CHIMERIC POLYPEPTIDES HAVING TARGETED BINDING SPECIFICITY

CROSS REFERENCE TO RELATED APPLICATION(S)

[0001] This application claims benefit of priority under 35 U.S.C. §119(e) of U.S. Serial No. 61/696,689, filed September 4, 2012; U.S. Serial No. 61/753,763, filed January 17, 2013; and U.S. Serial No. 61/818,364, filed May 1, 2013, the entire contents of which are incorporated herein by reference in their entireties.

BACKGROUND OF THE INVENTION

FIELD OF THE INVENTION

[0002] The present invention relates generally to the field of biotechnology, and more specifically to chimeric recombinases that recognize specific DNA sequences.

BACKGROUND INFORMATION

[0003] The ability of proteins to recognize DNA in a sequence-dependent manner is central to life, as a variety of protein domains have evolved to provide sequence-specific DNA recognition. DNA recognition by a select few of these domains is also the foundation for a wide variety of biotechnological applications. In particular, C_2H_2 zinc-finger proteins (ZFPs) were among the first DNA-binding proteins to be engineered to recognize user-defined DNA sequences and have been used with varying degrees of success for many applications, including transcriptional regulation, genome engineering and epigenetic modification. Modular assembly of ZFPs has facilitated these approaches. However, despite the advances and promise of ZFP technology, construction of specific, high-affinity ZFPs for certain sequences remains difficult and in select cases requires the use of time-consuming and labor-intensive selection systems not readily adopted by non-specialty laboratories.

[0004] Transcription activator-like effector (TALE) domains are a class of naturally occurring DNA-binding domains (DBDs) that represent a potential alternative to ZFP technology. TALEs, which are found in the plant pathogen *Xanthomonas*, contain a series of 33 to 35 amino acid repeats that function to selectively bind target DNA sequences. These repeats are identical with the exception of two adjacent repeat variable di-residues (RVDs) that confer DNA specificity by mediating binding to a single nucleotide. Arrays of over 30 repeats have been described that bind to DNA sites of similar numbers of base pairs (bps). Although there is inherent degeneracy in the binding of each RVD, recent reports have indicated that synthetic TALE proteins are specific enough to target single loci within the human genome.

[0005] The introduction of DNA double-strand breaks (DSBs) by chimeric nucleases,

such as zinc-finger nucleases (ZFNs) can be used to knockout gene function or in the presence of exogenously added DNA drive cassette integration at the targeted loci. ZFNs have been extensively studied over the last decade and in some cases are approaching clinical use for gene therapy. Recently, a number of groups have explored the use of TALE DNA-binding domains fused to nucleases (TALENs) for targeted genome editing. Indeed, much of the work with ZFNs has been replicated with TALE nucleases, as TALENs may have advantages over ZFNs in regards to DNA-binding modularity. However, despite impressive research with ZFNs and TALENs, questions remain about their safety and specificity. In particular, off-target cleavage events remain difficult to detect, as the most likely result of an off-target DSB is the introduction of small insertions or deletions. Additionally, repair of DSBs relies on cell machinery that varies with cell type.

[0006] An alternate approach for achieving targeted genomic modifications is the use of site-specific recombinases (SSRs). SSRs, such as the tyrosine recombinases Cre and Flp, are valuable molecular biology tools that are routinely used to manipulate chromosome structure inside cells. Because these enzymes rely on a number of complex protein-protein and protein-DNA interactions to coordinate catalysis, SSRs exhibit remarkable target site specificity. To date, however, altering the specificity of many SSRs has proven difficult. Serine recombinases of the resolvase/invertase type provide a versatile alternative to tyrosine recombinases for genome engineering. In nature, these enzymes function as multi-domain protein complexes that coordinate recombination in a highly modular manner. However, mutants of several serine recombinases have been identified that do not require accessory factors for recombination. Additionally, numerous studies have shown that the native DBDs of serine recombinases can be replaced with custom-designed ZFPs to generate chimeric zinc-finger recombinases (ZFRs). In principle, ZFRs capable of recognizing an extended number of sequences could be generated, however, the lack of zinc-finger domains capable of recognizing all possible DNA triplets limits the potential modular targeting capacity of these enzymes.

[0007] ZFRs are composed of an activated catalytic domain derived from the resolvase/invertase family of serine recombinases and a zinc- finger DNA-binding domain that can be custom-designed to recognize almost any DNA sequence (Figure 30A). ZFRs catalyze recombination between specific ZFR target sites that consist of two-inverted zinc-finger binding sites (ZFBS) flanking a central 20-bp core sequence recognized by the recombinase catalytic domain (Figure 30B). In contrast to zinc- finger nucleases (ZFNs) and

TAL effector nucleases (TALENs), ZFRs function autonomously and can excise and integrate transgenes in human and mouse cells without activating the cellular DNA damage response pathway. However, as with conventional site-specific recombinases, applications of ZFRs have been restricted by sequence requirements imposed by the recombinase catalytic domain, which dictate that ZFR target sites contain a 20-bp core derived from a native serine resolvase/invertase recombination site.

[0008] Site-specific DNA recombination systems such as Cre-loxP, FLP-FRT and ^C31-att have emerged as powerful tools for genetic engineering. The site-specific recombinases that promote these DNA rearrangements recognize short (30- to 40-bp) sequences and coordinate DNA cleavage, strand exchange and re-ligation by a mechanism that does not require DNA synthesis or a high-energy cofactor. This simplicity has allowed researchers to study gene function with extraordinary spatial and temporal sensitivity. However, the strict sequence requirements imposed by site-specific recombinases have limited their application to cells and organisms that contain artificially introduced recombination sites. In order to address this limitation, directed evolution has been used to alter the sequence specificity of several recombinases toward naturally occurring DNA sequences. Despite advances, the need for complex mutagenesis and selection strategies and the finding that re-engineered recombinase variants routinely exhibit relaxed substrate specificity have hindered the widespread adoption of this technology.

[0009] Accordingly, there is a need for a more generalized method of catalyzing targeted and site-specific recombination of the endogenous genome, particularly for gene therapy, as well as for enzymes that can catalyze such targeted and site-specific recombination. This is particularly useful for gene therapy, but would have many other applications in molecular biology, including in gene cloning and use in modification of industrial organisms and agricultural plants and animals.

SUMMARY OF THE INVENTION

[0010] Disclosed herein are targeted chimeric polypeptides, including compositions thereof, expression vectors, and methods of use thereof, for the generation of transgenic cells, tissues, plants, and animals. The compositions, vectors, and methods of the present invention are also useful in gene therapy techniques.

[0011] In one aspect, the invention provides a chimeric polypeptide. The polypeptide includes: a) a recombinase, nuclease or transcription factor, or fragment thereof; and b) a transcription activator-like effector (TALE) protein. In various embodiments, the TALE

protein is truncated and includes a C-terminal or N-terminal truncation. In embodiments, the TALE protein is AcrXa7, Tal1c, and PthXo1. In embodiments, the TALE protein includes all or a portion an amino acid sequence as set forth in SEQ ID NO: 2. In some embodiments, the TALE protein is truncated between amino acid residues 27 and 268, 92 and 134, 120 and 129, 74 and 147, or 87 and 120 of SEQ ID NO: 2. In some embodiments, the TALE protein is truncated at amino acid residue 28, 74, 87, 92, 95, 120, 124, 128, 129, 147 and 150 of SEQ ID NO: 2.

[0012] In another aspect, the invention provides a method of generating a transcription activator-like effector (TALE) protein binding domain which specifically binds a desired nucleotide. The method includes a) randomizing the amino acid sequence of the TALE protein binding domain by mutating an amino acid residue within a variable di-residue (RVD), or within 1 to 2 amino acid residues N-terminal or C-terminal of the RVD; and b) selecting for the randomized TALE protein binding domain of (a), wherein the TALE protein binding domain specifically binds to the desired nucleotide.

[0013] In another aspect, the invention provides an isolated polypeptide comprising a *Xanthamonus* derived transcription activator-like effector (TALE) protein, the TALE protein having an N-terminal domain (NTD) comprising an amino acid sequence as set forth in SEQ ID NO: 3 (VGKQWSGARAL) having one or more mutations or deletions selected from: Q is Y, Q is S, Q is R, W is R, W is G, W is deleted, S is R, S is H, S is A, S is N, and S is T.

[0014] In another aspect, the invention provides an isolated polypeptide including a *Ralstonia* derived transcription activator-like effector (TALE) protein, the TALE protein having an N-terminal domain (NTD) including an amino acid sequence as set forth in SEQ ID NO: 8 (IVDIAR$_1$QR$_2$SGDLA) having one or more mutations or deletions selected from: R$_1$ is K, Q is Y, Q is S, Q is R, R$_2$ is W, R$_2$ is G, R$_2$ is deleted, S is R, S is H, S is A, S is N, and S is T.

[0015] In another embodiment, the invention provides a method of generating a transcription activator-like effector (TALE) protein N-terminal domain (NTD). The method includes: a) randomizing an amino acid sequence of the NTD by mutating or deleting one or more amino acid residues within the NTD, wherein the amino acid sequence is SEQ ID NO: 14 (VGKXXXGAR) or SEQ ID NO: 15 (VDIAXXXXGDLA); and b) selecting for the randomized TALE protein NTD of (a), wherein the TALE protein NTD specifically binds to a desired nucleotide or exhibits enhanced activity.

[0016] Also disclosed herein are chimeric proteins including a serine recombinase and one or more zinc finger binding domains, methods of generating ZFRs, compositions thereof, expression vectors, and methods of use thereof, for the generation of transgenic cells, tissues, plants, and animals. The compositions, vectors, and methods of the present invention are also useful in gene therapy techniques.

[0017] In one aspect, the invention provides a method of generating a plurality of zinc finger recombinase (ZFRs) proteins having catalytic specificity greater than the corresponding wild type recombinase. The method includes performing random mutagenesis on a recombinase catalytic domain at positions equivalent to Gin Ile120, Thr123, Leu127, Ile136 and Gly137 or a combination thereof, mutating the DNA at positions 2 and 3 for each amino acid; fusing the recombinase catalytic domain with a plurality of zinc finger binding domains to form ZFRs, and enriching for ZFRs having catalytic specificity greater than the corresponding wild type recombinase. In embodiments the ZFRs have increased catalytic activity on DNA targets selected from GC, GT, CA, TT and AC. In one embodiment, the recombinase catalytic domain is mutagenized at Ile136 and/or Gly137.

[0018] In various aspects, the chimeric polypeptides described herein include a recombinase catalytic domain derived from or randomly mutagenized as disclosed herein from: a) Tn3, also known as EcoTn3; Hin, also known as StyHin; Gin, also known as MuGin; Sin; Beta; Pin; Min; Din; Cin; EcoTn21; SfaTn917; BmeTn5083; Bme53; Cpe; SauSK1; SauSK41; SauTn552; Ran; Aac; Lla; pMER05; Mlo92; Mlo90; Rrh; Pje; Req; PpsTn5501; Pae; Xan; ISXc5; Spy; RhizY4cG; SarpNL1; SsolSC1904a; SsolSC1904b; SsoISC1913; Aam606; MjaM0014; Pab; HpylS607; MtulS_Y349; MtuRv2792c; MtuRv2979c; MtuRv3828c; MtuRv0921; MceRv0921; TnpX; TndX; WwK; lactococcal phage TP901-1 serine recombinase; *S. pyogenes* phage ϕ370.1 serine recombinase; *S. pyogenes* phage ϕFC1 serine recombinase; *Listeria* phage A118 serine recombinase; *S. coelicolor* chromosome SC3C8.24 serine recombinase; *S. coelicolor* chromosome SC2E1.37 serine recombinase; *S. coelicolor* chromosome SCD78.04c serine recombinase; *S. coelicolor* chromosome SC8F4.15c serine recombinase; *S. coelicolor* chromosome SCD12A.23 serine recombinase; *S. coelicolor* chromosome SCH10.38c serine recombinase; *S. coelicolor* chromosome SCC88.14 serine recombinase; *Streptomyces* phage ϕC31 serine recombinase; *Streptomyces* phage R4 serine recombinase; *Bacillus* phage ϕ105 serine recombinase; *Bacillus* phage SPBc2 serine recombinase; *Bacillus* prophage SKIN serine recombinase; *S. aureus ccrA* serine recombinase; *S. aureus ccrB* serine recombinase; *M. tuberculosis* phage Bxb1 serine

recombinase; *M. tuberculosis* prophage ϕRV1 serine recombinase; YBCK_ECOLI; Y4bA; Bja; Spn; Cac 1956; and Cac 1954; or b) muteins of a).

[0019] In yet another aspect, the invention provides an isolated nucleic acid molecule encoding the chimeric polypeptide described herein.

[0020] In yet another aspect, the invention provides an expression cassette including the nucleic acid molecule the chimeric polypeptide described herein.

[0021] In yet another aspect, the invention provides a vector including the expression cassette described herein.

[0022] In yet another aspect, the invention provides an isolated host cell containing the vector described herein.

[0023] In yet another aspect, the invention provides a method for site-specific integration into a DNA sequence. The method includes contacting the DNA sequence with a chimeric polypeptide of the present invention, wherein the chimeric polypeptide catalyzes site-specific integration.

[0024] In yet another aspect, the invention provides a method for gene therapy. The method includes administering to a subject a composition comprising a nucleic acid molecule encoding the chimeric polypeptide described herein, wherein upon expression of the nucleic acid molecule, a gene present in the genome of the subject is specifically removed or inactivated.

[0025] In yet another aspect, the invention provides a pharmaceutical composition. The composition includes the chimeric polypeptide described herein; and a pharmaceutically acceptable carrier. In another aspect, the composition includes a nucleic acid molecule encoding the chimeric polypeptide described herein; and a pharmaceutically acceptable carrier.

[0026] In yet another aspect, the invention provides a transgenic organism produced by recombination catalyzed by the chimeric polypeptide of the present invention.

[0027] In yet another aspect, the invention provides a method for gene therapy. The method includes administering to a subject a cell comprising a nucleic acid molecule having the DNA sequence generated by the method of site-specific integration described herein.

[0028] In another aspect, the invention provides an isolated nucleic acid molecule encoding the chimeric protein described herein.

[0029] In another aspect, the invention provides a method for site-specific recombination. The method includes: a) providing a DNA sequence comprising at least two binding sites for

specifically interacting with the chimeric protein described herein; and b) reacting the DNA sequence with the chimeric protein, wherein the chimeric protein catalyzes a site-specific recombination event in which both strands of the DNA sequence are cleaved between the two sites specifically interacting with the chimeric protein.

BRIEF DESCRIPTION OF THE DRAWINGS

[0030] Figure 1 is a series of graphical and diagrammatic representations regarding TALER fusion orientation and activity. A) Cartoon illustrating the split β-lactamase system used to evaluate TALER activity. B) Schematic showing the fusion orientation of each TALER and its corresponding target site (1=SEQ ID NO: 288; 2=SEQ ID NO: 289; 3=SEQ ID NO: 290). C) Activity of each designed TALER fusion against its intended DNA target. Recombination was normalized to background (vector only control). D) Gin-Avr activity against cognate (Avr-20G) and non-cognate (Avr-20T, Avr-20GG, PthXo1-20G) DNA targets. Error bars indicate standard deviation (s.d.) (n = 3).

[0031] Figure 2 is a series of graphical and diagrammatic representations regarding recombination profiles of selected TALER truncations. A) Schematic illustrating the design of the 20-member TALER truncation library. B) Activity of selected TALER variants against DNA targets containing core sequences of increasing length (14, 20, 26, 32 and 44-bp). C) Gin-AvrXa7Δ120 activity against a diverse panel of substrates containing non-cognate cores sequences or core sites of increasing length. Error bars indicate s.d. (n = 3).

[0032] Figure 3 is a series of graphical representations regarding TALER variants selected from incremental truncation library. A) Frequency of selected TALER truncation variants. After 3 rounds of selection, incrementally truncated Gin-AvrXa7 variants were isolated and DNA sequencing was used to determine truncation length. B) Activity of incrementally truncated TALER variants (between Δ92 and Δ134 in length) against the Avr-32G DNA target. For reference, the shortest (Δ145) and longest (Δ74) truncation variants, as well as Δ87 were included. C) Activity of Gin-AvrΔ74, Gin-AvrΔ128 and Gin-AvrΔ145 against a diverse panel of cognate and non-cognate DNA targets. Error bars indicate s.d. (n = 3).

[0033] Figure 4 is a series graphical representations regarding activity of synthetic TALERs. A) Activity of synthetic Gin-Avr15Δ128, Gin-Avr15Δ120 and Gin-Pht15Δ120 variants against the DNA targets Avr-32G or Pth-32G. B) Activity of synthetic TALERs with DBDs between 15 and 20 repeats in length based on Gin-AvrΔ120 against Avr-32G and Avr-32T. Error bars indicate s.d. (n = 3).

[0034] Figure 5 is a series of graphical representations regarding TALER activity in mammalian cells. (A, B) Fold-reduction of luciferase expression in HEK293T cells co-transfected with (A) TALER or ZFR expression vectors (Gin-AvrΔ120 and GinC4) in the presence of reporter plasmid (Avr-32G, Avr-44G and C4-20G) or (B) TALER and ZFR expression vector in combination (Gin-AvrΔ120 + GinC4) with reporter plasmid (Avr-G-ZF). Error bars indicate s.d. (n = 3).

[0035] Figure 6 is a diagrammatic representation of location of primers for N-terminal designed truncations of AvrXa7 (SEQ ID NO: 1 DNA sequence; SEQ ID NO: 2 amino acid sequence). Star denotes the location of Δ120 fusion point.

[0036] Figure 7 is a diagrammatic representation of a comparison of native wild-type and synthetic RDV domains for the AvrXa7 target sequence (SEQ ID NOs: 16-18).

[0037] Figure 8 is a diagrammatic representation of TALE and TALER amino acid sequences of AvrXa7 protein (SEQ ID NO: 19).

[0038] Figure 9 is a diagrammatic representation of construct AvrXa7 DNA sequence (SEQ ID NO: 20).

[0039] Figure 10. is a diagrammatic representation of construct Gin-AvrΔ74 amino acid sequence (SEQ ID NO: 21).

[0040] Figure 11 is a diagrammatic representation of construct Gin-AvrΔ87 amino acid sequence (SEQ ID NO: 22).

[0041] Figure 12 is a diagrammatic representation of construct Gin-AvrΔ120 amino acid sequence (SEQ ID NO: 23).

[0042] Figure 13 is a diagrammatic representation of construct Gin-AvrΔ120* amino acid sequence (SEQ ID NO: 24).

[0043] Figure 14 is a diagrammatic representation of construct Gin-AvrΔ147 amino acid sequence (SEQ ID NO: 25).

[0044] Figure 15 is a diagrammatic representation of construct GinAvr15Δ128-synthetic protein amino acid sequence (SEQ ID NO: 26).

[0045] Figure 16 is a diagrammatic representation of construct Gin-Avr15Δ128-synthetic protein DNA sequence (SEQ ID NO: 27).

[0046] Figure 17 is a diagrammatic representation of construct GinAvr15Δ128-synthetic protein amino acid sequence (SEQ ID NO: 28).

[0047] Figure 18 is a series of pictorial and graphical representations pertaining to the specificity of the TALE N-terminal domain. A) Illustration of a TALE (SEQ ID NO: 29)

bound to its target DNA. B) Structural analysis suggests contact of the 5' T by W232 of the N-1 hairpin (N-0 - SEQ ID NO: 30; N-1 – SEQ ID NO: 31; and RVD – SEQ ID NO: 32). This hairpin shares significant sequence homology with RVD hairpins. C-F) Analyses of NT-T (wt) NTD in the context of C) AvrXa7 TALE-R, D) AvrXa7 TALE-TF, E) AvrXa7 MBPTALE, and F) a CCR5 targeting TALEN. (* = $p<0.05$, ** = $p< 0.01$, *** = $p<0.001$ compared to 5'T).

[0048] Figure 19 is a series of graphical and diagrammatic representations pertaining to recombinase variants. A-C) Activities of recombinase selection variants against substrates with A) 5' G, B) 5' A, and C) 5' C. Figure 18D is an alignment of optimized TALE NTDs SEQ ID NOs: 33-36), illustrating sequence differences in the N-1 hairpin. E) Comprehensive comparison of optimized NTD activities in the context of MBP-TALE AvrXa7. (* = $p<0.05$, ** = $p< 0.01$, *** = $p<0.001$, compared to wild type and 5'A/G/C).

[0049] Figure 20 is a series of diagrammatic and graphical representations of analysis of selected NTDs in the context of TALE-TFs. A) Illustration of 5xAvr promoter region (SEQ ID NO: 37) on the luciferase reporter plasmid used for transcription activation experiments. B) Relative luciferase activation of substrates with indicated 5' residues by TALE-TFs with NT-T, NT-G, NT-áN, and NT-âN domains. (* = $p<0.05$, ** = $p< 0.01$, *** = $p<0.001$, compared to NT-T and respective 5'A/G/C/T).

[0050] Figure 21 is a series of diagrammatic and graphical representations of design and activity of TALEN pairs with wild-type and evolved NTD's with varying 5' bases. A) The CCR5 gene (SEQ ID NOs: 38-39) expanded to highlight the target site (SEQ ID NOs: 40-47) for induction of the H32 mutation. B) Gene editing efficiency of the wild type (NT-T) TALEN, TALENs with domains optimized for non-T 5' residues, and dHax3 NTD. C) Fold enhancement of the TALEN pairs with optimized NTD vs. TALENs with 5' T specificity. The activity of each NTD is shown on each TALEN pair substrate.

[0051] Figure 22 is a diagrammatic representation showing alignment of N- and C-terminal domains SEQ ID NOs: 48-53).

[0052] Figure 23 is a schematic representation illustrating TALE-Recombinase selection protocol. A library of NTD was cloned into Avr15 TALE-R using Not1/Stu1 restriction enzymes and complementary ligation. Active TALE-R's result in more frequent recombination events that can be selected and amplified with antibiotics (carbenecillin). The resulting output plasmid was the digested Not1/Xba1 and ligated into the TALE-R backbone vector for further selection and amplification.

[0053] Figure 24 is a diagrammatic representation of a summary of variant populations discovered from library selections (Library XXXSGAR (SEQ ID NO: 39) and Library KXXGAR (SEQ ID NO: 291)).

[0054] Figure 25 is a diagrammatic representation showing alignment of NT-G (SEQ ID NO: 54) with NTD-Brg11 (SEQ ID NO: 55), a *Ralstonia* TALE domain. Alignment indicates Brg11 could exhibit specificity for 5' G bases.

[0055] Figure 26 is a series of graphical representations of relative binding affinity of MBP-TALE proteins to target 5' A/G/C/T Avr15 hairpin oligonucleotides as assayed by ELISA. Protein concentrations were ~ 75nM and plates were developed for 120 minutes.

[0056] Figure 27 is a series of pictorial and graphical representations of a cel1 assay of PCR amplified CCR5 after TALEN editing with % indels and indel populations shown on the right.

[0057] Figure 28 is a diagrammatic representation showing alignment indel sequencing of selected TALEN experiments from Figure 27 (SEQ ID NOs: 292-332 from top to bottom).

[0058] Figure 29 is a graphical representation of a comparison of the activity of two separate Goldy TALE-Transcription factor architectures, each targeting identical 5x AvrXa7 promoters varying only in the 5' residue.

[0059] Figure 30 is a series of diagrammatic representations relating to the structure of the zinc-finger recombinase dimer bound to DNA. A) Each zinc-finger recombinase (ZFR) monomer (blue or orange) consists of an activated serine recombinase catalytic domain linked to a custom-designed zinc-finger DNA-binding domain. Model was generated from crystal structures of the γδ resolvase and Aart zinc-finger protein (PDB IDs: 1GDT and 2I13, respectively). B) Cartoon of the ZFR dimer bound to DNA (SEQ ID NOs: 333-334). ZFR target sites consist of two-inverted zinc-finger binding sites (ZFBS) flanking a central 20-bp core sequence recognized by the ZFR catalytic domain. Zinc-finger proteins (ZFPs) can be designed to recognize 'left' or 'right' half-sites (blue and orange boxes, respectively). Abbreviations are as follows: N indicates A, T, C, or G; R indicates G or A; and Y indicates C or T.

[0060] Figure 31 is a series of graphical and diagrammatic representations of specificity of the Gin recombinase catalytic domain. A-D) Recombination was measured on DNA targets that contained (A, SEQ ID NO: 335) each possible two-base combination at the dinucleotide core, (B, SEQ ID NO: 336) each possible two-base combination at positions 3 and 2, (C, SEQ ID NO: 337) each possible single-base substitution at positions 6, 5, and 4, and (D, SEQ ID

NO: 338) each possible single-base substitution at positions 10, 9, 8, and 7. Substituted bases are boxed above each panel. Recombination was evaluated by split gene reassembly and measured as the ratio of carbenicillin-resistant to chloramphenicol-resistant transformants (Materials and Methods). Error bars indicate standard deviation (n = 3). (E) Interactions between the γδ resolvase dimer and DNA at (left) the dinucleotide core, (middle) positions 6, 5, and 4, and (right) positions 10, 9, 8, and 7 (PDB ID: 1GDT). Interacting residues are shown as purple sticks. Bases are colored as follows: A, yellow; T, blue; C, brown; and G, pink.

[0061] Figure 32 is a series of graphical and diagrammatic representations of re-engineering Gin recombinase catalytic specificity. A) The canonical 20-bp core recognized by the Gin catalytic domain. Positions 3 and 2 are boxed (SEQ ID NO: 339). B) (Top) Structure of the γδ resolvase in complex with DNA (PDB ID: 1GDT). Arm region residues selected for mutagenesis are shown as purple sticks. (Bottom) Sequence alignment of the γδ resolvase (SEQ ID NO: 341) and Gin recombinase (SEQ ID NO: 342) catalytic domains. Conserved residues are shaded orange. Black arrows indicate arm region positions selected for mutagenesis. C) Schematic representation of the split gene reassembly selection system. Expression of active ZFR variants leads to restoration of the β-lactamase reading frame and host-cell resistance to ampicillin. Solid lines indicate the locations and identity of the ZFR target sites. Positions 3 and 2 are underlined (SEQ ID NO: 340). D) Selection of Gin mutants that recombine core sites containing GC, GT, CA, TT, and AC base combinations at positions 3 and 2. Asterisks indicate selection steps in which incubation time was decreased from 16 hr to 6 hr (Materials and Methods, Example 5). E) Recombination specificity of the selected catalytic domains (β, γ, δ, ε, and ζ, wild-type Gin indicated by α) for each possible two-base combination at positions 3 and 2. Intended DNA targets are underlined. Recombination was determined by split gene reassembly and performed in triplicate.

[0062] Figure 33 is a series of graphical and diagrammatic representations illustrating the ability of ZFRs to recombine user-defined sequences in mammalian cells. A) Schematic representation of the luciferase reporter system used to evaluate ZFR activity in mammalian cells. ZFR target sites flank an SV40 promoter that drives luciferase expression. Solid lines denote the 44-bp consensus target sequence used to identify potential ZFR target sites. Underlined bases indicate zinc-finger targets and positions 3 and 2 (SEQ ID NO: 343). B) Fold-reduction of luciferase expression in HEK293T cells co-transfected with designed ZFR pairs and their cognate reporter plasmid. Fold-reduction was normalized to transfection with

empty vector and reporter plasmid. The sequence identity and chromosomal location of each ZFR target site (SEQ ID NOs: 344-362 top to bottom) and the catalytic domain composition of each ZFR pair are shown. Underlined bases indicate positions 3 and 2. Standard errors were calculated from three independent experiments. ZFR amino acid sequences are provided in Table 2. C) Specificity of ZFR pairs. Fold-reduction of luciferase expression was measured for ZFR pairs 1 through 9 and GinC4 for each non-cognate reporter plasmid. Recombination was normalized to the fold-reduction of each ZFR pair with its cognate reporter plasmid. Assays were performed in triplicate.

[0063] Figure 34 is a series of graphical and diagrammatic representations illustrating ZFRs ability to target integration into the human genome. A) Schematic representation of the donor plasmid (top) and the genomic loci targeted by ZFRs 1 (SEQ ID NO: 363), 2 (SEQ ID NO: 364), and 3 (SEQ ID NO: 365). Open boxes indicate neighboring exons. Arrows indicate transcript direction. The sequence and location of each ZFR target are shown. Underlined bases indicate zinc-finger targets and positions 3 and 2. B) Efficiency of ZFR-mediated integration. Data were normalized to data from cells transfected with donor plasmid only. Error bars indicate standard deviation (n = 3). C) PCR analysis of ZFR-mediated integration. PCR primer combinations amplified (top) unmodified locus or integrated plasmid in (middle) the forward or (bottom) the reverse orientation. D) Representative chromatograms of PCR-amplified integrated donor for ZFRs 1 (SEQ ID NO: 366) and 3 (SEQ ID NO: 367). Arrows indicate sequencing primer orientation. Shaded boxes denote genomic target sequences.

[0064] Figure 35 is a diagrammatic representation of recombinase DNA-binding residues are located outside the dimer interface. The γδ resolvase in complex with target DNA. Catalytic domain dimer is colored cyan. DNA is colored grey. Arm region residues are shown as red sticks. Residues at the dimer interface are shown as purple sticks (PDB ID: 1GDT).

[0065] Figure 36 is a diagrammatic representation of sequence analysis of selected recombinases. Pie charts showing the percentage of amino acid substitutions at each targeted arm position. After the 4^{th} round of selection, >20 clones were sequenced from each library. Sequence analysis of clones that recombine TT are described elsewhere(1).

[0066] Figure 37 is a table showing core specificity of isolated catalytic domains. After 4 rounds of selection, the ability of selected catalytic domains to recombine core sequences with substitutions at positions 3 and 2 was evaluated. Assigned DNA targets are underlined. Recombinase mutations are shown. Asterisks indicate catalytic domains selected for further analysis. Wild-type base combination at positions 3 and 2 is CC. Recombination was

determined by split gene reassembly(2) and performed in triplicate. Catalytic domains that recombine TT substitutions are described elsewhere(1).

[0067] Figure 38 is a series of graphical representations of position specificity of selected catalytic domains. Recombination assays between the α, β, γ, δ and ζ catalytic domains and symmetrically substituted target sites. Recombination was measured on a library DNA targets that contained (A (SEQ ID NO: 368)) >4,000 random strong base (S: G or C) substitutions at positions 6, 5 and 4 and (B (SEQ ID NO: 369)) $>10^6$ (of a possible 4.29 x 10^9) unique base combinations at positions 10, 9, 8 and 7 (N: A, T, C or G). Recombination was measured by split gene reassembly(2) ($n = 3$).

[0068] Figure 39 is a series of graphical representations of ZFR homodimer activity. HEK293T cells were co-transfected with 150 ng ZFR-L or 150 ng ZFR-R with 2.5 ng of corresponding pGL3 ZFR reporter plasmid. Recombination was normalized to co-transfection with 150 ng ZFR-L and 150 ng ZFR-R with 2.5 ng pGL3 ZFR reporter plasmid.

[0069] Figure 40 is a series of pictorial representations depicting clonal analysis of ZFR-modified cells. PCR primer combinations amplified either unmodified genomic target or integrated plasmid in the forward or reverse orientation.

DETAILED DESCRIPTION OF THE INVENTION

[0070] The present provides the first disclosure of a TALE recombinase (TALER). Using a library of incrementally truncated TALE domains, optimized TALER architecture that can be used to recombine DNA in bacterial and mammalian cells was identified. Any customized TALE repeat array can be inserted into the TALER architecture described herein, thus dramatically expanding the targeting capacity of engineered recombinases for applications in biotechnology and medicine.

[0071] Transcription activator-like effector (TALE) proteins can be designed to bind virtually any DNA sequence. General guidelines for design of TALE DNA-binding domains suggest that the 5'-most base of the DNA sequence bound by the TALE (the N_0 base) should be a thymine. The N_0 requirement was quantified by analysis of the activities of TALE transcription factors (TALE-TF), TALE recombinases (TALE-R) and TALE nucleases (TALENs) with each DNA base at this position. In the absence of a 5' T, decreases in TALE activity up to >1000-fold in TALE-TF activity, up to 100-fold in TALE-R activity and up to 10-fold reduction in TALEN activity compared with target sequences containing a 5' T was observed. To develop TALE architectures that recognize all possible N_0 bases, structure-guided library design coupled with TALE-R activity selections were used to evolve novel

TALE N-terminal domains to accommodate any N0 base. A G-selective domain and broadly reactive domains were isolated and characterized. The engineered TALE domains selected in the TALE-R format demonstrated modularity and were active in TALE-TF and TALEN architectures. Evolved N-terminal domains provide effective and unconstrained TALE-based targeting of any DNA sequence as TALE binding proteins and designer enzymes.

[0072] Additionally, in order to address sequence requirement limitations, a knowledge-base approach was described for re-engineering serine recombinase catalytic specificity. This strategy, which was based on the saturation mutagenesis of specificity-determining DNA-binding residues, was used to generate recombinase variants that showed a >10,000-fold shift in specificity. Importantly, this approach focused exclusively on amino acid residues located outside the recombinase dimer interface (Figure 35). As a result, it was determined that re-engineered catalytic domains could associate to form ZFR heterodimers and that these designed ZFR pairs recombine pre-determined DNA sequences with exceptional specificity. Together, these results led us to hypothesize that an expanded catalog of specialized catalytic domains developed by this method could be used to generate ZFRs with custom specificity. Here, a combination of substrate specificity analysis and directed evolution is used to develop a diverse collection of Gin recombinase catalytic domains that are capable of recognizing an estimated 4×10^8 unique 20-bp core sequences. It is shown that ZFRs assembled from these re-engineered catalytic domains recombine user-defined sequences with high specificity and integrate DNA into targeted endogenous loci in human cells. These results demonstrate the potential of ZFR technology for a wide variety of applications, including genome engineering and gene therapy.

[0073] Before the present compositions and methods are described, it is to be understood that this invention is not limited to the particular compositions, methods, and experimental conditions described, as such devices, methods, and conditions may vary. It is also to be understood that the terminology used herein is for purposes of describing particular embodiments only, and is not intended to be limiting, since the scope of the present invention will be limited only in the appended claims.

[0074] As used in this specification and the appended claims, the singular forms “a”, “an”, and “the” include plural references unless the context clearly dictates otherwise. Thus, for example, references to “the composition” or “the method” includes one or more compositions and methods, and/or steps of the type described herein which will become apparent to those persons skilled in the art upon reading this disclosure and so forth.

[0075] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of the invention, the preferred methods and materials are now described.

[0076] "Recombinases" are a family of enzymes that mediate site-specific recombination between specific DNA sequences recognized by the recombinase (Esposito, D., and Scocca, J. J., Nucleic Acids Research 25, 3605-3614 (1997); Nunes-Duby, S. E., et al., Nucleic Acids Research 26, 391-406 (1998); Stark, W. M., et al., Trends in Genetics 8, 432-439 (1992)).

[0077] As used herein, the term "chimeric TALE recombinase" includes without limitation recombinases having a TALE domain derived from a naturally-occurring TALE protein or a synthetically derived TALE protein or domain with sequence-specific binding activity.

[0078] As used herein, the term "chimeric zinc finger recombinase" includes without limitation recombinases having a zinc finger binding domain derived from a naturally-occurring zinc finger DNA binding protein or a synthetically derived zinc finger binding protein or domain with sequence-specific binding activity.

[0079] As used herein, the term "zinc finger," "zinc finger nucleotide binding domain," or similar terminology refers both to naturally occurring and artificially produced zinc fingers. Such zinc fingers can have various framework structures, such as, but not limited to, C2H2, C4, H4, H3C, C3X, H3X, C2X2, and H2X2, where X is a zinc ligating amino acid. In these framework structures, as is conventional in the recitation of zinc finger structures, "C" represents a cysteine residue and "H" represents a histidine residue. Zinc fingers having the framework C2H2 include, but are not limited to, zinc fingers described, for example, in International Publication Number WO2008/006028 to Barbas et al., United States Patent No. 7,101 ,972 to Barbas, United States Patent No. 7,067,617 to Barbas et al., United States Patent No. 6,790,941 to Barbas et al., United States Patent No. 6,610,512 to Barbas, United States Patent No. 6,242,568 to Barbas et al., United States Patent No. 6,140,466 to Barbas et al., United States Patent No. 6,140,081 to Barbas, United States Patent Application Publication No. 20060223757 by Barbas, United States Patent Application Publication No. 20060211846 by Barbas et al., United States Patent Application Publication No. 20060078880 by Barbas et al., United States Patent Application Publication No. 20050148075 by Barbas, United States Patent Application Publication No. 20050084885 by

Barbas et al., United States Patent Application Publication No. 20040224385 by Barbas et al., United States Patent Application Publication No. 20030059767 by Barbas et al., and United States Patent Application Publication No. 20020165356 by Barbas et al., all of which are incorporated herein by this reference. Other zinc fingers are described in: U.S. Patent No. 7,067,317 to Rebar et al.; U.S. Patent No. 7,030,215 to Liu et al.; U.S. Patent No. 7,026,462 to Rebar et al.; U.S. Patent No. 7,013,219 to Case et al.; U.S. Patent No. 6,979,539 to Cox III et al.; U.S. Patent No. 6,933,113 to Case et al.; U.S. Patent No. 6,824,978 to Cox III et al.; U.S. Patent No. 6,794,136 to Eisenberg et al.; U.S. Patent No. 6,785,613 to Eisenberg et al.; U.S. Patent No. 6,777,185 to Case et al.; U.S. Patent No. 6,706,470 to Choo et al.; U.S. Patent No. 6,607,882 to Cox IM et al.; U.S. Patent No. 6,599,692 to Case et al.; U.S. Patent No. 6,534,261 to Cox II et al.; U.S. Patent No. 6,503,717 to Case et al.; U.S. Patent No. 6,453,242 to Eisenberg et al.; United States Patent Application Publication No. 2006/0246588 to Rebar et al.; United States Patent Application Publication No. 2006/0246567 to Rebar et al.; United States Patent Application Publication No. 2006/0166263 to Case et al.; United States Patent Application Publication No. 2006/0078878 to Cox HI et at.; United States Patent Application Publication No. 2005/0257062 to Rebar et al.; United States Patent Application Publication No. 2005/0215502 to Cox III et al.; United States Patent Application Publication No. 2005/0130304 to Cox Ml et al.; United States Patent Application Publication No. 2004/0203064 to Case et al.; United States Patent Application Publication No. 2003/0166141 to Case et al.; United States Patent Application Publication No. 2003/0134318 to Case et al.; United States Patent Application Publication No. 2003/0105593 to Eisenberg et al.; United States Patent Application Publication No. 2003/0087817 to Cox IM et al.; United States Patent Application Publication No. 2003/0021776 to Rebar et al.; and United States Patent Application Publication No. 2002/0081614 to Case et al., all of which are incorporated herein by this reference. For example, one alternative described in these patents and patent publications involves the use of so-called "D-able sites" and zinc finger modules or zinc finger DNA binding domains that can bind to such sites. A "D-able" site is a region of a target site that allows an appropriately designed zinc finger module or zinc finger DNA binding domain to bind to four bases rather than three of the target strand. Such a zinc finger module or zinc finger DNA binding domain binds to a triplet of three bases on one strand of a double-stranded DNA target segment (target strand) and a fourth base on the other, complementary, strand. Binding of a single zinc

finger to a four base target segment imposes constraints both on the sequence of the target strand and on the amino acid sequence of the zinc finger.

[0080] As used herein, the amino acids, which occur in the various amino acid sequences appearing herein, are identified according to their well-known, three-letter or one-letter abbreviations. The nucleotides, which occur in the various DNA fragments, are designated with the standard single-letter designations used routinely in the art.

[0081] In a peptide or protein, suitable conservative substitutions of amino acids are known to those of skill in this art and may be made generally without altering the biological activity of the resulting molecule. Those of skill in this art recognize that, in general, single amino acid substitutions in non-essential regions of a polypeptide do not substantially alter biological activity (see, e.g. Watson et al. Molecular Biology of the Gene, 4th Edition, 1987, Benjamin/Cummings, p. 224). In particular, such a conservative variant has a modified amino acid sequence, such that the change(s) do not substantially alter the protein's (the conservative variant's) structure and/or activity, e.g., antibody activity, enzymatic activity, or receptor activity. These include conservatively modified variations of an amino acid sequence, i.e., amino acid substitutions, additions or deletions of those residues that are not critical for protein activity, or substitution of amino acids with residues having similar properties (e.g., acidic, basic, positively or negatively charged, polar or non-polar, etc.) such that the substitutions of even critical amino acids does not substantially alter structure and/or activity. Conservative substitution tables providing functionally similar amino acids are well known in the art. For example, one exemplary guideline to select conservative substitutions includes (original residue followed by exemplary substitution): Ala/Gly or Ser; Arg/Lys; Asn/Gln or His; Asp/Glu; Cys/Ser; Gln/Asn; Gly/Asp; Gly/Ala or Pro; His/Asn or Gln; Ile/Leu or Val; Leu/Ile or Val; Lys/Arg or Gln or Glu; Met/Leu or Tyr or Ile; Phe/Met or Leu or Tyr; Ser/Thr; Thr/Ser; Trp/Tyr; Tyr/Trp or Phe; Val/Ile or Leu. An alternative exemplary guideline uses the following six groups, each containing amino acids that are conservative substitutions for one another: (1) alanine (A or Ala), serine (S or Ser), threonine (T or Thr); (2) aspartic acid (D or Asp), glutamic acid (E or Glu); (3) asparagine (N or Asn), glutamine (Q or Gln); (4) arginine (R or Arg), lysine (K or Lys); (5) isoleucine (I or Ile), leucine (L or Leu), methionine (M or Met), valine (V or Val); and (6) phenylalanine (F or Phe), tyrosine (Y or Tyr), tryptophan (W or Trp); (see also, e.g., Creighton (1984) Proteins, W. H. Freeman and Company; Schulz and Schimer (1979) Principles of Protein Structure, Springer-Verlag). One of skill in the art will appreciate that the above-identified substitutions are not the only

possible conservative substitutions. For example, for some purposes, one may regard all charged amino acids as conservative substitutions for each other whether they are positive or negative. In addition, individual substitutions, deletions or additions that alter, add or delete a single amino acid or a small percentage of amino acids in an encoded sequence can also be considered "conservatively modified variations" when the three-dimensional structure and the function of the protein to be delivered are conserved by such a variation.

[0082] As used herein, the term "expression vector" refers to a plasmid, virus, phagemid, or other vehicle known in the art that has been manipulated by insertion or incorporation of heterologous DNA, such as nucleic acid encoding the fusion proteins herein or expression cassettes provided herein. Such expression vectors typically contain a promoter sequence for efficient transcription of the inserted nucleic acid in a cell. The expression vector typically contains an origin of replication, a promoter, as well as specific genes that permit phenotypic selection of transformed cells.

[0083] As used herein, the term "host cells" refers to cells in which a vector can be propagated and its DNA expressed. The term also includes any progeny of the subject host cell. It is understood that all progeny may not be identical to the parental cell since there may be mutations that occur during replication. Such progeny are included when the term "host cell" is used. Methods of stable transfer where the foreign DNA is continuously maintained in the host are known in the art.

[0084] As used herein, genetic therapy involves the transfer of heterologous DNA to the certain cells, target cells, of a mammal, particularly a human, with a disorder or conditions for which such therapy is sought. The DNA is introduced into the selected target cells in a manner such that the heterologous DNA is expressed and a therapeutic product encoded thereby is produced. Alternatively, the heterologous DNA may in some manner mediate expression of DNA that encodes the therapeutic product, or it may encode a product, such as a peptide or RNA that in some manner mediates, directly or indirectly, expression of a therapeutic product. Genetic therapy may also be used to deliver nucleic acid encoding a gene product that replaces a defective gene or supplements a gene product produced by the mammal or the cell in which it is introduced. The introduced nucleic acid may encode a therapeutic compound, such as a growth factor inhibitor thereof, or a tumor necrosis factor or inhibitor thereof, such as a receptor therefor, that is not normally produced in the mammalian host or that is not produced in therapeutically effective amounts or at a therapeutically useful time. The heterologous DNA encoding the therapeutic product may be modified prior to

introduction into the cells of the afflicted host in order to enhance or otherwise alter the product or expression thereof. Genetic therapy may also involve delivery of an inhibitor or repressor or other modulator of gene expression.

[0085] As used herein, heterologous DNA is DNA that encodes RNA and proteins that are not normally produced in vivo by the cell in which it is expressed or that mediates or encodes mediators that alter expression of endogenous DNA by affecting transcription, translation, or other regulatable biochemical processes. Heterologous DNA may also be referred to as foreign DNA. Any DNA that one of skill in the art would recognize or consider as heterologous or foreign to the cell in which is expressed is herein encompassed by heterologous DNA. Examples of heterologous DNA include, but are not limited to, DNA that encodes traceable marker proteins, such as a protein that confers drug resistance, DNA that encodes therapeutically effective substances, such as anti-cancer agents, enzymes and hormones, and DNA that encodes other types of proteins, such as antibodies. Antibodies that are encoded by heterologous DNA may be secreted or expressed on the surface of the cell in which the heterologous DNA has been introduced.

[0086] Hence, herein heterologous DNA or foreign DNA, includes a DNA molecule not present in the exact orientation and position as the counterpart DNA molecule found in the genome. It may also refer to a DNA molecule from another organism or species (i.e., exogenous).

[0087] As used herein, a therapeutically effective product is a product that is encoded by heterologous nucleic acid, typically DNA, that, upon introduction of the nucleic acid into a host, a product is expressed that ameliorates or eliminates the symptoms, manifestations of an inherited or acquired disease or that cures the disease. Typically, DNA encoding a desired gene product is cloned into a plasmid vector and introduced by routine methods, such as calcium-phosphate mediated DNA uptake (see, (1981) Somat. Cell. Mol. Genet. 7:603-616) or microinjection, into producer cells, such as packaging cells. After amplification in producer cells, the vectors that contain the heterologous DNA are introduced into selected target cells.

[0088] As used herein, an expression or delivery vector refers to any plasmid or virus into which a foreign or heterologous DNA may be inserted for expression in a suitable host cell--i.e., the protein or polypeptide encoded by the DNA is synthesized in the host cell's system. Vectors capable of directing the expression of DNA segments (genes) encoding one or more

proteins are referred to herein as "expression vectors". Also included are vectors that allow cloning of cDNA (complementary DNA) from mRNAs produced using reverse transcriptase.

[0089] As used herein, a gene refers to a nucleic acid molecule whose nucleotide sequence encodes an RNA or polypeptide. A gene can be either RNA or DNA. Genes may include regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

[0090] As used herein, the term "isolated" with reference to a nucleic acid molecule or polypeptide or other biomolecule means that the nucleic acid or polypeptide has been separated from the genetic environment from which the polypeptide or nucleic acid were obtained. It may also mean that the biomolecule has been altered from the natural state. For example, a polynucleotide or a polypeptide naturally present in a living animal is not "isolated," but the same polynucleotide or polypeptide separated from the coexisting materials of its natural state is "isolated," as the term is employed herein. Thus, a polypeptide or polynucleotide produced and/or contained within a recombinant host cell is considered isolated. Also intended as an "isolated polypeptide" or an "isolated polynucleotide" are polypeptides or polynucleotides that have been purified, partially or substantially, from a recombinant host cell or from a native source. For example, a recombinantly produced version of a compound can be substantially purified by the one-step method described in Smith et al. (1988) Gene 67:3140. The terms isolated and purified are sometimes used interchangeably.

[0091] Thus, by "isolated" is meant that the nucleic acid is free of the coding sequences of those genes that, in a naturally-occurring genome immediately flank the gene encoding the nucleic acid of interest. Isolated DNA may be single-stranded or double-stranded, and may be genomic DNA, cDNA, recombinant hybrid DNA, or synthetic DNA. It may be identical to a native DNA sequence, or may differ from such sequence by the deletion, addition, or substitution of one or more nucleotides.

[0092] "Isolated" or "purified" as those terms are used to refer to preparations made from biological cells or hosts means any cell extract containing the indicated DNA or protein including a crude extract of the DNA or protein of interest. For example, in the case of a protein, a purified preparation can be obtained following an individual technique or a series of preparative or biochemical techniques and the DNA or protein of interest can be present at various degrees of purity in these preparations. Particularly for proteins, the procedures may include for example, but are not limited to, ammonium sulfate fractionation, gel filtration, ion

exchange change chromatography, affinity chromatography, density gradient centrifugation, electrofocusing, chromatofocusing, and electrophoresis.

[0093] A preparation of DNA or protein that is "substantially pure" or "isolated" should be understood to mean a preparation free from naturally occurring materials with which such DNA or protein is normally associated in nature. "Essentially pure" should be understood to mean a "highly" purified preparation that contains at least 95% of the DNA or protein of interest.

[0094] A cell extract that contains the DNA or protein of interest should be understood to mean a homogenate preparation or cell-free preparation obtained from cells that express the protein or contain the DNA of interest. The term "cell extract" is intended to include culture media, especially spent culture media from which the cells have been removed.

[0095] As used herein, a promoter region of a gene includes the regulatory element or elements that typically lie 5′ to a structural gene; multiple regulatory elements can be present, separated by intervening nucleotide sequences. If a gene is to be activated, proteins known as transcription factors attach to the promoter region of the gene. This assembly resembles an "on switch" by enabling an enzyme to transcribe a second genetic segment from DNA into RNA. In most cases the resulting RNA molecule serves as a template for synthesis of a specific protein; sometimes RNA itself is the final product. The promoter region may be a normal cellular promoter or, for example, an onco-promoter. An onco-promoter is generally a virus-derived promoter. Viral promoters to which zinc finger binding polypeptides may be targeted include, but are not limited to, retroviral long terminal repeats (LTRs), and Lentivirus promoters, such as promoters from human T-cell lymphotrophic virus (HTLV) 1 and 2 and human immunodeficiency virus (HIV) 1 or 2.

[0096] As used herein, the term "truncated" or similar terminology refers to a polypeptide derivative that contains less than the full amino acid sequence of a native protein, such as a ZFP, TALE or serine recombinase.

[0097] As used herein, a polypeptide "variant" or "derivative" refers to a polypeptide that is a mutagenized form of a polypeptide or one produced through recombination but that still retains a desired activity, such as the ability to bind to a ligand or a nucleic acid molecule or to modulate transcription.

[0098] As used herein, the terms "pharmaceutically acceptable", "physiologically tolerable" and grammatical variations thereof, as they refer to compositions, carriers, diluents and reagents, are used interchangeably and represent that the materials are capable of

administration to or upon a human without the production of undesirable physiological effects such as nausea, dizziness, gastric upset and the like which would be to a degree that would prohibit administration of the composition.

[0099] As used herein, the term "vector" refers to a nucleic acid molecule capable of transporting between different genetic environments another nucleic acid to which it has been operatively linked. Preferred vectors are those capable of autonomous replication and expression of structural gene products present in the DNA segments to which they are operatively linked. Vectors, therefore, preferably contain the replicons and selectable markers described earlier. Vectors include, but are not necessarily limited to, expression vectors.

[0100] As used herein with regard to nucleic acid molecules, including DNA fragments, the phrase "operatively linked" means the sequences or segments have been covalently joined, preferably by conventional phosphodiester bonds, into one strand of DNA, whether in single or double-stranded form such that operatively linked portions function as intended. The choice of vector to which transcription unit or a cassette provided herein is operatively linked depends directly, as is well known in the art, on the functional properties desired, e.g., vector replication and protein expression, and the host cell to be transformed, these being limitations inherent in the art of constructing recombinant DNA molecules.

[0101] As used herein, administration of a therapeutic composition can be effected by any means, and includes, but is not limited to, oral, subcutaneous, intravenous, intramuscular, intrasternal, infusion techniques, intraperitoneal administration and parenteral administration.

[0102] Methods of transforming cells are well known in the art. By "transformed" it is meant a heritable alteration in a cell resulting from the uptake of foreign DNA. Suitable methods include viral infection, transfection, conjugation, protoplast fusion, electroporation, particle gun technology, calcium phosphate precipitation, direct microinjection, and the like. The choice of method is generally dependent on the type of cell being transformed and the circumstances under which the transformation is taking place (i.e. in vitro, ex vivo, or in vivo). A general discussion of these methods can be found in Ausubel, et al, Short Protocols in Molecular Biology, 3rd ed., Wiley & Sons, 1995.

[0103] The terms "nucleic acid molecule" and "polynucleotide" are used interchangeably and refer to a polymeric form of nucleotides of any length, either deoxyribonucleotides or ribonucleotides, or analogs thereof. Polynucleotides may have any three-dimensional structure, and may perform any function, known or unknown. Non-limiting examples of polynucleotides include a gene, a gene fragment, exons, introns, messenger RNA (mRNA),

transfer RNA, ribosomal RNA, ribozymes, cDNA, recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, nucleic acid probes, and primers.

[0104] An "expression cassette" comprises any nucleic acid construct capable of directing the expression of a gene/coding sequence of interest. Such cassettes can be constructed into a "vector," "vector construct," "expression vector," or "gene transfer vector," in order to transfer the expression cassette into target cells. Thus, the term includes cloning and expression vehicles, as well as viral vectors.

[0105] Techniques for determining nucleic acid and amino acid "sequence identity" also are known in the art. Typically, such techniques include determining the nucleotide sequence of the mRNA for a gene and/or determining the amino acid sequence encoded thereby, and comparing these sequences to a second nucleotide or amino acid sequence. In general, "identity" refers to an exact nucleotide-to-nucleotide or amino acid-to-amino acid correspondence of two polynucleotides or polypeptide sequences, respectively. Two or more sequences (polynucleotide or amino acid) can be compared by determining their "percent identity." The percent identity of two sequences, whether nucleic acid or amino acid sequences, is the number of exact matches between two aligned sequences divided by the length of the shorter sequences and multiplied by 100. An approximate alignment for nucleic acid sequences is provided by the local homology algorithm of Smith and Waterman, Advances in Applied Mathematics 2:482-489 (1981). This algorithm can be applied to amino acid sequences by using the scoring matrix developed by Dayhoff, Atlas of Protein Sequences and Structure, M. O. Dayhoff ed., 5 suppl. 3:353-358, National Biomedical Research Foundation, Washington, D.C., USA, and normalized by Gribskov, Nucl. Acids Res. 14(6):6745-6763 (1986). An exemplary implementation of this algorithm to determine percent identity of a sequence is provided by the Genetics Computer Group (Madison, Wis.) in the "BestFit" utility application. The default parameters for this method are described in the Wisconsin Sequence Analysis Package Program Manual, Version 8 (1995) (available from Genetics Computer Group, Madison, Wis.). A preferred method of establishing percent identity in the context of the present invention is to use the MPSRCE package of programs copyrighted by the University of Edinburgh, developed by John F. Collins and Shane S. Sturrok, and distributed by IntelliGenetics, Inc. (Mountain View, Calif.). From this suite of packages the Smith-Waterman algorithm can be employed where default parameters are used for the scoring table (for example, gap open penalty of 12, gap extension penalty of one, and

a gap of six). From the data generated the "Match" value reflects "sequence identity." Other suitable programs for calculating the percent identity or similarity between sequences are generally known in the art, for example, another alignment program is BLAST, used with default parameters. For example, BLASTN and BLASTP can be used using the following default parameters: genetic code=standard; filter=none; strand=both; cutoff=60; expect=10; Matrix=BLOSUM62; Descriptions=50 sequences; sort by=HIGH SCORE; Databases=non-redundant, GenBank+EMBL+DDBJ+PDB+GenBank CDS translations+Swiss protein+Spupdate+PIR.

[0106] Alternatively, homology can be determined by hybridization of polynucleotides under conditions that form stable duplexes between homologous regions, followed by digestion with single-stranded-specific nuclease(s), and size determination of the digested fragments. Two DNA, or two polypeptide sequences are "substantially homologous" to each other when the sequences exhibit at least about 80%-85%, preferably at least about 85%-90%, more preferably at least about 90%-95%, and most preferably at least about 95%-98% sequence identity over a defined length of the molecules, as determined using the methods above. As used herein, substantially homologous also refers to sequences showing complete identity to the specified DNA or polypeptide sequence. DNA sequences that are substantially homologous can be identified in a Southern hybridization experiment under, for example, stringent conditions, as defined for that particular system. Defining appropriate hybridization conditions is within the skill of the art. See, e.g., Sambrook et al., supra; DNA Cloning, supra; Nucleic Acid Hybridization, supra.

[0107] As such, the invention provides nucleic acid and amino acid sequences encoding chimeric polypeptides of the invention which are substantially homologous and encode polypeptides that retain equivalent biological activity.

[0108] Two nucleic acid fragments are considered to "selectively hybridize" as described herein. The degree of sequence identity between two nucleic acid molecules affects the efficiency and strength of hybridization events between such molecules. A partially identical nucleic acid sequence will at least partially inhibit a completely identical sequence from hybridizing to a target molecule. Inhibition of hybridization of the completely identical sequence can be assessed using hybridization assays that are well known in the art (e.g., Southern blot, Northern blot, solution hybridization, or the like, see Sambrook, et al., Molecular Cloning: A Laboratory Manual, Second Edition, (1989) Cold Spring Harbor, N.Y.). Such assays can be conducted using varying degrees of selectivity, for example, using

conditions varying from low to high stringency. If conditions of low stringency are employed, the absence of non-specific binding can be assessed using a secondary probe that lacks even a partial degree of sequence identity (for example, a probe having less than about 30% sequence identity with the target molecule), such that, in the absence of non-specific binding events, the secondary probe will not hybridize to the target.

[0109] When utilizing a hybridization-based detection system, a nucleic acid probe is chosen that is complementary to a target nucleic acid sequence, and then by selection of appropriate conditions the probe and the target sequence "selectively hybridize," or bind, to each other to form a hybrid molecule. A nucleic acid molecule that is capable of hybridizing selectively to a target sequence under "moderately stringent" typically hybridizes under conditions that allow detection of a target nucleic acid sequence of at least about 10-14 nucleotides in length having at least approximately 70% sequence identity with the sequence of the selected nucleic acid probe. Stringent hybridization conditions typically allow detection of target nucleic acid sequences of at least about 10-14 nucleotides in length having a sequence identity of greater than about 90-95% with the sequence of the selected nucleic acid probe. Hybridization conditions useful for probe/target hybridization where the probe and target have a specific degree of sequence identity, can be determined as is known in the art (see, for example, Nucleic Acid Hybridization: A Practical Approach, editors B. D. Hames and S. J. Higgins, (1985) Oxford; Washington, D.C.; IRL Press).

[0110] With respect to stringency conditions for hybridization, it is well known in the art that numerous equivalent conditions can be employed to establish a particular stringency by varying, for example, the following factors: the length and nature of probe and target sequences, base composition of the various sequences, concentrations of salts and other hybridization solution components, the presence or absence of blocking agents in the hybridization solutions (e.g., formamide, dextran sulfate, and polyethylene glycol), hybridization reaction temperature and time parameters, as well as, varying wash conditions. The selection of a particular set of hybridization conditions is selected following standard methods in the art (see, for example, Sambrook, et al., Molecular Cloning: A Laboratory Manual, Second Edition, (1989) Cold Spring Harbor, N.Y.)

[0111] A first polynucleotide is "derived from" a second polynucleotide if it has the same or substantially the same basepair sequence as a region of the second polynucleotide, its cDNA, complements thereof, or if it displays sequence identity as described above.

[0112] A first polypeptide is "derived from" a second polypeptide if it is (i) encoded by a first polynucleotide derived from a second polynucleotide, or (ii) displays sequence identity to the second polypeptides as described above.

[0113] Site-specific recombinases are powerful tools for genome engineering. Hyperactivated variants of the resolvase/invertase family of serine recombinases function without accessory factors, and thus can be re-targeted to sequences of interest by replacing native DNA-binding domains with engineered zinc-finger proteins (ZFPs).

[0114] The zinc finger recombinases described herein are chimeric enzymes composed of an activated catalytic domain derived from the resolvase/invertase family of serine recombinases and a custom- designed zinc-finger DNA-binding domain. The ZFRs assembled from engineered catalytic domains efficiently recombine user-defined DNA targets with high specificity and designed ZFRs integrate DNA into targeted endogenous loci in human cells.

[0115] In one aspect, the invention provides a method of generating a plurality of zinc finger recombinase (ZFRs) proteins having catalytic specificity greater than the corresponding wild type recombinase. The method includes performing random mutagenesis on a recombinase catalytic domain at positions equivalent to Gin Ile120, Thr123, Leu127, Ile136 and Gly137 or a combination thereof with reference to a wild-type Gin catalytic domain, mutating the DNA at positions 2 and 3 for each amino acid; fusing the recombinase catalytic domain with a plurality of zinc finger binding domains to form ZFRs, and enriching for ZFRs having catalytic specificity greater than the corresponding wild type recombinase. In embodiments the ZFRs have increased catalytic activity on DNA targets selected from GC, GT, CA, TT and AC. In one embodiment, the recombinase catalytic domain is mutagenized at Ile136 and/or Gly137.

[0116] As used herein, a wild-type Gin catalytic domain refers to a Gin catalytic domain including all or a portion of a polypeptide having the amino acid sequence set forth as SEQ ID NO: 56 as follows:

MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLKRALKRLQKGD
TLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTDSIDTSSPMGRFFFYVMGALAE
MERELIIERTMAGLAAARNKGRIGGRPPKLTKAEWEQAGRLLAQGIPRKQVALIYDV
ALSTLYKKHP

[0117] In various embodiments, the chimeric polypeptides of the invention include a Gin catalytic domain, such as those generated by the method of the invention. Particular Gin catalytic domains include those set forth in Table 1.

[0118] Table 1. Gin catalytic domains.

Gin catalytic domains.		
Variant	SEQ ID NO:	Sequence
Gin α	57	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLKRA LKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTDSIDTS SPMGRFFFYVMGALAEMERELIIERTMAGIAAARNKGRIGGRPPKSG
Gin β	58	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLKRA LKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTDSIDTS SPMGRFFFYVMGALAEMERELIIERTMAGIAAARNKGRIFGRPPKS
Gin γ	59	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLKRA LKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTDSIDTS SPMGRFFFYVMGALAEMERELIIERVMAGIAAARNKGRIVGRPPKSG
Gin δ	60	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLKRA LKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTDSIDTS SPMGRFFFYVMGALAEMERELIIERVMAGIAAARNKGRIGGRPPKSG
Gin ε	61	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLKRA LKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTDSIDTS SPMGRFFFYVMGALAEMER**LS**IIERIMAGIAAARNKGRIFGRPPKSG
Gin ζ	62	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLKRA LKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTDSIDTS SPMGRFFFYVMGALAEMERELIIERTSAGIAAASNKGRIVGRPPKSG
Targeted arm region positions are highlighted. Random substitutions are emboldened and underlined. The hyperactivating H106Y mutation is underlined.		

[0119] In various embodiments, the ZFRs generated by the method of the invention include a Gin catalytic domain operatively linked to a plurality of zinc finger binding domains. Exemplary ZFRs generated by the invention include those set forth in Table 2.

[0120] Table 2. ZFRs.

Amino acid sequences of exemplary ZFRs.		
ZFR-1 Left	SEQ ID NO: 63	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRTGG RPPKSGTGEKPYKCPECGKSFSTSGNLVRHQRTHTGEKPYKCPECG KSFSQSGDLRRHQRTHTGEKPYKCPECGKSFSTSGNLVRHQRTHTG EKPYKCPECGKSFSTSGELVRHQRTHTGKKTSGQAGQ
ZFR-1 Right	SEQ ID NO: 64	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRTGG RPPKSGTGEKPYKCPECGKSFSHRTTLTNHQRTHTGEKPYKCPECG KSFSQSGDLRRHQRTHTGEKPYKCPECGKSFSQSGDLRRHQRTHTG EKPYKCPECGKSFSQSGDLRRHQRTHTGKKTSGQAGQ
ZFR-2 Left	SEQ ID NO: 65	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRTGG RPPKSGTGEKPYKCPECGKSFSQSGDLRRHQRTHTGEKPYKCPECG KSFSQRAHLERHQRTHTGEKPYKCPECGKSFSTSGNLVRHQRTHTG EKPYKCPECGKSFSRSDELVRHQRTHTGKKTSGQAGQ
ZFR-2 Right	SEQ ID NO: 66	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRTGG RPPKSGTGEKPYKCPECGKSFSRSDKLVRHQRTHTGEKPYKCPECG KSFSRKDNLKNHQRTHTGEKPYKCPECGKSFSTSGELVRHQRTHTG EKPYKCPECGKSFSRSDKLVRHQRTHTGKKTSGQAGQ
ZFR-3 Left	SEQ ID NO: 67	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRTGG

		RPPKSGTGEKPYKCPECGKSFSTTGNLTVHQRTHTGEKPYKCPECG KSFSDPGALVRHQRTHTGEKPYKCPECGKSFSQSSNLVRHQRTHTG EKPYKCPECGKSFSRSDHLTNHQRTHTGKKTSGQAGQ
ZFR-3 Right	SEQ ID NO: 68	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRIGG RPPKSGTGEKPYKCPECGKSFSRKDNLKNHQRTHTGEKPYKCPECG KSFSRSDHLTNHQRTHTGEKPYKCPECGKSFSDPGNLVRHQRTHTG EKPYKCPECGKSFSRKDNLKNHQRTHTGKKTSGQAGQ
ZFR-4 Left	SEQ ID NO: 69	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRIAM GRPPKSGTGEKPYKCPECGKSFSQRANLRAHQRTHTGEKPYKCPEC GKSFSQSSSLVRHQRTHTGEKPYKCPECGKSFSTTGNLTVHQRTHT GEKPYKCPECGKSFSQRAHLERHQRTHTGKKTSGQAGQ
ZFR-4 Right	SEQ ID NO: 70	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRIAM GRPPKSGTGEKPYKCPECGKSFSQRANLRAHQRTHTGEKPYKCPEC GKSFSRRDELNVHQRTHTGEKPYKCPECGKSFSQLAHLRAHQRTHT GEKPYKCPECGKSFSQRAHLERHQRTHTGKKTSGQAGQ
ZFR-5 Left	SEQ ID NO: 71	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRIAM GRPPKSGTGEKPYKCPECGKSFSRRDELNVHQRTHTGEKPYKCPEC GKSFSRSDHLTNHQRTHTGEKPYKCPECGKSFSQLAHLRAHQRTHT GEKPYKCPECGKSFSQRAHLERHQRTHTGKKTSGQAGQ
ZFR-5 Right	SEQ ID NO: 72	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRIGG RPPKSGTGEKPYKCPECGKSFSTSGSLVRHQRTHTGEKPYKCPECG KSFSRSDKLVRHQRTHTGEKPYKCPECGKSFSQSGDLRRHQRTHTG EKPYKCPECGKSFSTSGELVRHQRTHTGKKTSGQAGQ

ZFR-6 Left	SEQ ID NO: 73	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTSAGLAAAINKGRLMGR PLKSGTGEKPYKCPECGKSFSQLAHLRAHQRTHTGEKPYKCPECGK SFSQLAHLRAHQRTHTGEKPYKCPECGKSFSDPGHLVRHQRTHTGE KPYKCPECGKSFSDSGNLRVHQRTHTGKKTSGQAGQ
ZFR-6 Right	SEQ ID NO: 74	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERVSMAGLAAARNKGRLVM GRPPKSGTGEKPYKCPECGKSFSQRAHLERHQRTHTGEKPYKCPEC GKSFSTTGNLTVHQRTHTGEKPYKCPECGKSFSDSGNLRVHQRTHT GEKPYKCPECGKSFSQSSNLVRHQRTHTGKKTSGQAGQ
ZFR-7 Left	SEQ ID NO: 75	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRLFGR PPKSGTGEKPYKCPECGKSFSTHLDLIRHQRTHTGEKPYKCPECGKS FSTTGNLTVHQRTHTGEKPYKCPECGKSFSQSSSLVRHQRTHTGEKP YKCPECGKSFSRSDNLVRHQRTHTGKKTSGQAGQ
ZFR-7 Right	SEQ ID NO: 76	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLAAARNKGRLFGR PPKSGTGEKPYKCPECGKSFSRSDKLVRHQRTHTGEKPYKCPECGK SFSRRDELNVHQRTHTGEKPYKCPECGKSFSQSSSLVRHQRTHTGE KPYKCPECGKSFSRSDHLTNHQRTHTGKKTSGQAGQ
ZFR-8 Left	SEQ ID NO: 77	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERLMAGLAAARNKGRLCG RPPKSGTGEKPYKCPECGKSFSQRAHLERHQRTHTGEKPYKCPECG KSFSTSGNLVRHQRTHTGEKPYKCPECGKSFSRSDELVRHQRTHTG EKPYKCPECGKSFSHKNALQNHQRTHTGKKTSGQAGQ
ZFR-8 Right	SEQ ID NO: 78	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD

		SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGIAAARNKGRIGG RPPKSGTGEKPYKCPECGKSFSRRDELNVHQRTHTGEKPYKCPECG KSFSQSSNLVRHQRTHTGEKPYKCPECGKSFSQSSSLVRHQRTHTGE KPYKCPECGKSFSTTGNLTVHQRTHTGKKTSGQAGQ
ZFR-9 Left	SEQ ID NO: 79	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGIAAARNKGRIAA GRPPKSGTGEKPYKCPECGKSFSTTGNLTVHQRTHTGEKPYKCPEC GKSFSQSSNLVRHQRTHTGEKPYKCPECGKSFSQRAHLERHQRTHT GEKPYKCPECGKSFSQKSSLIAHQRTHTGKKTSGQAGQ
ZFR-9 Right	SEQ ID NO: 80	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDRPGLK RALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSLTD SIDTSSPMGRFFFYVMGALAEMERELIIERTMAGIAAAINKGRIAGR PAKSGTGEKPYKCPECGKSFSDPGALVRHQRTHTGEKPYKCPECGK SFSQSSSLVRHQRTHTGEKPYKCPECGKSFSQLAHLRAHQRTHTGE KPYKCPECGKSFSQRANLRAHQRTHTGKKTSGQAGQ
Arm region mutations are highlighted. Specificity-determining α-helical zinc-finger residues are underlined.		

[0121] While the Examples illustrate generation of ZFRs having a Gin catalytic domain, the methods may be applied to catalytic domains of a number of other recombinases. Such recombinases include: a) Tn3, also known as EcoTn3; Hin, also known as StyHin; MuGin; Sin; Beta; Pin; Min; Din; Cin; EcoTn21; SfaTn917; BmeTn5083; Bme53; Cpe; SauSK1; SauSK41; SauTn552; Ran; Aac; Lla; pMER05; Mlo92; Mlo90; Rrh; Pje; Req; PpsTn5501; Pae; Xan; ISXc5; Spy; RhizY4cG; SarpNL1; SsolSC1904a; SsolSC1904b; SsoISC1913; Aam606; MjaM0014; Pab; HpylS607; MtulS_Y349; MtuRv2792c; MtuRv2979c; MtuRv3828c; MtuRv0921; MceRv0921; TnpX; TndX; WwK; lactococcal phage TP901-1 serine recombinase; *S. pyogenes* phage ϕ370.1 serine recombinase; *S. pyogenes* phage ϕFC1 serine recombinase; *Listeria* phage A118 serine recombinase; *S. coelicolor* chromosome SC3C8.24 serine recombinase; *S. coelicolor* chromosome SC2E1.37 serine recombinase; *S. coelicolor* chromosome SCD78.04c serine recombinase; *S. coelicolor* chromosome SC8F4.15c serine recombinase; *S. coelicolor* chromosome SCD12A.23 serine recombinase;

S. coelicolor chromosome SCH10.38c serine recombinase; *S. coelicolor* chromosome SCC88.14 serine recombinase; *Streptomyces* phage ϕC31 serine recombinase; *Streptomyces* phage R4 serine recombinase; *Bacillus* phage ϕ105 serine recombinase; *Bacillus* phage SPBc2 serine recombinase; *Bacillus* prophage SKIN serine recombinase; *S. aureus ccrA* serine recombinase; *S. aureus ccrB* serine recombinase; *M. tuberculosis* phage Bxb1 serine recombinase; *M. tuberculosis* prophage ϕRV1 serine recombinase; YBCK_ECOLI; Y4bA; Bja; Spn; Cac 1956; and Cac 1954; and b) muteins of a).

[0122] Imperfect modularity with particular domains, lack of high-affinity binding to all DNA triplets, and difficulty in construction has hindered the widespread adoption of ZFPs in unspecialized laboratories. The discovery of a novel type of DNA-binding domain in transcription activator-like effector (TALE) proteins from *Xanthomonas* provides an alternative to ZFPs. Described herein are chimeric TALE recombinases (TALERs): engineered fusions between a hyperactivated catalytic domain from the DNA invertase Gin and an optimized TALE architecture. A library of incrementally truncated TALE variants was identified to identify TALER fusions that modify DNA with efficiency and specificity comparable to zinc-finger recombinases in bacterial cells. Also shown in the Examples, TALERs recombine DNA in mammalian cells. The TALER architecture described herein provides a platform for insertion of customized TALE domains, thus significantly expanding the targeting capacity of engineered recombinases and their potential applications in biotechnology and medicine.

[0123] Transcription activator-like effector (TALE) proteins can be designed to bind virtually any DNA sequence. General guidelines for design of TALE DNA-binding domains suggest that the 5'-most base of the DNA sequence bound by the TALE (the N_0 base) should be a thymine. We quantified the N_0 requirement by analysis of the activities of TALE transcription factors (TALE-TF), TALE recombinases (TALE-R) and TALE nucleases (TALENs) with each DNA base at this position. In the absence of a 5' T, we observed decreases in TALE activity up to >1000-fold in TALE-TF activity, up to 100-fold in TALE-R activity and up to 10-fold reduction in TALEN activity compared with target sequences containing a 5' T. To develop TALE architectures that recognize all possible N_0 bases, a structure-guided library design coupled with TALE-R activity selections was used to evolve novel TALE N-terminal domains to accommodate any N_0 base. A G-selective domain and broadly reactive domains were isolated and characterized. The engineered TALE domains selected in the TALE-R format demonstrated modularity and were active in TALE-TF and

TALEN architectures. Evolved N-terminal domains provide effective and unconstrained TALE-based targeting of any DNA sequence as TALE binding proteins and designer enzymes.

[0124] In one aspect, the invention provides a method of generating a transcription activator-like effector (TALE) protein binding domain which specifically binds a desired nucleotide. As shown in the Examples, the method includes a) randomizing the amino acid sequence of the TALE protein binding domain by mutating an amino acid residue within a variable di-residue (RVD), or within 1 to 2 amino acid residues N-terminal or C-terminal of the RVD; and b) selecting for the randomized TALE protein binding domain of (a), wherein the TALE protein binding domain specifically binds to the desired nucleotide.

[0125] Sequence-specific nucleases, recombinases, nucleases and transcription factors are provided herein. The sequence-specific polypeptides include customized TAL effector DNA binding domains. As such, in another aspect, the invention provides a chimeric polypeptide. The polypeptide includes: a) a recombinase, a transcription factor or nuclease; and b) a transcription activator-like effector (TALE) protein.

[0126] TALEs are proteins of plant pathogenic bacteria that are injected by the pathogen into the plant cell, where they travel to the nucleus and function as transcription factors to turn on specific plant genes. The primary amino acid sequence of a TALE dictates the nucleotide sequence to which it binds. Thus, target sites can be predicted for TALE, and TALE also can be engineered and generated for the purpose of binding to particular nucleotide sequences, as described herein.

[0127] Fused to the TALE-encoding nucleic acid sequences are sequences encoding a nuclease, transcription factor or recombinase, or a portion thereof. Many such proteins are known in art that may be used in the present invention.

[0128] In various embodiments, the chimeric polypeptide includes a catalytic domain of a recombinase. As discussed above, catalytic domains of a number of recombinases may be utilized. Such recombinases include: a) Tn3, also known as EcoTn3; Hin, also known as StyHin; Gin, also known as MuGin; Sin; Beta; Pin; Min; Din; Cin; EcoTn21; SfaTn917; BmeTn5083; Bme53; Cpe; SauSK1; SauSK41; SauTn552; Ran; Aac; Lla; pMER05; Mlo92; Mlo90; Rrh; Pje; Req; PpsTn5501; Pae; Xan; ISXc5; Spy; RhizY4cG; SarpNL1; SsolSC1904a; SsolSC1904b; SsoISC1913; Aam606; MjaM0014; Pab; HpylS607; MtulS_Y349; MtuRv2792c; MtuRv2979c; MtuRv3828c; MtuRv0921; MceRv0921; TnpX; TndX; WwK; lactococcal phage TP901-1 serine recombinase; *S. pyogenes* phage ϕ370.1

serine recombinase; *S. pyogenes* phage ϕFC1 serine recombinase; *Listeria* phage A118 serine recombinase; *S. coelicolor* chromosome SC3C8.24 serine recombinase; *S. coelicolor* chromosome SC2E1.37 serine recombinase; *S. coelicolor* chromosome SCD78.04c serine recombinase; *S. coelicolor* chromosome SC8F4.15c serine recombinase; *S. coelicolor* chromosome SCD12A.23 serine recombinase; *S. coelicolor* chromosome SCH10.38c serine recombinase; *S. coelicolor* chromosome SCC88.14 serine recombinase; *Streptomyces* phage ϕC31 serine recombinase; *Streptomyces* phage R4 serine recombinase; *Bacillus* phage ϕ105 serine recombinase; *Bacillus* phage SPBc2 serine recombinase; *Bacillus* prophage SKIN serine recombinase; *S. aureus ccrA* serine recombinase; *S. aureus ccrB* serine recombinase; *M. tuberculosis* phage Bxb1 serine recombinase; *M. tuberculosis* prophage ϕRV1 serine recombinase; YBCK_ECOLI; Y4bA; Bja; Spn; Cac 1956; and Cac 1954; and b) muteins of a). In preferred embodiments, a highly active Gin catalytic domain is utilized. Such a domain may be generated using the methods of the present invention as described herein.

[0129] As described herein, TALEs may include a number of imperfect repeats that determine the specificity with which they interact with DNA. Each repeat binds to a single base, depending on the particular di-amino acid sequence at residues 12 and 13 of the repeat. Thus, by engineering the repeats within a TALE, particular DNA sites can be targeted. Such engineered TALEs can be used, for example, as transcription factors targeted to particular DNA sequences.

[0130] As illustrated in the Examples, the chimeric proteins of the present invention are exemplified by the variants and portions thereof (e.g., RVDs and NTDs) as set forth in Table 3.

[0131] Table 3.

Variant	SEQ ID NO:	Sequence
TALEN (Goldy) NT-T T1 Protein Sequence	81	MRSPKKKRKVQVDLRTLGYSQQQQEKIKPKVRSTVAQHH EALVGHGFTHAHIVALSQHPAALGTVAVTYQHIITALPEAT HEDIVGVGKQWSGARALEALLTDAGELRGPPLQLDTGQLV KIAKRGGVTAMEAVHASRNALTGAPLNLTPDQVVAIASNG GGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQA LETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQR LLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLC QDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLT PDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAI

		ASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGG KQALETVQRLLPVLCQDHGLTPDQVVAIVSHDGGKQALET VQRLLPVLCQDHGLTPDQVVAIVSHDGGKQALETVQRLLP VLCQDHGLTPDQVVAIVSNGGGKQALETVQRLLPVLCQD HGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPD QVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIA SNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGK QALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALESI VAQLSRPDPALAALTNDHLVALACLGGRPAMDAVKKGLP HAPELIRRVNRRIGERTSHRVAGSQLVKSELEEKKSELRHK LKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGK HLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQA DEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSG HFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGT LTLEEVRRKFNNGEINF ..... = N-Terminal Domain (NTD) *-varied as shown below*
TALEN RVD Sequences		
G1	82	NG-NN-HD-NG-HD-NI-NG-NG-NI-HD-NI-HD-HD-NG-NN-HD-NI targeting (TCTTCATTACACCTGCA; SEQ ID NO: 280)
G2	83	HD-NI-NN-NG-HD-NI-NN-NG-NI-NG-HD-NI-NI-NG-NG targeting (CAGTCAGTATCAATT; SEQ ID NO: 281)
A1	84	HD-HD-NG-NN-HD-NI-NN-HD-NG-HD-NG-HD-NI-NG-NG-NG-NG targeting (CCTGCAGCTCTCATTTT; SEQ ID NO: 282)
A2	85	NI-NG-NG-HD-NG-NG-HD-HD-NI-NN-NI-NG-NG-NN-NI targeting (ATTCTTCCAGAATTGA; SEQ ID NO: 283)
C2	86	HD-NI-NN-NI-NI-NG-NG-NN-NI-NG-NI-HD-NG-NN-NI-HD-NG targeting (CAGAATTGATACTGACT; SEQ ID NO: 284)
T1	87	NG-HD-NI-NG-NG-NI-HD-NI-HD-HD-NG-NN-HD-NI-NN-HD targeting (TCATTACACCTGCAGC; SEQ ID NO: 285)

T2	88	HD-NG-NG-HD-HD-NI-NN-NI-NI-NG-NG-NN-NI-NG-NI-HD-NG-NN targeting (CTTCCAGAATTGATACTG; SEQ ID NO: 286)
N-Terminal Domains		
NTD = dHax3-TALEN DNA Sequence	89	ATGAGATCTCCTAAGAAAAAGAGGAAGATGGTGGACTTGAGGACACTCGGTTATTCGCAACAGCAACAGGAGAAAATCAAGCCTAAGGTCAGGAGCACCGTCGCGCAACACCACGAGGCGCTTGTGGGGCATGGCTTCACTCATGCGCATATTGTCGCGCTTTCACAGCACCCTGCGGCGCTTGGGACGGTGGCTGTCAAATACCAAGATATGATTGCGGCCCTGCCCGAAGCCACGCACGAGGCAATTGTAGGGGTCGGTAAACAGTGGTCGGGAGCGCGAGCACTTGAGGCGCTGCTGACTGTGGCGGGTGAGCTTAGGGGGCCTCCGCTCCAGCTCGACACCGGGCAGCTGCTGAAGATCGCGAAGAGAGGGGGAGTAACAGCGGTAGAGGCAGTGCATGCATCGCGCAATGCACTGACGGGTGCCCCC
NTD = dHax3-TALEN	90	MRSPKKKRKMVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTAVEAVHASRNALTGAP… repeat variable diresidues
NTD = NT-βN TALEN DNA Sequence	91	ATGAGATCTCCTAAGAAAAAGAGGAAGGTGCAGGTGGATCTACGCACGCTCGGCTACAGTCAGCAGCAGCAAGAGAAGATCAAACCGAAGGTGCGTTCGACAGTGGCGCAGCACCACGAGGCACTGGTGGGCCATGGGTTTACACACGCGCACATCGTTGCGCTCAGCCAACACCCGGCAGCGTTAGGGACCGTCGCTGTCACGTATCAGCACATAATCACGGCGTTGCCAGAGGCGACACACGAAGACATCGTTGGCGTCGGCAAATATCATGGGGCACGCGCTCTGGAGGCCTTGCTCACGGATGCGGGGGAGTTGAGAGGTCCGCCGTTACAGTTGGACACAGGCCAACTTGTGAAGATTGCAAAACGTGGCGGCGTGACCGCAATGGAGGCAGTGCATGCATCGCGCAATGCACTGACGGGTGCCCCC
NTD = NT-βN TALEN	92	MRSPKKKRKVQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVTYQHIITALPEATHEDIVGVGKYHGARALEALLTDAGELRGPPLQLDTGQLVKIAKRGGVTAMEAVHASRNALTGAP…. repeat variable diresidues

NTD NT-G TALEN DNA Sequence	93	ATGAGATCTCCTAAGAAAAAGAGGAAGGTGCAGGTGGATC TACGCACGCTCGGCTACAGTCAGCAGCAGCAAGAGAAGAT CAAACCGAAGGTGCGTTCGACAGTGGCGCAGCACCACGAG GCACTGGTGGGCCATGGGTTTACACACGCGCACATCGTTGC GCTCAGCCAACACCCGGCAGCGTTAGGGACCGTCGCTGTCA CGTATCAGCACATAATCACGGCGTTGCCAGAGGCGACACAC GAAGACATCGTTGGCGTCGGCAAATCGCGGTCGGGGGCAC GCGCTCTGGAGGCCTTGCTCACGGATGCGGGGGAGTTGAGA GGTCCGCCGTTACAGTTGGACACAGGCCAACTTGTGAAGAT TGCAAAACGTGGCGGCGTGACCGCAATGGAGGCAGTGCAT GCATCGCGCAATGCACTGACGGGTGCCCCC
NTD NT-G TALEN Protein Sequence	94	MRSPKKKRKVQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEAL VGHGFTHAHIVALSQHPAALGTVAVTYQHIITALPEATHEDIV GVGKSRSGARALEALLTDAGELRGPPLQLDTGQLVKIAKRGG VTAMEAVHASRNALTGAP… repeat variable diresidues
NTD = NT-αN TALEN DNA Sequence	95	ATGAGATCTCCTAAGAAAAAGAGGAAGGTGCAGGTGGATC TACGCACGCTCGGCTACAGTCAGCAGCAGCAAGAGAAGAT CAAACCGAAGGTGCGTTCGACAGTGGCGCAGCACCACGGG GCACTGGTGGGCCATGGGTTTACACACGCGCACATCGTTGC GCTCAGCCAACACCCGGCAGCGTTAGGGACCGTCGCTGTCA CGTATCAGCACATAATCACGGCGTTGCCAGAGGCGACACAC GAAGACATCGTTGGCGTCGGCAAACGGGGGGCTGGTGCAC GCGCTCTGGAGGCCTTGCTCACGGATGCGGGGGAGTTGAGA GGTCCGCCGTTACAGTTGGACACAGGCCAACTTGTGAAGAT TGCAAAACGTGGCGGCGTGACCGCAATGGAGGCAGTGCAT GCATCGCGCAATGCACTGACGGGTGCCCCC
NTD = NT-αN TALEN Protein	96	MRSPKKKRKVQVDLRTLGYSQQQQEKIKPKVRSTVAQHHGA LVGHGFTHAHIVALSQHPAALGTVAVTYQHIITALPEATHEDI VGVGKRGAGARALEALLTDAGELRGPPLQLDTGQLVKIAKRG GVTAMEAVHASRNALTGAP…repeat variable diresidues
NTD = NT-T T-1 TALEN DNA	97	ATGAGATCTCCTAAGAAAAAGAGGAAGGTGCAGGTGGATC TACGCACGCTCGGCTACAGTCAGCAGCAGCAAGAGAAGAT CAAACCGAAGGTGCGTTCGACAGTGGCGCAGCACCACGAG

		GCACTGGTGGGCCATGGGTTTACACACGCGCACATCGTTGC GCTCAGCCAACACCCGGCAGCGTTAGGGACCGTCGCTGTCA CGTATCAGCACATAATCACGGCGTTGCCAGAGGCGACACAC GAAGACATCGTTGGCGTCGGCAAACAGTGGTCCGGCGCAC GCGCCCTGGAGGCCTTGCTCACGGATGCGGGGGAGTTGAGA GGTCCGCCGTTACAGTTGGACACAGGCCAACTTGTGAAGAT TGCAAAACGTGGCGGCGTGACCGCAATGGAGGCAGTGCAT GCATCGCGCAATGCACTGACGGGTGCCCCC
MBP-TALE Protein Sequence	98	MKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPD KLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLAEITPDKA FQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPK TWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAF KYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDY SIAEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKG QPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNK DKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMS AFWYAVRTAVINAASGRQTVDEALKDAQTNSSSNNNNNNNN NNLGIEGRISEFGSPARPPRAKPAPRRRSAQPSDASPAAQVDLR TLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQ HPAALGTVAVTYQHIITALPEATHEDIVGVGK[**XXX**]GARALE ALLTDAGELLRGPPLQLDTGQLVKIAKRGGVTAMEAVHASRN ALTGAPLNLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGL TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAI ASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQA LETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLP VLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGL TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAI ASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIVSHDGGKQ ALETVQRLLPVLCQDHGLTPDQVVAIVSHDGGKQALETVQRL LPVLCQDHGLTPDQVVAIVSNGGGKQALETVQRLLPVLCQDH GLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVV AIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGK QALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALESIVAQ LSRPDPALAALTNDHLVALACLG XXX: NT-T = QWS NT-G = SRS; NT-αN = RGA; NT-βN = Y-H

TALE-R Protein Sequence	99	MLIGYVRVSTNDQNTDLQRNALVCAGCEQIFEDKLSGTRTDR PGLKRALKRLQKGDTLVVWKLDRLGRSMKHLISLVGELRERG INFRSLTDSIDTSSPMGRFFFYVMGALAEMERELIIERTMAGLA AARNKGRIGGRPPKSGSPRPPRAKPAPRRRAAQPSDASPAAQV DLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVA LSQHPAALGTVAVTYQHIITALPEATHEDIVGVGK[**XXX**]GARA LEALLTDAGELRGPPLQLDTGQLVKIAKRGGVTAMEAVHASR NALTGAPLNLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVA IASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQ ALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLL PVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVA IASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIVSHDGGK QALETVQRLLPVLCQDHGLTPDQVVAIVSHDGGKQALETVQR LLPVLCQDHGLTPDQVVAIVSNGGGKQALETVQRLLPVLCQD HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGG KQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALESIVA QLSRPDPALAALTNDHLVALACLG XXX: NT-T = QWS NT-G = SRS; NT-αN = RGA; NT-βN = Y-H
Avr15 TALE-TF Protein Sequence	100	MAQAASGSPRPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYS QQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAAL GTVAVTYQHIITALPEATHEDIVGVGK[**XXX**]GARALEALLTDA GELRGPPLQLDTGQLVKIAKRGGVTAMEAVHASRNALTGAPL NLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVV AIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGK QALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRL LPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDH GLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVV AIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGG KQALETVQRLLPVLCQDHGLTPDQVVAIVSHDGGKQALETVQ RLLPVLCQDHGLTPDQVVAIVSHDGGKQALETVQRLLPVLCQ DHGLTPDQVVAIVSNGGGKQALETVQRLLPVLCQDHGLTPDQ VVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHD GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET

		VQRLLPVLCQDHGLTPDQVVAIASNIGGKQALESIVAQLSRPD PALAALTNDHLVALACLGGRPAMDAVKKGLPHAPELIRRVNR RIGERTSHRVADYAQVVRVLEFFQCHSHPAYAFDEAMTQFGM SGQAGQASPKKKRKVGRADALDDFDLDMLGSDALDDFDLDM LGSDALDDFDLDMLGSDALDDFDLDMLINYPYDVPDYAS XXX: NT-T = QWS NT-G = SRS; NT-αN = RGA; NT-βN = Y-H
Avr15 RVD Sequence (for TALE-R, TALE-TF, MBP-TALE)	101	NI-NG-NI-NI-NI-HD-HD-HD-HD-HD-NI-HD-HD-NI-NI targeting (ATAAACCCCCTCCAA; SEQ ID NO: 287)

[0132] In various embodiments, chimeric protein includes a TALE protein having a C-terminal or N-terminal truncation. For example, the TALE protein may include all or a portion of SEQ ID NO: 2. In embodiments, the TALE protein is truncated between amino acid residues 27 and 268, 92 and 134, 120 and 129, 74 and 147, or 87 and 120, such at amino acid residue 28, 74, 87, 92, 95, 120, 124, 128, 129, 147 and 150.

[0133] In another embodiment, a isolated polypeptide comprising a transcription activator-like effector (TALE) protein is provided in which the TALE protein has an N-terminal domain (NTD) comprising an amino acid sequence as set forth in SEQ ID NO: 3 (VGKQWSGARAL) having one or more mutations or deletions selected from: Q is Y, Q is S, Q is R, W is R, W is G, W is deleted, S is R, S is H, S is A, S is N, and S is T.

[0134] In some embodiments, the NTD comprises an amino acid sequence selected from: VGKYRGARAL (SEQ ID NO: 4), VGKSRSGARAL (SEQ ID NO: 5), VGKYHGARAL (SEQ ID NO: 6), and VGKRGAGARAL (SEQ ID NO: 7).

[0135] In another embodiment, an isolated polypeptide comprising a transcription activator-like effector (TALE) protein is provided in which the TALE protein has an N-terminal domain (NTD) comprising an amino acid sequence as set forth in SEQ ID NO: 8 (IVDIAR$_1$QR$_2$SGDLA) having one or more mutations or deletions selected from: R$_1$ is K, Q is Y, Q is S, Q is R, R$_2$ is W, R$_2$ is G, R$_2$ is deleted, S is R, S is H, S is A, S is N, and S is T.

[0136] In some embodiments, the NTD comprises an amino acid sequence selected from: IVDIARQWSGDLA (SEQ ID NO: 9), IVDIARYRGDLA (SEQ ID NO: 10),

IVDIARSRSGDLA (SEQ ID NO: 11), IVDIARYHGDLA (SEQ ID NO: 12), and IVDIARRGAGDLA (SEQ ID NO: 13).

[0137] In another embodiment, the TALE protein includes a modified N_0 domain having an amino acid sequence set forth as follows: LTPDQLVKIAKRGGTAMEAVHASRNALTGAPLN (SEQ ID NO: 102). In various embodiments, the TALE protein includes a mutated variant in which KRGG (SEQ ID NO: 103) of SEQ ID NO: 102 is selected from LDYE (SEQ ID NO: 104), INLV (SEQ ID NO: 105), YSKK (SEQ ID NO: 106), NMAH (SEQ ID NO: 107), SPTN (SEQ ID NO: 108), SNTR (SEQ ID NO: 109), LTTT (SEQ ID NO: 110), VADL (SEQ ID NO: 111), MVLS (SEQ ID NO: 112), YNGR (SEQ ID NO: 113), RIPR (SEQ ID NO: 114), YSKI (SEQ ID NO: 115), LTQY (SEQ ID NO: 116), YLSK (SEQ ID NO: 117), LRPN (SEQ ID NO: 118), LFTN (SEQ ID NO: 119), LLTN (SEQ ID NO: 120), EEDK (SEQ ID NO: 121), VTAM (SEQ ID NO: 122), CPSR (SEQ ID NO: 123), LTRV (SEQ ID NO: 124), KGDL (SEQ ID NO: 125), QKAL (SEQ ID NO: 126), LYLL (SEQ ID NO: 127), WISV (SEQ ID NO: 128), GDQV (SEQ ID NO: 129) and CPSR (SEQ ID NO: 130).

[0138] In another embodiment, the TALE protein includes a modified N_{-1} domain having an amino acid sequence set forth as follows: MRSPKKKRKVQVDLRTLGYSQQQQEKIKPKVRSTVAQHH EALVGHGFTHAHIVALSQHPAALGTVAVTYQHIITALPEATHEDIVGVGXXXXXARA LEALLTDAGELRGPPLQLDTGQLVKIAKRGGVTAMEAVHASRNALTGAP (SEQ ID NO: 131). In various embodiments, XXXXX of SEQ ID NO: 131 is KRPAG (SEQ ID NO: 132) or KRPSG (SEQ ID NO: 133). Additionally, the protein may include, a E40G mutation (with reference to SEQ ID NO: 131) that exhibits enhanced activity.

[0139] In another embodiment, the TALE protein includes a repeat domain having an amino acid sequence set forth as follows: LTPDVVAISNNGGKQALETVQRLLPVLCQDGH (SEQ ID NO: 134). In various embodiments, the TALE protein includes a mutated variant in which SNNG (SEQ ID NO: 135) of SEQ ID NO: 134 is selected from RGGG (SEQ ID NO: 136), RGGR (SEQ ID NO: 137), RGVR (SEQ ID NO: 138), KGGG (SEQ ID NO: 139), SGGG (SEQ ID NO: 140), GGRG (SEQ ID NO: 141), LGGS (SEQ ID NO: 142), MDNI (SEQ ID NO: 143), RVMA (SEQ ID NO: 144), LASV (SEQ ID NO: 145), VGTG (SEQ ID NO: 146) and QGGG (SEQ ID NO: 147).

[0140] The following examples are provided to further illustrate the advantages and features of the present invention, but are not intended to limit the scope of the invention. While they are typical of those that might be used, other procedures, methodologies, or techniques known to those skilled in the art may alternatively be used.

EXAMPLE 1

CHIMERIC TALE RECOMBINASES

[0141] Experimental Summary.

[0142] This study provides the first example of a TALE recombinase (TALER). Using a library of incrementally truncated TALE domains, an optimized TALER architecture was identified that can be used to recombine DNA in bacterial and mammalian cells. Any customized TALE repeat array can be inserted into the TALER architecture described herein, thus dramatically expanding the targeting capacity of engineered recombinases for applications in biotechnology and medicine.

[0143] The following Material and Methods were utilized in this Example.

[0144] Reagents.

[0145] All enzymes were purchased from New England BioLabs unless otherwise indicated. Primer sequences are provided in Table 4.

[0146] Table 4. Primers.

Primers used in this study		
Primers for pBLA substrate construction		
AvrXa7 lac target F	SEQ ID NO: 148	TTAATTAAGAGTCTAGAAATATAAACCCCCTCC AACCAGGTGCTAACTGTAAACCATGGTTTTGGA TTAGCACCTGGTTGGAGGGGGTTTATAAGATCT AGGAGGAATTTAAAATGAG
AvrXa7 lac target R	SEQ ID NO: 149	ACTGACCTAGAGAAGCTTATATAAACCCCCTCC AACCAGGTGCTAATCCAAAACCATGGTTTAC AGTTAGCACCTGGTTGGAGGGGGTTTATACTG CAGTTATTTGTACAGTTCATC
AvrXa7 N F	SEQ ID NO: 150	TTAATTAAGAGTCTAGATTAGCACCTGGTTGG AGGGGGTTTATAAGGTTTTGGTACCAAATGTC TATAAACCCCCTCCAACCAGGTGCTAAAGATC TAGGAGGAATTTAAAATGAG
AvrXa7 N R	SEQ ID NO: 152	ACTGACCTAGAGAAGCTTTTAGCACCTGGTTG GAGGGGGTTTATAGACATTTGGTACCAAAACC

		TTATAAACCCCCTCCAACCAGGTGCTAACTGC AGTTATTTGTACAGTTCATC
AvrXa7 N RC F	SEQ ID NO: 153	TTAATTAAGAGTCTAGATTAGCACCTGGTTGG AGGGGGTTTATATCCAAAACCATGGTTTACAG TATAAACCCCCTCCAACCAGGTGCTAAAGATC TAGGAGGAATTTAAAATGAG
AvrXa7 N RC R	SEQ ID NO: 154	ACTGACCTAGAGAAGCTTTTAGCACCTGGTTG GAGGGGGTTTATATCCAAAACCATGGTTTACA GTATAAACCCCCTCCAACCAGGTGCTAACTGC AGTTATTTGTACAGTTCATC
AvrXa7 N RC +3 F	SEQ ID NO: 155	TTAATTAAGAGTCTAGATTAGCACCTGGTTGG AGGGGGTTTATAGCTTCCAAAACCATGGTTTA CAGGGTTATAAACCCCCTCCAACCAGGTGCTA AAGATCTAGGAGGAATTTAAAATGAG
AvrXa7 N RC +3 R	SEQ ID NO: 277	ACTGACCTAGAGAAGCTTTTAGCACCTGGTTG GAGGGGGTTTATAACCCTGTAAACCATGGTTT TGGAAGCTATAAACCCCCTCCAACCAGGTGCT AACTGCAGTTATTTGTACAGTTCATC
AvrXa7 N RC +6 F	SEQ ID NO: 156	TTAATTAAGAGTCTAGATTAGCACCTGGTTGG AGGGGGTTTATAGCTTCATCCAAAACCATGGT TTACAGGGTTCCTATAAACCCCCTCCAACCAG GTGCTAAAGATCTAGGAGGAATTTAAAATGAG
AvrXa7 N RC +6 R	SEQ ID NO: 157	ACTGACCTAGAGAAGCTTTTAGCACCTGGTTG GAGGGGGTTTATAGCAACCCTGTAAACCATGG TTTTGGATGAAGCTATAAACCCCCTCCAACCA GGTGCTAACTGCAGTTATTTGTACAGTTCATC
AvrXa7 N RC +12 F	SEQ ID NO: 158	TTAATTAAGAGTCTAGATTAGCACCTGGTTGG AGGGGGTTTATAGCTTCAGCTTCATCCAAAAC CATGGTTTACAGGGTTCCGGTTCCTATAAACC CCCTCCAACCAGGTGCTAAAGATCTAGGAGGA ATTTAAAATGAG
AvrXa7 N RC +12 R	SEQ ID NO: 278	ACTGACCTAGAGAAGCTTTTAGCACCTGGTTG GAGGGGGTTTATAGCAACCGCAACCCTGTAAA CCATGGTTTTGGATGAAGCTGAAGCTATAAAC CCCCTCCAACCAGGTGCTAACTGCAGTTATTT GTACAGTTCATC
AvrXa7 N RC -3 F	SEQ ID NO: 160	TTAATTAAGAGTCTAGATTAGCACCTGGTTGG

		AGGGGGTTTATAAAAACCATGGTTTATATAAA CCCCCTCCAACCAGGTGCTAAAGATCTAGGAG GAATTTAAAATGAG
AvrXa7 N RC -3 R	SEQ ID NO: 161	ACTGACCTAGAGAAGCTTTTAGCACCTGGTTG GAGGGGGTTTATATAAACCATGGTTTTTATAA ACCCCCTCCAACCAGGTGCTAACTGCAGTTAT TTGTACAGTTCATC
AvrXa7 N RC GG F	SEQ ID NO: 162	TTAATTAAGAGTCTAGATTAGCACCTGGTTGG AGGGGGTTTATATCCAAAACCGGGGTTTACA GTATAAACCCCCTCCAACCAGGTGCTAAAGA TCTAGGAGGAATTTAAAATGAG
AvrXa7 N RC GG R	SEQ ID NO: 163	ACTGACCTAGAGAAGCTTTTAGCACCTGGTT GGAGGGGGTTTATACTGTAAACCCCGGTTTT GGATATAAACCCCCTCCAACCAGGTGCTAAC TGCAGTTATTTGTACAGTTCATC
AvrXa7 N 20t F	SEQ ID NO: 164	TTAATTAAGAGTCTAGATTAGCACCTGGTTG GAGGGGGTTTATACGAAATATTATAAATTA TCATATAAACCCCCTCCAACCAGGTGCTAA AGATCTAGGAGGAATTTAAAATGAG
AvrXa7 N RC 20t R	SEQ ID NO: 165	ACTGACCTAGAGAAGCTTTTAGCACCTGGTT GGAGGGGGTTTATATGATAATTTATAATATT TCGTATAAACCCCCTCCAACCAGGTGCTAAC TGCAGTTATTTGTACAGTTCATC
AvrXa7 32 GG F	SEQ ID NO: 166	TTAATTAAGAGTCTAGATTAGCACCTGGTTG GAGGGGGTTTATAGCTTCATCCAAAACCGG GGTTTACAGGGTTCCTATAAACCCCCTCCAA CCAGGTGCTAAAGATCTAGGAGGAATTTAA AATGAG
AvrXa7 32 GG R	SEQ ID NO: 167	ACTGACCTAGAGAAGCTTTTAGCACCTGGTT GGAGGGGGTTTATAGCAACCCTGTAAACCGG GGTTTTGGATGAAGCTATAAACCCCCTCCAA CCAGGTGCTAACTGCAGTTATTTGTACAGTTC ATC
AvrXa7 32t F	SEQ ID NO: 168	TTAATTAAGAGTCTAGATTAGCACCTGGTTG GAGGGGGTTTATAGCTTCACGAAATATTATA AATTATCAGGTTCCTATAAACCCCCTCCAAC CAGGTGCTAAAGATCTAGGAGGAATTTAAA

		ATGAG
AvrXa7 32t R	SEQ ID NO: 169	ACTGACCTAGAGAAGCTTTTAGCACCTGGTT GGAGGGGGTTTATAGCAACCTGATAATTTAT AATATTTCGTGAAGCTATAAACCCCCTCCAA CCAGGTGCTAACTGCAGTTATTTGTACAGTT CATC
Primers for pGL3Pro target site construction.		
5' pGL3 SV40 Avr.32G BglII	SEQ ID NO: 170	TTAATTAAGAGAGATCTTTAGCACCTGGTTG GAGGGGGTTTATAGCTTCATCCAAAACCATG GTTTACAGGGTTCCTATAAACCCCCTCCAAC CAGGTGCTAAGCGATCTGCATCTCAATTAGT CAGC
3' pGL3 SV40 Avr.20G HindIII	SEQ ID NO: 171	ACT GAC CTA GAG AAG CTT TTA GCA CCT GGT TGG AGG GGG TTT ATAGCAACC CTG TAA ACC ATG GTT TTG GATGAAGCT ATA AAC CCC CTC CAA CCA GGT GCT AAT TTG CAA AAG CCT AGG CCT CCA AA
5' pGL3 SV40 PH4.20G6Avr BglII	SEQ ID NO: 172	TTAATTAAGAGAGATCTGCGGGAGGCGTGTC CAAAACCATGGTTTACAGGGTTCCTATAAAC CCCCTCCAACCAGGTGCTAAGCGATCTGCAT CTCAATTAGTCAGC
3' pGL3 SV40 PH4.20G6Avr HindIII	SEQ ID NO: 173	ACT GAC CTA GAG AAG CTT TTA GCA CCT GGT TGG AGG GGG TTT ATAGCAACCCTGTA AACCATGGTTTTGGACACGCCTCCCGCTTTG CAAAAGCCTAGGCCTCCAAA
5' pGL3 SV40 Avr.44G BglII	SEQ ID NO: 174	TTAATTAAGAGAGATCTTTAGCACCTGGTTG GAGGGGGTTTATAGCTTCAGCTTCATCCAAA ACCATGGTTTACAGGGTTCCGGTTCCTATAA ACCCCCTCCAACCAGGTGCTAAGCGATCTGC ATCTCAATTAGTCAGC
3' pGL3 SV40 Avr.44G HindIII	SEQ ID NO: 175	ACT GAC CTA GAG AAG CTT TTA GCA CCT GGT TGG AGG GGG TTTATAGCAACCGCAA CCCTG TAA ACC ATG GTT TTG GATGAAGC TGAAGCT ATA AAC CCC CTC CAA CCA GGT GCT AAT TTG CAA AAG CCT AGG CCT CCA AA
Primers for BamHI fusions		

Gin_N-term F	SEQ ID NO: 176	AGTCAGTCGAGAGCTCATGGATCCCGGCTCTA TGCTGATTGGCTATGTAAGG
Gin_N-term R	SEQ ID NO: 177	ATGCTGATATCTAGACTATCCCGATTTAGGTGG GCGACC
Gin_C-term F	SEQ ID NO: 178	AGTCAGTCGAGAGCTCATGCTGATTGGCTATGT AAGG
Gin_C-term R	SEQ ID NO: 179	TCTAGACTACGGATCCACCGATTTACGCGGGC
Primers for designed truncations		
TalR+28 Xba	SEQ ID NO: 180	ATCGCGTATCTAGACTAGCCGAGGCAGGCCAA GGCGACG
TalR+95 Xba AvrX	SEQ ID NO: 181	ATCGCGTATCTAGACTAGCTCATCTCGAACTGC GTCATG
avr n 1	SEQ ID NO: 182	GTCGCCCGCGTAAATCGGGATCCACTGCAGAT CGGGGGGGGGC
avr n 2	SEQ ID NO: 183	GTCGCCCGCGTAAATCGGGATCCCCCTCGCCTG CGTTCTCGGC
avr n 3	SEQ ID NO: 184	GTCGCCCGCGTAAATCGGGATCCGATTCGATGC CTGCCGTCGG
avr n 4	SEQ ID NO: 185	GTCGCCCGCGTAAATCGGGATCCACCGTGCGT GTCGCTGTCACTG
avr n 5	SEQ ID NO: 186	GTCGCCCGCGTAAATCGGGATCCGTGGATCTAC GCACGCTCGGC
avr n 6	SEQ ID NO: 187	GTCGCCCGCGTAAATCGGGATCCACACACGCG CACATCGTTGC
avr n 7	SEQ ID NO: 188	GTCGCCCGCGTAAATCGGGATCCCACGAAGAC ATCGTTGGCGTCG
avr n 8	SEQ ID NO: 189	GTCGCCCGCGTAAATCGGGATCCAGCGCtCTGG AGGCCTTGCTC
avr n 9	SEQ ID NO: 190	GTCGCCCGCGTAAATCGGGATCCTTGGACACA GGCCAACTTCTC
avr n 10	SEQ ID NO: 191	GTCGCCCGCGTAAATCGGGATCCAGCGGCGTG ACCGCAgTGGA
GinNTALPCRfusR	SEQ ID NO: 192	GGATCCCGATTTACGCGGGC
Primers used for pcDNA cloning		
Nhe-SD-Gin F	SEQ ID NO: 193	ATCGTAGCAGCTAGCGCCACCATGCTGATTGGC TATGTAAG
GinGS R	SEQ ID NO: 194	GGATCCAGACCCCGATTTACGCGGGC

[0147] Plasmid construction.

[0148] In order to introduce a *Bam*H1 restriction site either 5' or 3' to the Gin coding sequence, the Gin catalytic domain was PCR amplified with primers 5' Gin_N-term and 3' Gin_N-term or 5' Gin_C-term and 3' Gin_C-term, respectively. PCR products were ligated into the *Sac*I and *Xba*I restriction sites of pBluescriptII (Fermentas) to generate pB-Bam-Gin and pB-Gin-Bam. To generate the C-terminal and N-terminal TALER fusions, the AvrXa7 gene (kindly provided by Dr. B. Yang, Iowa State University) was released from pWAvrXa7 with *Bam*H1 and ligated into *Bam*H1 sites of pB-Bam-Gin and pB-Gin-Bam (41) to establish pB-Avr-Bam-Gin and pB-Gin-Bam-Avr, respectively. Correct construction of each TALER was verified by sequence analysis (Figures 6-16).

[0149] To generate N-terminal truncations of AvrXa7, AvrXa7 was PCR amplified using the Expand High Fidelity PCR System (Roche) with 5' Avr-n-(1-10) and 3' Avr +28 or 3' Avr +95 primers with the following program: 1 cycle of 3 min at 94 °C, 16 cycles of 1 min at 94 °C, 1 min at 52 °C, 6 min at 68 °C; and a final cycle of 1 hr at 68 °C. The Gin catalytic domain was PCR amplified under standard PCR conditions with 5' Gin_C-term and 3' GinNTalPCRFus and fused to truncated AvrXa7 variants by overlap PCR using the PCR conditions described above. Purified Gin-Avr PCR products were mixed in an equimolar ratio and digested with *Sac*I and *Xba*I.

[0150] To generate designer TALEs, we used a TALEN kit (Addgene) with the following modification: pTAL1 was modified to include truncations at Δ120, Δ128, or +28. To achieve this, AvrXa7Δ120 and AvrXa7Δ128 fragments were PCR amplified with 5' Avr n4 or Avr n128 and 3' TalR Xba+28 and ligated into the *Bam*H1 restriction site of pTAL1 to generate pTALΔ120 and pTALΔ128. The plasmids pTALΔ120 and pTALΔ128 retained the *Esp*3I restriction sites for Golden Gate cloning. TALE arrays cloned into pTALΔ120 and pTALΔ128 were digested with *Bam*H1 and *Xba*I for ligation into pB-Gin-Bam.

[0151] To generate mammalian TALER expression vectors, the Gin catalytic domain was PCR amplified from pB-Gin-Avr with 5' Nhe-SD-Gin F and 3' GinGS R and ligated into the *Nhe*I and *Bam*HI restriction sites of pcDNA 3.1 (Invitrogen). Avr15 was digested from pTALΔ120 or pTALΔ128 with *Bam*H1 and *Xba*I and ligated into pcDNA-Gin-Bam to generate pcDNA-Gin-Avr expression vectors.

[0152] The pBLA substrate plasmids were constructed as previously described.

[0153] To generate pGL3 reporter plasmids, the SV40 promoter was PCR amplified from pGL3-Promoter (Promega) with the recombination site-containing primers 5' pGL3 SV40 BglII and 3' pGL3 SV40 HindIII and ligated into the *Bgl*II and *Hind*III restriction sites of pGL3-Promoter.

[0154] Bacterial recombination assays.

[0155] Bacterial recombination assays were performed as previously described.

[0156] Incremental truncation library.

[0157] The incremental truncation library was generated using a modified protocol previously described. Briefly, in order to protect the Gin coding sequence from exonuclease digestion, a stuffer fragment with a *Sma*I restriction site was inserted into *Bam*H1 to generate pB-Gin-SmaI-Bam-Avr. This plasmid was linearized with *Nhe*I and incubated with Exonuclease III for 2.5 min at 37 °C followed by heat inactivation at 75 °C for 25 min. pB-Gin-Bam-Avr was then incubated with Klenow Fragment (3' to 5' Exo) with 200 μM dNTPs and 5 μM [α]-S-dNTPs for 30 min at 37 °C followed by heat inactivation at 80 °C for 25 min. To generate the truncation library, pB-Gin-Bam-Avr was incubated with Exonuclease III for 2.5 min at 37 °C followed by heat inactivation and subsequent blunt-ending with Mung Bean Nuclease for 1 hr at 30 °C. After digestion with *Sma*I, the blunt 3' end of the recombinase coding sequence was ligated to the blunt-ended library of TALE fragments. After transformation and purification, the plasmids were digested with *Sac*I and *Xba*I to release Gin-ΔAvr.

[0158] Mammalian reporter assays.

[0159] HEK293T cells were seeded onto 96-well plates at a density of 4×10^4 cells per well and grown in a humidified 5% CO_2 atmosphere at 37 °C. At 24 hr after seeding, cells were transfected with 150 ng pcDNA TALER expression vector, 2.5 ng pGL3 reporter plasmid, and 1 ng pRL-CMV for expression of *Renilla* luciferase using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. At 48 hr after transfection, cells were lysed with Passive Lysis Buffer (Promega) and luciferase expression was determined using the Dual-Luciferase Reporter Assay System (Promega) according to the manufacturer's instructions. Luminescence was measured using a Veritas Microplate Luminometer (Turner Biosystems).

[0160] Results.

[0161] TALER architecture.

[0162] A quantitative system for the evaluation and directed evolution of recombinase activity has been described. In this system (Figure 1A), a GFPuv transgene flanked by recombination sites is inserted into the gene encoding TEM-1 β-lactamase. This alteration disrupts β-lactamase expression and renders *Escherichia coli* cells that harbor this plasmid (pBLA) susceptible to ampicillin. Expression of an active recombinase from the substrate-containing plasmid, however, leads to recombination between target sites and restoration of the β-lactamase reading frame. This modification establishes host-cell resistance to ampicillin and enables the isolation of active recombinase variants from the substrate-containing plasmid. By measuring the number of ampicillin-resistant transformants following plasmid purification and re-transformation, recombinase activity can be also directly assessed. Because the activity of a chimeric recombinase is dependent upon both the catalytic domain and the DBD, this split gene reassembly selection system can also be used to evaluate the effectiveness of individual DBDs. Thus, the system was adapted to determine an optimal TALER architecture.

[0163] Importantly, because the catalytic domain of the DNA invertase Gin and related serine recombinases have pre-defined catalytic specificities, TALER fusion proteins cannot be constructed using the design described for TALENs. Structural and functional studies with the γδ resolvase and designed enzymes have indicated that the C-terminal E-helix mediates serine recombinase DNA recognition. In ZFRs, this helix binds DNA from the C to the N-terminus, 5' to 3'. Thus, because TALEs bind DNA in the 5' to 3' direction, it was anticipated that recombination could only occur when the TALE binding site is positioned on the opposite strand of the 20-bp core (Figure 1B).

[0164] It was chosen to generate TALERs using AvrXa7, as this TALE protein has been previously used to generate TALE nucleases and transcription factors. Conveniently, *Bam*HI restriction sites flank many TALEs, including AvrXa7 and multiple groups have used this restriction site to generate synthetic TALE fusions. Notably, this *Bam*HI fragment leaves the N-terminus of the TALE intact but removes the native effector domain from the C-terminus. This strategy was adopted and generated a Gin-AvrXa7 fusion by *Bam*H1 restriction digestion.

[0165] Gin-AvrXa7 was cloned into a pBLA selection vector containing recombination sites composed of a central 20-bp core sequence, which is recognized by the Gin catalytic domain, and two flanking 26-bp AvrXa7 binding sites. As anticipated, the Gin-AvrXa7 fusion was unable to recombine DNA when AvrXa7 binding sites were positioned adjacent to

the 20-bp core (Figure 1C). However, when AvrXa7 binding sites were positioned on the opposite strand of the 20-bp core, recombination was evident (Figure 1C), indicating that recombination site orientation is a critical component for catalytic domain fusion to the TALE N-terminus. In order to further establish that N-terminal fusion is necessary for recombination, a C-terminal AvrXa7-Gin variant was constructed that contained a non-canonical fusion orientation predicted to constrain catalytic domain activity (Figure 1B and Table 5). As expected, it was determined that this C-terminal AvrXa7 fusion demonstrated negligible activity in bacterial cells (Figure 1C).

[0166] Table 5.

Variant	SEQ ID NO:	Sequence
Gin-Avr (#1) /Avr20G	195	TTAGCACCTGGTTGGAGGGGGTTTATA TCCAAAACCATGGTTTACAG TATAAACCCCCTCCAACCAGGTGCTAA
Gin-Avr (#2)	196	TTAGCACCTGGTTGGAGGGGGTTTATA AGGTTTTGGTACCAAATGTC TATAAACCCCCTCCAACCAGGTGCTAA
Avr-Gin (#3)	197	TATAAACCCCCTCCAACCAGGTGCTAA CTGTAAACCATGGTTTTGGA TTAGCACCTGGTTGGAGGGGGTTTATA
Avr14G	198	TTAGCACCTGGTTGGAGGGGGTTTATA AAAACCATGGTTTA TATAAACCCCCTCCAACCAGGTGCTAA
Avr26G	199	TTAGCACCTGGTTGGAGGGGGTTTATA GCTTCCAAAACCATGGTTTACAGGGT TATAAACCCCCTCCAACCAGGTGCTAA
Avr32G	200	TTAGCACCTGGTTGGAGGGGGTTTATA GCTTCATCCAAAACCATGGTTTACAGGGTTCC TATAAACCCCCTCCAACCAGGTGCTAA
Avr44G	201	TTAGCACCTGGTTGGAGGGGGTTTATA GCTTCAGCTTCATCCAAAACCATGGTTTACAGGGTTCCGGTTCC TATAAACCCCCTCCAACCAGGTGCTAA
Avr20GG	202	TTAGCACCTGGTTGGAGGGGGTTTATA TCCAAAACCGGGGTTTACAG TATAAACCCCCTCCAACCAGGTGCTAA
Avr20T	203	TTAGCACCTGGTTGGAGGGGGTTTATA

		CGAAATATTATAAATTATCA TATAAACCCCCTCCAACCAGGTGCTAA
Avr32GG	204	TTAGCACCTGGTTGGAGGGGGTTTATA GCTTCATCCAAAACCGGGGTTTACAGGGTTCC TATAAACCCCCTCCAACCAGGTGCTAA
Avr32T	205	TTAGCACCTGGTTGGAGGGGGTTTATA GCTTCACGAAATATTATAAATTATCAGGTTCC TATAAACCCCCTCCAACCAGGTGCTAA
Avr-G-ZF	206	GCGGGAGGCGTG TCCAAAACCATGGTTTACAGGGTTCC TATAAACCCCCTCCAACCAGGTGCTAA
PthXo1-20G	207	GTGGTGTACAGTAGGGGGAGATGCA TCCAAAACCATGGTTTACAG TGCATCTCCCCCTACTGTACACCAC
PthXo1-32G	208	GTGGTGTACAGTAGGGGGAGATGCA GCTGCTTCCAAAACCATGGTTTACAGGGTGGT TGCATCTCCCCCTACTGTACACCAC

[0167] Designed truncations.

[0168] Although the Gin-AvrXa7 fusion described above catalyzed recombination, the activity of this variant was considerably lower than that of engineered ZFRs. Further, specificity analysis revealed that the Gin-AvrXa7 fusion was unable to faithfully discriminate between recognition sites containing non-cognate DBD sites and non-native 20-bp core sequences, indicating that recombination might not be Gin-mediated (Figure 1D). Recent reports have shown that TALEN activity can be enhanced when the TALE portion of the fusion protein is truncated. Thus, in order to attempt to improve TALER activity, a series of N and C-terminal AvrXa7 truncations were generated (Figure 2A).

[0169] Ten N-terminal truncations were assembled at roughly equal intervals beginning at AvrXa7 Thr 27 (Δ27) and ending at AvrXa7 Gly 268 (Δ268) (Figure 6). AvrXa7 Δ150, which has been reported as an N-terminal truncation variant for TALENs, was also generated. Two C-terminal AvrXa7 truncations were generated at positions 28 (+28) and 95 (+95). Both +28 and +95 have been reported as stable fusion points in TALENs. Each TALE truncation variant was fused to the Gin catalytic domain and this 20-member TALER library was cloned into a pBLA selection vector containing Avr-20G recognition sites. Following one round of selection in bacterial cells (Materials and Methods), individual ampicillin-

resistant clones were sequences and it was found that all selected TALERs contained either one of two N-terminal truncations: Δ87 and Δ120. Each selected clone was also +28 on the C-terminus. With the exception of a single Δ120 clone with a spontaneous 12 amino acid deletion near the fusion point (Δ120*), the activity of these clones was quite low (Figure 2B). In this assay, Gin-based ZFRs routinely show 20-40% recombination, however, the highest activity observed amongst the selected TALER fusions was ~7% recombination (Gin-AvrXa7Δ120*). Because the TALE DBD is three times larger than a ZF domain (not including the required flanking peptide sequence), we reasoned that the 20-bp spacer used for these TALER constructs might not be the optimal length for recombination.

[0170] Core sequence length.

[0171] Next the effect core sequence length has on recombination was investigated by evaluating whether DNA targets containing 14 (Avr-14G), 26 (Avr-26G) and 32-bp (Avr-32G) core sites could be recombined by selected TALERs. In order to maintain the reading frame of the β-lactamase gene following recombinase-mediated reassembly, core half-sites were modified by ± 3-bps (Table 1). The 20-member TALER library described above was subjected to one round of selection against each target site variant. Although identification of TALER variants capable of recombining the shortest target was not possible, Avr-14G (data not shown), two Gin-ΔAvrXa7 variants were identified (based on the N-terminal TALE truncations Δ87 and Δ120 and the C-terminal truncation +28) that recombined Avr-26G and Avr-32G. In particular, clonal analysis revealed that the selected TALERs (Gin-AvrXa7Δ87 and Gin-AvrXa7Δ120) recombined DNA with longer cores (e.g., 26 and 32-bps) at least 100-fold more efficiently than shorter cores (e.g., 14 and 20-bps) (Figure 2B). Further, it was found that Gin-AvrXa7Δ120 recombined targets containing a cognate core sequence (Avr-26G and Avr-32G) >100-fold more efficiently than a non-cognate core (Avr-20T, Avr-20GG, Avr-32T and Avr-32GG) (Figure 2C). Interestingly, the Gin-AvrXa7Δ120 fusion was not as active on 44-bp cores (Avr-44G) (recombination was ~3-fold lower than Avr-32G) (Figure 2C), indicating that core lengths between 26 and 44-bp are likely optimal for recombination by Gin-AvrXa7Δ120 in *E. coli*.

[0172] Incremental truncation library.

[0173] Although Gin-AvrXa7Δ120 showed increased recombination in comparison to Gin-AvrXa7, it was suspected that Gin-AvrXa7Δ120 might not be an optimal TALE fusion architecture because: (i) ZFRs containing the Gin catalytic domain recombined DNA >2-fold more efficiently than Gin-AvrXa7Δ120 and (ii) Gin-AvrXa7Δ120 was not identified from a

comprehensive library of TALE truncation variants. Thus, in order to identify better fusion architectures, a screen was devised based on the generation of a library of incrementally truncated TALE DBDs.

[0174] To achieve this, a protocol was adapted as previously described to enable fusion of an unmodified N-terminal domain (Gin) to a library of truncated C-terminal fragments (AvrXa7) (Materials and Methods). N-terminal AvrXa7 truncations that spanned the region between the AvrXa7 N-terminus (Met 1) and the first AvrXa7 repeat (Leu 298) were generated by exonuclease digestion and fused to an unmodified copy of the Gin catalytic domain (theoretical number of protein variants: ~300). Because previous results indicated that +28 is the optimal C-terminal truncation, we incorporated this architecture into the truncation library. TALERs were cloned into a pBLA selection vector containing Avr-32G target sites and transformed into *E. coli* ($>1 \times 10^5$ transformants). Sequence analysis confirmed an equal distribution of truncations spanning the region of interest (data not shown).

[0175] Following three rounds of selection, individual ampicillin-resistant clones were sequences and a number of unique truncation variants were identified (Figure 3A). Consistent with the selections performed using the 20-member TALE truncation library, which suggested that the optimal N-terminal TALER fusion points were likely located in proximity to positions 87 and 120, all selected Gin-AvrXa7 variants were found to contain a truncation between positions 74 (Δ74) and 147 (Δ147). In particular, 26 of 73 (35.6%, $p < .001$) clones contained truncations between positions 124 (Δ124) and 129 (Δ129). From this population, truncations at position 128 (Δ128) were among the most represented.

[0176] In order to systematically determine whether selected AvrXa7 domains increased TALER activity, we evaluated the performance of isolated Gin-AvrXa7 variants against DNA substrates containing Avr-32G target sites in *E. coli*. We focused our analysis on clones containing N-terminal deletions between AvrXa7 position 92 (Δ92) and 134 (Δ134). Consistent with sequence analysis, it was found that TALERs containing N-terminal truncations between Δ120 and Δ129 recombined DNA more efficiently than variants based on comparatively longer or shorter truncations, although the Δ92 fusion was also quite active (Figure 3B). Three clones further characterized: Δ74 and Δ145 were chosen because they represented the boundaries of possible fusion points, and Δ128 was assayed because it was the most prevalent clone found in the selections. Five targets with spacer lengths from 14 to 44-bp were assayed along with three negative controls (Avr32T, Avr32GG, and PthXo1-

32G). It was determined that Gin-Avr32GΔ74 and Gin-Avr32GΔ145 had modest activity on spacers longer than 20-bp, whereas Gin-Avr32GΔ128 recombined DNA with efficiencies comparable to the ZFR GinC4 (Figure 3C). Furthermore, specificity analysis revealed that Gin-Avr32GΔ74, Gin-Avr32GΔ128, and Gin-Avr32GΔ145 could recombine substrates harboring cognate cores >100-fold more efficiently than non-cognate cores (Avr-32T, Avr-32GG and PthXo1-32G) (Figure 3C). Together, these results suggest that TALE proteins containing N-terminal deletions between Δ120 and Δ129 represent an optimal truncation for fusion to a recombinase.

[0177] Incorporation of synthetic TALE repeat arrays.

[0178] The studies described above used the native DBDs of the naturally occurring AvrXa7 TALE protein. In order to determine whether designed TALE repeat arrays can be incorporated into the selected Gin-ΔAvrXa7 frameworks, a series of synthetic TALE proteins (15 to 20 repeats in length) were generated designed to target the AvrXa7 binding site (Figure 7). TALE proteins were constructed using a publicly available TALEN plasmid set (Addgene). The cloning plasmid was modified to include the +28 C-terminal truncation and either the Δ120 or Δ128 N-terminal truncation. Designed TALEs were fused to the Gin catalytic domain (denoted as Gin-Avr15Δ120 and Gin-Avr15Δ128) and cloned into a pBLA selection vector containing Avr-32G or Avr-32T target sites.

[0179] Activity analysis in *E. coli* revealed that both Gin-Avr15Δ120 and Gin-Avr15Δ128 could be used to recombine DNA when fused to an active catalytic domain and that incorporation of synthetic repeats provided an increase in activity (Figure 4A). Importantly, each TALER displayed stringent selectivity, recombining target sites that contained cognate cores >1,000-fold more efficiently than non-cognate cores (Figure 4B). Surprisingly, TALERs based on the Δ120 truncation were also found to recombine DNA as effectively as TALEs based on the Δ128 architecture (Figure 4A), indicating that designed TALEs may be less sensitive to N-terminal truncation than those containing the native AvrXa7 DBD.

[0180] To further demonstrate that the TALER architecture described herein can be reprogrammed to target any DNA sequence, a synthetic enzyme was created designed to target the sequence recognized by the naturally occurring TALE protein PthXo1 (Gin-Pth15Δ120). It was found that Gin-Pth15Δ120 was highly active on its cognate substrate and that both Gin-Pth15Δ120 and Gin-Avr15Δ120 showed a >600 fold increase in recombination for targets with their cognate binding sites (Figure 4A). The activity of a series of designed TALERs containing DBDs between 15 and 20 repeats in length was also assessed and found

that each fusion catalyzed recombination with similarly high efficiency and specificity (Figure 4B), demonstrating that chimeric recombinases that incorporate synthetic TALE repeat arrays can be used for site-specific recombination.

[0181] TALER activity in mammalian cells.

[0182] It was also determined whether TALERs could modify DNA in mammalian cells. To achieve this, we used an episomal reporter assay that enables rapid assessment of recombinase activity in cell culture. In this assay, human embryonic kidney (HEK) 293T cells are co-transfected with a recombinase expression vector and a reporter plasmid (pGL3) that contains a luciferase gene under the control of a SV40 promoter flanked by recombination sites. Transient expression of the appropriate recombinase leads to excision of the SV40 promoter and reduced luciferase expression in cells. Recombinase activity is thus directly proportional to the fold-reduction in luciferase expression.

[0183] Co-transfection of Gin-Avr15Δ120 with a reporter plasmid harboring Avr-44G recognition sites (pGL3-Avr-44G) led to a ~20-fold reduction in luciferase expression as compared to transfection of pGL3-Avr-44G alone (Figure 5A). Despite the fact that Gin-Avr15Δ120 showed similar activity to the ZFR GinC4 in *E. coli*, we found that GinC4 reduced luciferase expression by >80-fold after co-transfection with its cognate target plasmid, pGL3-C4-20G (Figure 5A). This discrepancy may be due to the comparatively shorter intervening DNA sequence between recombinase target sites in pGL3 than pBLA or differential expression between TALERs and ZFRs in mammalian cells. The underlying cause for this disparity, however, remains unclear. Finally, although 32-bp was determined to be the optimal core sequence length for TALERs in *E. coli*, it was determined that co-transfection of Gin-Avr15Δ120 with pGL3-Avr-32G led to only a 6-fold reduction in luciferase expression (Figure 5A). The underlying cause behind this disparity also remains unclear.

[0184] Next whether a ZFR (GinC4) and a TALER (Gin-Avr15Δ120) could form a compatible heterodimer in mammalian cells was investigated. To evaluate this possibility, a hybrid recombination site was generated in which the AvrXa7 binding site and the C4 zinc-finger binding site (GCG GGA GGC GTG; SEQ ID NO: 279) flank the core sequence recognized by the Gin catalytic domain (pGL3-Avr-G-ZF) (see Table 2). Surprisingly, co-transfection of pGL3-Avr-G-ZF with GinC4 and Gin-Avr15Δ120 led to a >140-fold reduction in luciferase expression as compared to pGL3-Avr-G-ZF (Figure 5B), whereas transfection with either GinC4 or Gin-Avr15Δ120 with pGL3-Avr-G-ZF led to a negligible

decrease in reporter gene expression. These results demonstrate that generating ZF-TALE heterodimers represents a potentially effective approach for improving the targeting capacity of chimeric recombinases.

[0185] Discussion.

[0186] Unlike ZFPs, which contain a very minimal fusion architecture, TALE DBDs require native protein framework on either side of the DBD array to function. The so-called 0^{th} and 1^{st} repeats, which mediate binding of the thymidine residue at position 0 and are found in almost all known TALE recognition sites, represent such an N-terminal framework. A recent crystal structure provided a description of the binding of the position 0 thymine, yet there remains insufficient data to determine a minimal TALE architecture. Indeed, all studies to date have used an N-terminal truncation containing considerably more residues than those required to mediate binding at position 0. It remains uncertain what role this part of the protein has in enabling the proper DNA binding conformation or what might constitute a minimal TALE domain. Although initial attempts to generate functional TALE chimeras were based on fusion to full-length TALE proteins, more recent studies have focused on the identification of unique C-terminal truncations that improve effector domain function in the context of the Δ150 N-terminal architecture. A previous report indicated that deletion of N-terminal residues 2-153 (Δ150) of the AvrBs3 TALE removes the domain required for translocation of the TALE from its native bacteria to the target plant cell but does not compromise transcription factor activity.

[0187] Developing an active TALER, however, necessitated that unique N-terminal TALE variants be identified. A broad, systematic survey was initially conducted of N-terminal TALEs with the C-terminal truncations +28 and +95 and found that only two domains (Δ87 with +28 and Δ120 with +28) demonstrated sufficiently high activity for further analysis. A secondary analysis based on incremental truncation of the AvrXa7 N-terminus led to the identification of a broad cluster of truncation variants centered between AvrXa7 position 74 (Δ74) and position 145 (Δ145). Of the clones recovered in this experiment, 38% contained truncations between positions Δ119 and Δ128, and a survey of data obtained on TALERs with fusions in this region showed high activity. In particular, it was determined that TALERs based on N-terminal truncations from this region (Δ128 and Δ120) could be used to recombine DNA in bacteria and mammalian cells. The clustering of truncation variants between Δ119 and Δ128 may also be indicative of the intrinsic stability of this region.

[0188] ZFRs typically catalyze recombination between target sites 44 to 50-bp in length. Each target site contains a central 20-bp core sequence, which is recognized by the recombinase catalytic domain, and two adjacent ZFP binding sites. The fusion orientation of TALERs, however, necessitates that TALE binding sites are on the opposite strand relative to the central core sequence. This unique geometry led us to investigate the minimum core sequence requirements for recombination. Because of the length of TALE DBDs (TALE repeats are 3 to 4 times longer than ZFPs) and the extended N-terminal linker between the catalytic domain and the TALE domain, we reasoned that longer core sequences (32 or 44-bp) would be necessary for recombination. Indeed, with the exception of a TALE variant harboring a spontaneous deletion (Δ120*), most N-terminal truncation variants identified in this study demonstrated optimal performance against 32-bp cores. These results are consistent with those reported with TALENs, which unlike ZFNs require significantly longer spacer sequences (e.g. TALENs: 17 to 20-bp, ZFNs: 5 to 6-bp) to efficiently cleave DNA. In support of these observations, it was found that selection for unique N-terminal truncation variants against a short core sequence (14-bp) did not yield any clones.

[0189] Gin-AvrXa7Δ128 was identified as an optimal TALE fusion, but subsequent studies using synthetic TALE proteins generated using a publicly available TALE assembly kit indicated that Δ128 and Δ120-based TALERs showed similar activity in *E. coli*. These designed TALEs were based on a chimeric protein derived from the closely related and naturally occurring Tal1c and PthXo1 TALE proteins. Although TALEs share high homology, they are not identical. While polymorphisms in RVD repeats outside of residues 12 and 13 have been shown to have no affect on TALE fusion activities, to our knowledge no systematic evaluation of differences in TALE framework outside the DBDs has been reported. As demonstrated by the analysis of the incremental truncation library, minor amino acid alterations can significantly influence the activity of a particular fusion. Thus, some of the discrepancy in activity we observed between Gin-AvrXa7Δ120 and the synthetic Gin-Avr15Δ120 may be attributable to the sequence variations between AvrXa7 framework and the TALE framework architecture used previously.

[0190] The four RVDs (NI: A, HD: C, NG: T, and NN: G) favored for construction of synthetic TALEs are the most prevalent in nature; however, it remains to be determined whether these repeats represent the most specific RVD modules. For the 26-repeat AvrXa7 TALE, a synthetic version targeting the same sequence would have 16 changes in RVD composition (Figure 7). It was hypothesized that because they are more commonly found in

nature, the four RVDs selected for synthetic use might have a higher affinity for their cognate bases than other RVDs. If this were the case, it would be reasonable to assume that a TALE created with the synthetic RVD repeats could have higher DNA-binding affinity than a TALE using the native domains. Although the issue of RVD affinity was not directly addressed, it was determined that that TALERs containing synthetic repeat arrays were more active than constructs, which contained the native AvrXa7 DBD. TALERs with synthetic DBDs showed approximately two-fold higher activities than constructs containing the native repeats, despite containing significantly fewer DBDs. Additionally, the gain in activity observed with the synthetic arrays was not correlated with any increase in off-target recombination.

[0191] Several studies have shown that TALEs can tolerate some mismatches in their target sequence. These findings are unsurprising, as RVDs that are positively associated with particular bases have been shown to tolerate non-cognate bases in nature. The cooperative specificity afforded by TALERs could be used to circumvent potential limitations, however. Because the catalytic domain contributes specificity to recombination, it is envisioned that designer TALERs capable of selectively modifying highly homologous genomic sequences could be generated as well. Indeed, it has been recently demonstrated that recombinase catalytic specificity can be effectively reprogrammed to target unnatural core sites.

EXAMPLE 2

SELECTION OF NOVEL 0^{TH} RESIDUE SPECIFICITY

[0192] A new class of Tal-based DNA binding proteins was engineered. TAL (transcription activator-like) effectors constitute a novel class of DNA-binding proteins with predictable specificity. Tal effectors are employed by Gram-negative plant-pathogenic bacteria of the genus *Xanthomonas* which translocate a cocktail of different effector proteins via a type III secretion system (T3SS) into plant cells where they serve as virulence determinants. DNA-binding specificity of TALs is determined by a central domain of tandem repeats. Each repeat confers recognition of one base pair (bp) in the DNA. Rearrangement of repeat modules allows design of proteins with desired DNA-binding specificities with certain important limitations. For example, the most constraining feature of targeting a DNA sequence with a Tal domain is the requirement that the Tal DNA site start with the base T and sometimes C. Targeting a binding site starting with a G or A base has not been possible at the -1 position. Tal-recombinase activity selections were used to select for Tal DNA binding domains that lack this restriction by targeting mutations to the -1 and 0^{th} RVD regions. The practical consequences of this discovery are vast since now every DNA sequence can be

targeted with new Tal domains facilitating new unrestricted approaches to TAL transcription factors to turn transcription on/up or off/down, to target TAL nucleases to knock out gene function or to direct homologous recombination or to target our own TAL recombinases or other TAL enzymes.

[0193] For G specificity at the (-1) position, the amino acids QWSG (SEQ ID NO: 209) were first randomized using an NNK codon strategy within the (-1) domain of the GinAvr15Δ128-synthetic protein. Following 3 rounds of tal recombinase activity selection of the resulting library, novel tal binding domains with the selected sequences RSNG (SEQ ID NO: 210) and SRSG (SEQ ID NO: 211) in the targeted region were selected. These were then shown to bind G at the 0th position of the target sequence over the parental T recognized by the starting clone. The selection was repeated randomizing the KQW region shown below in red that overlaps with the QWSG (SEQ ID NO: 212) selected initially. Now clones with selected SSR, SRA, SRC, and KRC sequences were selected. All selected Tal binding domains were assayed in binding studies to defined oligos bearing the G substitution and shown to now preferentially bind the sequence G-ATAAACCCCCTCCAA (SEQ ID NO: 213). Note that the Tal recombinase activity selection was performed using this same sequence. The starting Tal binding protein the GinAvr15Δ128 binds T-ATAAACCCCCTCCAA (SEQ ID NO: 214). Subsequence testing of Tal nucleases bearing the selected mutations verify the G specify of these sequences allowing for this novel class of Tals to be developed for the first time. Selected sequences are portable to Tals derived from other species.

[0194] Table 6.

Selections	
SEQ ID NO: 215	ATHEDIVGVGKQWSGARALEALLTDAGELRGPPLQ (-1 domain)
SEQ ID NO: 216	ATHEDIVGVGK**QWSG**ARALEALLTDAGELRGPPLQ (randomized AA in bold)
SEQ ID NO: 217	KQWSG-starting clone sequence
SEQ ID NO: 218	K*RSNG-selected to bind G*
SEQ ID NO: 219	K*SRSG-selected to bind G*

SEQ ID NO: 220	ATHEDIVGVG**KQW**SGARALEALLTDAGELRGPPLQ
SEQ ID NO: 221	**KQW**SG- WT
	SSR-*selected to bind G*
	SRA-*selected to bind G*
	SRC-*selected to bind G*
	KRC-*selected to bind G*

[0195] Selections were also performed using this same library to target A. In this study, sequences PRG, PTR, and PKD were selected. All selected Tal binding domains were assayed in binding studies to defined oligos bearing the A substitution and shown to now preferentially bind the sequence A-ATAAACCCCCTCCAA (SEQ ID NO: 222). Note that the Tal recombinase activity selection was performed using this same sequence. The starting Tal binding protein the GinAvr15Δ128 binds T-ATAAACCCCCTCCAA (SEQ ID NO: 223). Subsequence testing of Tal nucleases bearing the selected mutations verify the A specify of these sequences allowing for this novel class of Tals to be developed for the first time. Subsequent refinements in binding activities can be achieved by random mutagenesis of the N-terminal domain or target mutagenesis of the KRGG (SEQ ID NO: 224) sequence within the 0th domain and reselection in the recombinase system.

EXAMPLE 3

SELECTIONS

[0196] For context dependent RVD selections and selections of RVDs with new specificities, libraries were created that randomize the HD sequence emboldened below. LTPDQVVAIAS**HD**GGKQALETVQRLLPVLCQDHG (prototype RVD sequence; SEQ ID NO: 225)

[0197] Typically the library allows all amino acids at these two positions, though libraries limited to N, D, H, K, and Q amino acids are often successful substitutes for the H residue. Alternatively larger libraries that randomized the S**HD**G (SEQ ID NO: 226) and AS**HD**GG (SEQ ID NO: 227) regions allow for the selection of unique RVD specificities with context dependent characteristics.

[0198] Tal recombinase activity selections then rapidly allow for the selection of new specificities within the targeted RVD domain. The resulting RVDs can be highly modular or context dependent in their sequence recognition and can be then used to create Tal nucleases and transcription factors.

[0199] Utility of this technology includes unrestricted approaches to TAL transcription factors to turn transcription on/up or off/down, to target TAL nucleases to knock out gene function or to direct homologous recombination or to target our own TAL recombinases or other TAL enzymes for use as tools and therapeutics.

[0200] Advantages and the practical consequences of this discovery are vast since now every DNA sequence can be targeted with our new Tal domains and their specificities can be readily optimized.

EXAMPLE 4

DIRECTED EVOLUTION OF TALE N-TERMINAL DOMAIN TO ACCOMMODATE 5' BASES OTHER THAN THYMINE

[0201] Transcription activator-like effector (TALE) proteins can be designed to bind virtually any DNA sequence of interest. The DNA binding sites for natural TALE transcription factors (TALE-TFs) that target plant avirulence genes have a 5' thymidine. Synthetic TALE-TFs also have this requirement. Recent structural data indicate that there is an interaction between the N-terminal domain (NTD) and a 5' T of the target sequence. A survey of the recent TALE nuclease (TALEN) literature yielded conflicting data regarding the importance of the first base of the target sequence, the N_0 residue. Additionally, there have been no studies regarding the impact of the N_0 base on the activities of TALE recombinases (TALE-Rs). Here, the impact of the N_0 base is quantified in the binding regions of TALE-Rs, TALE-TFs, TALE DNA-binding domains expressed as fusions with maltose binding protein (MBP-TALEs) and TALENs. Each of these TALE platforms have distinct N- and C-terminal architectures, but all demonstrated highest activity when the N_0 residue was a thymidine. To simplify the rules for constructing effective TALEs in these platforms, and allow precision genome engineering applications at any arbitrary DNA sequence, we devised a structure-guided activity selection using our recently developed TALE-R system. Novel NTD sequences were identified that provided highly active and selective TALE-R activity on TALE binding sites with 5' G, and additional domain sequences were selected that permitted general targeting of any 5' N_0 residue. These domains were imported into TALE-TF, MBP-TALE and TALEN architectures and consistently exhibited greater activity than did the wild-

type NTD on target sequences with non-T 5' residues. The novel NTDs are compatible with the golden gate TALEN assembly protocol and now make possible the efficient construction of TALE transcription factors, recombinases, nucleases and DNA-binding proteins that recognize any DNA sequence allowing for precise and unconstrained positioning of TALE-based proteins on DNA without regard to the 5' T rule that limits most natural TALE proteins.

[0202] The following Material and Methods were utilized in this Example.

[0203] Oligonucleotides.

[0204] Primers and other oligonucleotides (Table 4 below) were ordered from Integrated DNA Technologies (San Diego, CA).

[0205] Table 7. Primers.

Primer	SEQ ID NO:	Sequence
KXXG Lib Rev	228	TCTCAACTCCCCCGCCTCCGTGAGCAAGGCCTCCAGAGCGCGTGCC CCMNNMNNTTTGCCGACGCCAACGATGTCTTCGTG
KXXXX Lib Rev	229	TCT CAA CTC CCC CGC CTC CGT GAG CAA GGC CTC CAG AGC GCG TGC MNN MNN MNN MNN TTT GCC GAC GCC AAC GAT GTC TTC GTG
XXXSG Lib Rev	230	CCCGCCTCCGTGAGCAAGGCCTCCAGGGCGCGTGCGCCGGAMNNM NNMNNGCCGACGCCAACGATGTCTTCGTGTGTCGC
KRGG Lib Rev	231	GGC ACC CGT CAG TGC ATT GCG CCA TGC ATG CAC TGC CTC CAC TGC GGT CAC MNN MNN MNN MNN TGC AAT CTT GAG AAG TTG GCC TGT GTC
Goldy TALEN fwd	232	AGAGAGAGAAGAAAATGAGATCTCCTAAGAAAAAGAGGAAGGTGC AGGTGGATCTACGCACGCTCGGCTAC
NTD-dHax3 Fwd	233	AGGAAGAAGAGAAGCATGAGATCTCCTAAGAAAAAGAGGAAGGTG ATGGTGGACTTGAGGACACTCGGTTA
NTD-dHax3 Rev	234	AAGAGAAGAAGAAGAAGCATTGCGCCATGCATGCACTGCCTCTA
pTal127 Not1 fwd	235	CCC GCC ACC CAC CGT GC
N-Term Sph1	236	TGC TCT ATG CAT GCA CTG CCT CC
pTAL127-SFI Fwd	237	AGA GAA GAG AAG AGA AGG CGC CCG CGG CCC AGG CGG CCT CGG GAT CCC CTC GGC CTC CGC GCG CCA AG
pTAL127-SFI +95 Rev	238	AGA GAG AGA GAG AGA GTC TAG AGG CCG GCC TGG CCG CTC ATC CCG AAC TGC GTC ATG GCC TCA TC

pTAL127 Xba +28 Rev	239	GCC CCA GAT CCT GGT ACG CTC TAG AGG
Avr 5'A biotin hairpin	240	5'BiosgATC TTA GCA CCT GGT TGG AGG GGG TTT ATTGG GTT TTC CCAAT AAA CCC CCT CCA ACC AGG TGC TAA GAT
Avr 5'T biotin hairpin	241	5'Biosg/ATC TTA GCA CCT GGT TGG AGG GGG TTT ATAGG GTT TTC CCTAT AAA CCC CCT CCA ACC AGG TGC TAA GAT
Avr 5'G biotin hairpin	242	5'BiosgATC TTA GCA CCT GGT TGG AGG GGG TTT ATCGG GTT TTC CCGAT AAA CCC CCT CCA ACC AGG TGC TAA GAT
Avr 5'C biotin hairpin	243	5'BiosgATC TTA GCA CCT GGT TGG AGG GGG TTT ATGGG GTT TTC CCCAT AAA CCC CCT CCA ACC AGG TGC TAA GAT
CCR5-inner fwd	244	TTAAAAGCCAGGACGGTCAC
CCR5-inner rev	245	TGTAGGGAGCCCAGAAGAGA
CCR5-outer fwd	246	ACAGTTTGCATTCATGGAGGGC
CCR5-outer rev	247	CCGAGCGAGCAAGCTCAGTT
CCR5-indel fwd	248	CGCGGATCCCCGCCCAGTGGGACTTTG
CCR5-indel rev2	249	CCGGAATTCACCTGTTAGAGCTACTGC
pGL3 NTD stuffer fwd	250	AGA GAG AGA GAG AGG CGG CCG CCC TAC CAG GGA TTT CAG TCG ATG TAC ACG TTC
pGL3 NTD stuffer rev	251	AAG AAG AAG AAG GAA GAG AAG TAG GCC TGT CAT CGT CGG GAA GAC CTG CGA CAC CTG C
pgl3 5X Avr Xho1	252	ACTGCTATCCGAGTATAAACCCCCTCCAACCAGGTATAAACCCCCT CCAACCAGGTATAAACCCCCTCCAACCAGGTATAAACCCCCTCCAA CCAGGTATAAACCCCCTCCAACCAGGATCTGCGATCTAAGTAAGCT
AvrXa7 32G A F	253	TTAATTAAGAGTCTAGAttagcacctggttggaggggggtttatTgcttcaTCCAAAACC ATGGTTTACAGggttccAATAAACCCCCTCCAACCAGGTGCTAAAGAT CTAGGAGGAATTTAAAATGAG
AvrXa7 32G A R	254	ACTGACCTAGAGAAGCTTTTAGCACCTGGTTGGAGGGGGTTTATTgc aaccCTGTAAACCATGGTTTTGGAtgaagcAATAAACCCCCTCCAACCAG GTGCTAACTGCAGTTATTTGTACAGTTCATC
AvrXa7 32G G F	255	TTAATTAAGAGTCTAGAttagcacctggttggaggggggtttatCgcttcaTCCAAAACC ATGGTTTACAGggttccGATAAACCCCCTCCAACCAGGTGCTAAAGAT

		CTAGGAGGAATTTAAAATGAG
AvrXa7 32G G R	256	ACTGACCTAGAGAAGCTTTTAGCACCTGGTTGGAGGGGGTTTATCgc aaccCTGTAAACCATGGTTTTGGAtgaagcGATAAACCCCCTCCAACCAG GTGCTAACTGCAGTTATTTGTACAGTTCATC
AvrXa7 32G C F	257	TTAATTAAGAGTCTAGAttagcacctggttggaggggggtttatGgcttcaTCCAAAACC ATGGTTTACAGggttccCATAAACCCCCTCCAACCAGGTGCTAAAGAT CTAGGAGGAATTTAAAATGAG
AvrXa7 32G C R	258	ACTGACCTAGAGAAGCTTTTAGCACCTGGTTGGAGGGGGTTTATGgc aaccCTGTAAACCATGGTTTTGGAtgaagcCATAAACCCCCTCCAACCAG GTGCTAACTGCAGTTATTTGTACAGTTCATC
Luciferase. Vector = pgl3 basic. XhoI/SphI		
Forward target containing:	SEQ ID NO:	
5x Avr15 n-1c xhoF:	259	actgctatctcgagcTATAAACCCCCTCCAACCAGGcTATAAACCCCCTCCAACC AGGcTATAAACCCCCTCCAACCAGGcTATAAACCCCCTCCAACCAGGcTA TAAACCCCCTCCAACCAGGATCTGCGATCTAAGTAAGCT
5x Avr15 0=A n-1c	260	actgctatctcgagcAATAAACCCCCTCCAACCAGGcAATAAACCCCCTCCAACC AGGcAATAAACCCCCTCCAACCAGGcAATAAACCCCCTCCAACCAGGcA ATAAACCCCCTCCAACCAGGATCTGCGATCTAAGTAAGCT
5x Avr15 0=C n-1c	261	actgctatctcgagcCATAAACCCCCTCCAACCAGGcCATAAACCCCCTCCAACC AGGcCATAAACCCCCTCCAACCAGGcCATAAACCCCCTCCAACCAGGcC ATAAACCCCCTCCAACCAGGATCTGCGATCTAAGTAAGCT
5x Avr15 0=G n-1c	262	actgctatctcgagcGATAAACCCCCTCCAACCAGGcGATAAACCCCCTCCAACC AGGcGATAAACCCCCTCCAACCAGGcGATAAACCCCCTCCAACCAGGcG ATAAACCCCCTCCAACCAGGATCTGCGATCTAAGTAAGCT
Luciferase Reverse Primer:	263	TCAGAAACAGCTCTTCTTCAAATCT

[0206] Generation of TALE-R NTD evolution plasmids.

[0207] The TALE-R system previously reported was adapted for this study. Briefly, pBCS (containing chloramphenicol and carbenicillin resistance genes) was digested with HindIII/Spe1. The stuffer (Avr X, where X is the N0 base), containing twin recombinase

sites, was digested with HindIII/Xba1 and ligated into the vector to create a split *beta-lactamase* gene. pBCS AvrX was then digested with BamH1/Sac1, and Gin127-N-stuffer-Avr15 was digested with BamH1/Sac1 and ligated into the vector to create Gin127-N-stuffer-Avr15-X. The stuffer was digested with Not1/Stu1 for evolutions at the N_{-1} TALE hairpin and Not1/Sph1 for evolutions at the N_0 TALE hairpin.

[0208] Generation of TALE NTD evolution libraries.

[0209] Primer ptal127 Not1 fwd and reverse primers KXXG lib rev or KXXXX lib rev were used to generate N-terminal variants at the N_{-1} TALE hairpin and were subsequently digested with Not1/Stu1 then ligated into digested Gin127-AvrX. Forward primer ptal127 Not1 fwd and reverse primer KRGG Lib Rev were used to PCR amplify a library with mutations in the N_0 TALE hairpin. This was subsequently digested with Not1/Sph1 and ligated into Not1/Sph1-digested Gin127-AvrX.

[0210] TALE-R NTD evolution assay.

[0211] Round 1 ligations were ethanol precipitated and transformed into electrocompetent Top10 F' cells then recovered in SOC for 1 h. The cells were grown overnight in 100 ml Super Broth (SB) media containing 100 mg/ml chloramphenicol. DNA was isolated via standard procedures. The resulting plasmid DNA (Rd 1 input) was transformed into electrocompetent Top10F' cells; cells were grown overnight in 100 ml of SB containing 100 mg/ml carbenicillin and 100 mg/ml chloramphenicol. Plasmid DNA was isolated via standard procedures. Round 1 output was digested with Not1/Xba1 and ligated into the Gin127-AvrX vector with complementary sticky ends. This protocol was repeated three to four times when a consensus sequence was observed and clones were characterized.

[0212] Measurement of N-terminal TALEN activity.

[0213] Four TALEN pairs containing each possible base were generated using the golden gate protocol. Fusion A and B plasmids were directly ligated via second golden gate reaction into the Goldy TALEN (N Δ152/C +63) framework. The NTD was modified by digesting the pCAG vector with BglII/Nsi1 and ligating with PCR amplified NTD digested with BglII/Nsi1. TALEN pairs (50–75 ng each TALEN/well) were transfected into HeLa cells in wells of 96-well plates at a density of 1.5×10^4 cells/well. After transfection, cells were placed in a 37° C incubator for 24 h, then were moved to 30° C for 2 days and then moved to 37° C for 24 h. Genomic DNA was isolated according to a published protocol, and DNA mutation rates were quantified with the Cel1 Surveyor assay and by sequencing. For CelI assays, genomic DNA was amplified by nested PCR, first with primers CCR5 outer

fwd/CCR5 outer rev and then with CCR5 inner fwd/CCR5 inner rev. For sequencing of indels, the second PCR was performed with CCR5 indel fwd/CCR5 indel rev. Fragments were then digested with BamH1/EcoR1 and ligated into pUC19 with complementary digestion.

[0214] TALE-TFs and luciferase assay.

[0215] Variant NTDs from the recombinase selection were PCR amplified with primers ptal127 SFI fwd and N-Term Sph1. The PCR product was amplified and digested with Not1/Stu1 and ligated into pTAL127-SFI Avr15, which contains twin SFI-1 digestion sites facilitating transfer of the N-terminal-modified TALE from pTAL127-SFI Avr15 into pcDNA 3.0 VP64. Corresponding TALE binding sites were cloned into the pGL3 Basic vector (Promega) upstream of the luciferase gene. For each assay, 100 ng of pcDNA was co-transfected with 5 ng of pGL3 vector and 1 ng of pRL *Renilla* luciferase control vector into HEK293t cells in a well of a 96-well plate using Lipofectimine 2000 (Life Technology) according to manufacturer's specifications. After 48 h, cells were washed, lysed and luciferase activity assessed with the Dual-Luciferase reporter system (Promega) on a Veritas Microplate luminometer (Turner Biosystems). Transfections were done in triplicate and results averaged.

[0216] MBP-TALE assay.

[0217] Affinity assays of MBP-TALE binding to biotinylated oligonucleotides were performed using a protocol previously described. Briefly, AvrXa7 TALE domains were expressed from pMAL MBP-AvrXa7 plasmid in XL1-Blue cells and purified on amylose resin. Biotinylated oligonucleotides containing the target AvrXa7 target site with modified residues were used to determine TALE-binding activity in sandwich enzyme-linked immunosorbent assay format. Antibodies targeting the MBP substituent were used for assay development.

[0218] Results.

[0219] Preliminary analysis of the 5' T rule.

[0220] A recent crystal structure of a TALE protein bound to PthXo7 DNA sequence revealed a unique interaction between W232 in the N-1 hairpin with a thymidine at the 5' end of the contacted region of the DNA substrate (the N_0 base). This study provided a structural basis for the previously established 5' T rule reported when the TALE code was first deciphered (Figure 18A and B). There are conflicting data regarding the importance of the first base of the target sequence of TALENs. The requirement for a 5' T in the target DNA

was initially assesses in the context of TALE-Rs using four split *beta lactamase* TALE recombinase selection vectors containing four AvrXa7 binding sites with all possible 5' residues flanking a Gin32G core (Figure 18C). Recognition of the N_0 residue by TALE-TFs was then evaluated using four luciferase reporter vectors containing a pentamer AvrXa7 promoter region with recognition sites containing each possible 5' residue (Figure 18D). With bases other than a 5' T, we observed decreases in activity up to >100-fold in TALE-Rs and 1000-fold in TALE-TFs relative to the sequence with a 5' T (Figure 18C and D). These reductions were observed despite variations in the C-terminal architectures of these chimeras that reportedly remove the 5' T bias, especially in the presence of a greatly shortened C-terminal domain (CTD). Enzyme-linked immunosorbent assay also indicated decreased affinity of MBP-TALE DNA-binding proteins toward target oligonucleotides with non-T 5' residues (Figure 18E). Finally, examination of the activity of designed TALENs with wild-type NTDs on targets with non-T 5' nucleotides showed up to 10-fold decrease in activity versus those with a 5' T (Figure 18F). The results indicate that a 5' T is an important design parameter for maximally effective TALE domains in the context of recombinases, transcription factors, nucleases and simple DNA-binding proteins.

[0221] Evolution of the TALE NTD to accommodate non-T 5' residues.

[0222] To create a more flexible system for DNA recognition, it was hypothesized that the recently developed TALE-R selection system could be utilized to evolve the NTD of the TALE to remove the 5' T constraint (Figure 23). Libraries were generated with residues K230 through G234 randomized, and TALE-Rs with activity against each possible 5' base were isolated after several rounds of selection (Figure 19A-C). The most active selected clones exhibited strong conservation of K230 and G234; the former may contact the DNA phosphate backbone, and the latter may influence hairpin loop formation (Figure 24). In the case of library K230-W232, K230S was frequently observed but had much lower activity than K230R or K230 variants in nearly all variants assayed individually. One clone (NT-G) of several observed with a W232 to R232 mutation demonstrated a significant shift of selectivity from 5' T to 5' G; the sequence resembles that of the NTD of a recently described *Ralstonia* TALE protein in this region. The *Ralstonia* NTD, in the context of plant transcription factor reporter gene regulation, has been reported to prefer a 5' G in its substrate (see Figure 25 for a protein alignment). Residue R232 may contact the G base specifically, as indicated by the stringency of NT-G for 5' G. The preference of NT-G for a 5' G was comparable with the specificity of the wild-type domain for 5' T. NTD variants specific for 5'

A or 5' C were not able to be derived, but a permissive NTD, NT-αN, was obtained that resembles the K265-G268 N_0 hairpin that accepts substrates with any 5' residue and maintains high activity. It was hypothesized that this variant makes enhanced non-specific contacts with the DNA phosphate backbone compared with the wild-type NTD, enhancing the overall binding of the TALE-DNA complex without contacting a specific 5' residue. It was hypothesized that a shortened hairpin structure would allow selection of variants with specificity for 5' A or 5' C residues. A library with randomization at Q231-W232 and with residue 233 deleted was designed to shorten the putative DNA-binding loop. Recombinase selection revealed a highly conserved Q231Y mutation that had high activity in a number of clones (Figure 19D). In particular, NT-βN demonstrated improved activity on substrates with 5' A, C and G but diminished activity on 5' T substrates compared with TALEs with the wild-type NTD (Figure 19E).

[0223] Applications of evolved TALE NTDs.

[0224] To assess the portability of the evolved NTDs in designer TALE fusion protein applications, optimized NTDs were incorporated into TALE-TFs, MBP-TALEs and TALENs. TALE-TFs with NT-G, NT-αN and NT-βN domains demonstrated 400–1500-fold increases in transcriptional activation of a luciferase target gene bearing operator sites without a 5' T residue when compared with the TALE-TF with the NT-T domain. The NT-G-based TF retained the 5' G selectivity as observed in the TALE-R selection system. The activities of NT-αN- and NT-βN-based TFs against all 5' nucleotides tracked the relative activity observed in the recombinase format (Figure 20). MBP-TALEs also exhibited greater relative binding affinity for target oligonucleotides with sites that did not have a 5' T than did the wild-type MBP-TALE (Figure 26), providing further evidence that the selected domains enhanced recognition of or tolerance for non-thymine 5' bases.

[0225] Four of the optimized NTDs were then imported into the Goldy TALEN framework. For these experiments, four substrates were constructed within the context of the Δ32 locus of the CCR5 gene (Figure 21A). Each substrate contained a different 5' residue. Experiments included TALENs with wild-type (NT-T) and dHax3 NTDs (dHax3 is commonly used NTD variant isolated from *Xanthomonas campestris*) with specificity for 5' T, to benchmark gene editing activity. The substrate TALEN pairs were designed to retain as much RVD homology (50–90%) as possible to determine the activity enhancing contributions of the variant NTDs (Figure 21A).

[0226] Activities of the TALENs were analyzed both by sequencing and by using the Cel1 assay. The selected domains exhibited increases in gene editing activity between 2- and 9-fold for the non-T 5' residues when compared with activities of the TALEN containing the wild-type domain (Figure 21 and Figure 27). Activity was highest on TALEN pair T1/T2 with wild-type or dHax3 NTD. The TALEN pair substrate G1/G2 was processed most effectively by TALENs with NT- αN, NT-βN and NT-G, with 2.0–3.5-fold enhancement versus NT-T. NT-αN had activity 9- and 2-fold higher than the wild-type NT-T on TALEN pairs A1/A2 and C1/C2, respectively. Although the impact of a mismatch at the 5' residue is more modest in TALENs than in TALE-TF and TALE-R frameworks, the optimized NTDs greatly improved TALEN activity when used in gene editing experiments.

[0227] Discussion.

[0228] Most, but not all, previous studies have suggested that a thymidine is required as the 5'-most residue in design of optimal TALE DNA-binding domains. The analyses described here indicate that a thymidine is optimal, and in some cases critical, for building functional TALE fusion proteins. This requirement therefore imposes limitations on the sequences that can be effectively targeted with TALE transcription factor, nuclease and recombinase chimeras. Although this requirement theoretically imposes minor limitations on the use of TALENs for inducing gene knockout, given their broad spacer region tolerance, NTD's that can accommodate any 5' residue would further simplify the rules for effective TALE construction and greatly enhance applications requiring precise TALE placement for genome engineering and interrogation (e.g. precise cleavage of DNA at a defined base pair using TALENs, seamless gene insertion and exchange via TALE-Recombinases, displacement of natural DNA-binding proteins from specific endogenous DNA sequences to interrogate their functional role, the development of orthogonal transcription factors for pathway engineering, the synergistic activation of natural and synthetic genes wherein transcription factor placement is key and many other applications). Other uses in DNA-based nanotechnology include decorating DNA nanostructures/origami with specific DNA-binding proteins. Here, targeting to specific sites is constrained based on DNA folding/structure and thus being able to bind any site is critical. Elaboration of these structures and devices with DNA-binding proteins could be a fascinating approach to expanding function. Indeed, it is not difficult to imagine many applications for DNA binding proteins and their fusions when all targeting constraints are removed. Encouraged by these potential applications, we aimed to develop NTDs that enable targeting of sites initiated at any base.

[0229] The recently developed TALE-R system was used to evolve the NTD of the TALE to remove the 5'-T constraint. In three rounds of selection, an NTD was obtained with specificity for a 5' G. Numerous selections were performed in attempts to obtain variants that recognized either 5' A or 5' C. The G230-K234 hairpin was inverted, the K230-G234/ins232 hairpin extended, modification of the K265-G268 N_0 hairpin attempted, and random mutagenesis libraries evaluated. None of these strategies yielded NTDs with affinity for target sequences with 5' A or 5' C, although we did identify an NTD, NT-βN, with a deletion that recognized substrates with both 5' A and 5' C residues with acceptable affinity. The strong selection preference exhibited by the NTDs NT-T and NT-G and the importance of W232 in NT-T and R232 in NT-G are likely due to specific interactions of these amino acids with the 5' terminal residue of the DNA recognition sequence. It was recently reported that the *Ralstonia solanacearum* TALE stringently requires a 5' G, and a sequence alignment with NT-G shows what appears to be a comparable N-1 hairpin containing an arginine at the position analogous to 232 in NT-G (Figure 25). Owing to the high structural homology between the NTDs Brg11 and NT-T, it may be possible to modify the preference of the *Ralstonia* TALE NTD to thymine by a simple arginine to tryptophan mutation or to eliminate specificity by grafting NT-αN or NT-βN domains into this related protein. It is also interesting to note that arginine–guanine interactions are common in evolved zinc finger domains.

[0230] The variant NTDs selected were successfully imported into TALE-TFs, MBP-TALEs and TALENs and generally conferred the activity and specificity expected based on data from the recombinase evolution system. TALE-TFs with optimized NTDs enhanced TALE activation between 400- and 1500-fold relative to the activity of NT-T against AvrXa7 promoter sites with non-T 5' residues. When incorporated into TALENs, our NTD with non-T selectivity enhanced activity 2–9-fold relative to that of the NT-T domain on substrates with 5' A, C or G. The increases in TALEN gene editing generally correlated with increases in activity observed in TALE-R and TALE-TF constructs. The specificity and high activity of NT-G was maintained, as evidenced by the lower activity in assays with TALEN pairs A1/A2, C1/C2, and T1/T2, and the generally high activity of NT-αN and NT-βN was also imparted into the TALEN Δ152/+63 architecture.

[0231] It was recently reported that alternatively truncated TALEs with synthetic TALE RVD domains do not require a 5' T in the DNA substrate. The reported Δ143, +47 truncation was constructed as a Goldy TALE-TF and substantially lower activity on the AvrXa7

substrate was observed than for the Δ127, +95 truncation, which has been most commonly used by others and which is the truncation set used in our study (Figure 29). Thus, the difference in reported outcomes could be due to the truncated architectures used.

[0232] In summary, the importance of a 5' thymidine in the DNA substrate for binding and activity of designed TALEs was determined in the context of TALE-R, TALE-TF, MBP-TALEs and TALEN chimeras. Targeted mutagenesis and TALE-R selection were applied to engineer TALE NTDs that recognize bases other than thymine as the 5' most base of the substrate DNA. The engineered TALE domains developed here demonstrated modularity and were highly active in TALE-TF and TALEN architectures. These novel NTDs expand by ~15-fold the number of sites that can be targeted by current TALE-Rs, which have strict geometric requirements on their binding sites and which are highly sensitive to the identity of the N_0 base. Furthermore, they now allow for the precise placement of TALE DBDs and TALE-TFs at any DNA sequence to facilitate gene regulation, displacement of endogenous DNA-binding proteins and synthetic biology applications where precise binding might be key. Although TALENs based on the native NTD show varying degrees of tolerance of N_0 base substitutions, the data indicate that the novel NTDs reported here also facilitate higher efficiency gene editing with any N_0 base as compared with natural NTD-based TALENs.

EXAMPLE 5

CHIMERIC ZINC FINGER RECOMBINASES

[0233] The following materials and method were utilized.

[0234] The split gene reassembly vector (pBLA) was derived from pBluescriptII SK (-) (Stratagene) and modified to contain a chloramphenicol resistance gene and an interrupted TEM-1 p lactamase gene under the control of a *lac* promoter. ZFR target sites were introduced as previously described. Briefly, GFPuv (Clontech) was PCR amplified with the primers GFP- ZFR-XbaI-Fwd and GFP-ZFR-HindIII-Rev and cloned into the SpeI and HindIII restriction sites of pBLA to generate pBLA-ZFR substrates. All primer sequences are provided in Table 8.

[0235] Table 8. Primer Sequences.

Primer	SEQ ID NO:	Sequence
GFP-ZFR-20G-XbaI-Fwd	264	TTAATTAAGAGTCTAGAGGAGGCGTGTCCAAAACCATGGTTTACAGCACGCCTCCAGATCTAGGAGGAATTTAAAATGAG

GFP-ZFR-20G-HindIII-Rev	265	ACTGACCTAGAGAAGCTTGGAGGCGTGCTGTAAACCATGGTTTTGGACACGCCTCCCTGCAGTTATTTGTACAGTTCATC
SV40-ZFR-1-BglII-Fwd	266	TTAATTAAGAGAGATCTGCTGATGCAGATACAGAAACCAAGGTTTTCTTACTTGCTGCTGCGCGATCTGCATCTCAATTAGTCAGC
CMV-PstI-ZFR-1 Rev	267	CACCACCACGGATCCGCAGCAGCAAGTAAGAAAACCTTGGTTTCTGTATCTGCATCAGCAATTTCGATAAGCCAGTAAGCAG
5' Gin-HBS-Koz	268	CACCACCACGCGCGCAAGCTTAGATCTGGCCCAGGCGGCCACCATGCTGATTGGCTATGTAAGGG
3' Gin-AgeI-Rev	269	CACCACCACACCGGTTCCCGATTTAGGTGGGCGAC
ZFR-Target-1-Fwd	270	GTTCCTGCCAGGATCCACTAG
ZFR-Target-1-Rev	271	GCATGTGTCCAGATGCATAGG
ZFR-Target-2-Fwd	272	CACCTTCTCCCAGGATAAGG
ZFR-Target-2-Rev	273	GTTGGCCTGTATTCCTCTGG
ZFR-Target-3-Fwd	274	AATGAAGTTCCCTTGGCACTTC
ZFR-Target-3-Rev	275	CTGAAGGGTTTTAAGTGCAGAAG
CMV-Mid Prim-1	276	TGACGTCAATGACGGTAAATGG
ZFR targets are underlined.		

[0236] To generate luciferase reporter plasmids, the SV40 promoter was PCR amplified from pGL3- Prm (Promega) with the primers SV40-ZFR-BglIII-Fwd and SV40-ZFR-HindIII-Rev. PCR products were digested with BglII and HindIII and ligated into the same restriction sites of pGL3-Prm to generate pGL3-ZFR-1, 2, 3...18. The pBPS-ZFR donor

plasmid was constructed as previously described with the following exception: the ZFR-1, 2 and 3 recombination sites were encoded by primers 3' CMV-PstI-ZFR-1, 2 or 3-Rev. Correct construction of each plasmid was verified by sequence analysis.

[0237] Recombination assays.

[0238] ZFRs were assembled by PCR as previously described. PCR products were digested with SacI and XbaI and ligated into the same restrictions sites of pBLA. Ligations were transformed by electroporation into *Escherichia coli* T0P10F' (Invitrogen). After 1 hr recovery in SOC medium, cells were incubated with 5 mL SB medium with 30 ^g mL^{-1} chloramphenicol and cultured at 37°C. At 16 hr, cells were harvested; plasmid DNA was isolated by Mini-prep (Invitrogen) and 200 ng pBLA was used to transform *E. coli* T0P10F'. After 1 hr recovery in SOC, cells were plated on solid LB media with 30 ^g mL^{-1} chloramphenicol or 30 ^g mL^{-1} chloramphenicol and 100 ^g mL^{-1} carbenicillin, an ampicillin analogue. Recombination was determined as the number of colonies on LB media containing chloramphenicol and carbenicillin divided by the number of colonies on LB media containing chloramphenicol. Colony number was determined by automated counting using the GelDoc XR Imaging System (Bio-Rad).

[0239] Selections.

[0240] The ZFR library was constructed by overlap extension PCR as previously described. Mutations were introduced at positions 120, 123, 127, 136 and 137 with the degenerate codon NNK (N: A, T, C or G and K: G or T), which encodes all 20 amino acids. PCR products were digested with SacI and XbaI and ligated into the same restriction sites of pBLA. Ligations were ethanol precipitated and used to transform *E. coli* T0P10F'. Library size was routinely determined to be ~5 x 10^7. After 1 hr recovery in SOC medium, cells were incubated in 100 mL SB medium with 30 ^g mL^{-1} chloramphenicol at 37°C. At 16 hr, 30 mL of cells were harvested; plasmid DNA was isolated by Mini-prep and 3 ^g plasmid DNA was used to transform *E. coli* T0P10F'. After 1 hr recovery in SOC, cells were incubated with 100 mL SB medium with 30 ^g mL^{-1} chloramphenicol and 100 ^g mL^{-1} carbenicillin at 37°C. At 16 hr, cells were harvested and plasmid DNA was isolated by Maxi-prep (Invitrogen). Enriched ZFRs were isolated by SacI and XbaI digestion and ligated into fresh pBLA for further selection. After 4 rounds of selection, sequence analysis was performed on individual carbenicillin-resistant clones. Recombination assays were performed as described above.

[0241] ZFR construction.

[0242] Recombinase catalytic domains were PCR amplified from their respective pBLA selection vector with the primers 5' Gin-HBS-Koz and 3' Gin-AgeI-Rev. PCR products were digested with HindIII and AgeI and ligated into the same restriction sites of pBH to generate the SuperZiF- compitable subcloning plasmids: pBH-Gin-a, P, y, 5, s or Z. Zinc-fingers were assembled by SuperZiF and ligated into the AgeI and SpeI restriction sites of pBH-Gin-a, P, y, 5, s or Z to generate pBH-ZFR-L/R-1, 2, 3.18 (L: left ZFR; R: right ZFR). ZFR genes were released from pBH by SfiI digestion and ligated into pcDNA 3.1 (Invitrogen) to generate pcDNA-ZFR-L/R-1, 2, 3.18. Correct construction of each ZFR was verified by sequence analysis (Table 9).

[0243] Table 9. Catalytic domain substitutions and intended DNA targets.

		Positions				
Catalytic domain	Target	120	123	127	136	137
A	CC[a]	Ile	Thr	Leu	Ile	Gly
B	GC	Ile	Thr	Leu	Arg	Phe
Γ	GT	Leu	Val	Ile	Arg	Trp
Δ	CA	Ile	Val	Leu	Arg	Phe
ε[b]	AC	Leu	Pro	His	Arg	Phe
ζ[c]	TT	Ile	Thr	Arg	Ile	Phe

[a]Indicates wild-type DNA target.

[b]The ε catalytic domain also contains the substitutions E117L and L118S.

[c]The ζ catalytic domain also contains the substitutions M124S, R131I and P141R.

[0244] Luciferase assays.

[0245] Human embryonic kidney (HEK) 293 and 293T cells (ATCC) were maintained in DMEM containing 10% (vol/vol) FBS and 1% (vol/vol) Antibiotic-Antimycotic (Anti-Anti; Gibco). HEK293 cells were seeded onto 96-well plates at a density of 4×10^4 cells per well and established in a humidified 5% CO_2 atmosphere at 37°C. At 24 hr after seeding, cells were transfected with 150 ng pcDNA-ZFR-L 1-18, 150 ng pcDNA-ZFR-R 1-18, 2.5 ng pGL3-ZFR-1, 2, 3. or 18 and 1 ng pRL-CMV using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. At 48 hr after transfection, cells were lysed with Passive Lysis Buffer (Promega) and luciferase expression was determined with the Dual-

Luciferase Reporter Assay System (Promega) using a Veritas Microplate Luminometer (Turner Biosystems).

[0246] Integration assays.

[0247] HEK293 cells were seeded onto 6-well plates at a density of 5 x 10^5 cells per well and maintained in serum-containing media in a humidified 5% CO_2 atmosphere at 37°C. At 24 hr after seeding, cells were transfected with 1 ^g pcDNA-ZFR-L-1, 2 or 3 and 1 ^g pcDNA-ZFR- R-1, 2 or 3 and 200 ng pBPS-ZFR-1, 2 or 3 using Lipofectamine 2000 according to the manufacturer's instructions. At 48 hr after transfection, cells were split onto 6-well plates at a density of 5 x 10^4 cells per well and maintained in serum-containing media with 2 ^g mL" puromycin. Cells were harvested upon reaching 100% confluence and genomic DNA was isolated with the Quick Extract DNA Extraction Solution (Epicentre). ZFR targets were PCR amplified with the following primer combinations: ZFR-Target-1, 2 or 3-Fwd and ZFR-Target-1, 2 or 3-Rev (Unmodified target); ZFR-Target-1, 2 or 3-Fwd and CMV-Mid-Prim-1 (Forward integration); and CMV-Mid-Prim-1 and ZFR-Target-1, 2 or 3-Rev (Reverse integration) using the Expand High Fidelity Taq System (Roche). For clonal analysis, at 2 days post-transfection 1 x 10^5 cells were split onto a 100 mm dish and maintained in serum-containing media with 2 ^g mL^{-1} puromycin. Individual colonies were isolated with 10 mm X 10 mm open-ended cloning cylinders with sterile silicone grease (Millipore) and expanded in culture. Cells were harvested upon reaching 100% confluence and genomic DNA was isolated and used as template for PCR, as described above. For colony counting assays, at 2 days post-transfection cells were split into 6-well plates at a density of 1 x 10^4 cells per well and maintained in serum-containing media with or without 2 ^g mL^{-1} puromycin. At 16 days, cells were stained with a 0.2% crystal violet solution and integration efficiency was determined by counting the number of colonies formed in puromycin-containing media divided by the number of colonies formed in the absence of puromycin. Colony number was determined by automated counting using the GelDoc XR Imaging System (Bio-Rad).

[0248] Results.

[0249] Specificity profile of the Gin recombinase.

[0250] In order to re-engineer serine recombinase catalytic specificity, a detailed understanding was developed of the factors underlying substrate recognition by this family of enzymes. To accomplish this, the ability of an activated mutant of the catalytic domain of the DNA invertase Gin to recombine a comprehensive set of symmetrically substituted target

sites was evaluated. The Gin catalytic domain recombines a pseudo-symmetric 20-bp core that consists of two 10-bp half-site regions. This collection of recombination sites therefore contained each possible single-base substitution at positions 10, 9, 8, 7, 6, 5, and 4 and each possible two-base combination at positions 3 and 2 and in the dinucleotide core. Recombination was determined by split gene reassembly, a previously described method that links recombinase activity to antibiotic resistance.

[0251] In general, it was found that Gin tolerates (i) 12 of the 16 possible two-base combinations at the dinucleotide core (AA, AT, AC, AG, TA, TT, TC, TG, CA, CT, GA, GT); (ii) 4 of the 16 possible two-base combinations at positions 3 and 2 (CC, CG, GG and TG); (iii) a single A to T substitution at positions 6, 5, or 4; and (iv) all 12 possible single-base substitutions at positions 10, 9, 8, and 7 (Figure 31A-D). Further, it was found that Gin could recombine a target site library containing at least 10^6 (of a possible 4.29×10^9) unique base combinations at positions 10, 9, 8, and 7 (Figure 31D).

[0252] These findings are consistent with observations made from crystal structures of the yS resolvase, which indicate that (i) the interactions made by the recombinase dimer across the dinucleotide core are asymmetric and predominately non-specific; (ii) the interactions between an evolutionarily conserved Gly-Arg motif in the recombinase arm region and the DNA minor groove imposes a requirement for adenine or thymine at positions 6, 5, and 4; and (iii) there are no sequence-specific interactions between the arm region and the minor groove at positions 10, 9, 8, or 7 (Figure 31E). These results are also consistent with studies that focused on determining the DNA-binding properties of the closely related Hin recombinase.

[0253] Re-engineering Gin recombinase catalytic specificity.

[0254] Based on the finding that Gin tolerates conservative substitutions at positions 3 and 2 (i.e., CC, CG, GG, and TG), whether Gin catalytic specificity could be re- engineered to specifically recognize core sequences containing each of the 12 base combinations not tolerated by the native enzyme (Figure 32A) was investigated. In order to identify the specific amino acid residues involved in DNA recognition by Gin, the crystal structures of two related serine recombinases, the y6 resolvase and Sin recombinase, in complex with their respective DNA targets were examined. Based on these models, five residues were identified that contact DNA at positions 3 and 2: Leu 123, Thr 126, Arg 130, Val 139, and Phe 140 (numbered according to the y5 resolvase) (Figure 32B). Random mutagenesis was performed on the equivalent residues in the Gin catalytic domain (Ile 120, Thr 123, Leu 127, Ile 136,

and Gly 137) by overlap extension PCR and constructed a library of ZFR mutants by fusing these catalytic domain variants to an unmodified copy of the 'H1' ZFP. The theoretical size of this library was 3.3×10^7 variants.

[0255] The ZFR library was cloned into substrate plasmids containing one of the five base combinations not tolerated by the native enzyme (GC, GT, CA, AC, or TT) and enriched for active ZFRs by split gene reassembly (Figure 32C). After 4 rounds of selection, it was found that the activity of each ZFR population increased >1,000-fold on DNA targets containing GC, GT, CA, and TT substitutions and >100-fold on a DNA target containing AC substitutions (Figure 32D).

[0256] Individual recombinase variants were sequenced from each population and found that a high level of amino acid diversity was present at positions 120, 123, and 127 and that >80% of selected clones contained Arg at position 136 and Trp or Phe at position 137 (Figure 36). These results suggest that positions 136 and 137 play critical roles in the recognition of unnatural core sequences. The ability of each selected enzyme to recombine its target DNA was evaluated and it was found that nearly all recombinases showed activity (>10% recombination) and displayed a >1,000-fold shift in specificity toward their intended core sequence (Figure 37). As with the parental Gin, it was found that several recombinases tolerated conservative substitutions at positions 3 and 2 (i.e., cross-reactivity against GT and CT or AC and AG), indicating that a single re- engineered catalytic domain could be used to target multiple core sites (Figure 37).

[0257] In order to further investigate recombinase specificity, the recombination profiles were determined of five Gin variants (hereafter designated Gin p, y, 6, e and Z) shown to recognize nine of the 12 possible two-base combinations not tolerated by the parental enzyme (GC, TC, GT, CT, GA, CA, AG, AC, and TT) (Table 1). Gin p, 6, and e recombined their intended core sequences with activity and specificity comparable to that of the parental enzyme (hereafter referred to as Gin a) and that Gin y and Z were able to recombine their intended core sequences with specificity exceeding that of Gin a (Figure 32E). Each recombinase displayed >1,000-fold preference for adenine or thymine at positions 6, 5, and 4 and showed no base preference at positions 10, 9, 8, and 7 (Figure 38). These results indicate that mutagenesis of the DNA-binding arm did not compromise recombinase specificity. It was not possible to select for Gin variants capable of tolerating AA, AT, or TA substitutions at positions 3 and 2. One possibility for this result is that DNA targets containing >4

consecutive A-T bps might exhibit bent DNA conformations that interfere with recombinase binding and/or catalysis.

[0258] Engineering ZFRs to recombine user-defined sequences

[0259] Whether ZFRs composed of the re-engineered catalytic domains could recombine pre-determined sequences was investigated. To test this possibility, the human genome (GRCh37 primary reference assembly) was searched for potential ZFR target sites using a 44-bp consensus recombination site predicted to occur approximately once every 400,000 bp of random DNA (Fig. 4A). This ZFR consensus target site, which was derived from the core sequence profiles of the selected Gin variants, includes approximately 7×10^8 (of a possible 1.0955×10^{12}) unique 20-bp core combinations predicted to be tolerated by the 21 possible catalytic domain combinations and a conservative selection of modular zinc finger domains that excludes 5'- CNN-3' and 5'-TNN-3' triplets within each ZFBS. Using ZFP specificity as the primary determinant for selection, 18 possible ZFR target sites across 8 human chromosomes (Chr. 1, 2, 4, 6, 7, 11, 13 and X) at non-protein coding loci were identified. On average, each 20- bp core showed ~46% sequence identity to the core sequence recognized by the native Gin catalytic domain (Figure 33B). Each corresponding ZFR was constructed by modular assembly (see Materials and Methods).

[0260] To determine whether each ZFR pair could recombine its intended DNA target, a transient reporter assay was developed that correlates ZFR-mediated recombination to reduced luciferase expression (Figures 33A and 39). To accomplish this, ZFR target sites were introduced upstream and downstream an SV40 promoter that drives expression of a luciferase reporter gene. Human embryonic kidney (HEK) 293T cells were co-transfected with expression vectors for each ZFR pair and its corresponding reporter plasmid. Luciferase expression was measured 48 hr after transfection. Of the 18 ZFR pairs analyzed, 38% (7 of 18) reduced luciferase expression by >75-fold and 22% (4 of 18) decreased luciferase expression by >140-fold (Figure 33B). In comparison, GinC4, a positive ZFR control designed to target the core sequence recognized by the native Gin catalytic domain, reduced luciferase expression by 107 fold. Overall, it was found that 50% (9 of 18) of the evaluated ZFR pairs decreased luciferase expression by at least 20-fold. Importantly, virtually every catalytic domain that displayed significant activity in bacterial cells (>20% recombination) was successfully used to recombine at least one naturally occurring sequence in mammalian cells.

[0261] In order to evaluate ZFR specificity, separately HEK293T cells were co-transfected with expression plasmids for the nine most active ZFRs with each non-cognate reporter plasmid. Each ZFR pair demonstrated high specificity for its intended DNA target and 77% (7 of 9) of the evaluated ZFRs showed an overall recombination specificity nearly identical to that of the positive control GinC4 (Fig. 4C). To establish that reduced luciferase expression is the product of the intended ZFR heterodimer and not the byproduct of recombination-competent ZFR homodimers, the contribution of each ZFR monomer to recombination was measured. Co- transfection of the ZFR 1 'left' monomer with its corresponding reporter plasmid led to a modest reduction in luciferase expression (total contribution to recombination: ~22%), but the vast majority of individual ZFR monomers (16 of 18) did not significantly contribute to recombination (<10% recombination), and many (7 of 18) showed no activity (Figure 39). Taken together, these studies indicate that ZFRs can be engineered to recombine user-defined sequences with high specificity.

[0262] Engineered ZFRs mediate targeted integration into the human genome.

[0263] Whether ZFRs could integrate DNA into endogenous loci in human cells was evaluated next. To accomplish this, HEK293 cells were co-transfected with ZFR expression vectors and a corresponding DNA donor plasmid that contained a specific ZFR target site and a puromycin- resistance gene under the control of an SV40 promoter. For this analysis, ZFR pairs 1, 2, and 3, were used which were designed to target non-protein coding loci on human chromosomes 4, X, and 4, respectively (Figure 34A). At 2 days post-transfection, cells were incubated with puromycin- containing media and measured integration efficiency by determining the number of puromycin- resistant ($puro^R$) colonies. It was found that (i) co-transfection of the donor plasmid and the corresponding ZFR pair led to a >12-fold increase in $puro^R$ colonies in comparison to transfection with donor plasmid only and that (ii) co-transfection with both ZFRs led to a 6-to 9fold increase in $puro^R$ colonies in comparison to transfection with individual ZFR monomers (Figure 34B). In order to evaluate whether ZFR pairs correctly targeted integration, genomic DNA was isolated from $puro^R$ populations and amplified each targeted locus by PCR. The PCR products corresponding to integration in the forward and/or reverse orientations were observed at each locus targeted by these ZFR pairs (Figure 34C). Next, to determine the overall specificity of ZFR-mediated integration, genomic DNA was isolated from clonal cell populations and evaluated plasmid insertion by PCR. This analysis revealed targeting efficiencies of 8.3% (1 of 12 clones), 14.2% (5 of 35 clones), and 9.1% (1 of 11 clones) for ZFR pairs 1, 2, and 3, respectively (Fig. S6). Sequence

analysis of each PCR product confirmed ZFR-mediated integration (Figure 34D). Taken together, these results indicate that ZFRs can be designed to accurately integrate DNA into endogenous loci.

[0264] Finally, it is noted that the ZFR-1 'left' monomer was found to target integration into the ZFR-1 locus (Figure 34C). This result, which is consistent with the luciferase reporter studies described above (Figure 39) indicates that recombination-competent ZFR homodimers have the capacity to mediate off-target integration. Future development of an optimized heterodimeric ZFR architecture and a comprehensive evaluation of off-target integration should lead to the design of ZFRs that demonstrate greater targeting efficiency.

[0265] It is herein shown that ZFRs can be designed to recombine user-defined sequences with high specificity and that ZFRs can integrate DNA into pre-determined endogenous loci in human cells. By combining substrate specificity analysis and directed evolution, virtually all sequence requirements imposed by the ZFR catalytic domain were eliminated. Using the archive of 45 pre-selected zinc-finger modules, it is estimated that ZFRs can be designed to recognize $>1 \times 10^{22}$ unique 44-bp DNA sequences, which corresponds to approximately one potential ZFR target site for every 4,000 bp of random sequence. Construction of customized zinc-finger domains by selection would further extend targeting. The re-engineered catalytic domains described herein will be compatible with recently described TAL effector recombinases. This work demonstrates the feasibility of generating ZFRs with custom specificity and illustrates the potential utility of ZFRs for a wide range of applications, including genome engineering, synthetic biology, and gene therapy.

[0266] Although the invention has been described with reference to the above example, it will be understood that modifications and variations are encompassed within the spirit and scope of the invention. Accordingly, the invention is limited only by the following claims.

What is claimed is:

1. A chimeric polypeptide comprising:

a) a recombinase, a transcription factor or nuclease; and

b) a transcription activator-like effector (TALE) protein.

2. The chimeric protein of claim 1, wherein the TALE protein is truncated.

3. The chimeric protein of claim 2, wherein the TALE protein comprises a C-terminal or N-terminal truncation.

4. The chimeric protein of claim 3, wherein the TALE protein comprises a C-terminal truncation.

5. The chimeric protein of claim 1, wherein the TALE protein is selected from the group consisting of: AcrXa7, Tal1c, and PthXo1.

6. The chimeric protein of claim 1, wherein the TALE protein comprises an amino acid sequence as set forth in SEQ ID NO: 2.

7. The chimeric protein of claim 6, wherein the TALE protein comprises a C-terminal truncation.

8. The chimeric protein of claim 7, wherein the TALE protein is truncated between amino acid residues 27 and 268, 92 and 134, 120 and 129, 74 and 147, or 87 and 120.

9. The chimeric protein of claim 8, wherein the TALE protein is truncated at amino acid residue 28, 74, 87, 92, 95, 120, 124, 128, 129, 147 and 150.

10. The chimeric protein of claim 1, wherein the recombinase is selected from the group consisting of:

(a) Tn3, also known as EcoTn3; Hin, also known as StyHin; Gin, also known as MuGin; Sin; Beta; Pin; Min; Din; Cin; EcoTn21; SfaTn917; BmeTn5083; Bme53; Cpe; SauSK1; SauSK41; SauTn552; Ran; Aac; Lla; pMER05; Mlo92; Mlo90; Rrh; Pje; Req; PpsTn5501; Pae; Xan; ISXc5; Spy; RhizY4cG; SarpNL1; SsolSC1904a; SsolSC1904b; SsoISC1913; Aam606; MjaM0014; Pab; HpylS607; MtulS_Y349; MtuRv2792c; MtuRv2979c; MtuRv3828c; MtuRv0921; MceRv0921; TnpX; TndX; WwK; lactococcal phage TP901-1 serine recombinase; *S. pyogenes* phage ϕ370.1 serine recombinase; *S. pyogenes* phage ϕFC1 serine recombinase; *Listeria* phage A118 serine recombinase; *S. coelicolor* chromosome SC3C8.24 serine recombinase; *S. coelicolor* chromosome SC2E1.37 serine recombinase; *S. coelicolor* chromosome SCD78.04c serine recombinase; *S. coelicolor* chromosome SC8F4.15c serine recombinase; *S. coelicolor* chromosome SCD12A.23 serine recombinase; *S. coelicolor* chromosome SCH10.38c serine recombinase; *S. coelicolor*

chromosome SCC88.14 serine recombinase; *Streptomyces* phage ϕC31 serine recombinase; *Streptomyces* phage R4 serine recombinase; *Bacillus* phage ϕ105 serine recombinase; *Bacillus* phage SPBc2 serine recombinase; *Bacillus* prophage SKIN serine recombinase; *S. aureus ccrA* serine recombinase; *S. aureus ccrB* serine recombinase; *M. tuberculosis* phage Bxb1 serine recombinase; *M. tuberculosis* prophage ϕRV1 serine recombinase; YBCK_ECOLI; Y4bA; Bja; Spn; Cac 1956; and Cac 1954; and

(b) muteins of recombinases of (a).

11. The chimeric protein of claim 10, wherein the recombinase is selected from the group consisting of Gin, Hin, Tn3, Sin, Beta, Pin. Min, Din, and Cin and muteins of Gin, muteins of Hin, muteins of Sin, muteins of Beta, muteins of Pin, muteins of Min, muteins of Din, muteins of Cin, muteins of Tn3.

12. The chimeric protein of claim 10, wherein the recombinase is Gin.

13. The chimeric protein of claim 1, wherein the recombinase is Gin and the TALE protein is AcrXa7.

14. An isolated nucleic acid molecule encoding the chimeric protein of any of claims 1-13.

15. An expression cassette comprising the nucleic acid molecule of claim 14.

16. A vector comprising the expression cassette of claim 15.

17. A host cell transformed or transfected with the nucleic acid molecule of claim 14 or the vector of claim 16.

18. A method for site-specific recombination comprising:

(a) providing a DNA sequence comprising at least two binding sites for specifically interacting with the chimeric protein of claim 1; and

(b) reacting the DNA sequence with the chimeric protein, wherein the chimeric protein catalyzes a site-specific recombination event in which both strands of the DNA sequence are cleaved between the two sites specifically interacting with the chimeric protein.

19. The method of claim 18, wherein the site-specific recombination event is an inversion.

20. The method of claim 18, wherein the site-specific recombination event is an integration.

21. The method of claim 18, wherein the site-specific recombination event is a resolution.

22. A method for gene therapy comprising administering to a subject a composition comprising a nucleic acid molecule encoding the chimeric protein of claim 1, wherein upon

expression of the nucleic acid molecule, a gene present in the genome of the subject is specifically removed or inactivated.

23. The method of claim 22, further comprising administering to the subject a nucleic acid molecule comprising a functional replacement of the gene.

24. A pharmaceutical composition comprising:

a) the chimeric protein of claim 1; and

b) a pharmaceutically acceptable carrier.

25. A pharmaceutical composition comprising:

a) a nucleic acid molecule encoding the chimeric protein of claim 1; and

b) a pharmaceutically acceptable carrier.

26. A transgenic organism produced by an act of recombination catalyzed by the chimeric protein of claim 1.

27. A method for gene therapy comprising administering to a subject a cell comprising a nucleic acid molecule having the DNA sequence generated by the method of any of claims 18-21.

28. A method for modifying a genome of an organism, comprising modifying the genome of the organism by performing site-specific recombination on a nucleic acid molecule of the genome using the method of any of claims 18-22.

29. The method of claim 28, wherein the organism is a prokaryote, bacterium, virus or eukaryote.

30. A method of generating a transcription activator-like effector (TALE) protein binding domain which specifically binds a desired nucleotide, comprising:

a) randomizing the amino acid sequence of the TALE protein binding domain by mutating an amino acid residue within a variable di-residue (RVD), or within 1 to 2 amino acid residues N-terminal or C-terminal of the RVD;

b) selecting for the randomized TALE protein binding domain of (a), wherein the TALE protein binding domain specifically binds to the desired nucleotide.

31. An isolated protein comprising a transcription activator-like effector (TALE) protein binding domain generated by the method of claim 30.

32. The isolated protein of claim 31, wherein the protein comprises from 1 to 40 TALE protein binding domains.

33. The isolated protein of claim 32, wherein the protein specifically binds a target nucleotide sequence.

34. The isolated protein of claim 33, wherein the protein comprises nuclease or recombinase activity.

35. The isolated protein of claim 33, wherein the protein regulates gene expression.

36. An isolated nucleic acid molecule encoding a protein comprising the TALE protein binding domain of any of claims 31-35.

37. An expression cassette comprising the nucleic acid molecule of claim 36.

38. A vector comprising the expression cassette of claim 37.

39. A host cell transformed or transfected with the nucleic acid molecule of claim 37 or the vector of claim 38.

40. An isolated polypeptide comprising a *Xanthamonus* derived transcription activator-like effector (TALE) protein, the TALE protein having an N-terminal domain (NTD) comprising an amino acid sequence as set forth in SEQ ID NO: 3 (VGKQWSGARAL) having one or more mutations or deletions selected from: Q is Y, Q is S, Q is R, W is R, W is G, W is deleted, S is R, S is H, S is A, S is N, and S is T.

41. The polypeptide of claim 40, wherein the NTD comprises an amino acid sequence selected from: VGKYRGARAL (SEQ ID NO: 4), VGKSRSGARAL (SEQ ID NO: 5), VGKYHGARAL (SEQ ID NO: 6), and VGKRGAGARAL (SEQ ID NO: 7).

42. An isolated polypeptide comprising a *Ralstonia* derived transcription activator-like effector (TALE) protein, the TALE protein having an N-terminal domain (NTD) comprising an amino acid sequence as set forth in SEQ ID NO: 8 ($IVDIAR_1QR_2SGDLA$) having one or more mutations or deletions selected from: R_1 is K, Q is Y, Q is S, Q is R, R_2 is W, R_2 is G, R_2 is deleted, S is R, S is H, S is A, S is N, and S is T.

43. The polypeptide of claim 42, wherein the NTD comprises an amino acid sequence selected from: IVDIARQWSGDLA (SEQ ID NO: 9), IVDIARYRGDLA (SEQ ID NO: 10), IVDIARSRSGDLA (SEQ ID NO: 11), IVDIARYHGDLA (SEQ ID NO: 12), and IVDIARRGAGDLA (SEQ ID NO: 13).

44. The polypeptide of any of claims 40-43 further comprising a recombinase domain or a nuclease domain.

45. An isolated nucleic acid molecule encoding the polypeptide of any of claims 40-44.

46. An expression cassette comprising the nucleic acid molecule of claim 45.

47. A vector comprising the expression cassette of claim 46.

48. A host cell transformed or transfected with the nucleic acid molecule of claim 45 or the vector of claim 47.

49. A method of generating a transcription activator-like effector (TALE) protein N-terminal domain (NTD), comprising:

a) randomizing an amino acid sequence of the NTD by mutating or deleting one or more amino acid residues within the NTD, wherein the amino acid sequence is SEQ ID NO: 14 (VGKXXXGAR) or SEQ ID NO: 15 (VDIAXXXXGDLA); and

b) selecting for the randomized TALE protein NTD of (a), wherein the TALE protein NTD specifically binds to a desired nucleotide or exhibits enhanced activity.

50. A method of generating a plurality of zinc finger recombinase (ZFRs) proteins having catalytic specificity greater than the corresponding wild type recombinase comprising:

a) performing random mutagenesis on a recombinase catalytic domain at positions equivalent to Gin Ile120, Thr123, Leu127, Ile136 and Gly137 or a combination thereof, mutating the DNA at positions 2 and 3 for each amino acid;

b) fusing the recombinase catalytic domain of a) with a plurality of zinc finger binding domains to form ZFRs, and

c) enriching for ZFRs of b) having catalytic specificity greater than the corresponding wild type recombinase.

51. The method of claim 50, wherein the ZFRs have increased catalytic activity on DNA targets selected from GC, GT, CA, TT and AC.

52. The method of claim 50, wherein the recombinase catalytic domain is mutagenized at Ile136 and/or Gly137.

53. The method of claim 50, wherein the ZFRs have increased target specificity at chromosomes 1, 2, 4, 6, 7, 11, 13, and/or X.

54. The method of claim 50, wherein the ZFRs are in a vector.

55. The method of claim 50, wherein the recombinase catalytic domain is from a recombinase selected from the group consisting of:

a) Tn3, also known as EcoTn3; Hin, also known as StyHin; Gin, also known as MuGin; Sin; Beta; Pin; Min; Din; Cin; EcoTn21; SfaTn917; BmeTn5083; Bme53; Cpe; SauSK1; SauSK41; SauTn552; Ran; Aac; Lla; pMER05; Mlo92; Mlo90; Rrh; Pje; Req; PpsTn5501; Pae; Xan; ISXc5; Spy; RhizY4cG; SarpNL1; SsolSC1904a; SsolSC1904b; SsoISC1913; Aam606; MjaM0014; Pab; HpylS607; MtulS_Y349; MtuRv2792c; MtuRv2979c; MtuRv3828c; MtuRv0921; MceRv0921; TnpX; TndX; WwK; lactococcal phage TP901-1 serine recombinase; *S. pyogenes* phage ϕ370.1 serine recombinase; *S. pyogenes* phage ϕFC1 serine recombinase; *Listeria* phage A118 serine recombinase; *S.*

coelicolor chromosome SC3C8.24 serine recombinase; *S. coelicolor* chromosome SC2E1.37 serine recombinase; *S. coelicolor* chromosome SCD78.04c serine recombinase; *S. coelicolor* chromosome SC8F4.15c serine recombinase; *S. coelicolor* chromosome SCD12A.23 serine recombinase; *S. coelicolor* chromosome SCH10.38c serine recombinase; *S. coelicolor* chromosome SCC88.14 serine recombinase; *Streptomyces* phage ϕC31 serine recombinase; *Streptomyces* phage R4 serine recombinase; *Bacillus* phage ϕ105 serine recombinase; *Bacillus* phage SPBc2 serine recombinase; *Bacillus* prophage SKIN serine recombinase; *S. aureus ccrA* serine recombinase; *S. aureus ccrB* serine recombinase; *M. tuberculosis* phage Bxb1 serine recombinase; *M. tuberculosis* prophage ϕRV1 serine recombinase; YBCK_ECOLI; Y4bA; Bja; Spn; Cac 1956; and Cac 1954; and

b) muteins of a).

56. The method of claim 6, wherein the recombinase catalytic domain is from a mutein of Gin, Hin, Sin, Beta, Pin, Min, Din, Cin, or Tn3.

57. A chimeric polypeptide generated by the method of claim 50.

58. The chimeric polypeptide of claim 57, wherein the recombinase catalytic domain is selected from the group consisting of:

a) Tn3, also known as EcoTn3; Hin, also known as StyHin; Gin, also known as MuGin; Sin; Beta; Pin; Min; Din; Cin; EcoTn21; SfaTn917; BmeTn5083; Bme53; Cpe; SauSK1; SauSK41; SauTn552; Ran; Aac; Lla; pMER05; Mlo92; Mlo90; Rrh; Pje; Req; PpsTn5501; Pae; Xan; ISXc5; Spy; RhizY4cG; SarpNL1; SsolSC1904a; SsolSC1904b; SsoISC1913; Aam606; MjaM0014; Pab; HpylS607; MtulS_Y349; MtuRv2792c; MtuRv2979c; MtuRv3828c; MtuRv0921; MceRv0921; TnpX; TndX; WwK; lactococcal phage TP901-1 serine recombinase; *S. pyogenes* phage ϕ370.1 serine recombinase; *S. pyogenes* phage ϕFC1 serine recombinase; *Listeria* phage A118 serine recombinase; *S. coelicolor* chromosome SC3C8.24 serine recombinase; *S. coelicolor* chromosome SC2E1.37 serine recombinase; *S. coelicolor* chromosome SCD78.04c serine recombinase; *S. coelicolor* chromosome SC8F4.15c serine recombinase; *S. coelicolor* chromosome SCD12A.23 serine recombinase; *S. coelicolor* chromosome SCH10.38c serine recombinase; *S. coelicolor* chromosome SCC88.14 serine recombinase; *Streptomyces* phage ϕC31 serine recombinase; *Streptomyces* phage R4 serine recombinase; *Bacillus* phage ϕ105 serine recombinase; *Bacillus* phage SPBc2 serine recombinase; *Bacillus* prophage SKIN serine recombinase; *S. aureus ccrA* serine recombinase; *S. aureus ccrB* serine recombinase; *M. tuberculosis* phage Bxb1 serine recombinase; *M. tuberculosis* prophage ϕRV1 serine recombinase;

YBCK_ECOLI; Y4bA; Bja; Spn; Cac 1956; and Cac 1954; and

b) muteins of a).

59. The method of claim 9, wherein the recombinase catalytic domain is from a mutein of Gin, Hin, Sin, Beta, Pin, Min, Din, Cin, or Tn3.

60. An isolated nucleic acid molecule encoding the polypeptide of claim 57.

61. An expression cassette comprising the nucleic acid molecule of claim 60.

62. A vector comprising the expression cassette of claim 61.

63. An isolated host cell containing the vector of claim 62.

64. A method for site-specific integration into a DNA sequence comprising:

contacting the DNA sequence with a chimeric polypeptide of claim 57, wherein the chimeric polypeptide catalyzes site-specific integration.

65. A method for gene therapy comprising administering to a subject a composition comprising a nucleic acid molecule encoding the chimeric polypeptide of claim 57, wherein upon expression of the nucleic acid molecule, a gene present in the genome of the subject is specifically removed or inactivated.

66. The method of claim 65, further comprising administering to the subject a nucleic acid molecule comprising a functional replacement of the gene.

67. A pharmaceutical composition comprising:

a) the chimeric polypeptide of claim 57; and

b) a pharmaceutically acceptable carrier.

68. A pharmaceutical composition comprising:

a) a nucleic acid molecule encoding the chimeric polypeptide of claim 57; and

b) a pharmaceutically acceptable carrier.

69. A transgenic organism produced by recombination catalyzed by the chimeric polypeptide of claim 57.