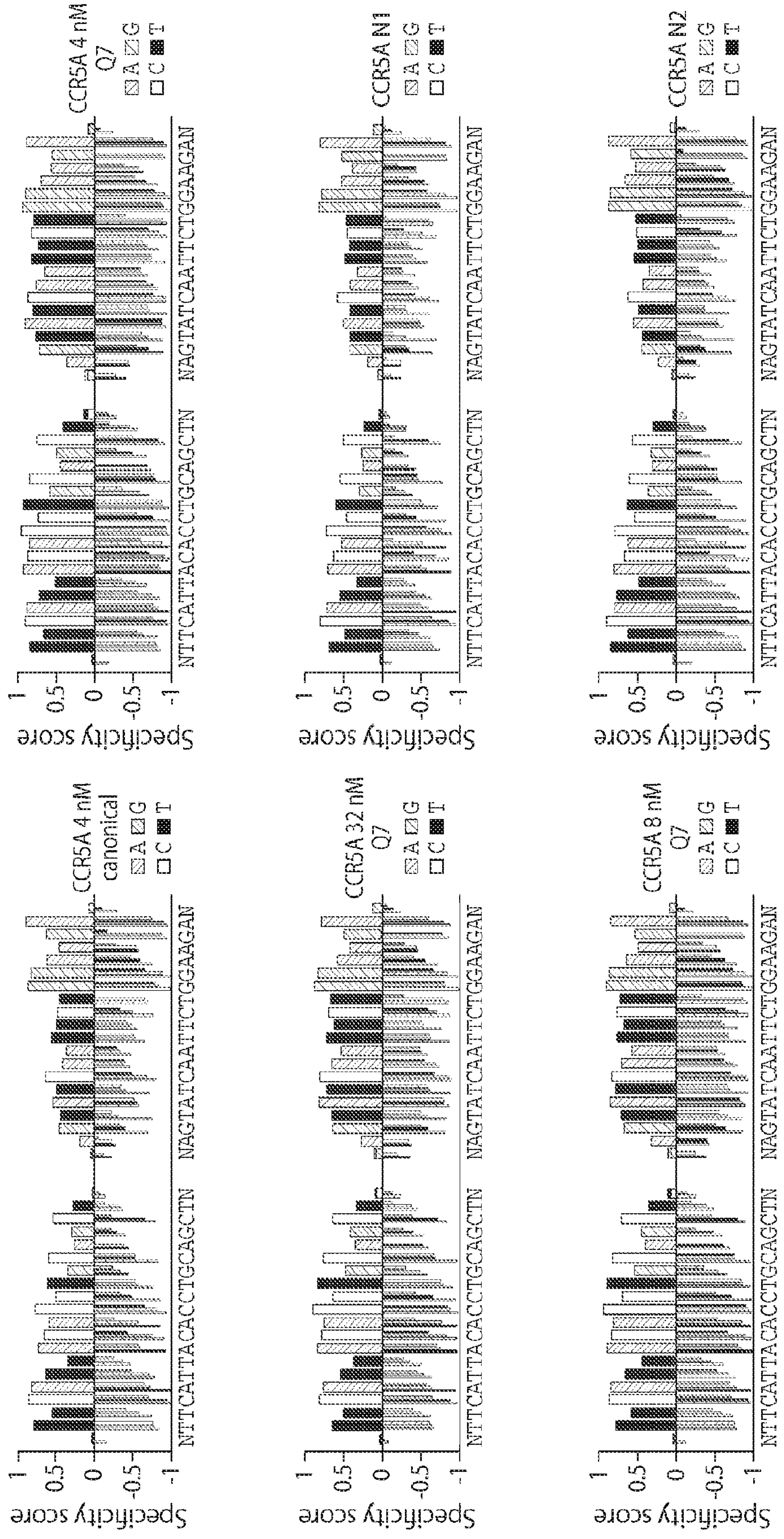

CCR5A 4 nM canonical
▨A ▨G
□C ■T
Specificity score
1
0.5
0
-0.5
-1
NTTCATTACACCTGCAGCTN
NAGTATCAATTCTGGAAGAN
CCR5A 32 nM Q7
▨A ▨G
□C ■T
Specificity score
NTTCATTACACCTGCAGCTN
NAGTATCAATTCTGGAAGAN
CCR5A 8 nM Q7
▨A ▨G
□C ■T
Specificity score
NTTCATTACACCTGCAGCTN
NAGTATCAATTCTGGAAGAN
CCR5A 4 nM Q7
▨A ▨G
□C ■T
Specificity score
NTTCATTACACCTGCAGCTN
NAGTATCAATTCTGGAAGAN
CCR5A N1
▨A ▨G
□C ■T
Specificity score
NTTCATTACACCTGCAGCTN
NAGTATCAATTCTGGAAGAN
CCR5A N2
▨A ▨G
□C ■T
Specificity score
NTTCATTACACCTGCAGCTN
NAGTATCAATTCTGGAAGAN

Fig. 9B

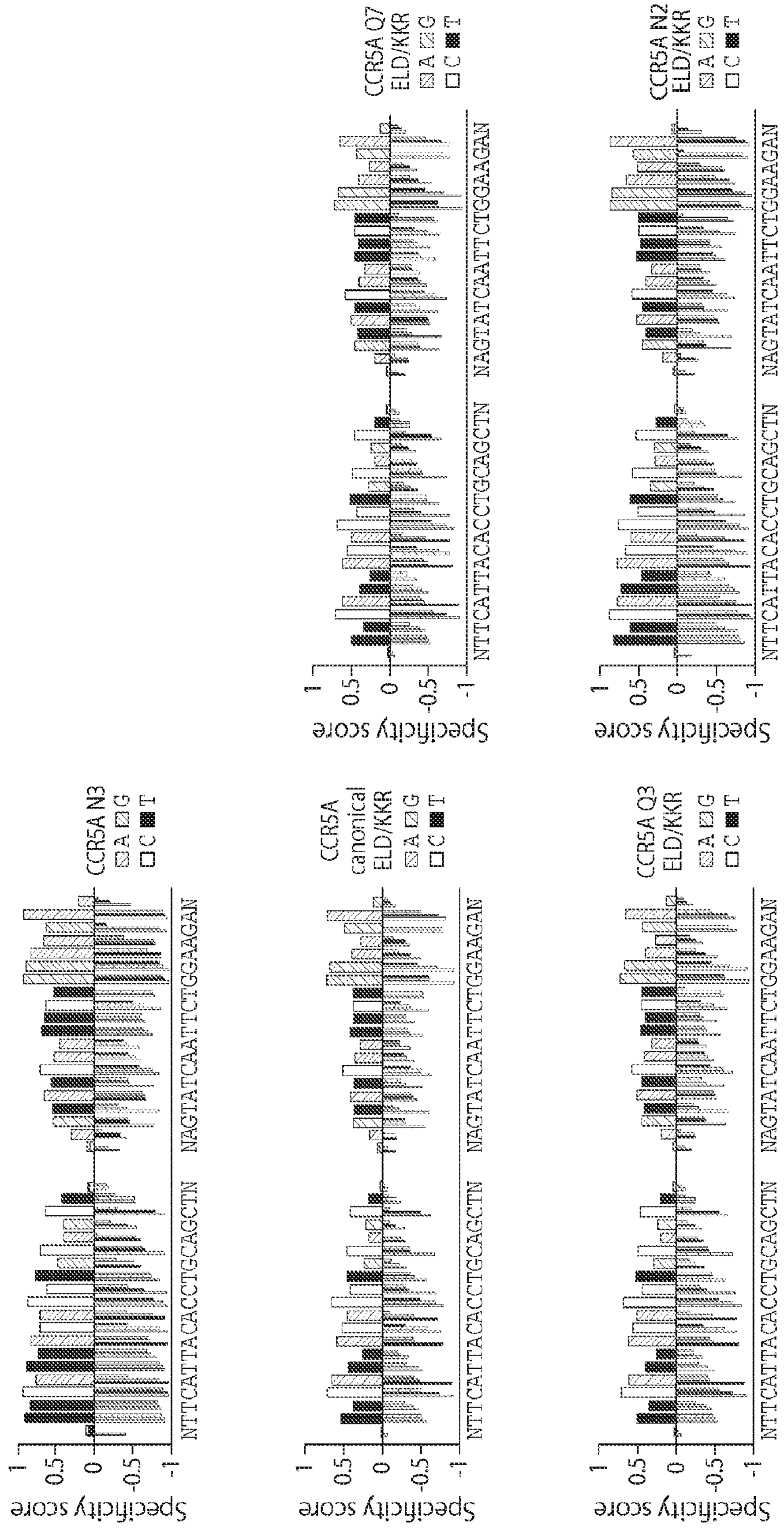

CCR5A N3
▨ A ▨ G □ C ■ T
CCR5A canonical ELD/KKR
▨ A ▨ G □ C ■ T
CCR5A Q3 ELD/KKR
▨ A ▨ G □ C ■ T
CCR5A Q7 ELD/KKR
▨ A ▨ G □ C ■ T
CCR5A N2 ELD/KKR
▨ A ▨ G □ C ■ T
Specificity score
1
0.5
0
-0.5
-1
NTTCATTACACCTGCAGCTN
NAGTATCAATTCTGGAAGAN

Fig. 9C

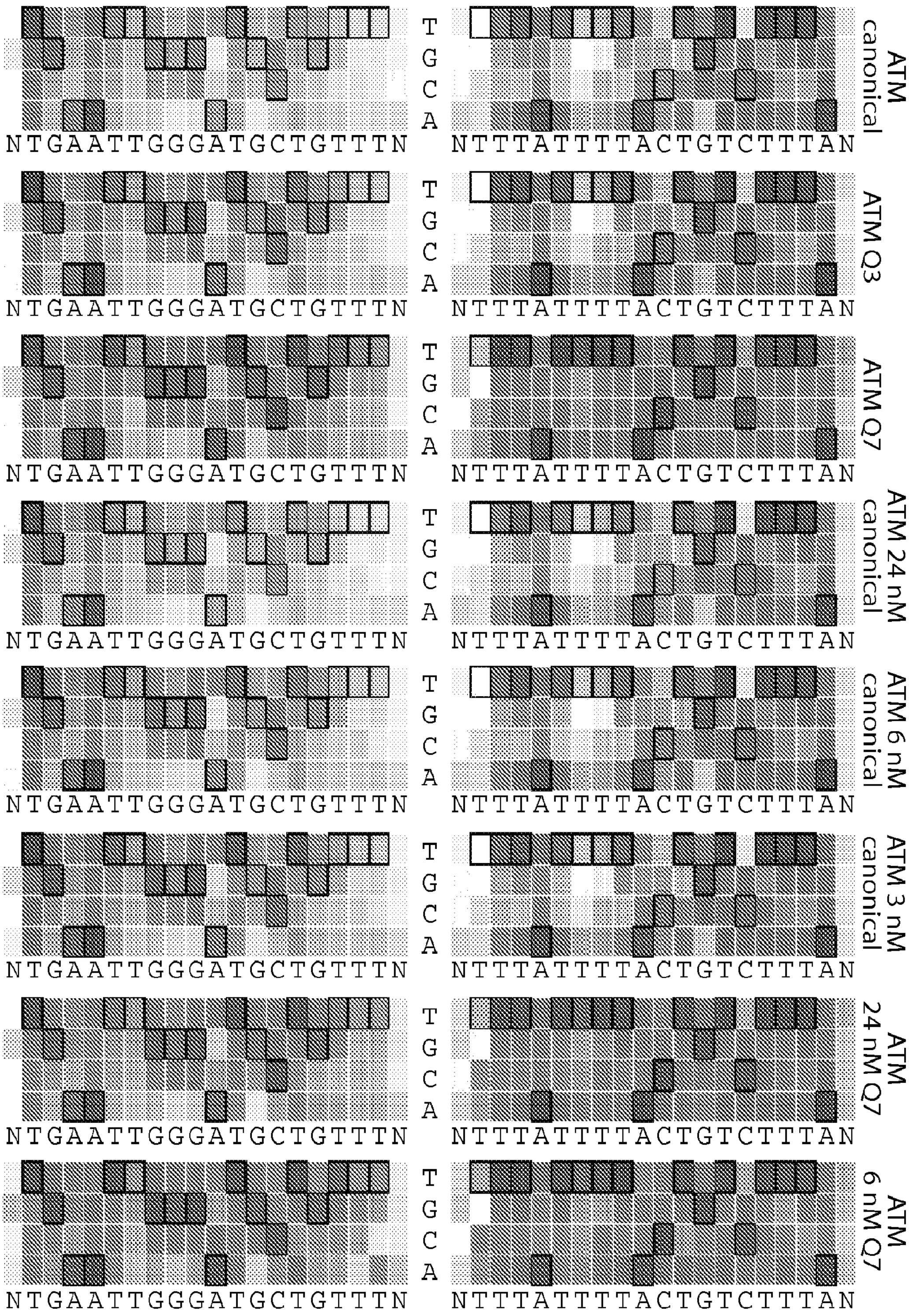

T
G
C
A
NTGAATTGGGATGCTGTTTN
NTTTATTTTACTGTCTTTAN
ATM canonical
ATM Q3
ATM Q7
ATM 24 nM canonical
ATM 6 nM canonical
ATM 3 nM canonical
ATM 24 nM Q7
ATM 6 nM Q7

Fig. 10A

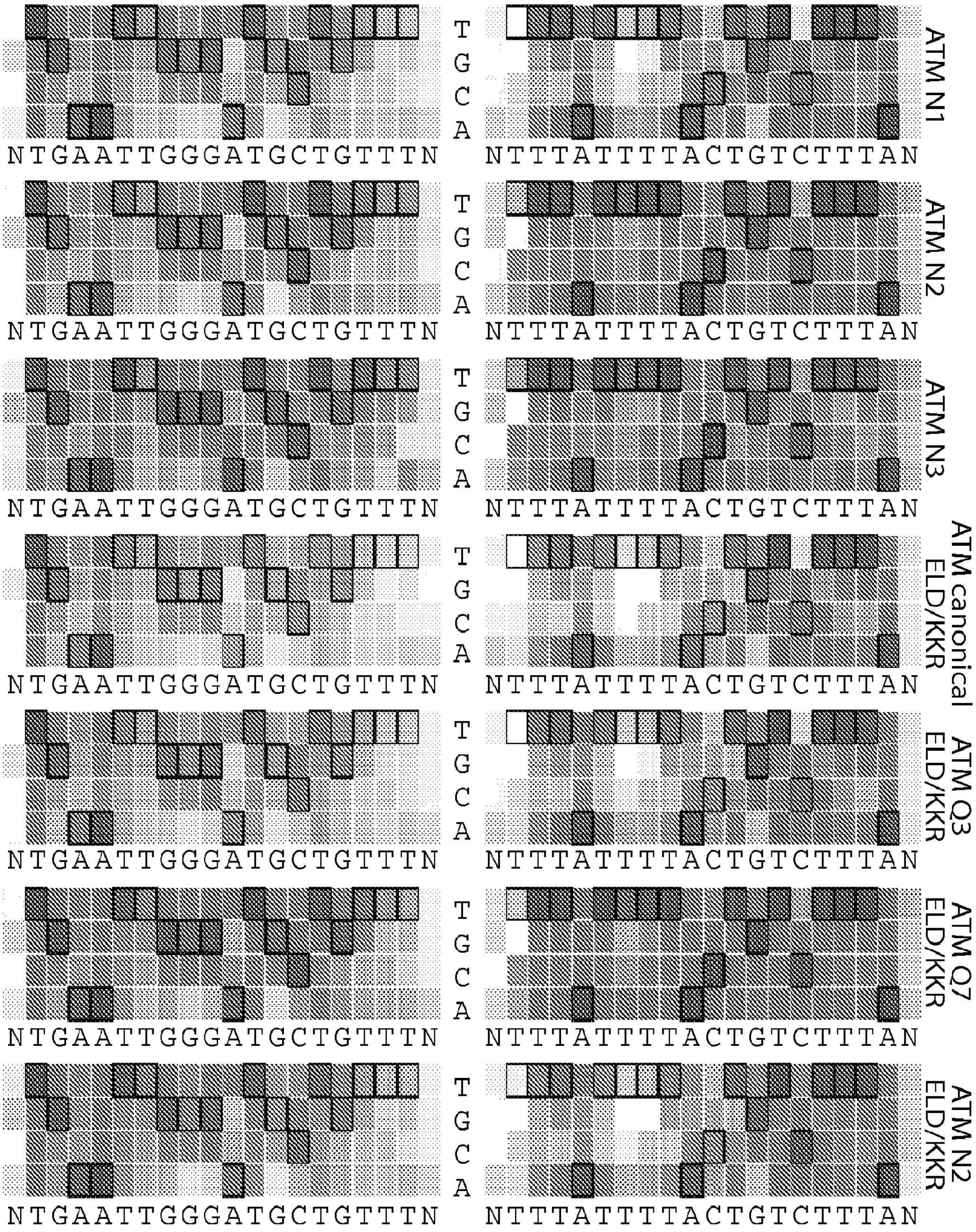

T
G
C
A
NTGAATTGGGATGCTGTTTN
NTTTATTTTACTGTCTTTAN
ATM N1
ATM N2
ATM N3
ATM canonical ELD/KKR
ATM Q3 ELD/KKR
ATM Q7 ELD/KKR
ATM N2 ELD/KKR

Fig. 10B

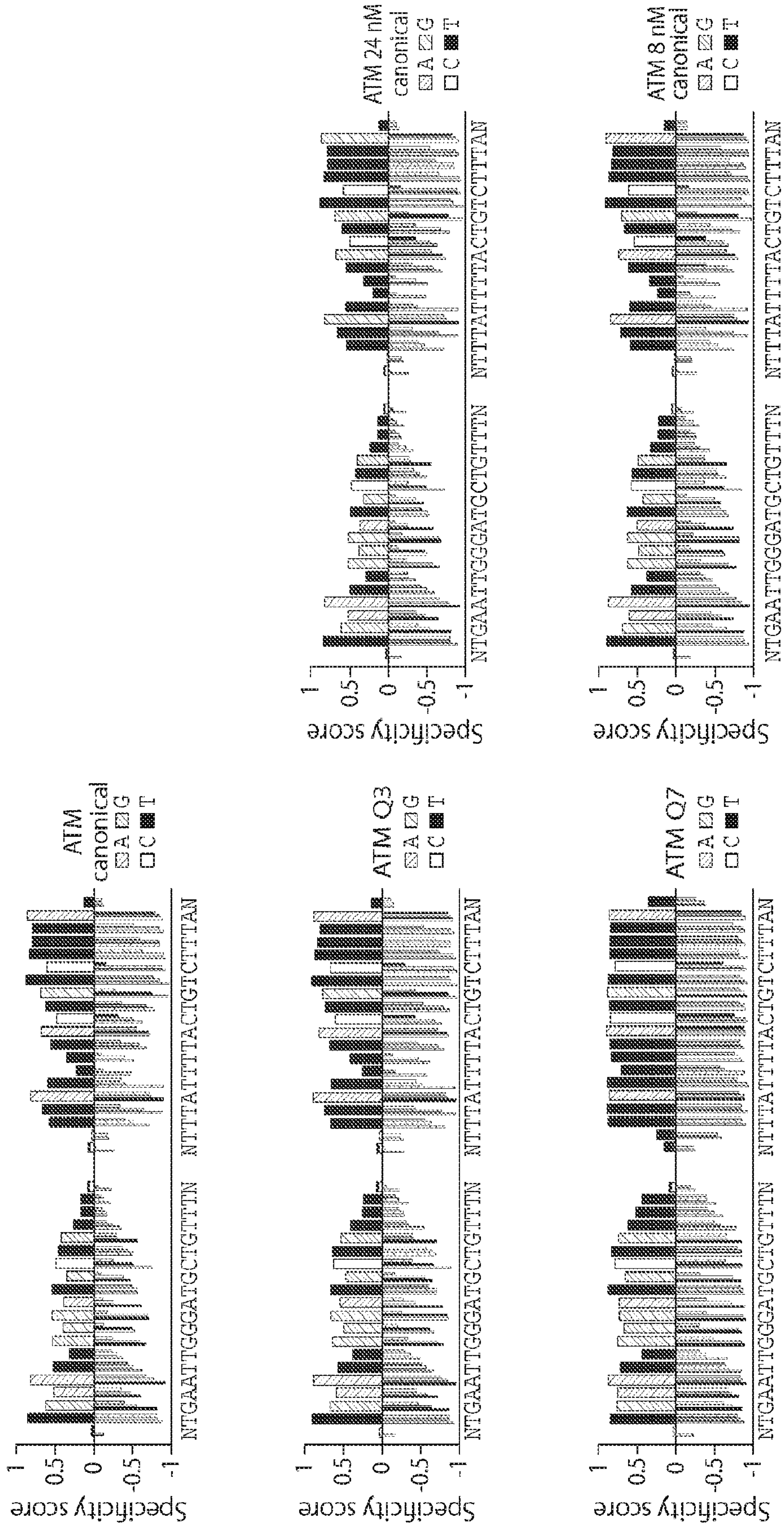

ATM canonical
▨A ▨G □C ■T
Specificity score
1
0.5
0
-0.5
-1
NTGAATTGGGATGCTGTTTN
NTTTATTTTACTGTCTTTAN
ATM Q3
▨A ▨G □C ■T
ATM Q7
▨A ▨G □C ■T
ATM 24 nM canonical
▨A ▨G □C ■T
ATM 8 nM canonical
▨A ▨G □C ■T

Fig. 11A

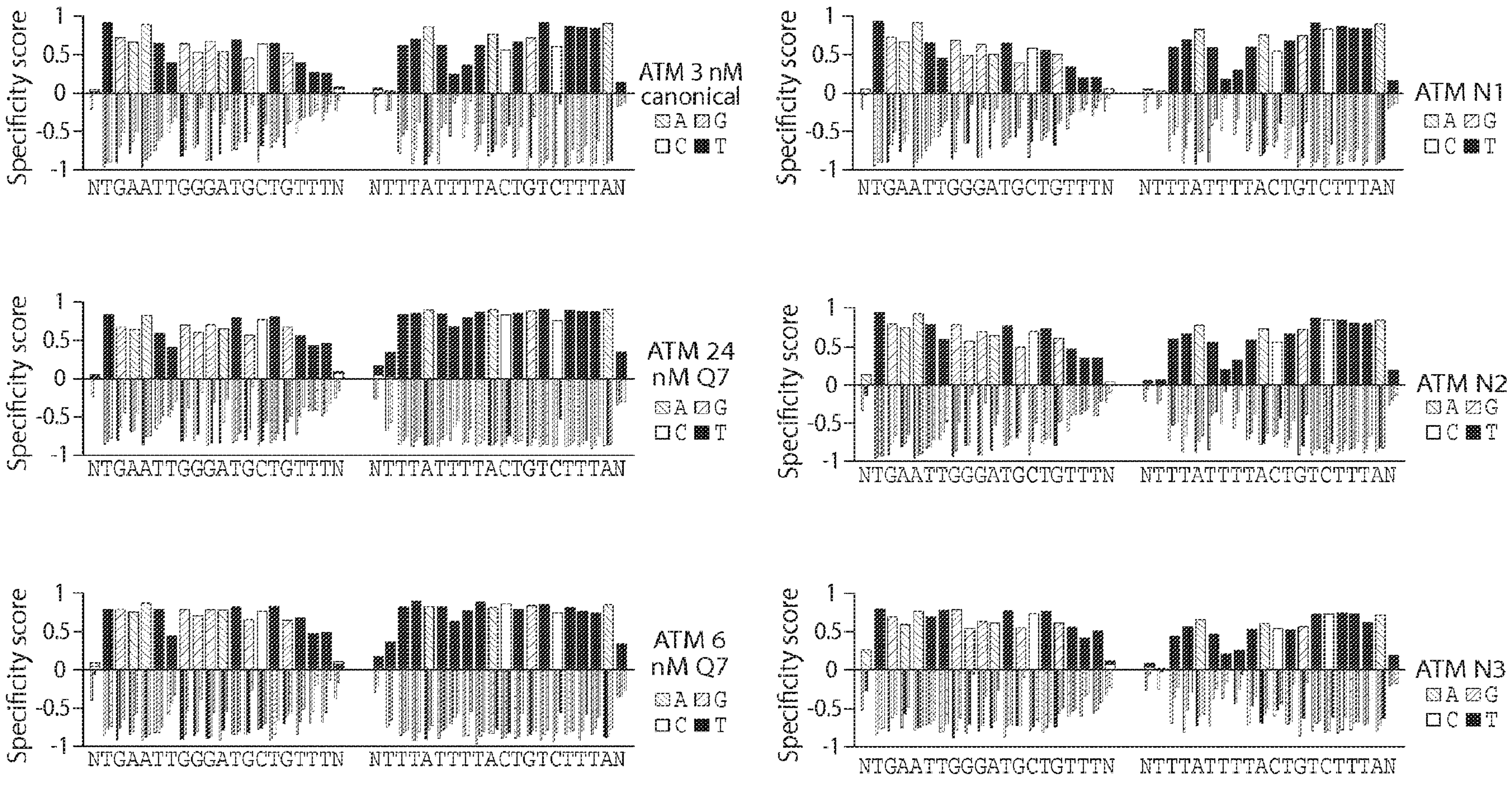

Specificity score
1
0.5
0
-0.5
-1
NTGAATTGGGATGCTGTTTN
NTTTATTTTACTGTCTTTAN
ATM 3 nM canonical
▨A ▨G
□C ■T
ATM N1
ATM 24 nM Q7
ATM N2
ATM 6 nM Q7
ATM N3

Fig. 11B

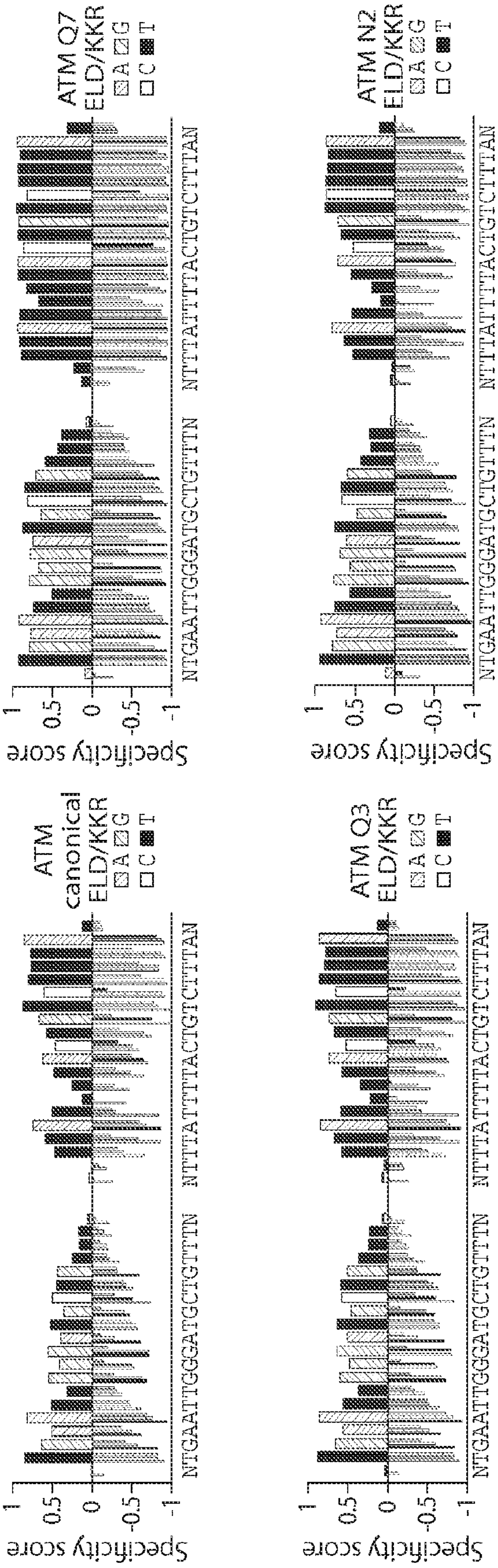

Specificity score
1
0.5
0
-0.5
-1
NTGAATTGGGATGCTGTTTN
NTTTATTTTACTGTCTTTAN
ATM canonical ELD/KKR
▨A ▨G
□C ■T
ATM Q3 ELD/KKR
ATM Q7 ELD/KKR
ATM N2 ELD/KKR

Fig. 11C

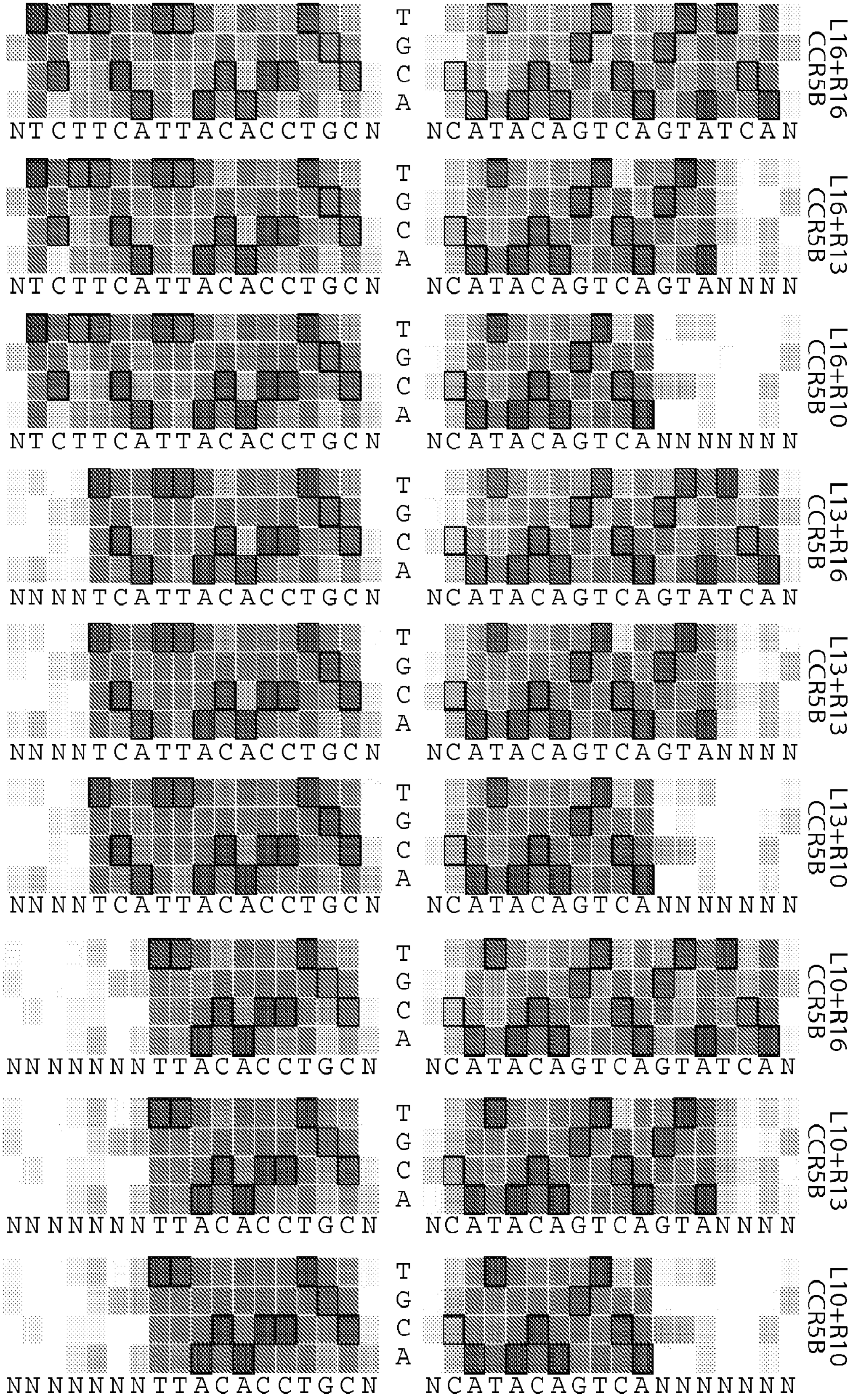

T
G
C
A
NTCTTCATTACACCTGCN
NCATACAGTCAGTATCAN
L16+R16
CCR5B
NTCTTCATTACACCTGCN
NCATACAGTCAGTANNNN
L16+R13
CCR5B
NTCTTCATTACACCTGCN
NCATACAGTCANNNNNNN
L16+R10
CCR5B
NNNNTCATTACACCTGCN
NCATACAGTCAGTATCAN
L13+R16
CCR5B
NNNNTCATTACACCTGCN
NCATACAGTCAGTANNNN
L13+R13
CCR5B
NNNNTCATTACACCTGCN
NCATACAGTCANNNNNNN
L13+R10
CCR5B
NNNNNNNTTACACCTGCN
NCATACAGTCAGTATCAN
L10+R16
CCR5B
NNNNNNNTTACACCTGCN
NCATACAGTCAGTANNNN
L10+R13
CCR5B
NNNNNNNTTACACCTGCN
NCATACAGTCANNNNNNN
L10+R10
CCR5B

Fig. 12

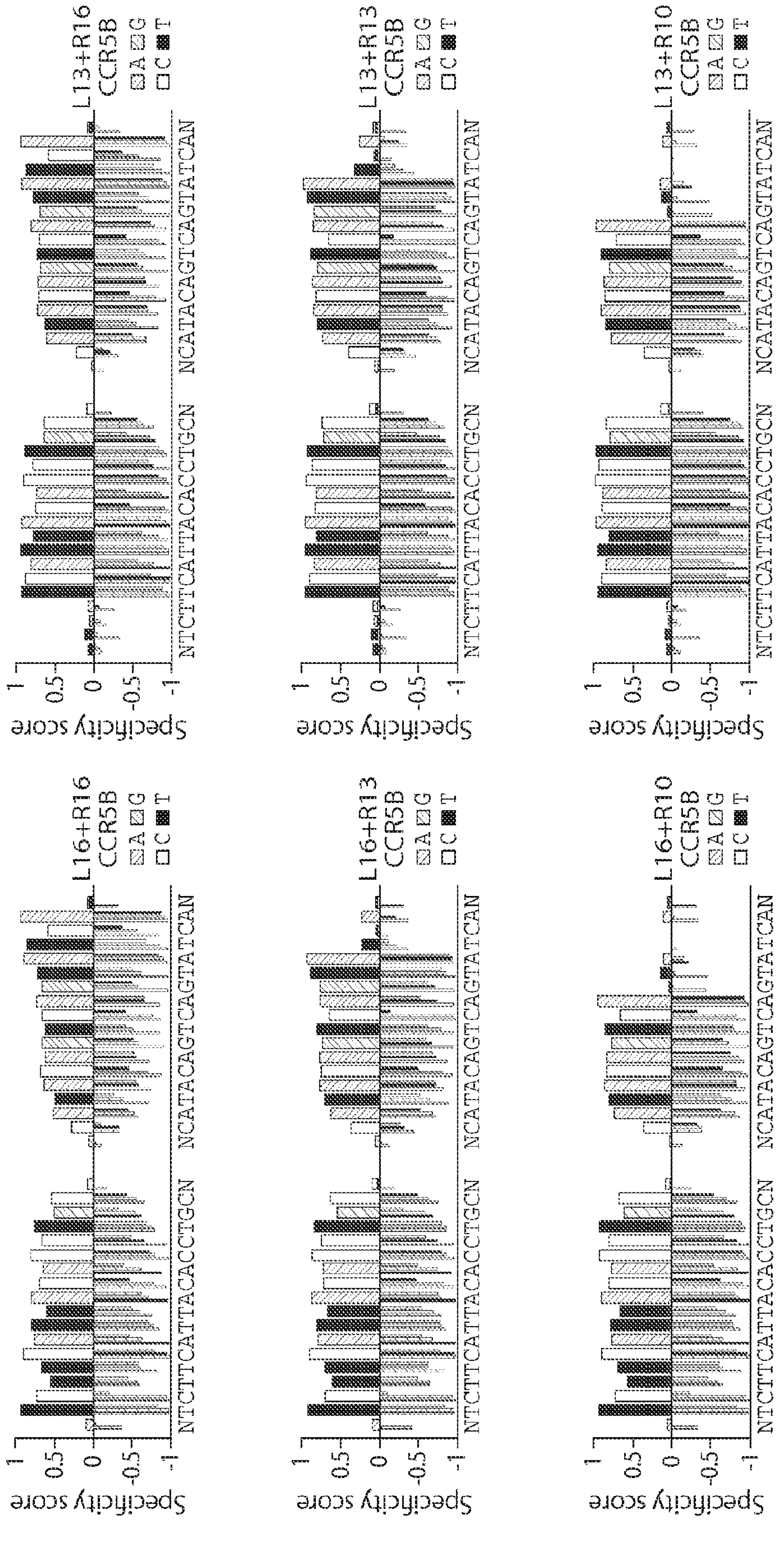

L16+R16
CCR5B
▨A ▨G
□C ■T
L16+R13
CCR5B
▨A ▨G
□C ■T
L16+R10
CCR5B
▨A ▨G
□C ■T
L13+R16
CCR5B
▨A ▨G
□C ■T
L13+R13
CCR5B
▨A ▨G
□C ■T
L13+R10
CCR5B
▨A ▨G
□C ■T
Specificity score
1
0.5
0
-0.5
-1
NTCTTCATTACACCTGCN
NCATACAGTCAGTATCAN

Fig. 13A

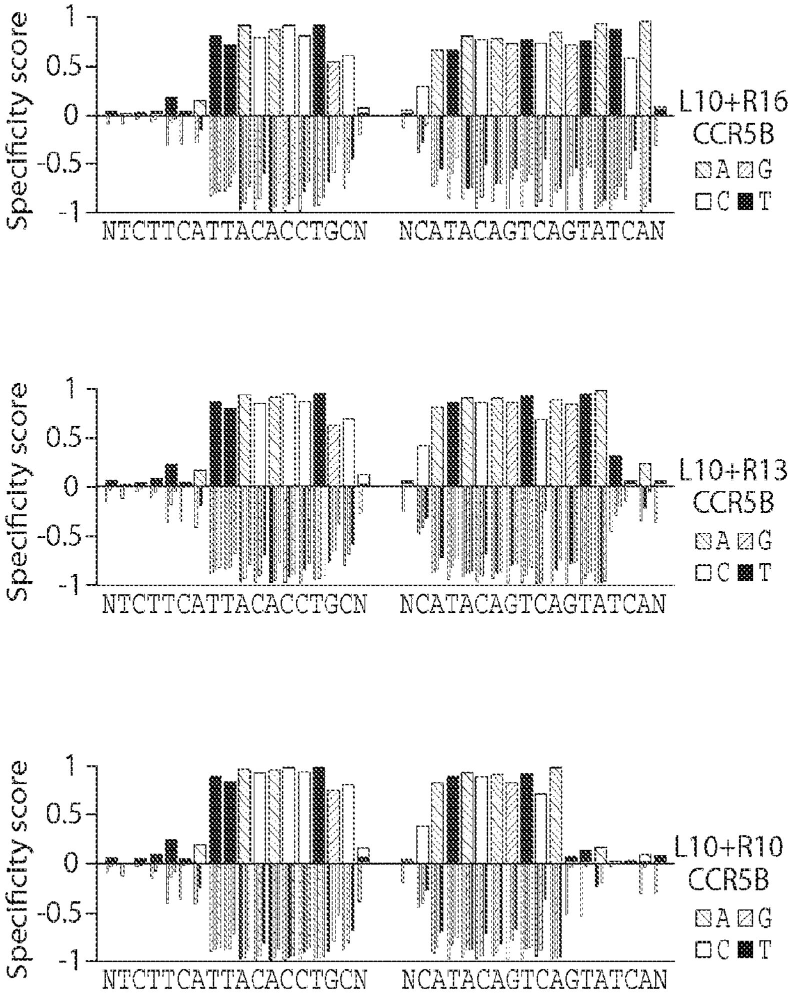

Specificity score
1
0.5
0
-0.5
-1
NTCTTCATTACACCTGCN
NCATACAGTCAGTATCAN
L10+R16
CCR5B
▨A ▨G
□C ■T
Specificity score
1
0.5
0
-0.5
-1
NTCTTCATTACACCTGCN
NCATACAGTCAGTATCAN
L10+R13
CCR5B
▨A ▨G
□C ■T
Specificity score
1
0.5
0
-0.5
-1
NTCTTCATTACACCTGCN
NCATACAGTCAGTATCAN
L10+R10
CCR5B
▨A ▨G
□C ■T

Fig. 13B

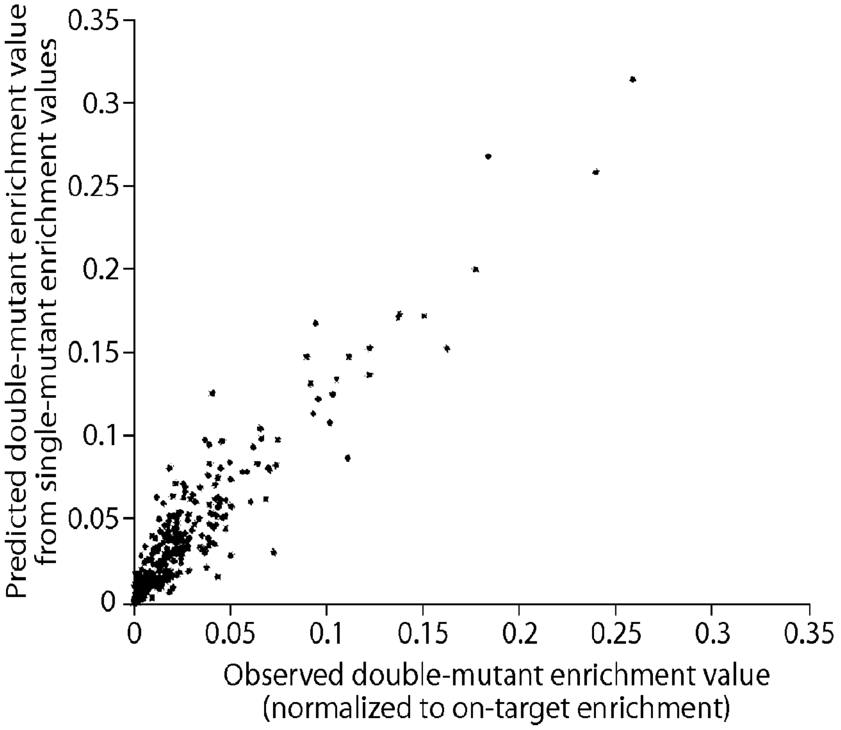


Fig. 14A

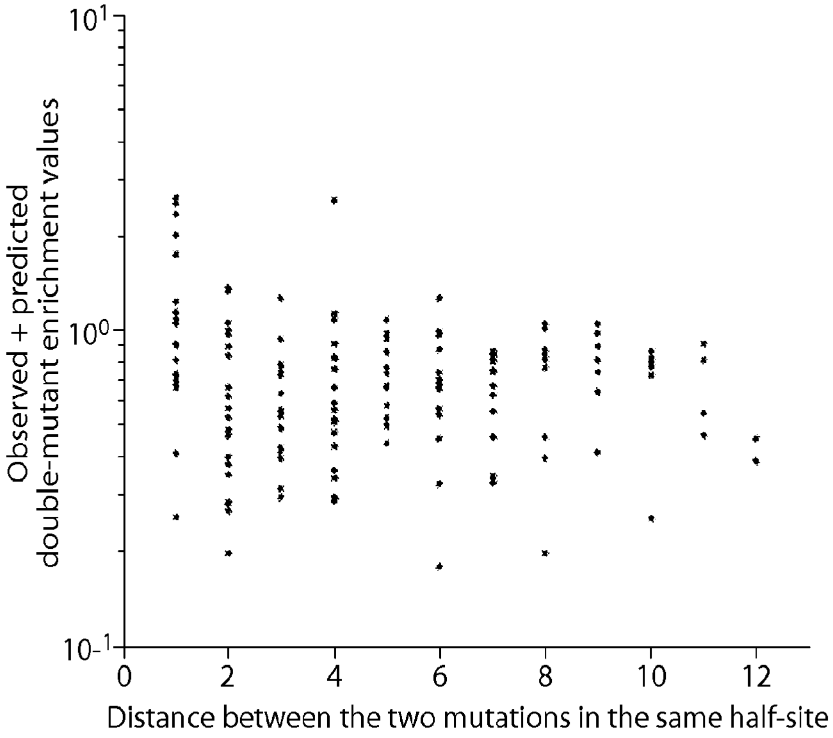


Fig. 14B

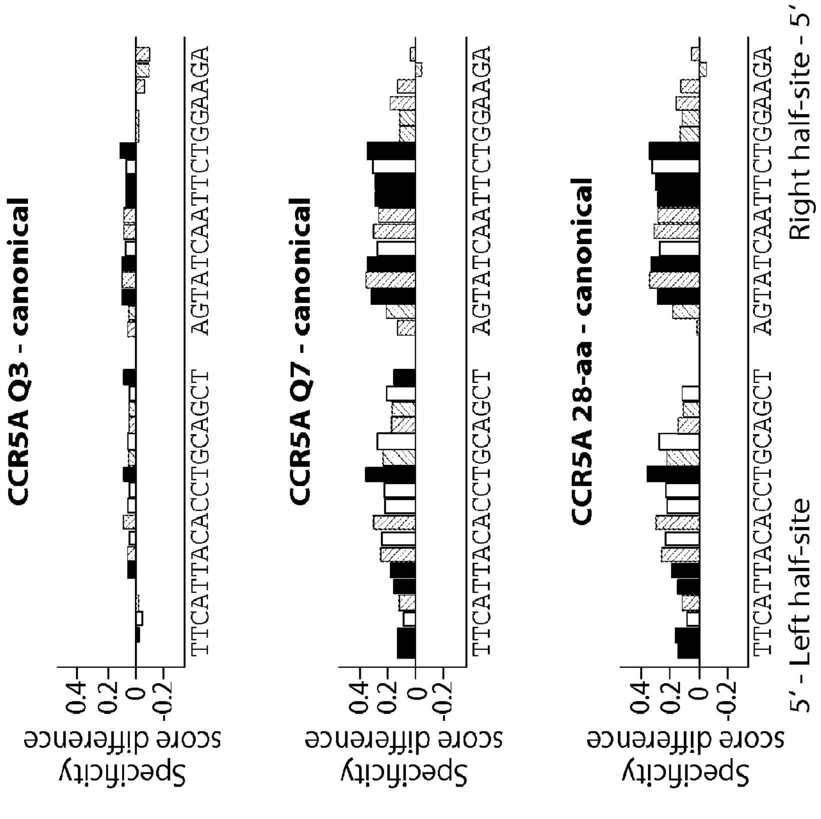


Fig. 15A

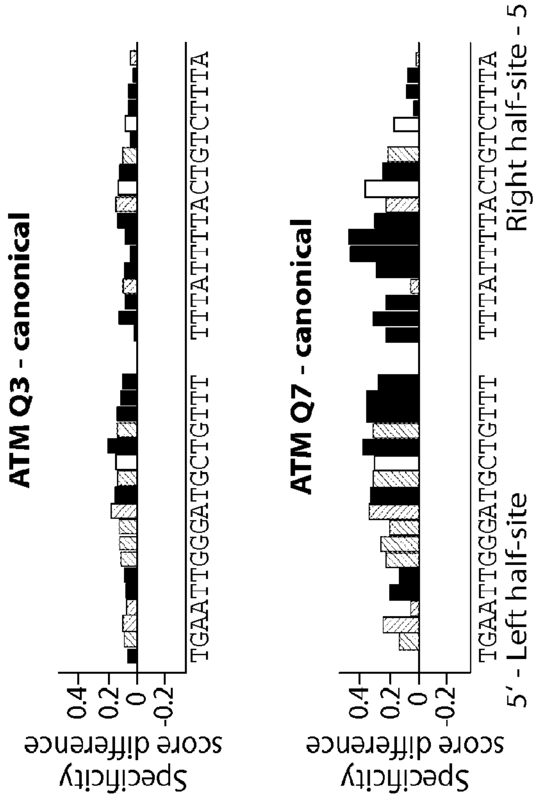


Fig. 15B

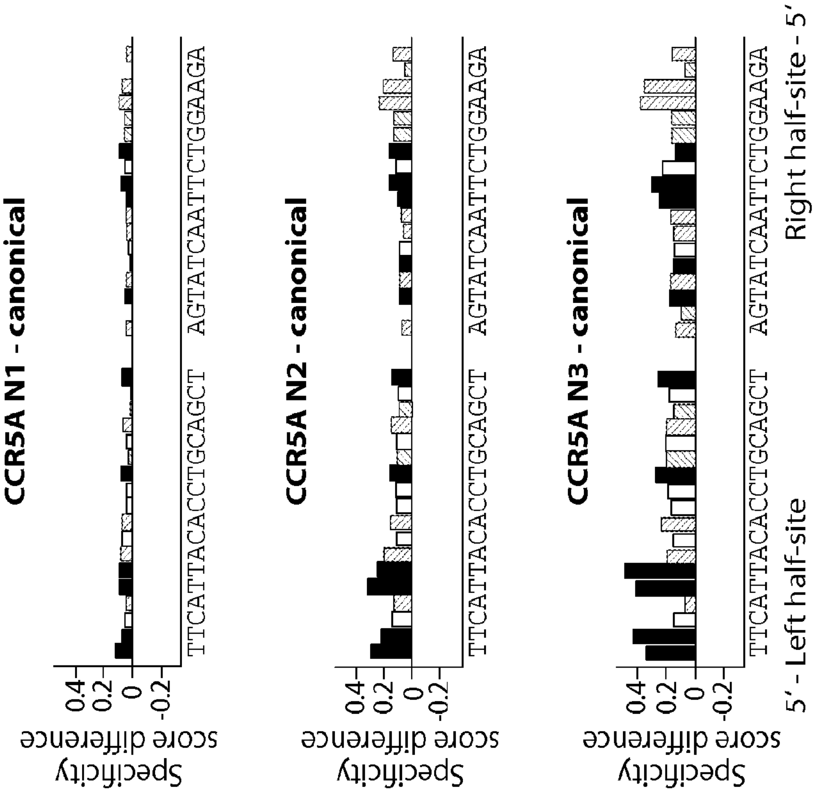


Fig. 15C

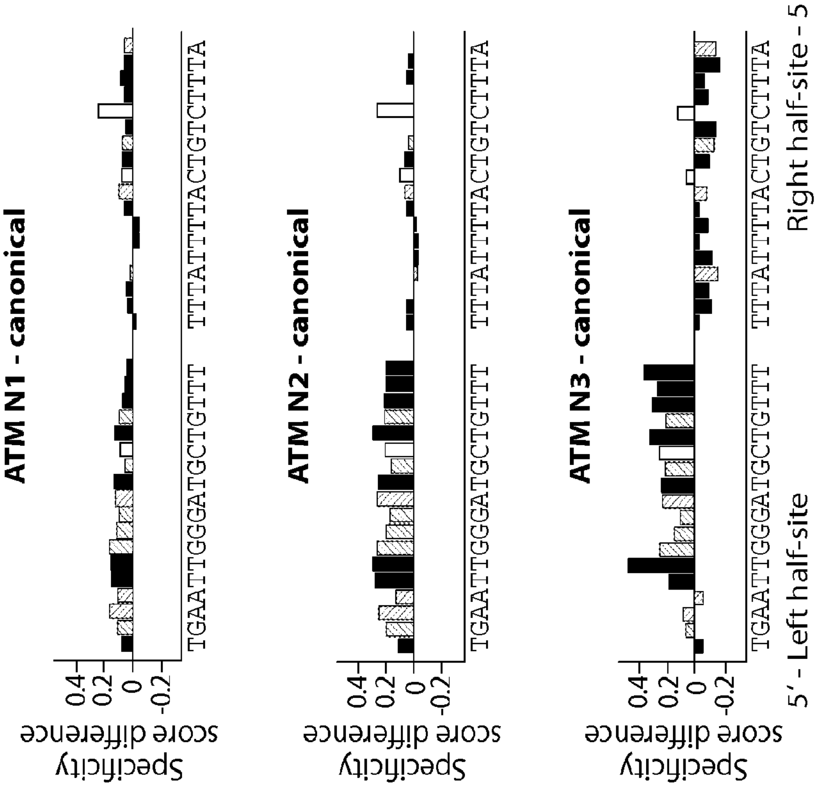


Fig. 15D

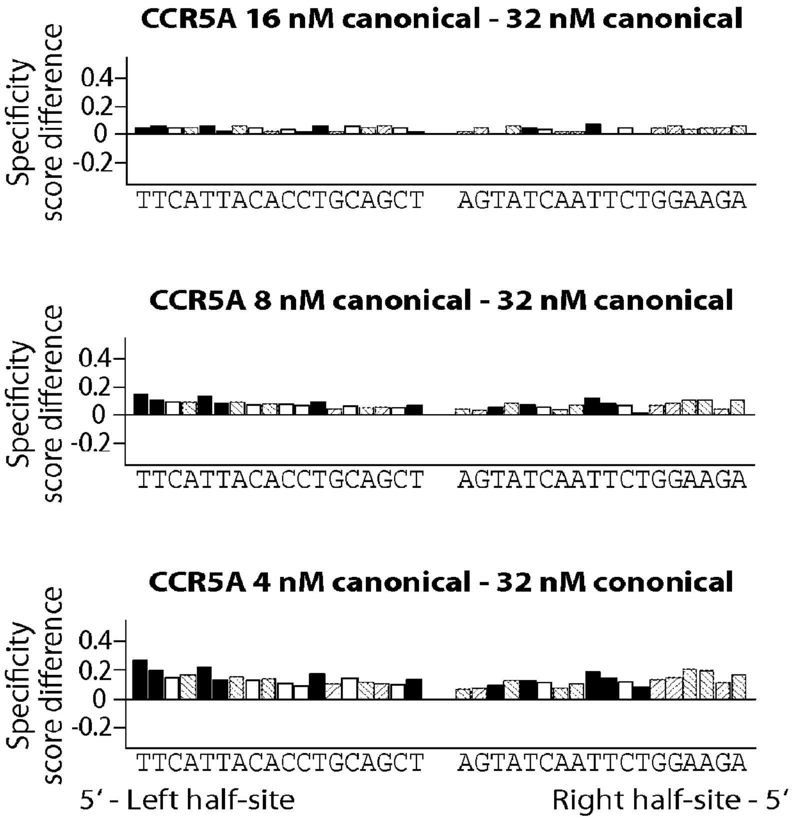


Fig. 15E

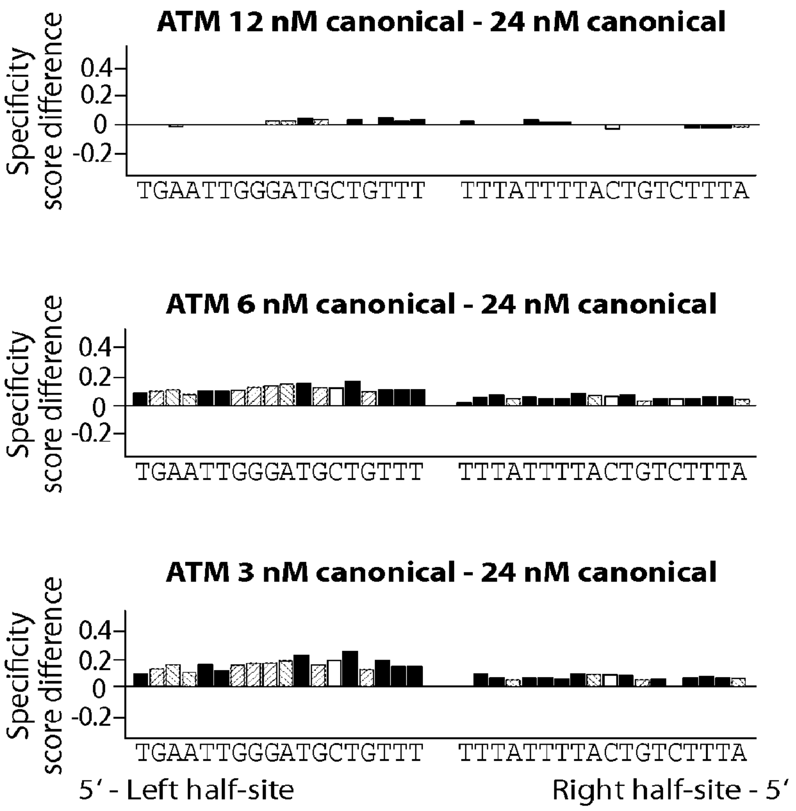


Fig. 15F

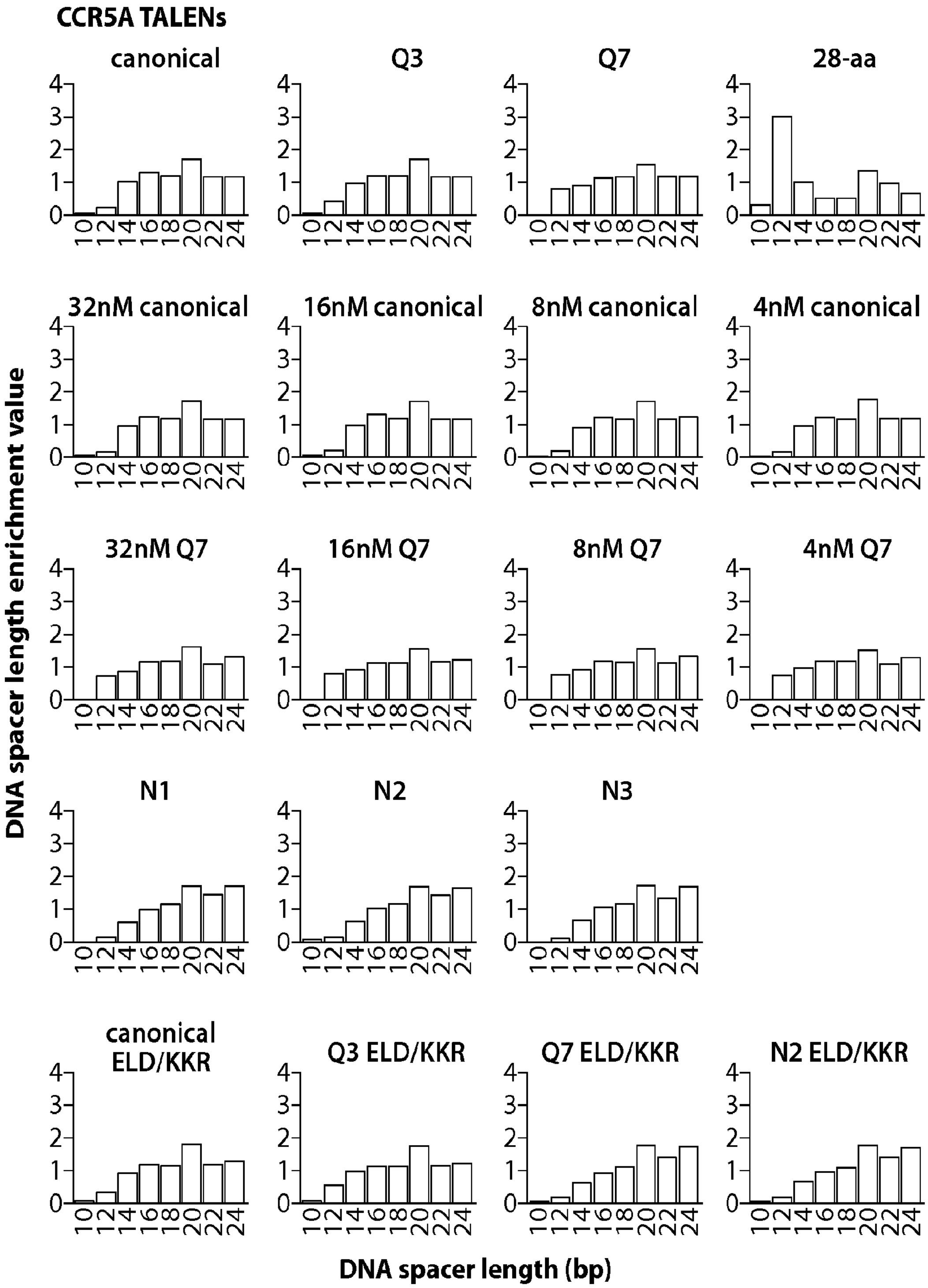

CCR5A TALENs
canonical
Q3
Q7
28-aa
32nM canonical
16nM canonical
8nM canonical
4nM canonical
32nM Q7
16nM Q7
8nM Q7
4nM Q7
N1
N2
N3
canonical ELD/KKR
Q3 ELD/KKR
Q7 ELD/KKR
N2 ELD/KKR
DNA spacer length enrichment value
0
1
2
3
4
10
12
14
16
18
20
22
24
DNA spacer length (bp)

Fig. 16A

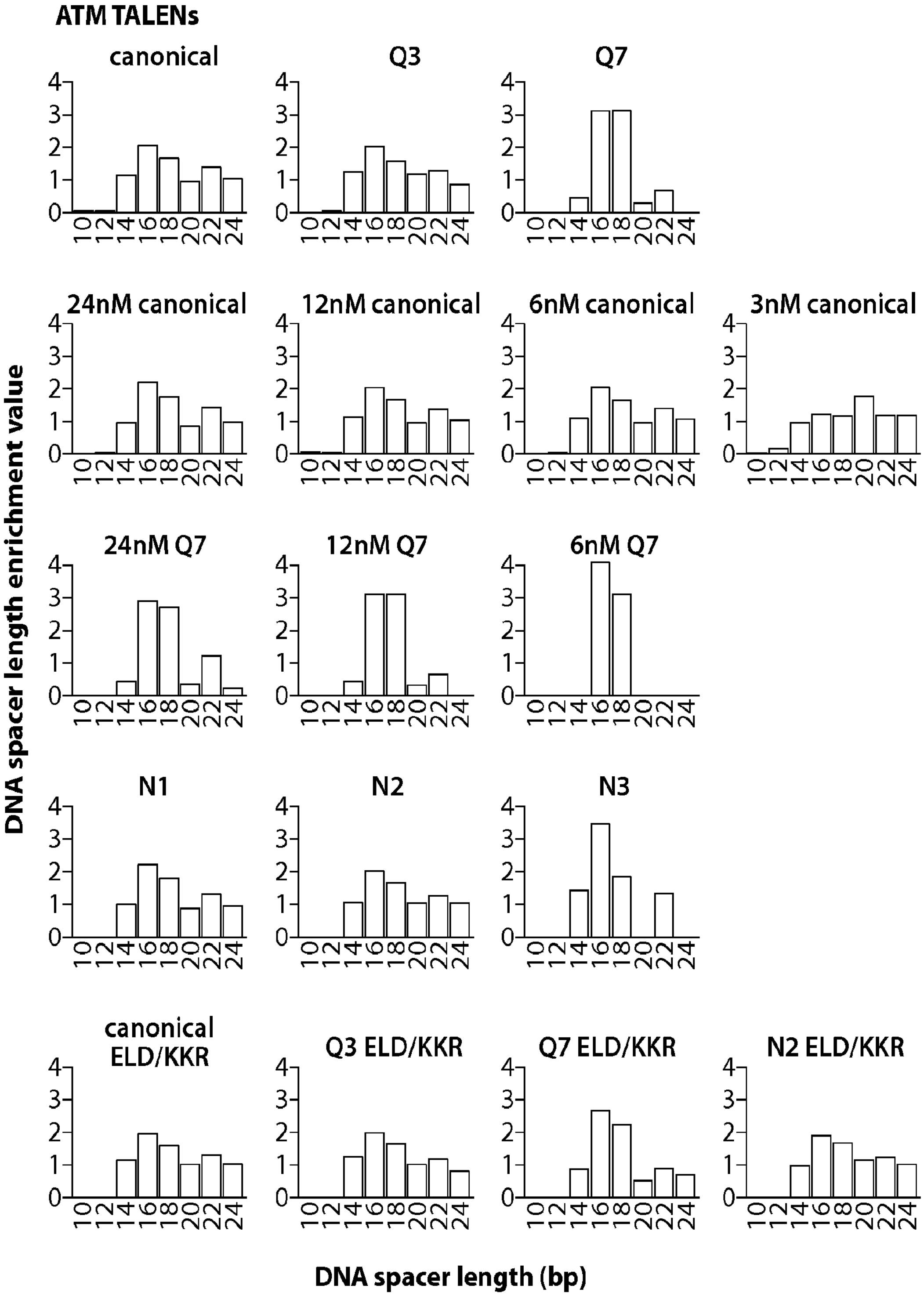

ATM TALENs
canonical
Q3
Q7
24nM canonical
12nM canonical
6nM canonical
3nM canonical
24nM Q7
12nM Q7
6nM Q7
N1
N2
N3
canonical ELD/KKR
Q3 ELD/KKR
Q7 ELD/KKR
N2 ELD/KKR
0
1
2
3
4
10
12
14
16
18
20
22
24
DNA spacer length enrichment value
DNA spacer length (bp)

Fig. 16B

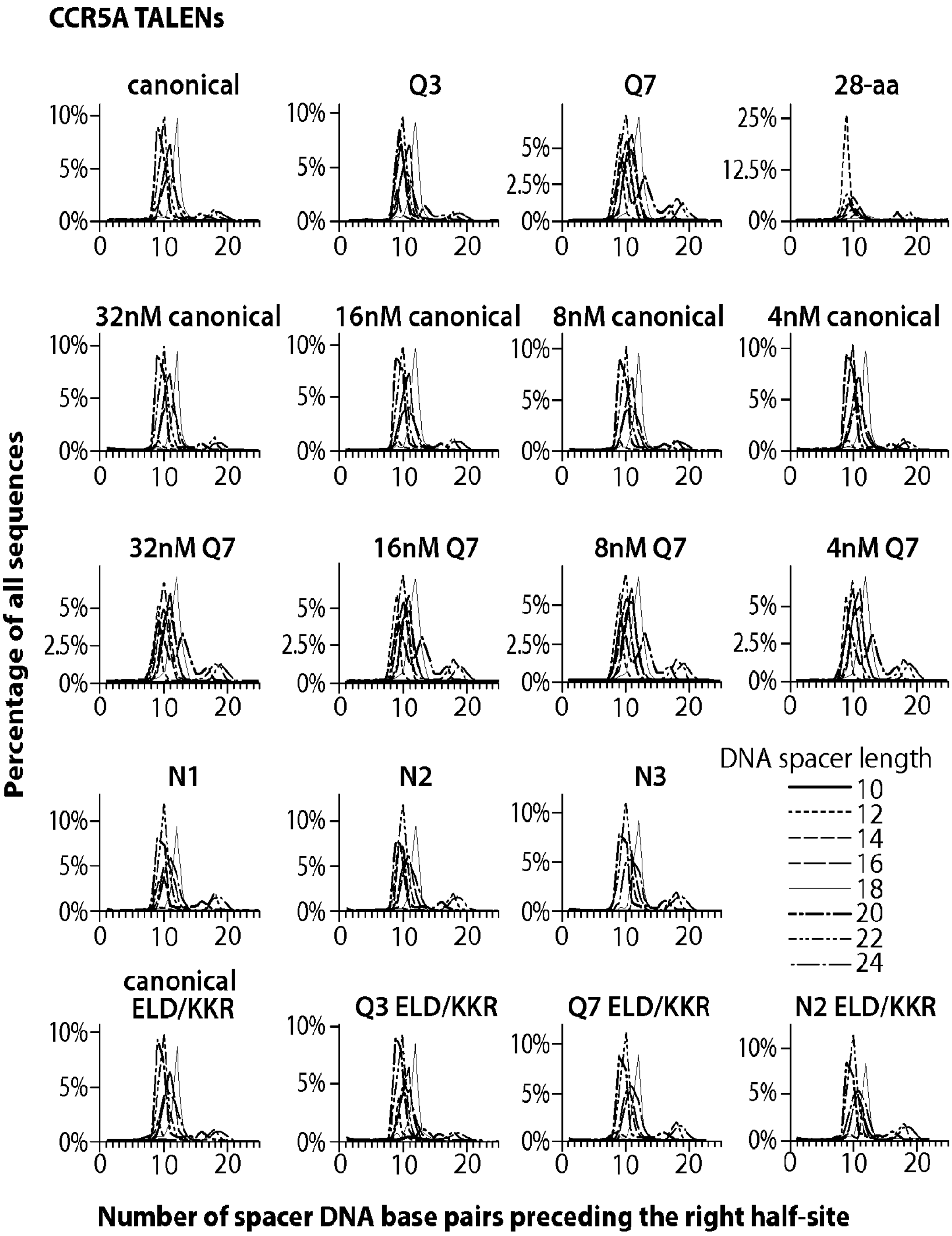

CCR5A TALENs
canonical
Q3
Q7
28-aa
32nM canonical
16nM canonical
8nM canonical
4nM canonical
32nM Q7
16nM Q7
8nM Q7
4nM Q7
N1
N2
N3
canonical ELD/KKR
Q3 ELD/KKR
Q7 ELD/KKR
N2 ELD/KKR
10%
5%
2.5%
0%
25%
12.5%
0
10
20
Percentage of all sequences
Number of spacer DNA base pairs preceding the right half-site
DNA spacer length
10
12
14
16
18
20
22
24

Fig. 17A

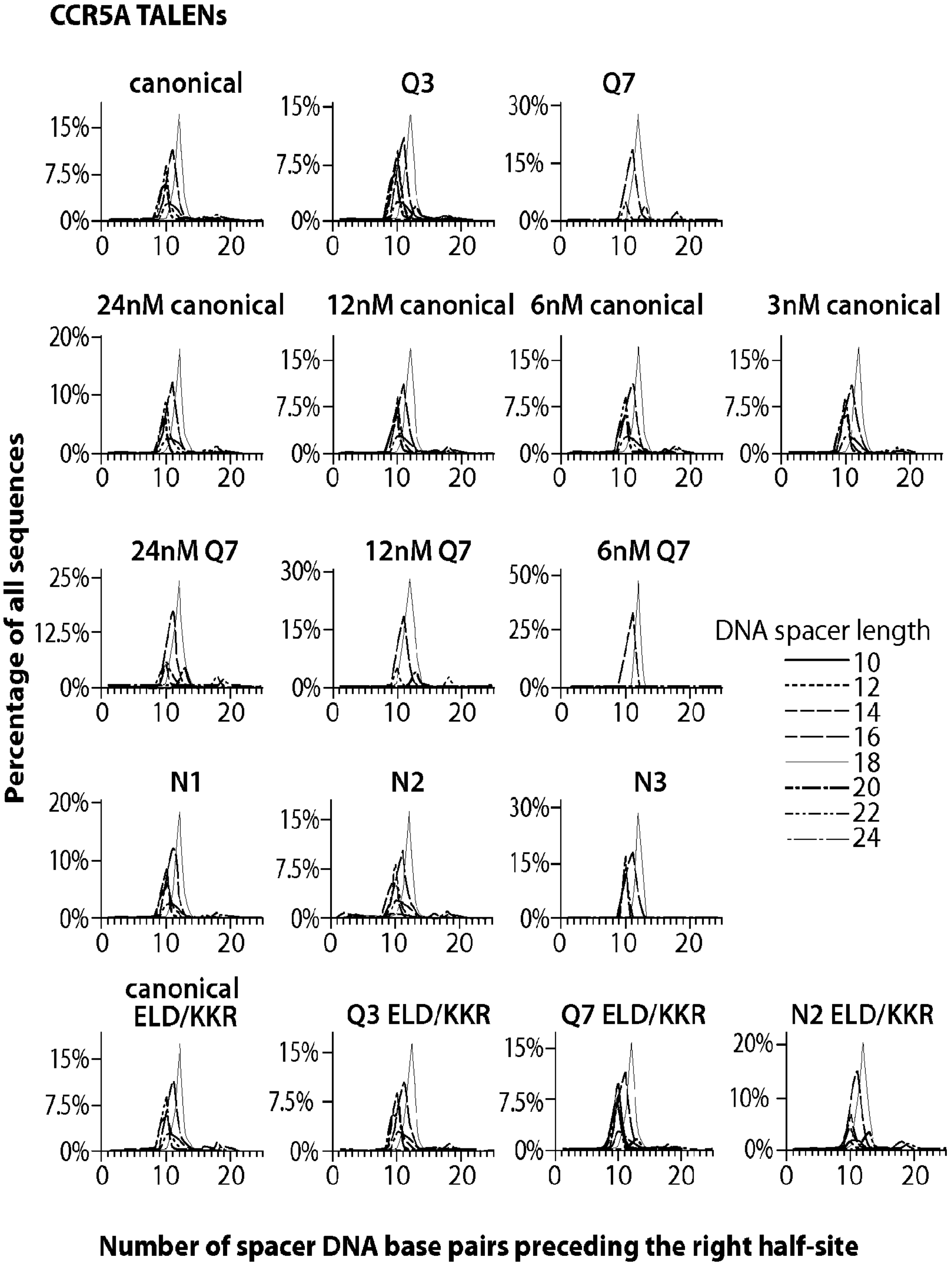
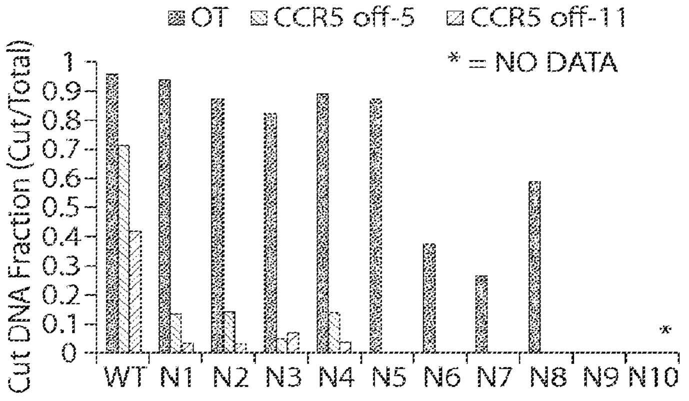
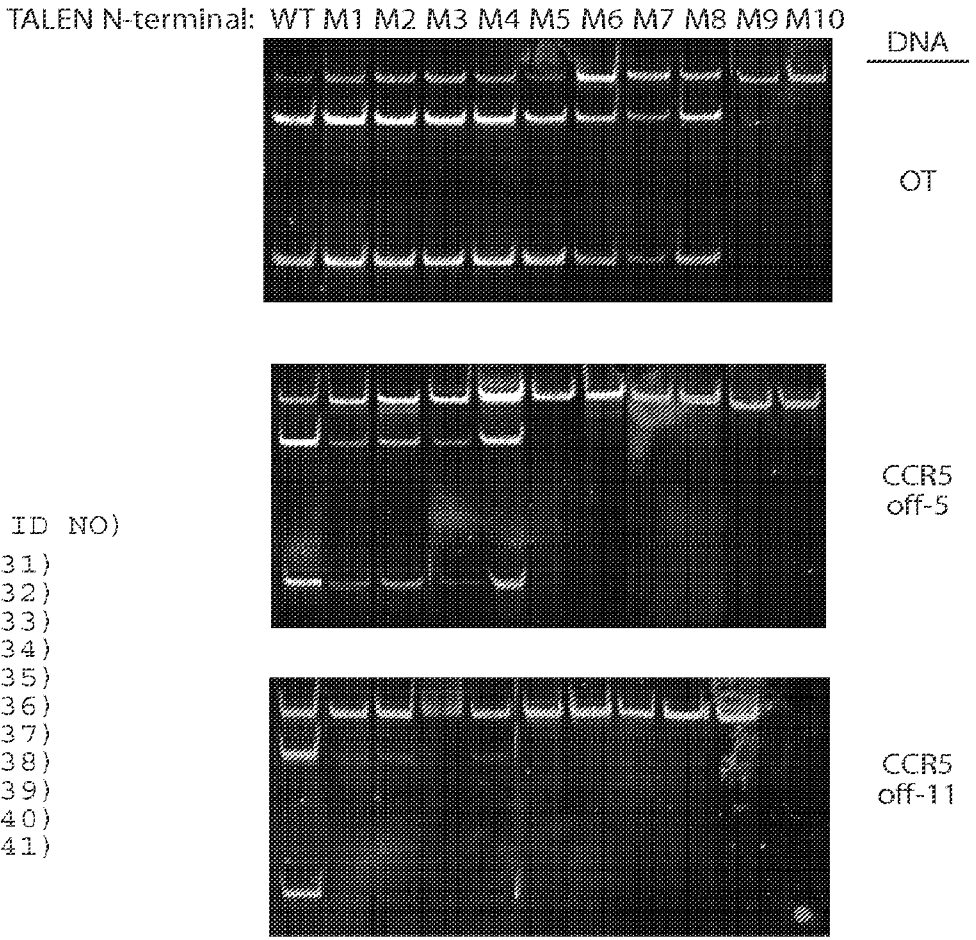

CCR5A TALENs
canonical
Q3
Q7
24nM canonical
12nM canonical
6nM canonical
3nM canonical
24nM Q7
12nM Q7
6nM Q7
N1
N2
N3
canonical ELD/KKR
Q3 ELD/KKR
Q7 ELD/KKR
N2 ELD/KKR
Percentage of all sequences
0%
7.5%
10%
12.5%
15%
20%
25%
30%
50%
0
10
20
DNA spacer length
10
12
14
16
18
20
22
24
Number of spacer DNA base pairs preceding the right half-site

Fig. 17B





TALEN N - terminus (141-162aa)		(SEQ ID NO)
WT	...LQLDTGQLL**K**IA**KR**GGVTAVEA...	(31)
M1	...LQLDTGQLL**Q**IAKRGGVTAVEA...	(32)
M2	...LQLDTGQLLKIA**Q**RGGVTAVEA...	(33)
M3	...LQLDTGQLLKIAK**Q**GGVTAVEA...	(34)
M4	...LQLDTGQLLKIAK**G**GGVTAVEA...	(35)
M5	...LQLDTGQLL**Q**IA**Q**RGGVTAVEA...	(36)
M6	...LQLDTGQLL**Q**IAK**Q**GGVTAVEA...	(37)
M7	...LQLDTGQLLKIA**QQ**GGVTAVEA...	(38)
M8	...LQLDTGQLL**G**IAK**G**GGVTAVEA...	(39)
M9	...LQLDTGQLL**Q**IA**QQ**GGVTAVEA...	(40)
M10	...LQLDTGQLL**G**IA**GG**GGVTAVEA...	(41)

Fig. 18-1

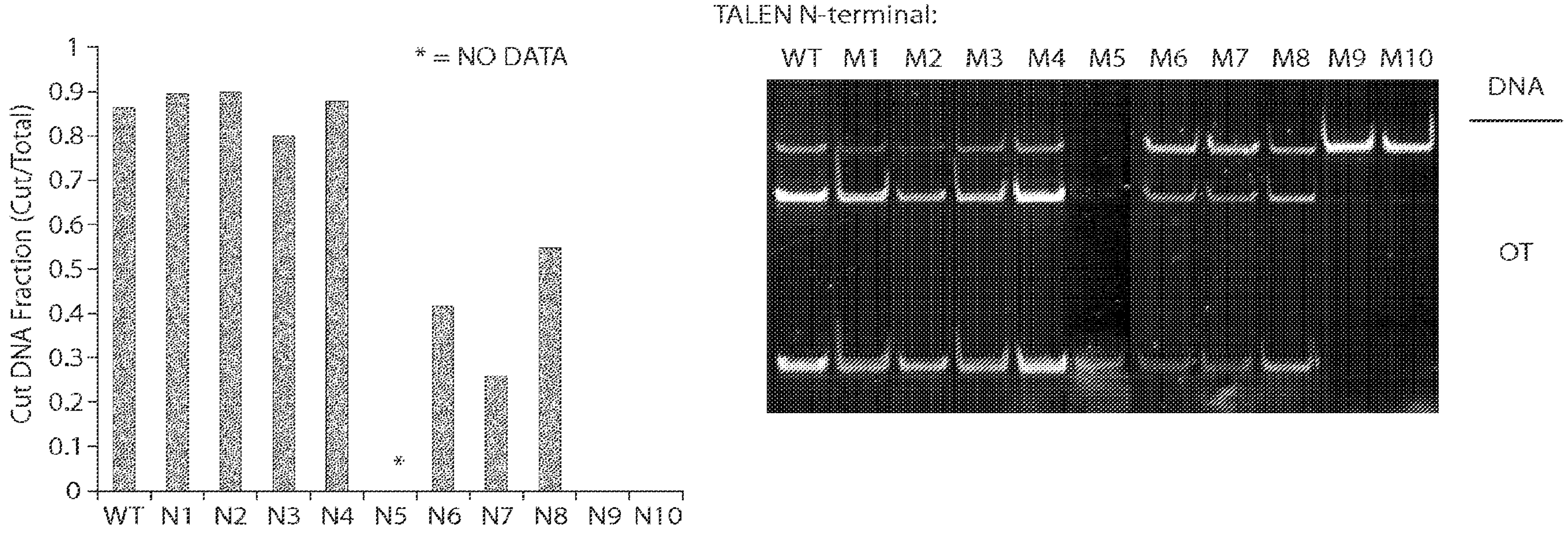

Cut DNA Fraction (Cut/Total)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
WT
N1
N2
N3
N4
N5
N6
N7
N8
N9
N10
*
* = NO DATA
TALEN N-terminal:
WT
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
DNA
OT

Fig. 18-2

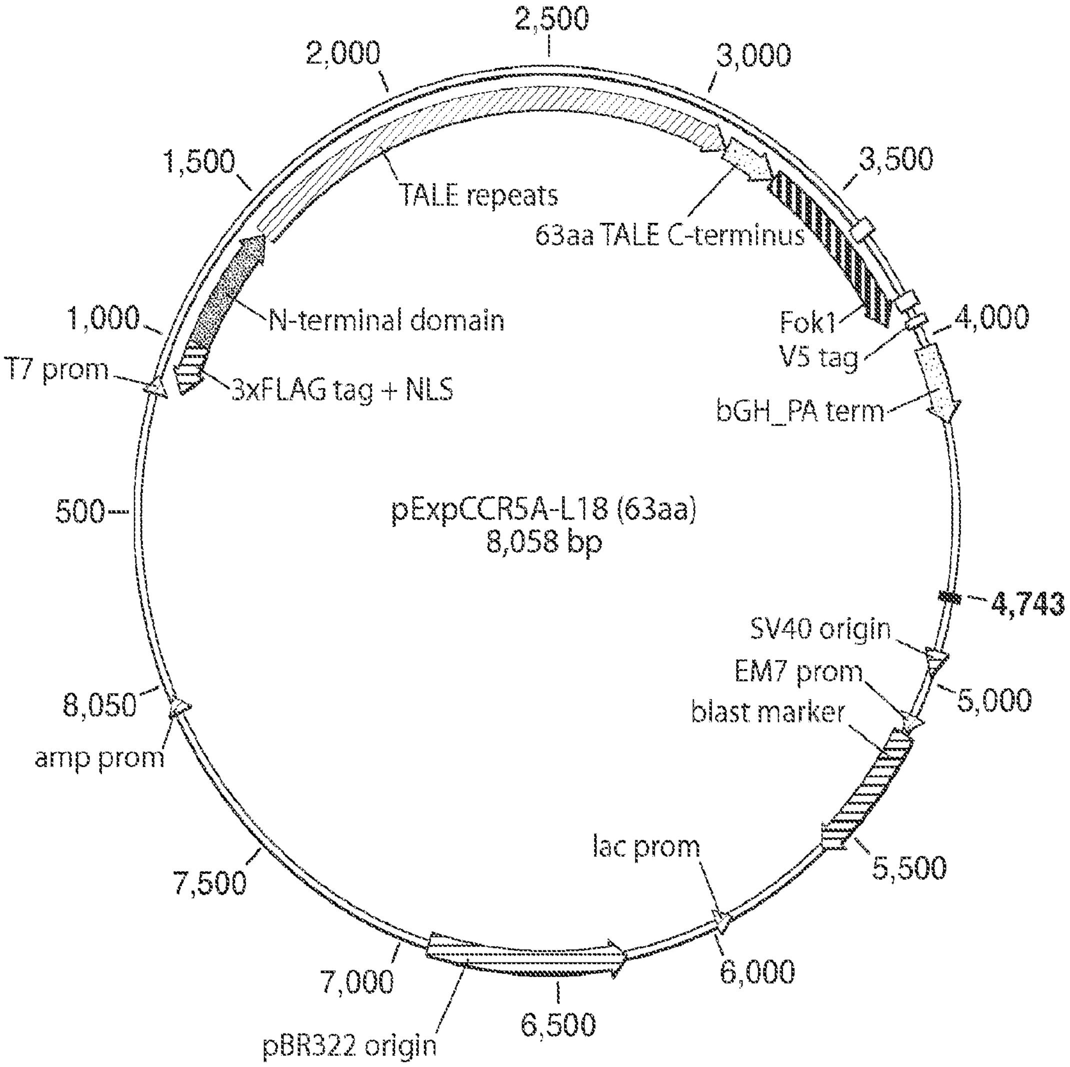

2,500
2,000
3,000
1,500
3,500
TALE repeats
63aa TALE C-terminus
1,000
N-terminal domain
Fok1
4,000
V5 tag
T7 prom
3xFLAG tag + NLS
bGH_PA term
500
pExpCCR5A-L18 (63aa)
8,058 bp
4,743
SV40 origin
EM7 prom
5,000
blast marker
8,050
amp prom
lac prom
5,500
7,500
6,000
7,000
6,500
pBR322 origin

Fig. 19