METHODS FOR TARGETED INSERTION OF DNA IN GENES

REFERENCE TO RELATED APPLICATIONS

This application claims the benefit of previously filed and co-pending applications USSN 62/746,497 filed October 16, 2018, USSN 62/830,654 filed April 8, 2019, and USSN 62/864,432 filed June 20, 2019, the contents of which are incorporated herein by reference in their entirety.

SEQUENCE LISTING

The instant application contains a Sequence Listing which has been submitted in ASCII format via EFS-Web and is hereby incorporated by reference in its entirety. Said ASCII copy, created on October 14, 2019 is named SEQUENCE_LISTING_BA2018-4WO_P12987WO00.txt and is 517,036 bytes in size.

TECHNICAL FIELD

The present document is in the field of genome editing. More specifically, this document relates to the targeted modification of endogenous genes using rare-cutting endonucleases or transposases.

BACKGROUND

Monogenic disorders are caused by one or more mutations in a single gene, examples of which include sickle cell disease (hemoglobin-beta gene), cystic fibrosis (cystic fibrosis transmembrane conductance regulator gene), and Tay-Sachs disease (beta-hexosaminidase A gene). Monogenic disorders have been an interest for gene therapy, as replacement of the defective gene with a functional copy could provide therapeutic benefits. However, one bottleneck for generating effective therapies includes the size of the functional copy of the gene. Many delivery methods, including those that use viruses, have size limitations which hinder the delivery of large transgenes. Further, many genes have alternative splicing patterns resulting in a single gene coding for multiple proteins. Methods to correct partial regions of a defective gene may provide an alternative means to treat monogenic disorders.

SUMMARY

Gene editing holds promise for correcting mutations found in genetic disorders; however, many challenges remain for creating effective therapies for individual disorders, including those that are caused by gain-of-function mutations, or where precise repair is required. These challenges are seen with disorders such as spinocerebellar ataxia 3 and spinocerebellar ataxia 6, wherein the disorder is caused by gain-of-function mutations (expanded trinucleotide repeat) at the 3' end of the genes.

The methods described herein provide novel approaches for correcting mutations found at the 3' end of genes. The disclosure herein is based at least in part on the design of bimodule transgenes compatible with integration through multiple repair pathways. The transgenes described herein can be integrated into genes by the homologous recombination pathway, the non-homologous end joining pathway, or both the homologous recombination and non-homologous end joining pathway, or through transposition. Further, the outcome of integration in any case (HR, NHEJ forward, NHEJ reverse; transposition forward, or transposition reverse) can result in precise correction/alteration of the target gene's protein product. The transgenes described herein can be used to fix or introduce mutations in the 3' region of genes-of-interest. The methods are particularly useful in cases where precise editing of genes is necessary, or where the mutated endogenous gene being targeted cannot be 'replaced' by a synthetic copy because it exceeds the size capacity of standard vectors or viral vectors. The methods described herein can be used for applied research (e.g., gene therapy) or basic research (e.g., creation of animal models, or understanding gene function).

The methods described herein are compatible with current in vivo delivery vehicles (e.g., adeno-associated virus vectors and lipid nanoparticles), and they address several challenges with achieving precise alteration of gene products.

In one embodiment, this document features a method for integrating a transgene into an endogenous gene. The method can include delivery of a transgene, where the transgene harbors a first and second splice acceptor sequence, a first and second partial coding sequence, and a first and second terminator. In some embodiments, the first and

second terminators can be replaced with a single bidirectional terminator. The method further includes administering one or more rare-cutting endonucleases targeted to a site within the endogenous gene, where the transgene is then integrated into the endogenous gene. The transgene can be targeted to a site within an intron or at an intron-exon junction. The first and second partial coding sequences can be oriented in a tail-to-tail orientation, such that integration of the transgene in either direction (i.e., forward or reverse) by NHEJ can result in precise alteration of the gene's protein product. In other embodiments, the transgene can include a left and right homology arm to enable integration by HR. These transgenes can be harbored within an adeno-associated virus vector (AAV), wherein the transgene can be integrated via HR (through the homology arms) or by NHEJ forward direction or NHEJ reverse direction (through direct integration of the AAV vector within a targeted double-strand break). In an embodiment, vectors with a first and second coding sequence and a left and right homology arm can further include a first and second site for cleavage by one or more rare-cutting endonucleases. Cleavage by the one or more rare-cutting endonucleases can result in liberation of a linear transgene with homology arms, capable of integrating into the genome through HR or NHEJ. In another embodiment, vectors with a first and second coding sequence can be flanked by a first and second site for cleavage by one or more rare-cutting endonucleases. Cleavage by the one or more rare-cutting endonucleases can result in liberation of a linear transgene, capable of integrating into the genome through NHEJ. In another embodiment, vectors with a first and second coding sequence can be flanked by a left and right transposon end. Delivery of a CRISPR-associated transposase (e.g., Cas6/7/8 along with TniQ, TnsA, TnsB, and TnsC) can result in integration of the transgene through transposition.

The methods can be used to alter the C-terminus of proteins produced by endogenous genes. In some embodiments, the endogenous gene can include the ATXN3 gene or CACNA1A gene. ATXN3 is a gene that encodes the enzyme ataxin-3. Ataxin-3 is a member in the ubiquitin-proteasome system which facilitates the destruction of excess or damaged proteins. Spinocerebellar ataxia type 3 is a genetic disorder caused by a trinucleotide repeat expansion within the 3' end of the ATXN3 gene. CACNA1A is a

gene that encodes proteins involved in the formation of calcium channels. Spinocerebellar ataxia type 6 is a genetic disorder caused by mutations in the CACNA1A gene. The mutations which cause SCA6 include a trinucleotide repeat expansion in the 3' end of the CACNA1A gene. In some embodiments, the methods provided herein can be used to alter the 3' end of the endogenous ATXN3 gene or CACNA1A gene. In specific embodiments, the target for integration of the transgenes described herein can be intron 9 of the ATXN3 gene or intron 46 of the CACNA1A gene.

Unless otherwise defined, 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 pertains. Although methods and materials similar or equivalent to those described herein can be used to practice the invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety for all purposes. In case of conflict, the present specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

The details of one or more embodiments of the invention are set forth in the description below. Other features, objects, and advantages of the invention will be apparent from the description and from the claims.

DESCRIPTION OF DRAWINGS

FIG. 1 is an illustration of the transgenes for the targeted insertion into endogenous genes. TS1, target site 1; SA1, splice acceptor site 1, CDS1, coding sequence 1; T1, terminator 1, TS2, target site 2; SA2, splice acceptor site 2, CDS2, coding sequence 2; T2, terminator 2; HA1, homology arm 1; HA2, homology arm 2; BT1, bidirectional terminator 1; AS1, additional sequence 1; AS2, additional sequence 2.

FIG. 2 is an illustration showing integration of a transgene into an exemplary gene. The transgene comprises two target sites for one or more rare-cutting endonucleases, two splice acceptor sequences, two coding sequences (3.1 and 3.2) and two terminators (T). Integration proceeds through non-homologous end joining (NHEJ).

FIG. 3 is an illustration showing integration of a transgene into an exemplary gene. The transgene comprises two homology arms, two target sites for one or more rare-cutting endonucleases, two splice acceptor sequences, two coding sequences (3.1 and 3.2) and two terminators. Integration proceeds through either homologous recombination (HR) or non-homologous end joining (NHEJ).

FIG. 4 is an illustration of exon 46, intron 46 and intron 47 of the CACNA1A gene. Also shown is the pB1011-D1 transgene for integration in the CACNA1A gene.

FIG. 5 is an illustration of the integration outcomes for the pB1011-D1 transgene within the CACNA1A gene.

FIG. 6 is an illustration of exon 9, intron 9, exon 10, intron 10 and exon 11 of the ATXN3 gene. Also shown is the pB1012-D1 transgene for integration in the ATXN3 gene.

FIG. 7 is an illustration of the integration outcomes for the pB1012-D1 transgene within the ATXN3 gene.

FIG. 8 are images of gels detecting integration of transgenes into the ATXN3 gene. 1, 100 bp ladder with top band running at 1,517 bp; 2, pBA1135 5' junction; 3, pBA1136 5' junction; 4, pBA1137 5' junction; 5, pBA1135 3' junction; 6, pBA1136 3' junction; 7, pBA1137 3' junction; 8, 1kb ladder with darker bands running at 500 bp, 1,000 bp and 3,000 bp; 9, 1kb ladder with darker bands running at 500 bp, 1,000 bp and 3,000 bp; 10, pBA1135 inverted 5' junction; 11, 1kb ladder with darker bands running at 500 bp, 1,000 bp and 3,000 bp; 12, pBA1136 inverted 5' junction; 13, 1kb ladder with darker bands running at 500 bp, 1,000 bp and 3,000 bp; 14;, primer pair oNJB156+oNJB113; 15, primer pair 114+162; 16, primer pair oNJB116+oNJB113; 17, primer pair oNJB114+oNJB170; 18, primer pair oNJB167+oNJB170; 19, 100 bp ladder with the dark band running at 500 bp; 20, genomic DNA from transfection with pBA1135 and nuclease; 21, genomic DNA from transfection with pBA1136 and nuclease; 22, genomic DNA from transfection with pBA1137 and nuclease; 23, genomic DNA from transfection with water; 24, no DNA control.

DETAILED DESCRIPTION

Disclosed herein are methods and compositions for modifying the coding sequence of endogenous genes. In some embodiments, the methods include inserting a transgene into an endogenous gene, wherein the transgene provides a partial coding sequence which substitutes for the endogenous gene's coding sequence.

In one embodiment, this document features a method of integrating a transgene into an endogenous gene, the method including administering a transgene, wherein the transgene comprises a first and second splice acceptor sequence, a first and second partial coding sequence, and one bidirectional terminator or a first and second terminator, and administering one or more rare-cutting endonuclease targeted to a site within the endogenous gene, wherein the transgene is integrated within the endogenous gene. The method can include designing the transgene to have the first splice acceptor operably linked to the first partial coding sequence and the second splice acceptor operably linked to the second partial coding sequence. The arrangement can also include having the first partial coding sequence operably linked to the first terminator, and the second partial coding sequence operably linked to the second terminator. In an embodiment, the two terminators can be replaced with a single bidirectional terminator. In an embodiment, transgenes with first and second splice acceptors, first and second partial coding sequences, and first and second terminators can be oriented in a tail-to-tail orientation. The transgenes with a tail-to-tail orientation of sequences can further comprise a first and second target site for one or more rare-cutting endonucleases, wherein the target sites flank the first and second splice acceptors. In another embodiment, the transgenes can comprise a left and right homology arm which flank the first and second splice acceptors. In this embodiment, the transgene can be harbored within an adeno-associated viral vector. In another embodiment, the transgene can further comprise a first and second target site for the one or more rare-cutting endonucleases, wherein the target sites flank the first and second splice acceptors. The first and second target sites can flank the first and second homology arms. In embodiments, the transgenes described herein can be integrated within an intron of the endogenous gene or at an intron-exon junction. The transgenes can be integrated within an intron, or at the intron-exon junction of the

ATXN3 gene or CACNA1A gene. The transgene can comprise a first and second partial coding sequence encoding the peptide produced by exon 10 of a non-pathogenic ATXN3 gene and can be targeted to intron 9, or the intron 9 exon 10 junction, of a pathogenic ATXN3 gene. The transgene can comprise a first and second partial coding sequence encoding the peptide produced by exon 47 of a non-pathogenic CACNA1A gene and can be targeted to intron 46, or the intron 46 exon 47 junction, of a pathogenic CACNA1A gene. In certain embodiments, the rare-cutting endonuclease can be a CRISPR/Cas12a nuclease or a CRISPR/Cas9 nuclease. The first and second partial coding sequences encode the same amino acids. In an embodiment, the first and second coding sequences can differ in nucleic acid sequence but encode the same amino acids. The transgene can be harbored on a vector, wherein the vector format is selected from double-stranded linear DNA, double-stranded circular DNA, or a viral vector. The viral vector can include an adenovirus vector, an adeno-associated virus vector, or a lentivirus vector. The methods described here can be used with a transgene equal to or less than 4.7 kb. The transgene can comprise a first and second partial coding sequence that encode a partial peptide from a functional protein produced by the target endogenous gene. The target endogenous gene can be aberrant.

In another embodiment, this document provides DNA polynucleotides with a first and second splice acceptor sequence, a first and second partial coding sequence, one bidirectional terminator or a first and second terminator, optionally, a first and second homology arm, and, optionally, a first and second rare-cutting endonuclease target site. The DNA polynucleotides can include a design having the first splice acceptor operably linked to the first partial coding sequence and the second splice acceptor operably linked to the second coding sequence. The arrangement can also include having the first partial coding sequence operably linked to the first terminator, and the second partial coding sequence operably linked to the second terminator. In an embodiment, the two terminators can be replaced with a single bidirectional terminator. In an embodiment, DNA polynucleotides with first and second splice acceptors, first and second coding sequences, and first and second terminators can be oriented in a tail-to-tail orientation. The DNA polynucleotides with a tail-to-tail orientation of sequences can further

comprise a first and second target site for one or more rare-cutting endonucleases, wherein the target sites flank the first and second splice acceptors. In another embodiment, the DNA polynucleotides can comprise a left and right homology arm which flank the first and second splice acceptors. In this embodiment, the DNA polynucleotide can be harbored within an adeno-associated viral vector. In another embodiment, the DNA polynucleotides can further comprise a first and second target site for one or more rare-cutting endonucleases, wherein the target sites flank the first and second splice acceptors. The first and second target sites can flank the first and second homology arms. In embodiments, the DNA polynucleotides described herein can be integrated within an intron of the endogenous gene or at an intron-exon junction. The DNA polynucleotides can be integrated within an intron, or at the intron-exon junction of the ATXN3 gene or CACNA1A gene. The DNA polynucleotide can comprise a first and second partial coding sequence encoding the peptide produced by exon 10 of a non-pathogenic ATXN3 gene. The DNA polynucleotide can comprise a first and second partial coding sequence encoding the peptide produced by exon 47 of a non-pathogenic CACNA1A gene. The first and second partial coding sequences encode the same amino acids. In an embodiment, the first and second coding sequences can differ in nucleic acid sequence but encode the same amino acids. The DNA polynucleotides can be harbored on a vector, wherein the vector format is selected from double-stranded linear DNA, double-stranded circular DNA, or a viral vector. The viral vector can be selected from an adenovirus vector, an adeno-associated virus vector, or a lentivirus vector. The DNA polynucleotides described here can be equal to or less than 4.7 kb.

In one embodiment, this document features a method of integrating a transgene into an endogenous gene, the method including administering a transgene, wherein the transgene comprises a left and right transposon end, a first and second splice acceptor sequence, a first and second partial coding sequence, and one bidirectional terminator or a first and second terminator, and administering a transposase targeted to the endogenous gene, where the transgene is integrated in the endogenous gene. The method can include designing the transgene to have the first splice acceptor operably linked to the first partial coding sequence and the second splice acceptor operably linked to the second coding

sequence. The arrangement can also include having the first partial coding sequence operably linked to the first terminator, and the second partial coding sequence operably linked to the second terminator. In an embodiment, the two terminators can be replaced with a single bidirectional terminator. In an embodiment, transgenes with first and second splice acceptors, first and second coding sequences, and first and second terminators can be oriented in a tail-to-tail orientation. The transgenes with a tail-to-tail orientation of sequences can further comprise a left and right transposon end flanking the first and second splice acceptors. In embodiments, the transgenes described herein can be integrated within an intron of the endogenous gene or at an intron-exon junction. The transgenes can be integrated within an intron, or at the intron-exon junction of the ATXN3 gene or CACNA1A gene. The transgene can comprise a first and second partial coding sequence encoding the peptide produced by exon 10 of a non-pathogenic ATXN3 gene and can be targeted to intron 9, or the intron 9 exon 10 junction, of a pathogenic ATXN3 gene. The transgene can comprise a first and second partial coding sequence encoding the peptide produced by exon 47 of a non-pathogenic CACNA1A gene and can be targeted to intron 46, or the intron 46 exon 47 junction, of a pathogenic CACNA1A gene. The transposase can be a CRISPR transposase, where the CRISPR transposase comprises the Cas12k or Cas6 protein. The first and second partial coding sequences encode the same amino acids. In an embodiment, the first and second coding sequences can differ in nucleic acid sequence but encode the same amino acids. The transgene can be harbored on a vector, wherein the vector format is selected from double-stranded linear DNA, double-stranded circular DNA, or a viral vector. The viral vector iscan include an adenovirus vector, an adeno-associated virus vector, or a lentivirus vector. The methods described here can be used with a transgene equal to or less than 4.7 kb. The left end can comprise the sequence shown in SEQ ID NO:41, and the right end can comprise the sequence shown in SEQ ID NO:13.

In another embodiment, this document provides DNA polynucleotides with a first and second splice acceptor sequence, a first and second partial coding sequence, one bidirectional terminator or a first and second terminator, and a left and right transposon end. The DNA polynucleotides can include a design having the first splice acceptor operably linked to the first partial coding sequence and the second splice acceptor

operably linked to the second coding sequence. The arrangement can also include having the first partial coding sequence operably linked to the first terminator, and the second partial coding sequence operably linked to the second terminator. In an embodiment, the two terminators can be replaced with a single bidirectional terminator. In an embodiment, DNA polynucleotides with first and second splice acceptors, first and second coding sequences, and first and second terminators can be oriented in a tail-to-tail orientation. The DNA polynucleotides with a tail-to-tail orientation of sequences can further comprise a left and right transposon end which flank the first and second splice acceptors. In embodiments, the DNA polynucleotides described herein can be integrated within an intron of the endogenous gene or at an intron-exon junction. The DNA polynucleotides can be integrated within an intron, or at the intron-exon junction of the ATXN3 gene or CACNA1A gene. The DNA polynucleotide can comprise a first and second partial coding sequence encoding the peptide produced by exon 10 of a non-pathogenic ATXN3 gene. The DNA polynucleotide can comprise a first and second partial coding sequence encoding the peptide produced by exon 47 of a non-pathogenic CACNA1A gene. The first and second partial coding sequences encode the same amino acids. In an embodiment, the first and second coding sequences can differ in nucleic acid sequence but encode the same amino acids. The DNA polynucleotides can be harbored on a vector, wherein the vector format is selected from double-stranded linear DNA, double-stranded circular DNA, or a viral vector. The viral vector can be selected from an adenovirus vector, an adeno-associated virus vector, or a lentivirus vector. The DNA polynucleotides described here can be equal to or less than 4.7 kb. The left end can comprise the sequence shown in SEQ ID NO:41, and the right end can comprise the sequence shown in SEQ ID NO:13.

In one embodiment, this document features a method of integrating a transgene into an endogenous gene, the method including administering a transgene, wherein the transgene comprises a first and second splice acceptor sequence, a first and second coding sequence, one bidirectional terminator or a first and second terminator, and a first and second homology arm, wherein the transgene is integrated within the endogenous gene. The method can include designing the transgene to have the first splice acceptor

operably linked to the first partial coding sequence and the second splice acceptor operably linked to the second coding sequence. The arrangement can also include having the first partial coding sequence operably linked to the first terminator, and the second partial coding sequence operably linked to the second terminator. In an embodiment, the two terminators can be replaced with a single bidirectional terminator. The homology arms can flank the first and second splice acceptor sequence, the first and second coding sequence, the one bidirectional terminator or the first and second terminator. The coding sequence can encode a full coding sequence or a partial coding sequence. In an embodiment, transgenes with first and second splice acceptors, first and second coding sequences, and first and second terminators can be oriented in a tail-to-tail orientation. The transgenes with a tail-to-tail orientation of sequences can further comprise a first and second target site for one or more rare-cutting endonucleases, wherein the target sites flank the first and second splice acceptors. In another embodiment, the transgenes can comprise a left and right homology arm which flank the first and second splice acceptors. In this embodiment, the transgene can be harbored within an adeno-associated viral vector. In another embodiment, the transgene can further comprise a first and second target site for the one or more rare-cutting endonucleases, wherein the target sites flank the first and second splice acceptors. The first and second target sites can flank the first and second homology arms. In embodiments, the transgenes described herein can be integrated within an intron of the endogenous gene or at an intron-exon junction.

Practice of the methods, as well as preparation and use of the compositions disclosed herein employ, unless otherwise indicated, conventional techniques in molecular biology, biochemistry, chromatin structure and analysis, computational chemistry, cell culture, recombinant DNA and related fields as are within the skill of the art. These techniques are fully explained in the literature. See, for example, Sambrook et al. MOLECULAR CLONING: A LABORATORY MANUAL, Second edition, Cold Spring Harbor Laboratory Press, 1989 and Third edition, 2001; Ausubel et al., CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, John Wiley & Sons, New York, 1987 and periodic updates; the series METHODS IN ENZYMOLOGY, Academic Press, San Diego; Wolffe, CHROMATIN STRUCTURE AND FUNCTION, Third

edition, Academic Press, San Diego, 1998; METHODS IN ENZYMOLOGY, Vol. 304, "Chromatin" (P. M. Wassarman and A. P. Wolffe, eds.), Academic Press, San Diego, 1999; and METHODS IN MOLECULAR BIOLOGY, Vol. 119, "Chromatin Protocols" (P. B. Becker, ed.) Humana Press, Totowa, 1999.

As used herein, the terms "nucleic acid" and "polynucleotide," can be used interchangeably. Nucleic acid and polynucleotide can refer to a deoxyribonucleotide or ribonucleotide polymer, in linear or circular conformation, and in either single- or double-stranded form. These terms are not to be construed as limiting with respect to the length of a polymer. The terms can encompass known analogues of natural nucleotides, as well as nucleotides that are modified in the base, sugar and/or phosphate moieties.

The terms "polypeptide," "peptide" and "protein" can be used interchangeably to refer to amino acid residues covalently linked together. The term also applies to proteins in which one or more amino acids are chemical analogues or modified derivatives of corresponding naturally occurring amino acids.

The terms "operatively linked" or "operably linked" are used interchangeably and refer to a juxtaposition of two or more components (such as sequence elements), in which the components are arranged such that both components function normally and allow the possibility that at least one of the components can mediate a function that is exerted upon at least one of the other components. By way of illustration, a transcriptional regulatory sequence, such as a promoter, is operatively linked to a coding sequence if the transcriptional regulatory sequence controls the level of transcription of the coding sequence in response to the presence or absence of one or more transcriptional regulatory factors. A transcriptional regulatory sequence is generally operatively linked in cis with a coding sequence but need not be directly adjacent to it. For example, an enhancer is a transcriptional regulatory sequence that is operatively linked to a coding sequence, even though they are not contiguous. Further, by way of example, a splice acceptor can be operably linked to a partial coding sequence if the splice acceptor enables delineation of an intron's 3' boundary, and if translation of the resulting mature mRNA results in

incorporation of the peptide sequence encoded by the partial coding sequence into the final protein product.

As used herein, the term "cleavage" refers to the breakage of the covalent backbone of a nucleic acid molecule. Cleavage can be initiated by a variety of methods including, but not limited to, enzymatic or chemical hydrolysis of a phosphodiester bond. Cleavage can refer to both a single-stranded nick and a double-stranded break. A double-stranded break can occur as a result of two distinct single-stranded nicks. Nucleic acid cleavage can result in the production of either blunt ends or staggered ends. In certain embodiments, rare-cutting endonucleases are used for targeted double-stranded or single-stranded DNA cleavage.

An "exogenous" molecule can refer to a small molecule (e.g., sugars, lipids, amino acids, fatty acids, phenolic compounds, alkaloids), or a macromolecule (e.g., protein, nucleic acid, carbohydrate, lipid, glycoprotein, lipoprotein, polysaccharide), or any modified derivative of the above molecules, or any complex comprising one or more of the above molecules, generated or present outside of a cell, or not normally present in a cell. Exogenous molecules can be introduced into cells. Methods for the introduction or "administering" of exogenous molecules into cells can include lipid-mediated transfer, electroporation, direct injection, cell fusion, particle bombardment, calcium phosphate co-precipitation, DEAE-dextran-mediated transfer and viral vector-mediated transfer. As defined herein, "administering" can refer to the delivery, the providing, or the introduction of exogenous molecules into a cell. If a transgene or a rare-cutting endonuclease is administered to a cell, then the transgene or rare-cutting endonuclease is delivered to, provided, or introduced into the cell. The rare-cutting endonuclease can be administered as purified protein, nucleic acid, or a mixture of purified protein and nucleic acid. The nucleic acid (i.e., RNA or DNA), can encode for the rare-cutting endonuclease, or a part of a rare-cutting endonuclease (e.g., a gRNA). The administering can be achieved though methods such as lipid-mediated transfer, electroporation, direct injection, cell fusion, particle bombardment, calcium phosphate co-precipitation, DEAE-dextran-mediated transfer, viral vector-mediated transfer, or any means suitable of

delivering purified protein or nucleic acids, or a mixture of purified protein and nucleic acids, to a cell.

An “endogenous” molecule is a molecule that is present in a particular cell at a particular developmental stage under particular environmental conditions. An endogenous molecule can be a nucleic acid, a chromosome, the genome of a mitochondrion, chloroplast or other organelle, or a naturally occurring episomal nucleic acid. Additional endogenous molecules can include proteins, for example, transcription factors and enzymes.

As used herein, a “gene,” refers to a DNA region encoding that encodes a gene product, including all DNA regions which regulate the production of the gene product. Accordingly, a gene includes, but is not necessarily limited to, promoter sequences, terminators, translational regulatory sequences such as ribosome binding sites and internal ribosome entry sites, enhancers, silencers, insulators, boundary elements, replication origins, matrix attachment sites and locus control regions. As used herein, a “wild type gene” refers to a form of the gene that is present at the highest frequency in a particular population.

An "endogenous gene" refers to a DNA region normally present in a particular cell that encodes a gene product as well as all DNA regions which regulate the production of the gene product.

“Gene expression” refers to the conversion of the information, contained in a gene, into a gene product. A gene product can be the direct transcriptional product of a gene. For example, the gene product can be, but not limited to, mRNA, tRNA, rRNA, antisense RNA, ribozyme, structural RNA, or a protein produced by translation of an mRNA. Gene products also include RNAs which are modified, by processes such as capping, polyadenylation, methylation, and editing, and proteins modified by, for example, methylation, acetylation, phosphorylation, ubiquitination, ADP-ribosylation, myristilation, and glycosylation.

"Encoding" refers to the conversion of the information contained in a nucleic acid, into a product, wherein the product can result from the direct transcriptional product of a nucleic acid sequence. For example, the product can be, but not limited to, mRNA, tRNA, rRNA, antisense RNA, ribozyme, structural RNA, or a protein produced by translation of an mRNA. Gene products also include RNAs which are modified, by processes such as capping, polyadenylation, methylation, and editing, and proteins modified by, for example, methylation, acetylation, phosphorylation, ubiquitination, ADP-ribosylation, myristilation, and glycosylation.

A "target site" or "target sequence" defines a portion of a nucleic acid to which a rare-cutting endonuclease or CRISPR-associated transposase will bind, provided sufficient conditions for binding exist.

As used herein, the term "recombination" refers to a process of exchange of genetic information between two polynucleotides. The term "homologous recombination (HR)" refers to a specialized form of recombination that can take place, for example, during the repair of double-strand breaks. Homologous recombination requires nucleotide sequence homology present on a "donor" molecule. The donor molecule can be used by the cell as a template for repair of a double-strand break. Information within the donor molecule that differs from the genomic sequence at or near the double-strand break can be stably incorporated into the cell's genomic DNA.

The term "integrating" as used herein refers to the process of adding DNA to a target region of DNA. As described herein, integration can be facilitated by several different means, including non-homologous end joining, homologous recombination, or targeted transposition. By way of example, integration of a user-supplied DNA molecule into a target gene can be facilitated by non-homologous end joining. Here, a targeted-double strand break is made within the target gene and a user-supplied DNA molecule is administered. The user-supplied DNA molecule can comprise exposed DNA ends to facilitate capture during repair of the target gene by non-homologous end joining. The exposed ends can be present on the DNA molecule upon administration (i.e., administration of a linear DNA molecule) or created upon administration to the cell (i.e.,

a rare-cutting endonuclease cleaves the user-supplied DNA molecule within the cell to expose the ends). Additionally, the user-supplied DNA molecule can be harbored on a viral vector, including an adeno-associated virus vector. In another example, integration occurs though homologous recombination. Here, the user-supplied DNA can harbor a left and right homology arm. In another example, integration occurs through transposition. Here, the user-supplied DNA harbors a transposon left and right end.

The term "transgene" as used herein refers to a sequence of nucleic acids that can be transferred to an organism or cell. The transgene may comprise a gene or sequence of nucleic acids not normally present in the target organism or cell. Additionally, the transgene may comprise a copy of a gene or sequence of nucleic acids that is normally present in the target organism or cell. A transgene can be an exogenous DNA sequence introduced into the cytoplasm or nucleus of a target cell. In one embodiment, the transgenes described herein contain partial coding sequences, wherein the partial coding sequences encodes a portion of a protein produced by a gene in the host cell.

As used herein, the term "pathogenic" refers to anything that can cause disease. A pathogenic mutation can refer to a modification in a gene which causes disease. A pathogenic gene refers to a gene comprising a modification which causes disease. By means of example, a pathogenic ATXN3 gene in patients with spinocerebellar ataxia 3 refers to an ATXN3 gene with an expanded CAG trinucleotide repeat, wherein the expanded CAG trinucleotide repeat causes the disease.

As used herein, the term "tail-to-tail" refers to an orientation of two units in opposite and reverse directions. The two units can be two sequences on a single nucleic acid molecule, where the 3' end of each sequence are placed adjacent to each other. For example, a first nucleic acid having the elements, in a 5' to 3' direction, [splice acceptor 1] – [partial coding sequence 1] – [terminator 1] and a second nucleic acid having the elements [splice acceptor 2] – [partial coding sequence 2] – [terminator 2] can be placed in tail-to-tail orientation resulting in [splice acceptor 1] – [partial coding sequence 1] – [terminator 1] – [terminator 2 RC] – [partial coding sequence 2 RC] - [splice acceptor 2 RC], where RC refers to reverse complement.

The term "intron-exon junction" refers to a specific location within a gene. The specific location is between the last nucleotide in an intron and the first nucleotide of the following exon. When integrating a transgene described herein, the transgene can be integrated within the "intron-exon junction." If the transgene comprises cargo, the cargo will be integrated immediately following the last nucleotide in the intron. In some cases, integrating a transgene within the intron-exon junction can result in removal of sequence within the exon (e.g., integration via HR and replacement of sequence within the exon with the cargo within the transgene).

The term "homologous" as used herein refers to a sequence of nucleic acids or amino acids having similarity to a second sequence of nucleic acids or amino acids. In some embodiments, the homologous sequences can have at least 80% sequence identity (e.g., 81%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% sequence identity) to one another.

The term "partial coding sequence" as used herein refers to a sequence of nucleic acids that encodes a partial protein. The partial coding sequence can encode a protein that comprises one or less amino acids as compared to the wild type protein or functional protein. The partial coding sequence can encode a partial protein with homology to the wild type protein or functional protein. The term "partial coding sequence" when referring to ATXN3 refers to a sequence of nucleic acids that encodes a partial ATXN3 protein. The partial ATXN3 protein has one or less amino acids compared to a wild type ATXN3 protein. If modifying the 3' end of the gene, the one or less amino acids can be from the N-terminus end of the protein. If the ATXN3 gene has 11 exons, then the partial coding sequence can comprise sequence encoding the peptide produced by exons 2-11, or 3-11 or 4-11, or 5-11, or 6-11, or 7-11, or 8-11, or 9-11, or 10-11, or 11.

The methods and compositions described in this document can use transgenes having a cargo sequence. The term "cargo" can refer to elements such as the complete or partial coding sequence of a gene, a partial sequence of a gene harboring single-nucleotide polymorphisms relative to the WT or altered target, a splice acceptor, a terminator, a transcriptional regulatory element, purification tags (e.g., glutathione-S-transferase, poly(His), maltose binding protein, Strep-tag, Myc-tag, AviTag, HA-tag, or

chitin binding protein) or reporter genes (e.g., GFP, RFP, lacZ, cat, luciferase, puro, neomycin). As defined herein, "cargo" can refer to the sequence within a transgene that is integrated at a target site. For example, "cargo" can refer to the sequence on a transgene between two homology arms, two rare-cutting endonuclease target sites, or a left and right transposon end.

The term "homology sequence" refers to a sequence of nucleic acids that comprises homology to a second nucleic acid. Homology sequence, for example, can be present on a donor molecule as an "arm of homology" or "homology arm." A homology arm can be a sequence of nucleic acids within a donor molecule that facilitates homologous recombination with the second nucleic acid. As defined herein, a homology arm can also be referred to as an "arm". In a donor molecule with two homology arms, the homology arms can be referred to as "arm 1" and "arm 2." In one aspect, a cargo sequence can be flanked with first and second homology arm.

The term "bidirectional terminator" refers to a terminator that can terminate RNA polymerase transcription in either the sense or antisense direction. In contrast to two unidirectional terminators in tail-to-tail orientation, a bidirectional terminator can comprise a non-chimeric sequence of DNA. Examples of bidirectional terminators include the ARO4, TRP1, TRP4, ADH1, CYC1, GAL1, GAL7, and GAL10 terminator.

A 5' or 3' end of a nucleic acid molecule references the directionality and chemical orientation of the nucleic acid. As defined herein, the "5' end of a gene" can comprise the exon with the start codon, but not the exon with the stop codon. As defined herein, the "3' end of a gene" can comprise the exon with the stop codon, but not the exon with the start codon.

The term "ATXN3" gene refers to a gene that encodes the enzyme ataxin-3. A representative sequence of the ATXN3 gene can be found with NCBI Reference Sequence: NG_008198.2 and corresponding SEQ ID NO:42. The exon and intron boundaries can be defined with the sequence provided in SEQ ID NO:42. Specifically, exon 1 includes the sequence from 1 to 54. Exon 2 includes the sequence from 9745 to 9909. Exon 3 includes the sequence from 10446 to 10490. Exon 4 includes the sequence

from 12752 to 12837. Exon 5 includes the sequence from 13265 to 13331. Exon 6 includes the sequence from 17766 to 17853. Exon 7 includes the sequence from 23325 to 23457. Exon 8 includes the sequence from 24117 to 24283. Exon 9 includes the sequence from 25522 to 25618. Exon 10 includes the sequence from 35530 to 35648. Exon 11 includes the sequence from 42169 to 48031. Intron 1 includes the sequence from 55 to 9744. Intron 2 includes the sequence from 9910 to 10445. Intron 3 includes the sequence from 10491 to 12751. Intron 4 includes the sequence from 12838 to 13264. Intron 5 includes the sequence from 13332 to 17765. Intron 6 includes the sequence from 17854 to 23324. Intron 7 includes the sequence from 23458 to 24116. Intron 8 includes the sequence from 24284 to 25521. Intron 9 includes the sequence from 25619 to 35529. Intron 10 includes the sequence from 35649 to 42168.

The term "CACNA1A" gene refers to a gene that encodes the calcium voltage-gated channel subunit alpha1A protein. A representative sequence of the CACNA1A gene can be found with NCBI Reference Sequence: NG_011569.1 and corresponding SEQ ID NO:43. The exon and intron boundaries can be defined with the sequence provided in SEQ ID NO:43. Specifically, exon 1 includes the sequence from 1 to 529. Exon 2 includes the sequence from 51249 to 51354. Exon 3 includes the sequence from 53446 to 53585. Exon 4 includes the sequence from 134682 to 134773. Exon 5 includes the sequence from 140992 to 141144. Exon 6 includes the sequence from 146662 to 146855. Exon 7 includes the sequence from 170552 to 170655. Exon 8 includes the sequence from 171968 to 172083. Exon 9 includes the sequence from 173536 to 173592. Exon 10 includes the sequence from 176125 to 176217. Exon 11 includes the sequence from 189140 to 189349. Exon 12 includes the sequence from 193680 to 193792. Exon 13 includes the sequence from 197933 to 198045. Exon 14 includes the sequence from 198210 to 198341. Exon 15 includes the sequence from 198607 to 198679. Exon 16 includes the sequence from 202577 to 202694. Exon 17 includes the sequence from 202848 to 202915. Exon 18 includes the sequence from 205805 to 205911. Exon 19 includes the sequence from 207108 to 207917. Exon 20 includes the sequence from 219495 to 219958. Exon 21 includes the sequence from 221255 to 221393. Exon 22 includes the sequence from 223065 to 223194. Exon 23 includes the sequence from

229333 to 229392. Exon 24 includes the sequence from 230505 to 230611. Exon 25 includes the sequence from 243628 to 243727. Exon 26 includes the sequence from 244851 to 245011. Exon 27 includes the sequence from 246760 to 246897. Exon 28 includes the sequence from 248910 to 249111. Exon 29 includes the sequence from 251202 to 251366. Exon 30 includes the sequence from 253360 to 253470. Exon 31 includes the sequence from 261196 to 261279. Exon 32 includes the sequence from 270731 to 270847. Exon 33 includes the sequence from 271187 to 271252. Exon 34 includes the sequence from 271425 to 271540. Exon 35 includes the sequence from 274601 to 274751. Exon 36 includes the sequence from 276252 to 276379. Exon 37 includes the sequence from 277666 to 277762. Exon 38 includes the sequence from 281689 to 281794. Exon 39 includes the sequence from 291853 to 291960. Exon 40 includes the sequence from 292128 to 292228. Exon 41 includes the sequence from 293721 to 293830. Exon 42 includes the sequence from 293939 to 294077. Exon 43 includes the sequence from 294245 to 294358. Exon 44 includes the sequence from 295809 to 295844. Exon 45 includes the sequence from 296963 to 297149. Exon 46 includes the sequence from 297452 to 297705. Exon 47 includes the sequence from 298413 to 300019. Intron 1 includes the sequence from 530 to 51248. Intron 2 includes the sequence from 51355 to 53445. Intron 3 includes the sequence from 53586 to 134681. Intron 4 includes the sequence from 134774 to 140991. Intron 5 includes the sequence from 141145 to 146661. Intron 6 includes the sequence from 146856 to 170551. Intron 7 includes the sequence from 170656 to 171967. Intron 8 includes the sequence from 172084 to 173535. Intron 9 includes the sequence from 173593 to 176124. Intron 10 includes the sequence from 176218 to 189139. Intron 11 includes the sequence from 189350 to 193679. Intron 12 includes the sequence from 193793 to 197932. Intron 13 includes the sequence from 198046 to 198209. Intron 14 includes the sequence from 198342 to 198606. Intron 15 includes the sequence from 198680 to 202576. Intron 16 includes the sequence from 202695 to 202847. Intron 17 includes the sequence from 202916 to 205804. Intron 18 includes the sequence from 205912 to 207107. Intron 19 includes the sequence from 207918 to 219494. Intron 20 includes the sequence from 219959 to 221254. Intron 21 includes the sequence from 221394 to 223064. Intron 22 includes the sequence from 223195 to 229332. Intron 23 includes the sequence from

229393 to 230504. Intron 24 includes the sequence from 230612 to 243627. Intron 25 includes the sequence from 243728 to 244850. Intron 26 includes the sequence from 245012 to 246759. Intron 27 includes the sequence from 246898 to 248909. Intron 28 includes the sequence from 249112 to 251201. Intron 29 includes the sequence from 251367 to 253359. Intron 30 includes the sequence from 253471 to 261195. Intron 31 includes the sequence from 261280 to 270730. Intron 32 includes the sequence from 270848 to 271186. Intron 33 includes the sequence from 271253 to 271424. Intron 34 includes the sequence from 271541 to 274600. Intron 35 includes the sequence from 274752 to 276251. Intron 36 includes the sequence from 276380 to 277665. Intron 37 includes the sequence from 277763 to 281688. Intron 38 includes the sequence from 281795 to 291852. Intron 39 includes the sequence from 291961 to 292127. Intron 40 includes the sequence from 292229 to 293720. Intron 41 includes the sequence from 293831 to 293938. Intron 42 includes the sequence from 294078 to 294244. Intron 43 includes the sequence from 294359 to 295808. Intron 44 includes the sequence from 295845 to 296962. Intron 45 includes the sequence from 297150 to 297451. Intron 46 includes the sequence from 297706 to 298412.

The percent sequence identity between a particular nucleic acid or amino acid sequence and a sequence referenced by a particular sequence identification number is determined as follows. First, a nucleic acid or amino acid sequence is compared to the sequence set forth in a particular sequence identification number using the BLAST 2 Sequences (Bl2seq) program from the stand-alone version of BLASTZ containing BLASTN version 2.0.14 and BLASTP version 2.0.14. This stand-alone version of BLASTZ can be obtained online at fr.com/blast or at ncbi.nlm.nih.gov. Instructions explaining how to use the Bl2seq program can be found in the readme file accompanying BLASTZ. Bl2seq performs a comparison between two sequences using either the BLASTN or BLASTP algorithm. BLASTN is used to compare nucleic acid sequences, while BLASTP is used to compare amino acid sequences. To compare two nucleic acid sequences, the options are set as follows: -i is set to a file containing the first nucleic acid sequence to be compared (e.g., C:\seq1.txt); -j is set to a file containing the second nucleic acid sequence to be compared (e.g., C:\seq2.txt); -p is set to blastn; -o is set to

any desired file name (e.g., C:\output.txt); -q is set to -1; -r is set to 2; and all other options are left at their default setting. For example, the following command can be used to generate an output file containing a comparison between two sequences: C:\Bl2seq -i c:\seq1.txt -j c:\seq2.txt -p blastn -o c:\output.txt -q -1 -r 2. To compare two amino acid sequences, the options of Bl2seq are set as follows: -i is set to a file containing the first amino acid sequence to be compared (e.g., C:\seq1.txt); -j is set to a file containing the second amino acid sequence to be compared (e.g., C:\seq2.txt); -p is set to blastp; -o is set to any desired file name (e.g., C:\output.txt); and all other options are left at their default setting. For example, the following command can be used to generate an output file containing a comparison between two amino acid sequences: C:\Bl2seq -i c:\seq1.txt -j c:\seq2.txt -p blastp -o c:\output.txt. If the two compared sequences share homology, then the designated output file will present those regions of homology as aligned sequences. If the two compared sequences do not share homology, then the designated output file will not present aligned sequences.

Once aligned, the number of matches is determined by counting the number of positions where an identical nucleotide or amino acid residue is presented in both sequences. The percent sequence identity is determined by dividing the number of matches either by the length of the sequence set forth in the identified sequence, or by an articulated length (e.g., 100 consecutive nucleotides or amino acid residues from a sequence set forth in an identified sequence), followed by multiplying the resulting value by 100. The percent sequence identity value is rounded to the nearest tenth.

In one embodiment, this document features methods for modifying the 3' end of endogenous genes, where endogenous genes have at least one intron between two coding exons. The intron can be any intron which is removed from precursor messenger RNA by normal messenger RNA processing machinery. The intron can be between 20 bp and >500 kb and comprise elements including a splice donor site, branch sequence, and acceptor site. The transgenes disclosed herein for the modification of the 3' end of endogenous genes can comprise multiple functional elements, including target sites for rare-cutting endonucleases, homology arms, splice acceptor sequences, coding sequences, and transcription terminators (FIG. 1).

In one embodiment, the transgene comprises two target sites for one or more rare-cutting endonucleases. The target sites can be a suitable sequence and length for cleavage by a rare-cutting endonuclease. The target site can be amenable to cleavage by CRISPR systems, TAL effector nucleases, zinc-finger nucleases or meganucleases, or a combination of CRISPR systems, TALE nucleases, zinc finger nucleases or meganucleases, or any other site-specific nuclease. The target sites can be positioned such that cleavage by the rare-cutting endonuclease results in liberation of a transgene from a vector. The vector can include viral vectors (e.g., adeno-associated vectors) or non-viral vectors (e.g., plasmids, minicircle vectors). If the transgene comprises two target sites, the target sites can be the same sequence (i.e., targeted by the same rare-cutting endonuclease) or they can be different sequences (i.e., targeted by two or more different rare-cutting endonucleases).

In one embodiment, the transgene comprises a first and second target site for one or more rare-cutting endonucleases along with a first and second homology arm. The first and second homology arms can include sequence that is homologous to a genomic sequence at or near the desired site of integration. The homology arms can be a suitable length for participating in homologous recombination with sequence at or near the desired site of integration. The length of each homology arm can be between 20 nt and 10,000 nt (e.g., 20 nt, 30 nt, 40 nt, 50 nt, 100 nt, 200 nt, 300 nt, 400 nt, 500 nt, 600 nt, 700 nt, 800 nt, 900 nt, 1,000 nt, 2,000 nt, 3,000 nt, 4,000 nt, 5,000 nt, 6,000 nt, 7,000 nt, 8,000 nt, 9,000 nt, 10,000 nt). In one embodiment, a homology arms can comprise functional elements, including a target site for a rare-cutting endonuclease and/or a splice acceptor sequence. In one embodiment, a first homology arm (e.g., a left homology arm) can comprise sequence homologous to the intron being targeted, which includes the splice acceptor site of the intron being targeted. In another embodiment, a second homology arm can comprise sequence homologous to genomic sequence downstream of the intron being targeted (e.g., exon sequence, 3' UTR sequence). However, the second homology arm must not possess splice acceptor functions in the reverse complement direction. To determine if a sequence comprises splice acceptor functions, several steps can be taken, including in silico analysis and experimental tests. To determine if there is

potential for splice acceptor functions, the sequence desired for second homology arm can be searched for consensus branch sequences (e.g., YTRAC) and splice acceptor sites (e.g., Y-rich NCAGG). If branch or splice acceptor sequences are present, single nucleotide polymorphisms can be introduced to destroy function, or a different but adjacent sequence not comprising such sequences can be selected. Preferably, the window of sequence that can be used for a second homology arm extends from 1 bp to 10kb downstream of the intron being targeted for integration. To experimentally determine if the second homology possesses splice acceptor function, a synthetic construct comprising the second homology arm within an intron within a reporter gene can be constructed. The construct can then be administered to an appropriate cell type and monitored for splicing function.

In one embodiment, the transgene comprises two splice acceptor sequences, referred to herein as the first and second splice acceptor sequence. The first and second splice acceptor sequences are positioned within the transgene in opposite directions (i.e., in tail-to-tail orientations) and flanking internal sequences (i.e., coding sequences and terminators). When the transgene is integrated into an intron in forward or reverse directions, the splice acceptor sequences facilitate the removal of the adjacent/upstream intron sequence during mRNA processing. The first and second splice acceptor sequences can be the same sequences or different sequences. One or both splice acceptor sequences can be the splice acceptor sequence of the intron where the transgene is to be integrated. One or both splice acceptor sequences can be a synthetic splice acceptor sequence or a splice acceptor sequence from an intron from a different gene.

In one embodiment, the transgene comprises a first and second coding sequence operably linked to the first and second splice acceptor sequences. The first and second coding sequences are positioned within the transgene in opposite directions (i.e., in tail-to-tail orientations). When the transgene is integrated into an endogenous gene in forward or reverse directions, the first or second coding sequence is transcribed into mRNA by the endogenous gene's promoter. The coding sequences can be designed to correct defective coding sequences, introduce mutations, or introduce novel peptide sequences. The first and second coding sequence can be the same nucleic acid sequence and code for

the same protein. Alternatively, the first and second coding sequence can be different nucleic acid sequences and code for the same protein (i.e., using the degeneracy of codons). The coding sequence can encode purification tags (e.g., glutathione-S-transferase, poly(His), maltose binding protein, Strep-tag, Myc-tag, AviTag, HA-tag, or chitin binding protein) or reporter proteins (e.g., GFP, RFP, lacZ, cat, luciferase, puro, neomycin). In one embodiment, the transgene comprises a first and second partial coding sequence operably linked to a first and second splice acceptor sequence, and the transgene does not comprise a promoter.

In one embodiment, the transgene can comprise a bidirectional terminator, or a first and second terminator, operably linked to a first and second coding sequence. The bidirectional terminator, or the first and second terminators are positioned within the transgene in opposite directions (i.e., in tail-to-tail orientations). When the transgene is integrated into an endogenous gene in forward or reverse directions, the bidirectional terminator, or first and second terminators, terminate transcription from the endogenous gene's promoter. The first and second terminators can be the same terminators or different terminators.

In one embodiment, this document provides a transgene comprising a first and second rare-cutting endonuclease target site, a first and second splice acceptor sequence, a first and second coding sequence, and one bidirectional terminator or a first and second terminator. The transgene can be integrated in endogenous genes via non-homology dependent methods, including non-homologous end joining and alternative non-homologous end joining or by microhomology-mediated end joining. In one aspect, the transgene is integrated into an intron within the endogenous gene (FIG. 2).

In another embodiment, this document provides a transgene comprising a first and second homology arm, a first and second rare-cutting endonuclease target site, a first and second splice acceptor sequence, a first and second coding sequence, and one bidirectional terminator or a first and second terminator. The transgene can be integrated in endogenous genes via both homology dependent methods (e.g., synthesis dependent strand annealing and microhomology-mediated end joining) and non-homology

dependent methods (e.g., non-homologous end joining and alternative non-homologous end joining). In one aspect, the transgene is integrated into an intron within the endogenous gene (FIG. 3). In another aspect, the transgene is integrated at the end of the intron or the starting of the downstream exon (FIG. 3).

In another embodiment, this document provides a transgene comprising a first and second homology arm, a first and second coding sequence, a first and second splice acceptor sequence, and one bidirectional terminator or a first and second terminator (FIG. 1). In another embodiment, this document provides a transgene comprising, a first and second coding sequence, a first and second splice acceptor sequence, and one bidirectional terminator or a first and second terminator.

In another embodiment, this document provides a transgene comprising a first and second homology arm, a first and second coding sequence, a first and second splice acceptor sequence, one bidirectional terminator or a first and second terminator, and a first and second additional sequence (FIG. 1). In certain embodiments, the additional sequence can be any additional sequence that is present on the transgene at the 5' and 3' ends, however, the additional sequence should not comprise any element that functions as a splice acceptor. The additional sequence can be, for example, inverted terminal repeats of a virus genome. The additional sequence can be present on a transgene having a linear format. The linear format permits integration by NHEJ. For example, a transgene harbored in an adeno-associated virus vector, wherein the additional sequence is the inverted terminal repeats, can be directly integrated by NHEJ at a target site after cleavage by a rare-cutting endonuclease (i.e., no processing of the transgene is required). In another example, the additional sequence is a left and right transposon end.

In another embodiment, this document provides transgenes within viral vectors, including adeno-associated viruses and adenoviruses, where the transgene comprises a first and second splice acceptor sequence, a first and second coding sequence, and one bidirectional terminator or a first and second terminator. Due to the inverted terminal repeats of the viral vectors, the transgenes also comprise a first and second additional sequence.

In another embodiment, this document provides transgenes within viral vectors, including adeno-associated viruses and adenoviruses, where the transgene comprises a first and second homology arm, a first and second splice acceptor sequence, a first and second coding sequence, and one bidirectional terminator or a first and second terminator. Due to the inverted terminal repeats of the viral vectors, the transgenes also comprise a first and second additional sequence.

In some embodiments, the transgenes provided herein can be integrated with transposases. The transposases can include CRISPR transposases (Strecker et al., *Science* 10.1126/science.aax9181, 2019; Klompe et al., *Nature*, 10.1038/s41586-019-1323-z, 2019). The transposases can be used in combination with a transgene comprising, a first and second splice acceptor sequence, a first and second coding sequence, one bidirectional terminator or a first and second terminator (FIG. 1), and a transposon left end and right end. The CRISPR transposases can include the TypeV-U5, C2C5 CRISPR protein, Cas12k, along with proteins tnsB, tnsC, and tniQ. In some embodiments, the Cas12k can be from *Scytonema hofmanni* (SEQ ID NO:30) or *Anabaena cylindrica* (SEQ ID NO:31). In one embodiment, the transgenes described herein comprising a left (SEQ ID NO:32) and right transposon end (SEQ ID NO:33) can be delivered to cells along with ShCas12k, tnsB, tnsC, TniQ and a gRNA (SEQ ID NO:14). Alternatively, the CRISPR transposase can include the Cas6 protein, along with helper proteins including Cas7, Cas8 and TniQ. In one embodiment, the transgenes described herein comprising a left (SEQ ID NO:41) and right transposon end (SEQ ID NO:13) can be delivered to eukaryotic cells along with Cas6 (SEQ ID NO:37), Cas7 (SEQ ID NO:37), Cas8 (SEQ ID NO:37), TniQ (SEQ ID NO:37), TnsA (SEQ ID NO:37), TnsB (SEQ ID NO:37), TnsC (SEQ ID NO:37) and a gRNA (SEQ ID NO:12). The proteins can be administered to cells directly as purified protein or encoded on RNA or DNA. If encoded on RNA or DNA, the sequence can be codon optimized for expression in eukaryotic cells. The gRNA (SEQ ID NO:12) can be placed downstream of an RNA polIII promoter and terminated with a poly(T) terminator.

In some embodiments, the transgenes described herein can have a combination of elements including splice acceptors, partial coding sequences, terminators, homology

arms, left and right transposase ends, and sites for cleavage by rare-cutting endonucleases. In one embodiment, the combination can be, from 5' to 3', [splice acceptor 1] – [partial coding sequence 1] – [terminator 1] – [terminator 2 RC] – [partial coding sequence 2 RC] – [splice acceptor 2 RC], where RC stands for reverse complement. This combination can be harbored on a linear DNA molecule or AAV molecule and can be integrated by NHEJ through a targeted break in the target gene. In another embodiment, the combination can be, from 5' to 3', [rare-cutting endonuclease cleavage site 1] – [splice acceptor 1] – [partial coding sequence 1] – [terminator 1] – [terminator 2 RC] – [partial coding sequence 2 RC] – [splice acceptor 2 RC] – [rare-cutting endonuclease cleavage site 1]. In another embodiment, the combination can be, from 5' to 3', [rare-cutting endonuclease cleavage site 1] – [homology arm 1] – [splice acceptor 1] – [partial coding sequence 1] – [terminator 1] – [terminator 2 RC] – [partial coding sequence 2 RC] – [splice acceptor 2 RC] – [homology arm 2] – [rare-cutting endonuclease cleavage site 2]. In this combination one or more rare-cutting endonucleases can be used to facilitate HR and NHEJ. For example, a single rare-cutting nuclease can cleave the target gene (i.e., a desired intron) and the cleavage sites flanking the homology arms can be designed to be the same target sequence within the intron. In another embodiment, the combination can be, from 5' to 3', [homology arm 1 + rare-cutting endonuclease cleavage site 1] – [splice acceptor 1] – [partial coding sequence 1] – [terminator 1] – [terminator 2 RC] – [partial coding sequence 2 RC] – [splice acceptor 2 RC] – [homology arm 2] – [rare-cutting endonuclease cleavage site 1]. In this combination, one or more rare-cutting endonucleases can facilitate HR and NHEJ. For example, a single-rare cutting nuclease can cleave within homology arm 1, downstream of homology arm 2, and at the genomic target site (i.e., at the site with homology to the sequence in the homology arm 1). In another embodiment, the combination can be from 5' to 3', [left end for a transposase] – [splice acceptor 1] – [partial coding sequence 1] – [terminator 1] – [terminator 2 RC] – [partial coding sequence 2 RC] – [splice acceptor 2 RC] – [right end for a transposase]. In all embodiments, the splice acceptor 1 and splice acceptor 2 can be the same or different sequences; the partial coding sequence 1 and partial coding sequence 2 can be the same or different sequences; the terminator 1 and terminator 2 can be the same or different sequences.

In embodiments, a transgene comprising the structure [rare-cutting endonuclease cleavage site 1] – [homology arm 1] – [splice acceptor 1] – [partial coding sequence 1] – [terminator 1] – [terminator 2 RC] – [partial coding sequence 2 RC] – [splice acceptor 2 RC] – [homology arm 2] – [rare-cutting endonuclease cleavage site 2] can be integrated into the DNA through delivery of one or more rare-cutting endonucleases. If one rare-cutting endonuclease is delivered, the rare-cutting endonuclease can liberate the transgene by cleavage at the rare-cutting endonuclease cleavage site 1 and 2. Further, the same rare-cutting endonuclease can create a break within the target gene, simulating insertion through HR or NHEJ.

In other embodiments, a transgene comprising the structure [homology arm 1 + rare-cutting endonuclease cleavage site 1] – [splice acceptor 1] – [partial coding sequence 1] – [terminator 1] – [terminator 2 RC] – [partial coding sequence 2 RC] – [splice acceptor 2 RC] – [homology arm 2] – [rare-cutting endonuclease cleavage site 1] can be integrated into the DNA thorough delivery of one or more rare-cutting endonucleases. If one rare-cutting endonuclease is delivered, the rare-cutting endonuclease can liberate the transgene by cleavage at the rare-cutting endonuclease cleavage site 1 and 2. Further, the same rare-cutting endonuclease can create a break within the target gene, simulating insertion through HR or NHEJ. Integration by HR can occur when cleavage is upstream of the site of integration (i.e., within a homology arm).

In embodiments, the location for integration of transgenes can be an intron or an intron-exon junction. When targeting an intron, the partial coding sequence can comprise sequence encoding the peptide produced by the following exons within the endogenous gene. For example, if the transgene is designed to be integrated in intron 9 of an endogenous gene with 11 exons, then the partial coding sequence can comprise sequence encoding the peptide produced by exons 10 and 11 of the endogenous gene. When targeting an intron-exon junction, the transgene can be designed to comprise homology arms with sequence homologous to the 3' of said intron.

In some embodiments, the partial coding sequences can be full coding sequences. The full coding sequence can encode an endogenous gene (e.g., Factor VIII, Factor IX, or

INS), or reporter genes (e.g., RFP, GFP, cat, lacZ, luciferase). The full coding sequences can be operably linked to splice acceptors and terminators and placed in a transgene in a tail-to-tail orientation.

The methods and compositions provided herein can be used within to modify endogenous genes within cells. The endogenous genes can include, fibrinogen, prothrombin, tissue factor, Factor V, Factor VII, Factor VIII, Factor IX, Factor X, Factor XI, Factor XII (Hageman factor), Factor XIII (fibrin-stabilizing factor), von Willebrand factor, prekallikrein, high molecular weight kininogen (Fitzgerald factor), fibronectin, antithrombin III, heparin cofactor II, protein C, protein S, protein Z, protein Z-related protease inhibitor, plasminogen, alpha 2-antiplasmin, tissue plasminogen activator, urokinase, plasminogen activator inhibitor-1, plasminogen activator inhibitor-2, glucocerebrosidase (GBA), α-galactosidase A (GLA), iduronate sulfatase (IDS), iduronidase (IDUA), acid sphingomyelinase (SMPD1), MMAA, MMAB, MMACHC, MMADHC (C2orf25), MTRR, LMBRD1, MTR, propionyl-CoA carboxylase (PCC) (PCCA and/or PCCB subunits), a glucose-6-phosphate transporter (G6PT) protein or glucose-6-phosphatase (G6Pase), an LDL receptor (LDLR), ApoB, LDLRAP-1, a PCSK9, a mitochondrial protein such as NAGS (N-acetylglutamate synthetase), CPS1 (carbamoyl phosphate synthetase I), and OTC (ornithine transcarbamylase), ASS (argininosuccinic acid synthetase), ASL (argininosuccinase acid lyase) and/or ARG1 (arginase), and/or a solute carrier family 25 (SLC25A13, an aspartate/glutamate carrier) protein, a UGT1A1 or UDP glucuronsyltransferase polypeptide A1, a fumarylacetoacetate hydrolyase (FAH), an alanine-glyoxylate aminotransferase (AGXT) protein, a glyoxylate reductase/hydroxypyruvate reductase (GRHPR) protein, a transthyretin gene (TTR) protein, an ATP7B protein, a phenylalanine hydroxylase (PAH) protein, an USH2A protein, an ATXN protein, and a lipoprotein lyase (LPL) protein.

The transgene may include sequence for modifying the sequence encoding a polypeptide that is lacking or non-functional or having a gain-of-function mutation in the subject having a genetic disease, including but not limited to the following genetic diseases: achondroplasia, achromatopsia, acid maltase deficiency, adenosine deaminase deficiency, adrenoleukodystrophy, aicardi syndrome, alpha-1 antitrypsin deficiency,

alpha-thalassemia, androgen insensitivity syndrome, pert syndrome, arrhythmogenic right ventricular dysplasia, ataxia telangictasia, barth syndrome, beta-thalassemia, blue rubber bleb nevus syndrome, canavan disease, chronic granulomatous diseases (CGD), cri du chat syndrome, cystic fibrosis, dercum's disease, ectodermal dysplasia, fanconi anemia, fibrodysplasia ossificans progressive, fragile X syndrome, galactosemis, Gaucher's disease, generalized gangliosidoses (e.g., GM1), hemochromatosis, the hemoglobin C mutation in the 6th codon of beta-globin (HbC), hemophilia, Huntington's disease, Hurler Syndrome, hypophosphatasia, Klinefleter syndrome, Krabbes Disease, Langer-Giedion Syndrome, leukocyte adhesion deficiency, leukodystrophy, long QT syndrome, Marfan syndrome, Moebius syndrome, mucopolysaccharidosis (MPS), nail patella syndrome, nephrogenic diabetes insipdius, neurofibromatosis, Neimann-Pick disease, osteogenesis imperfecta, porphyria, Prader-Willi syndrome, progeria, Proteus syndrome, retinoblastoma, Rett syndrome, Rubinstein-Taybi syndrome, Sanfilippo syndrome, severe combined immunodeficiency (SCID), Shwachman syndrome, sickle cell disease (sickle cell anemia), Smith-Magenis syndrome, Stickler syndrome, Tay-Sachs disease, Thrombocytopenia Absent Radius (TAR) syndrome, Treacher Collins syndrome, trisomy, tuberous sclerosis, Turner's syndrome, urea cycle disorder, von Hippel-Landau disease, Waardenburg syndrome, Williams syndrome, Wilson's disease, Wiskott-Aldrich syndrome, X-linked lymphoproliferative syndrome, lysosomal storage diseases (e.g., Gaucher's disease, GM1, Fabry disease and Tay-Sachs disease), mucopolysaccahidosis (e.g. Hunter's disease, Hurler's disease), hemoglobinopathies (e.g., sickle cell diseases, HbC, α-thalassemia, β-thalassemia) and hemophilias.

Additional diseases that can be treated by targeted integration include von Willebrand disease, usher syndrome, polycystic kidney disease, spinocerebellar ataxia type 3, and spinocerebellar ataxia type 6.

In one embodiment, the genomic modification is the insertion of a transgene in the endogenous CACNA1A genomic sequence. The transgene can include a synthetic and partial coding sequence for the CACNA1A protein. The partial coding sequence can be homologous to coding sequence within a wild type CACNA1A gene, or a functional variant of the wild type CACNA1A gene, or a mutant of the wild type CACNA1A gene.

In one embodiment, the transgene encoding the partial CACNA1A protein is inserted into intron 46 or the beginning of exon 47.

In another embodiment, the genomic modification is the insertion of a transgene in the endogenous ATXN3 genomic sequence. The transgene can include a synthetic and partial coding sequence for the ATXN3 protein. The partial coding sequence can be homologous to coding sequence within a wild type ATXN3 gene, or a functional variant of the wild type ATXN3 gene, or a mutant of the wild type ATXN3 gene. In one embodiment, the transgene encoding the partial ATXN3 protein is inserted into intron 9 or the beginning of exon 10.

In one embodiment, the methods and compositions described herein can be used to modify the 3' end of an endogenous gene, thereby resulting in modification of the C-terminus of the protein encoded by the endogenous gene. The modification of the 3' end of the endogenous gene's coding sequence can include the replacement of the final coding exon (i.e., the exon comprising the stop codon), up to an exon that is between the exon with the start coding and the final exon. As defined herein "replacement" refers to the insertion of DNA in a gene, wherein the inserted DNA provides the information for producing the mRNA and protein of 1 or more exons. Replacement can occur by integrating a transgene into the endogenous gene, wherein the transgene comprises one or more coding sequences operably linked to a splice acceptor. The insertion may or may not result in the deletion of sequence within the endogenous gene (e.g., deletion of introns and exons). For example, if a gene comprises 72 exons, and the start codon is within exon 1, the modification can include replacement of exons 2-72, 3-72, 4-72, 5-72, 6-72, 7-72, 8-72, 9-72, 10-72, 11-72, 12-72, 13-72, 14-72, 15-72, 16-72, 17-72, 18-72, 19-72, 20-72, 21-72, 22-72, or 23-72, or 24-72, or 25-72, or 26-72, or 27-72, or 28-72, or 29-72, or 30-72, or 31-72, or 32-72, or 33-72, or 34-72, or 35-72, or 36-72, or 37-72, or 38-72, or 39-72, or 40-72, or 41-72, or 42-72, or 43-72, or 44-72, or 45-72, or 46-72, or 47-72, or 48-72, or 49-72, or 50-72, or 51-72, or 52-72, or 53-72, or 54-72, or 55-72, or 56-72, or 57-72, or 58-72, or 59-72, or 60-72, or 61-72, or 62-72, or 63-72, or 64-72, or 65-72, or 66-72, or 67-72, or 68-72, or 69-72, or 70-72, or 71-72 or 72. In one embodiment, the endogenous gene's exons can be replaced by integrating a transgene

into the endogenous gene, wherein the transgene comprises a first and second partial coding sequence, wherein the first and second partial coding sequence encodes a peptide produced by the endogenous genes exons. For example, the transgene's first and second coding sequence can encode a peptide that is produced by the endogenous gene's exons 2-72, 3-72, 4-72, 5-72, 6-72, 7-72, 8-72, 9-72, 10-72, 11-72, 12-72, 13-72, 14-72, 15-72, 16-72, 17-72, 18-72, 19-72, 20-72, 21-72, 22-72, or 23-72, or 24-72, or 25-72, or 26-72, or 27-72, or 28-72, or 29-72, or 30-72, or 31-72, or 32-72, or 33-72, or 34-72, or 35-72, or 36-72, or 37-72, or 38-72, or 39-72, or 40-72, or 41-72, or 42-72, or 43-72, or 44-72, or 45-72, or 46-72, or 47-72, or 48-72, or 49-72, or 50-72, or 51-72, or 52-72, or 53-72, or 54-72, or 55-72, or 56-72, or 57-72, or 58-72, or 59-72, or 60-72, or 61-72, or 62-72, or 63-72, or 64-72, or 65-72, or 66-72, or 67-72, or 68-72, or 69-72, or 70-72, or 71-72 or 72. The transgene can be integrated within the endogenous gene in the upstream intron or at the beginning of the exon corresponding to the first exon within the transgene's partial coding sequence (FIG. 2). The transgene can be designed to be 4.7kb or less, and incorporated into an AAV vector and particle, and delivered in vivo to target cells.

In an embodiment, the transgene is a sequence of DNA that harbors a first and second partial coding sequence, wherein the partial coding sequences encode a partial protein, wherein the partial protein is homologous to a corresponding region in a functional protein produced from a wild type gene. The host gene or endogenous gene is one in which expression of the protein is aberrant, in other words, is not expressed, is expressed at low levels, or is expressed but the mRNA or protein product or portion thereof is non-functional, has reduced function, or has a gain-of-function, resulting in a disorder in the host.

As described herein, the donor molecule can be in a viral or non-viral vector. The vectors can be in the form of circular or linear double-stranded or single stranded DNA. The donor molecule can be conjugated or associated with a reagent that facilitates stability or cellular update. The reagent can be lipids, calcium phosphate, cationic polymers, DEAE-dextran, dendrimers, polyethylene glycol (PEG) cell penetrating peptides, gas-encapsulated microbubbles or magnetic beads. The donor molecule can be incorporated into a viral particle. The virus can be retroviral, adenoviral, adeno-

associated vectors (AAV), herpes simplex, pox virus, hybrid adenoviral vector, epstein-bar virus, lentivirus, or herpes simplex virus.

In certain embodiments, the AAV vectors as described herein can be derived from any AAV. In certain embodiments, the AAV vector is derived from the defective and nonpathogenic parvovirus adeno-associated type 2 virus. All such vectors are derived from a plasmid that retains only the AAV 145 bp inverted terminal repeats flanking the transgene expression cassette. Efficient gene transfer and stable transgene delivery due to integration into the genomes of the transduced cell are key features for this vector system. (Wagner et al., Lancet 351:9117 1702-3, 1998; Kearns et al., Gene Ther. 9:748-55, 1996). Other AAV serotypes, including AAV1, AAV2, AAV3, AAV4, AAV5, AAV6, AAV7, AAV8, AAV9 and AAVrh.10 and any novel AAV serotype can also be used in accordance with the present invention. In some embodiments, chimeric AAV is used where the viral origins of the long terminal repeat (LTR) sequences of the viral nucleic acid are heterologous to the viral origin of the capsid sequences. Non-limiting examples include chimeric virus with LTRs derived from AAV2 and capsids derived from AAV5, AAV6, AAV8 or AAV9 (i.e. AAV2/5, AAV2/6, AAV2/8 and AAV2/9, respectively).

The constructs described herein may also be incorporated into an adenoviral vector system. Adenoviral based vectors are capable of very high transduction efficiency in many cell types and do not require cell division. With such vectors, high titer and high levels of expression can been obtained.

The methods and compositions described herein are applicable to any eukaryotic organism in which it is desired to alter the organism through genomic modification. The eukaryotic organisms include plants, algae, animals, fungi and protists. The eukaryotic organisms can also include plant cells, algae cells, animal cells, fungal cells and protist cells.

Exemplary mammalian cells include, but are not limited to, oocytes, K562 cells, CHO (Chinese hamster ovary) cells, HEP-G2 cells, BaF-3 cells, Schneider cells, COS cells (monkey kidney cells expressing SV40 T-antigen), CV-1 cells, HuTu80 cells, NTERA2 cells, NB4 cells, HL-60 cells and HeLa cells, 293 cells (see, e.g., Graham et al.

(1977) J. Gen. Virol. 36:59), and myeloma cells like SP2 or NS0 (see, e.g., Galfre and Milstein (1981) Meth. Enzymol. 73(B):3 46). Peripheral blood mononucleocytes (PBMCs) or T-cells can also be used, as can embryonic and adult stem cells. For example, stem cells that can be used include embryonic stem cells (ES), induced pluripotent stem cells (iPSC), mesenchymal stem cells, hematopoietic stem cells, liver stem cells, skin stem cells and neuronal stem cells.

The methods and compositions of the invention can be used in the production of modified organisms. The modified organisms can be small mammals, companion animals, livestock, and primates. Non-limiting examples of rodents may include mice, rats, hamsters, gerbils, and guinea pigs. Non-limiting examples of companion animals may include cats, dogs, rabbits, hedgehogs, and ferrets. Non-limiting examples of livestock may include horses, goats, sheep, swine, llamas, alpacas, and cattle. Non-limiting examples of primates may include capuchin monkeys, chimpanzees, lemurs, macaques, marmosets, tamarins, spider monkeys, squirrel monkeys, and vervet monkeys. The methods and compositions of the invention can be used in humans.

Exemplary plants and plant cells which can be modified using the methods described herein include, but are not limited to, monocotyledonous plants (e.g., wheat, maize, rice, millet, barley, sugarcane), dicotyledonous plants (e.g., soybean, potato, tomato, alfalfa), fruit crops (e.g., tomato, apple, pear, strawberry, orange), forage crops (e.g., alfalfa), root vegetable crops (e.g., carrot, potato, sugar beets, yam), leafy vegetable crops (e.g., lettuce, spinach); vegetative crops for consumption (e.g. soybean and other legumes, squash, peppers, eggplant, celery etc.), flowering plants (e.g., petunia, rose, chrysanthemum), conifers and pine trees (e.g., pine fir, spruce); poplar trees (e.g. P. tremula×P. alba); fiber crops (cotton, jute, flax, bamboo) plants used in phytoremediation (e.g., heavy metal accumulating plants); oil crops (e.g., sunflower, rape seed) and plants used for experimental purposes (e.g., Arabidopsis). The methods disclosed herein can be used within the genera Asparagus, Avena, Brassica, Citrus, Citrullus, Capsicum, Cucurbita, Daucus, Erigeron, Glycine, Gossypium, Hordeum, Lactuca, Lolium, Lycopersicon, Malus, Manihot, Nicotiana, Orychophragmus, Oryza, Persea, Phaseolus, Pisum, Pyrus, Prunus, Raphanus, Secale, Solanum, Sorghum, Triticum, Vitis, Vigna, and

Zea. The term plant cells include isolated plant cells as well as whole plants or portions of whole plants such as seeds, callus, leaves, and roots. The present disclosure also encompasses seeds of the plants described above wherein the seed has the has been modified using the compositions and/or methods described herein. The present disclosure further encompasses the progeny, clones, cell lines or cells of the transgenic plants described above wherein said progeny, clone, cell line or cell has the transgene or gene construct. Exemplary algae species include microalgae, diatoms, Botryococcus braunii, Chlorella, Dunaliella tertiolecta, Gracileria, Pleurochrysis carterae, Sorgassum and Ulva.

The methods described in this document can include the use of rare-cutting endonucleases for stimulating homologous recombination or non-homologous integration of a transgene molecule into an endogenous gene. The rare-cutting endonuclease can include CRISPR, TALENs, or zinc-finger nucleases (ZFNs). The CRISPR system can include CRISPR/Cas9 or CRISPR/Cas12a (Cpf1). The CRISPR system can include variants which display broad PAM capability (Hu et al., *Nature* 556, 57-63, 2018; Nishimasu et al., *Science* DOI: 10.1126, 2018) or higher on-target binding or cleavage activity (Kleinstiver et al., *Nature* 529:490-495, 2016). The gene editing reagent can be in the format of a nuclease (Mali et al., *Science* 339:823-826, 2013; Christian et al., *Genetics* 186:757-761, 2010), nickase (Cong et al., *Science* 339:819-823, 2013; Wu et al., *Biochemical and Biophysical Research Communications* 1:261-266, 2014), CRISPR-*Fok*I dimers (Tsai et al., *Nature Biotechnology* 32:569-576, 2014), or paired CRISPR nickases (Ran et al., *Cell* 154:1380-1389, 2013).

The methods and compositions described in this document can be used in a circumstance where it is desired to modify the 3' end of the coding sequence of an endogenous gene. For example, patients with SCA3 or SCA6 have expanded CAG repeats in exons 10 (second to last exon) and exon 47 (last exon), respectively. Patients with SCA3 or SCA6 may benefit from replacement of exons 10-11 and exon 47, respectively. In other examples, patients with genetic disorders due to loss of function mutations within the 3' end of an endogenous gene could benefit from replacement of the final exons of said gene.

The invention will be further described in the following examples, which do not limit the scope of the invention described in the claims.

EXAMPLES

Example 1: Targeted Integration of DNA in the ATXN3 gene

Three plasmids were constructed with transgenes designed to integrate into the ATXN3 gene in human cells. All transgenes were designed to be inserted within intron 9 or the junction of intron 9 and exon 10 of the ATXN3 gene and all transgenes were designed to insert at least one splice acceptor and at least one functional coding sequence for exons 10 and 11 of the ATXN3 gene. The first plasmid, designated pBA1135, comprised a left and right homology arm with sequence homologous to the 3' end of intron 9 and 5' end of intron 10 (i.e., successful gene targeting would result in removal of exon 10 and replacement with the cargo sequence within pBA1135). Between the homology arms, from 5' to 3', was a splice acceptor (splice acceptor from ATXN3 intron 9), coding sequence for exons 10 and 11 of ATXN3, SV40 terminator, reverse BGH terminator, reverse coding sequence for exons 10 and 11 (codon adjusted), and reverse splice acceptor. The sequence for the pBA1135 transgene is shown in SEQ ID NO:17. A corresponding Cas9 nuclease was designed to cleave i) within intron 9 of the ATXN3 gene, ii) within the left homology arm of pBA1135, and iii) at the 3' end of the right homology arm of pBA1135. Successful cleavage of the plasmid was expected to liberate the transgene, thereby enabling the sequence to be used as a template for HR or for integration via NHEJ. The Cas9 gRNA target site is shown in SEQ ID NO:18. The individual elements within pBA1135 are shown in SEQ ID NOS:44-51. SEQ ID NO:44 comprises the left homology arm, nuclease target site, and splice acceptor. SEQ ID NO:45 comprises the partial coding sequence (exon 10 and 11) of a non-pathogenic ATXN3 gene. SEQ ID NO:46 comprises the SV40 p(A) terminator sequence. SEQ ID NO:47 comprises the BGH terminator in reverse complement. SEQ ID NO:48 comprises the reverse complement, codon adjusted partial coding sequence (exon 10 and 11) of a non-pathogenic ATXN3 gene. SEQ ID NO:49 comprises the sequence for the splice acceptor. SEQ ID NO:50 comprises the sequence for the right homology arm. SEQ ID

NO:51 comprises the target site sequence for the nuclease. The second plasmid, designated pBA1136, comprised the same cargo as pBA1135, however, the homology arms were removed. Nuclease target sites were kept to facilitate liberation of the transgene from the plasmid. Successful cleavage of the plasmid was expected to liberate the transgene, thereby enabling the sequence to be used for integration by NHEJ into the ATXN3 gene. The sequence of pBA1136 is shown in SEQ ID NO:19. The third plasmid, designated pBA1137, comprised the same sequence as pBA1135, except for the reverse sequences and nuclease target site (i.e., reverse terminator, reverse coding sequence and reverse splice acceptor). Plasmid pBA1137 was used as a control for conventional HR based methods. The sequence of pBA1137 is shown in SEQ ID NO:20.

Transfection was performed using HEK293T cells. HEK293T cells were maintained at 37°C and 5% CO2 in DMEM high supplemented with 10% fetal bovine serum (FBS). HEK293T cells were transfected with 2 ug of donor, 2 ug of guide RNA (RNA format) and 2 ug of Cas9 (RNA format). Transfections were performed using electroporation. Genomic DNA was isolated 72 hours post transfection and assessed for integration events. A list of primers used to detect integration or genomic DNA is shown in Table 1.

Table 1: Primers for detecting integration of transgenes in ATXN3.

Primer Name	Sequence (5' to 3')	SEQ ID NO:
oNJB043	CAAAGGTGCCCTTGAGGTT	21
oNJB044	AGGAGAAGTCTGCCGTTACT	22
oNJB113	GGACAAACCACAACTAGAATGC	23
oNJB114	TAGGAAAGGACAGTGGGAGT	24
oNJB116	CCATTATGTCTCAGTTGTTCAGTG	25
oNJB156	CCAGACCATCTCAGACACC	26
oNJB162	GGCTGGGCTTCCACTTAC	27
oNJB167	GTGGTTTGTCCAAACTCATCAA	28
oNJB170	AGTAACTCTGCACTTCCCATTG	29

To detect the integration of pBA1135, pBA1136 and pBA1137, PCRs were performed on the genomic DNA. Regarding pBA1137, the transgene was designed to be integrated precisely by HR. Accordingly, bands were detected in the 5' and 3' junction PCRs, which indicate precise insertion into exon 10 (FIG. 8 lanes 4 and 7). Expected band sizes were 1,520 bp for the 5' junction and 786 bp for the 3' junction. Primers oNJB113 and oNJB116 were used for the 5' junction PCR. Primers oNJB167 and oNJB170 were used for the 3' junction PCR. Regarding pBA1136, as no homology arms were present, the transgene was predicted to insert via NHEJ insertion. Appropriate size bands were observed for the transgene integrating in the forward and reverse directions. Integration in the forward direction can be seen in FIG. 8 lanes 3 (expected size approximately 1,520 bp) and 6 (expected size approximately 1,519 bp). Integrating in the reverse direction can be seen in FIG. 8 lane 12 (expected size approximately 1,520 bp). Primers oNJB113 and oNJB116 were used for the 5' junction PCR. Primers oNJB114 and oNJB170 were used for the 3' junction PCR. Primers oNJB116 and oNJB114 were used for the inverse 5' junction PCR. Regarding ppBA1135, both homology arms and nuclease cleavage sites were present on the transgene. Integration by HR was observed by detecting bands in the 5' and 3' junction PCRs (FIG. 8 lane 2 and 5). Further, integration by NHEJ was observed by detecting bands in an inverse 5' junction PCR (FIG. 8 lane 10). Expected size for the 5' junction PCR was 1,520 bp. Expected size for the 3' junction PCR was 1,157 bp. Expected size for the inverse 5' junction PCR was approximately 1,520 bp. Primers oNJB113 and oNJB116 were used for the 5' junction PCR. Primers oNJB114 and oNJB170 were used for the 3' junction PCR. Primers oNJB116 and oNJB114 were used for the inverse 5' junction PCR.

The results show that the described transgenes comprising bidirectional partial coding sequences can be integrated into genomic DNA through multiple different repair pathways.

Example 2: Targeted Integration of DNA in the CACNA1A gene

A CACNA1A-targeting transgene is designed to replace the 3' end of the CACNA1A coding sequence. A plasmid is constructed with a transgene designed to

integrate WT coding sequence into intron 46 or the start of exon 47 (FIG. 4). The transgene comprises a first homology arm which is homologous to sequence immediately following the splice donor site in intron 46. The first homology arm also comprises the target site for a nuclease (SEQ ID NO:9) and a splice acceptor sequence. The first homology arm is followed by a first coding sequence comprising the CACNA1A exon 47 and a non-expanded CAG repeat sequence (SEQ ID NO:3). Following the first coding sequence is a SV40 poly(A) termination sequence (SEQ ID NO:4). In a tail-to-tail orientation, a second set of functional elements is present. The beginning of the second set of elements comprises a target site for the nuclease (SEQ ID NO:9) followed by a second homology arm. The second homology arm harbors 446 bp which is homologous to sequence immediately following the stop coding (SEQ ID NO:8). This sequence was determined to be free of consensus branch or splice acceptor sequences via in silico analysis. Following the second homology arm is a second splice acceptor from carp beta-actin intron 1 (SEQ ID NO:7). Following the splice acceptor is a codon optimized version of the CACNA1A exon 47 (SEQ ID NO:6) and a bGH poly(A) terminator (SEQ ID NO:5).

A corresponding Cas12a nuclease is designed to create three double-strand breaks following transfection of the plasmid: i) within intron 46 of the endogenous CACNA1A gene, 2) within the first homology arm in the pBA1011-D1 transgene, and 3) following the second homology arm in the pBA1011-D1 transgene. The target sequence for the Cas12a nuclease is shown in SEQ ID NO:9.

Confirmation of the function of the transgene and CRISPR vectors is achieved by transfection of HEK293 cells. HEK293 cells are maintained at 37°C and 5% CO2 in DMEM high glucose without L-glutamine without sodium pyruvate medium supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin (PS) solution 100X. HEK293 cells are transfected with each of the plasmid constructs and combinations thereof using Lipofectamine 3000. Two days post transfection, DNA is extracted and assessed for mutations and targeted insertions within the CACNA1A gene. Nuclease activity is analyzed using the Cel-I assay or by deep sequencing of amplicons

comprising the CRISPR/Cas12a target sequence. Successful integration of the transgene is analyzed using PCR (FIG. 5).

Example 3: Targeted Integration of DNA in the ATXN3 gene

An ATXN3-targeting transgene is designed to replace the 3' end of the ATXN coding sequence (exons 10 and 11). A plasmid is constructed with a transgene designed to integrate WT coding sequence into intron 9 or the start of exon 10 (FIG. 5). The transgene comprises a first homology arm which is homologous to sequence intron 9 (SEQ ID NO:10). The first homology arm also comprises the target site for a Cas12a nuclease and a splice acceptor sequence. The first homology arm is followed by a first coding sequence comprising the ATXN3 exon 10 and 11 and a non-expanded CAG repeat sequence. Following the first coding sequence is a SV40 poly(A) termination sequence. In a tail-to-tail orientation, a second set of functional elements is present. The beginning of the second set of elements comprises a target site for the Cas12a nuclease followed by a second homology arm. The second homology arm harbors 379 bp which is homologous to sequence immediately following the end of exon 10 (i.e., the start of intron 10). This sequence was determined via in silico analysis to have a limited number of potential branch or splice acceptor sequences. Following the second homology arm is a second splice acceptor from carp beta-actin intron 1. Following the splice acceptor is a codon optimized version of the ATXN3 exons 10 and 11 and a bGH poly(A) terminator.

A corresponding Cas12a nuclease is designed to create three double-strand breaks following transfection of the plasmid: i) within intron 9 of the endogenous ATXN3 gene, 2) within the first homology arm in the pBA1012-D1 transgene, and 3) following the second homology arm in the pBA1012-D1 transgene. The target sequence for the Cas12a nuclease is shown in SEQ ID NO:11.

Confirmation of the function of the transgene and CRISPR vectors is achieved by transfection of HEK293 cells. HEK293 cells are maintained at 37°C and 5% CO2 in DMEM high glucose without L-glutamine without sodium pyruvate medium supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin (PS) solution 100X. HEK293 cells are transfected with each of the plasmid constructs and

combinations thereof using Lipofectamine 3000. Two days post transfection, DNA is extracted and assessed for mutations and targeted insertions within the ATXN3 gene. Nuclease activity is analyzed using the Cel-I assay or by deep sequencing of amplicons comprising the CRISPR/Cas12a target sequence. Successful integration of the transgene is analyzed using PCR (FIG. 7).

Example 4: Targeted Integration of DNA in the ATXN3 gene using Cas12k transposases

An ATXN3-targeting transgene is designed to replace the 3' end of the ATXN coding sequence (exons 10 and 11). A plasmid is constructed with a transgene designed to integrate WT coding sequence into intron 9 or the start of exon 10. The transgene comprises a transposon right end and left end, a first and second splice acceptor, a first and second coding sequence (encoding amino acids from exons 10 and 11), and a first and second terminator. The sequence between the transposon right and left ends is shown in SEQ ID NO: 17.

Plasmids are engineered to express the *Scytonema hofmanni* tnsB, tnsC, tniQ and Cas12k (SEQ ID NO:30) using eukaryotic promoters. A second plasmid is engineered to express the corresponding Cas12k guide RNA (SEQ ID NO:14). The guide RNA targeted sequence CCGCCCGACCTTTCACTTTC (SEQ ID NO:15). The Cas12k transposon plasmids is cotransformed in HEK293 cells with a plasmid harboring the ATXN3-targeting transgene. HEK293 cells are maintained at 37°C and 5% CO2 in DMEM high glucose without L-glutamine without sodium pyruvate medium supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin (PS) solution 100X. HEK293 cells are transfected with each of the plasmid constructs and combinations thereof using Lipofectamine 3000. Two days post transfection, DNA is extracted and assessed for targeted insertions within the ATXN3 gene. Integration of the transgene is analyzed using PCR.

Example 5: Targeted Integration of DNA in the CACNA1A gene

A CACNA1A-targeting transgene is designed to replace the 3' end of the CACNA1A coding sequence. A plasmid is constructed with a transgene designed to

integrate WT coding sequence into intron 46 or the start of exon 47. The transgene comprises a transposon right end and left end, a first and second splice acceptor, a first and second coding sequence (encoding amino acids from exon 47), and a first and second terminator.

Plasmids are engineered to express the Scytonema hofmanni tnsB, tnsC, tniQ and Cas12k (SEQ ID NO:30) using eukaryotic promoters. A second plasmid is engineered to express the corresponding Cas12k guide RNA (SEQ ID NO:14). The guide RNA is designed to target sequence CCCGGATCCCGGCTGTGACC (SEQ ID NO: 16). The Cas12k transposon plasmids are cotransformed in HEK293 cells with a plasmid harboring the ATXN3-targeting transgene. HEK293 cells are maintained at 37°C and 5% CO2 in DMEM high glucose without L-glutamine without sodium pyruvate medium supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin (PS) solution 100X. HEK293 cells are transfected with each of the plasmid constructs and combinations thereof using Lipofectamine 3000. Two days post transfection, DNA is extracted and assessed for targeted insertions within the ATXN3 gene. Integration of the transgene is analyzed using PCR.

OTHER EMBODIMENTS

It is to be understood that while the invention has been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended claims. Other aspects, advantages, and modifications are within the scope of the following claims.

WHAT IS CLAIMED IS:

1. A method of integrating a transgene into an endogenous gene, the method comprising:
    a. administering a transgene, wherein the transgene comprises
        i. a first and second splice acceptor sequence,
        ii. a first and second partial coding sequence, and
        iii. one bidirectional terminator or a first and second terminator;
    b. administering at least one rare-cutting endonuclease targeted to a site within the endogenous gene,
    wherein the transgene is integrated within the endogenous gene.
2. The method of claim 1, wherein the first splice acceptor is operably linked to the first partial coding sequence and the second splice acceptor is operably linked to the second partial coding sequence.
3. The method of claim 2, wherein the first partial coding sequence is operably linked to the first terminator, and the second partial coding sequence is operably linked to the second terminator.
4. The method of claim 2, wherein the first and second partial coding sequences are operably linked to the bidirectional terminator.
5. The method of claim 3, wherein the first and second splice acceptors, first and second coding sequences, and first and second terminators are oriented in a tail-to-tail orientation.
6. The method of claim 5, wherein the transgene further comprises a first and second target site for one or more rare-cutting endonucleases, wherein the target sites flank the first and second splice acceptors.
7. The method of claim 5, wherein the transgene further comprises a left and right homology arm which flank the first and second splice acceptors.
8. The method of claim 7, wherein the transgene is harbored within an adeno-associated viral vector.
9. The method of claim 7, wherein the transgene further comprises a first and second target site for the one or more rare-cutting endonucleases, wherein the target sites flank the first and second splice acceptors.

10. The method of claim 9, wherein the first and second target sites flank the first and second homology arms.
11. The method of claim 1, wherein the transgene is integrated within an intron of the endogenous gene or at an intron-exon junction.
12. The method of claim 1, wherein the transgene is integrated within an intron, or at the intron-exon junction of the ATXN3 gene or CACNA1A gene.
13. The method of claim 12, wherein the transgene comprises a first and second partial coding sequence encoding the peptide produced by exon 10 of a non-pathogenic ATXN3 gene and is targeted to intron 9, or the intron 9 exon 10 junction, of a pathogenic ATXN3 gene.
14. The method of claim 12, wherein the transgene comprises a first and second partial coding sequence encoding the peptide produced by exon 47 of a non-pathogenic CACNA1A gene and is targeted to intron 46, or the intron 46 exon 47 junction, of a pathogenic CACNA1A gene.
15. The method of claim 1, wherein the nuclease is a CRISPR/Cas12a nuclease or a CRISPR/Cas9 nuclease.
16. The method of claim 1, wherein the first and second partial coding sequences encode the same amino acids.
17. The method of claim 1, wherein the first and second coding sequences differ in nucleic acid sequence but encode the same amino acids.
18. The method of claim 1, wherein the transgene is harbored on a vector, wherein the vector format is selected from double-stranded linear DNA, double-stranded circular DNA, or a viral vector.
19. The method of claim 18, wherein the viral vector is selected from an adenovirus vector, an adeno-associated virus vector, or a lentivirus vector.
20. The method of claim 19, wherein the transgene is equal to or less than 4.7 kb.
21. The method of claim 1, wherein said endogenous gene is the wild type gene of said partial coding sequences.
22. The method of claim 21, wherein said endogenous gene is aberrant and the partial coding sequences encode a partial protein from a functional version of said endogenous gene.
23. A DNA polynucleotide comprising:

a. a first and second splice acceptor sequence,
b. a first and second partial coding sequence,
c. one bidirectional terminator or a first and second terminator,
d. optionally, a first and second homology arm, and
e. optionally, a first and second rare-cutting endonuclease target site.

24. A method of integrating a transgene into an endogenous gene, the method comprising:
a. administering a transgene, wherein the transgene comprises
i. a left and right transposon end,
ii. a first and second splice acceptor sequence,
iii. a first and second partial coding sequence, and
iv. one bidirectional terminator or a first and second terminator;
b. administering a transposase
wherein the transgene is integrated within the endogenous gene.

25. A method of integrating a transgene into an endogenous gene, the method comprising:
a. administering a transgene, wherein the transgene comprises
i. a first and second splice acceptor sequence,
ii. a first and second coding sequence, and
iii. one bidirectional terminator or a first and second terminator;
b. administering at least one rare-cutting endonuclease targeted to a site within the endogenous gene,
wherein the transgene is integrated within the endogenous gene.

26. A method of integrating a transgene into an endogenous gene, the method comprising:
a. administering a transgene, wherein the transgene comprises
i. a first and second splice acceptor sequence,
ii. a first and second coding sequence,
iii. one bidirectional terminator or a first and second terminator, and
iv. a first and second homology arm
wherein the transgene is integrated within the endogenous gene.