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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
1 September 2016 (01.09.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/135507 A1

(51) International Patent Classification:
C12N 15/11 (2006.01)

(21) International Application Number:
PCT/GB2016/050508

(22) International Filing Date:
26 February 2016 (26.02.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1503381.4 27 February 2015 (27.02.2015) GB
1521454.7 4 December 2015 (04.12.2015) GB

(71) Applicant: **UNIVERSITY OF EDINBURGH** [GB/GB]; Old College South Bridge, Edinburgh EH8 9YL (GB).

(72) Inventor: **AXTON, Richard**; MRC Centre for Regenerative Medicine, SCRM Building, The University of Edinburgh, Edinburgh Bioquarter 5 Little France Drive, Edinburgh EH16 4UU (GB).

(74) Agent: **GIBBS, Richard**; 120 Bothwell Street, Glasgow G2 7JS (GB).

(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, 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))*

WO 2016/135507 A1

(54) Title: NUCLEIC ACID EDITING SYSTEMS

(57) Abstract: The present invention provides improved methods for generating nucleic acid/genome editing systems and/or tools. In particular, the invention provides methods and tools for generating Transcription Activator Like-Effector (TALE) molecules.

NUCLEIC ACID EDITING SYSTEMS

FIELD OF THE INVENTION

The present invention provides novel methods for the production of nucleic acid or genome editing tools/systems and methods for assessing the same.

BACKGROUND OF THE INVENTION

Genome editing systems such as zinc fingers, Clustered Regularly Interspersed Short Palindromic Repeats (CRISPR) and Transcription Activator Like Effectors (TALEs) have become powerful tools for transcriptional activation and genome editing[1-5]. In particular, CRISPR and TALE technologies utilise relatively simple molecular biology techniques and toolkits are readily available for end users[6-11]. However, the production of these genome-editing tools is laborious.

It is among the objects of this invention to provide alternate methods for the production of genome editing systems (including, for example TALENs). The methods exploit a minimal number of steps and are significantly less complex, time consuming and error prone than prior art protocols.

SUMMARY OF THE INVENTION

The present invention provides improved methods for generating nucleic acid/genome editing systems and/or tools. In particular, the invention provides methods for generating **T**ranscription **A**ctivator **L**ike-**E**ffector (TALE) molecules.

Compared to prior art methods, the methods described herein comprise fewer steps, are significantly less laborious and error prone.

TALE molecules comprise a number of TALE units and the improved methods of this invention exploit a cohort of newly designed TALE unit encoding nucleic acid sequences. These TALE unit encoding nucleic sequences may be combined to provide TALE molecule encoding nucleic acid sequences which may then be synthesised for use. For example, as will be described in more detail below, synthesised TALE molecule nucleic acid encoding sequences may be used to facilitate the expression of TALE protein molecules and TALE fusions for use.

The methods of this invention significantly reduce the reliance on polymerase chain reaction (PCR) based amplification techniques and may find particular application in the

production of **T**ranscription **A**ctivator **L**ike-**E**ndo**n**uclease (TALENs) molecules which are widely regarded as efficient genome editing tools.

Throughout this specification the term "TALE molecule" is used and it should be understood that this term relates both to nucleic acid and/or amino acid sequences which encode or provide complete TALE molecules. A TALE molecule exhibits specificity/affinity for a target nucleic acid sequence and comprises multiple TALE units. A TALE unit exhibits specificity/affinity for a single nucleotide and the term "TALE unit encoding sequence" relates to those sequences (again nucleic acid or amino acid) which encode or provide TALE units.

The term TALE fusion relates to a nucleic acid and/or amino acid sequence encoding or providing a fusion comprising a TALE molecule and a heterologous (i.e. non-TALE) moiety. For example a TALEN may be regarded as a fusion between a TALE molecule and an endonuclease.

The TALE unit encoding nucleic acid sequences of this invention have been designed to avoid problems which prevent prior art methods from being used to efficiently generate TALE molecules. Specifically, the methods used to design the TALE unit encoding nucleic acid sequences presented herein use the degeneracy of the genetic code to ensure that when multiple TALE unit encoding nucleic acid sequences are combined (as would be required in order to form a TALE molecule encoding sequence), the incidence of repetitive DNA sequences within the resulting TALE molecule encoding sequence is (substantially) avoided or (significantly) reduced. In other words, by exploiting the degeneracy of amino acid codons, the inventors have been able to design a series of TALE unit encoding sequences which, when combined or assembled to generate a TALE molecule encoding sequence, yield a sequence with much less internal repetition.

Additionally and to further increase the heterogeneity, additional codon alterations may be made to certain specific positions. For example, certain of the TALE unit encoding sequences presented herein have been subjected to codon alterations at one or more of the positions encoding amino acid residues 4, 11 and/or 32. Further detail regarding these additional modifications is given below.

The reduction in the incidence of repetitive DNA sequences within TALE molecule encoding sequences generated by the methods of this invention is, for example, in comparison to the incidence of repetitive sequences in TALE molecule encoding sequences prepared using multiple copies of TALE units encoding substantially the same sequence.

One of skill will appreciate that it is easier to synthesise nucleic acids with minimal sequence repetition. Indeed, reducing the incidence of repetitive DNA sequences in sequences encoding TALE molecules allows these sequences to be synthesised "in one go" or as multiple (perhaps two or three) large fragments which can be joined or ligated by some suitable technique. As such, the methods of this invention offer a vastly simplified means of generating TALE molecules.

Any given TALE unit encodes a protein sequence which has specificity for a particular nucleotide. It is known that within each TALE unit, the amino acid residues at positions 12 and 13 determine DNA binding specificity. For example, the sequence "NN" binds nucleotide G (guanine), "NI" binds nucleotide A (adenosine), "NG" binds nucleotide T (thymine) and "HD" binds nucleotide C (cytosine).

As such, it is possible to design TALE molecules with specificity for any given nucleic acid sequence (a "target" sequence) by selecting and combining those TALE units with specificity for the nucleic acid residues within the target sequence. For example, if one required a TALE molecule with specificity for a target nucleic acid sequence comprising 10 residues, one would create a TALE comprising 10 TALE units, each unit having specificity for one of the residues in the target nucleic acid sequence.

As such, each of the TALE unit encoding sequences provided by this invention encodes or contains a di-variable amino acid which specifies or determines the binding affinity/specificity of that unit. Thus the TALE unit encoding sequences of this invention encode TALE units which bind one of nucleotides G, A, T or C.

TALE UNIT			
TALE unit	**Nucleotide specificity**	**Nucleic acid sequence (SEQ ID NOS: 1-64)**	**Amino acid sequence (SEQ ID NOS: 65—128)**
1	G	CTTACTCCTGATCAAGTTGTGGCTATTGC GTCTAATAACGGTGGTAAACAAGCTCTTG AAACTGTTCAACGTCTCCTCCCTGTTTTA TGTCAAGATCATGGT	LTPDQVVAIASNNGGKQALETVQRLLPVL CQDHG
	A	CTTACTCCTGATCAAGTTGTGGCTATTGC GTCTAATATTGGTGGTAAACAAGCTCTTG AAACTGTTCAACGTCTCCTCCCTGTTTTA TGTCAAGATCATGGT	LTPDQVVAIASNIGGKQALETVQRLLPVL CQDHG
	T	CTTACTCCTGATCAAGTTGTGGCTATTGC GTCTAATGGAGGTGGTAAACAAGCTCTTG AAACTGTTCAACGTCTCCTCCCTGTTTTA TGTCAAGATCATGGT	LTPDQVVAIASNGGGKQALETVQRLLPVL CQDHG

	C	CTTACTCCTGATCAAGTTGTGGCTATTGC GTCTCATGATGGTGGTAAACAAGCTCTTG AAACTGTTCAACGTCTCCTCCCTGTTTTA TGTCAAGATCATGGT	LTPDQVVAIASHDGGKQALETVQRLLPVL CQDHG
2	G	CTTACCCCTGAGCAGGTCGTAGCCATCGC ATCCAACAATGGCGGCAAGCAGGCCCTAG AGACAGTCCAGCGCCTCCTTCCCGTCTTG TGCCAGGACCACGGC	LTPEQVVAIASNNGGKQALETVQRLLPVL CQDHG
	A	CTTACCCCTGAGCAGGTCGTAGCCATCGC ATCCAACATAGGCGGCAAGCAGGCCCTAG AGACAGTCCAGCGCCTCCTTCCCGTCTTG TGCCAGGACCACGGC	LTPEQVVAIASNIGGKQALETVQRLLPVL CQDHG
	T	CTTACCCCTGAGCAGGTCGTAGCCATCGC ATCCAACGGCGGCGGCAAGCAGGCCCTAG AGACAGTCCAGCGCCTCCTTCCCGTCTTG TGCCAGGACCACGGC	LTPEQVVAIASNGGGKQALETVQRLLPVL CQDHG
	C	CTTACCCCTGAGCAGGTCGTAGCCATCGC ATCCCACGATGGCGGCAAGCAGGCCCTAG AGACAGTCCAGCGCCTCCTTCCCGTCTTG TGCCAGGACCACGGC	LTPEQVVAIASHDGGKQALETVQRLLPVL CQDHG
3	G	CTAACACCAGCTCAAGTGGTTGCAATAGC CTCAAACAATGGTGGAAAACAAGCACTAG AAACAGTACAGCGACTACTACCAGTATTG TGTCAAGCTCACGGA	LTPAQVVAIASNNGGKQALETVQRLLPVL CQAHG
	A	CTAACACCAGCTCAAGTGGTTGCAATAGC CTCAAACATTGGTGGAAAACAAGCACTAG AAACAGTACAGCGACTACTACCAGTATTG TGTCAAGCTCACGGA	LTPAQVVAIASNIGGKQALETVQRLLPVL CQAHG
	T	CTAACACCAGCTCAAGTGGTTGCAATAGC CTCAAACGGAGGTGGAAAACAAGCACTAG AAACAGTACAGCGACTACTACCAGTATTG TGTCAAGCTCACGGA	LTPAQVVAIASNGGGKQALETVQRLLPVL CQAHG
	C	CTAACACCAGCTCAAGTGGTTGCAATAGC CTCACATGATGGTGGAAAACAAGCACTAG AAACAGTACAGCGACTACTACCAGTATTG TGTCAAGCTCACGGA	LTPAQVVAIASHDGGKQALETVQRLLPVL CQAHG
4	G	CTGACGCCGGCCCAGGTAGTCGCGATTGC TAATAACAACGGTGGCAAGCAAGCGCTGG AGACGGTGCAACGGCTGCTGCCGGTGTTA TGCCAAGCCCATGGG	LTPAQVVAIANNNGGKQALETVQRLLPVL CQAHG
	A	CTGACGCCGGCCCAGGTAGTCGCGATTGC TAATAATATTGGTGGCAAGCAAGCGCTGG AGACGGTGCAACGGCTGCTGCCGGTGTTA TGCCAAGCCCATGGG	LTPAQVVAIANNIGGKQALETVQRLLPVL CQAHG
	T	CTGACGCCGGCCCAGGTAGTCGCGATTGC TAATAATGGAGGTGGCAAGCAAGCGCTGG AGACGGTGCAACGGCTGCTGCCGGTGTTA TGCCAAGCCCATGGG	LTPAQVVAIANNGGGKQALETVQRLLPVL CQAHG
	C	CTGACGCCGGCCCAGGTAGTCGCGATTGC TAATCATGATGGTGGCAAGCAAGCGCTGG	LTPAQVVAIANHDGGKQALETVQRLLPVL CQAHG

		AGACGGTGCAACGGCTGCTGCCGGTGTTA TGCCAAGCCCATGGG	
5	G	TTGACTCCCGCACAAGTGGTAGCTATAGC TTCCAACAATGGCGGAAAGCAGGCATTGG AGACTGTACAGAGATTGCTCCCGGTTCTC TGCCAGGCACACGGT	LTPAQVVAIASNNGGKQALETVQRLLPVL CQAHG
	A	TTGACTCCCGCACAAGTGGTAGCTATAGC TTCCAACATCGGCGGAAAGCAGGCATTGG AGACTGTACAGAGATTGCTCCCGGTTCTC TGCCAGGCACACGGT	LTPAQVVAIASNIGGKQALETVQRLLPVL CQAHG
	T	TTGACTCCCGCACAAGTGGTAGCTATAGC TTCCAACGGTGGCGGAAAGCAGGCATTGG AGACTGTACAGAGATTGCTCCCGGTTCTC TGCCAGGCACACGGT	LTPAQVVAIASNGGGKQALETVQRLLPVL CQAHG
	C	TTGACTCCCGCACAAGTGGTAGCTATAGC TTCCCATGATGGCGGAAAGCAGGCATTGG AGACTGTACAGAGATTGCTCCCGGTTCTC TGCCAGGCACACGGT	LTPAQVVAIASHDGGKQALETVQRLLPVL CQAHG
6	G	TTAACCCCAGCGCAGGTTGTCGCCATTGC CAATAACAACGGCGGTAAGCAAGCGTTAG AAACGGTTCAAAGGTTACTGCCTGTATTG TGTCAAGCGCATGGC	LTPAQVVAIANNNGGKQALETVQRLLPVL CQAHG
	A	TTAACCCCAGCGCAGGTTGTCGCCATTGC CAATAACATTGGCGGTAAGCAAGCGTTAG AAACGGTTCAAAGGTTACTGCCTGTATTG TGTCAAGCGCATGGC	LTPAQVVAIANNIGGKQALETVQRLLPVL CQAHG
	T	TTAACCCCAGCGCAGGTTGTCGCCATTGC CAATAACGGTGGCGGTAAGCAAGCGTTAG AAACGGTTCAAAGGTTACTGCCTGTATTG TGTCAAGCGCATGGC	LTPAQVVAIANNGGGKQALETVQRLLPVL CQAHG
	C	TTAACCCCAGCGCAGGTTGTCGCCATTGC CAATCATGATGGCGGTAAGCAAGCGTTAG AAACGGTTCAAAGGTTACTGCCTGTATTG TGTCAAGCGCATGGC	LTPAQVVAIANHDGGKQALETVQRLLPVL CQAHG
7	G	CTTACCCCTGAACAAGTCGTGGCAATCGC GTCGAACAATGGAGGTAAACAAGCTTTAG AAACCGTTCAGCGTCTCCTCCCAGTGTTA TGTCAAGACCATGGT	LTPEQVVAIASNNGGKQALETVQRLLPVL CQAHG
	A	CTTACCCCTGAACAAGTCGTGGCAATCGC GTCGAACATCGGAGGTAAACAAGCTTTAG AAACCGTTCAGCGTCTCCTCCCAGTGTTA TGTCAAGACCATGGT	LTPEQVVAIASNIGGKQALETVQRLLPVL CQAHG
	T	CTTACCCCTGAACAAGTCGTGGCAATCGC GTCGAACGGGGGAGGTAAACAAGCTTTAG AAACCGTTCAGCGTCTCCTCCCAGTGTTA TGTCAAGACCATGGT	LTPEQVVAIASNGGGKQALETVQRLLPVL CQAHG
	C	CTTACCCCTGAACAAGTCGTGGCAATCGC GTCGCACGACGGAGGTAAACAAGCTTTAG AAACCGTTCAGCGTCTCCTCCCAGTGTTA TGTCAAGACCATGGT	LTPEQVVAIASHDGGKQALETVQRLLPVL CQAHG

8	G	CTAACACCAGAGCAGGTGGTGGCGATCGC CAATAATAATGGAGGGAAGCAAGCTCTGG AAACAGTCCAACGCCTTCTTCCGGTTCTT TGTCAAGATCACGGG	LTPEQVVAIANNNGGKQALETVQRLLPVL CQDHG
	A	CTAACACCAGAGCAGGTGGTGGCGATCGC CAATAATATCGGAGGGAAGCAAGCTCTGG AAACAGTCCAACGCCTTCTTCCGGTTCTT TGTCAAGATCACGGG	LTPEQVVAIANNIGGKQALETVQRLLPVL CQDHG
	T	CTAACACCAGAGCAGGTGGTGGCGATCGC CAATAACGGTGGAGGGAAGCAAGCTCTGG AAACAGTCCAACGCCTTCTTCCGGTTCTT TGTCAAGATCACGGG	LTPEQVVAIANNGGGKQALETVQRLLPVL CQDHG
	C	CTAACACCAGAGCAGGTGGTGGCGATCGC CAATCACGACGGAGGGAAGCAAGCTCTGG AAACAGTCCAACGCCTTCTTCCGGTTCTT TGTCAAGATCACGGG	LTPEQVVAIANHDGGKQALETVQRLLPVL CQDHG
9	G	CTGACTCCAGATCAAGTTGTTGCCATAGC ATCGAACAATGGTGGCAAACAGGCTCTGG AAACCGTCCAAAGATTACTTCCAGTTTTA TGCCAAGCCCACGGT	LTPDQVVAIASNNGGKQALETVQRLLPVL CQAHG
	A	CTGACTCCAGATCAAGTTGTTGCCATAGC ATCGAACATTGGTGGCAAACAGGCTCTGG AAACCGTCCAAAGATTACTTCCAGTTTTA TGCCAAGCCCACGGT	LTPDQVVAIASNIGGKQALETVQRLLPVL CQAHG
	T	CTGACTCCAGATCAAGTTGTTGCCATAGC ATCGAACGGTGGTGGCAAACAGGCTCTGG AAACCGTCCAAAGATTACTTCCAGTTTTA TGCCAAGCCCACGGT	LTPDQVVAIASNGGGKQALETVQRLLPVL CQAHG
	C	CTGACTCCAGATCAAGTTGTTGCCATAGC ATCGCACGATGGTGGCAAACAGGCTCTGG AAACCGTCCAAAGATTACTTCCAGTTTTA TGCCAAGCCCACGGT	LTPDQVVAIASHDGGKQALETVQRLLPVL CQAHG
10	G	TTGACCCCAGACCAGGTCGTCGCTATTGC TAACAATAACGGGGGGCAAACAAGCCTTGG AGACAGTACAAAGGCTTCTCCCCGTTCTA TGCCAGGATCACGGC	LTPDQVVAIANNNGGKQALETVQRLLPVL CQDHG
	A	TTGACCCCAGACCAGGTCGTCGCTATTGC TAACAATATTGGGGGGCAAACAAGCCTTGG AGACAGTACAAAGGCTTCTCCCCGTTCTA TGCCAGGATCACGGC	LTPDQVVAIANNIGGKQALETVQRLLPVL CQDHG
	T	TTGACCCCAGACCAGGTCGTCGCTATTGC TAACAATGGCGGGGGGCAAACAAGCCTTGG AGACAGTACAAAGGCTTCTCCCCGTTCTA TGCCAGGATCACGGC	LTPDQVVAIANNGGGKQALETVQRLLPVL CQDHG
	C	TTGACCCCAGACCAGGTCGTCGCTATTGC TAACCACGATGGGGGGCAAACAAGCCTTGG AGACAGTACAAAGGCTTCTCCCCGTTCTA TGCCAGGATCACGGC	LTPDQVVAIANHDGGKQALETVQRLLPVL CQDHG
11	G	CTTACGCCAGCTCAAGTAGTAGCGATAGC CTCTAATAATGGTGGGAAGCAGGCGCTCG	LTPAQVVAIASNNGGKQALETVQRLLPVL CQDHG

		AGACAGTTCAACGACTACTCCCGGTATTA TGTCAAGATCATGGG	
	A	CTTACGCCAGCTCAAGTAGTAGCGATAGC CTCTAATATAGGTGGGAAGCAGGCGCTCG AGACAGTTCAACGACTACTCCCGGTATTA TGTCAAGATCATGGG	LTPAQVVAIASNIGGKQALETVQRLLPVL CQDHG
	T	CTTACGCCAGCTCAAGTAGTAGCGATAGC CTCTAATGGAGGTGGGAAGCAGGCGCTCG AGACAGTTCAACGACTACTCCCGGTATTA TGTCAAGATCATGGG	LTPAQVVAIASNGGGKQALETVQRLLPVL CQDHG
	C	CTTACGCCAGCTCAAGTAGTAGCGATAGC CTCTCATGACGGTGGGAAGCAGGCGCTCG AGACAGTTCAACGACTACTCCCGGTATTA TGTCAAGATCATGGG	LTPAQVVAIASHDGGKQALETVQRLLPVL CQDHG
12	G	CTCACACCCGCCCAGGTTGTAGCAATTGC CTCGAACAACGGCGGCAAGCAAGCACTTG AGACTGTCCAGCGGCTCTTGCCAGTTCTC TGCCAGGCACACGGC	LTPAQVVAIASNNGGKQALETVQRLLPVL CQAHG
	A	CTCACACCCGCCCAGGTTGTAGCAATTGC CTCGAACATTGGCGGCAAGCAAGCACTTG AGACTGTCCAGCGGCTCTTGCCAGTTCTC TGCCAGGCACACGGC	LTPAQVVAIASNIGGKQALETVQRLLPVL CQAHG
	T	CTCACACCCGCCCAGGTTGTAGCAATTGC CTCGAACGGTGGCGGCAAGCAAGCACTTG AGACTGTCCAGCGGCTCTTGCCAGTTCTC TGCCAGGCACACGGC	LTPAQVVAIASNGGGKQALETVQRLLPVL CQAHG
	C	CTCACACCCGCCCAGGTTGTAGCAATTGC CTCGCATCATGGCGGCAAGCAAGCACTTG AGACTGTCCAGCGGCTCTTGCCAGTTCTC TGCCAGGCACACGGC	LTPAQVVAIASHDGGKQALETVQRLLPVL CQAHG
13	G	CTAACTCCAGCACAAGTCGTTGCTATCGC TAACAACAATGGTGGCAAACAGGCATTAG AAACCGTTCAACGTCTTTTACCGGTCCTG TGCCAAGCTCACGGC	LTPAQVVAIANNNGGKQALETVQRLLPVL CQAHG
	A	CTAACTCCAGCACAAGTCGTTGCTATCGC TAACAACATCGGTGGCAAACAGGCATTAG AAACCGTTCAACGTCTTTTACCGGTCCTG TGCCAAGCTCACGGC	LTPAQVVAIANNIGGKQALETVQRLLPVL CQAHG
	T	CTAACTCCAGCACAAGTCGTTGCTATCGC TAACAACGGTGGTGGCAAACAGGCATTAG AAACCGTTCAACGTCTTTTACCGGTCCTG TGCCAAGCTCACGGC	LTPAQVVAIANNGGGKQALETVQRLLPVL CQAHG
	C	CTAACTCCAGCACAAGTCGTTGCTATCGC TAACCACGACGGTGGCAAACAGGCATTAG AAACCGTTCAACGTCTTTTACCGGTCCTG TGCCAAGCTCACGGC	LTPAQVVAIANHDGGKQALETVQRLLPVL CQAHG
14	G	CTGACCCCTGCGCAGGTTGTAGCGATAGC CAACAATAATGGCGGTAAGCAAGCCCTGG AAACAGTACAACGTCTACTGCCTGTGTTG TGCCAAGCTCATGGT	LTPAQVVAIANNNGGKQALETVQRLLPVL CQAHG

	A	CTGACCCCTGCGCAGGTTGTAGCGATAGC CAACAATATTGGCGGTAAGCAAGCCCTGG AAACAGTACAACGTCTACTGCCTGTGTTG TGCCAAGCTCATGGT	LTPAQVVAIANNIGGKQALETVQRLLPVL CQAHG
	T	CTGACCCCTGCGCAGGTTGTAGCGATAGC CAACAATGGAGGCGGTAAGCAAGCCCTGG AAACAGTACAACGTCTACTGCCTGTGTTG TGCCAAGCTCATGGT	LTPAQVVAIANNGGGKQALETVQRLLPVL CQAHG
	C	CTGACCCCTGCGCAGGTTGTAGCGATAGC CAACCATGATGGCGGTAAGCAAGCCCTGG AAACAGTACAACGTCTACTGCCTGTGTTG TGCCAAGCTCATGGT	LTPAQVVAIANHDGGKQALETVQRLLPVL CQAHG
15	G	TTGACTCCAGAACAAGTAGTCGCCATCGC CAACAATAATGGAGGTAAACAGGCTTTAG AGACTGTGCAAAGACTTCTTCCTGTATTA TGTCAGGCCCATGGT	LTPEQVVAIANNNGGKQALETVQRLLPVL CQAHG
	A	TTGACTCCAGAACAAGTAGTCGCCATCGC CAACAATATTGGAGGTAAACAGGCTTTAG AGACTGTGCAAAGACTTCTTCCTGTATTA TGTCAGGCCCATGGT	LTPEQVVAIANNNGGKQALETVQRLLPVL CQAHG
	T	TTGACTCCAGAACAAGTAGTCGCCATCGC CAACAATGGTGGAGGTAAACAGGCTTTAG AGACTGTGCAAAGACTTCTTCCTGTATTA TGTCAGGCCCATGGT	LTPEQVVAIANNGGGKQALETVQRLLPVL CQAHG
	C	TTGACTCCAGAACAAGTAGTCGCCATCGC CAACCATGATGGAGGTAAACAGGCTTTAG AGACTGTGCAAAGACTTCTTCCTGTATTA TGTCAGGCCCATGGT	LTPEQVVAIANHDGGKQALETVQRLLPVL CQAHG
16	G	TTAACGCCAGAGCAGGTTGTTGCAATAGC AAACAACAACGGAGGTAAACAAGCGCTCG AAACGGTCCAACGTCTCTTGCCCGTCCTT TGTCAAGCGCACGGA	LTPEQVVAIANNNGGKQALETVQRLLPVL CQAHG
	A	TTAACGCCAGAGCAGGTTGTTGCAATAGC AAACAATATCGGAGGTAAACAAGCGCTCG AAACGGTCCAACGTCTCTTGCCCGTCCTT TGTCAAGCGCACGGA	LTPEQVVAIANNIGGKQALETVQRLLPVL CQAHG
	T	TTAACGCCAGAGCAGGTTGTTGCAATAGC AAACAATGGTGGAGGTAAACAAGCGCTCG AAACGGTCCAACGTCTCTTGCCCGTCCTT TGTCAAGCGCACGGA	LTPEQVVAIANNGGGKQALETVQRLLPVL CQAHG
	C	TTAACGCCAGAGCAGGTTGTTGCAATAGC AAATCACGATGGAGGTAAACAAGCGCTCG AAACGGTCCAACGTCTCTTGCCCGTCCTT TGTCAAGCGCACGGA	LTPEQVVAIANHDGGKQALETVQRLLPVL CQAHG

A novel cohort of TALE units is presented in Table 1 below.

Table 1: A cohort of TALE units

As such, a first aspect of this invention relates to one or more of the TALE unit sequences presented in TABLE 1. Specifically, the invention relates both to one or more of the TALE unit nucleic acid sequences and/or one or more of the TALE unit amino acid sequences of Table 1 (i.e. each of SEQ ID NO: 1-128).

It should be understood that any one of the TALE unit sequences presented in Table 1 and especially the TALE unit encoding nucleic acid sequences, may be modified to include additional sequences. The additional sequences may be 5' and/or 3' additional sequences. The additional sequences may provide or encode sequences which facilitate, for example, restriction, joining/ligation (of one TALE unit to another) amplification and/or purification. As such, the invention relates to a TALE unit sequence conforming to the following consensus:

$\mathbf{A_1}$ – [**TU**] – $\mathbf{A_2}$

wherein $\mathbf{A_1}$ represents an optional additional sequence or modification;

TU represents any one of the 64 TALE unit encoding nucleic acid sequences presented in Table 1; and

$\mathbf{A_2}$ represents an optional additional sequence or modification.

Optional sequences A_1 and A_2 may comprise, for example, restriction site sequences, primer binding sites, sequences which facilitate the ligation or joining of one TALE unit sequence (or molecule comprising the same) to another. In some cases, optional sequences A_1 and A_2 may comprise or further comprise sequences encoding parts of other TALE unit sequences. Some suitable additional sequences are identified in this application and it should be understood that any of these sequences may (subject to minor modification) be appended or added to any of the sequences presented in Table 1. However, one of skill will be familiar with the types of sequence that can be added to the TALE unit encoding sequences of this invention.

A second aspect provides a TALE molecule (either a nucleic acid or amino acid molecule) comprising two or more of the TALE unit (nucleic acid or amino acid) sequences provided by the first aspect of this invention. A TALE molecule according to this second aspect of the invention may be a TALE molecule encoding nucleic acid sequence or a TALE molecule amino acid sequence (namely, the product of the TALE molecule encoding nucleic acid sequence).

A third aspect provides a sequence encoding a TALE fusion, wherein the TALE fusion comprises a TALE unit sequence or a TALE molecule sequence fused (optionally via

a linker moiety) to a heterologous (i.e. non-TALE type) sequence. For example, the TALE fusion may be nucleic acid fusion (comprising a TALE molecule encoding sequence fused to a heterologous nucleic acid sequence) or an amino acid fusion comprising a TALE molecule amino acid sequence fused to a heterologous amino acid sequence. A TALE fusion may, for example, encode or provide a TALEN.

In a further (fourth) aspect, the invention provides a method of generating a TALE molecule sequence, said method comprising combining or assembling one or more of the TALE unit sequences provided by the first aspect of the invention (for example those presented in Table 1) to provide a TALE molecule encoding sequence or a TALE molecule.

The method of generating a TALE molecule sequence may comprise generating a TALE molecule encoding nucleic acid sequence. The method may require the user to combine or assemble together one or more TALE unit encoding nucleic acid sequences in order to provide a larger TALE molecule encoding nucleic acid sequence.

The method according to the fourth aspect of this invention may further require the selection and/or analysis of a target nucleic acid sequence; that is a nucleic acid sequence to which the TALE molecule is to exhibit some binding specificity/affinity. Thus, the method of the fourth aspect of this invention provides TALE molecule sequences which have binding specificity/affinity for predetermined target nucleic acid sequences.

For example, using the TALE sequences provided in Table 1, it is possible to select and combine/assemble those sequences which have binding specificity for some or all of the nucleotides within a target sequence. Thus the methods of this invention may comprise (on the basis of target sequence analysis/information) combining the relevant or required number of TALE unit encoding nucleic acid sequences.

As stated, a TALE molecule may comprise any number of TALE units. For example, a TALE molecule may comprise 10-30 TALE units, for example 15-20 TALE units. A TALE molecule may comprise 16 TALE units. The skilled person will however appreciate that while a TALE molecule may comprise almost any number of TALE units, the actual number of TALE units used may be determined by the length of the target sequence.

For example, where the target sequence comprises 16 nucleotides, the TALE molecule may also comprise 16 TALE units, each having specificity for a nucleotide of the target sequence. Thus, in this case, the method of this invention may require the selection and combination of 16 of the TALE unit encoding sequences presented in Table 1.

It should be understood that when generating a TALE molecule, while it is desirable to use TALE unit encoding sequences which minimise the amount of sequence repetition across the full length of the nucleic acid encoding the TALE molecule, the same TALE unit may be used multiple (two or more) times. In any case, when selecting TALE units to combine, the user will make their selection while all the time trying to minimise incidences of sequence repetition within the generated TALE molecule sequence.

The method according to the fourth aspect of this invention may comprise computationally assembling the TALE unit encoding sequences to provide a TALE molecule sequence.

The term "computationally assembling" should be taken to encompass the act of using a computer (or other automated device) to provide a suitable (i.e. target sequence specific) TALE molecule sequence. To create a target molecule specific sequence, the computer may be imputed with target sequence information - for example the sequence of the target sequence or region of the target sequence that the TALE molecule is to bind. Thereafter, the computer may interrogate a database comprising at least the TALE unit sequences described in the first aspect of the invention/Table 1 so as to provide a suitable TALE molecule sequence. Again, any method of computationally assembling a TALE molecule sequence may take into account the need to minimise sequence repetition within the computationally assembled TALE molecule.

A TALE molecule sequence as provided by a method according to the fourth aspect of this invention may be synthesised, for example chemically synthesised, to provide a TALE molecule sequence for use. The synthesised sequences may be nucleic acid sequences encoding TALE molecules for use.

A method according to the fourth aspect of this invention may yield a nucleic acid sequence which encodes a complete TALE molecule. In such cases, the TALE molecule encoding nucleic acid sequence may be synthesised as a single sequence corresponding to the full length sequence of the required TALE molecule.

Alternatively, the methods of this invention may be exploited in order to provide multiple (for example 2, 3, 4 or more) fragments each providing part of a complete TALE molecule encoding nucleic acid sequence. These fragments can be joined together or ligated by some suitable method.

Where the TALE molecule encoding nucleic acid sequence is to be synthesised as one or more fragments for joining or ligation, each fragment may comprise 5' and/or 3'

modifications to permit the joining to or ligation with, other fragments. For example, the 5' and/or 3' ends of any of the fragments may comprise sequences which facilitate joining, ligation, amplification, restriction and/or cloning. Suitable 5' and/or 3' modifications are described in relation to the first aspect of this invention and the same definitions apply here.

For example, where the 5' and/or 3' ends of any of the fragments comprise sequences which harbour restriction sites, by treating with the appropriate restriction enzyme, it is possible to provide fragments with ends which can be joined to, or ligated with, the corresponding (restricted) end of another fragment. Additionally or alternatively, fragments for joining or ligation may comprise other 5' and/or 3' modifications and/or sequences which facilitate joining and/or ligation.

Fragments with 5' and/or 3' modifications may be prepared by combining TALE unit sequences in which the first and last unit sequences have 5' and 3' modifications respectively. Additionally, or alternatively, the modifications may be added to the 5' and/or 3' end of the fragment later using suitable molecular techniques.

The fragments may be designed so that they are suitable for joining by Gibson assembly. One of skill in this field will be familiar with Gibson assembly which is a method allowing the joining of multiple DNA fragments in a single, isothermal reaction. Using this method, it is possible to simultaneously combine numerous (>10) DNA fragments based on sequence identity. Gibson assembly methods generally require that the DNA fragments to be joined contain approximately a 20-40 base pair overlap with adjacent DNA fragments. The DNA fragments are then mixed with a cocktail of enzymes (for example three enzymes) and other buffer components. Compared to conventional restriction enzyme/ligation cloning of recombinant DNA Gibson assembly avoids the need for restriction digest of the DNA fragments after amplification by PCR. Moreover, DNA joining by Gibson assembly is simpler, requires fewer steps and takes less time. A further advantage of joining by Gibson assembly is that the process yields no restriction site scar (i.e. the process is "scarless"). It is also possible to combine multiple DNA fragments simultaneously in a single-tube reaction.

As such, the methods of this invention may provide multiple (for example 2, 3, 4 or more) fragments, each fragment providing part of a complete TALE molecule encoding nucleic acid sequence, wherein the fragments are designed to permit joining by Gibson assembly.

Further information regarding ligation and other nucleic acid joining procedures which may be useful in the methods of this invention are described in Molecular Cloning: A Laboratory Manual (Hughes & Joseph Sambrook: CSHLP; Fourth Edition) – the contents of this publication are incorporated herein by reference.

It should be noted that while the present invention may, in part rely on Gibson assembly, over use of the technique (as occurs with some prior art methods) can be deleterious. Gibson assembly tolerates mismatches and therefore, when creating large molecules, the more units joined by Gibson assembly, the more errors can occur. Indeed, if one attempts to join too many individual units by Gibson assembly based methods, errors can lead to the generation of partial, rather than full length, clones.

The present invention offers an advantage over prior art methods as it minimises the use of Gibson assembly. Large fragments for joining are first synthesised as complete "blocks" and only these fragments are joined by Gibson assembly methods. Indeed, the inventors have shown that methods in which the fragments are themselves created by the Gibson assembly of multiple units yield incomplete and partial length TALEN.

A further advantage associated with this invention (which methods exploit minimal steps and a reduced reliance on Gibson assembly) is that is easily adaptable to accommodate advances in the field.

The methods of this invention also require minimal use of consumable products. Methods which are over reliant on Gibson assembly and other methods of joining nucleic acid sequences, may require large stocks of components – some of which may be consumed faster than others. The present invention allows for the rapid design and synthesis of TALE molecules of any length (i.e. comprising any number of TALE units) without the need to modify complex protocols and procedures. Once the required number of TALE molecules has been generated, a complete TALE molecule may be generated by simply assembling together the various synthesised TALE molecules.

Where multiple fragments are assembled by ligation, PCR may then be used to generate a complete TALE molecule encoding amplicon.

Once generated, a TALE encoding nucleic acid sequence (or a TALE molecule encoding amplicon) may be introduced into a vector, for example an expression vector. A TALE encoding nucleic acid sequence may be introduced into a vector using standard cloning procedures; again, useful protocols are summarised in Molecular Cloning: A Laboratory Manual (Hughes & Joseph Sambrook: CSHLP; Fourth Edition). By way of example, restriction enzyme based cloning methods, Golden Gate Cloning and/or Gibson assembly based methods may be used to facilitate and/or affect the introduction of the TALE

molecule encoding nucleic acid sequence into a suitable vector. One of skill will appreciate that the precise method of cloning used may depend on the 5' and/or 3' features (restriction sites, sequence and the like) of the TALE molecule encoding nucleic acid sequence.

The selection of vector may depend on whether or not the TALE molecule is to be joined or fused to a heterologous moiety. If the TALE molecule is to be joined or fused to a heterologous moiety, then the selected vector may include a sequence encoding said moiety. The moiety may be an endonuclease and therefore the vector may contain an endonuclease encoding sequence.

In order to introduce a TALE molecule encoding nucleic acid sequence into a vector for use in this invention, the vector may be cut or restricted with a suitable enzyme and the cut vector and TALE molecule sequence to be introduced, incubated together under conditions which facilitate the introduction (cloning) of the TALE molecule encoding sequence into the vector.

Introduction of a TALE molecule encoding nucleic acid sequence into a vector which contains an endonuclease encoding sequence may facilitate the generation of **T**ranscription **A**ctivator **L**ike-**E**ndo**n**uclease (TALEN) molecules. Molecules of this type exploit the DNA binding specificity of the TALE part to provide efficient genome editing tools.

Expression of a TALE molecule, a TALE molecule fusion (i.e. a TALE **::** heterologous moiety fusion) or a TALEN, may be achieved by introduction of the vector into a suitable host cell. Host cells may be transfected and/or transformed with vectors by any suitable means including, for example, heat shock, electroporation and/or chemical based techniques. Prokaryotic and/or eukaryotic cells may be transformed or transfected with vectors – including the vectors provided by this invention. As such, bacterial and/or mammalian, for example human, cells may be transformed and/or transfected.

A transformed/transfected host cell may be maintained under and/or in conditions which are suitable for the expression of the TALE and any fused, associated or joined heterologous sequence. For example, the conditions may include the use of agents which induce expression and/or agents which facilitate the selection of transformed/transfected cells.

A suitable vector may include a mammalian expression vector such as the FOK1 endonuclease expression vector. As such, a TALE molecule encoding nucleic acid sequence of this invention and/or prepared according to a method of this invention, may be introduced into a FOK1 endonuclease expression vector.

A suitable host cell may be any competent mammalian cell including cells from established cell lines. The skilled man will be aware of the array of cells that can be used and such cells may be obtained from culture collections such as those held and catalogued at http://www.phe-culturecollections.org.uk/. Suitable cells include, but are not limited to, 273FT cells.

In view of the above, the invention provides a method of generating a TALE molecule, said method comprising the steps of:

combining two or more of the TALE unit encoding nucleic acid sequences provided in Table 1 to yield a TALE molecule encoding nucleic acid sequence specific for a predetermined target sequence;

synthesising the TALE molecule encoding nucleic acid sequence;

introducing the TALE molecule encoding nucleic acid sequence into a vector; and

introducing the vector into a host cell and maintaining the host cell under conditions which facilitate the expression of the TALE molecule encoding nucleic acid sequence.

A TALE molecule prepared by a method of this invention may be purified by any suitable means. For example, a TALE molecule may be purified using, for example, affinity chromatography. For example, A TALE molecule may be modified to include a fusion tag (for example a His tag) at its 5' and/or 3' end. The fused tag may then be used as a means to purify or extract the TALE molecule from, for example, a heterogeneous protein mix. For example, a fusion tagged (His tagged) TALE molecule may be expressed in a bacterial cell and harvested or purified from the cell lysate. The fused tag (in particular fused 5' or 3' His tags) should not affect TALE binding and therefore may not need to be cleaved.

Additionally, or alternatively, a TALE molecule may be expressed with an N-terminal leader sequence - in this way, it may be possible to facilitate secretion of the TALE molecule from the cell.

A TALE molecule may be further modified or supplemented with sequences, for example, viral (TAT) sequences which facilitate, permit or enhance cellular uptake.

In addition, the invention provides a method of generating a TALEN molecule, said method comprising the steps of:

combining two or more of the TALE unit encoding nucleic acid sequences provided in Table 1 to yield a TALE molecule encoding nucleic acid sequence specific for a predetermined target sequence;

synthesising the TALE molecule encoding nucleic acid sequence;

introducing the TALE molecule encoding nucleic acid sequence into a vector, which vector comprises an endonuclease encoding nucleic acid to provide a vector which encodes a TALEN molecule; and

introducing the vector into a host cell and maintaining the host cell under conditions which facilitate the expression of a TALEN encoding nucleic acid molecule.

Optionally, the expressed protein product of the TALEN encoding nucleic acid molecule – namely the TALEN molecule protein, may be harvested or purified by any suitable means.

As described above, a TALEN molecule generated by any of the methods described herein may be modified to include one or more fusion tags to facilitate purification by, for example affinity chromatography techniques. Moreover, the TALEN molecule may further comprise one or more sequences or motifs (for example leader sequences or viral derived (TAT) sequences) which facilitate TALEN cell export/secretion and/or import (cell uptake).

The TALEN may be expressed *in situ* (i.e. within the cell in which it is expressed). The TALEN may be used to genome edit the cell. For example, the TALEN may be used as a means to affect a mutation by NHEJ in a specific gene and/or to make a reporter line by introducing a donor cassette vector with a selectable marker or fluorescent protein.

In a fifth aspect, the invention further provides TALE and/or TALEN molecules obtainable by any of the methods described herein.

It should be noted that when compared to TALE or TALEN molecules prepared by other (prior art) methods, the TALE and/or TALEN molecules generated (or obtainable) by any of the methods described herein demonstrate comparable activity.

A method of designing, generating or providing a TALE and/or TALEN molecule (as described or disclosed herein) may exploit a Computer Aided Design (CAD) program, wherein the CAD program may perform a method of determining an appropriate or suitable assembly of TALE unit sequences. For example, a CAD program may be exploited as a means to produce a TALE and/or TALEN molecule specific for a predetermined target sequence.

A program for use may require a user to input data (for example sequence data) relating to a target DNA sequence. The data may be input into a computer and the computer may execute the CAD program such that it selects an appropriate cohort of TALE units for the generation of a TALE or TALEN molecule with specificity for the target sequence. The

computer may inform the user of which units are to be used and in which order they are to be used to generate a suitable TALE/TALEN molecule.

The computer may comprise or be loaded with the information presented in Table 1 above. As such, upon receipt of data input by a user, for example sequence data, the computer can select the most suitable TALE units from the library presented in Table 1. When selecting a Tale unit suitable to bind the first residue in the target sequence, the computer may select one of the four units presented in Table 1 as unit 1. The same process may be repeated for the second, third and all subsequent residues of the target sequence.

Where the TALE/TALEN molecule comprises 16 TALE unit encoding sequences presented in Table 1, the computer program will perform operations to determine the required assembly of the necessary 16 TALE unit sequences from the information presented in Table 1. For each necessary TALE unit sequence, the computer program will select a TALE unit sequences (from the data presented in Table 1) that exhibits the necessary specificity for a nucleotide of the target DNA sequence.

The computer program may be configured to optimise the TALE unit encoding sequences or the selection of TALE unit encoding sequences, so as to minimise the amount of sequence repetition across the full length of a resulting TALE molecule encoding.

A computer program may comprise elements that are executed at least one of sequentially, in-parallel, in-order or out-of-order. The computer program may be written, created or synthesised in a language such as “R”, “S”, “S-Plus”, “C”, “C++”, or the like. The computer program may include algorithms and/or library components for functionality which comprises table look-up, table search, string operations, matrix operations, vector operations, statistical operations, or the likes. Further, the term “computer program” may refer to a computer program, or a macro, script, or any other sequence of operations executed directly by at least one computer or within another computer program executing on at least one computer, such as Microsoft Excel™ or the like. The computer program, macro or script or any other sequence of operations may reside and/or execute on a computer local to the user, or may reside and/or execute remotely from the user on at least one computer.

In a sixth aspect the invention provides a method of providing TALE units for use. For example, the TALE units may be for use in a method according to the fourth aspect of this invention. The method may comprise providing a plurality of TALE units having, relative to a reference TALE unit encoding sequence, one or more conservative codon modifications.

In the context of this invention, a conservative codon modification may be taken to be any modification which, through the degeneracy of the genetic code, preserves the encoded wild type amino acid residue. For example, where the wild type amino acid residue is alanine, conservative codon modifications may include selection of any one of the codons "GCT"; "GCC"; "GCA" or "GCG" - all of which encode alanine.

The TALE unit encoding sequences generated or provided by the method according to the sixth aspect of this invention may each comprise a nucleic acid sequence encoding a TALE unit with the following sequence (SEQ ID NO: 129):

LTPX$_1$QVVAIAX$_2$X$_3$X$_4$GGKQALETVQRLLPVLCQX$_5$HG

The amino acids at positions 4 (X_1), 11 (X_2) and 32 (X_5) are variable. The skilled person will appreciate that any number of modifications, in particular conservative modifications may be made at these positions. Table 2 details a range of conservative substitutions that may be exploited in this invention. For example the amino acid alanine (Ala) may be replaced with any one of Pro, Gly, Glu, Asp, Gln, Asn, Ser or Thr.

Hydrophilic:	Ala, Pro, Gly, Glu, Asp, Gln, Asn, Ser, Thr
Sulphydryl:	Cys
Aliphatic:	Val, Ile, Leu, Met
Basic:	Lys, Arg, His
Aromatic:	Phe, Tyr, Trp.

Table 2 – conservative amino acid substitutions

Without wishing to be bound by theory, the amino acid residue at position 4 may be D (Aspartic acid), E (Glutamic acid) or A (alanine). The amino acid residue at position 11 may be S (serine) or N (Asparagine). The amino acid residue at position 32 may be A (alanine) or D (aspartic acid).Suitable alterations may manifest as alterations which modulate (for example improve or enhance) binding between a TALE unit and the nucleotide it has binding specificity/affinity for.

Residues 12 and 13 (marked X_3X_4 above) are also variable and the exact sequence will depend on the intended binding specificity of the TALE unit. For example, the residues selected to occupy positions X_3 and X_4 may be any suitable to bestow or impart the desired nucleotide binding specificity to the TALE unit. For example, while there exist many possibilities and without wishing to be bound by any particular theory, X_3 and X_4 may be selected from the group consisting of NN; NI; NK; NG and HD.

A codon encoding residue "N" may be AAT or AAC, a codon encoding residue "I" may be ATT, ATC or ATA, a codon encoding residue G may be GGT, GGC, GGA or GGG, a codon encoding residue "H" may be CAT or CAC; a codon encoding residue "K" may be AAA or AAG and a codon encoding "D" may be GAT or GAC.

Table 3 below shows available codon selections at each of positions 1-11 and 14-34 in an example TALE unit encoding sequence. It should be noted that at positions 4, 11 and 32, the encoded amino acid is variable and so each variant (together with the codon options) is presented.

Residue No:	Wild type amino acid residue	CODON selection
1	L	CTT CTC CTA CTG
2	T	ACT ACC ACA ACG
3	P	CCT CCC CCA CCG
4a	E	GAA GAG
4b	D	GAT GAC
4c	A	GCT GCC GCA GCG
5	Q	CAA CAG
6	V	GTT CTC GTA GTG

7	**V**	GTT CTC GTA GTG
8	**A**	GCT GCC GCA GCG
9	**I**	ATT ATC ATA
10	**A**	GCT GCC GCA GCG
11a	**N**	AAT AAC
11b	**S**	TCT TCC TCA TCG
12	**Xaa**	Dependent on specificity of TALE unit
13	**Xaa**	Dependent on specificity of TALE unit
14	**G**	GGT GGC GGA GGG
15	**G**	GGT GGC GGA GGG
16	**K**	AAA AAG

17	Q	CAA CAG
18	A	GCT GCC GCA GCG
19	L	CTT CTC CTA CTG
20	E	GAA GAG
21	T	ACT ACC ACA ACG
22	V	GTT CTC GTA GTG
23	Q	CAA CAG
24	R	CGT CGC CGA CGG
25	L	CTT CTC CTA CTG
26	L	CTT CTC CTA CTG
27	P	CCT CCC CCA

		CCG
28	**V**	GTT CTC GTA GTG
29	**L**	CTT CTC CTA CTG
30	**C**	TGT TGC
31	**Q**	CAA CAG
32a	**A**	GCT GCC GCA GCG
32b	**D**	GAT GAC
33	**H**	CAT CAC
34	**G**	GGT GGC GGA GGG

TALE unit encoding nucleic acid sequences prepared according to the sixth aspect of this invention may be modified to include additional sequences. The additional sequences may be included at the 5' and/or 3' ends of any of the TALE unit encoding nucleic acid sequences.

The additional sequences may provide or comprise primer binding sites and/or restriction sites. The additional sequences may also encode or provide parts of the sequences of other TALE units. As described above, the additional sequences may comprise or may further comprise, sequences which encode fusion tags (for purification by, for example, affinity chromatography), leader sequences or other moieties which facilitate cell uptake and/or secretion.

When creating multiple fragments containing combined TALE units, the 5' and/or 3' end of each fragment may contain sequences or additional sequences, which facilitate the ligation and/or joining of the fragments. For example, the additional sequences may comprise sequences which facilitate Gibson assembly, restriction sites and/or primer binding sites.

By way of (non-limiting) example, a TALE molecule encoding nucleic acid sequence may be generated by first assembling or compiling two, three or more fragments for synthesis. The first and any subsequent fragments may be compiled/assembled using, for example, the TALE unit encoding nucleic acid sequences described herein – including those encompassed by the first aspect of this invention. For example, the first fragment may comprise, for example, two or more, for example three, four, five, six, seven, eight or more TALE unit encoding nucleic acid sequences. For example, a fragment comprising seven TALE unit encoding sequences may be compiled as follows:

one TALE unit is selected from the group consisting of TALE unit sequences 1A, 1C, 1G and 1T (as identified in Table 1); the selected unit is combined with:

one TALE unit is selected from the group consisting of TALE unit sequences 2A, 2C, 2G and 2T (as identified in Table 1):

one TALE unit selected from the group consisting of TALE unit sequences 3A, 3C, 3G and 3T (as identified in Table 1);

one TALE unit selected from the group consisting of TALE unit sequences 4A, 4C, 4G and 4T (as identified in Table 1);

one TALE unit selected from the group consisting of TALE unit sequences 5A, 5C, 5G and 5T (as identified in Table 1);

one TALE unit selected from the group consisting of TALE unit sequences 6A, 6C, 6G and 6T (as identified in Table 1); and

one TALE unit selected from the group consisting of 7A, 7C, 7G and 7T (as identified in Table 1):

to yield a sequence encoding a TALE molecule comprising 7 TALE units (this may represent part of a complete TALE molecule comprising, for example, 16 or more TALE units).

As with other methods in this invention, the step of combining may be done computationally and the first (and other) fragment(s) may be assembled or compiled computationally before synthesis.

Each of the TALE unit encoding sequences 1A, 1C, 1G and 1T and/or TALE unit encoding sequences 7A, 7C, 7G and 7T may be modified to comprise additional sequences which facilitate or permit subsequent cloning, ligation, amplification and/or joining protocols. For example, the additional sequence may provide a restriction site and/or primer binding site at the 5' and/or 3' end of the relevant TALE unit sequence.

For example, TALE unit encoding sequence 1A may comprise, consist essentially of or consist of the following sequence (SEQ ID NO: 130):

TCGGACGAGCTGCACCCGCCACTAGCCTATCTAGT**GAAGACAAGAAC**CTTACTCCTGATCAAGTTGTG
GCTATTGCGTCTAATATTGGTGGTAAACAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGTTTTATG
TCAAGATCATGGT

TALE unit encoding sequence 1C may comprise, consist essentially of or consist of the following sequence (SEQ ID NO: 131):

TCGGACGAGCTGCACCCGCCACTAGCCTATCTAGT**GAAGACAAGAAC**CTTACTCCTGATCAAGTTGTG
GCTATTGCGTCTCATGATGGTGGTAAACAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGTTTTATG
TCAAGATCATGGT

TALE unit encoding sequence 1G may comprise, consist essentially of or consist of the following sequence (SEQ ID NO: 132):

TCGGACGAGCTGCACCCGCCACTAGCCTATCTAGT**GAAGACAAGAAC**CTTACTCCTGATCAAGTTGTG
GCTATTGCGTCTAATAACGGTGGTAAACAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGTTTTATG
TCAAGATCATGGT

TALE unit encoding sequence 1T may comprise, consist essentially of or consist of the following sequence (SEQ ID NO: 133):

TCGGACGAGCTGCACCCGCCACTAGCCTATCTAGT**GAAGACAAGAAC**CTTACTCCTGATCAA
GTTGTGGCTATTGCGTCTAATGGAGGTGGTAAACAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGT
TTTATGTCAAGATCATGGT

In each case, a 5' primer binding sequence (underlined) and a (Bbs1) restriction site (bold) are shown. The TALE unit encoding sequence is shown in grey highlight. One of skill

will appreciate that any of the TALE unit encoding sequences may be amended to include this or any other sequence providing primer binding sequences and/or restriction sites. Such sequences may be added to either the 5' and/or 3' ends of any of the TALE unit encoding sequences described herein. The precise technical features of any additional sequence added to the 3' and/or 5' ends of the TALE unit encoding sequences of this invention may depend on the sequence of any primers used in later stages of the methods and/or the type of any restriction enzymes used.

Depending on the size of the first fragment and the number and size of additional fragments to be used, one or more (for example two further) fragments may be compiled and/or assembled from the sequences encoding TALE units 8-16.

For example, a second fragment (to be combined with a first fragment encoding 7 TALE units) may comprise 1, 2, 3, 4, 5, 6, 7, 8 or 9 TALE encoding units. Methods exploiting the assembly of three fragments may exploit a second fragment encoding 3, 4 or 5 TALE units and a third fragment encoding 6, 5 or 4 fragments respectively.

The skilled person will understand that where the TALE molecule is to comprise, for example 16 TALE units, the various fragments will together encode 16 TALE units.

By way of example, a second fragment comprising four TALE unit encoding sequences may be compiled as follows:

one TALE unit is selected from the group consisting of TALE unit sequences 8A, 8C, 8G and 8T (as identified in Table 1); this is combined with:

one TALE unit selected from the group consisting of TALE unit sequences 9A, 9C, 9G and 9T (as identified in Table 1);

one TALE unit selected from the group consisting of TALE unit sequences 10A, 10C, 10G and 10T (as identified in Table 1); and

one TALE unit selected from the group consisting of TALE unit sequences 11A, 11C, 11G and 11T (as identified in Table 1);

to yield a sequence encoding a TALE molecule comprising 4 TALE units (this may represent part of a complete TALE molecule comprising, for example, 16 or more TALE units).

Each of TALE unit encoding sequences 8A, 8C, 8G and 8T may be modified to comprise additional sequences which facilitate or permit subsequent cloning, ligation,

amplification and/or joining protocols. For example, the additional sequence may provide a restriction site and/or primer binding site at the 5' end.

For example, the sequence encoding TALE unit encoding sequence 8G may comprise, consist essentially of or consist of (SEQ ID NO: 134):

TGGCAATCGCGTCGAACGGGGGAGGTAAAC**AAGCTT**TAGAAACCGTTCAGCGTCTCCTCCCAGTGTTA
TGTCAAGACCATGGTCTAACACCAGAGCAGGTGGTGGCGATCGCCAATAATAATGGAGGGAAGCAAGC
TCTGGAAACAGTCCAACGCCTTCTTCCGGTTCTTTGTCAAGATCACGGG

The sequence encoding TALE unit encoding sequence 8A may comprise, consist essentially of or consist of (SEQ ID NO: 135):

TGGCAATCGCGTCGAACGGGGGAGGTAAAC**AAGCTTT**AGAAACCGTTCAGCGTCTCCTCCCAGTGTTA
TGTCAAGACCATGGTCTAACACCAGAGCAGGTGGTGGCGATCGCCAATAATATCGGAGGGAAGCAAGC
TCTGGAAACAGTCCAACGCCTTCTTCCGGTTCTTTGTCAAGATCACGGG

The sequence encoding TALE unit encoding sequence 8T may comprise, consist essentially of or consist of (SEQ ID NO: 136):

TGGCAATCGCGTCGAACGGGGGAGGTAAAC**AAGCTTT**AGAAACCGTTCAGCGTCTCCTCCCAGTGTTA
TGTCAAGACCATGGTCTAACACCAGAGCAGGTGGTGGCGATCGCCAATAACGGTGGAGGGAAGCAAGC
TCTGGAAACAGTCCAACGCCTTCTTCCGGTTCTTTGTCAAGATCACGGG

The sequence encoding TALE unit encoding sequence 8C may comprise, consist essentially of or consist of (SEQ ID NO: 137):

TGGCAATCGCGTCGAACGGGGGAGGTAAAC**AAGCTTT**AGAAACCGTTCAGCGTCTCCTCCCAGTGTTA
TGTCAAGACCATGGTCTAACACCAGAGCAGGTGGTGGCGATCGCCAATCACGACGGAGGGAAGCAAGC
TCTGGAAACAGTCCAACGCCTTCTTCCGGTTCTTTGTCAAGATCACGGG

In each case, a (Hind111) restriction site (bold/underline) is shown with the TALE unit encoding sequence in grey highlight.

Residues 50-55 of the sequences encoding TALE units 7G, 7A, 7T and 7C define a HindIII restriction site (AAGCTT). Thus, HindIII treated fragments comprising sequences encoding TALE unit 7 (G, A, T or C) and sequences encoding TALE unit 8 (G, A, T or C: see above) may be joined or ligated together.

A third fragment comprising five TALE unit encoding sequences may be compiled as follows:

one TALE unit is selected from the group consisting of TALE unit sequences 12A, 12C, 12G and 12T (as identified in Table 1); this is combined with:

one TALE unit selected from the group consisting of TALE unit sequences 13A, 13C, 13G and 13T (as identified in Table 1);

one TALE unit selected from the group consisting of TALE unit sequences 14A, 14C, 14G and 14T (as identified in Table 1);

one TALE unit selected from the group consisting of TALE unit sequences 15A, 15C, 15G and 15T (as identified in Table 1); and

one TALE unit is selected from the group consisting of TALE unit sequences 16A, 16C, 16G and 16T:

to yield a sequence encoding a TALE molecule comprising 5 TALE units (this may represent part of a complete TALE molecule comprising, for example, 16 or more TALE units).

The first (comprising 7 TALE unit encoding sequences), second (comprising 4 TALE unit encoding sequences) and third (comprising 5 TALE unit encoding sequences) fragments may then be individually synthesised and joined (for example ligated) together to provide a complete TALE molecule encoding sequence (comprising 16 TALE units). For example, the synthesised fragments may be joined by ligation protocols and/or Gibson assembly

The sequence encoding TALE unit encoding sequence 12G may comprise, consist essentially of or consist of (SEQ ID NO: 138):

GCGATAGCCTCTCATGACGGTGGGAAGCAGGCG**CTCGAG**ACAGTTCAACGACTACTCCCGGTATTATG
TCAAGATCATGGGCTCACACCCGCCCAGGTTGTAGCAATTGCCTCGAACAACGGCGGCAAGCAAGCAC
TTGAGACTGTCCAGCGGCTCTTGCCAGTTCTCTGCCAGGCACACGGC

The sequence encoding TALE unit encoding sequence 12A may comprise, consist essentially of or consist of (SEQ ID NO: 139):

GCGATAGCCTCTCATGACGGTGGGAAGCAGGCG**CTCGAG**ACAGTTCAACGACTACTCCCGGTATTATG
TCAAGATCATGGGCTCACACCCGCCCAGGTTGTAGCAATTGCCTCGAACATTGGCGGCAAGCAAGCAC
TTGAGACTGTCCAGCGGCTCTTGCCAGTTCTCTGCCAGGCACACGGC

The sequence encoding TALE unit encoding sequence 12T may comprise, consist essentially of or consist of (SEQ ID NO: 140):

GCGATAGCCTCTCATGACGGTGGGAAGCAGGCG**CTCGAG**ACAGTTCAACGACTACTCCCGGTATTATG
TCAAGATCATGGGCTCACACCCGCCCAGGTTGTAGCAATTGCCTCGAACGGTGGCGGCAAGCAAGCAC
TTGAGACTGTCCAGCGGCTCTTGCCAGTTCTCTGCCAGGCACACGGC

The sequence encoding TALE unit encoding sequence 12C may comprise, consist essentially of or consist of (SEQ ID NO: 141):

GCGATAGCCTCTCATGACGGTGGGAAGCAGGCG**CTCGAG**ACAGTTCAACGACTACTCCCGGTATTATG
TCAAGATCATGGGCTCACACCCGCCCAGGTTGTAGCAATTGCCTCGCATGATGGCGGCAAGCAAGCAC
TTGAGACTGTCCAGCGGCTCTTGCCAGTTCTCTGCCAGGCACACGGC

In each case, a (Xho1) restriction site (bold/underline) is shown with the TALE unit encoding sequence in grey highlight.

Residues 55-60 of the sequences encoding TALE units 11G, 11A, 11T and 11C define a Xho restriction site (CTCGAG). Thus, Xho1 treated fragments comprising sequences encoding TALE unit 11 (G, A, T or C) and sequences encoding TALE unit 12 (G, A, T or C: see above) may be joined or ligated together.

Once the required number of fragments have been compiled/assembled and synthesised, they may be treated with the appropriate restriction enzyme so as to yield restricted ends which may be ligated together.

Fragment 1 may comprise seven TALE unit encoding sequences, each encoding one of TALE units 1-7 (with G, A, T or C specificity as required). Fragment 2 may comprise four TALE unit encoding sequences, each encoding one of TALE units 8-11 (with G, A, T or C specificity as required). Fragment 3 may comprise five TALE unit encoding sequences, each encoding one of TALE units 12-116 (with G, A, T or C specificity as required).

Fragments 1 and 2 may be contacted with HindIII. This will yield two fragments with restricted HindIII ends which can be ligated together. Fragment 1 may be further contacted with a restriction enzyme suitable to yield a fragment compatible with a restricted vector.

Fragments 2 and 3 may be contacted with Xho1. This will yield two fragments with restricted XhoI ends which can be ligated together. Fragment 3 may be further contacted with a restriction enzyme suitable to yield a fragment compatible with a restricted vector.

After treatment with the appropriate restriction enzyme, fragments 1, 2 and 3 may be ligated together to provide a TALE molecule encoding sequence. In this example, the TALE molecule comprises (in total) 16 TALE units.

As stated, TALE fragments may additionally or alternatively be joined by Gibson assembly.

Prior to cloning into a vector, the TALE molecule encoding sequence may be contacted with other restriction enzymes in order that it's 5' and 3' ends are rendered compatible with a restricted vector.

A method as substantially described in the description and Figures. For example, one aspect of the invention provides the method detailed in Figure 1 (as Method 1 or Method 2).

A data carrier or digital medium or storage device comprising (containing, carrying or loaded with) the information presented in Table 1.

DETAILED DESCRIPTION

The present invention will now be described in detail with reference to the following Figures which show:

Figure 1: Assembly of AxTALENs The detail of TALEN generation Methods 1 and 2.

Figure 1a: In Method 1 three fragments (F1, F2 and F3) are synthesised to the target sequence. Fragment 1 is digested with Hind III (H), Fragment 2 with Hind III (H) and Xho 1 (X) and Fragment 3 with XhoI (X). Restriction enzymes are heat denatured and equal amounts of the three fragments ligated. DNA ligase is heat denatured and the complete AxTALEN is amplified by PCR with primer gb1 and primer gb2. The resulting amplicon is then TA cloned. Bacterial colonies are picked and plasmids isolated and validated by sequencing. Plasmids are then cut with Bbs1 (Bb) and Bsa1 (B) prior to Golden Gate cloning[12] into BsmB1-digested destination FOK1 endonuclease vector. The entire procedure takes 3 days.

Figure 1b: Single full length AAVS1 AxTALEN R PCR product generated with primer gb1 and primer gb2. M is 1KB plus DNA ladder.

Figure 1c: Western blot analysis using an anti-FLAG antibody confirming expression of full length AAVS1 R AxTALEN protein in transfected 293FT cells. Un-transfected 293FT cells (UT), and cells transfected with Reverse AAVS1 AxTALEN (AR). Loading control, anti GAPDH.

Figure 1d: In Method 2 the three fragments 1, 2 and 3 and the BsmB1 destination FOK1 endonuclease vector are joined by Gibson Assembly in a single step then transformed into bacteria. Colonies are picked, then plasmids isolated and validated by sequencing. This method reduces the time to just 2 days

Figure 1e Western blot analysis using anti-FLAG antibody confirming expression of full length AAVS1 AxTALEN F (AF), OCT4 AxTALEN F and R (OF and OR, respectively) assembled using Method 2. Lysates from un-transfected 293FT cells (UT) and 293FT cells transfected with AF, OF and OR. Loading control, anti GAPDH.

Figure 2: Heterogeneity of TALEs and schematic AxTALE design strategy. Figure 2a) the 34 amino acid TALE repeat showing di-variable residues at position 12 and 13. Underlined NN bind the DNA base G, NI binds A, NG binds T and HD binds C. Alternative amino acid preferences at position 4 (E, D or A), position 11 (N or S) and position 32 (A or D). Figure 2b) the schematic outlines the iterative process of AxTALE design and computer analysis.

Figure 3: GFP-SplitAx, A novel assay for the functional validation of AxTALENs, zinc fingers and CRISPR.

Figure 3a: Schematic of the GFP-SplitAx system. The GFP-SplitAx vector consisting of the N-terminus of GFP (1-157), a genome editing binding site and the C-terminus (158-end) which is out of frame with the N-terminus. GFP-SplitAx vector with its corresponding AxTALENs AF, AR, OF, OR, zinc fingers, CRISPR are co-transfected into 293FT cells. The creation of a double strand break and error prone repair by NHEJ can result in deletions or insertions that generate the full length open reading frame of GFP.

Figures 3b-3g: Representative flow cytometry plots of 293FT cells 48 hours after transfection with AAVS1-GFP-SplitAx only (b), co-transfection of AAVS1-GFP-SplitAx and AF/AR (c), AAVS1-GFP-SplitAx and an AAVS ZFN (d), AAVS1-GFP-SplitAx and AAVS1 CRISPR (e), OCT4-GFP-SplitAx and OF/OR (f) and AAVS1-Zeis Green-SplitAx and AAVS ZFN (g).

Figures 3h-3l: Graphical representation of flow cytometry data for the GFP-SplitAx and Zeis Green-SplitAx with their respective AxTALENs, Zinc Fingers or CRISPR. Graphical

plots show % GFP or % Zeis Green 293FT cells against cells transfected with a plasmid (+), cells not transfected with a plasmid (-). Data shown as +STDev (n=3).

Figure 4: Targeting of AAVS1 and OCT4 loci in 293FT cells using AxTALENs. Figure 2a: Schematic overview of the targeting strategy for the AAVS1 locus. The AAVS1 donor plasmid consists of homology arms left (grey box) and right (yellow box) splice acceptor (SA), self-cleaving peptide (4A), puromycin resistance gene (Puro), polyadenylation sequence (PA), pCAG promoter and a fluorescent reporter Zeis Green. Vector specific (A1, A2) and genomic (A3) PCR primers are indicated. Figure 4b: PCR analysis of genomic DNA isolated from 293FT cells in which the AAVS locus was targeted using the donor plasmid (a) and forward (AF) and reverse (AR) AAVS1 AxTALENs. Primers pairs designed to amplify a fragment within the donor vector (A1/A2) or from the vector to an external sequence (A1/A3) were used to confirm the correct targeting event. Un-transfected cells (UT), targeting vector only (V), 2 independent experiments with targeting vector and AAVS1 AxTALENs (V, AF, AR) and negative water control (-ve). Figure 4c: Schematic overview of the targeting strategy for the OCT4 locus. The OCT4 donor plasmid consists of homology arms left (grey box), right (yellow box), exon 5 in frame with the eGFP reporter, Lox P sites (black triangles) encompass a PGK promoter and puromycin resistance gene (Puro). Vector specific primers O1, O2 and an external genomic primer, O3 are indicated. Figure 4d: PCR analysis of genomic DNA isolated from 293FT OCT4 targeted cells using the donor plasmid and OCT4 AxTALENs OF and OR. Primer pairs designed to amplify within the donor vector (O1/O2) or from the vector to an external sequence (O1/O3) were used to the correct targeting event. 293FT cells transfected with single AxTALEN (OF), Single AxTALEN (OR), Vector only (V), co-transfection of OCT4 targeting vector, AxTALEN OCT4-F and -R (OF, OR). Un-transfected cells (UT) and negative water control (-ve).

Figure 5: Schematic of AAVS1-GFP-SplitAx and validation of SplitAx technology with AAVS1 AxTALENs, Zinc Fingers and CRISPR. Figure 5a). The GFP-SplitAx vector consisting of the N-terminus GFP (1-157), AAVS1 binding site and the C-terminus GFP (158-end) which is out of frame with the N-terminus. Co-transfection of the AAVS1 GFP-SplitAx vector with AAVS1 AxTALEN F and R (rectangle boxes), zinc fingers L and R, CRISPR (T2) into 293FT cells. The creation of a double strand break and error prone repair by NHEJ can result in deletions or insertions that restore the GFP open reading frame of the C-terminus with N-terminus. N and X are Not1 and Xho1 restriction sites that allow the binding site to be exchanged for an alternative binding site. Figures 5b-l) Flow cytometry of 293FT cells at 48 hours post transfection of the AAVS1-GFP-SplitAx vector with AAVS1 AxTALENS, AAVS1 Zinc Fingers and AAVS1 CRISPR.

Figure 6. Schematic of OCT4-GFP-SplitAx and validation of SplitAx technology with OCT4 AxTALENs. 6a). The GFP-SplitAx vector consisting of the N-terminus GFP (1-157), OCT4 binding site and the C-terminus GFP (158-end) which is out of frame with the N-terminus. Co-transfection of the OCT4 -GFP-SplitAx vector with OCT4 AxTALEN F and R (rectangle boxes) into 293FT cells. The creation of a double strand break and error prone repair by NHEJ can result in deletions or insertions that restore the GFP open reading frame of the C-terminus with N-terminus. N and X are Not1 and Xho1 restriction sites that allow the binding site to be exchanged for an alternative binding site. 6b-l) Flow cytometry of 293FT cells at 48 hours post transfection of the OCT4-GFP-SplitAx vector with OCT4 AxTALENS.

Materials and Methods

Computational assembly of AxTALE fragments.

To increase the variability of the TALE DNA sequence, each amino acid codon was varied using the naturally occurring degeneracy (for example Glutamic acid may be coded by GAA or GAG). This step wise approach increased the differences between the naturally occurring TALEs to generate synthetic TALEs which we called AxTALEs. Heterogeneity was further increased by using alternative amino acids at positions 4, 11 and 32 (**Figure 2a**). In total 16 major DNA sequence files were created and then the di-variable repeat for G, A, T and C was included to generate a total of 64 DNA sequence files (see Table 1)

Publicly available ZiFIT software (http://zifit.partners.org/ZiFiT/) was used to identify TALE target sequences for the AAVS1 and OCT4 locus. A computational build was manually generated from the 64 AxTALE files against the ZiFit target to generate 3 separate fragments (fragment 1, 2, and 3). These were chemically generated as Gene Blocks by Integrated DNA Technologies (http://eu.idtdna.com/site).

AxTALEN Method 1

200ng of the AxTALE fragment Oligo was re-suspended in 20ul TE. In separate tubes 30ng of AxTALE Fragment 1 was cut with restriction enzyme Hind111 (Roche), 30 ng AxTALE Fragment 2 was cut with the restriction enzymes Hind111 and Xho1 (Roche) and 30ng of AxTALE Fragment 3 was cut with Xho1 (Roche) for 30 minutes at 37°C. All restriction digests were prepared in 10µl volumes with 10 units of the respective enzyme. The enzymes were heat denatured at 65°C for 20 minutes.

Rapid DNA ligation (NEB) with equal amounts of the restriction cut fragment 1, 2 and 3 was carried out at room temperature for 5 minutes. The ligase was denatured at 94°C for 5 minutes.

PCR with primers gb1 and gb2 was performed with 2µl ligation product and High Fidelity polymerase (Roche) using the following cycling conditions. 2 minutes at 94°C (1 cycle), followed by 15 seconds at 94°C, 30 seconds at 60°C, 1 minute 30 seconds at 68°C (35 Cycles). PCR products were gel purified and TA cloned (Life Sciences). Colonies were picked and the plasmid DNA isolated prior to sequencing (see primer list below). Validated clones were cut with the restriction enzymes Bbs1 and Bsa1 and Golden Gate cloned BsmB1 cut FOK1 endonuclease destination vector (JDS 70, 71, 74 or 78 Joung Addgene). The completed AxTALENs were verified by Asp718i, BamH1 restriction digest and sequencing.

AxTALEN Method 2

200ng of the AxTALE fragment oligo was re-suspended in 20ul TE. 30ng of AxTALE Fragment 1, 30 ng AxTALE Fragment 2 and 30ng of AxTALE Fragment 3 and BsmB1 cut FOK1 endonuclease destination vector (JDS 70, 71, 74 or 78 Joung Addgene) were joined by Gibson Assembly (modified protocol) for 1 hour at 50°C. Bacteria were transformed, plasmid DNA isolated. Full length AxTALENs were verified by Asp718i, BamH1 restriction digest and sequencing (see primer list below).

Primer list

Table 1: List of primers

Sequencing primers

M13 F GTAAAACGACGGCCAG (SEQ ID NO: 142)

M13 R CAGGAAACAGCTATGAC (SEQ ID NO: 143)

JDS2978 TTGAGGCGCTGCTGACTG (SEQ ID NO: 144)

JDS2980 TTAATTCAATATATTCATGAGGCAC (SEQ ID NO: 145)

Gbfrag2_for TGGCAATCGCGTCGAACGGGGGAG (SEQ ID NO: 146)

Gbfrag3_for GCGATAGCCTCTCATGACGGTGGGA (SEQ ID NO: 147)

Gbfrag1rev GAGACGCTGAACGGTTTCTAAAGCT (SEQ ID NO: 148)

Gbfrag2_rev TGGCAATCGCGTCGAACGGGGGAG (SEQ ID NO: 149)

Method 1 PCR primers

gb1 GACGAGCTGCACCCGCCACTAGCCTATC (SEQ ID NO: 150)

gb2 TCATGGCTAACTGCCTTGGTACTGAGC (SEQ ID NO: 151)

AAVS1 genomic targeting PCR assay primers

A1 CCGTCGACGCTCTCTAGAGCTAG (SEQ ID NO: 152)

A2 TCTCCTGGGCTTGCCAAGGACTCAAAC (SEQ ID NO: 153)

A3 CACACCCACACCTGACCCAAACCCAG (SEQ ID NO: 154)

Oct 4 genomic targeting PCR assay primers

O1 CCACTTTGTGGTTCTAAGTACTGTGGTTTC (SEQ ID NO: 155)

O2 GGCAAGAGAAAGCCTGGTAAACCAGCTAC (SEQ ID NO: 156)

O3 AACAGGTAACAGCTACATGGTGACT (SEQ ID NO: 157)

Nucleic acid construct (designated "Split-Ax") for assessment of genome editing system function

The AAVS1-GFP-SplitAx and AAVS1- Zeis GreenSplitAx were generated as a single double stranded DNA oligos (http://eu.idtdna.com/site). 50ng was incubated at 72°C with dNTP and Taq polymerase (Clontech) to add adenine bases for TA cloning (Life Science). Colonies were grown and plasmid DNA extracted and verified by DNA sequence. The GFP-SplitAx was sub cloned by EcoR1 digest into EcoR1 pCAG-ASIP.

OCT4-SplitAx design strategy

OCT4 GFP-SplitAx was made by overlapping PCR using Hi Fidelity Taq polymerase (Roche) TA cloned and then sub-cloned by Not1/Xho1 digest into a pCAG-GFP-SplitAx Not1/Xho1 cut vector

OCT4-SplitAx_for GCGGCCGCGTCACCTGCAGCTGCCCAGACCTGGC (SEQ ID NO: 158)

Not1

Oct4-SplitAx_rev CTCGAGCTGACCCTGCCTGCTCCTCTCCTGGGTGCCAGGTCTGGGC (SEQ ID NO: 159)

Xho1

OCT4-GFP-SplitAx: Amino acid Position 158 marked in redbold underline, followed by Not1 restriction site (under lined). OCT4 genome editing binding site (shaded yellowgrey) followed by Xho1 restriction site (broken line: (SEQ ID NO: 160)).

ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGC
CACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACC
ACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTAC
CCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTC
TTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAG
CTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAAC
GTCTATATCATGGCCGACAAG**CAG**GCGGCCGCGTCACCTGCAGCTGCCCAGACCTGGCACCCAGGAGAGGAGCAG
GCAGGGTCAGCTCGAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCA
GCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAG
CACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGC
CGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAA

Plasmid cloning

The pCAG promoter was cloned into the MCS of the plasmid pZDonor-AAVS1 puromycin (Sigma Aldrich) with EcoRV. The Zeis Green-poly A was cloned into the EcoR1 site pZDonor-AAVS1 puromycin pCAG. Note the orientation of the pCAG cassette is in the opposite direction to AAVS1 (**Figure 5a**).

Transfection Protocols

1000ng of AxTALENs, Zinc Fingers, CRISPR, hCAS9 were transfected (Xfect, Clontech) with 500ng respective SplitAx vector into 293FT cells. Flow cytometry was carried out 48 hours post transfection (BD LSR Fortessa) and analysed with FlowJo data analysis Software.

Genome Targeting of the AAVS1/OCT4 loci

AAVS1-AxTALENs F and R with AAVS1 pZDonor-pCAGASIP-Zeis Green targeting vector and OCT4 AxTALENs F and R with OCT4-eGFP-PGK-PURO targeting vector[1] were introduced into 293 FT cells by Xfect transfection (Clontech). At 72 hours genomic DNA was isolated (using Qiagen DNA extraction kit) for PCR validation assays. Primers used were as follows: For AAVS1, random insertion A1-A2 and gene targeted events A1-A3. For OCT4, random insertion O1-O2 and gene targeted events O1-O3.

Results and discussion

Described herein are novel methods for generating TALEN genome editing systems. As described elsewhere, TALENs are a fusion between a TALE molecule and an endonuclease, for example a FOK1 endonuclease domain. The methods have a minimal number of steps and are significantly less laborious than prior art protocols.

Using the degeneracy of amino acid codons to introduce DNA changes and to systematically reduce the repetitive DNA sequence, sixteen TALEs (TALE 1-16) were generated. To increase the heterogeneity further we introduced additional amino acid changes at position 4, position 11 and or 32. To each TALE 1-16 the DNA sequence that specifies the di-variable residue and binding was added to give TALE1 G, TALE1 A, TALE1 T and TALE1 C. This was repeated for all 16 TALEs. As such, in total 64 different TALE unit encoding sequences (referred to as "TALE sequence files") were generated (See Table 1). These TALE sequence files can be computationally assembled to generate designer TALEs to a desired target DNA sequence. For example, we have designed and manufactured TALENs specific to the AAVS1 and OCT4 locus. The activity and/or function of theee TALENs has also been tested using our novel reporter assay.

Two methods for TALE (targeting the AAVS1 locus) assembly were tested (**Figure 1:** Method 1 and 2). In the first method, a TALE molecule was generated by assembling three fragments (labelled below as F1-3).

F1 (SEQ ID NO: 161)

TCGGACGAGCTGCACCCGCCACTAGCCTATCTAGTGAAGACAAGAACCTTACTCCTGATCAAGTTGTGGCTATTG
CGTCTAATGGAGGTGGTAAACAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGTTTTATGTCAAGATCATGGTC
TTACCCCTGAGCAGGTCGTAGCCATCGCATCCAACGGCGGCGGCAAGCAGGCCCTAGAGACAGTCCAGCGCCTCC
TTCCCGTCTTGTGCCAGGACCACGGCCTAACACCAGCTCAAGTGGTTGCAATAGCCTCAAACGGAGGTGGAAAAC
AAGCACTAGAAACAGTACAGCGACTACTACCAGTATTGTGTCAAGCTCACGGACTGACGCCGGCCCAGGTAGTCG
CGATTGCTAATCATGATGGTGGCAAGCAAGCGCTGGAGACGGTGCAACGGCTGCTGCCGGTGTTATGCCAAGCCC
ATGGGTTGACTCCCGCACAAGTGGTAGCTATAGCTTCCAACGGTGGCGGAAAGCAGGCATTGGAGACTGTACAGA
GATTGCTCCCGGTTCTCTGCCAGGCACACGGTTTAACCCCAGCGCAGGTTGTCGCCATTGCCAATAACAACGGCG

GTAAGCAAGCGTTAGAAACGGTTCAAAGGTTACTGCCTGTATTGTGTCAAGCGCATGGCCTTACCCCTGAACAAG
TCGTGGCAATCGCGTCGAACGGGGGAGGTAAACAAGCTTTAGAAACCGTTCAGCGTCTC

F2 (SEQ ID NO: 162)

TGGCAATCGCGTCGAACGGGGGAGGTAAACAAGCTTTAGAAACCGTTCAGCGTCTCCTCCCAGTGTTATGTCAAG
ACCATGGTCTAACACCAGAGCAGGTGGTGGCGATCGCCAATCACGACGGAGGGAAGCAAGCTCTGGAAACAGTCC
AACGCCTTCTTCCGGTTCTTTGTCAAGATCACGGGCTGACTCCAGATCAAGTTGTTGCCATAGCATCGAACATTG
GTGGCAAACAGGCTCTGGAAACCGTCCAAAGATTACTTCCAGTTTTATGCCAAGCCCACGGTTTGACCCCAGACC
AGGTCGTCGCTATTGCTAACCACGATGGGGGCAAACAAGCCTTGGAGACAGTACAAAGGCTTCTCCCCGTTCTAT
GCCAGGATCACGGCCTTACGCCAGCTCAAGTAGTAGCGATAGCCTCTCATGACGGTGGGAAGCAGGCGCTCGAGA
CAGTTCAACGACTACTCCCGGTATTATGTCAA

F3 (SEQ ID NO: 163)

GCGATAGCCTCTCATGACGGTGGGAAGCAGGCGCTCGAGACAGTTCAACGACTACTCCCGGTATTATGTCAAGAT
CATGGGCTCACACCCGCCCAGGTTGTAGCAATTGCCTCGAACATTGGCGGCAAGCAAGCACTTGAGACTGTCCAG
CGGCTCTTGCCAGTTCTCTGCCAGGCACACGGCCTAACTCCAGCACAAGTCGTTGCTATCGCTAACAACATCGGT
GGCAAACAGGCATTAGAAACCGTTCAACGTCTTTTACCGGTCCTGTGCCAAGCTCACGGCCTGACCCCTGCGCAG
GTTGTAGCGATAGCCAACAATGGAGGCGGTAAGCAAGCCCTGGAAACAGTACAACGTCTTTTGCCTGTGTTGTGC
CAAGCTCATGGTTTGACTCCAGAACAAGTAGTCGCCATCGCCAACCATGATGGAGGTAAACAGGCTTTAGAGACT
GTGCAAAGACTTCTTCCTGTATTATGTCAGGCCCATGGTTTAACGCCAGAGCAGGTTGTTGCAATAGCAAATCAC
GATGGAGGTAAACAAGCGCTCGAAACGGTCCAACGTCTCTTGCCCGTCCTTTGTCAAGCGCACGGACTGAAGAGA
CCGGATCCGTACCCGGCTCAGTACCAAGGCAGTTAGCCATGAAT

After Hind111 and Xho1 restriction, fragments F1, F2 and F3 are then “stitched” together using suitable ligation protocols. The complete ligated sequence is given below (SEQ ID NO: 164).

TCGGACGAGCTGCACCCGCCACTAGCCTATCTAGTGAAGACAAGAACCTTACTCCTGATCAAGTTGTGGCTATTG
CGTCTAATGGAGGTGGTAAACAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGTTTTATGTCAAGATCATGGTC
TTACCCCTGAGCAGGTCGTAGCCATCGCATCCAACGGCGGCGGCAAGCAGGCCCTAGAGACAGTCCAGCGCCTCC
TTCCCGTCTTGTGCCAGGACCACGGCCTAACACCAGCTCAAGTGGTTGCAATAGCCTCAAACGGAGGTGGAAAAC
AAGCACTAGAAACAGTACAGCGACTACTACCAGTATTGTGTCAAGCTCACGGACTGACGCCGGCCCAGGTAGTCG
CGATTGCTAATCATGATGGTGGCAAGCAAGCGCTGGAGACGGTGCAACGGCTGCTGCCGGTGTTATGCCAAGCCC
ATGGGTTGACTCCCGCACAAGTGGTAGCTATAGCTTCCAACGGTGGCGGAAAGCAGGCATTGGAGACTGTACAGA
GATTGCTCCCGGTTCTCTGCCAGGCACACGGTTTAACCCCAGCGCAGGTTGTCGCCATTGCCAATAACAACGGCG
GTAAGCAAGCGTTAGAAACGGTTCAAAGGTTACTGCCTGTATTGTGTCAAGCGCATGGCCTTACCCCTGAACAAG
TCGTGGCAATCGCGTCGAACGGGGGAGGTAAACAAGCTTTAGAAACCGTTCAGCGTCTCCTCCCAGTGTTATGTC
AAGACCATGGTCTAACACCAGAGCAGGTGGTGGCGATCGCCAATCACGACGGAGGGAAGCAAGCTCTGGAAACAG
TCCAACGCCTTCTTCCGGTTCTTTGTCAAGATCACGGGCTGACTCCAGATCAAGTTGTTGCCATAGCATCGAACA

TTGGTGGCAAACAGGCTCTGGAAACCGTCCAAAGATTACTTCCAGTTTTATGCCAAGCCCACGGTTTGACCCCAG
ACCAGGTCGTCGCTATTGCTAACCACGATGGGGGCAAACAAGCCTTGGAGACAGTACAAAGGCTTCTCCCCGTTC
TATGCCAGGATCACGGCCTTACGCCAGCTCAAGTAGTAGCGATAGCCTCTCATGACGGTGGGAAGCAGGCGCTCG
AGACAGTTCAACGACTACTCCCGGTATTATGTCAAGATCATGGGCTCACACCCGCCCAGGTTGTAGCAATTGCCT
CGAACATTGGCGGCAAGCAAGCACTTGAGACTGTCCAGCGGCTCTTGCCAGTTCTCTGCCAGGCACACGGCCTAA
CTCCAGCACAAGTCGTTGCTATCGCTAACAACATCGGTGGCAAACAGGCATTAGAAACCGTTCAACGTCTTTTAC
CGGTCCTGTGCCAAGCTCACGGCCTGACCCCTGCGCAGGTTGTAGCGATAGCCAACAATGGAGGCGGTAAGCAAG
CCCTGGAAACAGTACAACGTCTTTTGCCTGTGTTGTGCCAAGCTCATGGTTTGACTCCAGAACAAGTAGTCGCCA
TCGCCAACCATGATGGAGGTAAACAGGCTTTAGAGACTGTGCAAAGACTTCTTCCTGTATTATGTCAGGCCCATG
GTTTAACGCCAGAGCAGGTTGTTGCAATAGCAAATCACGATGGAGGTAAACAAGCGCTCGAAACGGTCCAACGTC
TCTTGCCCGTCCTTTGTCAAGCGCACGGACTGAAGAGACC**GGATCCGTACCCGGCTCAGTACCAAGGCAGTTAGC**
CATGAAT

AAVS1 has the sequence shown below together with the divariable TALE residues required to ensure TALE specificity.

The AxTALEN-R AAVS1 sequence: T T T C T G T C A C C A A T C C

Di-variable Residue NG NG NG HD NG NN NG HD NI HD HD NI NI NG HD HD

The translated AAVS1 specific TALE is shown below with the di-variable repeats highlighted. Please note the Bbs1 restriction site is shown in bold text and the Bsa1 restriction site as bold underlined text (SEQ ID NO: 165).

tcggacgagctgcacccgccactagcctatctagt**gaagacaagaac**cttactcctgatcaa
G R A A P A T S L S S E D K N L T P D Q
gttgtggctattgcgtctaatggaggtggtaaacaagctcttgaaactgttcaacgtctc
V V A I A S N G G G K Q A L E T V Q R L
ctccctgttttatgtcaagatcatggtcttacccctgagcaggtcgtagccatcgcatcc
L P V L C Q D H G L T P E Q V V A I A S
aacggcggcggcaagcaggccctagagacagtccagcgcctccttcccgtcttgtgccag
N G G G K Q A L E T V Q R L L P V L C Q
gaccacggcctaacaccagctcaagtggttgcaatagcctcaaacggaggtggaaaacaa
D H G L T P A Q V V A I A S N G G G K Q
gcactagaaacagtacagcgactactaccagtattgtgtcaagctcacggactgacgccg
A L E T V Q R L L P V L C Q A H G L T P
gcccaggtagtcgcgattgctaatcatgatggtggcaagcaagcgctggagacggtgcaa
A Q V V A I A N H D G G K Q A L E T V Q
cggctgctgccggtgttatgccaagcccatgggttgactcccgcacaagtggtagctata
R L L P V L C Q A H G L T P A Q V V A I
gcttccaacggtggcggaaagcaggcattggagactgtacagagattgctcccggttctc
A S N G G G K Q A L E T V Q R L L P V L
tgccaggcacacggtttaaccccagcgcaggttgtcgccattgccaataacaacggcggt
C Q A H G L T P A Q V V A I A N N G G

```
aagcaagcgttagaaacggttcaaaggttactgcctgtattgtgtcaagcgcatggcctt
 K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L
acccctgaacaagtcgtggcaatcgcgtcgaacgggggaggtaaacaagctttagaaacc
 T  P  E  Q  V  V  A  I  A  S  N  G  G  G  K  Q  A  L  E  T
gttcagcgtctcctcccagtgttatgtcaagaccatggtctaacaccagagcaggtggtg
 V  Q  R  L  L  P  V  L  C  Q  D  H  G  L  T  P  E  Q  V  V
gcgatcgccaatcacgacggagggaagcaagctctggaaacagtccaacgccttcttccg
 A  I  A  N  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P
gttctttgtcaagatcacgggctgactccagatcaagttgttgccatagcatcgaacatt
 V  L  C  Q  D  H  G  L  T  P  D  Q  V  V  A  I  A  S  N  I
ggtggcaaacaggctctggaaaccgtccaaagattacttccagttttatgccaagcccac
 G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  A  H
ggtttgaccccagaccaggtcgtcgctattgctaaccacgatgggggcaaacaagccttg
 G  L  T  P  D  Q  V  V  A  I  A  N  H  D  G  G  K  Q  A  L
gagacagtacaaaggcttctccccgttctatgccaggatcacggccttacgccagctcaa
 E  T  V  Q  R  L  L  P  V  L  C  Q  D  H  G  L  T  P  A  Q
gtagtagcgatagcctctcatgacggtgggaagcaggcgctcgagacagttcaacgacta
 V  V  A  I  A  S  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L
ctcccggtattatgtcaagatcatgggctcacacccgcccaggttgtagcaattgcctcg
 L  P  V  L  C  Q  D  H  G  L  T  P  A  Q  V  V  A  I  A  S
aacattggcggcaagcaagcacttgagactgtccagcggctcttgccagttctctgccag
 N  I  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q
gcacacggcctaactccagcacaagtcgttgctatcgctaacaacatcggtggcaaacag
 A  H  G  L  T  P  A  Q  V  V  A  I  A  N  N  I  G  G  K  Q
gcattagaaaccgttcaacgtcttttaccggtcctgtgccaagctcacggcctgacccct
 A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P
gcgcaggttgtagcgatagccaacaatggaggcggtaagcaagccctggaaacagtacaa
 A  Q  V  V  A  I  A  N  N  G  G  G  K  Q  A  L  E  T  V  Q
cgtcttttgcctgtgttgtgccaagctcatggtttgactccagaacaagtagtcgccatc
 R  L  L  P  V  L  C  Q  A  H  G  L  T  P  E  Q  V  V  A  I
gccaaccatgatggaggtaaacaggctttagagactgtgcaaagacttcttcctgtatta
 A  N  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L
tgtcaggcccatggtttaacgccagagcaggttgttgcaatagcaaatcacgatggaggt
 C  Q  A  H  G  L  T  P  E  Q  V  V  A  I  A  N  H  D  G  G
aaacaagcgctcgaaacggtccaacgtctcttgcccgtcctttgtcaagcgcacggactg
 K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L
aagagaccggatccgtacccggctcagtaccaaggcagttagccatgaat
 K  R  P  D  P  Y  P  A  Q  Y  Q  G  S  -  P  -
```

PCR using primers, gb1 and gb2 (**Figure 1a**) generated a single TALE amplicon **(Figure 1b)**. Subsequent TA cloning, plasmid isolation and sequencing showed that 10% of the TALE clones were correct. Bbs1/Bsa1restriction digest and Golden Gate Cloning[12] into the BsmB1 digested FOK1 endonuclease destination vector then generated the completed

TALEN (Joung Laboratory Addgene). Expression of the full length TALEN protein was verified in 293FT cells by Western blot analysis using an anti-FLAG antibody (**Figure 1c)**.

The efficiency of TALEN production with the correct sequence using the first method was approximately 10% - this may be due to errors introduced during synthesis of DNA fragments and at subsequent PCR steps; this despite using a high fidelity polymerase.

In an attempt to improve efficiency, we developed and tested an alternative approach (Method 2). In this method TALE fragments were generated with complimentary ends that allowed the joining of fragments and the destination vector by Gibson Assembly[13] in a single step **(Figure 1d)**.

To achieve this, we designed new TALE unit 1, 7, 8, 12 and 16 encoding sequences (with specificity for each of the nucleotides: G A T C). The sequence of each of these modified TALE unit encoding sequences is given below. Please note, these sequences are to be considered as encompassed within the first aspect of this invention.

Sequences 1 (G, A, T and C) and 7 (G, A, T and C) are used to generate Fragment 1. Sequences 8 (G, A, T and C) and 11 (G, A, T and C) are used to generate Fragment 2. Sequences 12 (G, A, T and C) and 16 (G, A, T and C) are used to generate Fragment 3.

1 G (SEQ ID NO: 166)

```
aatgcgctcaccggggccccttgaaccttactcctgatcaagttgtggctattgcgtct
 N  A  L  T  G  A  P  L  N  L  T  P  D  Q  V  V  A  I  A  S
aataacggtggtaaacaagctcttgaaactgttcaacgtctcctccctgttttatgtcaa
 N  N  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q
gatcatggt
 D  H  G
```

1 A (SEQ ID NO: 167)

```
aatgcgctcaccggggccccttgaaccttactcctgatcaagttgtggctattgcgtct
 N  A  L  T  G  A  P  L  N  L  T  P  D  Q  V  V  A  I  A  S
aatattggtggtaaacaagctcttgaaactgttcaacgtctcctccctgttttatgtcaa
 N  I  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q
gatcatggt
 D  H  G
```

1 T (SEQ ID NO: 168)

```
aatgcgctcaccggggccccttgaaccttactcctgatcaagttgtggctattgcgtct
 N  A  L  T  G  A  P  L  N  L  T  P  D  Q  V  V  A  I  A  S
aatggaggtggtaaacaagctcttgaaactgttcaacgtctcctccctgttttatgtcaa
```

N G G G K Q A L E T V Q R L L P V L C Q
gatcatggt
D H G

1 C (SEQ ID NO: 169)

aatgcgctcaccggggccccccttgaaccttactcctgatcaagttgtggctattgcgtct
N A L T G A P L N L T P D Q V V A I A S
catgatggtggtaaacaagctcttgaaactgttcaacgtctcctccctgttttatgtcaa
H D G G K Q A L E T V Q R L L P V L C Q
gatcatggt
D H G

7 G (SEQ ID NO: 170)

cttacccctgaacaagtcgtggcaatcgcgtcgaacaatggaggtaaacaagctttagaa
L T P E Q V V A I A S N N G G K Q A L E
accgttc
T V

7 A (SEQ ID NO: 171)

cttacccctgaacaagtcgtggcaatcgcgtcgaacatcggaggtaaacaagctttagaa
L T P E Q V V A I A S N I G G K Q A L E
accgttc
T V

7 T (SEQ ID NO: 172)

cttacccctgaacaagtcgtggcaatcgcgtcgaacgggggaggtaaacaagctttagaa
L T P E Q V V A I A S N G G G K Q A L E
accgttc
T V

7 C (SEQ ID NO: 173)

cttacccctgaacaagtcgtggcaatcgcgtcgcacgacggaggtaaacaagctttagaa
L T P E Q V V A I A S H D G G K Q A L E
accgttc
T V

8 G (SEQ ID NO: 174)

ggaggtaaacaagctttagaaaccgttcagcgtctcctcccagtgttatgtcaagaccat
G G K Q A L E T V Q R L L P V L C Q D H
ggtctaacaccagagcaggtggtggcgatcgccaataaataatggagggaagcaagctctg

G L T P E Q V V A I A N N N G G K Q A L
gaaacagtccaacgccttcttccggttctttgtcaagatcacggg
E T V Q R L L P V L C Q D H G

8 A (SEQ ID NO: 175)

ggaggtaaacaagctttagaaaccgttcagcgtctcctcccagtgttatgtcaagaccat
G G K Q A L E T V Q R L L P V L C Q D H
ggtctaacaccagagcaggtggtggcgatcgccaataatatcggagggaagcaagctctg
G L T P E Q V V A I A N N I G G K Q A L
gaaacagtccaacgccttcttccggttctttgtcaagatcacggg
E T V Q R L L P V L C Q D H G

8 T (SEQ ID NO: 176)

ggaggtaaacaagctttagaaaccgttcagcgtctcctcccagtgttatgtcaagaccat
G G K Q A L E T V Q R L L P V L C Q D H
ggtctaacaccagagcaggtggtggcgatcgccaataacggtggagggaagcaagctctg
G L T P E Q V V A I A N N G G G K Q A L
gaaacagtccaacgccttcttccggttctttgtcaagatcacggg
E T V Q R L L P V L C Q D H G

8 C (SEQ ID NO: 177)

ggaggtaaacaagctttagaaaccgttcagcgtctcctcccagtgttatgtcaagaccat
G G K Q A L E T V Q R L L P V L C Q D H
ggtctaacaccagagcaggtggtggcgatcgccaatcacgacggagggaagcaagctctg
G L T P E Q V V A I A N H D G G K Q A L
gaaacagtccaacgccttcttccggttctttgtcaagatcacggg
E T V Q R L L P V L C Q D H G

11 G (SEQ ID NO: 178)

cttacgccagctcaagtagtagcgatagcctctaataatggtgggaagcaggcgctcgag
L T P A Q V V A I A S N N G G K Q A L E
acagttcaacgactactc
T V Q R L L

11 A (SEQ ID NO: 179)

cttacgccagctcaagtagtagcgatagcctctaaiataggtgggaagcaggcgctcgag
L T P A Q V V A I A S N I G G K Q A L E
acagttcaacgactactc
T V Q R L L

11 T (SEQ ID NO: 180)

```
cttacgccagctcaagtagtagcgatagcctctaatggaggtgggaagcaggcgctcgag
 L  T  P  A  Q  V  V  A  I  A  S  N  G  G  G  K  Q  A  L  E
acagttcaacgactactc
 T  V  Q  R  L  L
```

11 C (SEQ ID NO: 181)

```
cttacgccagctcaagtagtagcgatagcctctcatgacggtgggaagcaggcgctcgag
 L  T  P  A  Q  V  V  A  I  A  S  H  D  G  G  K  Q  A  L  E
acagttcaacgactactc
 T  V  Q  R  L  L
```

12 G (SEQ ID NO: 182)

```
ggcgctcgagacagttcaacgactactcccggtattatgtcaagatcatgggctcacaccc
  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  D  H  G  L  T  P
gcccaggttgtagcaattgcctcgaacaacggcggcaagcaagcacttgagactgtccag
 A  Q  V  V  A  I  A  S  N  N  G  G  K  Q  A  L  E  T  V  Q
cggctcttgccagttctctgccaggcacacggc
 R  L  L  P  V  L  C  Q  A  H  G
```

12 A (SEQ ID NO: 183)

```
ggcgctcgagacagttcaacgactactcccggtattatgtcaagatcatgggctcacaccc
  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  D  H  G  L  T  P
gcccaggttgtagcaattgcctcgaacattggcggcaagcaagcacttgagactgtccag
 A  Q  V  V  A  I  A  S  N  I  G  G  K  Q  A  L  E  T  V  Q
cggctcttgccagttctctgccaggcacacggc
 R  L  L  P  V  L  C  Q  A  H  G
```

12 T (SEQ ID NO: 184)

```
ggcgctcgagacagttcaacgactactcccggtattatgtcaagatcatgggctcacaccc
  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  D  H  G  L  T  P
gcccaggttgtagcaattgcctcgaacggtggcggcaagcaagcacttgagactgtccag
 A  Q  V  V  A  I  A  S  N  G  G  G  K  Q  A  L  E  T  V  Q
cggctcttgccagttctctgccaggcacacggc
 R  L  L  P  V  L  C  Q  A  H  G
```

12 C (SEQ ID NO: 185)

```
ggcgctcgagacagttcaacgactactcccggtattatgtcaagatcatgggctcacaccc
  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  D  H  G  L  T  P
gcccaggttgtagcaattgcctcgcatgatggcggcaagcaagcacttgagactgtccag
 A  Q  V  V  A  I  A  S  H  D  G  G  K  Q  A  L  E  T  V  Q
```

```
cggctcttgccagttctctgccaggcacacggc
 R  L  L  P  V  L  C  Q  A  H  G
```

16 G (SEQ ID NO: 186)

```
ttaacgccagagcaggttgttgcaatagcaaacaacaacggaggtaaacaagcgctcgaa
 L  T  P  E  Q  V  V  A  I  A  N  N  N  G  G  K  Q  A  L  E
acggtccaacgtctcttgcccgtcctttgtcaagcgcacggactgacacccgaacaggtg
 T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P  E  Q  V
gtcgccattgcttctaa
 V  A  I  A  S
```

16 A (SEQ ID NO: 187)

```
ttaacgccagagcaggttgttgcaatagcaaacaatatcggaggtaaacaagcgctcgaa
 L  T  P  E  Q  V  V  A  I  A  N  N  I  G  G  K  Q  A  L  E
acggtccaacgtctcttgcccgtcctttgtcaagcgcacggactgacacccgaacaggtg
 T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P  E  Q  V
gtcgccattgcttctaa
 V  A  I  A  S
```

16 T (SEQ ID NO: 188)

```
ttaacgccagagcaggttgttgcaatagcaaacaatggtggaggtaaacaagcgctcgaa
 L  T  P  E  Q  V  V  A  I  A  N  N  G  G  G  K  Q  A  L  E
acggtccaacgtctcttgcccgtcctttgtcaagcgcacggactgacacccgaacaggtg
 T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P  E  Q  V
gtcgccattgcttctaa
 V  A  I  A  S
```

16 C (SEQ ID NO: 189)

```
ttaacgccagagcaggttgttgcaatagcaaatcacgatggaggtaaacaagcgctcgaa
 L  T  P  E  Q  V  V  A  I  A  N  H  D  G  G  K  Q  A  L  E
acggtccaacgtctcttgcccgtcctttgtcaagcgcacggactgacacccgaacaggtg
 T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P  E  Q  V
gtcgccattgcttctaa
 V  A  I  A  S
```

The above sequences (and others) were then used to generate a TALE molecule encoding sequence with specificity for the AAVS1 locus

The sequences of Fragment 1, Fragment 2 and Fragment 3 (F1, F2 and F3) as used in assembly Method 2 are shown below. The complimentary overlapping ends for fragment 1, 2, and 3 as required for Gibson Assembly are underlined. The complimentary overlapping

ends for fragment 1 and fragment 3 to join to FOK1 endonuclease destination vector using Gibson Assembly broken underlined.

F1 (SEQ ID NO: 190)

AATGCGCTCACCGGGGCCCCCTTGAACCTTACTCCTGATCAAGTTGTGGCTATTGCGTCTCATGATGGTGGAAAG
CAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGTTTTATGTCAAGATCATGGTCTTACCCCTGAGCAGGTCGTA
GCCATCGCATCCCACGATGGCGGCAAGCAGGCCCTAGAGACAGTCCAGCGCCTCCTTCCCGTCTTGTGCCAGGAC
CACGGCCTAACACCAGCTCAAGTGGTTGCAATAGCCTCACATGATGGTGGAAAACAAGCACTAGAAACAGTACAG
CGACTACTACCAGTATTGTGTCAAGCTCACGGACTGACGCCGGCCCAGGTAGTCGCGATTGCTAATCATGATGGT
GGCAAGCAAGCGCTGGAGACGGTGCAACGGCTGCTGCCGGTGTTATGCCAAGCCCATGGGTTGACTCCCGCACAA
GTGGTAGCTATAGCTTCCAACGGTGGCGGAAAGCAGGCATTGGAGACTGTACAGAGATTGCTCCCGGTTCTCTGC
CAGGCACACGGTTTAACCCCAGCGCAGGTTGTCGCCATTGCCAATCATGATGGCGGTAAGCAAGCGTTAGAAACG
GTTCAAAGGTTACTGCCTGTATTGTGTCAAGCGCATGGCCTTACCCCTGAACAAGTCGTGGCAATCGCGTCGCAC
GACGGAGGTAAACAAGCTTTAGAAACCGTTC

F2 (SEQ ID NO: 191)

GGAGGTAAACAAGCTTTAGAAACCGTTCAGCGTCTCCTCCCAGTGTTATGTCAAGACCATGGTCTAACACCAGAG
CAGGTGGTGGCGATCGCCAATAATATCGGAGGGAAGCAAGCTCTGGAAACAGTCCAACGCCTTCTTCCGGTTCTT
TGTCAAGATCACGGGCTGACTCCAGATCAAGTTGTTGCCATAGCATCGCACGATGGTGGCAAACAGGCTCTGGAA
ACCGTCCAAAGATTACTTCCAGTTTTATGCCAAGCCCACGGTTTGACCCCAGACCAGGTCGTCGCTATTGCTAAC
CACGATGGGGGCAAACAAGCCTTGGAGACAGTACAAAGGCTTCTCCCCGTTCTATGCCAGGATCACGGCCTTACG
CCAGCTCAAGTAGTAGCGATAGCCTCTCATGACGGTGGGAAGCAGGCGCTCGAGACAGTTCAACGACTACTC

F3 (SEQ ID NO: 192)

GGCGCTCGAGACAGTTCAACGACTACTCCCGGTATTATGTCAAGATCATGGGCTCACACCCGCCCAGGTTGTAGC
AATTGCCTCGCATGATGGCGGCAAGCAAGCACTTGAGACTGTCCAGCGGCTCTTGCCAGTTCTCTGCCAGGCACA
CGGCCTAACTCCAGCACAAGTCGTTGCTATCGCTAACAACATCGGTGGCAAACAGGCATTAGAAACCGTTCAACG
TCTTTTACCGGTCCTGTGCCAAGCTCACGGCCTGACCCCTGCGCAGGTTGTAGCGATAGCCAACCATGATGGCGG
TAAGCAAGCCCTGGAAACAGTACAACGTCTACTGCCTGTGTTGTGCCAAGCTCATGGTTTGACTCCAGAACAAGT
AGTCGCCATCGCCAACAATATTGGAGGTAAACAGGCTTTAGAGACTGTGCAAAGACTTCTTCCTGTATTATGTCA
GGCCCATGGTTTAACGCCAGAGCAGGTTGTTGCAATAGCAAACAACAACGGAGGTAAACAAGCGCTCGAAACGGT
CCAACGTCTCTTGCCCGTCCTTTGTCAAGCGCACGGACTGACACCCGAACAGGTGGTCGCCATTGCTTCTAA

Fragments 1, 2 and 3 were then stitched/joined by Gibson Assembly to yield the following sequence. Again, the complementary ends to join to FOK1 endonuclease destination vector by Gibson Assembly underlined (SEQ ID NO: 193).

AATGCGCTCACCGGGGCCCCCTTGAACCTTACTCCTGATCAAGTTGTGGCTATTGCGTCTCATGATGGTGGAAAG
CAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGTTTTATGTCAAGATCATGGTCTTACCCCTGAGCAGGTCGTA
GCCATCGCATCCCACGATGGCGGCAAGCAGGCCCTAGAGACAGTCCAGCGCCTCCTTCCCGTCTTGTGCCAGGAC
CACGGCCTAACACCAGCTCAAGTGGTTGCAATAGCCTCACATGATGGTGGAAAACAAGCACTAGAAACAGTACAG
CGACTACTACCAGTATTGTGTCAAGCTCACGGACTGACGCCGGCCCAGGTAGTCGCGATTGCTAATCATGATGGT
GGCAAGCAAGCGCTGGAGACGGTGCAACGGCTGCTGCCGGTGTTATGCCAAGCCCATGGGTTGACTCCCGCACAA
GTGGTAGCTATAGCTTCCAACGGTGGCGGAAAGCAGGCATTGGAGACTGTACAGAGATTGCTCCCGGTTCTCTGC
CAGGCACACGGTTTAACCCCAGCGCAGGTTGTCGCCATTGCCAATCATGATGGCGGTAAGCAAGCGTTAGAAACG
GTTCAAAGGTTACTGCCTGTATTGTGTCAAGCGCATGGCCTTACCCCTGAACAAGTCGTGGCAATCGCGTCGCAC
GACGGAGGTAAACAAGCTTTAGAAACCGTTCAGCGTCTCCTCCCAGTGTTATGTCAAGACCATGGTCTAACACCA
GAGCAGGTGGTGGCGATCGCCAATAATATCGGAGGGAAGCAAGCTCTGGAAACAGTCCAACGCCTTCTTCCGGTT
CTTTGTCAAGATCACGGGCTGACTCCAGATCAAGTTGTTGCCATAGCATCGCACGATGGTGGCAAACAGGCTCTG
GAAACCGTCCAAAGATTACTTCCAGTTTTATGCCAAGCCCACGGTTTGACCCCAGACCAGGTCGTCGCTATTGCT
AACCACGATGGGGGCAAACAAGCCTTGGAGACAGTACAAAGGCTTCTCCCCGTTCTATGCCAGGATCACGGCCTT
ACGCCAGCTCAAGTAGTAGCGATAGCCTCTCATGACGGTGGGAAGCAGGCGCTCGAGACAGTTCAACGACTACTC
CCGGTATTATGTCAAGATCATGGGCTCACACCCGCCCAGGTTGTAGCAATTGCCTCGCATGATGGCGGCAAGCAA
GCACTTGAGACTGTCCAGCGGCTCTTGCCAGTTCTCTGCCAGGCACACGGCCTAACTCCAGCACAAGTCGTTGCT
ATCGCTAACAACATCGGTGGCAAACAGGCATTAGAAACCGTTCAACGTCTTTTACCGGTCCTGTGCCAAGCTCAC
GGCCTGACCCCTGCGCAGGTTGTAGCGATAGCCAACCATGATGGCGGTAAGCAAGCCCTGGAAACAGTACAACGT
CTACTGCCTGTGTTGTGCCAAGCTCATGGTTTGACTCCAGAACAAGTAGTCGCCATCGCCAACAATATTGGAGGT
AAACAGGCTTTAGAGACTGTGCAAAGACTTCTTCCTGTATTATGTCAGGCCCATGGTTTAACGCCAGAGCAGGTT
GTTGCAATAGCAAACAACAACGGAGGTAAACAAGCGCTCGAAACGGTCCAACGTCTCTTGCCCGTCCTTTGTCAA
GCGCACGGACTGACACCCGAACAGGTGGTCGCCATTGCTTCTAA

Translated AAVS1 specific

AAVS1 has the sequence shown below together with the divariable TALE residues required to ensure TALE specificity.

The AxTALEN-F AAVS1 sequence: C C C C T C C A C C C C A C A G

Di-variable Residue HD HD HD HD NG HD HD NI HD HD HD HD NI HD NI NN

The translated AAVS1 specific TALE is shown below with the di-variable repeats highlighted (SEQ ID NO: 194).

aatgcgctcaccggggccccttgaaccttactcctgatcaagttgtggctattgcgtct
N A L T G A P L N L T P D Q V V A I A S
catgatggtggaaagcaagctcttgaaactgttcaacgtctcctccctgttttatgtcaa
H D G G K Q A L E T V Q R L L P V L C Q
gatcatggtcttacccctgagcaggtcgtagccatcgcatcccacgatggcggcaagcag
D H G L T P E Q V V A I A S H D G G K Q

```
gccctagagacagtccagcgcctccttcccgtcttgtgccaggaccacggcctaacacca
 A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  D  H  G  L  T  P
gctcaagtggttgcaatagcctcacatgatggtggaaaacaagcactagaaacagtacag
 A  Q  V  V  A  I  A  S  H  D  G  G  K  Q  A  L  E  T  V  Q
cgactactaccagtattgtgtcaagctcacggactgacgccggcccaggtagtcgcgatt
 R  L  L  P  V  L  C  Q  A  H  G  L  T  P  A  Q  V  V  A  I
gctaatcatgatggtggcaagcaagcgctggagacggtgcaacggctgctgccggtgtta
 A  N  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L
tgccaagcccatgggttgactcccgcacaagtggtagctatagcttccaacggtggcgga
 C  Q  A  H  G  L  T  P  A  Q  V  V  A  I  A  S  N  G  G  G
aagcaggcattggagactgtacagagattgctcccggttctctgccaggcacacggttta
 K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L
accccagcgcaggttgtcgccattgccaatcatgatggcggtaagcaagcgttagaaacg
 T  P  A  Q  V  V  A  I  A  N  H  D  G  G  K  Q  A  L  E  T
gttcaaaggttactgcctgtattgtgtcaagcgcatggccttacccctgaacaagtcgtg
 V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P  E  Q  V  V
gcaatcgcgtcgcacgacggaggtaaacaagctttagaaaccgttcagcgtctcctccca
 A  I  A  S  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P
gtgttatgtcaagaccatggtctaacaccagagcaggtggtggcgatcgccaataatatc
 V  L  C  Q  D  H  G  L  T  P  E  Q  V  V  A  I  A  N  N  I
ggagggaagcaagctctggaaacagtccaacgccttcttccggttctttgtcaagatcac
 G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  D  H
gggctgactccagatcaagttgttgccatagcatcgcacgatggtggcaaacaggctctg
 G  L  T  P  D  Q  V  V  A  I  A  S  H  D  G  G  K  Q  A  L
gaaaccgtccaaagattacttccagttttatgccaagcccacggtttgaccccagaccag
 E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P  D  Q
gtcgtcgctattgctaaccacgatgggggcaaacaagccttggagacagtacaaaggctt
 V  V  A  I  A  N  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L
ctccccgttctatgccaggatcacggccttacgccagctcaagtagtagcgatagcctct
 L  P  V  L  C  Q  D  H  G  L  T  P  A  Q  V  V  A  I  A  S
catgacggtgggaagcaggcgctcgagacagttcaacgactactcccggtattatgtcaa
 H  D  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q
gatcatgggctcacacccgcccaggttgtagcaattgcctcgcatgatggcggcaagcaa
 D  H  G  L  T  P  A  Q  V  V  A  I  A  S  H  D  G  G  K  Q
gcacttgagactgtccagcggctcttgccagttctctgccaggcacacggcctaactcca
 A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P
gcacaagtcgttgctatcgctaacaacatcggtggcaaacaggcattagaaaccgttcaa
 A  Q  V  V  A  I  A  N  N  I  G  G  K  Q  A  L  E  T  V  Q
cgtcttttaccggtcctgtgccaagctcacggcctgacccctgcgcaggttgtagcgata
 R  L  L  P  V  L  C  Q  A  H  G  L  T  P  A  Q  V  V  A  I
gccaaccatgatggcggtaagcaagccctggaaacagtacaacgtctactgcctgtgttg
 A  N  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L
tgccaagctcatggtttgactccagaacaagtagtcgccatcgccaacaatattggaggt
 C  Q  A  H  G  L  T  P  E  Q  V  V  A  I  A  N  N  I  G  G
aaacaggctttagagactgtgcaaagacttcttcctgtattatgtcaggcccatggttta
```

K Q A L E T V Q R L L P V L C Q A H G L
acgccagagcaggttgttgcaatagcaaacaacaacggaggtaaacaagcgctcgaaacg
T P E Q V V A I A N N N G G K Q A L E T
gtccaacgtctcttgcccgtcctttgtcaagcgcacggactgacacccgaacaggtggtc
V Q R L L P V L C Q A H G L T P E Q V V
gccattgcttctaa
A I A S

Subsequent transformation, isolation of plasmid DNA and restriction digestion with Asp718i and BamH1 confirmed full length assembled TALENs in the destination vector. Full length plasmids were validated by sequencing and a significant improvement in accuracy was observed: 30% had the correct sequence using Method 2 compared to 10% using Method 1.

Furthermore the single step assembly results in a faster protocol, reducing the time required from 3 days (Method 1) to 2 days (Method 2).

We also assembled TALENs specific to the OCT 4 locus using Method 2. The various sequences used are shown below.

Synthetic OCT4 AxTALEN Forward.

The sequence of Fragment 1, Fragment 2 and Fragment 3 (F1, F2 and F3) for use in assembly Method 2. Complimentary overlapping ends for Gibson Assembly are underlined

F1 (SEQ ID NO: 195)

AATGCGCTCACCGGGGCCCCCTTGAACCTTACTCCTGATCAAGTTGTGGCTATTGCGTCTCATGATGGTGGAAAG
CAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGTTTTATGTCAAGATCATGGTCTTACCCCTGAGCAGGTCGTA
GCCATCGCATCCAACATAGGCGGCAAGCAGGCCCTAGAGACAGTCCAGCGCCTCCTTCCCGTCTTGTGCCAGGAC
CACGGCCTAACACCAGCTCAAGTGGTTGCAATAGCCTCACATGATGGTGGAAAACAAGCACTAGAAACAGTACAG
CGACTACTACCAGTATTGTGTCAAGCTCACGGACTGACGCCGGCCCAGGTAGTCGCGATTGCTAATCATGATGGT
GGCAAGCAAGCGCTGGAGACGGTGCAACGGCTGCTGCCGGTGTTATGCCAAGCCCATGGGTTGACTCCCGCACAA
GTGGTAGCTATAGCTTCCAACGGTGGCGGAAAGCAGGCATTGGAGACTGTACAGAGATTGCTCCCGGTTCTCTGC
CAGGCACACGGTTTAACCCCAGCGCAGGTTGTCGCCATTGCCAATAACAACGGCGGTAAGCAAGCGTTAGAAACG
GTTCAAAGGTTACTGCCTGTATTGTGTCAAGCGCATGGCCTTACCCCTGAACAAGTCGTGGCAATCGCGTCGCAC
GACGGAGGTAAACAAGCTTTAGAAACCGTTC

F2 (SEQ ID NO: 196)

GGAGGTAAACAAGCTTTAGAAACCGTTCAGCGTCTCCTCCCAGTGTTATGTCAAGACCATGGTCTAACACCAGAG
CAGGTGGTGGCGATCGCCAATAATATCGGAGGGAAGCAAGCTCTGGAAACAGTCCAACGCCTTCTTCCGGTTCTT
TGTCAAGATCACGGGCTGACTCCAGATCAAGTTGTTGCCATAGCATCGAACAATGGTGGCAAACAGGCTCTGGAA
ACCGTCCAAAGATTACTTCCAGTTTTATGCCAAGCCCACGGTTTGACCCCAGACCAGGTCGTCGCTATTGCTAAC

CACGATGGGGGCAAACAAGCCTTGGAGACAGTACAAAGGCTTCTCCCCGTTCTATGCCAGGATCACGGCCTTACG
CCAGCTCAAGTAGTAGCGATAGCCTCTAATGGAGGTGGGAAGCAGGCGCTCGAGACAGTTCAACGACTACTC

F3 (SEQ ID NO: 197)

GGCGCTCGAGACAGTTCAACGACTACTCCCGGTATTATGTCAAGATCATGGGCTCACACCCGCCCAGGTTGTAGC
AATTGCCTCGAACAACGGCGGCAAGCAAGCACTTGAGACTGTCCAGCGGCTCTTGCCAGTTCTCTGCCAGGCACA
CGGCCTAACTCCAGCACAAGTCGTTGCTATCGCTAACCACGACGGTGGCAAACAGGCATTAGAAACCGTTCAACG
TCTTTTACCGGTCCTGTGCCAAGCTCACGGCTTGACTCCCGCACAAGTGGTAGCTATAGCTTCCCATGATGGCGG
AAAGCAGGCATTGGAGACTGTACAGAGATTGCTCCCGGTTCTCTGCCAGGCACACGGTTTGACTCCAGAACAAGT
AGTCGCCATCGCCAACCATGATGGAGGTAAACAGGCTTTAGAGACTGTGCAAAGACTTCTTCCTGTATTATGTCA
GGCCCATGGTTTAACGCCAGAGCAGGTTGTTGCAATAGCAAACAATATCGGAGGTAAACAAGCGCTCGAAACGGT
CCAACGTCTCTTGCCCGTCCTTTGTCAAGCGCACGGACTGACACCCGAACAGGTGGTCGCCATTGCTTCTAA

The sequence of the OCT4 specific TALEN (AxTALEN-F) generated after stitching of fragments 1, 2 and 3 by Gibson Assembly. Complementary ends to join to FOK1 endonuclease destination vector by Gibson Assembly underlined (SEQ ID NO: 198).

AATGCGCTCACCGGGGCCCCCTTGAACCTTACTCCTGATCAAGTTGTGGCTATTGCGTCTCATGATGGTGGAAAG
CAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGTTTTATGTCAAGATCATGGTCTTACCCCTGAGCAGGTCGTA
GCCATCGCATCCAACATAGGCGGCAAGCAGGCCCTAGAGACAGTCCAGCGCCTCCTTCCCGTCTTGTGCCAGGAC
CACGGCCTAACACCAGCTCAAGTGGTTGCAATAGCCTCACATGATGGTGGAAAACAAGCACTAGAAACAGTACAG
CGACTACTACCAGTATTGTGTCAAGCTCACGGACTGACGCCGGCCCAGGTAGTCGCGATTGCTAATCATGATGGT
GGCAAGCAAGCGCTGGAGACGGTGCAACGGCTGCTGCCGGTGTTATGCCAAGCCCATGGGTTGACTCCCGCACAA
GTGGTAGCTATAGCTTCCAACGGTGGCGGAAAGCAGGCATTGGAGACTGTACAGAGATTGCTCCCGGTTCTCTGC
CAGGCACACGGTTTAACCCCAGCGCAGGTTGTCGCCATTGCCAATAACAACGGCGGTAAGCAAGCGTTAGAAACG
GTTCAAAGGTTACTGCCTGTATTGTGTCAAGCGCATGGCCTTACCCCTGAACAAGTCGTGGCAATCGCGTCGCAC
GACGGAGGTAAACAAGCTTTAGAAACCGTTCAGCGTCTCCTCCCAGTGTTATGTCAAGACCATGGTCTAACACCA
GAGCAGGTGGTGGCGATCGCCAATAATATCGGAGGGAAGCAAGCTCTGGAAACAGTCCAACGCCTTCTTCCGGTT
CTTTGTCAAGATCACGGGCTGACTCCAGATCAAGTTGTTGCCATAGCATCGAACAATGGTGGCAAACAGGCTCTG
GAAACCGTCCAAAGATTACTTCCAGTTTTATGCCAAGCCCACGGTTTGACCCCAGACCAGGTCGTCGCTATTGCT
AACCACGATGGGGGCAAACAAGCCTTGGAGACAGTACAAAGGCTTCTCCCCGTTCTATGCCAGGATCACGGCCTT
ACGCCAGCTCAAGTAGTAGCGATAGCCTCTAATGGAGGTGGGAAGCAGGCGCTCGAGACAGTTCAACGACTACTC
CCGGTATTATGTCAAGATCATGGGCTCACACCCGCCCAGGTTGTAGCAATTGCCTCGAACAACGGCGGCAAGCAA
GCACTTGAGACTGTCCAGCGGCTCTTGCCAGTTCTCTGCCAGGCACACGGCCTAACTCCAGCACAAGTCGTTGCT
ATCGCTAACCACGACGGTGGCAAACAGGCATTAGAAACCGTTCAACGTCTTTTACCGGTCCTGTGCCAAGCTCAC
GGCTTGACTCCCGCACAAGTGGTAGCTATAGCTTCCCATGATGGCGGAAAGCAGGCATTGGAGACTGTACAGAGA
TTGCTCCCGGTTCTCTGCCAGGCACACGGTTTGACTCCAGAACAAGTAGTCGCCATCGCCAACCATGATGGAGGT
AAACAGGCTTTAGAGACTGTGCAAAGACTTCTTCCTGTATTATGTCAGGCCCATGGTTTAACGCCAGAGCAGGTT

GTTGCAATAGCAAACAATATCGGAGGTAAACAAGCGCTCGAAACGGTCCAACGTCTCTTGCCCGTCCTTTGTCAA
GCGCACGGACTGACACCCGAACAGGTGGTCGCCATTGCTTCTAA

OCT4 has the sequence shown below together with the divariable TALE residues required to ensure TALE specificity.

The OCT4-F sequence:	C	A	C	C	T	G	C	A	G	C	T	G	C	C	C	A
Di-variable Residue	HD	NI	HD	HD	NG	NN	HD	NI	NN	HD	NG	NN	HD	HD	HD	NI

The translated OCT4 specific TALE is shown below with the di-variable repeats highlighted (SEQ ID NO: 199).

```
aatgcgctcaccggggccccttgaaccttactcctgatcaagttgtggctattgcgtct
 N  A  L  T  G  A  P  L  N  L  T  P  D  Q  V  V  A  I  A  S
catgatggtggaaagcaagctcttgaaactgttcaacgtctcctccctgttttatgtcaa
 H  D  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q
gatcatggtcttacccctgagcaggtcgtagccatcgcatccaacataggcggcaagcag
 D  H  G  L  T  P  E  Q  V  V  A  I  A  S  N  I  G  G  K  Q
gccctagagacagtccagcgcctccttcccgtcttgtgccaggaccacggcctaacacca
 A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  D  H  G  L  T  P
gctcaagtggttgcaatagcctcacatgatggtggaaaacaagcactagaaacagtacag
 A  Q  V  V  A  I  A  S  H  D  G  G  K  Q  A  L  E  T  V  Q
cgactactaccagtattgtgtcaagctcacggactgacgccggcccaggtagtcgcgatt
 R  L  L  P  V  L  C  Q  A  H  G  L  T  P  A  Q  V  V  A  I
gctaatcatgatggtggcaagcaagcgctggagacggtgcaacggctgctgccggtgtta
 A  N  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L
tgccaagcccatgggttgactcccgcacaagtggtagctatagcttccaacggtggcgga
 C  Q  A  H  G  L  T  P  A  Q  V  V  A  I  A  S  N  G  G  G
aagcaggcattggagactgtacagagattgctcccggttctctgccaggcacacggttta
 K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L
accccagcgcaggttgtcgccattgccaataacaacggcggtaagcaagcgttagaaacg
 T  P  A  Q  V  V  A  I  A  N  N  N  G  G  K  Q  A  L  E  T
gttcaaaggttactgcctgtattgtgtcaagcgcatggccttacccctgaacaagtcgtg
 V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P  E  Q  V  V
gcaatcgcgtcgcacgacggaggtaaacaagctttagaaaccgttcagcgtctcctccca
 A  I  A  S  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P
gtgttatgtcaagaccatggtctaacaccagagcaggtggtggcgatcgccaataatatc
 V  L  C  Q  D  H  G  L  T  P  E  Q  V  V  A  I  A  N  N  I
ggagggaagcaagctctggaaacagtccaacgccttcttccggttctttgtcaagatcac
 G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  D  H
gggctgactccagatcaagttgttgccatagcatcgaacaatggtggcaaacaggctctg
 G  L  T  P  D  Q  V  V  A  I  A  S  N  N  G  G  K  Q  A  L
```

```
gaaaccgtccaaagattacttccagttttatgccaagcccacggtttgaccccagaccag
 E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P  D  Q
gtcgtcgctattgctaaccacgatgggggcaaacaagccttggagacagtacaaaggctt
 V  V  A  I  A  N  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L
ctccccgttctatgccaggatcacggccttacgccagctcaagtagtagcgatagcctct
 L  P  V  L  C  Q  D  H  G  L  T  P  A  Q  V  V  A  I  A  S
aatggaggtgggaagcaggcgctcgagacagttcaacgactactcccggtattatgtcaa
 N  G  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q
gatcatgggctcacacccgcccaggttgtagcaattgcctcgaacaacggcggcaagcaa
 D  H  G  L  T  P  A  Q  V  V  A  I  A  S  N  N  G  G  K  Q
gcacttgagactgtccagcggctcttgccagttctctgccaggcacacggcctaactcca
 A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P
gcacaagtcgttgctatcgctaaccacgacggtggcaaacaggcattagaaaccgttcaa
 A  Q  V  V  A  I  A  N  H  D  G  G  K  Q  A  L  E  T  V  Q
cgtcttttaccggtcctgtgccaagctcacggcttgactcccgcacaagtggtagctata
 R  L  L  P  V  L  C  Q  A  H  G  L  T  P  A  Q  V  V  A  I
gcttcccatgatggcggaaagcaggcattggagactgtacagagattgctcccggttctc
 A  S  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L
tgccaggcacacggtttgactccagaacaagtagtcgccatcgccaaccatgatggaggt
 C  Q  A  H  G  L  T  P  E  Q  V  V  A  I  A  N  H  D  G  G
aaacaggctttagagactgtgcaaagacttcttcctgtattatgtcaggcccatggttta
 K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L
acgccagagcaggttgttgcaatagcaaacaatatcggaggtaaacaagcgctcgaaacg
 T  P  E  Q  V  V  A  I  A  N  N  I  G  G  K  Q  A  L  E  T
gtccaacgtctcttgcccgtcctttgtcaagcgcacggactgacacccgaacaggtggtc
 V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P  E  Q  V  V
gccattgcttctaa
 A  I  A  S
```

Synthetic OCT4 AxTALEN Reverse.

The sequence of fragments 1, 2 and 3 (F1, F2 and F3) as used in assembly Method 2 are shown below.

F1 (SEQ ID NO: 200)

AATGCGCTCACCGGGGCCCCCTTGAACCTTACTCCTGATCAAGTTGTGGCTATTGCGTCTAATAACGGTGGAAAG
CAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGTTTTATGTCAAGATCATGGTCTTACCCCTGAGCAGGTCGTA
GCCATCGCATCCAACATAGGCGGCAAGCAGGCCCTAGAGACAGTCCAGCGCCTCCTTCCCGTCTTGTGCCAGGAC
CACGGCCTAACACCAGCTCAAGTGGTTGCAATAGCCTCACATGATGGTGGAAAACAAGCACTAGAAACAGTACAG
CGACTACTACCAGTATTGTGTCAAGCTCACGGACTGACGCCGGCCCAGGTAGTCGCGATTGCTAATCATGATGGT
GGCAAGCAAGCGCTGGAGACGGTGCAACGGCTGCTGCCGGTGTTATGCCAAGCCCATGGGTTGACTCCCGCACAA
GTGGTAGCTATAGCTTCCCATGATGGCGGAAAGCAGGCATTGGAGACTGTACAGAGATTGCTCCCGGTTCTCTGC
CAGGCACACGGTTTAACCCCAGCGCAGGTTGTCGCCATTGCCAATAACGGTGGCGGTAAGCAAGCGTTAGAAACG

GTTCAAAGGTTACTGCCTGTATTGTGTCAAGCGCATGGCCTTACCCCTGAACAAGTCGTGGCAATCGCGTCGAAC
AATGGAGGTAAACAAGCTTTAGAAACCGTTC

F2 (SEQ ID NO: 201)

GGAGGTAAACAAGCTTTAGAAACCGTTCAGCGTCTCCTCCCAGTGTTATGTCAAGACCATGGTCTAACACCAGAG
CAGGTGGTGGCGATCGCCAATCACGACGGAGGGAAGCAAGCTCTGGAAACAGTCCAACGCCTTCTTCCGGTTCTT
TGTCAAGATCACGGGCTGACTCCAGATCAAGTTGTTGCCATAGCATCGCACGATGGTGGCAAACAGGCTCTGGAA
ACCGTCCAAAGATTACTTCCAGTTTTATGCCAAGCCCACGGTTTGACCCCAGACCAGGTCGTCGCTATTGCTAAC
AATGGCGGGGGCAAACAAGCCTTGGAGACAGTACAAAGGCTTCTCCCCGTTCTATGCCAGGATCACGGCCTTACG
CCAGCTCAAGTAGTAGCGATAGCCTCTAATAATGGTGGGAAGCAGGCGCTCGAGACAGTTCAACGACTACTC

F3 (SEQ ID NO: 202)

GGCGCTCGAGACAGTTCAACGACTACTCCCGGTATTATGTCAAGATCATGGGCTCACACCCGCCCAGGTTGTAGC
AATTGCCTCGCATGATGGCGGCAAGCAAGCACTTGAGACTGTCCAGCGGCTCTTGCCAGTTCTCTGCCAGGCACA
CGGCCTAACTCCAGCACAAGTCGTTGCTATCGCTAACAACGGTGGTGGCAAACAGGCATTAGAAACCGTTCAACG
TCTTTTACCGGTCCTGTGCCAAGCTCACGGCTTGACTCCCGCACAAGTGGTAGCTATAGCTTCCCATGATGGCGG
AAAGCAGGCATTGGAGACTGTACAGAGATTGCTCCCGGTTCTCTGCCAGGCACACGGTTTAACGCCAGAGCAGGT
TGTTGCAATAGCAAATCACGATGGAGGTAAACAAGCGCTCGAAACGGTCCAACGTCTCTTGCCCGTCCTTTGTCA
AGCGCACGGACTGACACCCGAACAGGTGGTCGCCATTGCTTCTAA

The sequence of the OCT4 specific TALEN (AxTALEN-R) generated after stitching of fragments 1, 2 and 3 by Gibson Assembly. Complementary ends to join to FOK1 endonuclease destination vector by Gibson Assembly underlined (SEQ ID NO: 203).

AATGCGCTCACCGGGGCCCCCTTGAACCTTACTCCTGATCAAGTTGTGGCTATTGCGTCTAATAACGGTGGAAAG
CAAGCTCTTGAAACTGTTCAACGTCTCCTCCCTGTTTTATGTCAAGATCATGGTCTTACCCCTGAGCAGGTCGTA
GCCATCGCATCCAACATAGGCGGCAAGCAGGCCCTAGAGACAGTCCAGCGCCTCCTTCCCGTCTTGTGCCAGGAC
CACGGCCTAACACCAGCTCAAGTGGTTGCAATAGCCTCACATGATGGTGGAAAACAAGCACTAGAAACAGTACAG
CGACTACTACCAGTATTGTGTCAAGCTCACGGACTGACGCCGGCCCAGGTAGTCGCGATTGCTAATCATGATGGT
GGCAAGCAAGCGCTGGAGACGGTGCAACGGCTGCTGCCGGTGTTATGCCAAGCCCATGGGTTGACTCCCGCACAA
GTGGTAGCTATAGCTTCCCATGATGGCGGAAAGCAGGCATTGGAGACTGTACAGAGATTGCTCCCGGTTCTCTGC
CAGGCACACGGTTTAACCCCAGCGCAGGTTGTCGCCATTGCCAATAACGGTGGCGGTAAGCAAGCGTTAGAAACG
GTTCAAAGGTTACTGCCTGTATTGTGTCAAGCGCATGGCCTTACCCCTGAACAAGTCGTGGCAATCGCGTCGAAC
AATGGAGGTAAACAAGCTTTAGAAACCGTTCAGCGTCTCCTCCCAGTGTTATGTCAAGACCATGGTCTAACACCA
GAGCAGGTGGTGGCGATCGCCAATCACGACGGAGGGAAGCAAGCTCTGGAAACAGTCCAACGCCTTCTTCCGGTT
CTTTGTCAAGATCACGGGCTGACTCCAGATCAAGTTGTTGCCATAGCATCGCACGATGGTGGCAAACAGGCTCTG
GAAACCGTCCAAAGATTACTTCCAGTTTTATGCCAAGCCCACGGTTTGACCCCAGACCAGGTCGTCGCTATTGCT
AACAATGGCGGGGGCAAACAAGCCTTGGAGACAGTACAAAGGCTTCTCCCCGTTCTATGCCAGGATCACGGCCTT
ACGCCAGCTCAAGTAGTAGCGATAGCCTCTAATAATGGTGGGAAGCAGGCGCTCGAGACAGTTCAACGACTACTC

CCGGTATTATGTCAAGATCATGGGCTCACACCCGCCCAGGTTGTAGCAATTGCCTCGCATGATGGCGGCAAGCAA
GCACTTGAGACTGTCCAGCGGCTCTTGCCAGTTCTCTGCCAGGCACACGGCCTAACTCCAGCACAAGTCGTTGCT
ATCGCTAACAACGGTGGTGGCAAACAGGCATTAGAAACCGTTCAACGTCTTTTACCGGTCCTGTGCCAAGCTCAC
GGCTTGACTCCCGCACAAGTGGTAGCTATAGCTTCCCATGATGGCGGAAAGCAGGCATTGGAGACTGTACAGAGA
TTGCTCCCGGTTCTCTGCCAGGCACACGGTTTAACGCCAGAGCAGGTTGTTGCAATAGCAAATCACGATGGAGGT
AAACAAGCGCTCGAAACGGTCCAACGTCTCTTGCCCGTCCTTTGTCAAGCGCACGGACTGACACCCGAACAGGTG
GTCGCCATTGCTTCTAA

OCT4 has the sequence shown below together with the divariable TALE residues required to ensure TALE specificity.

The OCT4-R sequence:	G	A	C	C	C	T	G	C	C	T	G	C	T	C	C
Di-variable Residue	NN	NI	HD	HD	HD	NG	NN	HD	HD	NG	NN	HD	NG	HD	HD

The translated OCT4 specific TALE is shown below with the di-variable repeats highlighted (SEQ ID NO: 204).

aatgcgctcaccggggccccctt gaaccttactcctgatcaagttgtggctattgcgtct
N A L T G A P L N L T P D Q V V A I A S
aataacggtggaaagcaagctcttgaaactgttcaacgtctcctccctgttttatgtcaa
N N G G K Q A L E T V Q R L L P V L C Q
gatcatggtcttacccctgagcaggtcgtagccatcgcatccaacataggcggcaagcag
D H G L T P E Q V V A I A S N I G G K Q
gccctagagacagtccagcgcctccttcccgtcttgtgccaggaccacggcctaacacca
A L E T V Q R L L P V L C Q D H G L T P
gctcaagtggttgcaatagcctcacatgatggtggaaaacaagcactagaaacagtacag
A Q V V A I A S H D G G K Q A L E T V Q
cgactactaccagtattgtgtcaagctcacggactgacgccggcccaggtagtcgcgatt
R L L P V L C Q A H G L T P A Q V V A I
gctaatcatgatggtggcaagcaagcgctggagacggtgcaacggctgctgccggtgtta
A N H D G G K Q A L E T V Q R L L P V L
tgccaagcccatgggttgactcccgcacaagtggtagctatagcttcccatgatggcgga
C Q A H G L T P A Q V V A I A S H D G G
aagcaggcattggagactgtacagagattgctcccggttctctgccaggcacacggttta
K Q A L E T V Q R L L P V L C Q A H G L
accccagcgcaggttgtcgccattgccaataacggtggcggtaagcaagcgttagaaacg
T P A Q V V A I A N N G G G K Q A L E T
gttcaaaggttactgcctgtattgtgtcaagcgcatggccttacccctgaacaagtcgtg
V Q R L L P V L C Q A H G L T P E Q V V

```
gcaatcgcgtcgaacaatggaggtaaacaagctttagaaaccgttcagcgtctcctccca
 A  I  A  S  N  N  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P
gtgttatgtcaagaccatggtctaacaccagagcaggtggtggcgatcgccaatcacgac
 V  L  C  Q  D  H  G  L  T  P  E  Q  V  V  A  I  A  N  H  D
ggagggaagcaagctctggaaacagtccaacgccttcttccggttctttgtcaagatcac
 G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  D  H
gggctgactccagatcaagttgttgccatagcatcgcacgatggtggcaaacaggctctg
 G  L  T  P  D  Q  V  V  A  I  A  S  H  D  G  G  K  Q  A  L
gaaaccgtccaaagattacttccagttttatgccaagcccacggtttgaccccagaccag
 E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P  D  Q
gtcgtcgctattgctaacaatggcgggggcaaacaagccttggagacagtacaaaggctt
 V  V  A  I  A  N  N  G  G  G  K  Q  A  L  E  T  V  Q  R  L
ctccccgttctatgccaggatcacggccttacgccagctcaagtagtagcgatagcctct
 L  P  V  L  C  Q  D  H  G  L  T  P  A  Q  V  V  A  I  A  S
aataatggtgggaagcaggcgctcgagacagttcaacgactactcccggtattatgtcaa
 N  N  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q
gatcatgggctcacacccgcccaggttgtagcaattgcctcgcatgatggcggcaagcaa
 D  H  G  L  T  P  A  Q  V  V  A  I  A  S  H  D  G  G  K  Q
gcacttgagactgtccagcggctcttgccagttctctgccaggcacacggcctaactcca
 A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L  T  P
gcacaagtcgttgctatcgctaacaacggtggtggcaaacaggcattagaaaccgttcaa
 A  Q  V  V  A  I  A  N  N  G  G  G  K  Q  A  L  E  T  V  Q
cgtcttttaccggtcctgtgccaagctcacggcttgactcccgcacaagtggtagctata
 R  L  L  P  V  L  C  Q  A  H  G  L  T  P  A  Q  V  V  A  I
gcttcccatgatggcggaaagcaggcattggagactgtacagagattgctcccggttctc
 A  S  H  D  G  G  K  Q  A  L  E  T  V  Q  R  L  L  P  V  L
tgccaggcacacggtttaacgccagagcaggttgttgcaatagcaaatcacgatggaggt
 C  Q  A  H  G  L  T  P  E  Q  V  V  A  I  A  N  H  D  G  G
aaacaagcgctcgaaacggtccaacgtctcttgcccgtcctttgtcaagcgcacggactg
 K  Q  A  L  E  T  V  Q  R  L  L  P  V  L  C  Q  A  H  G  L
acacccgaacaggtggtcgccattgcttctaa
 T  P  E  Q  V  V  A  I  A  S
```

It was noted that 100% of one TALEN construct (designated "AxTALEN-F") and 40% of another (designated "AxTALEN-R") clones had the correct sequence. TALENs specific to AAVS1 and OCT 4 generated using Method 2 were also validated by Western blotting using an anti-FLAG antibody **(Figure 1e)**.

Assay for determining and/or assessing the function and/or activity of a genome editing system.

Results and discussion

Several methods have been developed to assess the functional application of genome editing systems including the surveyor assay and episomal gene repair assays[14,15]. We have developed a novel quantitative system termed GFP-SplitAx.

This system has been used to assess the function of TALENs designed to target the AAVS1 and OCT4 loci.

The principle of the assay is that eGFP is split into two fragments consisting of a fragment encoding the N-terminus (amino acid 1-157) and a fragment encoding the C-terminus (amino acid 158-end)[16]. These N- and C-terminal fragments are separated by a TALEN binding site such that the C-terminus is out of frame with its N-terminus of GFP see SEQ ID NOS: 1-8 above).

Transfection of the eGFP-SplitAx vector and TALENs introduce double strand breaks which are repaired by error prone non homologous end joining (NHEJ) resulting in deletions or insertions of DNA[8,17]. A change in the frame shift in the AAVS1 of -1, -4 or +1 or +4 or any triplet combination of this will restore the open reading frame with the C-terminal eGFP fragment and generate a fluorescent signal within the cell (**Figure 3a**).

We have used this novel GFP-SplitAx assay to confirm the function of the AAVS1 AxTALENs that were generated using two distinct methods.

To evaluate whether the GFP-SplitAx could also be used to assess other genome editing tools, we tested AAVS1 Zinc finger nucleases[18] (**Figures 3d and 3i**) and an AAVS1 CRISPR[7] (**Figures 3e and 3j**). Each of these genome editing systems performed well in the GFP-SplitAx assay and the nature of the assay allowed us to quantify their activity (**Figure 5**). In this assay we demonstrate that the AAVS1 zinc finger nuclease had a higher activity followed by CRISPR and then TALENs.

To demonstrate that this system could be applied to other loci, we designed a GFP-SplitAx vector containing an OCT4 TALEN binding site.

The translated OCT4-GFP-SplitAx (SEQ ID NO: 9 nucleic acid) and 10 (amino acid): Amino acid position 158 (bold underlined) and C-terminus out of frame with the N-terminal GFP (SEQ ID NO: 205).

```
atggtgagcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggac
 M  V  S  K  G  E  E  L  F  T  G  V  V  P  I  L  V  E  L  D
ggcgacgtaaacggccacaagttcagcgtgtccggcgagggcgagggcgatgccacctac
 G  D  V  N  G  H  K  F  S  V  S  G  E  G  E  G  D  A  T  Y
ggcaagctgaccctgaagttcatctgcaccaccggcaagctgcccgtgccctggcccacc
```

```
 G  K  L  T  L  K  F  I  C  T  T  G  K  L  P  V  P  W  P  T
ctcgtgaccaccctgacctacggcgtgcagtgcttcagccgctaccccgaccacatgaag
 L  V  T  T  L  T  Y  G  V  Q  C  F  S  R  Y  P  D  H  M  K
cagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttc
 Q  H  D  F  F  K  S  A  M  P  E  G  Y  V  Q  E  R  T  I  F
ttcaaggacgacggcaactacaagacccgcgccgaggtgaagttcgagggcgacaccctg
 F  K  D  D  G  N  Y  K  T  R  A  E  V  K  F  E  G  D  T  L
gtgaaccgcatcgagctgaagggcatcgacttcaaggaggacggcaacatcctggggcac
 V  N  R  I  E  L  K  G  I  D  F  K  E  D  G  N  I  L  G  H
aagctggagtacaactacaacagccacaacgtctatatcatggccgacaagcaggcggcc
 K  L  E  Y  N  Y  N  S  H  N  V  Y  I  M  A  D  K  Q  A  A
gcgtcacctgcagctgcccagacctggcacccaggagaggagcaggcagggtcagctcga
 A  S  P  A  A  A  Q  T  W  H  P  G  E  E  Q  A  G  S  A  R
gaagaacggcatcaaggtgaacttcaagatccgccacaacatcgaggacggcagcgtgca
 E  E  R  H  Q  G  E  L  Q  D  P  P  Q  H  R  G  R  Q  R  A
gctcgccgaccactaccagcagaacacccccatcggcgacggccccgtgctgctgcccga
 A  R  R  P  L  P  A  E  H  P  H  R  R  R  P  R  A  A  A  R
caaccactacctgagcacccagtccgccctgagcaaagaccccaacgagaagcgcgatca
 Q  P  L  P  E  H  P  V  R  P  E  Q  R  P  Q  R  E  A  R  S
catggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgta
 H  G  P  A  G  V  R  D  R  R  R  D  H  S  R  H  G  R  A  V
caagtaa
 Q  V
```

The Translated OCT4-GFP-SplitAx with a 1bp deletion: SEQ ID NO: 11 (nucleic acid) and 12 (amino acid) - this restores GFP N-terminal open reading frame with the C-terminal GFP (SEQ ID NO: 206).

```
atggtgagcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggac
 M  V  S  K  G  E  E  L  F  T  G  V  V  P  I  L  V  E  L  D
ggcgacgtaaacggccacaagttcagcgtgtccggcgagggcgagggcgatgccacctac
 G  D  V  N  G  H  K  F  S  V  S  G  E  G  E  G  D  A  T  Y
ggcaagctgaccctgaagttcatctgcaccaccggcaagctgcccgtgccctggcccacc
 G  K  L  T  L  K  F  I  C  T  T  G  K  L  P  V  P  W  P  T
ctcgtgaccaccctgacctacggcgtgcagtgcttcagccgctaccccgaccacatgaag
 L  V  T  T  L  T  Y  G  V  Q  C  F  S  R  Y  P  D  H  M  K
cagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttc
 Q  H  D  F  F  K  S  A  M  P  E  G  Y  V  Q  E  R  T  I  F
ttcaaggacgacggcaactacaagacccgcgccgaggtgaagttcgagggcgacaccctg
 F  K  D  D  G  N  Y  K  T  R  A  E  V  K  F  E  G  D  T  L
```

```
gtgaaccgcatcgagctgaagggcatcgacttcaaggaggacggcaacatcctggggcac
 V  N  R  I  E  L  K  G  I  D  F  K  E  D  G  N  I  L  G  H
aagctggagtacaactacaacagccacaacgtctatatcatggccgacaagcaggcggcc
 K  L  E  Y  N  Y  N  S  H  N  V  Y  I  M  A  D  K  Q  A  A
gcgtcacctgcagctgcccagacctggccccaggagaggagcaggcagggtcagctcgag
 A  S  P  A  A  A  Q  T  W  P  Q  E  R  S  R  Q  G  Q  L  E
aagaacggcatcaaggtgaacttcaagatccgccacaacatcgaggacggcagcgtgcag
 K  N  G  I  K  V  N  F  K  I  R  H  N  I  E  D  G  S  V  Q
ctcgccgaccactaccagcagaacaccccatcggcgacggccccgtgctgctgcccgac
 L  A  D  H  Y  Q  Q  N  T  P  I  G  D  G  P  V  L  L  P  D
aaccactacctgagcacccagtccgccctgagcaaagaccccaacgagaagcgcgatcac
 N  H  Y  L  S  T  Q  S  A  L  S  K  D  P  N  E  K  R  D  H
atggtcctgctggagttcgtgaccgccgcgggatcactctcggcatggacgagctgtac
 M  V  L  L  E  F  V  T  A  A  G  I  T  L  G  M  D  E  L  Y
aagtaa
 K  -
```

Co-transfection of OCT4 AxTALENs (OF and OR) with the OCT4 GFP-SplitAx vector resulted in restoration of GFP fluorescence in a significant proportion of cells that could be monitored by flow cytometry **(Figures 3f and 3k and Figure 6)**

To demonstrate that the Split-Ax technology could be applicable to other fluorescent proteins we designed and tested a Zeis Green Split-Ax AAVS1 vector. The Zeis Green Fluorescent protein was split into two fragments, (N-terminus amino acid 1-157 and the C-terminus amino acid 158-end) that were separated by the AAVS1 genome editing binding site. The relevant sequences are shown below as SEQ ID NOS: 13.

SEQ ID NO 13 (nucleic acid) 14 (amino acid): Zeis Green sequence and translated protein (SEQ ID NO: 207)

```
atggcccagtccaagcacggcctgaccaaggagatgaccatgaagtac
 M  A  Q  S  K  H  G  L  T  K  E  M  T  M  K  Y
cgcatggagggctgcgtggacggccacaagttcgtgatcaccggcgagggcatcggctac
 R  M  E  G  C  V  D  G  H  K  F  V  I  T  G  E  G  I  G  Y
cccttcaagggcaagcaggccatcaacctgtgcgtggtggagggcggccccttgcccttc
 P  F  K  G  K  Q  A  I  N  L  C  V  V  E  G  G  P  L  P  F
gccgaggacatcttgtccgccgccttcatgtacggcaaccgcgtgttcaccgagtacccc
 A  E  D  I  L  S  A  A  F  M  Y  G  N  R  V  F  T  E  Y  P
caggacatcgtcgactacttcaagaactcctgccccgccggctacacctgggaccgctcc
```

```
 Q  D  I  V  D  Y  F  K  N  S  C  P  A  G  Y  T  W  D  R  S
ttcctgttcgaggacggcgccgtgtgcatctgcaacgccgacatcaccgtgagcgtggag
 F  L  F  E  D  G  A  V  C  I  C  N  A  D  I  T  V  S  V  E
gagaactgcatgtaccacgagtccaagttctacggcgtgaacttccccgccgacggcccc
 E  N  C  M  Y  H  E  S  K  F  Y  G  V  N  F  P  A  D  G  P
gtgatgaagaagatgaccgacaactgggagccctcctgcgagaagatcatccccgtgccc
 V  M  K  K  M  T  D  N  W  E  P  S  C  E  K  I  I  P  V  P
aagcagggcatcttgaagggcgacgtgagcatgtacctgctgctgaaggacggtggccgc
 K  Q  G  I  L  K  G  D  V  S  M  Y  L  L  L  K  D  G  G  R
ttgcgctgccagttcgacaccgtgtacaaggccaagtccgtgccccgcaagatgcccgac
 L  R  C  Q  F  D  T  V  Y  K  A  K  S  V  P  R  K  M  P  D
tggcacttcatccagcacaagctgacccgcgaggaccgcagcgacgccaagaaccagaag
 W  H  F  I  Q  H  K  L  T  R  E  D  R  S  D  A  K  N  Q  K
tggcacctgaccgagcacgccatcgcctccggctccgccttgccctga
 W  H  L  T  E  H  A  I  A  S  G  S  A  L  P  -
```

SEQ ID NO: 15 (nucleic acid) and 16 (amino acid): AAVS1- ZeisGreen-SplitAx with AAVS1 genome editing binding site downstream of N-terminus (highlighted grey). Stop codons shown as dashes (-): (SEQ ID NO: 208).

```
atggcccagtccaagcacggcctgaccaaggagatgaccatgaagtaccgcatggagggc
 M  A  Q  S  K  H  G  L  T  K  E  M  T  M  K  Y  R  M  E  G
tgcgtggacggccacaagttcgtgatcaccggcgagggcatcggctacccttcaagggc
 C  V  D  G  H  K  F  V  I  T  G  E  G  I  G  Y  P  F  K  G
aagcaggccatcaacctgtgcgtggtggagggcggcccttgcccttcgccgaggacatc
 K  Q  A  I  N  L  C  V  V  E  G  G  P  L  P  F  A  E  D  I
ttgtccgccgccttcatgtacggcaaccgcgtgttcaccgagtaccccaggacatcgtc
 L  S  A  A  F  M  Y  G  N  R  V  F  T  E  Y  P  Q  D  I  V
gactacttcaagaactcctgccccgccggctacacctgggaccgctccttcctgttcgag
 D  Y  F  K  N  S  C  P  A  G  Y  T  W  D  R  S  F  L  F  E
gacggcgccgtgtgcatctgcaacgccgacatcaccgtgagcgtggaggagaactgcatg
 D  G  A  V  C  I  C  N  A  D  I  T  V  S  V  E  E  N  C  M
taccacgagtccaagttctacggcgtgaacttccccgccgacggccccgtgatgaagaag
 Y  H  E  S  K  F  Y  G  V  N  F  P  A  D  G  P  V  M  K  K
atgaccgacaactgggagccctcctgcgagaagatcatccccgtgcccaaggcggccgca
 M  T  D  N  W  E  P  S  C  E  K  I  I  P  V  P  K  A  A  A
agcttatctgtcccctccaccccacagtggggccactagggacaggattggtgacagaaa
 S  L  S  V  P  S  T  P  Q  W  G  H  -  G  Q  D  W  -  Q  K
agccccatccttggatccctcgagacagggcatcttgaagggcgacgtgagcatgtacct
 S  P  I  L  G  S  L  E  T  G  H  L  E  G  R  R  E  H  V  P
gctgctgaaggacggtggccgcttgcgctgccagttcgacaccgtgtacaaggccaagtc
```

```
 A  A  E  G  R  W  P  L  A  L  P  V  R  H  R  V  Q  G  Q  V
cgtgccccgcaagatgcccgactggcacttcatccagcacaagctgacccgcgaggaccg
 R  A  P  Q  D  A  R  L  A  L  H  P  A  Q  A  D  P  R  G  P
cagcgacgccaagaaccagaagtggcacctgaccgagcacgccatcgcctccggctccgc
 Q  R  R  Q  E  P  E  V  A  P  D  R  A  R  H  R  L  R  L  R
cttgccctga
 L  A  L
```

SEQ ID NO: 209 (nucleic acid) and 210 (amino acid): Translated AAVS1-ZeisGreen-SplitAx with a 1bp deletion which restores Zeis Green N-terminal open reading frame with the C-terminal Zeis Green (SEQ ID NO: 209).

```
atggcccagtccaagcacggcctgaccaaggagatgaccatgaagtaccgcatggagggc
 M  A  Q  S  K  H  G  L  T  K  E  M  T  M  K  Y  R  M  E  G
tgcgtggacggccacaagttcgtgatcaccggcgagggcatcggctacccttcaagggc
 C  V  D  G  H  K  F  V  I  T  G  E  G  I  G  Y  P  F  K  G
aagcaggccatcaacctgtgcgtggtggagggcggcccttgcccttcgccgaggacatc
 K  Q  A  I  N  L  C  V  V  E  G  G  P  L  P  F  A  E  D  I
ttgtccgccgccttcatgtacggcaaccgcgtgttcaccgagtaccccaggacatcgtc
 L  S  A  A  F  M  Y  G  N  R  V  F  T  E  Y  P  Q  D  I  V
gactacttcaagaactcctgccccgccggctacacctgggaccgctccttcctgttcgag
 D  Y  F  K  N  S  C  P  A  G  Y  T  W  D  R  S  F  L  F  E
gacggcgccgtgtgcatctgcaacgccgacatcaccgtgagcgtggaggagaactgcatg
 D  G  A  V  C  I  C  N  A  D  I  T  V  S  V  E  E  N  C  M
taccacgagtccaagttctacggcgtgaacttccccgccgacggccccgtgatgaagaag
 Y  H  E  S  K  F  Y  G  V  N  F  P  A  D  G  P  V  M  K  K
atgaccgacaactgggagccctcctgcgagaagatcatccccgtgcccaaggcggccgca
 M  T  D  N  W  E  P  S  C  E  K  I  I  P  V  P  K  A  A  A
agcttatctgtcccctccaccccacagtggggccatagggacaggattggtgacagaaaa
 S  L  S  V  P  S  T  P  Q  W  G  H  R  D  R  I  G  D  R  K
gccccatccttggatccctcgagacagggcatcttgaagggcgacgtgagcatgtacctg
 A  P  S  L  D  P  S  R  Q  G  I  L  K  G  D  V  S  M  Y  L
ctgctgaaggacggtggccgcttgcgctgccagttcgacaccgtgtacaaggccaagtcc
 L  L  K  D  G  G  R  L  R  C  Q  F  D  T  V  Y  K  A  K  S
gtgccccgcaagatgcccgactggcacttcatccagcacaagctgacccgcgaggaccgc
 V  P  R  K  M  P  D  W  H  F  I  Q  H  K  L  T  R  E  D  R
agcgacgccaagaaccagaagtggcacctgaccgagcacgccatcgcctccggctccgcc
 S  D  A  K  N  Q  K  W  H  L  T  E  H  A  I  A  S  G  S  A
ttgccctga
 L  P  -
```

SEQ ID NO: 211: AAVS1-ZeisGreen-SplitAx synthesised fragment

ATGGCCCAGTCCAAGCACGGCCTGACCAAGGAGATGACCATGAAGTACCGCATGGAGGGCTGCGTGGACGGCCAC
AAGTTCGTGATCACCGGCGAGGGCATCGGCTACCCCTTCAAGGGCAAGCAGGCCATCAACCTGTGCGTGGTGGAG
GGCGGCCCCTTGCCCTTCGCCGAGGACATCTTGTCCGCCGCCTTCATGTACGGCAACCGCGTGTTCACCGAGTAC
CCCCAGGACATCGTCGACTACTTCAAGAACTCCTGCCCCGCCGGCTACACCTGGGACCGCTCCTTCCTGTTCGAG
GACGGCGCCGTGTGCATCTGCAACGCCGACATCACCGTGAGCGTGGAGGAGAACTGCATGTACCACGAGTCCAAG
TTCTACGGCGTGAACTTCCCCGCCGACGGCCCCGTGATGAAGAAGATGACCGACAACTGGGAGCCCTCCTGCGAG
AAGATCATCCCCGTGCCCAAGGCGGCCGCAAGCTTATCTGTCCCCTCCACCCCACAGTGGGGCCACTAGGGACAG
GATTGGTGACAGAAAAGCCCCATCCTTGGATCCCTCGAGACAGGGCATCTTGAAGGGCGACGTGAGCATGTACCT
GCTGCTGAAGGACGGTGGCCGCTTGCGCTGCCAGTTCGACACCGTGTACAAGGCCAAGTCCGTGCCCCGCAAGAT
GCCCGACTGGCACTTCATCCAGCACAAGCTGACCCGCGAGGACCGCAGCGACGCCAAGAACCAGAAGTGGCACCT
GACCGAGCACGCCATCGCCTCCGGCTCCGCCTTGCCCTGA

Co-transfection of this vector with the AAVS1 zinc fingers restored Zeis Green fluorescence in a significant proportion of cells (**Figure 3g and 3i and Figure 6)**

We then tested the AAVS1 TALENs to show that they could be used to edit the genome in human 293FT cells using a pZDonor-pCAG-Zeis Green targeting vector (**Figure 4a)**. The 293FT cells were transfected with their respective pair of AxTALENs and a targeting vector with homologous ends. Genomic DNA PCR with diagnostic primers demonstrated that homologous recombination at the AAVS1 locus had occurred (**Figure 4b**). We performed similar experiments using the OCT4 AxTALENs to show that they could also be used to edit the genome in human cells (**Figure 4c)**. 293FT cells were co-transfected with their respective pair of OCT4 AxTALENs and an OCT4 targeting vector[4]. Genomic DNA PCR with diagnostic primers demonstrated that targeting of the OCT4 locus had occurred (**Figure 4d**). Diagnostic PCR products for both AAVS1 and OCT4 targeting experiments were sequenced and this confirmed that homologous recombination had occurred.

The GFP-SplitAx described herein can be used to monitor, assess or determine for example, endonuclease, recomninase, TALEN, zinc finger and/or CRISPR function/activity and/or efficacy. The assay represents a significant improvement to the Surveyor Assay. For example, reporter activity can be monitored in real time and the assay provides a quantitative analysis of activity using flow cytometry.

References

1. Urnov, F.D., et al. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature 435, 646-651 (2005).

2. Wang, H., et al. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell 153, 910-918 (2013).

3. Gao, X., et al. Reprogramming to pluripotency using designer TALE transcription factors targeting enhancers. Stem cell reports 1, 183-197 (2013).

4. Hockemeyer, D., et al. Genetic engineering of human ES and iPS cells using TALE nucleases. (2011).

5. Perez-Pinera, P., et al. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nature methods 10, 973-976 (2013).

6. Cong, L., et al. Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819-823 (2013).

7. Mali, P., et al. RNA-guided human genome engineering via Cas9. Science 339, 823-826 (2013).

8. Sander, J.D., et al. Targeted gene disruption in somatic zebrafish cells using engineered TALENs. Nature biotechnology 29, 697 (2011).

9. Reyon, D., Khayter, C., Regan, M.R., Joung, J.K. & Sander, J.D. Engineering designer transcription Activator--Like effector nucleases (TALENs) by REAL or REAL-Fast assembly. Current protocols in molecular biology, 12.15. 11-12.15. 14 (2012).

10. Reyon, D., et al. FLASH assembly of TALENs for high-throughput genome editing. Nature biotechnology 30, 460-465 (2012).

11. Cermak, T., et al. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic acids research, gkr218 (2011).

12. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. PloS one 4, e5553 (2009).

13. Gibson, D.G., et al. Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. Science 319, 1215-1220 (2008).

14. Miller, J.C., et al. A TALE nuclease architecture for efficient genome editing. Nature biotechnology 29, 143-148 (2011).

15. Alwin, S., et al. Custom zinc-finger nucleases for use in human cells. Molecular Therapy 12, 610-617 (2005).

16. Crone, D.E., et al. GFP-based biosensors. (2013).

17. Gaj, T., Guo, J., Kato, Y., Sirk, S.J. & Barbas III, C.F. Targeted gene knockout by direct delivery of zinc-finger nuclease proteins. Nature methods 9, 805-807 (2012).

18. DeKelver, R.C., et al. Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. Genome research 20, 1133-1142 (2010).

19. Sakuma, T., et al. Repeating pattern of non-RVD variations in DNA-binding modules enhances TALEN activity. Scientific reports 3(2013).

20. Sanjana, N.E., et al. A transcription activator-like effector toolbox for genome engineering. Nature protocols 7, 171-192 (2012).

Claims

1. A method of generating one or more **T**ranscription **A**ctivator **L**ike-**E**ffector (TALE) molecule encoding sequence(s) or TALE molecule(s), said method comprising combining or assembling one or more of the TALE unit sequences presented in TABLE 1 or one or more of the TALE unit nucleic acid sequences and/or TALE unit amino acid sequences of any SEQ ID NOS: 1-128, to provide one or more TALE molecule encoding sequences or TALE molecule(s).

2. The method of claim 1, wherein the method is a method of generating a TALE molecule encoding nucleic acid sequence and the method requires the user to combine or assemble together one or more of the TALE unit encoding nucleic acid sequences presented in Table 1 or as SEQ ID NOS: 1-64 to provide a TALE molecule encoding nucleic acid sequence.

3. The method of any one of claims 1 or 2, wherein the method is a method of providing TALE molecule sequences which have binding specificity/affinity for predetermined target nucleic acid sequences.

4. The method of claim 3, wherein the method comprises the selection and/or analysis of a target nucleic acid sequence to which the TALE molecule is to exhibit some binding specificity/affinity.

5. The method of any one of claims 1-4, wherein the method comprises computationally combining or assembling TALE unit encoding sequences to provide a TALE molecule sequence.

6. The method of any one of claims 1-5, wherein the generated TALE molecule sequence is synthesised for use.

7. The method of claim 6, wherein the TALE molecule is chemically synthesised.

8. The method of claim 6 or 7, wherein the step of synthesising does not require the use of Gibson assembly or PCR.

9. The method of claims 6-8, wherein the synthesised sequence is a nucleic acid sequence encoding a TALE molecule for use.

10. The method of any one of claims 1-9, wherein the method is used to generate multiple TALE molecule encoding sequences for joining or ligating together.

11. The method of claim 10, wherein one or more of the TALE molecule encoding sequences comprises 5' and/or 3' modifications to permit the joining to or ligation with, other TALE molecule encoding sequences.

12. The method of claim 10 or 11, wherein one or more of the TALE molecule encoding sequences comprises is suitable for joining to another by Gibson assembly.

13. A method of generating a sequence encoding a complete **T**ranscription **A**ctivator **L**ike-**E**ffector (TALE) molecule specific for a target sequence, said method comprising:

(a) selecting and/or analysing the target nucleic acid sequence to determine the required number and type of TALE units of the complete TALE molecule to be encoded;

(b) combining or assembling the relevant one or more TALE unit sequences presented in TABLE 1 or the relevant one or more TALE unit nucleic acid sequences and/or TALE unit amino acid sequences of SEQ ID NOS: 1-128 to provide:

(i) a TALE molecule encoding sequence which encodes the complete TALE molecule; or

(ii) a plurality of TALE molecule encoding sequences which each encode a different part of the complete TALE molecule;

(c) synthesising the products of (b); and where step (c) provides a plurality of synthesised TALE molecule encoding sequences, joining these sequences by Gibson assembly so as to provide a complete TALE molecule encoding sequence specific for the target sequence.

14. The method of claim 13, wherein the sequence encoding a complete **T**ranscription **A**ctivator **L**ike-**E**ffector (TALE) molecule is fused to an endonuclease to provide a TALEN.

15. A **T**ranscription **A**ctivator **L**ike-**E**ffector (TALE) molecule encoding sequence obtainable by the method of claim 13.

16. A method of generating a TALE molecule, said method comprising the steps of:

combining two or more of the TALE unit encoding nucleic acid sequences presented in TABLE 1 or of SEQ ID NOS: 1-64 to provide a TALE molecule encoding nucleic acid sequence specific for a predetermined target sequence;

synthesising the TALE molecule encoding nucleic acid sequence;

introducing the synthesised TALE molecule encoding nucleic acid sequence into a vector; and

introducing the vector into a host cell and maintaining the host cell under conditions which facilitate the expression of the TALE molecule encoding nucleic acid sequence.

17. A method of generating a **T**ranscription **A**ctivator **L**ike-**E**ndo**n**uclease (TALEN) molecule, said method comprising the steps of:

combining two or more of the TALE unit encoding nucleic acid sequences in TABLE 1 or of SEQ ID NOS: 1-64 to provide a TALE molecule encoding nucleic acid sequence specific for a predetermined target sequence;

synthesising the TALE molecule encoding nucleic acid sequence;

introducing the synthesised TALE molecule encoding nucleic acid sequence into a vector, which vector comprises an endonuclease encoding nucleic acid to provide a vector which encodes a TALEN molecule; and

introducing the vector into a host cell and maintaining the host cell under conditions which facilitate the expression of a TALEN encoding nucleic acid molecule.

18. A TALE and/or TALEN molecule obtainable by any of the methods of claims 1-17.

19. A **T**ranscription **A**ctivator **L**ike-**E**ffector (TALE) unit sequence presented in TABLE 1.

20. A TALE unit nucleic acid sequence and/or one or more of the TALE unit amino acid sequences of any SEQ ID NO: 1-128

21. A TALE unit sequence conforming to the following consensus:

$\mathbf{A_1}$ – [**TU**] – $\mathbf{A_2}$

wherein $\mathbf{A_1}$ represents an optional additional sequence or modification;

TU represents any one of the 64 TALE unit encoding nucleic acid sequences presented in Table 1; and

$\mathbf{A_2}$ represents an optional additional sequence or modification.

22. The TALE unit sequence of claim 21, wherein optional sequences A_1 and A_2 comprise restriction site sequences, primer binding sites and/or sequences which facilitate the ligation or joining of one TALE unit sequence to another.

23. A TALE nucleic acid or amino acid molecule comprising two or more of the TALE unit (nucleic acid or amino acid) sequences according to claim 19 or 20.

24. The TALE molecule of claim 23, wherein the molecule further comprises a heterologous sequence.

25. The TALE molecule of claim 24, wherein the heterologous sequence encodes an endonuclease.

26. A data carrier or digital medium or storage device comprising, containing, carrying or loaded with, the information presented in Table 1.

27. A method of designing, generating or providing a TALE and/or TALEN molecule , said method comprising a Computer Aided Design (CAD) program, wherein the CAD program performs a method of determining an appropriate or suitable assembly of TALE unit sequences from those disclosed in Table 1 and/or presented as SEQ ID NOS: 1-128.

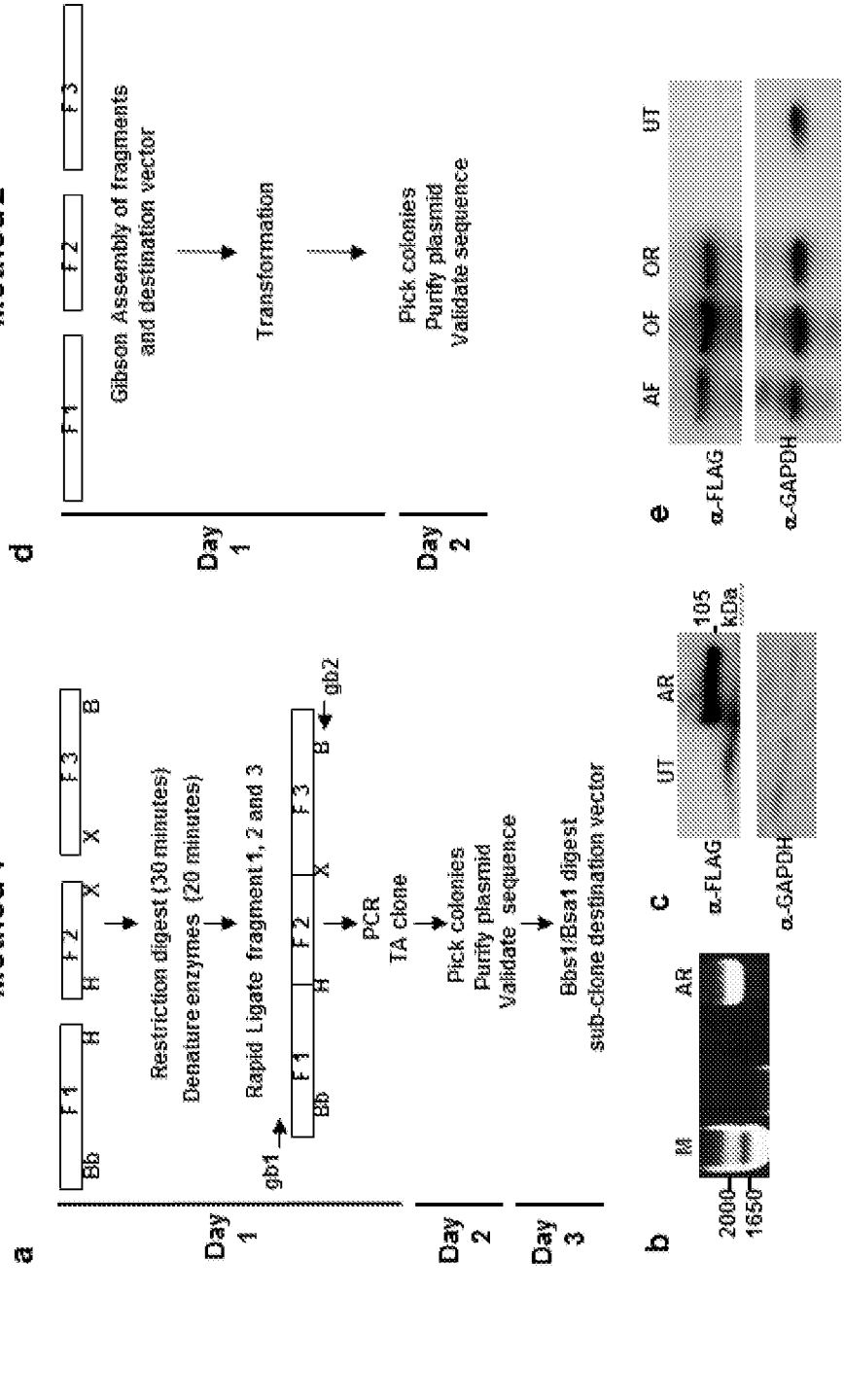
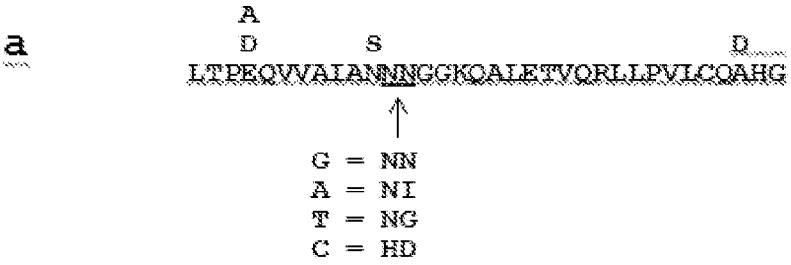

a
Method 1
F 1
Bb
H
F 2
H
X
F 3
X
B
Restriction digest (30 minutes)
Denature enzymes (20 minutes)
Rapid Ligate fragment 1, 2 and 3
gb1
gb2
PCR
TA clone
Pick colonies
Purify plasmid
Validate sequence
Bbs1/Bsa1 digest
sub-clone destination vector
Day 1
Day 2
Day 3
b
M
AR
2000
1650
c
UT
AR
105 kDa
α-FLAG
α-GAPDH
d
Method 2
F 1
F 2
F 3
Gibson Assembly of fragments and destination vector
Transformation
Pick colonies
Purify plasmid
Validate sequence
Day 1
Day 2
e
AF
OF
OR
UT
α-FLAG
α-GAPDH

Figure 1

a


A
D
S
D
LTPEQVVAIANNNGGKQALETVQRLLPVLCQAHG
G = NN
A = NI
T = NG
C = HD

b

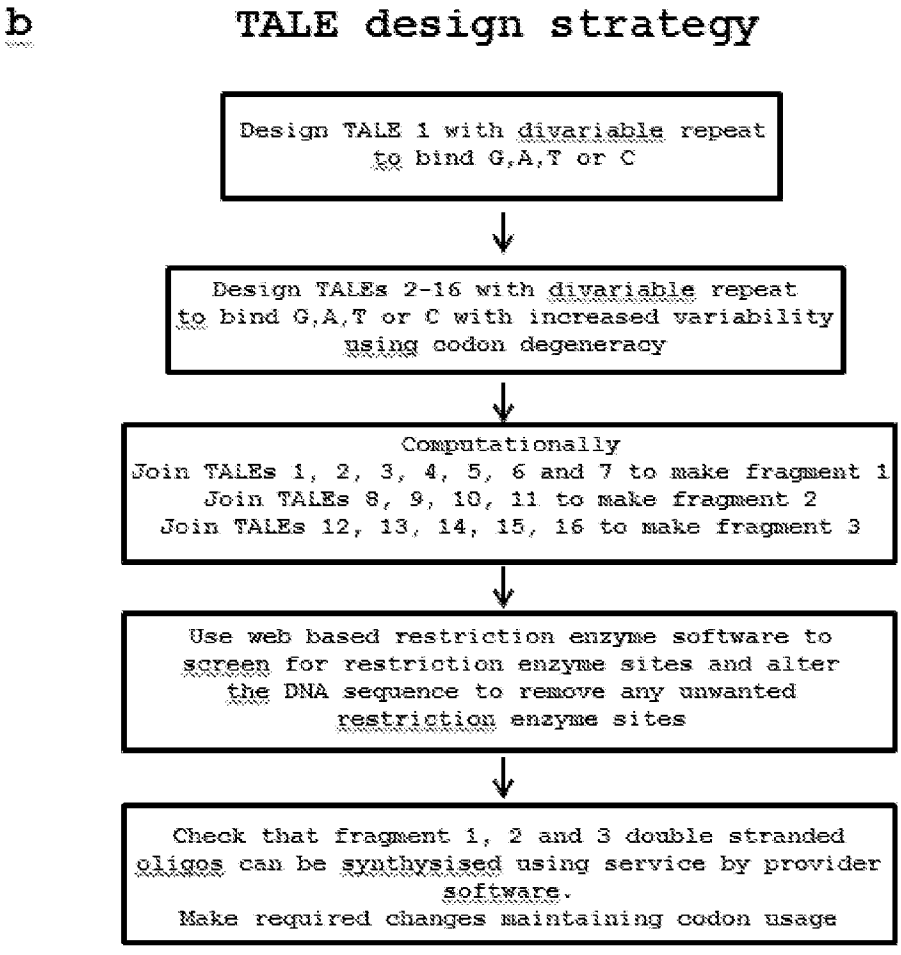

TALE design strategy
Design TALE 1 with divariable repeat to bind G,A,T or C
Design TALEs 2-16 with divariable repeat to bind G,A,T or C with increased variability using codon degeneracy
Computationally
Join TALEs 1, 2, 3, 4, 5, 6 and 7 to make fragment 1
Join TALEs 8, 9, 10, 11 to make fragment 2
Join TALEs 12, 13, 14, 15, 16 to make fragment 3
Use web based restriction enzyme software to screen for restriction enzyme sites and alter the DNA sequence to remove any unwanted restriction enzyme sites
Check that fragment 1, 2 and 3 double stranded oligos can be synthysised using service by provider software.
Make required changes maintaining codon usage

Figure 2

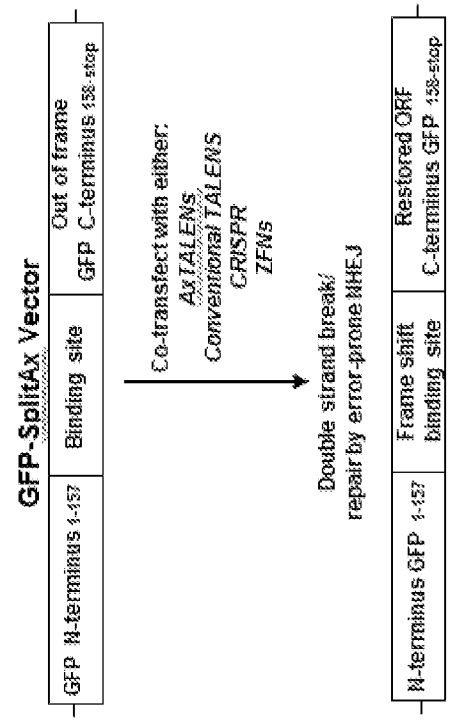

a
GFP-SplitAx Vector
GFP N-terminus 1-157
Binding site
Out of frame
GFP C-terminus 158-stop
Co-transfect with either:
AxTALENs
Conventional TALENs
CRISPR
ZFNs
Double strand break/
repair by error-prone NHEJ
N-terminus GFP 1-157
Frame shift
binding site
Restored ORF
C-terminus GFP 158-stop

Figure 3a

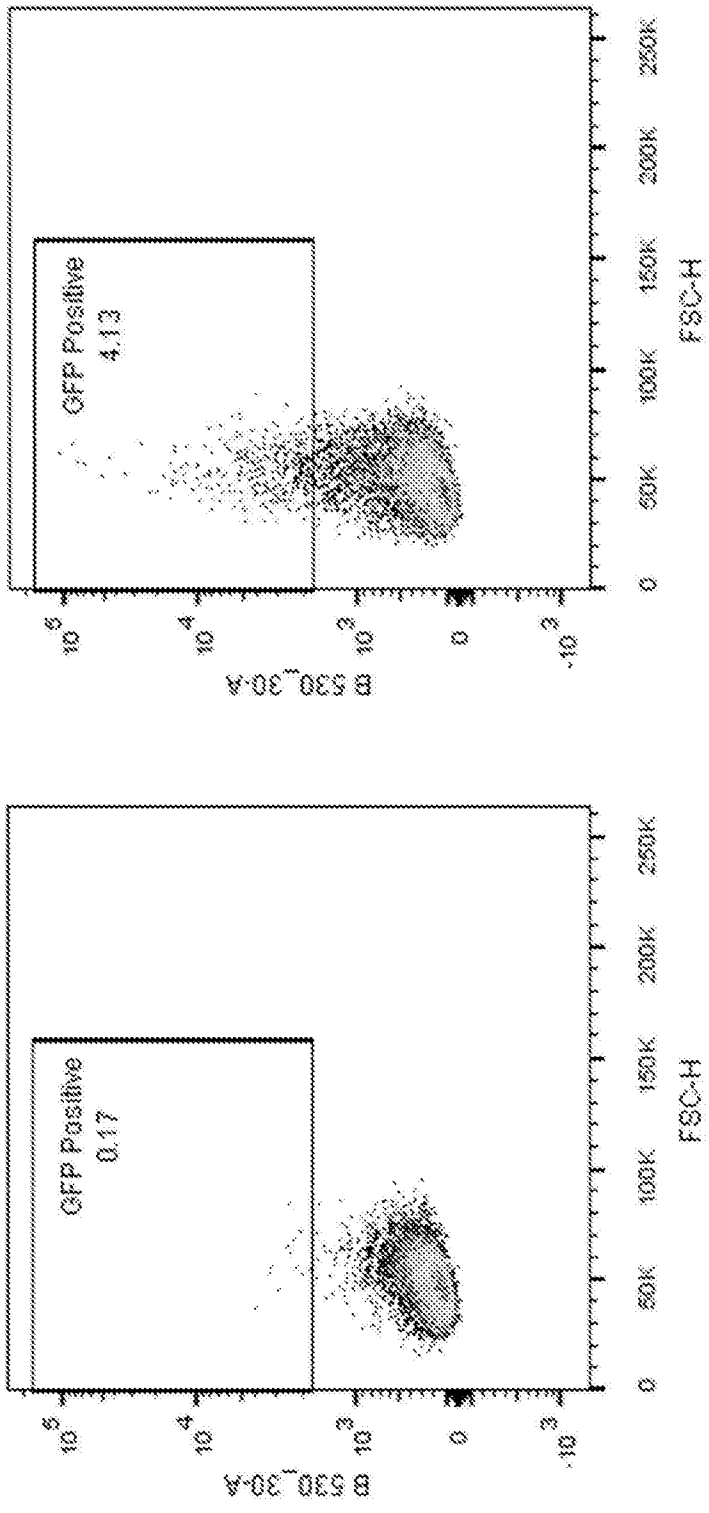


Figure 3b and 3c

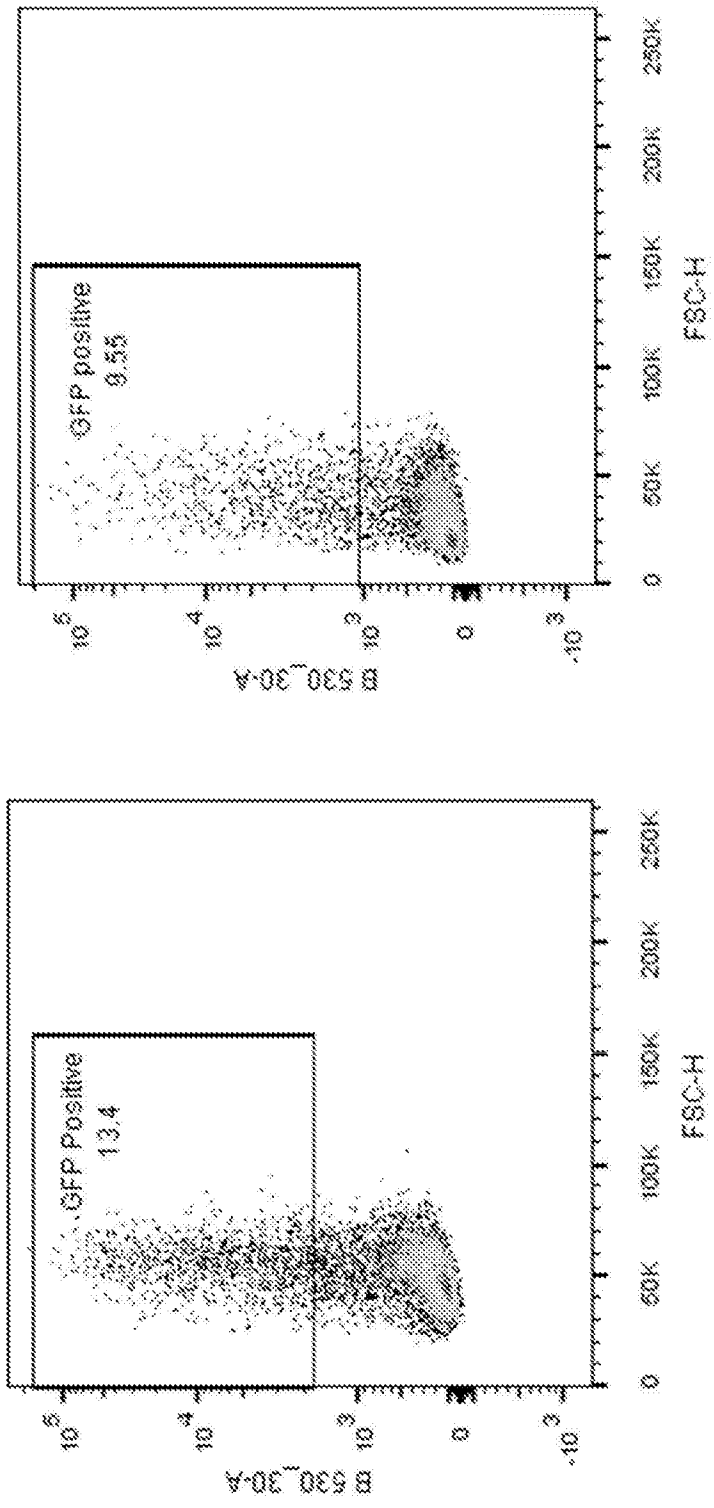


Figure 3d and 3e

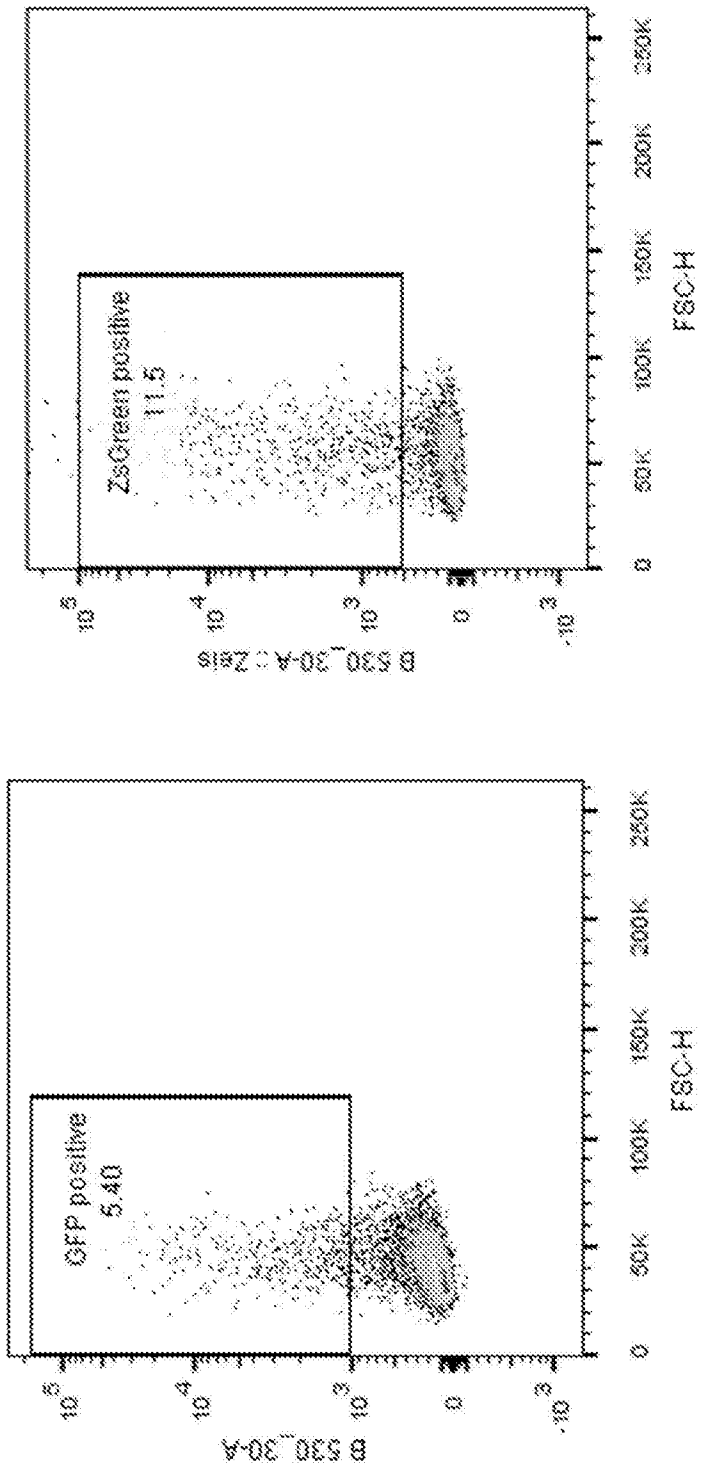


Figure 3f and 3g

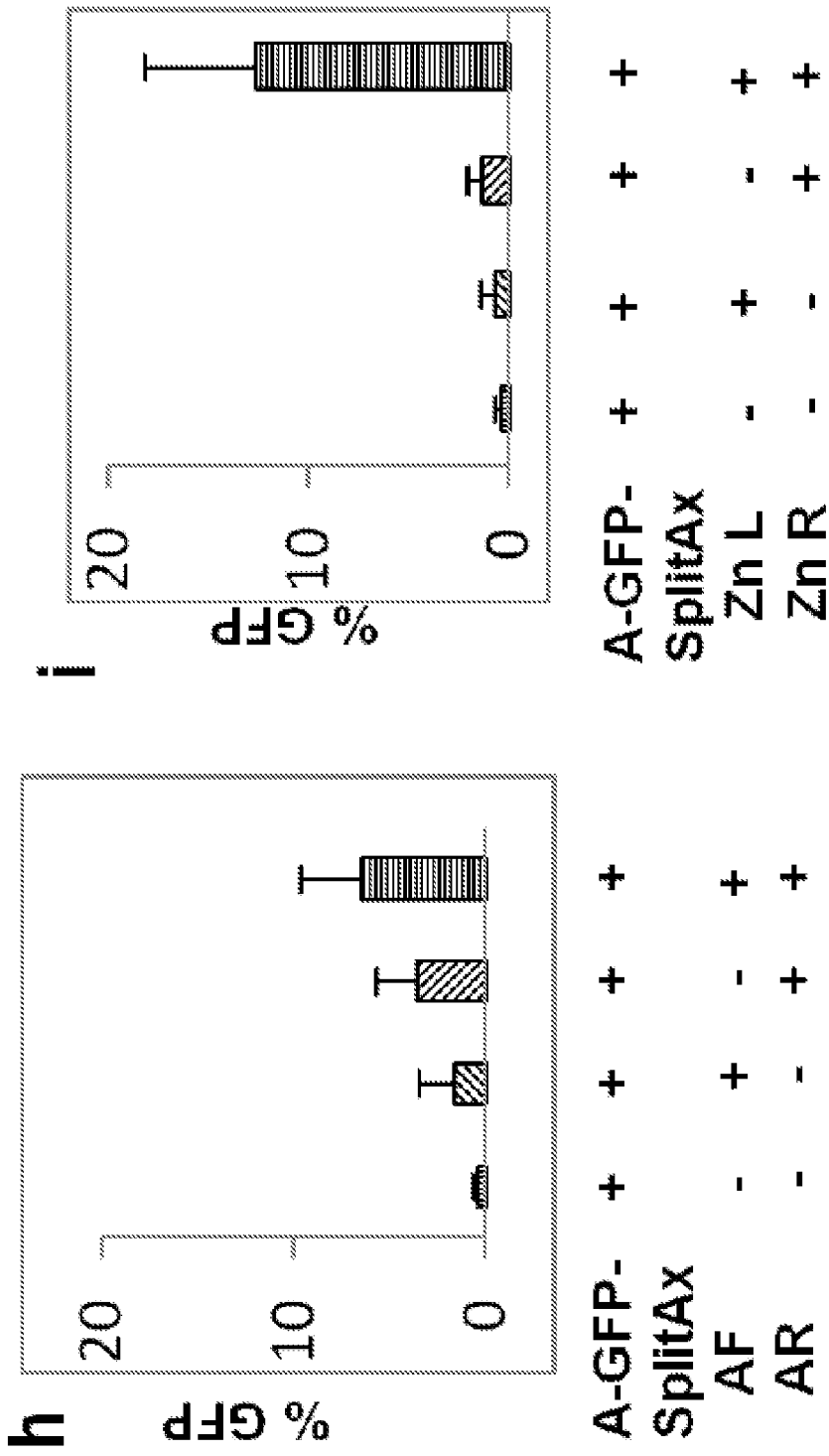

h
% GFP
20
10
0
A-GFP- + + + +
SplitAx + + - +
AF - + - +
AR - - + +
i
% GFP
20
10
0
A-GFP- + + + +
SplitAx - + - +
Zn L - + + +
Zn R - - + +

Figure 3h & i

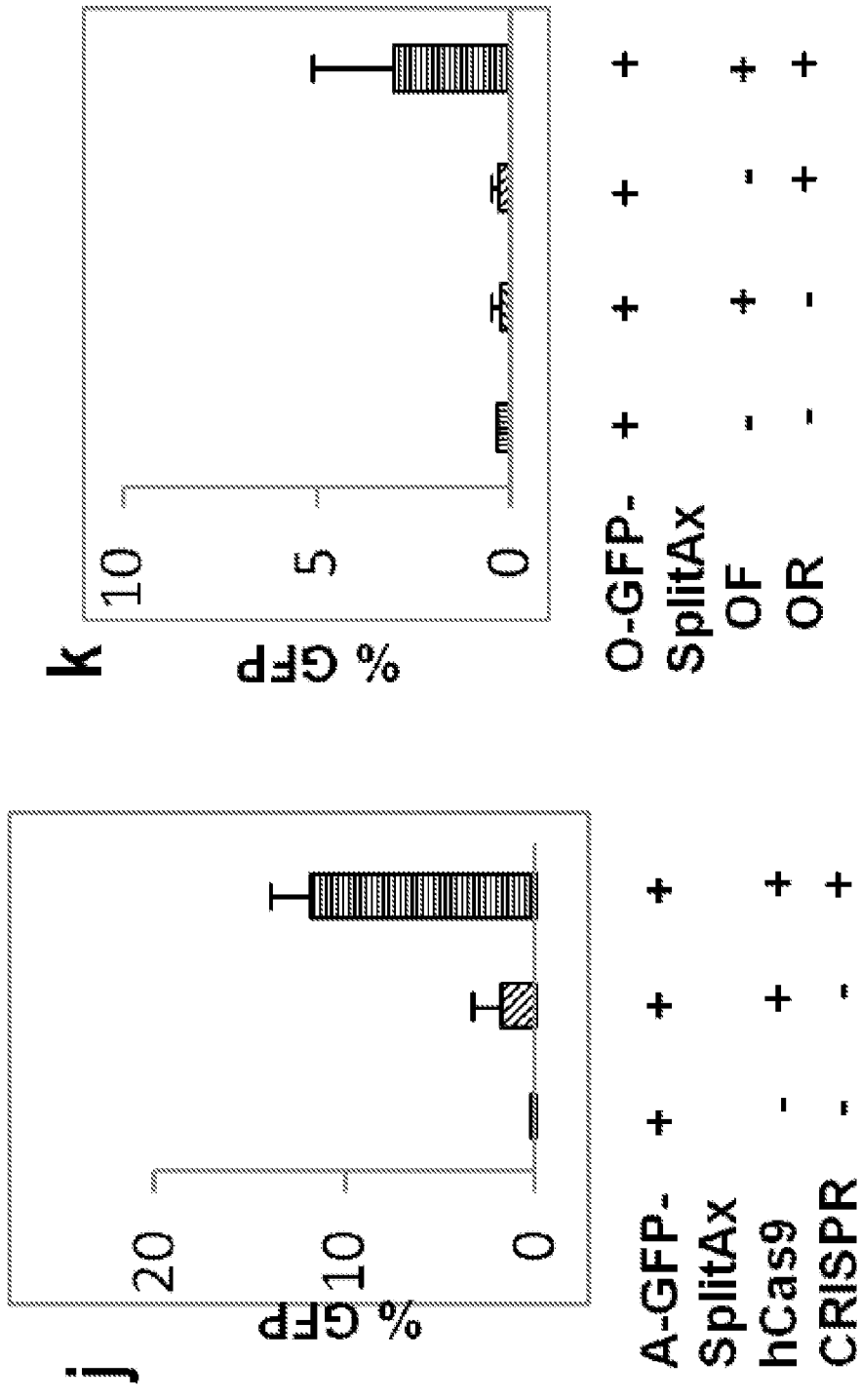

k
% GFP
10
5
0
O-GFP-
SplitAx
OF
OR
j
% GFP
20
10
0
A-GFP-
SplitAx
hCas9
CRISPR

Figure 3J & K

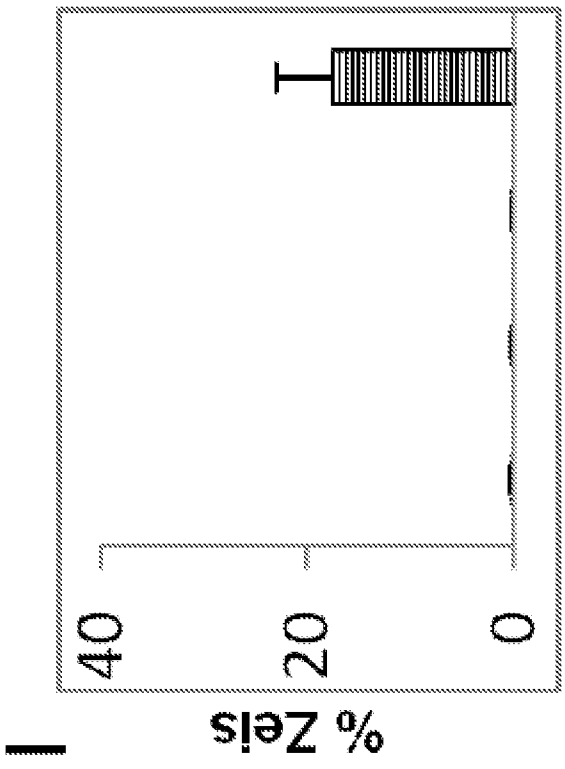


Figure 3I

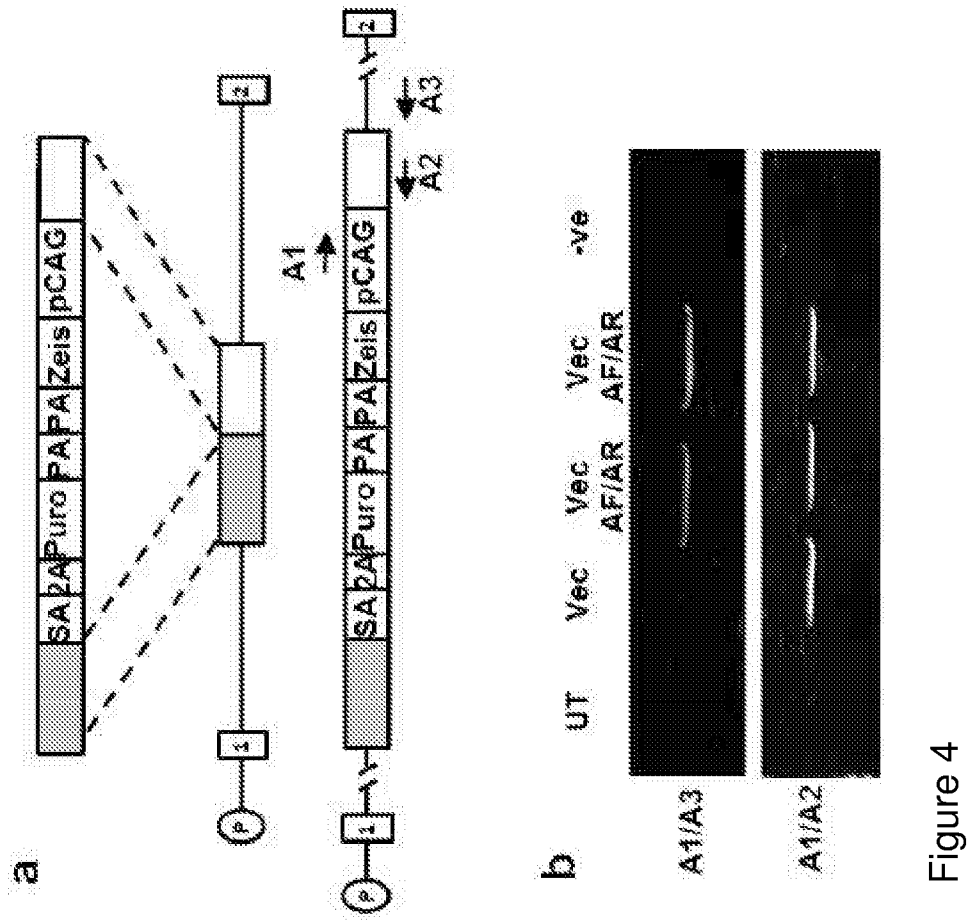

a
SA
pA
Puro
PA
PA
Zeis
pCAG
A1
A3
A2
b
UT
Vec
Vec AF/AR
Vec AF/AR
-ve
A1/A3
A1/A2

Figure 4

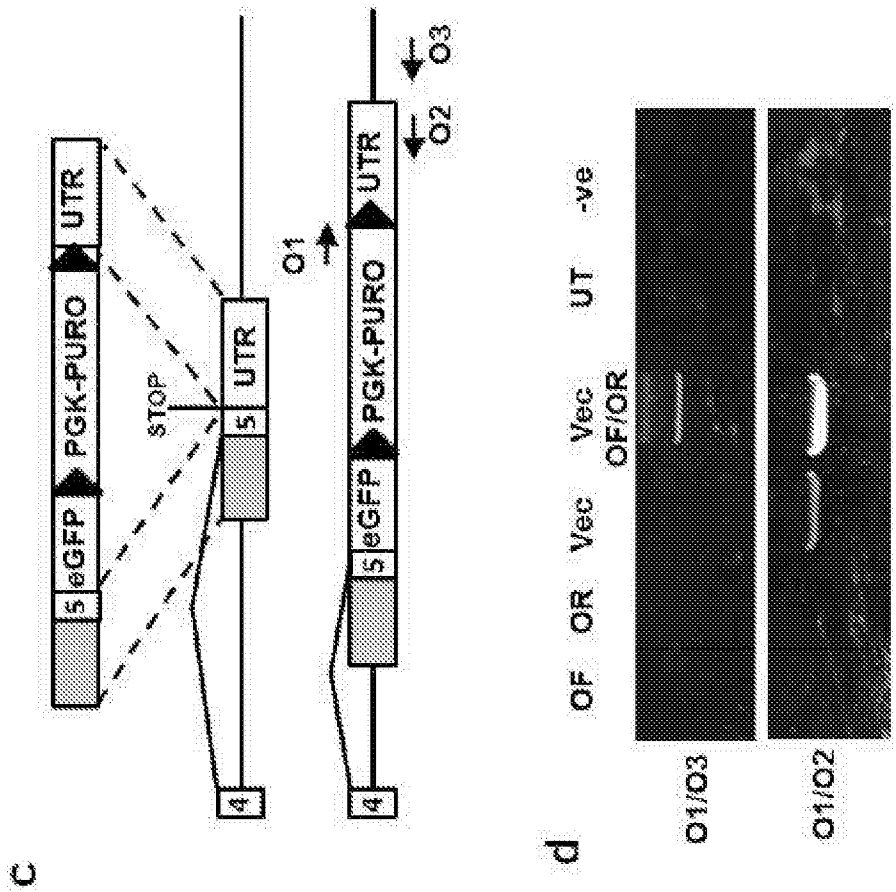


Figure 4 cont.

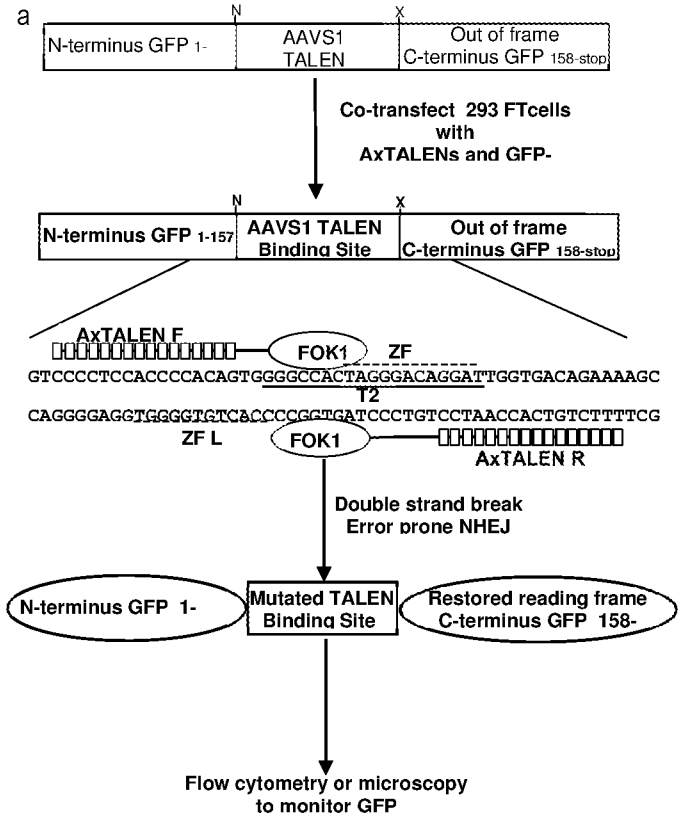

a
N
X
N-terminus GFP 1-
AAVS1 TALEN
Out of frame C-terminus GFP 158-stop
Co-transfect 293 FTcells with AxTALENs and GFP-
N-terminus GFP 1-157
AAVS1 TALEN Binding Site
Out of frame C-terminus GFP 158-stop
AxTALEN F
FOK1
ZF
GTCCCCTCCACCCCACAGTGGGGCCACTAGGGACAGGATTGGTGACAGAAAAGC
T2
CAGGGGAGGTGGGGTGTCACCCCGGTGATCCCTGTCCTAACCACTGTCTTTTCG
ZF L
FOK1
AxTALEN R
Double strand break Error prone NHEJ
N-terminus GFP 1-
Mutated TALEN Binding Site
Restored reading frame C-terminus GFP 158-
Flow cytometry or microscopy to monitor GFP

Figure 5

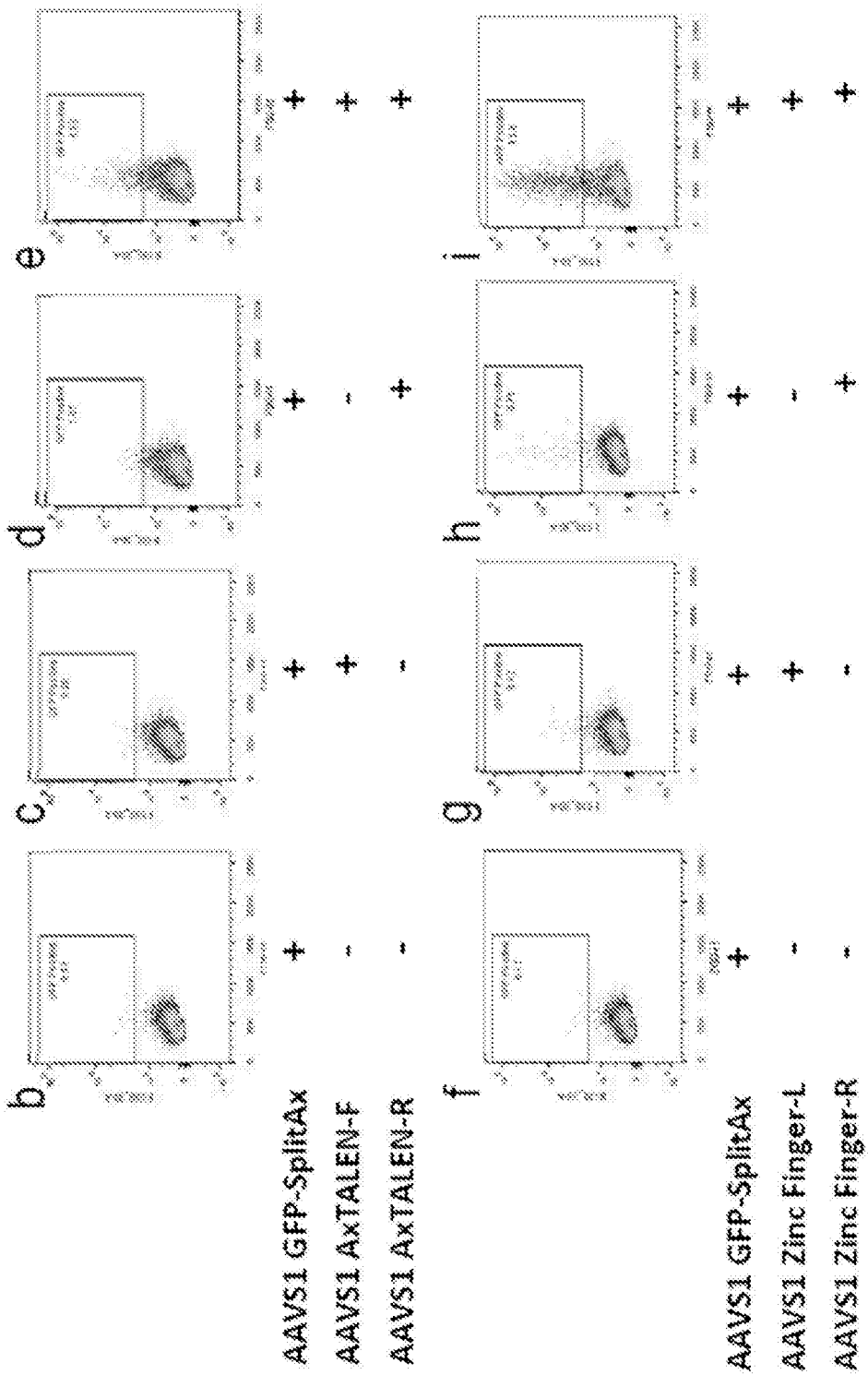


Figure 5 cont.

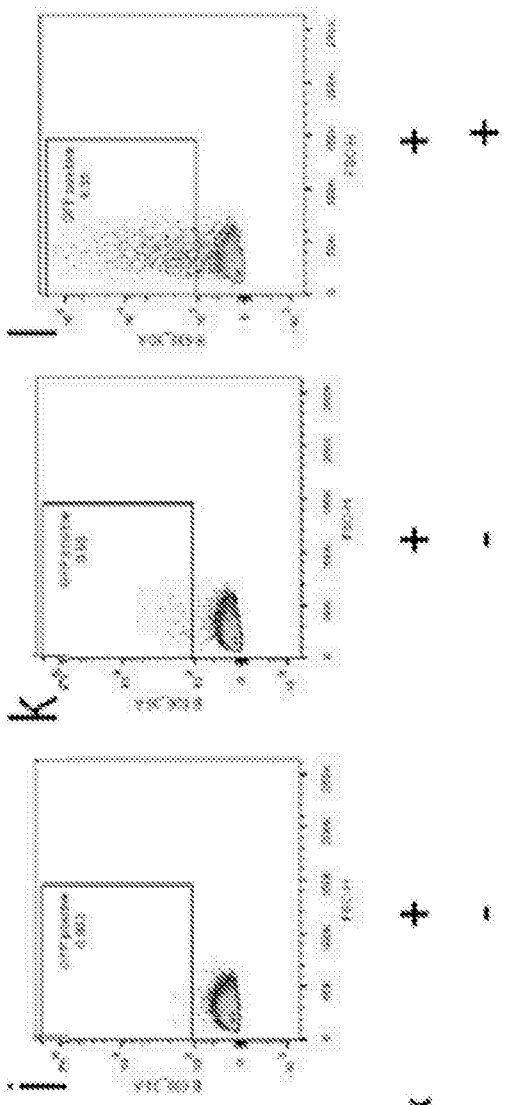


Figure 5 cont.

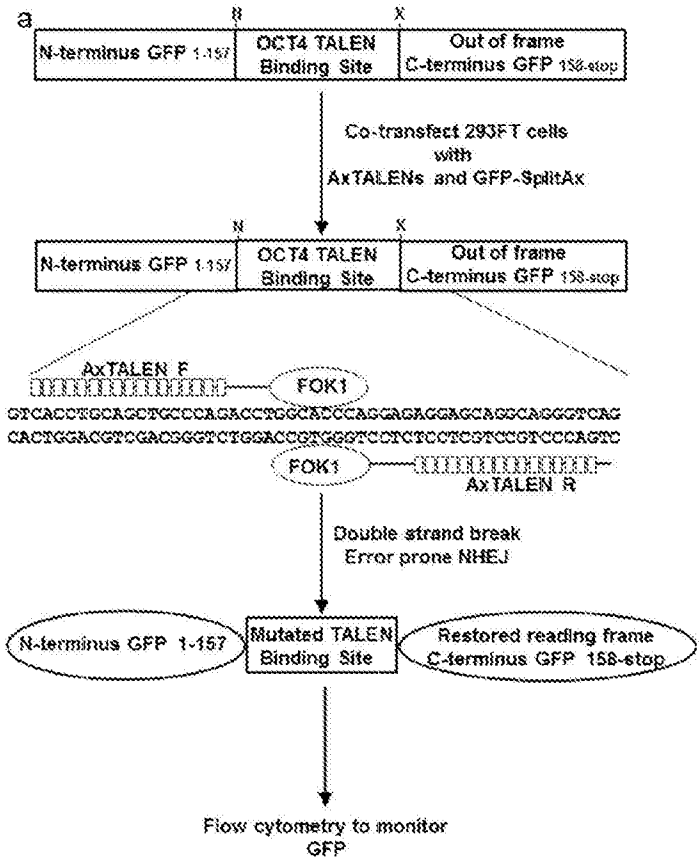

a
N-terminus GFP 1-157
OCT4 TALEN Binding Site
Out of frame C-terminus GFP 158-stop
Co-transfect 293FT cells with AxTALENs and GFP-SplitAx
AxTALEN F
FOK1
FOK1
AxTALEN R
Double strand break Error prone NHEJ
N-terminus GFP 1-157
Mutated TALEN Binding Site
Restored reading frame C-terminus GFP 158-stop
Flow cytometry to monitor GFP

Figure 6

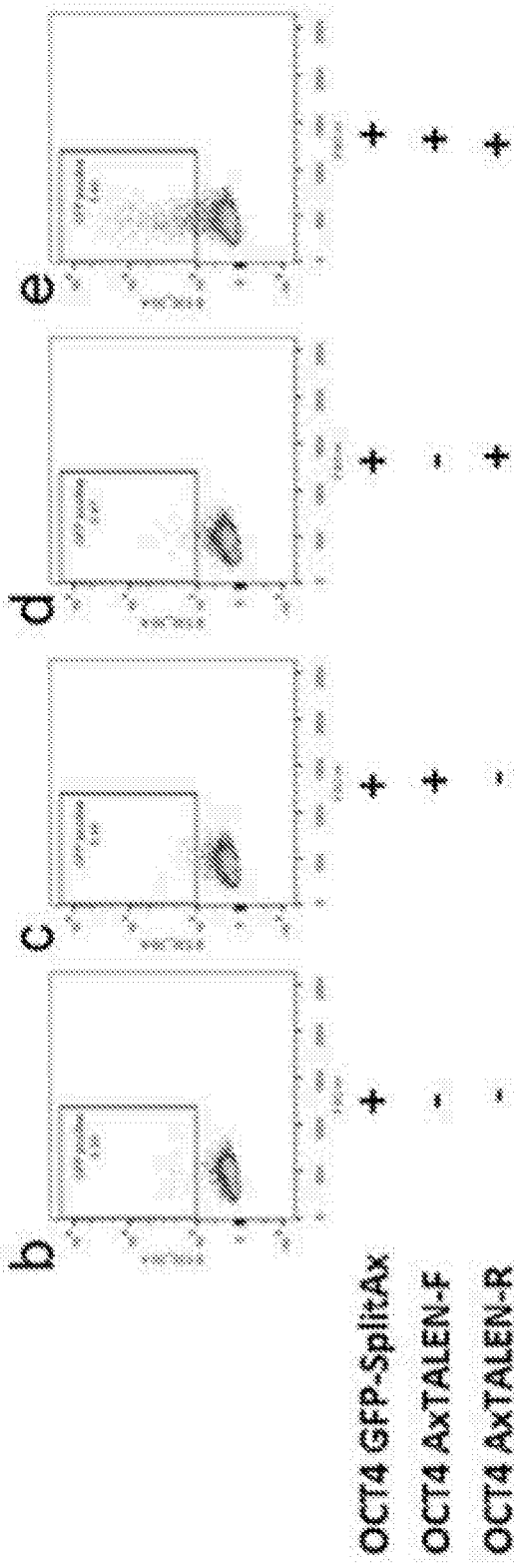


Figure 6 cont.

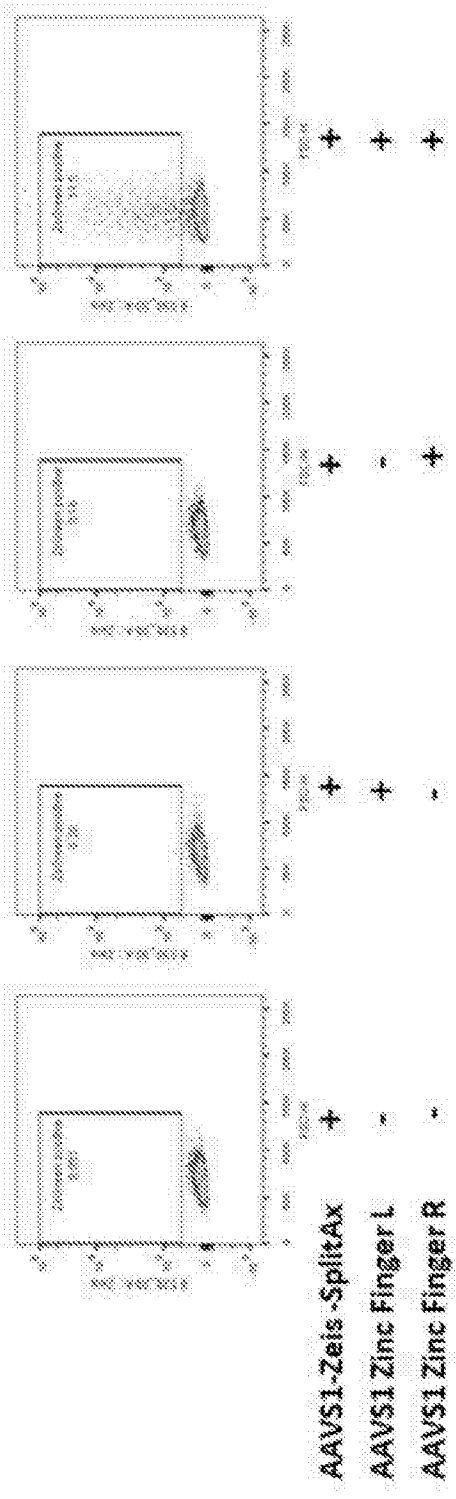


Figure 6

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2016/050508

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12N15/11
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data, EMBASE, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2011/146121 A1 (SANGAMO BIOSCIENCES, INC. [US]) 24 November 2011 (2011-11-24) positions 12 and 13 of the TALE unit defines specifity of binding to nucleotides; page 4 page 27 page 40; sequence 19 page 81 - page 85; example 5 page 98 - page 112; examples 11-14 page 143; sequence 333 the whole document -----	1-27
X	WO 2013/191769 A1 (MAYO FOUNDATION [US]) 27 December 2013 (2013-12-27) page 11, line 3 - line 9 the whole document -----	1-27
	-/--	

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
19 May 2016	02/06/2016

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Macchia, Giovanni

4

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2016/050508

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2012/270273 A1 (PRESIDENT AND FELLOWS OF HARVARD COLLEGE [US]) 25 October 2012 (2012-10-25) the whole document	1-27
X	WO 2013/082519 A2 (THE BROAD INSTITUTE INC. [US]; MASSACHUSETTS INSTITUTE TECHNOLOGY [US]) 6 June 2013 (2013-06-06) the whole document	1-27
X	R. JANKELE ET AL.: "TAL effectors: tools for DNA targeting", BRIEFINGS IN FUNCTIONAL GENOMICS, vol. 13, no. 5, 6 June 2014 (2014-06-06), pages 409-419, XP055273314, Oxford, UK ISSN: 2041-2649, DOI: 10.1093/bfgp/elu013 the whole document	1-27
A	HOMMELSHEIM CARL MAXIMILIAN ET AL.: "PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications.", SCIENTIFIC REPORTS, vol. 4, 5052, 2014, pages 1-13, XP002757803, ISSN: 2045-2322 the whole document	1-27
A	-& CARL MAXIMILIAN HOMMELSHEIM ET AL: "Supplementary information to PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications", SCIENTIFIC REPORTS, vol. 4, 23 May 2014 (2014-05-23), pages 1-23, XP055273331, DOI: 10.1038/srep05052 the whole document	1-27
A	C. MUSSOLINO ET AL.: "A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity", NUCLEIC ACIDS RESEARCH, vol. 39, no. 21, 3 August 2011 (2011-08-03), pages 9283-9293, XP055021128, ISSN: 0305-1048, DOI: 10.1093/nar/gkr597 the whole document	1-27

4

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/GB2016/050508

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2011146121 A1	24-11-2011	AU 2011256838 A1	06-12-2012
		CA 2798988 A1	24-11-2011
		CN 103025344 A	03-04-2013
		EP 2571512 A1	27-03-2013
		JP 2013529083 A	18-07-2013
		KR 20130111219 A	10-10-2013
		US 2011301073 A1	08-12-2011
		US 2014134723 A1	15-05-2014
		US 2014134740 A1	15-05-2014
		US 2014134741 A1	15-05-2014
		WO 2011146121 A1	24-11-2011
WO 2013191769 A1	27-12-2013	US 2015291951 A1	15-10-2015
		WO 2013191769 A1	27-12-2013
US 2012270273 A1	25-10-2012	NONE	
WO 2013082519 A2	06-06-2013	EP 2785831 A2	08-10-2014
		EP 2821413 A1	07-01-2015
		EP 2821505 A2	07-01-2015
		HK 1202892 A1	09-10-2015
		HK 1207650 A1	05-02-2016
		HK 1207669 A1	05-02-2016
		US 8450107 B1	28-05-2013
		US 2013137160 A1	30-05-2013
		US 2013137161 A1	30-05-2013
		US 2013149780 A1	13-06-2013
		US 2013149781 A1	13-06-2013
		WO 2013082519 A2	06-06-2013