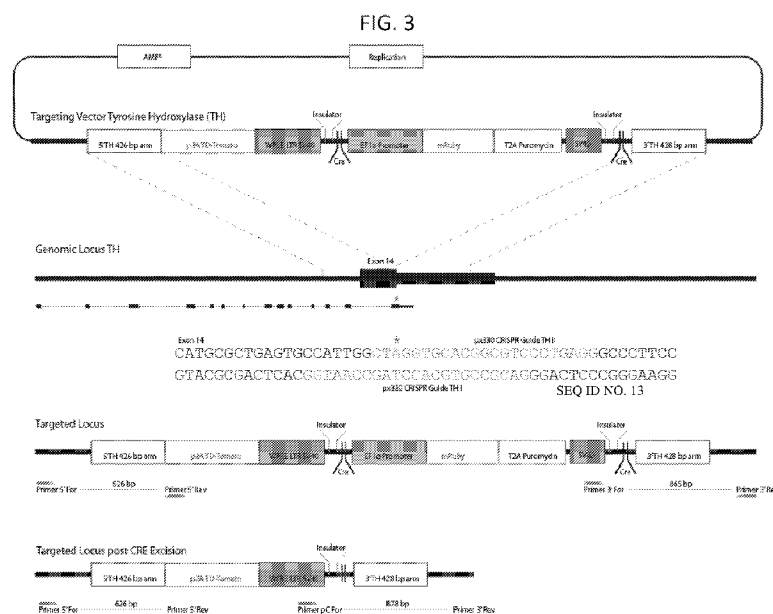




- (51) International Patent Classification:
C12N 15/85 (2006.01) *C07H 21/04* (2006.01)
- (21) International Application Number:
PCT/US2018/012849
- (22) International Filing Date:
08 January 2018 (08.01.2018)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
62/443,543 06 January 2017 (06.01.2017) US
- (71) Applicant: **PRESIDENT AND FELLOWS OF HARVARD COLLEGE** [US/US]; 17 Quincy Street, Cambridge, MA 02138 (US).
- (72) Inventors: **AHFELDT, Tim, D.**; 2651 HYMC 1 Oxford Street, Cambridge, MA 02138 (US). **RUBIN, Lee, L.**; 2 Barnstable Road, Wellesley, MA 02418 (US).
- (74) Agent: **WARREN, Lisa, M.**; Morse, Barnes-brown & Pendleton, P.C., Citypoint, 230 Third Avenue, 4th Floor, Waltham, MA 02451 (US).
- (81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

(54) Title: COMPOSITION AND METHODS FOR ENHANCED KNOCK-IN REPORTER GENE EXPRESSION



(57) Abstract: Provided are methods, vectors, transgenic cells, and compositions for expressing a marker with a gene of interest. In particular, provided are methods, vectors, transgenic cells and compositions for high expression of a fluorescent protein such as TDTomato with a gene of interest such as tyrosine hydroxylase in order to assess the expression of the gene of interest in vivo and in vitro.

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
KM, ML, MR, NE, SN, TD, TG).

Published:

- *without international search report and to be republished
upon receipt of that report (Rule 48.2(g))*

COMPOSITIONS AND METHODS FOR ENHANCED KNOCK-IN REPORTER GENE EXPRESSION

RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 62/443,543, filed January 6, 2017, the entire teachings of which are incorporated herein by reference.

BACKGROUND OF THE INVENTION

[0002] Co-expression of a gene of interest and a reporter gene is of great value for the study of cell differentiation and cellular biology. Techniques for enhanced expression of the reporter gene with a gene of interest are important for biological and biomedical research. For example, detecting expression of tyrosine hydroxylase in pluripotent stem cells by detecting the co-expression of a reporter gene is of great value for investigators studying midbrain neurons, dopaminergic neurons and Parkinson's disease pathology.

SUMMARY OF THE INVENTION

[0003] The present invention relates to compositions and methods useful for making reporter cells (i.e., cells co-expressing a reporter and a genetic locus of interest). The compositions and methods described herein provide a cell with a knock-in reporter (e.g., a fluorescent protein) and a downstream WPRE element, wherein the cell co-expresses a genetic locus of interest and a reporter gene.

[0004] In some aspects, the invention relates to a nucleic acid targeting vector comprising, in the 5' to 3' direction, a 5' homology arm homologous to a first target sequence in a cell, a reporter gene, an expression enhancer comprising a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE) operably linked to the reporter gene, and a 3' homology arm homologous to a second target sequence downstream of the first target sequence.

[0005] In other aspects, the invention relates to a composition comprising a cell (e.g., transgenic cell, a cell line) having a genome comprising a nucleotide sequence comprising a reporter gene and a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE) operably linked to the reporter gene, wherein the reporter gene is co-expressed with a target sequence of interest (e.g., genetic locus of interest, gene of interest).

[0006] In other aspects, the invention relates to a method of generating a cell that co-expresses a reporter gene with a gene of interest comprising providing a targeting vector; providing a cell in which at least a portion of the gene of interest is located between the first target sequence and the second target sequence; introducing the targeting vector into the cell; and maintaining the cell under conditions appropriate for integration of the reporter gene and WPRE into the genome of the cell such that the reporter gene is co-expressed with the gene of interest, wherein said portion of the gene of interest is cleaved prior to or subsequent to introducing the targeting vector into the cell.

[0007] Other aspects of the invention relate to a method of making a targeting vector capable of co-expressing a reporter gene with a gene of interest in a cell comprising providing a vector comprising, in the 5' to 3' direction, a reporter gene and an expression enhancer comprising a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE), wherein the expression enhancer and the reporter gene are operably linked; incorporating upstream of the reporter gene a 5' homology arm homologous to a first target

sequence; and incorporating downstream of the WPRE a 3' homology arm homologous to a second target sequence, wherein the second target sequence is located downstream of the first target sequence and wherein the first target sequence and the second target sequence flank at least a portion of the gene of interest.

[0008] The practice of the present invention will typically employ, unless otherwise indicated, conventional techniques of molecular biology, cell culture, recombinant nucleic acid (e.g., DNA) technology, immunology, nucleic acid and polypeptide synthesis, detection, manipulation, and quantification that are within the skill of the art. See, e.g., Ausubel, F., et al., (eds.), *Current Protocols in Molecular Biology*, *Current Protocols in Immunology*, *Current Protocols in Protein Science*, and *Current Protocols in Cell Biology*, all John Wiley & Sons, N.Y., edition as of December 2008; Sambrook, Russell, and Sambrook, *Molecular Cloning: A Laboratory Manual*, 3rd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2001; Harlow, E. and Lane, D., *Antibodies – A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1988. All patents, patent applications, publications, references, etc., cited in the instant patent application are incorporated by reference in their entirety. In the event of a conflict or inconsistency with the specification, the specification shall control. The Applicants reserve the right to amend the specification based on any of the incorporated references and/or to correct obvious errors. None of the content of the incorporated references shall limit the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawings will be provided by the Office upon request and payment of the necessary fee.

[0010] **FIG. 1** is a schematic illustrating the creation of targeting vectors and the insertion of a reporter gene and WPRE element. Commercially available HR120-PA-1 vector is shown at the top. To create a new multicistronic vector (HR120-p2A-TD-TOM), the copGFP-polyA cassette was removed using EcoRI and NruI. Vector sequence was restored via G-Block cloning which introduced a P2A cassette. Then, the XhoI site and Gibson assembly was used to introduce the TD-tomato followed by the WPRE element. To create the fusion protein vector (HR120-Clover), the copGFP-polyA cassette was removed using EcoRI and NruI. Vector sequence was restored via G-Block cloning and then the XhoI site and Gibson assembly was used to introduce the Clover followed by the WPRE element.

[0011] Homology arms were added to the multicistronic and fusion vectors via restriction digest and Gibson assembly. For the 5' homology addition in plasmid HR120-p2A-TD-TOM, NheI for the 5' arm and BamHI for the 3' arm were used. For the 5' homology addition in plasmid pHR120-clover, EcoRI for the 5' arm and BamHI for the 3' arm were used. The reporter gene and WPRE were inserted into the genomic locus shown (Genomic Locus TH) via CRISPR. A region of exon 14 TH gene with PAM sequences and suitable for targeting with a targetable nuclease (e.g., Cas9) is shown as sequences labeled px330 CRISPR Guide TH II and px330 CRISPR Guide TH I. The light blue areas of the sequences correspond to the guide RNA sequences and the PAM sequence. The asterisk shows the stop codon in the tyrosine hydroxylase exon 14. The resulting added sequences are shown in "Targeted Locus." Correct insertion direction was confirmed with an 865 nt PCR product for the 3' arm of the insert. Following selection of successful clones, the CRE

cassette was excised to result in the “targeted Locus post CRE excision.” Correct insertion was confirmed with an 878 nt PCR product for the 3’ arm of the insert.

[0012] **FIGS. 2A-2C** shows that correctly targeted clones were differentiated using a protocol based on a previously published protocol (FIG. 2A). Reporter expression can be seen as early as day 7. Cells were stained in embryoid bodies (EBs) using tyrosine hydroxylase (TH) antibody. The number of TH positive cells increased over time in the culture, as seen in the comparison between day 7 and day 30 EBs (FIG. 2B). Near perfect overlap between reporter expression and antibody staining was observed. Cells could be dissociated and plated or purified via FACS sorting which led to a strong enrichment with close to 100% of the cells expression the TD-tomato reporter (FIG. 2C).

[0013] **FIG. 3** is another schematic illustrating the insertion of a reporter gene and WPRE element into the tyrosine hydroxylase genomic locus (Genomic Locus TH). The reporter gene and WPRE element were inserted into the genomic locus shown (Genomic Locus TH) via CRISPR. A region of exon 14 TH gene with PAM sequences and suitable for targeting with a targetable nuclease (e.g., Cas9) is shown as sequences labeled px330 CRISPR Guide TH II and px330 CRISPR Guide TH I. The light blue areas of the sequences correspond to the guide RNA sequences and the PAM sequence. The asterisk shows the stop codon in the tyrosine hydroxylase exon 14. The resulting added sequences are shown in “Targeted Locus.” Correct insertion direction was confirmed with an 865 nt PCR product for the 3’ arm of the insert and a 626 bp PCR product for the 5’ arm of the insert. Following selection of successful clones, the CRE cassette was excised to result in the “targeted Locus post CRE excision.” Correct insertion was confirmed with an 878 nt PCR product for the 3’ arm of the insert and a 626 nt PCR product for the 5’ arm of the insert.

[0014] **FIGS. 4A-4F**- FIGS. 4A through 4F are illustrations of CRISPR mediated knock-out mutagenesis to create isogenic PD lines.

[0015] **FIGS. 5A-5G-** FIGS. 5A through 5G are illustrations of early onset PD mutations could result in increased rate of cell death in midbrain DANs in basal culture conditions.

[0016] **FIGS 6A-6E-** FIGS 6A through 6E are illustrations showing that global transcriptional analysis identifies overlapping dysregulated genes and pathways between *PARKIN*^{-/-} and *ATP13A2*^{-/-} cell lines.

[0017] **FIGS. 7A-7H-** FIGS. 7A through 7H are illustrations showing differential expression analysis showed a strong increase in the number of differentially expressed proteins during the time course of differentiation in the WT versus *PARKIN*^{-/-} comparison

[0018] **FIGS 8A-8C-** FIGS 8A through 8C are illustrations showing knockout of *DJ-1* leads to the dysregulation of proteins involved in cell cycle as well as proteins involved in the development of Charcot-Marie-Tooth disease.

[0019] **FIGS. 9A-9E-** FIGS. 9A through 9E are illustrations showing generation and characterization of isogenic tyrosine hydroxylase knock-in reporter cell lines carrying three distinct PD mutations.

[0020] **FIGS. 10A-10D-** FIGS. 10A through 10D are illustrations showing that targeting vector was designed to retain a largely unaltered endogenous TH gene product using a bicistronic targeting vector containing tdTomato.

[0021] **FIGS 11A-11B-** Illustration showing there was a significant increase of *PARKIN* mRNA relative to d0 in all differentiated cell types (Figure 11A). Illustration showing that most cells stained with an anti-TH antibody also expressed the tdTomato TH reporter (Figure 11B).

[0022] **FIGS. 12A-12E-** FIGS. 12A through 12E are illustration showing loss of *PARKIN* decreases the number of TH-positive neurons.

[0023] **FIGS. 13A-13C-** FIGS. 13A through 13C are illustrations showing Global transcriptional analysis identifies overlapping dysregulated genes and pathways between PARKIN^{-/-} and ATP13A2^{-/-} cell lines.

[0024] **FIGS 14A-14C-** FIGS. 14A through 14C are illustrations showing quantitative proteomics reveals overlap in dysregulated pathways in isogenic PD lines.

[0025] **FIGS 15A-15C-** FIGS 15A through 15C are illustrations showing quantitative proteomics reveals overlap in dysregulated pathways in isogenic PD lines.

DETAILED DESCRIPTION OF THE INVENTION

[0026] The practice of the present invention will typically employ, unless otherwise indicated, conventional techniques of cell biology, cell culture, molecular biology, transgenic biology, microbiology, recombinant nucleic acid (e.g., DNA) technology, immunology, and RNA interference (RNAi) which are within the skill of the art. Non-limiting descriptions of certain of these techniques are found in the following publications: Ausubel, F., et al., (eds.), Current Protocols in Molecular Biology, Current Protocols in Immunology, Current Protocols in Protein Science, and Current Protocols in Cell Biology, all John Wiley & Sons, N.Y., edition as of December 2008; Sambrook, Russell, and Sambrook, Molecular Cloning: A Laboratory Manual, 3rd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2001; Harlow, E. and Lane, D., Antibodies – A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1988; Freshney, R.I., “Culture of Animal Cells, A Manual of Basic Technique”, 5th ed., John Wiley & Sons, Hoboken, NJ, 2005. Non-limiting information regarding therapeutic agents and human diseases is found in Goodman and Gilman’s The Pharmacological Basis of Therapeutics, 11th Ed., McGraw Hill, 2005, Katzung, B. (ed.) Basic and Clinical Pharmacology, McGraw-Hill/Appleton & Lange; 10th ed. (2006) or 11th edition (July 2009). Non-limiting information regarding genes and genetic

disorders is found in McKusick, V.A.: Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. Baltimore: Johns Hopkins University Press, 1998 (12th edition) or the more recent online database: Online Mendelian Inheritance in Man, OMIM™. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), as of May 1, 2010, ncbi.nlm.nih.gov/omim/ and in Online Mendelian Inheritance in Animals (OMIA), a database of genes, inherited disorders and traits in animal species (other than human and mouse), at omia.angis.org.au/contact.shtml. All patents, patent applications, and other publications (e.g., scientific articles, books, websites, and databases) mentioned herein are incorporated by reference in their entirety. In case of a conflict between the specification and any of the incorporated references, the specification (including any amendments thereof, which may be based on an incorporated reference), shall control. Standard art-accepted meanings of terms are used herein unless indicated otherwise. Standard abbreviations for various terms are used herein.

[0027] The compositions and methods disclosed herein generally relate to compositions and methods useful for making and using reporter cells (i.e., cells co-expressing a reporter and a genetic locus of interest). In particular, such compositions and methods can provide a cell with a knock-in reporter (e.g., a fluorescent protein) and a downstream WPRE element. This combination dramatically and unexpectedly improves the co-expression of the reporter and genetic locus of interest versus standard methods in the art. In some embodiments, the reporter and WPRE element are inserted into embryonic stem cells to co-express with tyrosine hydroxylase, a protein indicative of dopaminergic neurons. Differentiation of the embryonic stem cells into dopaminergic neurons can then be tracked or differentiated cells can be isolated using the properties of the reporter (e.g., fluorescence). In contrast to traditional protocols, the compositions and methods disclosed herein provide

much higher co-expression of the reporter with the genetic locus of interest, dramatically increasing the robustness and sensitivity of identifying cells expressing the genetic locus of interest.

[0028] *Targeting Vector*

[0029] In certain embodiments, the compositions disclosed herein relate to a nucleic acid targeting vector having homology arms flanking a reporter gene and an expression enhancer comprising a WPRE operably linked to the reporter gene.

[0030] The term “nucleic acid” refers to polynucleotides such as deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). The terms “nucleic acid” and “polynucleotide” are used interchangeably herein and should be understood to include double-stranded polynucleotides, single-stranded (such as sense or antisense) polynucleotides, and partially double-stranded polynucleotides. A nucleic acid often comprises standard nucleotides typically found in naturally occurring DNA or RNA (which can include modifications such as methylated nucleobases), joined by phosphodiester bonds. In some embodiments a nucleic acid may comprise one or more non-standard nucleotides, which may be naturally occurring or non-naturally occurring (i.e., artificial; not found in nature) in various embodiments and/or may contain a modified sugar or modified backbone linkage. Nucleic acid modifications (e.g., base, sugar, and/or backbone modifications), non-standard nucleotides or nucleosides, etc. may be incorporated in various embodiments. Such modifications may, for example, increase stability (e.g., by reducing sensitivity to cleavage by nucleases), decrease clearance in vivo, increase cell uptake, or confer other properties that improve the translation, potency, efficacy, specificity, or otherwise render the nucleic acid more suitable for an intended use. Various non-limiting examples of nucleic acid modifications are described in, e.g., Deleavey GF, et al., Chemical modification of siRNA. Curr. Protoc. Nucleic Acid Chem. 2009; 39:16.3.1-16.3.22; Crooke, ST (ed.) Antisense drug technology: principles, strategies, and

applications, Boca Raton: CRC Press, 2008; Kurreck, J. (ed.) Therapeutic oligonucleotides, RSC biomolecular sciences. Cambridge: Royal Society of Chemistry, 2008; U. S. Patent Nos. 4,469,863; 5,536,821 ; 5,541,306; 5,637,683; 5,637,684; 5,700,922; 5,717,083; 5,719,262; 5,739,308; 5,773,601; 5,886,165; 5,929, 226; 5,977,296; 6,140,482; 6,455,308 and/or in PCT application publications WO 00/56746 and WO 01/14398. Different modifications may be used in the two strands of a double-stranded nucleic acid. A nucleic acid may be modified uniformly or on only a portion thereof and/or may contain multiple different modifications. Where the length of a nucleic acid or nucleic acid region is given in terms of a number of nucleotides (nt) it should be understood that the number refers to the number of nucleotides in a single-stranded nucleic acid or in each strand of a double-stranded nucleic acid unless otherwise indicated. An “oligonucleotide” is a relatively short nucleic acid, typically between about 5 and about 100 nt long.

[0031] The term “targeting vector” refers to a vector comprising a polynucleotide having homology regions (i.e., homology arms) with sequences that are homologous to sequences present in a host cell genetic locus. The homology arms flank a polynucleotide region (e.g., region containing a reporter gene and WPRE) which becomes integrated into a host cell genetic locus. The term “vector” as used herein refers to a nucleic acid or a virus or portion thereof (e.g., a viral capsid or genome) capable of mediating entry of, e.g., transferring, transporting, etc., a nucleic acid into a cell. Where the vector is a nucleic acid, the nucleic acid to be transferred is generally linked to, e.g., present in, the vector. A nucleic acid vector may include sequences that direct autonomous replication (e.g., an origin of replication). Useful nucleic acid vectors include, for example, naturally occurring or modified viral genomes or portions thereof or nucleic acids (DNA or RNA) that can be packaged into viral capsids, DNA or RNA plasmids, and transposons. Plasmid vectors typically include an origin of replication and may include one or more selectable marker

genes. Viruses or portions thereof that can be used to introduce nucleic acid molecules into cells are referred to as viral vectors. Useful viral vectors include adenoviruses, adeno-associated viruses, retroviruses, lentiviruses, vaccinia virus and other poxviruses, herpesviruses (e.g., herpes simplex virus), and others. In some embodiments a virus having tropism for a particular cell type (e.g., neurons or a particular type of neuron) may be used. Examples of expression vectors that may be used in mammalian cells include, e.g., the pcDNA vector series, pSV2 vector series, pCMV vector series, pRSV vector series, pEF1 vector series, Gateway® vectors, and PrecisionX™ HR Targeting Vectors, etc. One of ordinary skill in the art appreciates how to use a viral vector, plasmid, or other vector to introduce a DNA sequence of interest into a cell.

[0032] The term “expression enhancer” is intended to refer to a polynucleotide region or regions that binds proteins (e.g., transcription factors) to enhance (increase) transcription of a gene. Enhancers may be located some distance away from the promoters and transcription start site (TSS) of genes whose transcription they regulate and may be located upstream or downstream of the TSS. In some embodiments, the expression enhancer is located downstream of the TSS.

[0033] Woodchuck Posttranscriptional Regulatory Element (WPRE) is a posttranscriptional regulatory element (Donello, et al., “Woodchuck hepatitis virus contain a tripartite posttranscriptional regulatory element,” J. Virol., 72 (1998), pp. 5085–5092, incorporated herein by reference in its entirety) from the Woodchuck Virus that facilitates nucleocytoplasmic transport of RNA mediated by several alternative pathways that may be cooperative (Popa, et al., “CRM1-dependent function of a cis-acting RNA export element,” Mol. Cell. Biol., 22 (2002), pp. 2057–2067, incorporated herein by reference in its entirety). In addition, the WPRE has been shown to act on additional posttranscriptional mechanisms to stimulate expression of heterologous cDNAs (Zufferey et al., “Woodchuck hepatitis virus

posttranscriptional regulatory element enhances expression of transgenes delivered by retroviral vectors,' J. Virol., 73 (1999), pp. 2886–2892, incorporated herein by reference in its entirety).

[0034] The term “operably linked” refers to a nucleic acid regulatory element and a nucleic acid sequence being appropriately positioned relative to each other so as to place expression of the nucleic acid under the influence or control of the regulatory element(s). For example, an expression enhancer and a reporter gene are considered “operably linked” if they are positioned in such a way in a DNA molecule that the expression enhancer region enhances (increases) transcription of the reporter gene under appropriate conditions. As used herein, “operably linked” refers to the positional relationship between the regulatory element(s) (e.g., WPRE) and the nucleic acid sequence (e.g., reporter gene). It will be understood that whether a particular expression enhancer does in fact enhance transcription of an operably linked nucleic acid molecule (e.g., reporter gene), may depend on a variety of factors, such as the presence or absence of appropriate factors and/or the presence or absence of inhibitory substances.

[0035] The term “reporter” refers to a molecule that can be used as an indicator of the occurrence or level of a particular biological process, activity, event, or state in a cell or organism. Reporters typically have one or more properties or enzymatic activities that allow them to be readily measured or that allow selection of a cell that expresses the reporter molecule. In general, a cell can be assayed for the presence of a reporter by measuring the reporter itself or an enzymatic activity of the reporter protein. Detectable characteristics or activities that a reporter may have include, e.g., fluorescence, bioluminescence, ability to catalyze a reaction that produces a fluorescent or colored substance in the presence of a suitable substrate, or other readouts based on emission and/or absorption of photons (light).

Typically, a reporter is a molecule that is not endogenously expressed by a cell or organism in which the reporter is used.

[0036] The term “reporter gene” refers to a nucleic acid that encodes a reporter. The reporter construct may be assembled in or inserted into a vector. The reporter construct or vector may be transferred into one or more cells. The reporter gene may be integrated into the genome. After transfer, cells are assayed for the presence of the reporter by measuring the reporter or the activity (e.g., enzymatic activity) of the reporter. In some embodiments, a reporter gene is codon-optimized for expression in mammalian cells. In some embodiments, a reporter gene is codon-optimized for expression in human cells.

[0037] The term “homologous” means two or more nucleic acid sequences that are either identical or similar enough that they are able to hybridize to each other or undergo intermolecular exchange. As used herein, sequences are homologous if they are either identical or similar enough that they are able to hybridize to each other under physiological conditions present in a cell (e.g., a mammalian cell). As used herein, a “homology arm” refers to a region of a nucleic acid targeting vector homologous to a genomic region.

[0038] In some embodiments, at least one of homology arms is homologous to a region of a genetic locus (e.g., gene of interest). In some embodiments, the homology arms comprise a 5' homology arm homologous to a first target sequence in a cell and a 3' homology arm homologous to a second target sequence downstream of the first target sequence. The 5' and 3' homology arms may be homologous to a contiguous region of the genome of the cell or homologous to discontinuous regions of the genome of the cell. Using homology arms that are homologous to contiguous genomic regions enables knock-in of a reporter gene without removal of endogenous genomic nucleotide sequence. Using homology arms that are homologous to discontinuous genomic regions may enable both knock-in of the reporter gene and knock-out of an endogenous genomic nucleotide sequence.

“Knock-in” is a genetic modification resulting from the addition of the genetic information encoded in a chromosomal locus with further DNA sequence. “Knock-out” is a genetic modification resulting from the disruption or removal of the genetic information encoded in a chromosomal locus.

[0039] In some embodiments, the 5' homology arm is homologous to a target sequence immediately upstream of a stop codon of a genetic locus (e.g., gene of interest) and the 3' homology arm is homologous to a target sequence comprising the stop codon of the genetic locus (e.g., gene of interest), thereby enabling incorporation of the reporter gene and expression enhancer into a chromosome so that the reporter gene is co-expressed with the genetic locus (e.g., gene of interest) without changing the primary structure of a gene product. In some embodiments, the homology arms are both homologous to target sequences partially or fully upstream of a stop codon of the genetic locus (e.g., gene of interest). In some instances, insertion of a reporter gene and expression enhancer within the sequence encoding a gene product does not disrupt the function of the gene product.

[0040] Each of the homology arms may comprise about 40 or more nucleotides. In some embodiments, each homology arm comprises about 50-1000 nucleotides, about 100-800 nucleotides, about 200-500 nucleotides, or about 300-400 nucleotides. In some embodiments, each homology arm comprises about 350 nucleotides. In some embodiments the homology arms are between about 100 nt – 200 nt, about 200 nt – 300 nt, about 300 nt – 400 nt, about 400 nt – 500 nt, about 500 nt – 750 nt, about 750 nt -1000 nt, about 1 kb – 1.5 kb, or more. The two homology arms may be about the same length (e.g., within about 50 – 100 nt of each other) or may differ in length by more than about 100 nt. Either or both homology arms can independently fall within any of the afore-mentioned ranges. One of ordinary skill in the art appreciates that the homology arms need not be perfectly homologous to the genomic DNA. In some embodiments the homologous region(s) of a donor nucleic

acid have at least 50% 60%, 70%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99.5%, 99.9% or more sequence identity to a genomic sequence with which homologous recombination is desired.

[0041] In some embodiments, the homology arms are homologous to regions flanking a targeted nuclease cut site. As used herein, “flanking” indicates that the homology arms are located on either side of the targeted nuclease cut site, the flanking homology arms may be directly on either side of the cut site (contiguous) or one or both of the flanking homology arms may be some distance away from the cut site (non-contiguous). In some embodiments, the targeted nucleic acid cut site is located at the junction between contiguous flanking regions homologous to the homology arms. In some embodiments, the targeted nuclease cut site is within about 100 nt of the region homologous to the 3' end of the 5' homology arm. In some embodiments, the targeted nuclease cut site is within about 100 nt of the region homologous to the 5' end of the 3' homology arm. In some embodiments, the targeted nuclease cut site is within about 50 nt of the region homologous to the 3' end of the 5' homology arm. In some embodiments, the targeted nuclease cut site is within about 50 nt of the region homologous to the 5' end of the 3' homology arm. In some embodiments, the targeted nuclease cut site is within about 10 nt of the region homologous to the 3' end of the 5' homology arm. In some embodiments, the targeted nuclease cut site is within about 10 nt of the region homologous to the 5' end of the 3' homology arm. In some embodiments, the targeted nuclease cut site is within about 5 nt of the region homologous to the 3' end of the 5' homology arm. In some embodiments, the targeted nuclease cut site is within about 5 nt of the region homologous to the 5' end of the 3' homology arm. In some embodiments, the targeted nuclease cut site is within about 0-100 nt of the region homologous to the 5' end of the 3' homology arm and about 0-100 nt of the region homologous to the 3' end of the 5' homology arm. In some embodiments, a guide sequence for the targetable nuclease is not

homologous to the targeting vector. In some embodiments, a guide sequence for the targetable nuclease is not homologous to a genomic sequence comprising the inserted reporter gene and WPRE.

[0042] In some embodiments, the homology arms are homologous to one or more regions of the human tyrosine hydroxylase locus (Gene ID: 7054; NCBI). In some embodiments, the 5' homology arm is homologous to the 5' end of exon 14 of the human tyrosine hydroxylase locus and the 3' homology arm is homologous to a region comprising the human tyrosine hydroxylase stop codon with the region homologous to the 3' homology arm. In some embodiments, the 5' homology arm is homologous to the 5' end of exon 14 of the human tyrosine hydroxylase locus and the 3' homology arm is homologous to a region comprising the human tyrosine hydroxylase stop codon that is contiguous with the region homologous to the 3' homology arm. In some embodiments, the 5' homology arm comprises, consists essentially, or consists of the nucleotide sequence of SEQ ID NO: 1. In some embodiments, the 3' homology arm comprises, consists essentially, or consists of the nucleotide sequence of SEQ ID NO: 2.

5 Prime TH homology region (426 nt):

**TTCCTGGAGGAGGCCAGTGGAGGTTGAGGGAGGGATGGGGTGCCC
GGCAGTCTCTAGTGGAAAAGGCGCCTAGCCTATCTCCCCCATGAACC
CCCTCACCCAGCCCTGGAAGAGGCCTCAGTGTCCCGCCTGTGACCAG
TTGGCTCAGAAAAGCCCTGGGAGCTCTGAGCCACTGTGAAGGTGGAA
ACGCGGCCCCCTGGCCTCCCCTCTCCTGGAGGCTGCAGACTCTGCCCCG
CCAGTTGACGAGGGCTCTGCCGCTCTCCTCCCCAGGAGCTATGCCTC
ACGCATCCAGCGCCCCCTTCTCCGTGAAGTTCGACCCGTACACGCTGG
CCATCGACGTGCTGGACAGCCCCCAGGCCGTGCGGCGCTCCCTGGAG**

**GGTGTCCAGGATGAGCTGGACACCCTTGCCCATGCGCTGAGTGCCAT
TGGC (SEQ ID NO: 1)**

3 Prime TH homology region (428 nt):

**GTGCACGGCGTCCCTGAGGGCCCTTCCCAACCTCCCCTGGTCCTGCA
CTGTCCCGGAGCTCAGGCCCTGGTGAGGGGCTGGGTCCCGGGTGCCC
CCCATGCCCTCCCTGCTGCCAGGCTCCCACTGCCCCCTGCACCTGCTTC
TCAGCGCAACAGCTGTGTGTGCCCCGTGGTGAGGTTGTGCTGCCTGTG
GTGAGGTCCTGTCCTGGCTCCCAGGGTCCTGGGGGCTGCTGCACTGC
CCTCCGCCCTTCCCTGACACTGTCTGCTGCCCCAATCACCGTCACAAT
AAAAGAACTGTGGTCTCTACACCTGCCTGGCCCCACATCTGTGCCAC
AGAGACAGACCCTGGGATCCTCAGACTCCCACACCCCCACCCCAGCC
TCACTCAGAGGTTTCGCCCTGGCCTCCTTCCTCCTCTGGGAGATGGCT
G (SEQ ID NO: 2)**

[0043] In some embodiments, the nucleic acid targeting vector (i.e., targeting vector) comprises, in the 5' to 3' direction, (i) a 5' homology arm homologous to a first target sequence in a cell, (ii) a reporter gene, (iii) an expression enhancer comprising a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE) operably linked to the reporter gene, and (iv) a 3' homology arm homologous to a second target sequence downstream of the first target sequence.

[0044] In some embodiments, the expression enhancer region of the targeting vector may comprise the WPRE element and further transcription enhancer elements (e.g., SV40 enhancer, LTR). Transcription enhancers increase the likelihood of transcription of a particular gene. Any suitable transcription enhancer may be included in the expression enhancer region with the WPRE.

[0045] In some embodiments, the reporter gene encodes a fluorescent protein. Any suitable fluorescent protein may be used. For instance, fluorescent proteins which may be suitable can be found on the world-wide web at nic.ucsf.edu/dokuwiki/doku.php?id=fluorescent_proteins. In some embodiments the fluorescent protein is a green fluorescent protein, a red fluorescent protein, or an infrared fluorescent protein. Examples of fluorescent proteins that may be used include, e.g., GFP, EGFP, Sirius, Azurite, EBFP2, BFP, mTurquoise, ECFP, Cerulean, mTFP1, mUkG1, mAG1, AcGFP, mWasabi, EmGFP, YPF, EYFP, Topaz, SYFP2, Venus, Citrine, mKO, mKO2, mOrange, mOrange2, LSSmOrange, PSmOrange, and PSmOrange2, mStrawberry, mRuby, mCherry, mRaspberry, tdTomato, mKate, mKate2, mPlum, mNeptune, T-Sapphire, mAmetrine, mKeima, E2-Orange, E2-Red/Green, and E2-Crimson, ZsGreen. See, e.g., See, e.g., Chalfie, M. and Kain, SR (eds.) Green fluorescent protein: properties, applications, and protocols (Methods of biochemical analysis, v. 47) Wiley-Interscience, Hoboken, N.J., 2006; Chudakov, DM, et al., *Physiol Rev.* 90(3):1103-63, 2010, US Pat. Pub. Nos. 20030170911, 20060194282, 20070099175, 20090203035, 20100227400; 20100184954; 20110020784; 20140237632 for further description of various reporter molecules that may be used. In some embodiments, the fluorescent protein is CLOVER or TD-TOMATO. In some embodiments, the reporter gene further comprises a stop codon downstream of the sequence encoding a guide protein.

[0046] In some embodiments, the targeting vector has an IRES element or a sequence encoding a self-cleaving peptide located between the 5' homology arm and the reporter gene. In some embodiments, the sequence encoding a self-cleaving peptide encodes p2A, t2A, e2A, f2A. In some embodiments, the sequence encoding the self-cleaving peptide also encodes for a GSG sequence at the amino terminus to enhance cleavage efficiency. In some

embodiments, the targeting vector does not have an IRES element or sequence encoding a self-cleaving peptide between the 5' homology arm and the reporter gene.

[0047] In some embodiments, the targeting vector has an insulator sequence located between the WPRE and the 3' homology arm. The length of the insulator sequence is not limited. In some embodiments, the insulator sequence is about 1-10 nt, about 1-50 nt, about 1-100 nt, or about 1-500 nt in length. In some embodiments, the insulator sequence blocks transcription of the 3' homology arm. In some embodiments, the targeting vector may have one or more restriction sites. In some embodiments, the insulator sequence has one or more restriction sites.

[0048] In some embodiments, the targeting vector does not comprise a promoter sequence upstream of and/or operably linked to the reporter gene or WPRE element.

[0049] In some embodiments, the targeting vector may include an expression cassette having positive and/or negative selection or screening markers. Positive selection markers are those polynucleotides that encode a product that enables only cells that carry and express the gene to survive and/or grow under certain conditions. For example, cells that express neomycin resistance (Neo^R) gene are resistant to the compound G418, while cells that do not express Neo^R are killed by G418. Positive selection markers are not limited and can include hygromycin resistance, ZeocinTM resistance, and/or Puromycin resistance. Negative selection markers are those polynucleotides that encode a produce that enables only cells that carry and express the gene to be killed under certain conditions. For example, cells that express thymidine kinase (e.g., herpes simplex virus thymidine kinase, HSV-TK) are killed when gancyclovir is added. Any known negative selection marker is contemplated and is not limited. Screening markers that may be used can be, for example, flourescent proteins or luciferases (e.g., GFP, mRUBY), or beta-galactosidase. Other screening markers may include sequences encoding polypeptides that will be expressed on the cell surface, allowing for

identification with specific antibodies or other ligands to that surface expressed polypeptide. The antibodies or ligands in these assays may be tagged in some manner, for example with a fluorophore, to allow rapid cell screening. In some embodiments, the expression cassette having positive and/or negative selection or screening markers further comprises LoxP sites flanking the selection and/or screening markers.

[0050] *Cell Compositions*

[0051] In some embodiments, the invention is directed towards a composition comprising a cell having a genome comprising a nucleotide sequence having a reporter gene and a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE) operably linked to the reporter gene, wherein the reporter gene is co-expressed with a target sequence of interest.

[0052] In some embodiments, the cell is a non-naturally occurring transgenic cell. In some embodiments the cell is a mammalian cell, e.g. a human, non-human primate, rodent (e.g., mouse, rat, rabbit, hamster), ungulate (e.g., ovine, bovine, equine, caprine species), canine, or feline cell. In some embodiments, the cell is an avian cell (e.g., chicken). In some embodiments the cell is a somatic cell. In some embodiments the cell is a pluripotent stem cell, and induced pluripotent stem cell or a multipotent stem cell. In some embodiments the cell is a germ cell, stem cell, or zygote. In some embodiments the cell is a primary cell. In some embodiments the cell is a diseased cell. In some embodiments the cell is a cancer cell. In some embodiments the cell is a white blood cell or fibroblast. In some embodiments the cell is a cell that has been isolated from an embryo. In some embodiments, the cell is an embryonic stem cell. Cells of the invention include, but are not limited to, hepatocytes, epithelial cells, hematopoietic cells, epithelial cells, endothelial cells, lung cells, bone cells, stem cells, mesenchymal cells, neural cells, cardiac cells, adipocytes, vascular smooth muscle cells, cardiomyocytes, skeletal muscle cells, beta cells, pituitary cells, synovial lining cells,

ovarian cells, testicular cells, fibroblasts, B cells, T cells, reticulocytes, leukocytes, granulocytes and tumor cells. In certain embodiments, the cell is a neural cell (e.g., meninges, astrocyte, motor neuron, a cell of the dorsal root ganglia or anterior horn motor neuron), a neural lineage cell or a neural stem cell. In some embodiments, the cell is a pluripotent stem cell or an induced pluripotent stem cell. In some embodiments, the cell is a human pluripotent stem cell or a human induced pluripotent stem cell. In some embodiments, the cell is in a non-human transgenic animal. In some embodiments, the transgenic animal is a mouse, rat or non-human primate.

[0053] The phrase “co-expressed with a target sequence of interest” refers to expression of the reporter gene with the target sequence of interest, usually on the same mRNA. The co-expression may result in a fusion protein or a separate reporter protein and target sequence product. The reporter gene may be any reporter gene as described herein. As used herein co-expression is intended to mean that expression of the reporter gene substantially matches the expression of target sequence of interest.

[0054] The target sequence of interest (e.g., genomic locus of interest) is not limited. In some embodiments, the target sequence of interest is a gene of interest. In some embodiments, the gene of interest encodes a transcription factor, a transcriptional co-activator or co-repressor, an enzyme, a chaperone, a heat shock factor, a heat shock protein, a receptor, a secreted protein, a transmembrane protein, a histone (e.g., H1, H2A, H2B, H3, H4), a peripheral membrane protein, a soluble protein, a nuclear protein, a mitochondrial protein, a growth factor, a cytokine (e.g., an interleukin, e.g., any of IL-1 - IL-33), an interferon (e.g., alpha, beta, or gamma), a chemokine (e.g., a CXC, CX3C, C (or XC), or CX3C chemokine). A chemokine may be CCL1 - CCL28, CXCL1 - CXCL17, XCL1 or XCL2, or CXC3L1). In some embodiments the gene of interest encodes a colony-stimulating factor, a hormone (e.g., insulin, thyroid hormone, growth hormone, estrogen, progesterone, testosterone), an

extracellular matrix protein (e.g., collagen, fibronectin), a motor protein (e.g., dynein, myosin), cell adhesion molecule, a major or minor histocompatibility (MHC) gene, a transporter, a channel (e.g., an ion channel), an immunoglobulin (Ig) superfamily (IgSF) gene (e.g., a gene encoding an antibody, T cell receptor, B cell receptor), tumor necrosis factor, an NF-kappaB protein, an integrin, a cadherin superfamily member (e.g., a cadherin), a selectin, a clotting factor, a complement factor, a plasminogen, plasminogen activating factor. Growth factors include, e.g., members of the vascular endothelial growth factor (VEGF, e.g., VEGF-A, VEGF-B, VEGF-C, VEGF-D), epidermal growth factor (EGF), insulin-like growth factor (IGF; IGF-1, IGF-2), fibroblast growth factor (FGF, e.g., FGF1 - FGF22), platelet derived growth factor (PDGF), or nerve growth factor (NGF) families. It will be understood that the afore-mentioned protein families comprise multiple members. Any such member may be used in various embodiments. In some embodiments a growth factor promotes proliferation and/or differentiation of one or more hematopoietic cell types. For example, a growth factor may be CSF1 (macrophage colony- stimulating factor), CSF2 (granulocyte macrophage colony- stimulating factor, GM-CSF), or CSF3 (granulocyte colony-stimulating factors, G-CSF). In some embodiments the gene of interest encodes erythropoietin (EPO). In some embodiments, the gene of interest encodes a neurotrophic factor, i.e., a factor that promotes survival, development and/or function of neural lineage cells (which term as used herein includes neural progenitor cells, neurons, and glial cells, e.g., astrocytes, oligodendrocytes, microglia). For example, in some embodiments, the protein is a factor that promotes neurite outgrowth. In some embodiments, the protein is ciliary neurotrophic factor (CNTF) or brain-derived neurotrophic factor (BDNF).

[0055] In some embodiments, the gene of interest is a human tyrosine hydroxylase gene. In some embodiments, the gene (e.g., human gene) of interest is SLC6A3, AGRP, POMC, HB9, GFAP, SCN10A, SCN9A, or TRPV1.

[0056] In some embodiments, the cell comprises two or more nucleotide sequences each having a reporter gene and a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE) operably linked to the reporter gene, wherein the reporter gene of each nucleotide sequence is co-expressed with a different target sequence of interest. In some embodiments, the cell comprises 2, 3, 4, 5 or more nucleotide sequences each having a reporter gene and a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE) operably linked to the reporter gene, wherein the reporter gene of each nucleotide sequence is co-expressed with a different target sequence of interest. The reporter gene of each nucleotide sequence may express a different reporter.

[0057] In addition to the reporter gene and WPRE element, the nucleotide sequence may further comprise expression enhancers, isolator sequences, restriction sites, sequence encoding selection markers, and/or sequence encoding screening markers as described herein. In some embodiments, the reporter gene encodes TD-Tomato or Clover.

[0058] In some embodiments, the expression of the reporter gene is under control of the endogenous promoter of the gene of interest. The promoter (e.g., human gene promoter, mouse gene promoter) is not limited. In some embodiments, the promoter is the human tyrosine hydroxylase gene promoter (*see* Kessler, et al., Brain Res Mol Brain Res. 2003 Apr 10;112(1-2):8-23). In some embodiments, the promoter (e.g., human gene promoter) is a SLC6A3 gene promoter, an AGRP gene promoter, a POMC gene promoter, an HB9 gene promoter, a GFAP gene promoter, a SCN10A gene promoter, a SCN9A gene promoter, or a TRPV1 gene promoter.

[0059] The location within the genome of the nucleotide sequence comprising the reporter gene and WPRE element are not limited as long as the reporter gene is co-expressed with the target sequence of interest (e.g., gene of interest). In some embodiments, the target sequence of interest (e.g., gene of interest) is a region of a gene encoding for a polypeptide.

In some embodiments, the nucleotide sequence comprising the reporter gene and WPRE element is located upstream of the 5' end of a stop-codon of the target sequence of interest (e.g., gene of interest). In some embodiments, the nucleotide sequence comprising the reporter gene and WPRE element is located at the 3' end of an open reading frame of the target sequence of interest (e.g., gene of interest). In some embodiments, the nucleotide sequence comprising the reporter gene and WPRE element is located at the 3' end of an open reading frame of the target sequence of interest (e.g., gene of interest) and at the 5' end of the stop-codon of the target sequence of interest (e.g., gene of interest). In some embodiments, the nucleotide sequence comprising the reporter gene and WPRE element is located within or adjacent to a human gene (e.g., tyrosine hydroxylase, SLC6A3, AGRP, POMC, HB9, GFAP, SCN10A, SCN9A, or TRPV1). In some embodiments, the nucleotide sequence comprising the reporter gene and WPRE element is located upstream of the 5' end of the human gene (e.g., tyrosine hydroxylase, SLC6A3, AGRP, POMC, HB9, GFAP, SCN10A, SCN9A, or TRPV1) stop-codon. In some embodiments, the nucleotide sequence comprising the reporter gene and WPRE element is located upstream of the 5' end of the stop-codon human of a gene (e.g., tyrosine hydroxylase, SLC6A3, AGRP, POMC, HB9, GFAP, SCN10A, SCN9A, or TRPV1) and downstream of the 3' end of the open-reading frame of the human gene (e.g., tyrosine hydroxylase, SLC6A3, AGRP, POMC, HB9, GFAP, SCN10A, SCN9A, or TRPV1). In some embodiments, the nucleotide sequence comprising the reporter gene and WPRE element is located in exon 14 of the human gene (e.g., tyrosine hydroxylase, SLC6A3, AGRP, POMC, HB9, GFAP, SCN10A, SCN9A, or TRPV1).

[0060] *Methods of generating a cell co-expressing a reporter gene and a target sequence of interest*

[0061] Another embodiment of the invention is directed towards a method of generating a cell that co-expresses a reporter gene with a target sequence of interest (e.g., gene of interest). The method may comprise providing a nucleic acid targeting vector as disclosed herein and providing a cell with target sequences homologous to the homology arms of the targeting vector, introducing the targeting vector into the cell; and maintaining the cell under conditions appropriate for integration of the reporter gene and WPRE into the genome of the cell such that the reporter gene is co-expressed with the target sequence (e.g., gene) of interest.

[0062] The cell may be non-naturally occurring or naturally occurring. The cell may be any cell type disclosed herein. In some embodiments, the cell may be a pluripotent stem cell or an induced pluripotent stem cell. In some embodiments, the cell may be a human pluripotent stem cell or a human induced pluripotent stem cell. In some embodiments, the cell is an embryonic stem cell (e.g., human embryonic stem cell).

[0063] The sequences of the 3' homology arm and 5' homology arm are not limited. The homology arms may be any homology arm described herein.

[0064] In some embodiments, at least one of the homology arms is homologous to a region of a genetic locus of interest (e.g., gene of interest). The genetic locus of interest (e.g., gene of interest) is not limited. The genetic locus of interest (e.g., gene of interest) may be any gene disclosed herein. In some embodiments, both homology arms are homologous to regions of a genetic locus of interest (e.g., gene of interest). In some embodiments, the 5' homology arm is homologous to a region of a genetic locus of interest (e.g., gene of interest). In some embodiments, the 5' homology arm is homologous to a region of genetic locus of interest (e.g., gene of interest) upstream of, proximate to, or adjacent to a stop codon. In

some embodiments, a portion of the 3' homology arm is homologous to the stop codon or a portion of the stop codon. In some embodiments, the 5' homology arm is homologous to a region of a genetic locus of interest (e.g., gene of interest) adjacent to and upstream of the stop codon and the 3' homology arm is homologous to a region of the genetic locus of interest (e.g., gene of interest) contiguous with the region homologous to the 5' homology arm and including the stop codon.

[0065] The step of introducing the target vector into the cell is not limited and may be performed by any method known in the art. Suitable techniques include calcium phosphate or lipid-mediated transfection, electroporation, and transduction or infection using a viral vector. In some embodiments, the electroporation is via Nucleofector™ Technology (Lonza Group, Basel, Switzerland).

[0066] The step of maintaining the cell under conditions appropriate for integration of the reporter gene and WPRE into the genome of the cell is not limited and may be performed by any method known in the art. In some embodiments, the conditions comprise providing the cell with a targetable nuclease to generate a DNA break at a target site and incorporating the reporter gene and WPRE into the genome of the cell by homology directed repair (HDR).

[0067] Targetable nucleases (e.g., site specific nucleases) generate DNA breaks in the genome at a selected target site and can be used to produce precise genomic modifications. DNA breaks, e.g., double-stranded DNA breaks, can be repaired by various DNA repair pathways. Homologous recombination (HR) mediated repair (also termed homology-directed repair (HDR)) uses homologous donor DNA as a template to repair the break. If the sequence of the donor DNA differs from the genomic sequence, this process leads to the introduction of sequence changes into the genome. Precise modifications to the genome can be made by providing donor DNA comprising an appropriate sequence. Modifications that can be generated using targetable nucleases include insertions, deletions, or substitutions of

one or more nucleotides, or introducing an exogenous DNA segment such as an expression cassette (a nucleic acid comprising a sequence to be expressed and appropriate expression control elements, such as a promoter, to cause the sequence to be expressed in a cell) or tag at a selected location in the genome.

[0068] There are currently four main types of targetable nuclease in use: zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and RNA-guided nucleases (RGNs) such as the Cas proteins of the CRISPR/Cas Type II system, and engineered meganucleases. ZFNs and TALENs comprise the nuclease domain of the restriction enzyme *FokI* (or an engineered variant thereof) fused to a site-specific DNA binding domain (DBD) that is appropriately designed to target the protein to a selected DNA sequence. In the case of ZFNs, the DNA binding domain comprises a zinc finger DBD. In the case of TALENs, the site-specific DBD is designed based on the DNA recognition code employed by transcription activator–like effectors (TALEs), a family of site-specific DNA binding proteins found in plant-pathogenic bacteria such as *Xanthomonas* species. The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) Type II system is a bacterial adaptive immune system that has been modified for use as an RNA-guided endonuclease technology for genome engineering. The bacterial system comprises two endogenous bacterial RNAs called crRNA and tracrRNA and a CRISPR-associated (Cas) nuclease, e.g., Cas9. The tracrRNA has partial complementarity to the crRNA and forms a complex with it. The Cas protein is guided to the target sequence by the crRNA/tracrRNA complex, which forms a RNA/DNA hybrid between the crRNA sequence and the homologous sequence in the target. For use in genome modification, the crRNA and tracrRNA components are often combined into a single chimeric guide RNA (sgRNA or gRNA) in which the targeting specificity of the crRNA and the properties of the tracrRNA are combined into a single transcript that localizes the Cas protein to the target sequence so

that the Cas protein can cleave the DNA. The sgRNA often comprises an approximately 20 nucleotide guide sequence complementary to the desired target sequence followed by about 80 nt of hybrid crRNA/tracrRNA. One of ordinary skill in the art appreciates that the guide RNA need not be perfectly complementary to the target sequence. For example, in some embodiments it may have one or two mismatches.

[0069] In some embodiments, one or more guide sequences (e.g., guide RNA) is a naturally occurring RNA sequence, a modified RNA sequence (e.g., a RNA sequence comprising one or more modified bases), a synthetic RNA sequence, or a combination thereof. As used herein a "modified RNA" is an RNA comprising one or more modifications (e.g., RNA comprising one or more non-standard and/or non-naturally occurring bases and/or modifications to the backbone, internucleoside linkage(s) and/or sugar). Methods of modifying bases of RNA are well known in the art. Examples of such modified bases include those contained in the nucleosides 5-methylcytidine (5mC), pseudouridine (Ψ), 5-methyluridine, 2'-O-methyluridine, 2-thiouridine, N-6 methyladenosine, hypoxanthine, dihydrouridine (D), inosine (I), and 7-methylguanosine (m7G). It should be noted that any number of bases, sugars, or backbone linkages in a RNA sequence can be modified in various embodiments. It should further be understood that combinations of different modifications may be used. In some embodiments an RNA comprises one or more modifications selected from: phosphorothioate, 2'-OMe, 2'-F, 2'-constrained ethyl (2'-cEt), 2'-OMe 3' phosphorothioate (MS), and 2'-OMe 3-thioPACE (MSP) modifications. In some embodiments a modification may stabilize the RNA and/or increase its binding affinity to a complementary sequence.

[0070] In some embodiments, the one or more guide sequences comprise at least one locked nucleic acid (LNA) unit, such as 1, 2, 3, 4, 5, 6, 7, or 8 LNA units, such as from about 3-7 or 4-8 LNA units, or 3, 4, 5, 6 or 7 LNA units. In some embodiments, all the

nucleotides of the one or more guide sequences are LNA. In some embodiments, the one or more guide sequences may comprise both beta-D-oxy-LNA, and one or more of the following LNA units: thio-LNA, amino-LNA, oxy-LNA, and/or ENA in either the beta-D or alpha-L configurations or combinations thereof. In some embodiments all LNA cytosine units are 5'methyl-cytosine.

[0071] In some aspects, the one or more guide sequences is a morpholino. Morpholinos are typically synthetic molecules, of about 25 bases in length and bind to complementary sequences of RNA by standard nucleic acid base-pairing. Morpholinos have standard nucleic acid bases, but those bases are bound to morpholine rings instead of deoxyribose rings and are linked through phosphorodiamidate groups instead of phosphates.

[0072] In some embodiments, a guide sequence can vary in length from about 8 base pairs (bp) to about 200 bp. In some embodiments, each of one or more guide sequences can be about 9 to about 190 bp; about 10 to about 150 bp; about 15 to about 120 bp; about 20 to about 100 bp; about 30 to about 90 bp; about 40 to about 80 bp; about 50 to about 70 bp in length.

[0073] Chemical modifications and methods of synthesizing guide RNAs (guide sequences) are known in the art. See WO/2016/164356, herein incorporated by reference in its entirety.

[0074] The portion of each genomic sequence (e.g., target sequence of interest, gene of interest) to which each guide sequence is complementary or homologous to can also vary in size. In particular aspects, the portion of each genomic sequence to which the guide sequence is complementary or homologous to can be about 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 81,

92, 93, 94, 95, 96, 97, 98, or 100 nucleotides (contiguous nucleotides) in length. In some embodiments, each guide sequence can be at least about 70%, 75%, 80%, 85%, 90%, 95%, 100%, etc. identical, complementary or similar to the portion of each genomic sequence. In some embodiments, each guide sequence is completely or partially identical, complementary or similar to each genomic sequence. For example, each guide sequence can differ from perfect complementarity or homology to the portion of the genomic sequence by about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, etc. nucleotides. In some embodiments, one or more guide sequences are perfectly complementary or homologous (100%) across at least about 10 to about 25 (e.g., about 20) nucleotides of the genomic sequence.

[0075] The genomic target sequence (e.g., genomic locus of interest, gene of interest, target sequence of interest) should also be immediately followed by a Protospacer Adjacent Motif (PAM) sequence. The PAM sequence is present in the DNA target sequence but not in an guide sequence. The Cas protein will be directed to any DNA sequence with the correct target sequence followed by the PAM sequence. The PAM sequence varies depending on the species of bacteria from which the Cas protein was derived. In some embodiments, the targetable nuclease comprises a Cas9 protein. For example, Cas9 from *Streptococcus pyogenes* (Sp), *Neisseria meningitidis*, *Staphylococcus aureus*, *Streptococcus thermophiles*, or *Treponema denticola* may be used. The PAM sequences for these Cas9 proteins are NGG, NNNNGATT, NNAGAA, NAAAAC, respectively. A number of engineered variants of the site-specific nucleases have been developed and may be used in certain embodiments. For example, engineered variants of Cas9 and FokI are known in the art. Furthermore, it will be understood that a biologically active fragment or variant can be used. Other variations include the use of hybrid targetable nucleases. For example, in CRISPR RNA-guided FokI nucleases (RFNs) the FokI nuclease domain is fused to the amino-terminal end of a

catalytically inactive Cas9 protein (dCas9) protein. RFNs act as dimers and utilize two guide RNAs (Tsai, QS, et al., *Nat Biotechnol.* 2014; 32(6): 569– 576). Site-specific nucleases that produce a single-stranded DNA break are also of use for genome editing. Such nucleases, sometimes termed “nickases” can be generated by introducing a mutation (e.g., an alanine substitution) at key catalytic residues in one of the two nuclease domains of a targetable nuclease that comprises two nuclease domains (such as ZFNs, TALENs, and Cas proteins). Examples of such mutations include D10A, N863A, and H840A in SpCas9 or at homologous positions in other Cas9 proteins. A nick can stimulate HDR at low efficiency in some cell types. Two nickases, targeted to a pair of sequences that are near each other and on opposite strands can create a single-stranded break on each strand (“ double nicking”), effectively generating a DSB, which can be repaired by HDR using a donor DNA template (Ran, F. A. et al. *Cell* 154, 1380-1389 (2013)).

[0076] The term “ donor nucleic acid” or “donor” refers to an exogenous nucleic acid segment that, when provided to a cell, e.g., along with a targetable nuclease, can be used as a template for DNA repair by homologous recombination and thereby cause site-specific genome modification (sometimes termed “ genome editing”). The modifications can include insertions, deletions, or substitutions of one or more nucleotides, or introducing an exogenous DNA segment such as an expression cassette or tag at a selected location in the genome. A donor nucleic acid typically comprises sequences that have homology to the region of the genome at which the genomic modification is to be made. The donor may contain one or more single base changes, insertions, deletions, or other alterations with respect to the genomic sequence, so long as it has sufficient homology to allow for homology-directed repair. In the present invention, the donor nucleic acid is the nucleic acid sequence comprising the reporter gene and WPRE flanked by the homology arms. The homology arms are homologous to genomic sequences flanking a location in genomic DNA at which the

insertion is to be made (e.g., DNA break). One of ordinary skill in the art also appreciates that the homology need not extend all the way to the DNA break. For example, in some embodiments the homology begins no more than 100bp away from the break, e.g., between 1 and 100bp away, e.g., 1 – 50 bp away, e.g., 1-15 bp away, from the break.

[0077] Donor nucleic acid can be provided, for example, in the form of DNA plasmids, PCR products, or chemically synthesized oligonucleotides, and may be double-stranded or single-stranded in various embodiments. The size of the donor nucleic acid can vary from as small as about 40 base pairs (bp) to about 10 kilobases (kb), or more. In some embodiments the donor nucleic acid is between about 1 kb and about 5 kb long.

[0078] Those of ordinary skill in the art are aware of methods for performing site-specific genome modification using targetable nucleases and will be able to apply such methods to introduce a nucleotide sequence comprising a reporter gene and WPRE into the genome at a location of choice. Those of ordinary skill in the art can, for example, design appropriate guide RNAs, TALENs, or ZFNs to generate a DNA break at a selected location in the genome, can design a targeting vector (e.g., comprising homology arms) to promote HDR at a DNA break generated by a targetable nuclease, and are aware of appropriate methods that can be used to introduce a targetable nuclease into cells and, where appropriate, a donor nucleic acid, and/or guide RNA. A targetable nuclease may be targeted to a unique site in the genome of a mammalian cell by appropriate design of the nuclease or guide RNA. A nuclease or guide RNA may be introduced into cells by introducing a nucleic acid that encodes it into the cell. Standard methods such as plasmid DNA transfection, viral vector delivery, transfection with synthetic mRNA (e.g., capped, polyadenylated mRNA), or microinjection can be used. If DNA encoding the nuclease or guide RNA is introduced, the coding sequences should be operably linked to appropriate regulatory elements for expression, such as a promoter and termination signal. In some embodiments a sequence

encoding a guide RNA is operably linked to an RNA polymerase III promoter such as U6 or tRNA promoter. In some embodiments one or more guide RNAs and Cas protein coding sequences are transcribed from the same nucleic acid (e.g., plasmid). In some embodiments multiple guide RNAs are transcribed from the same plasmid or from different plasmids or are otherwise introduced into the cell. The multiple guide RNAs may direct Cas9 to different target sequences in the genome, allowing for multiplexed genome editing. In some embodiments a nuclease protein (e.g., Cas9) may comprise or be modified to comprise a nuclear localization signal (e.g., SV40 NLS). A nuclease protein may be introduced into cells, e.g., using protein transduction. Nuclease proteins, guide RNAs, or both, may be introduced using microinjection. Methods of using targetable nucleases, e.g., to perform genome editing, are described in numerous publications, such as *Methods in Enzymology*, Doudna JA, Sontheimer EJ. (eds), The use of CRISPR/Cas9, ZFNs, and TALENs in generating site-specific genome alterations. *Methods Enzymol.* 2014, Vol. 546 (Elsevier); Carroll, D., Genome Editing with Targetable Nucleases, *Annu. Rev. Biochem.* 2014. 83:409–39, and references in either of these. See also U.S. Pat. Pub. Nos. 20140068797, 20140186919, 20140170753 and/or PCT/US2014/034387 (WO/2014/172470). Each of these references is incorporated by reference in its entirety.

[0079] In some embodiments of the invention, clustered regularly interspaced short palindromic repeats-associated (Cas) protein and from one to two ribonucleic acid guide sequences (gRNAs) are present in the cell and the gRNAs direct Cas protein to create a double stranded break in a region between the regions homologous to the 5' homology arm and the 3' homology arm. The reporter gene and WPRE are then integrated into the genome of the cell by homology directed repair. In some embodiments, the gRNA sequences do not hybridize with the targeting vector or the genome after integration of the nucleic acid comprising the reporter gene and WPRE.

[0080] In some embodiments, the target polynucleotide sequence is cleaved such that a double-strand break results. In some embodiments, more than one target polynucleotide sequence is cleaved such that a double-strand break results.

[0081] In some embodiments, the method comprises selecting cells with homologous recombination events over non-homologous recombination events via an enrichment step. The enrichment step is not limited. At least two enrichment methods have been developed: the positive-negative selection (PNS) method and the "promoterless" selection method. Briefly, PNS, the first method, is in genetic terms a negative selection: it selects against recombination at the incorrect (non-homologous) loci by relying on the use of a negatively selectable gene that is placed on the flanks of a targeting vector. On the other hand, the second method, the "promoterless" selection, is a positive selection in genetic terms: it selects for recombination at the correct (homologous) locus by relying on the use of a positively selectable gene whose expression is made conditional on recombination at the homologous target site. See, e.g., Mortensen R., *Curr Protoc Mol Biol*. Chapter 23 :Unit 23.1, 2006 for description of mammalian gene targeting in the context of mouse cells. See also Waldman T, et al., *Human somatic cell gene targeting*. *Curr Protoc Mol Biol*. Chapter 9:Unit 9.15, 2003; and Rago C., *Nat Protoc*. 2(11):2734-46, 2007. See, e.g., Irion, S. et al., *Nat. Biotechnol.* 25, 1477–1482 (2007); Costa, M. et al., *Nat. Protoc.* 2, 792–796 (2007); Suzuki, K. et al., *Proc. Natl. Acad. Sci. USA* 105, 13781–13786 (2008); and Zwaka, T.P. & Thomson, J.A., *Nat. Biotechnol.* 21, 319–321 (2003) for examples of gene targeting in hESCs.

[0082] See, e.g., PCT/US2003/009081 (WO/2003/080809); Urnov, F.D. et al., *Nature* 435, 646–651 (2005); Carroll, D., *Gene Ther.* 15, 1463–1468 (2008); Moehle, E.A. et al., *Proc. Natl. Acad. Sci. USA* 104, 3055–3060 (2007); and Lombardo, A. et al., *Nat. Biotechnol.* 25, 1298–1306 (2007). In some embodiments zinc finger DNA-binding domains with alterations in at least one zinc coordinating residue, such as CCHC zinc fingers. See,

e.g., PCT/US2007/025455 (WO/2008/076290). Each of these references is incorporated by reference in its entirety.

[0083] In some embodiments of the invention, a cell (e.g., a human embryonic stem cell, a human induced pluripotent stem cell) is provided having Cas9 and a guide RNA homologous to a target sequence of a genomic region encoding a protein of interest under conditions such that Cas9 cleaves the genomic region. A targeting vector as disclosed herein is introduced to the cell, wherein the targeting vector comprises a reporter gene, a WPRE element, a 5' homology arm and a 3' homology arm and wherein one of the homology arms is homologous to a region on one side of the cleavage site of the Cas9 and the other homology arm is homologous to a region on the other side of the cleavage site of the Cas9. The reporter gene and WPRE are integrated into the genome of the cell by homologous recombination upstream of the stop codon of the nucleotide sequence encoding the protein of interest. Correct orientation of the inserted reporter gene and WPRE are confirmed by checking the length of a PCR product from PCR with primers to a region of the inserted sequence and a genomic region.

[0084] *Method of Making a Targeting Vector*

[0085] Another embodiment of the invention is directed towards a method of making a targeting vector for integrating a reporter gene and a WPRE in a cell wherein the reporter gene is co-expressed with the target sequence of interest (e.g., gene of interest) comprising: providing a vector comprising, in the 5' to 3' direction, a reporter gene and an expression enhancer comprising a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE), wherein the expression enhancer and the reporter gene are operably linked; incorporating upstream of the reporter gene a 5' homology arm homologous to a first target sequence; and incorporating downstream of the WPRE a 3' homology arm homologous to a second target sequence; wherein the second target sequence is located downstream of the first

target sequence and wherein the first target sequence and the second target sequence flank at least a portion of the target sequence of interest (e.g., gene of interest).

[0086] In some embodiments, the method further comprises making a vector comprising, in the 5' to 3' direction, a reporter gene and an expression enhancer comprising a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE), wherein the expression enhancer and the reporter gene are operably linked. In some embodiments, a nucleotide sequence coding for a self-cleaving peptide as described herein or an IPER element may also be incorporated upstream of the reporter gene. The vector may include an origin of replication and may include one or more selectable marker genes. The vector may be any appropriate vector as described herein. The vector may be created by any technique known in the art and is not limited.

[0087] In other embodiments, the method comprises providing a vector comprising, in the 5' to 3' direction, a reporter gene and an expression enhancer comprising a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE), wherein the expression enhancer and the reporter gene are operably linked. The vector is any suitable vector described herein. The vector may include a nucleotide sequence coding for a self-cleaving peptide as described herein or an IPER element upstream of the reporter gene. The vector may also include an origin of replication and may include one or more selectable marker genes. In some embodiments, the vector includes a cassette encoding a reporter gene and/or a selectable marker.

[0088] The terms “co-expressed,” “gene of interest,” “expression enhancer,” and “operably linked” are defined as described herein.

[0089] The step of incorporating the homology arms into the vector are by any suitable method known in the art and not limited. In some embodiments, G-Block Gibson assembly is utilized to add one or both of the homology arms. G-block Gibson assembly can

be performed via the method described in Gibson, et al. (2009) “Enzymatic assembly of DNA molecules up to several hundred kilobases,” Nature Methods, 6(5):343–345. In some embodiments, the vector is digested with a restriction enzyme at a desired location. A double stranded nucleotide sequence comprising the homology arm and about 18-40 bp ends having the same sequence as a cut ends of the vector is provided and both the vector and double stranded nucleotide sequence are subject to 5’ exonuclease digestion. The resulting single stranded ends of the homology arm vector are annealed and DNA polymerase is utilized to fill in any missing sequence. Ligase then covalently joins the DNA of adjacent segments, removing any nicks in the DNA. In some embodiments, the nucleotide sequences present at either side of the vector restriction site for making the overlapping sequence on the homology arm are shown in Table 1:

HR120-p2A-TD-Tomato:	NheI and BamHI
GACGTTGTAAAACGACGGCCAGTGAATTCAGCTAG (SEQ ID NO: 3)	5 Prime 1 st overhang (NheI digest)
GCTAGCGGAAGCGGAGCTACTAACTTCAGCCTGTTGA (SEQ ID NO: 4)	5 Prime 2 nd overhang (NheI digest)
CACGTAAGTAGAACATGAAATAACCTAGATCGGATC (SEQ ID NO: 5)	3 Prime 1 st overhang (BamHI digest)
GATCCCCGTCGACTGCATGCAAGCTTGGCGTAATC (SEQ ID NO: 6)	3 Prime 2 nd overhang (BamHI digest)
HR120-Fusion Clover:	EcoRI and BamHI
TTCCAGTCACGACGTTGTAAAACGACGGCCAGTG (SEQ ID NO: 7)	5 Prime 1 st overhang (EcoRI digest)
CTCATGGTGAGCAAGGGCGAGGAGCTGTTACCGGGG (SEQ ID NO: 8)	5 Prime 2 nd overhang (EcoRI digest)
TATCACGTAAGTAGAACATGAAATAACCTAGATCG (SEQ ID NO: 9)	3 Prime 1 st overhang (BamHI digest)
GATCCCCGTCGACTGCATGCAAGCTTGGCGTA (SEQ ID NO: 10)	3 Prime 2 nd overhang (BamHI digest)

[0090] In some embodiments, both double stranded nucleotide sequence comprising the 5’ homology arm and the double stranded nucleotide sequence comprising the 3’ homology arm are incorporated by Gibson assembly. In some embodiments, the vector is digested with a first restriction enzyme and the double stranded nucleotide sequence

comprising the 5' homology arm is incorporated. Then the vector is digested with a second restriction enzyme and the double stranded nucleotide sequence comprising the 3' homology arm is incorporated. In other embodiments, the vector is digested with a first restriction enzyme and the double stranded nucleotide sequence comprising the 3' homology arm is incorporated. Then the vector is digested with a second restriction enzyme and the double stranded nucleotide sequence comprising the 5' homology arm is incorporated. In some embodiments, the first and second restriction enzymes are selected from NheI, BamHI, and EcoRI but the restriction enzyme is not limited.

[0091] As used herein, flanking at least a portion of a gene of interest is intended to mean that at least a portion of the 5' homology arm is homologous to a portion of the gene of interest. In some embodiments, the 5' homology arm is homologous to a portion of the gene of interest that is upstream of the stop codon. In some embodiments, the 5' homology arm is homologous to a region of the gene of interest comprising a 3' end of the last exon of the gene of interest and not comprising a stop codon. In some embodiments, the 5' homology arm is homologous to a region of the gene of interest immediately upstream of a stop codon and the 3' homology arm is homologous to a region comprising the stop codon. In some embodiments, the homology arms have sequences to enable insertion of the reporter gene and expression enhancer immediately after the final exon of the gene of interest and prior to the stop codon to enable co-expression of the entire gene of interest followed by expression of the reporter gene.

[0092] The articles "a" and "an" as used herein in the specification and in the claims, unless clearly indicated to the contrary, should be understood to include the plural referents. Claims or descriptions that include "or" between one or more members of a group are considered satisfied if one, more than one, or all of the group members are present in, employed in, or otherwise relevant to a given product or process unless indicated to the

contrary or otherwise evident from the context. The invention includes embodiments in which exactly one member of the group is present in, employed in, or otherwise relevant to a given product or process. The invention also includes embodiments in which more than one, or all of the group members are present in, employed in, or otherwise relevant to a given product or process. Furthermore, it is to be understood that the invention provides all variations, combinations, and permutations in which one or more limitations, elements, clauses, descriptive terms, etc., from one or more of the listed claims is introduced into another claim dependent on the same base claim (or, as relevant, any other claim) unless otherwise indicated or unless it would be evident to one of ordinary skill in the art that a contradiction or inconsistency would arise. It is contemplated that all embodiments described herein are applicable to all different aspects of the invention where appropriate. It is also contemplated that any of the embodiments or aspects can be freely combined with one or more other such embodiments or aspects whenever appropriate. Where elements are presented as lists, e.g., in Markush group or similar format, it is to be understood that each subgroup of the elements is also disclosed, and any element(s) can be removed from the group. It should be understood that, in general, where the invention, or aspects of the invention, is/are referred to as comprising particular elements, features, etc., certain embodiments of the invention or aspects of the invention consist, or consist essentially of, such elements, features, etc. For purposes of simplicity those embodiments have not in every case been specifically set forth in so many words herein. It should also be understood that any embodiment or aspect of the invention can be explicitly excluded from the claims, regardless of whether the specific exclusion is recited in the specification. For example, any one or more nucleic acids, polypeptides, cells, species or types of organism, disorders, subjects, or combinations thereof, can be excluded.

[0093] Where the claims or description relate to a composition of matter, e.g., a nucleic acid, polypeptide, cell, or non-human transgenic animal, it is to be understood that methods of making or using the composition of matter according to any of the methods disclosed herein, and methods of using the composition of matter for any of the purposes disclosed herein are aspects of the invention, unless otherwise indicated or unless it would be evident to one of ordinary skill in the art that a contradiction or inconsistency would arise. Where the claims or description relate to a method, e.g., it is to be understood that methods of making compositions useful for performing the method, and products produced according to the method, are aspects of the invention, unless otherwise indicated or unless it would be evident to one of ordinary skill in the art that a contradiction or inconsistency would arise.

[0094] Where ranges are given herein, the invention includes embodiments in which the endpoints are included, embodiments in which both endpoints are excluded, and embodiments in which one endpoint is included and the other is excluded. It should be assumed that both endpoints are included unless indicated otherwise. Furthermore, it is to be understood that unless otherwise indicated or otherwise evident from the context and understanding of one of ordinary skill in the art, values that are expressed as ranges can assume any specific value or subrange within the stated ranges in different embodiments of the invention, to the tenth of the unit of the lower limit of the range, unless the context clearly dictates otherwise. It is also understood that where a series of numerical values is stated herein, the invention includes embodiments that relate analogously to any intervening value or range defined by any two values in the series, and that the lowest value may be taken as a minimum and the greatest value may be taken as a maximum. Numerical values, as used herein, include values expressed as percentages. For any embodiment of the invention in which a numerical value is prefaced by “about” or “approximately”, the invention includes an embodiment in which the exact value is recited. For any embodiment of the invention in

which a numerical value is not prefaced by “about” or “approximately”, the invention includes an embodiment in which the value is prefaced by “about” or “approximately”. “Approximately” or “about” generally includes numbers that fall within a range of 1% or in some embodiments within a range of 5% of a number or in some embodiments within a range of 10% of a number in either direction (greater than or less than the number) unless otherwise stated or otherwise evident from the context (except where such number would impermissibly exceed 100% of a possible value). It should be understood that, unless clearly indicated to the contrary, in any methods claimed herein that include more than one act, the order of the acts of the method is not necessarily limited to the order in which the acts of the method are recited, but the invention includes embodiments in which the order is so limited. It should also be understood that unless otherwise indicated or evident from the context, any product or composition described herein may be considered “isolated”.

[0095] Specific examples of these methods are set forth below in the Examples.

EXAMPLES

Example 1

[0096] The creation of the vectors and the targeting workflow are described schematically in FIG. 1.

[0097] The final vectors were derived through modifications of the commercially available HR120-PA-1 vector (www.systembio.com/genome-engineering-precisionx-HR-vectors/gene-tagging). To create a new multicistronic vector, the copGFP-polyA cassette was removed using EcoRI and NruI. Vector sequence was restored via G-Block cloning which introduced a P2A cassette. The following GBlock sequences were used:

[0098] **5Prime Gblock**

[0099] GACGTTGTAAAACGACGGCCAGTGAATTCAGCTAGTTCCTGGAGGA
GGCCCAGTGGAGGTTTCAGGGAGGGATGGGGTGGCCGGCAGTCTCTAGTGGA
GGCGCCTAGCCTATCTCCCCCATGAACCCCCTCACCCAGCCCTGGAAGAGGCCTC
AGTGTCCCGCCTGTGACCAGTTGGCTCAGAAAAGCCCTGGGAGCTCTGAGCCACT
GTGAAGGTGGAAACGCGGCCCTGGCCTCCCCTCTCCTGGAGGCTGCAGACTCTG
CCCGCCAGTTGACGAGGGCTCTGCCGCTCTCCTCCCCAGGAGCTATGCCTCACGC
ATCCAGCGCCCCTTCTCCGTGAAGTTCGACCCGTACACGCTGGCCATCGACGTGC
TGGACAGCCCCCAGGCCGTGCGGCGCTCCCTGGAGGGTGTCCAGGATGAGCTGG
ACACCCTTGCCCATGCGCTGAGTGCCATTGGCGCTAGCGGAAGCGGAGCTACTA
ACTTCAGCCTGTTGA (SEQ ID NO: 11)

[0100] **3Prime Gblock**

[0101] CACGTAAGTAGAACATGAAATAACCTAGATCGGATCGTGACGGC
GTCCCTGAGGGCCCTTCCCAACCTCCCCTGGTCCTGCACTGTCCCGGAGCTCAGG
CCCTGGTGAGGGGCTGGGTCCCGGGTGCCCCCATGCCCTCCCTGCTGCCAGGCT
CCCACTGCCCCTGCACCTGCTTCTCAGCGCAACAGCTGTGTGTGCCCCGTGGTGAG
GTTGTGCTGCCTGTGGTGAGGTCCTGTCCTGGCTCCCAGGGTCCTGGGGGCTGCT
GCACTGCCCTCCGCCCTTCCCTGACACTGTCTGCTGCCCCAATCACCGTCACAAT
AAAAGAAACTGTGGTCTCTACACCTGCCTGGCCCCACATCTGTGCCACAGAGACA
GACCCTGGGATCCTCAGACTCCCACACCCCCACCCCAGCCTCACTCAGAGGTTTC
GCCCTGGCCTCCTTCCTCCTCTGGGAGATGGCTGGATCCCCGTCGACTGCATGCA
AGCTTGGCGTAATC (SEQ ID NO: 12)

[0102] In subsequent steps, the XhoI site was cleaved via restriction digest and Gibson assembly was used to introduce the fluorophore CDS followed by the WPRE element. For example, the TD-tomato CDS followed by the WPRE element was amplified using pFUGW-TD-Tomato as a template. The PCR product was inserted via Gibson cloning.

The P2A-TD-Tomato-WPRE cassette can be excised through EcoRI restriction digest. Subsequently, Gibson assembly was used to introduce the fluorescent protein CDS followed by the WPRE element. For example, the Clover CDS followed by the WPRE element was amplified using pFUGW-Clover as a template. The PCR product was inserted via Gibson assembly. In subsequent steps, homology arms for any given gene can be added. To add 5' homology arms to the HR120-p2A-TD-TOM, the vector was cut with the restriction endonuclease NheI. A homology arm can be created either by PCR or DNA synthesis utilizing 18-40 bp long overlap sequences between the vector and insert. A template for the DNA generation is given below. Vector and homology arm are enzymatically assembled using Gibson reaction. In a subsequent step, BamHI restriction endonuclease linearize the vector to add a 3' homology arm. For the 5' homology addition in plasmid pHR120-clover, EcoRI for the 5' arm and BamHI for the 3' arm were used. The 5' homology arm ends before the stop codon of the gene of interest to allow for a fusion protein or multicistronic expression. The necessary overhangs are shown in detail in Table 2. Further CRISPRs can be designed with overlap in both the 5' and 3' arms to avoid cleavage of the homology construct during the targeting.

Table 2: Nucleotide sequence that can be used to create overlapping sequence for Gibson assembly

HR120-p2A-TD-Tomato:	NheI and BamHI
GACGTTGTAAAACGACGGCCAGTGAATTCAGCTAG (SEQ ID NO: 3)	5 Prime 1 st overhang (NheI digest)
GCTAGCGGAAGCGGAGCTACTAACTTCAGCCTGTTGA (SEQ ID NO: 4)	5 Prime 2 nd overhang (NheI digest)
CACGTAAGTAGAACATGAAATAACCTAGATCGGATC (SEQ ID NO: 5)	3 Prime 1 st overhang (BamHI digest)
GATCCCCGTCGACTGCATGCAAGCTTGGCGTAATC (SEQ ID NO: 6)	3 Prime 2 nd overhang (BamHI digest)
HR120-Fusion Clover:	EcoRI and BamHI
TTCCAGTCACGACGTTGTAAAACGACGGCCAGTG (SEQ ID NO: 7)	5 Prime 1 st overhang (EcoRI digest)
CTCATGGTGAGCAAGGGCGAGGAGCTGTTACCGGGG (SEQ ID NO: 8)	5 Prime 2 nd overhang (EcoRI digest)
TATCACGTAAGTAGAACATGAAATAACCTAGATCG (SEQ ID NO: 9)	3 Prime 1 st overhang (BamHI)

	digest)
GATCCCCGTCGACTGCATGCAAGCTTGGCGTA (SEQ ID NO: 10)	3 Prime 2 nd overhang (BamHI digest)

[0103] **Results:**

[0104] Using this method, we created a targeting vector for Tyrosine Hydroxylase (TH). We found that about 25% to greater than 50% of Puromycin-resistant colonies had the correct integration. Excision of the vector cassette by CRE recombinase works in >90% of analyzed clones. We have also performed this method using other cell lines with a targeting efficiency of at least 50%.

[0105] Correctly targeted clones were differentiated using a protocol based on a previously published protocol (Kriks et al., 2011) (FIG. 2A). Reporter expression can be seen as early as day 7. Cells were stained in embryoid bodies (EBs) using anti-TH antibody (FIG. 2B). The number of TH positive cells increased over time in the culture, as seen in the comparison between day 7 and day 30 EBs (Figure 2B). We observed near perfect overlap between reporter expression and antibody staining. Cells could be dissociated and plated or purified via FACS sorting which led to a strong enrichment with close to 100% of the cells expression the TD-tomato reporter (FIG. 2C).

[0106] **Discussion:**

[0107] Investigators studying midbrain neurons, dopaminergic neurons and PD pathology have long sought the creation of reliable TH reporters as a key research tool. To our knowledge, the approach described here is the first successful attempt at a live cell stage homology reporter that shows strong expression. More importantly, however, this cloning technique can be applied to any gene-specific reporter construct, and thus has broad implications throughout the fields of biological and biomedical research.

[0108] **Example 2- Generation of isogenic lines**

[0109] **Targeting Experiments:**

[0110] **Nucleofection:**

[0111] hESCs mTESR + 4 μ M ROCK inhibitor (Y-27632 EMD Millipore) was incubated for at least 1 h prior to electroporation. An appropriate number of 10 cm recovery dishes using geltrex/matrigel coating (1x10 cm per reaction) were prepared. Cells were harvested using Accutase, and gently triturated to get a single cell suspension. Cells were counter, if a new cell line (HUES9 $\sim 1 \times 10^7$ cells). 1×10^6 cells were transferred into an appropriate number of conical 15 ml tubes. Cells were pelleted (1100 RPM, 3min, 25°C). Cells were resuspended using DNA nucleofection solution mix and transfer into Amaxa cuvettes. The following Table provides the required amounts of the DNA nucleofection solution mix (AMAXA 4D-NUCLEOFECTOR (Lonza # V4XP3024))

[0112]

	Knockout	Targeting Vector	ssOligo
CRISPR1	5 µg	2.5 µg	2.5 µg
CRISPR2	5 µg	2.5 µg	2.5 µg
TV	-	10 µg	-
ssOligo 25 uM	-	-	5 -10 µl
Maximum volume	50 µl	50 µl	50 µl
Amaya sol mix	100 µl	100 µl	100 µl

[0113] Cuvettes were incubated for 5 min at room temperature. Nucleofection of cells was performed using program CB150. Completed reactions were transferred onto new coated 10 cm dishes containing warm mTESR + ROCK inhibitor. The medium was changed the next day to remove ROCK inhibitor.

[0114] A: If targeting vector was used; 48 hr Puromycin selection was begun using 1 µg/ml Puromycin. Post selection cells grown to confluence and nucleofection procedure repeated using 20 ug pCAG-CRE GFP – available on the world-wide web at www.addgene.org/13776/. FACS sort performed 24 hrs. later

[0115] B: If knockout mutagenesis or ssOligo targeting was the goal; FACS sort performed 24 hrs. post nucleofection for CRISPR GFP.

[0116] **FACS Sorting:**

[0117] Cells prepared for FACS sorting 24 hours post nucleofection. hESCs mTESR + ROCK inhibitor incubation for 3~4 h prior to electroporation. Cells harvested using Accutase (1:2.5 dilution); washed with PBS; filtered through 40 micron membrane into new 50ml conical tube. Cells were then resuspended in 300ul mTESR + ROCK (FACS buffer). Cells kept on ice during FACS sort and a negative control for gating was used.

[0118] 10000-15000 sorted cells per 10 cm geltrex coated dish were plated (medium mTESR + ROCK inhibitor). Colonies grow up and can usually be picked between 9-15 days post sorting.

[0119] **Picking Colonies:**

[0120] Medium aspirated and replace with fresh medium mTESR+RhoK 1h prior to picking. Cells incubated for 5 minutes on ice. 96 well dishes (without coating) filled with 100 µl mTESR + ROCK inhibitor. Colonies picked in open hood using a dissection scope. Single colonies circled with 200 µl pipette tip. Cells scrapped off in sheets (they will usually stick to the pipette tip). Cell suspension aspirated in 50 µl volume and transferred to a pre-labeled 96 well plate (Total volume per well =150 µl). Cells picked fast and kept in 96 well plates on ice. Cells broken into small chunks by aspirating up and down using a 200 µl multichannel pipette. 100 µl (2/3) of picked colonies were transferred to geltrex coated 96 well dishes and cells were grown to confluency (about 5 days). Geltrex coated 96 well dishes were prepared by coating with geltrex coating solution: 100µl of geltrex/Matrigel in about 10 ml DMEM. 50 µl (1/3) of picked colonies were transferred into pre-labeled PCR tubes for gDNA extraction.

[0121] **5.2.4 Genomic DNA extraction and Gel analysis**

[0122] Appropriate amount of PCR cells direct lysis mix was prepared (add 50 µl Proteinase K to 1 ml of Viagen cell lysis reagent). 100 µl of lysis reagent was added to each well containing 50 µl picked colony. Plate sealed with sticky lid and placed on a rocking plate in PCR machine at 55°C for 6h followed by incubation at 85°C for 45 min and then incubation at 4°C. Lysates were stored in a refrigerator.

[0123] **PCR amplification**

[0124] Phusion Hifi mastermix Polymerase was used for PCR Amplification:

[0125] Primers designed to amplify region of interest (size 50-250 bp). gDNA concentration/quality determined through a test PCR using 1-5 µl of lysate in a 15 µl PCR reaction. For gDNA primers, NEB TM calculator for Phusion Hifi was used. The same amount of gDNA for all PCRs was used and a touchdown PCR was always performed.

[0126] PCR products were analyzed for the appearance of one single band and sequencing was performed using forward PCR primers.

[0127] Example 3

[0128] **Generation and characterization of isogenic tyrosine hydroxylase knock-in reporter cell lines carrying three distinct PD mutations**

[0129] We used CRISPR mediated knock-out mutagenesis to create isogenic PD lines from a WT donor line, HUES1 (Figure 4A-top), for three distinct PD loci: *PARKIN*, *DJ-1* and *ATP13A2* (Figure 4B and Figure 9 AD). To specifically study DANs, we used CRISPR-Cas9 genome editing to introduce a fluorescent reporter into the locus for the TH gene, the rate limiting enzyme in the synthesis of dopamine (Figure 4A bottom). Knock-in efficiency, as determined by 5' genotyping PCR, was 60% (Figure 4B). Isolated clones had a small cytoplasm to nucleus ratio and stained positive for the pluripotency markers OCT4 and TRA-160 (Figure 9E). The targeting vector was designed to retain a largely unaltered endogenous TH gene product using a bicistronic targeting vector containing tdTomato (Shaner et al., 2004) (Figure 4C and Figure 10A-D). We generated midbrain DANs in spin-culture. Our differentiation scheme is based on a modified version of the dual SMAD inhibition protocol followed by patterning by modulating sonic hedgehog and WNT signaling (Figure 5D) (Kriks et al., 2011; Valente et al., 2004).

[0130] To analyze differentiation in isogenic cell lines, we performed a time-course qPCR analysis. *OCT4* expression was decreased at day 7 and was barely detectable after day 14. We observed an initial increase in *FOXA2*, as well as *NESTIN*, which was strongest in

day 7 neural progenitor cells (NPCs). There was a significant increase of *PARKIN* mRNA relative to d0 in all differentiated cell types (Figure 11A). We found that most cells stained with an anti-TH antibody also expressed the tdTomato TH reporter (Figure 11B). TdTomato was expressed diffusely throughout the cell body and less strongly within neuronal projections, allowing it to double as a cytoplasmic marker. The TH⁺ neurons generally showed multipolar soma, elaborate dendrites, and axons (Figure 4E).

[0131] Midbrain DANs exhibit two characteristic firing patterns, single spikes and bursting (Shi, 2005). To confirm that our hPSC derived TD-tomato positive neurons are functionally consistent with their putative identification as DANs, we carried out electrophysiological measurements. Flow-sorted neurons (d35 of differentiation) from all isogenic lines were plated on mouse glial cells and cultured for up to two months. We detected voltage-gated sodium and potassium currents, evoked repetitive action potentials, and frequent spontaneous potentials in TH-RFP DA neurons derived from both wild-type and isogenic PD lines (Figure 4F). Furthermore, live-cell calcium imaging also showed spontaneous TTXsensitive activity in neurons from all the lines (Supp. Fig. 7A-C). Thus, all isogenic lines differentiated into functional DANs with electrophysiological properties similar to those described previously (Jiang et al., 2012; Kriks et al., 2011).

[0132] Loss of *PARKIN* decreases the number of TH-positive neurons

[0133] PD is characterized by the disproportionate death of midbrain DANs. However, *PARKIN* is a ubiquitously expressed protein. To validate the increased *PARKIN* expression found in qPCR experiments during differentiation, we conducted a time-course western blot analysis of *PARKIN* protein expression. The western blot confirmed an accumulation of *PARKIN* protein during differentiation in both WT as well as *ATP13A2*^{-/-} lines and the complete loss of *PARKIN* protein in our *PARKIN*^{-/-} lines at all developmental time points (Figure 5A and Figure 12D).

[0134] We hypothesized that early onset PD mutations could result in increased rate of cell death in midbrain DANs in basal culture conditions. We dissociated spheres at different times during differentiation and analyzed the population for the percentage of TH⁺ cells via fluorescence flow cytometry (Figure 5B and Figure 12E). The average percentage of TH⁺ neurons in WT, DJ-1^{-/-} and ATP13A2^{-/-} line was ~40%, indicative of no substantial defects in TH⁺ neuron accumulation in these lines. However, the TH⁺ fraction was significantly smaller (16.7%) in the PARKIN^{-/-} line (Figure 5C). A timecourse of TH⁺ neuron emergence showed no significant differences in the onset of reporter expression or the percentage of TH⁺ cells during the first 17 days of differentiation in WT and PARKIN^{-/-} lines (Figure 5D). We observed a substantial increase in both lines in the number of TH⁺ cells between day 17 and day 21. Differences between WT and PARKIN^{-/-} line became apparent after day 21.

[0135] Focusing on the critical window between day 15 and day 21, we conducted a time lapse experiment using an automated live imager (Nikon Biostation) to examine the fate of newly generated TH⁺ cells, as well as track those cells over time, with the goal of determining whether the differences were due to differentiation or survival. We first noted faint TH expression in cells that showed no obvious neuritic processes. These might be TH expressing NPCs or DANs that lost processes during the enzymatic dissociation. Over time, faintly TH:tdTomato positive cells gained in fluorescence intensity, grew processes, and exhibited a morphology typical of DANs (Figure 5E, top panel). Additionally, we observed the birth of new TH:tdTomato positive cells over time (Supplemental Video 1-no shown). Terminal differentiation was characterized by a decrease in cell body size accompanied by the growth of neuronal processes and a substantial increase in fluorescence intensity. Cells in the PARKIN^{-/-} line exhibited a greatly altered morphology. Faintly TH⁺ cells in the early stages of dopaminergic differentiation appeared larger and displayed cytoplasmic vacuoles

(Figure 5E, bottom panel). Of those, many disappeared from culture during the imaging period, displaying cellular fragmentation suggestive of apoptosis (Supplemental Video 1-not shown). Consistent with our flow cytometry data, fewer tdTomato positive cells in the PARKIN line remained at the end of the experiment.

[0136] Loss of PARKIN, DJ-1 or ATP13A2 leads to an increase of oxidative stress in TH+ neurons

[0137] Several postmortem brain analyses of PD patients, as well as of PD animal models, have implicated increased oxidative stress through mitochondrial dysfunction in DANs as a common feature of PD pathology (Dias et al., 2013). To test the role of PD mutations in the regulation of reactive oxygen species (ROS), we analyzed the levels of ROS using the live cell dye CellROX Green in all our isogenic cell lines following differentiation (Figure 5F), and quantified the percentage of cells exhibiting high green fluorescence (mROS-G⁺) cells in each population (Figure 5G). Few mROS-G⁺ events were recorded in either TH:tdTomato negative (TH⁻) and TH:tdTomato positive (TH⁺) cells in the control line: 0.62% and 1.73% respectively. In contrast, a consistent increase in the average percentage was observed in all isogenic PD lines. In the PARKIN^{-/-} line: TH⁻: 2.77% and TH⁺: 24.97%, the DJ-1^{-/-} line: TH⁻: 1.46% and TH⁺: 4.97% and the ATP13A2^{-/-} line: TH⁻: 1.77% and TH⁺: 4.95%. Consistently, the relevant cell type, TH:tdTomato positive DANs, showed significantly higher ROS accumulation than their TH:tdTomato negative counterparts. The increase was significant in all PD lines, but the PARKIN^{-/-} line showed the strongest increase, consistent with the observed cell death phenotype.

[0138] Global transcriptional analysis identifies overlapping dysregulated genes and pathways between PARKIN^{-/-} and ATP13A2^{-/-} cell lines.

[0139] Oxidative stress was a common phenotype observed in all our isogenic cell lines. However, many different disease mechanisms have been proposed to play a role in the

development of PD. We set out to identify common and distinct dysregulated genes and networks in our three isogenic PD lines in an unbiased way, with the goal of grouping similar forms of PD. To this end, we performed global transcriptional analyses using RNA-sequencing. At day 35, spheres were dissociated and DANs were purified by flow sorting for tdTomato expression. RNA-sequencing profiles were generated in three separate differentiation experiments (n=12). We observed consistent coverage depths among all samples (Figure 13A). To visualize the relationship between each individual sample in our dataset, we performed unsupervised clustering as a multidimensional scaling plot, in which the distances correspond to leading log-fold changes between each pair of RNA samples (Figure 6A). The plot revealed substantial differences between the PARKIN^{-/-} versus all other isogenic lines, which led to the separation by dimension 1. Dimension 2 separated biological replicate two from the other replicates (Figure 13B). Dimension 3 separated the ATP13A2^{-/-} line from the WT and DJ-1^{-/-} line. The PARKIN^{-/-} line is most dissimilar from all other lines.

[0140] We performed differential gene expression analysis and analyzed the results for overlap of genes and pathways. We list genes that were differentially regulated by at least 2-fold, with a p-value < 0.01 between WT and all isogenic PD lines. We conducted an overlap analysis to determine whether differentially expressed genes between WT and all isogenic disease lines are present in common pathways or are independent from one another (Figure 6B). Only six genes are differentially regulated in all three isogenic lines. 37 genes are differentially expressed between DJ-1^{-/-} and WT. However, of those 37, 24 are also differentially expressed in the PARKIN^{-/-} lines. Shared genes were dysregulated in opposite directions, suggesting disparate disease mechanisms in DJ1 lines compared to PARKIN, despite similar molecular players (Figure 6C). We found no significant dysregulated pathways in our KEGG pathway enrichment analysis of this comparison. 141 genes are

differentially expressed between the ATP13A2^{-/-} and WT lines. Of those, 64 are also differentially expressed between the PARKIN^{-/-} and WT lines. The heat map analysis demonstrates that in PARKIN^{-/-} and ATP13A2^{-/-} DANs, most expression changes have the same direction when compared to WT or DJ^{-/-} DANs (Figure 6D). To analyze the differential expression results in a network context, we performed KEGG pathway enrichment analysis on all differentially expressed genes between WT and isogenic PD lines. We found that the PARKIN^{-/-} line, and the ATP13A2^{-/-} line show similar patterns of dysregulation. The two top KEGG pathways found dysregulated in both lines are hsa04512, extracellular matrix (ECM) receptor interaction, and hsa04974, protein digestion and absorption (Figure 6E).

[0141] Evidence for altered composition of DANs in the PARKIN^{-/-} line.

[0142] In addition to the common KEGG pathways, we found several KEGG pathways with direct relevance to DANs that were significantly dysregulated in the PARKIN^{-/-} line. These pathways included hsa05032, 'Morphine addiction', hsa04726, 'Serotonergic synapse', hsa04727, 'GABAergic synapse', hsa05030, 'Cocaine addiction', and hsa04080, 'Neuroactive ligand-receptor interaction'. The midbrain contains several DAN populations and selective cell death in the PARKIN^{-/-} line could change its composition relative to that of the other lines. For example, ventral tegmental area (VTA) DANs are known to play a primary role in the reward system and addiction. We plotted genes contributing to any addiction pathways in the KEGG analysis ('path:hsa05032', 'path:hsa05031', 'path:hsa05034', 'path:hsa05033'). We visualized the contributing genes in a heatmap. The PARKIN line forms a distinct cluster compared to the other isogenic lines, showing dysregulation of many transcripts with prominent roles in dopaminergic function (Figure 13C). This could be caused by altered function of DANs or by altered composition of the surviving DANs in the PARKIN^{-/-} line.

[0143] Quantitative proteomics analysis confirms importance of differentiation status in PARKIN^{-/-} line

[0144] Western blot analysis revealed the high abundance of PARKIN in differentiated dopaminergic neurons that was not reflected at the transcriptional level. The correlation between mRNA and protein abundance is generally poor (Maier et al., 2009). Given this discrepancy, we performed comprehensive quantitative proteomic analyses to further investigate proteins that could be responsible for phenotypic differences in the PD cell lines. To profile proteome alterations quantitatively in our isogenic PD lines, we used multiplexed tandem mass tag (TMT)-based quantitative mass spectrometry. Global proteomics analysis requires substantial amounts of input material. After sorting cells for several hours, we noted obvious transcriptional changes in our RNA sequencing experiments (Figure 13A replicate 2) and increased cell death. Cell death after sorting was particularly problematic in the sensitive PARKIN^{-/-} line. Given these two problems, we decided to perform this study in mixed cultures without using FACS purification. We created datasets for the WT and PARKIN^{-/-} isogenic lines in pluripotent cells and NPCs at the end of the patterning protocol at day 12 and for the WT, PARKIN^{-/-} and DJ-1^{-/-} lines as mixed cultures at day 35 in differentiation technical triplicates. We quantified 7060 proteins in the pluripotent state, 8256 proteins in the NPC state and 7140 proteins at day 35. In an independent proteomics experiment, we quantified 8116 proteins in WT and ATP13A2 lines at day 35 in differentiation.

[0145] Differential expression analysis showed a strong increase in the number of differentially expressed proteins during the time course of differentiation in the WT versus PARKIN^{-/-} comparison (Figure 7A). We found 32 proteins that were at least 2-fold dysregulated between WT and PARKIN^{-/-} in the pluripotent state, 76 in the NPC state at day 12, and 346 in differentiated cells at day 35. We performed overlap analysis to determine if

loss of PARKIN function would lead to dysregulation of the same proteins in all three developmental time points (Figure 7A). Only the Orphan G protein-coupled receptor 50 (GPR50) was significantly upregulated in pluripotent stem cells, NPCs and day 35 DANs.

[0146] Quantitative proteomics reveals overlap in dysregulated pathways in isogenic PD lines

[0147] Given that we analyzed mixed cultures at day 35 and that we had determined a significantly lower percentage of TH⁺ cells in the PARKIN^{-/-} line, we set out to analyze the differentiation and overall composition of our spheres. We performed label-free analysis of singlet samples analyzed sequentially by liquid chromatography–mass spectrometry (LC/MS). We identified the 30 proteins that showed the highest relative abundances in all lines. The 30 proteins are comprised of highly expressed housekeeping genes, proteins with roles in mitochondria, and proteins with known functions in neuronal differentiation and neuron maintenance. We found high similarity among the isogenic cell lines. Only one protein, lactate dehydrogenase B (LDHB), showed more than a 2-fold enrichment in the PARKIN^{-/-} line (Figure 14A). Overall, this indicates a similar composition of the differentiating spheres.

[0148] In line with our transcriptomics results, the number of 2-fold differentially expressed proteins according to TMT-based quantification between the WT and DJ-1^{-/-} lines was 6-fold smaller (58 proteins) than the difference between the WT and PARKIN^{-/-} lines (346 proteins) (Table 4B). We noted that enriched proteins were three times more numerous than depleted proteins in the PARKIN^{-/-} line when compared to WT cells.

[0149] The overlap between the dysregulated proteins seen in the PARKIN^{-/-} and DJ-1^{-/-} lines was substantial (Figure 4C). Several proteins were upregulated in both isogenic PD lines, including ALDH1A1 (aldehyde dehydrogenase 1A1) which participates in the metabolism of dopamine (DA) converting 3,4- dihydroxyphenylacetaldehyde, a potentially

toxic aldehyde, to 3,4-dihydroxyphenylacetic acid, a nontoxic metabolite (Anderson et al., 2011). Heat map analysis of the 25 overlapping genes demonstrated that protein expression changes in the PARKIN^{-/-} and DJ-1^{-/-} lines are in the same direction when compared to WT cell line and the dendrogram confirms this (Figure 14B). PARKIN's role in metabolic and mitochondrial pathways is well established. The role of DJ-1 is less clearly understood, but it has been implicated in mitochondrial dysfunction and oxidative stress. To focus on mitochondrial dysfunction, we utilized data from MitoCarta2.0, a database of genes with a known mitochondrial role. We compared this list with proteins that were at least 2-fold differentially expressed between PARKIN^{-/-} and WT lines. Despite the previously established substantial overlap, there was no overlap in dysregulated mitochondrial genes between the PARKIN^{-/-} and DJ-1^{-/-} lines. The clustering and heat map shows the PARKIN^{-/-} line is distinct from WT and DJ-1^{-/-} lines (Figure 14C).

[0150] Reduced activity of complex I, the NADH dehydrogenase complex, in mitochondria has been observed in midbrain tissue of sporadic PD patients (Keeney et al., 2006). Inhibitors of complex I activity, like MPTP and rotenone, can cause dopaminergic cell death, and have been used to generate animal models of PD. Inhibition of complex I activity leads to increased generation of ROS. We found increased amounts of ROS in DA neurons in all our isogenic PD lines (Figure 5F and 5G). We analyzed the cells for mitochondrial dysfunction by focusing on important structural components in the mitochondrial matrix. Complex I is comprised of 30 NADH: ubiquinone oxidoreductase subunits (NDUF-gene symbols). We analyzed the protein levels of all quantified subunits and plotted them in a heatmap. We found that none of the subunits was dysregulated, at the cut-off 2-fold level. However, analyzing the mean fold changes we saw consistent and highly significant downregulation of complex 1 components in the human mitochondrial respiratory chain in both the PARKIN^{-/-} ($P = 1.2E-10$), as well as the DJ-1^{-/-} ($P = 1.2E-9$) lines (Figure 7D). This

is consistent with data from PD patients indicating lowered complex I activity. KEGG pathway analysis of dysregulated proteins in both PARKIN^{-/-} as well as the DJ-1^{-/-} lines also confirmed dysregulation of ECM-receptor interaction pathways that we observed in the transcriptional analysis (Figure 14D).

[0151] We performed an additional quantitative proteomics experiment using two independent differentiation replicates for the WT control and the ATP13A2 line. We found that 187 proteins were at least 2-fold differentially expressed between the two lines (Figure 7B). Using proteins detected in both proteomics experiments, we performed an overlap analysis between all WT to disease lines comparisons (Figure 7E). We did not find a set of proteins that was dysregulated in all three isogenic PD lines. However, KEGG pathway analysis of dysregulated proteins in the ATP13A2^{-/-} line revealed dysregulation of ECM-receptor interaction pathways as the top hit (Figure 14D). Transcripts and dysregulated proteins that affect ECM-receptor interaction have been indicated in a graphic display of the involved pathways (Figure 14E).

[0152] Knockout of PARKIN and ATP13A2 leads to the significant dysregulation of other PD relevant proteins.

[0153] We hypothesized that loss of function of either DJ-1 or PARKIN could lead to significant dysregulation (2-fold change) in other known PD associated proteins, either as a direct effect or through compensatory mechanisms. For example, PD relevant proteins could be PARKIN substrates and loss of PARKIN could lead to a reduction of ubiquitination and subsequent regulation or degradation. We compared proteins contained within the UCL high priority set of 48 PD relevant genes (Foulger et al., 2016). We found that loss of PARKIN leads to dysregulation of several PD relevant genes (Figure 7F). For example, alpha synuclein (SNCA) and the microtubule associated protein tau (MAPT) were more than 2-fold enriched. Several protein/peptide properties influence detection ability in our method and PARKIN was

not detected. We used isotope labeled peptides (AQUA peptides) as an internal standard to allow absolute quantification of PARKIN in those samples (Stemmann et al., 2001). We detected no PARKIN in the PARKIN^{-/-} and no significant changes of PARKIN quantities in the other tested lines WT and DJ^{-/-} (Figure 7G).

[0154] Using the same list of PD relevant proteins, we compared protein levels in the ATP13A2^{-/-} line. In this set, we could detect the proteins PARKIN, lysosomal associated membrane protein 3 (LAMP3) and transmembrane glycoprotein NMB (GPNMB). We found 1.5-fold upregulation of SNCA and significant enrichment of LAMP3 and GPNMB in the ATP13A2^{-/-} line (Figure 7H).

[0155] **Knockout of DJ-1 leads to the dysregulation of proteins involved in cell cycle as well as proteins involved in the development of Charcot-Marie-Tooth disease.**

[0156] DJ-1 is a multifunctional protein. The role it plays in the development of PD is presently unclear. We analyzed all proteins that were at least 2-fold differentially expressed between WT and the DJ-1 line (Figure 8A). We performed GO enrichment analysis using Gorilla (Eden et al., 2007; Eden et al., 2009) and used the option to run two unranked lists and used every detected protein as the background (Figure 8B). Several minichromosome maintenance complexes (MCM) family members were significantly enriched in the DJ-1^{-/-} line and the associated GO terms relate to DNA conformational changes and DNA replication. We also analyzed depleted proteins and found GO-terms related to oxidative stress, including peptide cross-linking, the regulation of reactive oxygen species, and superoxide radicals. Twenty-six proteins were significantly down-regulated in the DJ-1^{-/-} line. GO term analysis of these proteins revealed functional terms for intermediate and neurofilament bundle assembly. Four of the 26 depleted proteins are associated with the development of Charcot-Marie-Tooth disease (Figure 8C).

[0157] DJ-1 may act as a cysteine protease, and a valine-lysine-valine-alanine (VKVA) recognition sequence has been identified in target proteins (Mitsugi et al., 2013). We used the Uniprot database to create a list of peptides that contain a VKVA motif and compared this list with proteins identified in our experiments. We plotted normalized relative intensities as a heat map. None of the proteins carrying a VKVA motif was significantly dysregulated (more than 2-fold). Analyzed in compound, the almost universal upregulation of these proteins in the DJ-1^{-/-} line, was significant (pValue = 0.0013) (Figure 8D).

[0158] **Discussion:**

[0159] In this work, we have generated an isogenic *in vitro* system to study cellular disease processes in WT and mutant DANs carrying severe autosomal recessive PD mutations in either *PARKIN*, *DJ-1* or *ATP13A2*. We have knocked-in and validated a sensitive TH-reporter in all our lines. Further, we have established a robust large-scale 3D spin-culture protocol that allows for the derivation of midbrain dopaminergic neurons. Using our system, we found evidence for shared common disease pathways present in all PD lines as well as dysregulated genes and pathways unique to each specific mutation. We used this system to corroborate presumed players in PD etiology and identify new ones.

[0160] **Creation and characterization of a large-scale fluorescent reporter model:**

[0161] Several groups have tried to utilize cell surface receptors, transgenic reporters, and cell type specific knock-in reporters to analyze a specific cell type in a heterogenous population (Goulburn et al., 2011; Hockemeyer et al., 2009; Liu et al., 2011; Merkle et al., 2015; Ruby and Zheng, 2009; Xia et al., 2017). We have improved on existing protocols and developed a highly efficient targeting method to insert a bright fluorescent reporter, correctly targeting >60% of analyzed colonies. Large-scale spinner flask bioreactors have been used to generate large numbers of functionally mature differentiated cell types (Ismadi et al., 2014; Otsuji et al., 2014; Rigamonti et al., 2016). To accommodate high input technologies — like

screening assays and proteomics analyses — we adapted existing culture protocols (Kriks et al., 2011; Xia et al., 2017) to 3D conditions. The average differentiation efficiency among all lines was ~40% at day 35. DANs showed high intensity TdTomato expression and molecular and functional electrophysiological characteristics of midbrain DANs. TH:tdTomato signal enabled FACS purification and allowed for live cell tracking. Spheres were cultured for up to 75 days and there was no indication that the cultures cannot be maintained for much longer time periods. Overall our spin-culture protocol is a cost effective, highly reproducible method to generate large quantities of mDANs and allows for extended culture periods.

[0162] Observation of increased cell death in PARKIN^{-/-} mDANs highlights the interplay of cell-type and genotype

[0163] PARKIN is broadly expressed throughout the body, including heart, testis, liver and kidney, as well as brain (Kuhn et al., 2004). However, PD patients in general and those carrying loss of PARKIN mutations specifically exhibit the dysfunction and death of midbrain DANs. Several recent publications have shown that PARKIN ubiquitinates many proteins (Bingol et al., 2014; Ordureau et al., 2015; Rose et al., 2016; Sarraf et al., 2013). In our isogenic system, we have been able to study the effect of PARKIN loss on cellular proteomes at three developmental time points. Only one protein, G protein-coupled receptor 50 (GPR50), was significantly enriched at all three-time points suggesting that it is the cell-type specific environment that is critical in disease progression. Broad dysregulation of protein abundances increased from pluripotent cells to NPCs, but was strongest in the DANs, illustrating the importance of studying loss of PARKIN in the most relevant cell type.

[0164] We found that knockout mutations in the PARKIN gene resulted in a cell-type specific phenotype culminating in the selectively diminished number of TH positive neurons in basal culture conditions. This phenotype mirrored the findings described in a recent publication (Shaltouki et al., 2015), in which several PARKIN^{-/-} lines exhibited fewer TH-

positive neurons after 28 days of differentiation when compared to a WT line. Our differentiation protocol generates midbrain DANs (mDANs). However, mDANs are a heterogeneous population. In PD patients, a subtype of nigral DANs shows enhanced vulnerability, while other populations, such as VTA DANs, are much less affected. Similarly, in our cultures, some, but not all, DANs, died. Our live imaging analysis revealed highly vacuolated TH:tdTomato positive cells that exhibited large cell bodies and reduced viability during early differentiation, when compared to WT cells. It is unclear whether this correlates to any human PD pathology. However, defects in DAN differentiation may confer increased vulnerability to PD. No other isogenic PD lines showed a decreased percentage of TH positive cells in basal conditions, regardless of clinical similarity. A recent study has shown that differentiated cells derived from a DJ-1 knockout line show elevated levels of oxidative stress and stress induced cell death in response to dopamine oxidation and accumulation starting after more than 70 days in culture (Burbulla et al., 2017). We did not analyze cell death in the DJ-1^{-/-} line at this time-point. However, the observed mechanism appears to be different from the one observed in our PARKIN^{-/-} line.

[0165] **Oxidative Stress is a shared phenotype in all EO-PD DANs**

[0166] One advantage of our *in vitro* disease model lies in the possibility of studying early disease-relevant molecular events before cell death occurs, facilitating efforts to find common and distinct pathways between the three isogenic lines. There is evidence that oxidative stress is a component of the pathophysiology of familial and sporadic forms of PD. Mitochondrial reactive oxygen species are a known source of cellular stress and have been widely examined in the context of PD and in hPSC models of PD (Blesa et al., 2015; Cooper et al., 2012; Csobonyei et al., 2016; Dias et al., 2013). Despite the presence of antioxidants in the differentiation medium, analysis using flow cytometry revealed significantly increased levels of ROS in the basal state in all mutated lines when compared to

the isogenic WT control. This increase was most pronounced in the TH-positive cells, demonstrating the cell-type specific vulnerability of DANs, and highlighting the importance of modeling using disease relevant cells. Deficiencies of the respiratory chain in complex I are critical for the neuronal degeneration seen in PD (Blesa and Przedborski, 2014; Schapira et al., 1990). We found subtle, but consistent and highly significant, downregulation of proteins in most subunits that comprise the large mitochondrial complex I. This might explain an increase in oxidative stress in both the DJ-1^{-/-} line as well as in the PARKIN^{-/-} line. We found the highest ROS levels in TH-positive cells in the PARKIN^{-/-} line, consistent with the observed cell death of those mDANs.

[0167] Dysregulation of ECM-receptor pathways is a common pathway in isogenic PD lines.

[0168] Numerous etiologies have been proposed and investigated for their role in PD neurodegeneration. We analyzed the three isogenic cell lines for evidence of common disease mechanisms in an unbiased manner using KEGG pathway mapping. In PARKIN^{-/-} and ATP13A2^{-/-} lines, dysregulation of the ‘extracellular matrix (ECM)-receptor interaction’ pathways were, respectively, the first and second most significant KEGG pathways found in the differential transcriptomics analysis. There were too few dysregulated transcripts in the DJ-1^{-/-} line compared to the WT line to allow meaningful pathway analysis. However, when analyzing global proteomics data, KEGG pathway analysis flagged ‘dysregulated ECM-receptor interaction’ in all isogenic lines, when compared to WT cells. ECM interactions have been implicated in Alzheimers Disease and PD, and multiple strategies exist to intervene pharmaceutically with ECM interactions, or their metabolizing enzymes (Berezin et al., 2014). However, little is known about the specific dysregulated genes found in our study and their implications in disease. It will be important to analyze if the same genes are found

to be dysregulated in PD patients and to understand the role these genes play in the development of PD phenotypes.

[0169] **Analysis of PD associated genes in all isogenic lines reveals common disease pathways.**

[0170] PD disease pathways are often explained in a network context (Trinh and Farrer, 2013; Verstraeten et al., 2015) but few studies have attempted to investigate, at the molecular level, how mutations in such diverse proteins can all lead to PD. Several studies, including our own, have used transcriptomics data to identify and understand PD relevant genes and pathways. However, protein stability and degradation are independent of transcriptional activity and strongly contribute to the regulation of protein levels. These processes are widely implicated in neurodegenerative diseases including PD (Caudle et al., 2010; Tai and Schuman, 2008). Dysregulation of a PD protein, that in turn, leads to the dysregulation of another PD relevant gene can be seen as a common pathway. Using the UCL high priority set of 48 PD relevant genes, we analyzed PD relevant proteins in all our isogenic cell lines. For example, α -synuclein (SNCA), as well as tau protein (MAPT,) were both more than 2-fold enriched in PARKIN^{-/-} DANs when compared to WT DANs. We also found a 1.6-fold enrichment of SNCA in the ATP13A2^{-/-} line. SNCA was the first specific genetic aberration to have been linked to the development of PD (Polymeropoulos et al., 1997), and accumulation of SNCA aggregates and the formation of Lewy bodies are hallmarks of PD. Several SNCA mutations in PD patients have been investigated, and a gene dosage effect exists. Patients with SNCA triplications show early onset of PD and severe disease progression (Fuchs et al., 2007; Singleton et al., 2003). In addition, several publications have presented data consistent with an increase of SNCA caused by loss of PARKIN (Chung et al., 2016; Imaizumi et al., 2012; Shaltouki et al., 2015). While tau accumulation is classically found in dementia and neuropathies that are described as

tauopathies (Moussaud et al., 2014), tau has been identified as one of the most significant hits in genome-wide association studies of PD (Labbe and Ross, 2014). Both synuclein and tau proteins might work together in a pathological pathway relating to protein aggregation. In addition, we found significant increases in LAMP3 and GPNMB in the ATP13A2^{-/-} line. GPNMB and LAMP3 might function in the autophagy-lysosome pathway (Li et al., 2010; Nagelkerke et al., 2014). Increased expression of GPNMB as a gain of function in PD (Murthy et al., 2017) has also been associated with lysosomal storage disorders such as Gaucher disease and Niemann-pick type C disease (Kramer et al., 2016; Marques et al., 2016). Neither protein was detected in the proteomics experiment involving the PARKIN^{-/-} and DJ-1^{-/-} lines for technical reasons, and, therefore, it is unclear if they were dysregulated in those lines. This analysis highlights how our isogenic model can be used to find direct and indirect downstream targets of PD associated gene mutations.

[0171] Loss of DJ-1 leads to dysregulation of distinct pathways involving the cell cycle and the neuropathology of Charcot-Marie-Tooth disease

[0172] We found no significantly changed UCL high priority proteins in the DJ-^{-/-} line, which appears phenotypically distinct from the other PD lines. We set out to analyze dysregulated proteins between the WT and DJ-1^{-/-} lines to understand other PD relevant pathways. Database for Annotation, Visualization and Integrated Discovery (DAVID) functional annotation clustering of dysregulated proteins showed enrichment of genes associated with neuropathy and Charcot-Marie-Tooth disease (CMT). Four of the 26 significantly depleted proteins, namely N-myc downstream regulated 1 (NDRG1), neurofilament light polypeptide (NEFL), heat shock protein beta-8 (HSPB8) and mid-sized neurofilament (NEFM), are genes that have been linked to development of CMT (Hoyle et al., 2015). NEFL is a major component of neurofilaments and, together with NEFM and heavy neurofilament (NEFH) subunits, form the major intermediate filament in neurons.

Mutations in this locus lead to disruption of axonal neurofilament translocation, which affects the transport of mitochondria in axons (Brownlee et al., 2002). Mutant forms of HSP27 induce CMT through deficient retrograde axonal transport of mitochondria (Kalmar et al., 2017). Defects in mitochondrial transport have been suggested to play a role in the pathogenesis of PD, but this has not been conclusively demonstrated. Here we suggest a connection between loss of function mutations in *DJ-1* and genes that are known to cause CMT. Future studies will be required to determine if and how these proteins contribute to disease pathology seen in CMT and PD. CMT is also genetically heterogeneous, and, recently, a different CMT mutation in the *LRSAM1* gene was linked to the development of PD in three patients (Aerts et al., 2016).

[0173] Familial CMT type 2 and PD have been linked in an older study, but in that case no molecular analysis exists (Tranchant et al., 1994). DJ-1 is a multi-functional protein and DJ-1 protease activity has been studied using recombinant DJ-1 and a peptide library. A valine-lysine-valine-alanine (VKVA) recognition sequence was identified in target proteins (Mitsugi et al., 2013). We set out to analyze if DJ-1 can act as a cysteine protease in DANs, regulating many target proteins. We generated a list of 162 proteins containing a validated VKVA site, of which 43 were identified in our dataset, and we compared these protein levels in WT and DJ-1^{-/-} lines. Loss of DJ-1 lead to enrichment in most VKVA-proteins. Analyzed in compound this enrichment was highly significant (pValue = 0.0013), Of the 43 identified peptides, several play important PD relevant roles in cell organelle trafficking or mitochondria trafficking, neuronal development, and apoptosis. It is unclear why loss of DJ-1 causes these subtle changes in the abundance of potential substrates. It could be explained by either weak DJ-1 activity in the WT cells in basal conditions or through existing compensatory mechanisms. However, proteins with a VKVA recognition sequence will be cell-type specific and could provide an avenue to explain the specific vulnerability of DANs.

[0174] The isogenic PD model increases our understanding of common and distinct disease pathways

[0175] Overall, our experimental paradigm using a knock-in fluorescence reporter in the TH locus, together with the results of the phenotyping, transcriptomics and proteomic analysis of isogenic lines, provides an improved understanding of *PARKIN*, *DJ-1* and *ATP13A2* function. The model allows us to compare the impact of PD relevant genes on the differentiation and health of the most vulnerable cell type in PD. Our study demonstrates how isogenic PD hPSC models can be used to increase the understanding of PD relevant disease pathology, determine affected pathways and enhance our understanding of genetic interactions in PD pathology. We have demonstrated DAN specific disease relevant phenotypes in the *PARKIN*^{-/-} line and identified oxidative stress as a common pathology and a shared dysregulated pathway in all isogenic cell lines. Conversely, and despite similar clinical presentations, we found evidence for at least two etiologic subtypes of PD between the *PARKIN*^{-/-} and the *DJ-1*^{-/-} lines. We found an overlap between the synucleopathies and tauopathies, and provide evidence for related pathologies in PD and CMT. On the other hand, loss of *ATP13A2* and loss of *PARKIN* appear more similar in their disease etiology at the molecular levels despite being clinically distinct. Our results emphasize that precise delineation of PD subtypes will require evaluation of both molecular and clinical data. In future studies, we will expand our system to include all relevant PD mutations in the same isogenic background and late stage time-points to study the hierarchy of molecular events that eventually lead to death of DANs.

[0176] Citations:

[0177] Aerts, M.B., Weterman, M.A., Quadri, M., Schelhaas, H.J., Bloem, B.R., Esselink, R.A., Baas, F., Bonifati, V., and van de Warrenburg, B.P. (2016). A LRSAM1 mutation links Charcot-Marie-Tooth type 2 to Parkinson's disease. *Ann Clin Transl Neurol* 3, 146-149.

- [0178] Anderson, D.W., Schray, R.C., Duester, G., and Schneider, J.S. (2011). Functional significance of aldehyde dehydrogenase ALDH1A1 to the nigrostriatal dopamine system. *Brain Res 1408*, 81-87.
- [0179] Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J., and Gygi, S.P. (2006). A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol 24*, 1285-1292.
- [0180] Berezin, V., Walmod, P.S., Filippov, M., and Dityatev, A. (2014). Targeting of ECM molecules and their metabolizing enzymes and receptors for the treatment of CNS diseases. *Prog Brain Res 214*, 353-388.
- [0181] Bingol, B., Tea, J.S., Phu, L., Reichelt, M., Bakalarski, C.E., Song, Q., Foreman, O., Kirkpatrick, D.S., and Sheng, M. (2014). The mitochondrial deubiquitinase USP30 opposes parkin-mediated mitophagy. *Nature 510*, 370-375.
- [0182] Blesa, J., and Przedborski, S. (2014). Parkinson's disease: animal models and dopaminergic cell vulnerability. *Front Neuroanat 8*, 155.
- [0183] Blesa, J., Trigo-Damas, I., Quiroga-Varela, A., and Jackson-Lewis, V.R. (2015). Oxidative stress and Parkinson's disease. *Front Neuroanat 9*, 91.
- [0184] Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H., *et al.* (2011). Reference Maps of human ES and iPS cell variation enable highthroughput characterization of pluripotent cell lines. *Cell 144*, 439-452.
- [0185] Brownlees, J., Ackerley, S., Grierson, A.J., Jacobsen, N.J., Shea, K., Anderton, B.H., Leigh, P.N., Shaw, C.E., and Miller, C.C. (2002). Charcot-Marie-Tooth disease neurofilament mutations disrupt neurofilament assembly and axonal transport. *Human molecular genetics 11*, 2837-2844.

[0186] Burbulla, L.F., Song, P., Mazzulli, J.R., Zampese, E., Wong, Y.C., Jeon, S., Santos, D.P., Blanz, J., Obermaier, C.D., Strojny, C., *et al.* (2017). Dopamine oxidation mediates mitochondrial and lysosomal dysfunction in Parkinson's disease. *Science* 357, 1255-1261.

[0187] Caudle, W.M., Bammler, T.K., Lin, Y., Pan, S., and Zhang, J. (2010). Using 'omics' to define pathogenesis and biomarkers of Parkinson's disease. *Expert Rev Neurother* 10, 925-942.

[0188] Chai, C., and Lim, K.L. (2013). Genetic insights into sporadic Parkinson's disease pathogenesis. *Curr Genomics* 14, 486-501.

[0189] Chung, S.Y., Kishinevsky, S., Mazzulli, J.R., Graziotto, J., Mrejeru, A., Mosharov, E.V., Puspita, L., Valiulahi, P., Sulzer, D., Milner, T.A., *et al.* (2016). Parkin and PINK1 Patient iPSC-Derived Midbrain Dopamine Neurons Exhibit Mitochondrial Dysfunction and alpha-Synuclein Accumulation. *Stem cell reports* 7, 664- 677.

[0190] Cooper, O., Seo, H., Andrabi, S., Guardia-Laguarta, C., Graziotto, J., Sundberg, M., McLean, J.R., Carrillo- Reid, L., Xie, Z., Osborn, T., *et al.* (2012). Pharmacological rescue of mitochondrial deficits in iPSC-derived neural cells from patients with familial Parkinson's disease. *Science translational medicine* 4, 141ra190.

[0191] Cowan, C.A., Klimanskaya, I., McMahon, J., Atienza, J., Witmyer, J., Zucker, J.P., Wang, S., Morton, C.C., McMahon, A.P., Powers, D., *et al.* (2004). Derivation of embryonic stem-cell lines from human blastocysts. *N Engl J Med* 350, 1353-1356.

[0192] Csobonyeiova, M., Danisovic, L., and Polak, S. (2016). Induced pluripotent stem cells for modeling and cell therapy of Parkinson's disease. *Neural regeneration research* 11, 727-728.

[0193] Dauer, W., and Przedborski, S. (2003). Parkinson's disease: mechanisms and models. *Neuron* 39, 889- 909.

[0194] Dias, V., Junn, E., and Mouradian, M.M. (2013). The role of oxidative stress in Parkinson's disease. *J Parkinsons Dis* 3, 461-491.

[0195] Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. *PLoS computational biology* 3, e39.

[0196] Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics* 10, 48.

[0197] Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* 4, 207-214.

[0198] Elias, J.E., and Gygi, S.P. (2010). Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol* 604, 55-71.

[0199] Foulger, R.E., Denny, P., Hardy, J., Martin, M.J., Sawford, T., and Lovering, R.C. (2016). Using the Gene Ontology to Annotate Key Players in Parkinson's Disease. *Neuroinformatics* 14, 297-304.

[0200] Fuchs, J., Nilsson, C., Kachergus, J., Munz, M., Larsson, E.M., Schule, B., Langston, J.W., Middleton, F.A., Ross, O.A., Hulihan, M., *et al.* (2007). Phenotypic variation in a large Swedish pedigree due to SNCA duplication and triplication. *Neurology* 68, 916-922.

[0201] Goulburn, A.L., Alden, D., Davis, R.P., Micallef, S.J., Ng, E.S., Yu, Q.C., Lim, S.M., Soh, C.L., Elliott, D.A., Hatzistavrou, T., *et al.* (2011). A targeted NKX2.1 human embryonic stem cell reporter line enables identification of human basal forebrain derivatives. *Stem Cells* 29, 462-473.

[0202] Greene, A.W., Grenier, K., Aguilera, M.A., Muise, S., Farazifard, R., Haque, M.E., McBride, H.M., Park, D.S., and Fon, E.A. (2012). Mitochondrial processing peptidase regulates PINK1 processing, import and Parkin recruitment. *EMBO Rep* 13, 378-385.

[0203] Hampshire, D.J., Roberts, E., Crow, Y., Bond, J., Mubaidin, A., Wriekat, A.L., Al-Din, A., and Woods, C.G. (2001). Kufor-Rakeb syndrome, pallido-pyramidal degeneration with supranuclear upgaze paresis and dementia, maps to 1p36. *Journal of medical genetics* 38, 680-682.

[0204] Hockemeyer, D., Soldner, F., Beard, C., Gao, Q., Mitalipova, M., DeKever, R.C., Katibah, G.E., Amora, R., Boydston, E.A., Zeitler, B., *et al.* (2009). Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nat Biotechnol* 27, 851-857.

[0205] Hoyle, J.C., Isfort, M.C., Roggenbuck, J., and Arnold, W.D. (2015). The genetics of Charcot-Marie-Tooth disease: current trends and future implications for diagnosis and management. *Appl Clin Genet* 8, 235- 243.

[0206] Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S.A., Villen, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* 143, 1174-1189.

[0207] Imaizumi, Y., Okada, Y., Akamatsu, W., Koike, M., Kuzumaki, N., Hayakawa, H., Nihira, T., Kobayashi, T., Ohyama, M., Sato, S., *et al.* (2012). Mitochondrial dysfunction associated with increased oxidative stress and alpha-synuclein accumulation in PARK2 iPSC-derived neurons and postmortem brain tissue. *Mol Brain* 5, 35.

[0208] Ismadi, M.Z., Gupta, P., Fouras, A., Verma, P., Jadhav, S., Bellare, J., and Hourigan, K. (2014). Flow characterization of a spinner flask for induced pluripotent stem cell culture application. *PLoS One* 9, e106493.

[0209] Jiang, H., Ren, Y., Yuen, E.Y., Zhong, P., Ghaedi, M., Hu, Z., Azabdaftari, G., Nakaso, K., Yan, Z., and Feng, J. (2012). Parkin controls dopamine utilization in human midbrain dopaminergic neurons derived from induced pluripotent stem cells. *Nature communications* 3, 668.

[0210] Kalmar, B., Innes, A., Wanisch, K., Koyen Kolaszynska, A., Pandraud, A., Kelly, G., Abramov, A.Y., Reilly, M.M., Schiavo, G., and Greensmith, L. (2017). Mitochondrial deficits and abnormal mitochondrial retrograde axonal transport play a role in the pathogenesis of mutant Hsp27 induced Charcot Marie Tooth Disease. *Human molecular genetics*.

[0211] Keeney, P.M., Xie, J., Capaldi, R.A., and Bennett, J.P., Jr. (2006). Parkinson's disease brain mitochondrial

[0212] complex I has oxidatively damaged subunits and is functionally impaired and misassembled. *The Journal*

[0213] of neuroscience : the official journal of the Society for Neuroscience 26, 5256-5264. Klein, C., and Westenberger, A. (2012). Genetics of Parkinson's disease. *Cold Spring Harbor perspectives in medicine* 2, a008888.

[0214] Kramer, G., Wegdam, W., Donker-Koopman, W., Ottenhoff, R., Gaspar, P., Verhoek, M., Nelson, J., Gabriel, T., Kallemeyjn, W., Boot, R.G., *et al.* (2016). Elevation of glycoprotein nonmetastatic melanoma protein B in type 1 Gaucher disease patients and mouse models. *FEBS Open Bio* 6, 902-913.

[0215] Kriks, S., Shim, J.W., Piao, J., Ganat, Y.M., Wakeman, D.R., Xie, Z., Carrillo-Reid, L., Auyeung, G., Antonacci, C., Buch, A., *et al.* (2011). Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson's disease. *Nature* 480, 547-551.

[0216] Kuhn, K., Zhu, X.R., Lubbert, H., and Stichel, C.C. (2004). Parkin expression in the developing mouse. *Brain research Developmental brain research* 149, 131-142.

[0217] Labbe, C., and Ross, O.A. (2014). Association studies of sporadic Parkinson's disease in the genomic era. *Curr Genomics* 15, 2-10.

[0218] Lewis, S.J., Foltynie, T., Blackwell, A.D., Robbins, T.W., Owen, A.M., and Barker, R.A. (2005). Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. *J Neurol Neurosurg Psychiatry* 76, 343-348.

[0219] Li, B., Castano, A.P., Hudson, T.E., Nowlin, B.T., Lin, S.L., Bonventre, J.V., Swanson, K.D., and Duffield, J.S. (2010). The melanoma-associated transmembrane glycoprotein Gpnmb controls trafficking of cellular debris for degradation and is essential for tissue repair. *FASEB J* 24, 4767-4781.

[0220] Liu, Y., Jiang, P., and Deng, W. (2011). OLIG gene targeting in human pluripotent stem cells for motor neuron and oligodendrocyte differentiation. *Nature protocols* 6, 640-655.

[0221] Maier, T., Guell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Lett* 583, 3966-3973.

[0222] Marques, A.R., Gabriel, T.L., Aten, J., van Roomen, C.P., Ottenhoff, R., Claessen, N., Alfonso, P., Irun, P., Giraldo, P., Aerts, J.M., *et al.* (2016). Gpnmb Is a Potential Marker for the Visceral Pathology in Niemann- Pick Type C Disease. *PLoS One* 11, e0147208.

[0223] Matsuda, T., and Cepko, C.L. (2007). Controlled expression of transgenes introduced by in vivo electroporation. *Proc Natl Acad Sci U S A* 104, 1027-1032.

[0224] McAlister, G.C., Huttlin, E.L., Haas, W., Ting, L., Jedrychowski, M.P., Rogers, J.C., Kuhn, K., Pike, I., Grothe, R.A., Blethrow, J.D., *et al.* (2012). Increasing the

multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses.

Analytical chemistry 84, 7469-7478.

[0225] McAlister, G.C., Nusinow, D.P., Jedrychowski, M.P., Wuhr, M., Huttlin, E.L., Erickson, B.K., Rad, R., Haas, W., and Gygi, S.P. (2014). MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Analytical chemistry* 86, 7150-7158.

[0226] McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* 40, 4288-4297.

[0227] Merkle, F.T., Neuhausser, W.M., Santos, D., Valen, E., Gagnon, J.A., Maas, K., Sandoe, J., Schier, A.F., and Eggan, K. (2015). Efficient CRISPR-Cas9-mediated generation of knockin human pluripotent stem cells lacking undesired mutations at the targeted locus. *Cell reports* 11, 875-883.

[0228] Mitsugi, H., Niki, T., Takahashi-Niki, K., Tanimura, K., Yoshizawa-Kumagaye, K., Tsunemi, M., Iguchi-Ariga, S.M., and Ariga, H. (2013). Identification of the recognition sequence and target proteins for DJ-1 protease. *FEBS Lett* 587, 2493-2499.

[0229] Moussaud, S., Jones, D.R., Moussaud-Lamodiere, E.L., Delenclos, M., Ross, O.A., and McLean, P.J. (2014). Alpha-synuclein and tau: teammates in neurodegeneration? *Mol Neurodegener* 9, 43.

[0230] Murthy, M.N., Blauwendraat, C., Ukec, Guelfi, S., Ipdgc, Hardy, J., Lewis, P.A., and Trabzuni, D. (2017). Increased brain expression of GPNMB is associated with genome wide significant risk for Parkinson's disease on chromosome 7p15.3. *Neurogenetics*.

[0231] Nagelkerke, A., Sieuwerts, A.M., Bussink, J., Sweep, F.C., Look, M.P., Foekens, J.A., Martens, J.W., and Span, P.N. (2014). LAMP3 is involved in tamoxifen

resistance in breast cancer cells through the modulation of autophagy. *Endocr Relat Cancer* 21, 101-112.

[0232] Ordureau, A., Heo, J.M., Duda, D.M., Paulo, J.A., Olszewski, J.L., Yanishevski, D., Rinehart, J., Schulman, B.A., and Harper, J.W. (2015). Defining roles of PARKIN and ubiquitin phosphorylation by PINK1 in mitochondrial quality control using a ubiquitin replacement strategy. *Proc Natl Acad Sci U S A* 112, 6637- 6642.

[0233] Otsuji, T.G., Bin, J., Yoshimura, A., Tomura, M., Tateyama, D., Minami, I., Yoshikawa, Y., Aiba, K., Heuser, J.E., Nishino, T., *et al.* (2014). A 3D sphere culture system containing functional polymers for large-scale human pluripotent stem cell production. *Stem cell reports* 2, 734-745.

[0234] Paisan-Ruiz, C., Guevara, R., Federoff, M., Hanagasi, H., Sina, F., Elahi, E., Schneider, S.A., Schwingenschuh, P., Bajaj, N., Emre, M., *et al.* (2010). Early-onset L-dopa-responsive parkinsonism with pyramidal signs due to ATP13A2, PLA2G6, FBXO7 and spatacsin mutations. *Movement disorders : official journal of the Movement Disorder Society* 25, 1791-1800.

[0235] Park, I.H., Arora, N., Huo, H., Maherali, N., Ahfeldt, T., Shimamura, A., Lensch, M.W., Cowan, C., Hochedlinger, K., and Daley, G.Q. (2008). Disease-specific induced pluripotent stem cells. *Cell* 134, 877- 886.

[0236] Park, J.S., Koentjoro, B., Veivers, D., Mackay-Sim, A., and Sue, C.M. (2014). Parkinson's disease-associated human ATP13A2 (PARK9) deficiency causes zinc dyshomeostasis and mitochondrial dysfunction. *Human molecular genetics* 23, 2802-2815.

[0237] Paulo, J.A., O'Connell, J.D., and Gygi, S.P. (2016). A Triple Knockout (TKO) Proteomics Standard for Diagnosing Ion Interference in Isobaric Labeling Experiments. *Journal of the American Society for Mass Spectrometry* 27, 1620-1625.

[0238] Pickrell, A.M., and Youle, R.J. (2015). The roles of PINK1, parkin, and mitochondrial fidelity in Parkinson's disease. *Neuron* 85, 257-273.

[0239] Polymeropoulos, M.H., Lavedan, C., Leroy, E., Ide, S.E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R., *et al.* (1997). Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* 276, 2045-2047.

[0240] Postuma, R.B., Berg, D., Adler, C.H., Bloem, B.R., Chan, P., Deuschl, G., Gasser, T., Goetz, C.G., Halliday, G., Joseph, L., *et al.* (2016). The new definition and diagnostic criteria of Parkinson's disease. *Lancet Neurol* 15, 546-548.

[0241] Rakovic, A., Seibler, P., and Klein, C. (2015). iPS models of Parkin and PINK1. *Biochemical Society transactions* 43, 302-307.

[0242] Reinhardt, P., Schmid, B., Burbulla, L.F., Schondorf, D.C., Wagner, L., Glatza, M., Hoing, S., Hargus, G., Heck, S.A., Dhingra, A., *et al.* (2013). Genetic correction of a LRRK2 mutation in human iPSCs links parkinsonian neurodegeneration to ERK-dependent changes in gene expression. *Cell stem cell* 12, 354- 367.

[0243] Rigamonti, A., Repetti, G.G., Sun, C., Price, F.D., Reny, D.C., Rapino, F., Weisinger, K., Benkler, C., Peterson, Q.P., Davidow, L.S., *et al.* (2016). Large-Scale Production of Mature Neurons from Human Pluripotent Stem Cells in a Three-Dimensional Suspension Culture System. *Stem cell reports* 6, 993-1008.

[0244] Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139-140.

[0245] Rose, C.M., Isasa, M., Ordureau, A., Prado, M.A., Beausoleil, S.A., Jedrychowski, M.P., Finley, D.J., Harper, J.W., and Gygi, S.P. (2016). Highly Multiplexed Quantitative Mass Spectrometry Analysis of Ubiquitylomes. *Cell Syst* 3, 395-403 e394.

[0246] Rouhani, F., Kumasaka, N., de Brito, M.C., Bradley, A., Vallier, L., and Gaffney, D. (2014). Genetic background drives transcriptional variation in human induced pluripotent stem cells. *PLoS genetics* 10, e1004432.

[0247] Ruby, K.M., and Zheng, B. (2009). Gene targeting in a HUES line of human embryonic stem cells via electroporation. *Stem Cells* 27, 1496-1506.

[0248] Ryan, S.D., Dolatabadi, N., Chan, S.F., Zhang, X., Akhtar, M.W., Parker, J., Soldner, F., Sunico, C.R., Nagar, S., Talantova, M., *et al.* (2013). Isogenic human iPSC Parkinson's model shows nitrosative stress-induced dysfunction in MEF2-PGC1alpha transcription. *Cell* 155, 1351-1364.

[0249] Sarraf, S.A., Raman, M., Guarani-Pereira, V., Sowa, M.E., Huttlin, E.L., Gygi, S.P., and Harper, J.W. (2013). Landscape of the PARKIN-dependent ubiquitylome in response to mitochondrial depolarization. *Nature* 496, 372-376.

[0250] Schapira, A.H., Cooper, J.M., Dexter, D., Clark, J.B., Jenner, P., and Marsden, C.D. (1990). Mitochondrial complex I deficiency in Parkinson's disease. *Journal of neurochemistry* 54, 823-827.

[0251] Schinzel, R.T., Ahfeldt, T., Lau, F.H., Lee, Y.K., Cowley, A., Shen, T., Peters, D., Lum, D.H., and Cowan, C.A. (2011). Efficient culturing and genetic manipulation of human pluripotent stem cells. *PLoS One* 6, e27495.

[0252] Schulte, C., and Gasser, T. (2011). Genetic basis of Parkinson's disease: inheritance, penetrance, and expression. *Appl Clin Genet* 4, 67-80.

[0253] Seibler, P., Graziotto, J., Jeong, H., Simunovic, F., Klein, C., and Krainc, D. (2011). Mitochondrial Parkin recruitment is impaired in neurons derived from mutant PINK1 induced pluripotent stem cells. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 31, 5970-5976.

[0254] Shaltouki, A., Sivapatham, R., Pei, Y., Gerencser, A.A., Momcilovic, O., Rao, M.S., and Zeng, X. (2015). Mitochondrial alterations by PARKIN in dopaminergic neurons using PARK2 patient-specific and PARK2 knockout isogenic iPSC lines. *Stem cell reports* 4, 847-859.

[0255] Shaner, N.C., Campbell, R.E., Steinbach, P.A., Giepmans, B.N., Palmer, A.E., and Tsien, R.Y. (2004).

[0256] Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nat Biotechnol* 22, 1567-1572.

[0257] Shi, W.X. (2005). Slow oscillatory firing: a major firing pattern of dopamine neurons in the ventral tegmental area. *J Neurophysiol* 94, 3516-3522.

[0258] Singleton, A.B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R., *et al.* (2003). alpha-Synuclein locus triplication causes Parkinson's disease. *Science* 302, 841.

[0259] Stemmann, O., Zou, H., Gerber, S.A., Gygi, S.P., and Kirschner, M.W. (2001). Dual inhibition of sister chromatid separation at metaphase. *Cell* 107, 715-726.

[0260] Tai, H.C., and Schuman, E.M. (2008). Ubiquitin, the proteasome and protein degradation in neuronal function and dysfunction. *Nat Rev Neurosci* 9, 826-838.

[0261] Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131, 861-872.

[0262] Thomas, B., and Beal, M.F. (2007). Parkinson's disease. *Human molecular genetics* 16 Spec No. 2, R183- 194.

[0263] Torrent, R., De Angelis Rigotti, F., Dell'Era, P., Memo, M., Raya, A., and Consiglio, A. (2015). Using iPS Cells toward the Understanding of Parkinson's Disease. *Journal of clinical medicine* 4, 548-566.

[0264] Toth, C., Brown, M.S., Furtado, S., Suchowersky, O., and Zochodne, D. (2008). Neuropathy as a potential complication of levodopa use in Parkinson's disease. *Movement disorders : official journal of the Movement Disorder Society* 23, 1850-1859.

[0265] Tranchant, C., Ruh, D., and Warter, J.M. (1994). [Type II Charcot-Marie-Tooth and dopa-sensitive Parkinson disease]. *Rev Neurol (Paris)* 150, 72-74.

[0266] Trinh, J., and Farrer, M. (2013). Advances in the genetics of Parkinson disease. *Nat Rev Neurol* 9, 445- 454.

[0267] Uhl, G.R., Hedreen, J.C., and Price, D.L. (1985). Parkinson's disease: loss of neurons from the ventral tegmental area contralateral to therapeutic surgical lesions. *Neurology* 35, 1215-1218.

[0268] Valente, E.M., Abou-Sleiman, P.M., Caputo, V., Muqit, M.M., Harvey, K., Gispert, S., Ali, Z., Del Turco, D., Bentivoglio, A.R., Healy, D.G., *et al.* (2004). Hereditary early-onset Parkinson's disease caused by mutations in PINK1. *Science* 304, 1158-1160.

[0269] van der Merwe, C., Jalali Sefid Dashti, Z., Christoffels, A., Loos, B., and Bardien, S. (2015). Evidence for a common biological pathway linking three Parkinson's disease-causing genes: parkin, PINK1 and DJ-1. *The European journal of neuroscience* 41, 1113-1125.

[0270] Verstraeten, A., Theuns, J., and Van Broeckhoven, C. (2015). Progress in unraveling the genetic etiology of Parkinson disease in a genomic era. *Trends Genet* 31, 140-149.

[0271] Wang, Y., Yang, F., Gritsenko, M.A., Clauss, T., Liu, T., Shen, Y., Monroe, M.E., Lopez-Ferrer, D., Reno, T., Moore, R.J., *et al.* (2011). Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* 11, 2019-2026.

[0272] Xia, N., Fang, F., Zhang, P., Cui, J., Tep-Cullison, C., Hamerley, T., Lee, H.J., Palmer, T., Bothner, B., Lee, J.H., *et al.* (2017). A Knockin Reporter Allows Purification and Characterization of mDA Neurons from Heterogeneous Populations. *Cell reports* 18, 2533-2546.

Claims

1. A nucleic acid targeting vector comprising, in the 5' to 3' direction:
 - (i) a 5' homology arm homologous to a first target sequence in a cell,
 - (ii) a reporter gene,
 - (iii) an expression enhancer comprising a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE) operably linked to the reporter gene, and
 - (iv) a 3' homology arm homologous to a second target sequence downstream of the first target sequence.
2. The targeting vector of claim 1, wherein the expression enhancer further comprises at least one of a Long Term Repeat (LTR) and a Simian Virus 40 (SV40) terminator.
3. The targeting vector of claims 1 or 2, wherein the 5' homology arm and the 3' homology arm each comprise at least about 350 nucleotide bases.
4. The targeting vector of any one of claims 1-3, wherein the reporter gene encodes a fluorescent protein.
5. The targeting vector of claim 4, wherein the fluorescent protein is tdTomato.
6. The targeting vector of any one of claims 1-5, further comprising an IRES element or a sequence encoding a self-cleaving peptide located between the 5' homology arm and the reporter gene.
7. The targeting vector of any one of claims 1-6, further comprising an insulator sequence located between the WPRE and the 3' homology arm.
8. A composition comprising a cell having a genome comprising a nucleotide sequence having a reporter gene and a Woodchuck Hepatitis Virus Posttranscriptional Regulatory

Element (WPRE) operably linked to the reporter gene, wherein the reporter gene is co-expressed with a target sequence of interest.

9. The composition of claim 8, wherein the nucleotide sequence is located at the 3' end of an open reading frame of a gene of interest.

10. The composition of claim 8 or 9, wherein the cell is a human pluripotent stem cell.

11. The composition of any one of claims 8-10, wherein the reporter gene encodes tdTomato.

12. The composition of any one of claims 8-11, wherein the target sequence of interest encodes tyrosine hydroxylase.

13. A method of generating a cell that co-expresses a reporter gene with a gene of interest, the method comprising:

- (i) providing the targeting vector of any one of claims 1-6;
- (ii) providing a cell in which at least a portion of the gene of interest is located between the first target sequence and the second target sequence;
- (iii) introducing the targeting vector into the cell; and
- (iv) maintaining the cell under conditions appropriate for integration of the reporter gene and WPRE into the genome of the cell such that the reporter gene is co-expressed with the gene of interest,

wherein said portion of the gene of interest is cleaved prior to or subsequent to introducing the targeting vector into the cell.

14. A method of making a targeting vector capable of co-expressing a reporter gene with a gene of interest in a cell, the method comprising:

- (i) providing a vector comprising, in the 5' to 3' direction, a reporter gene and an expression enhancer comprising a Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE), wherein the expression enhancer and the reporter gene are operably linked;
- (ii) incorporating upstream of the reporter gene a 5' homology arm homologous to a first target sequence by G-Block Gibson assembly; and
- (iii) adding downstream of the WPRE a 3' homology arm homologous to a second target sequence by G-Block Gibson assembly,

wherein the second target sequence is located downstream of the first target sequence and wherein the first target sequence and the second target sequence flank at least a portion of the gene of interest.

FIG. 1

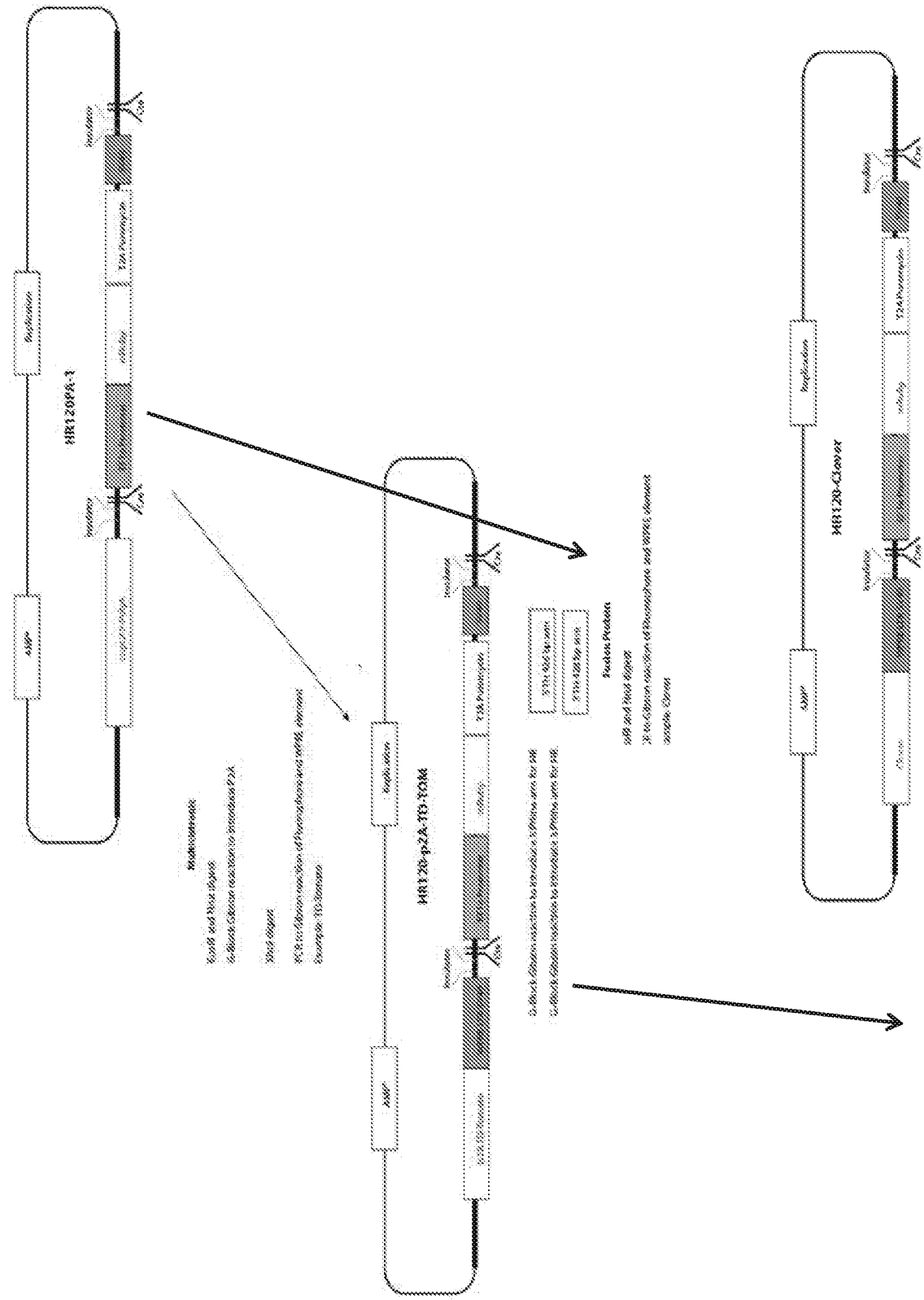
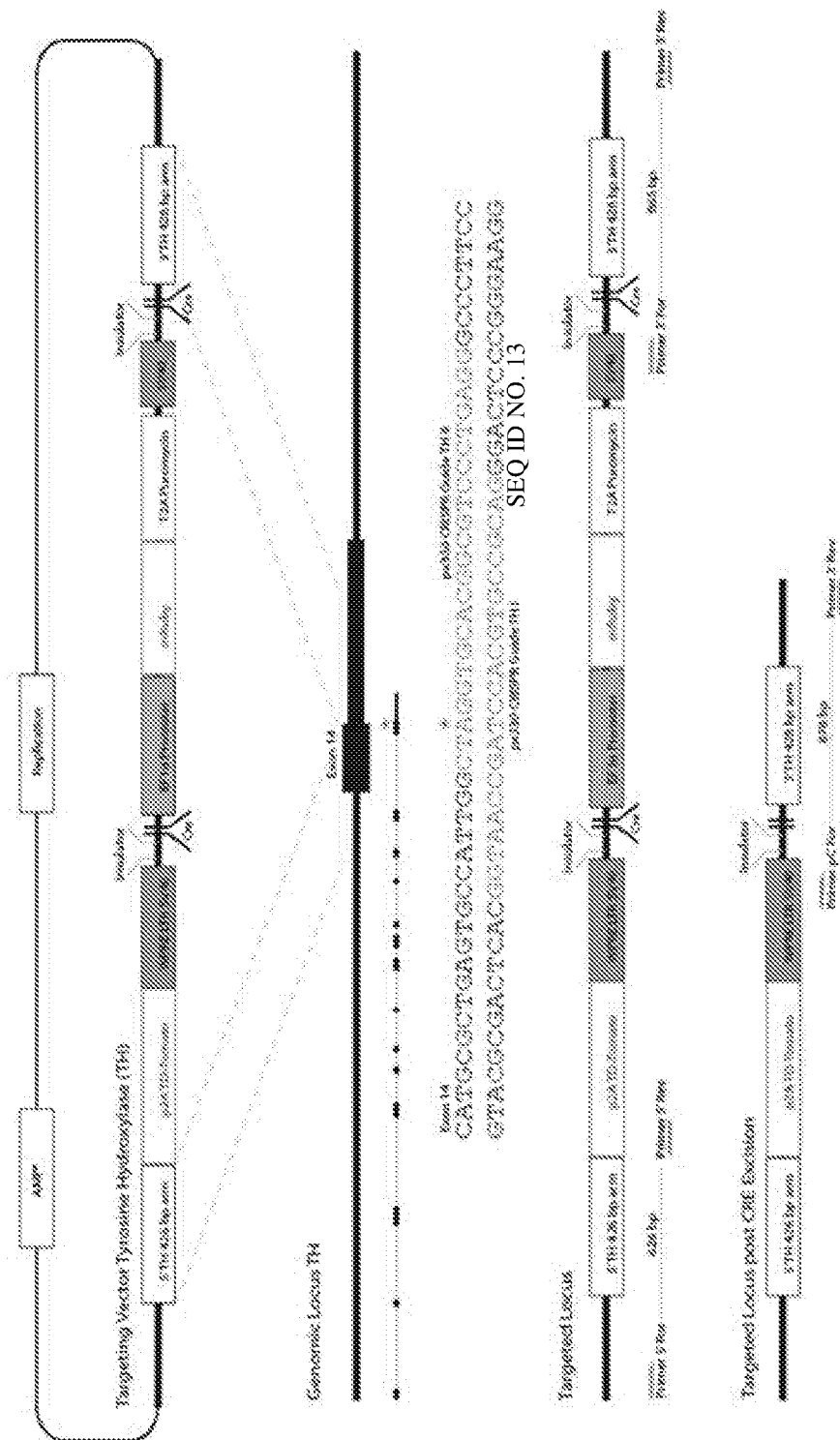


FIG. 1 (cont.)



FIGS. 2A-2B

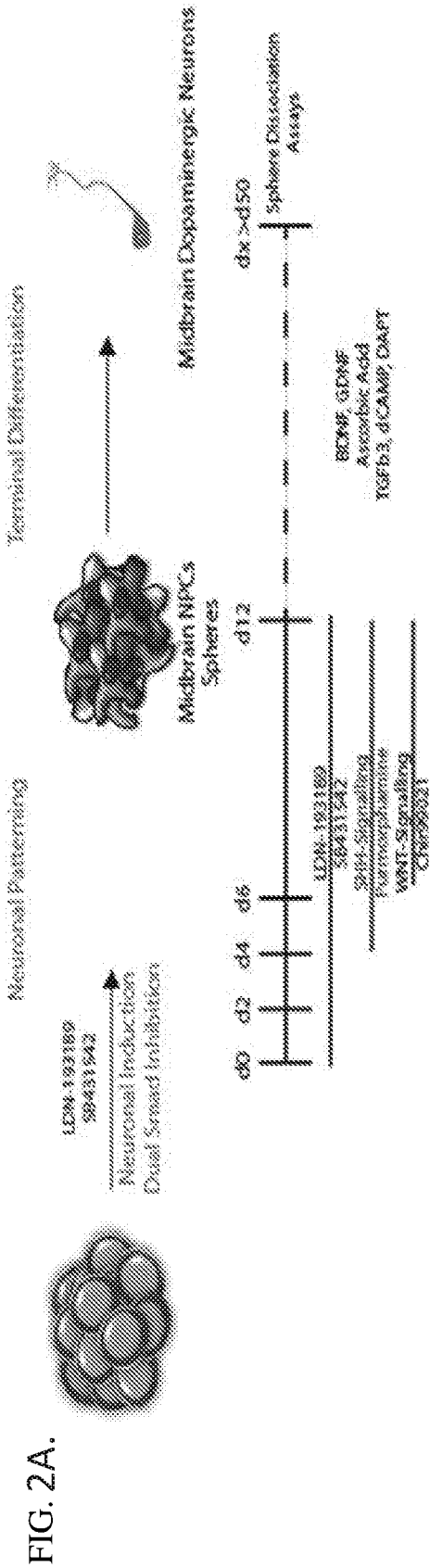
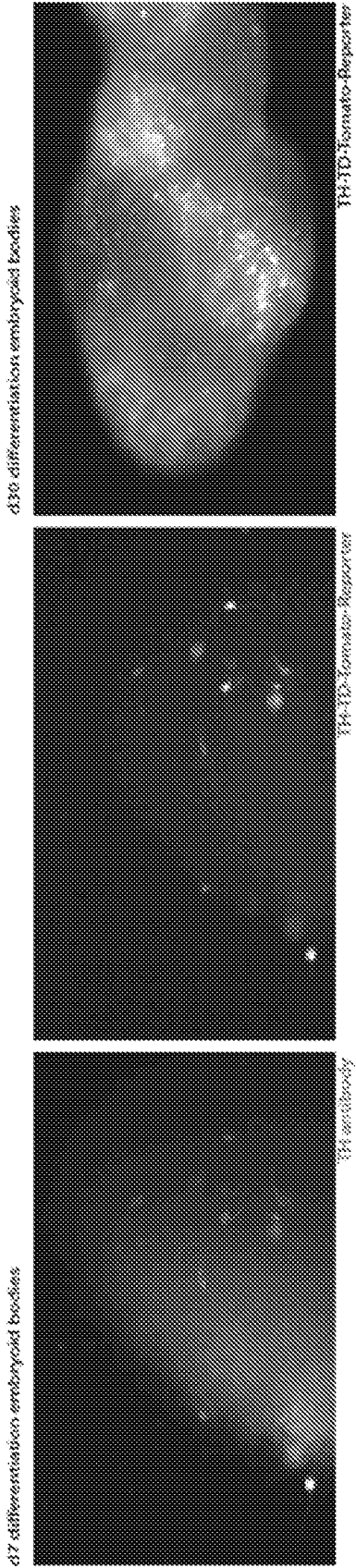


FIG. 2B.



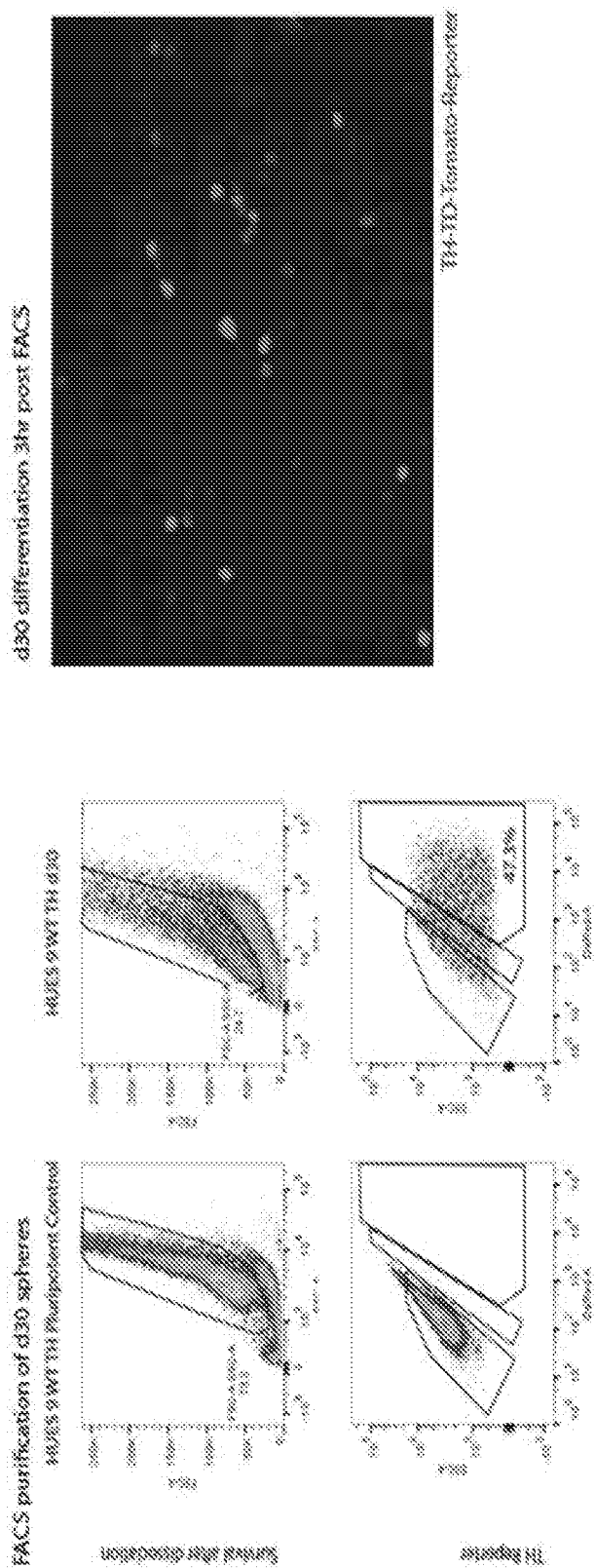
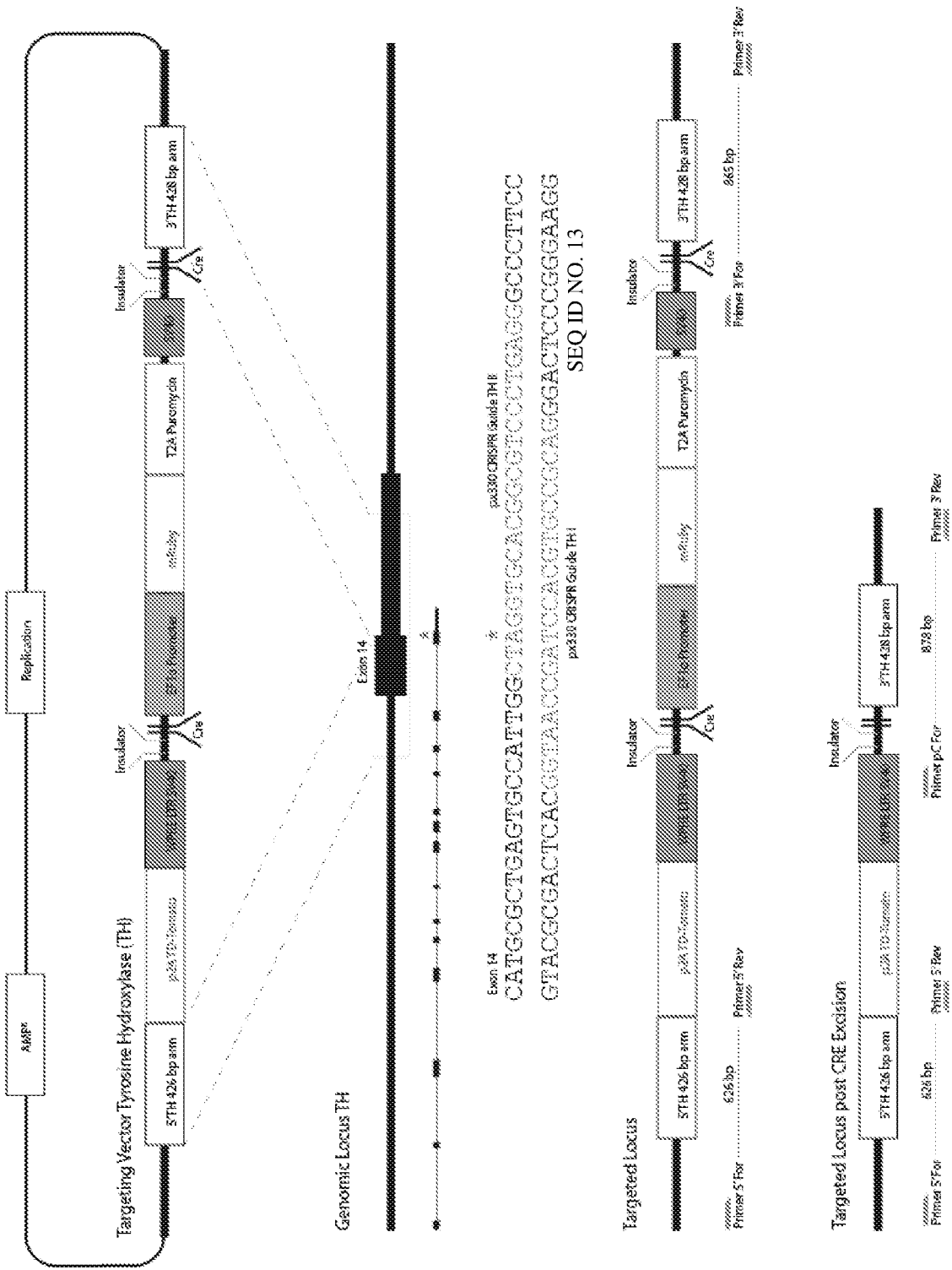
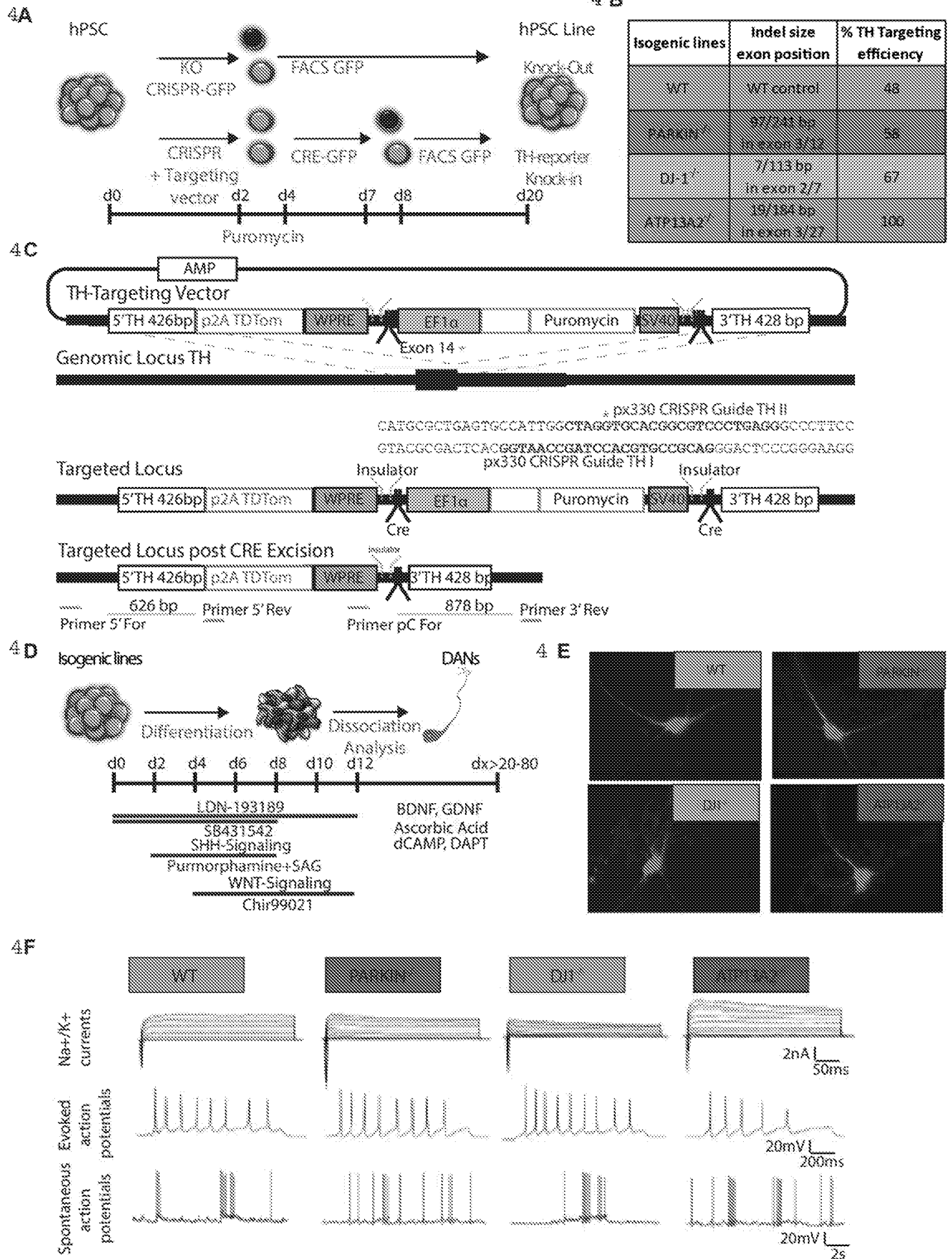
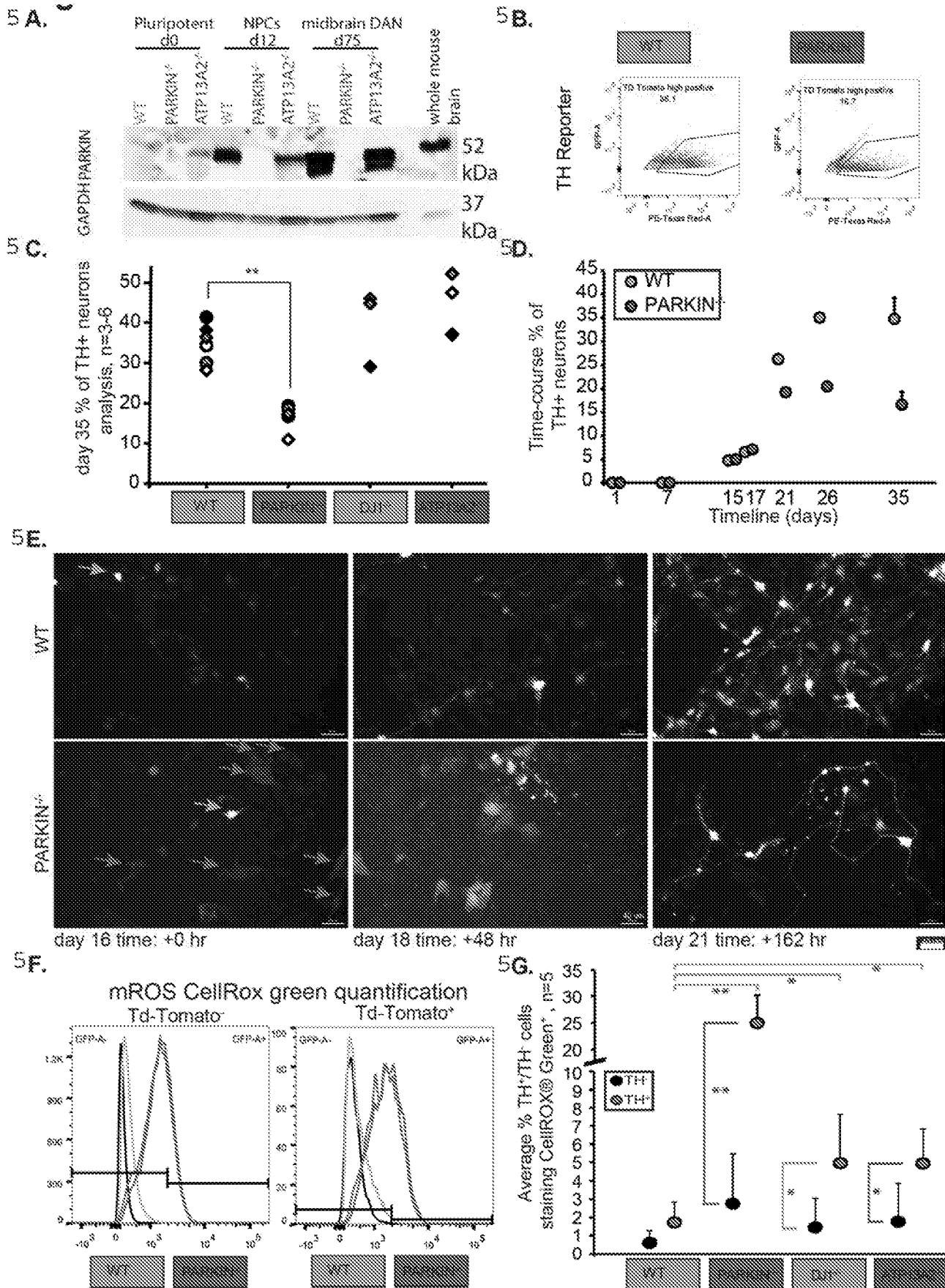


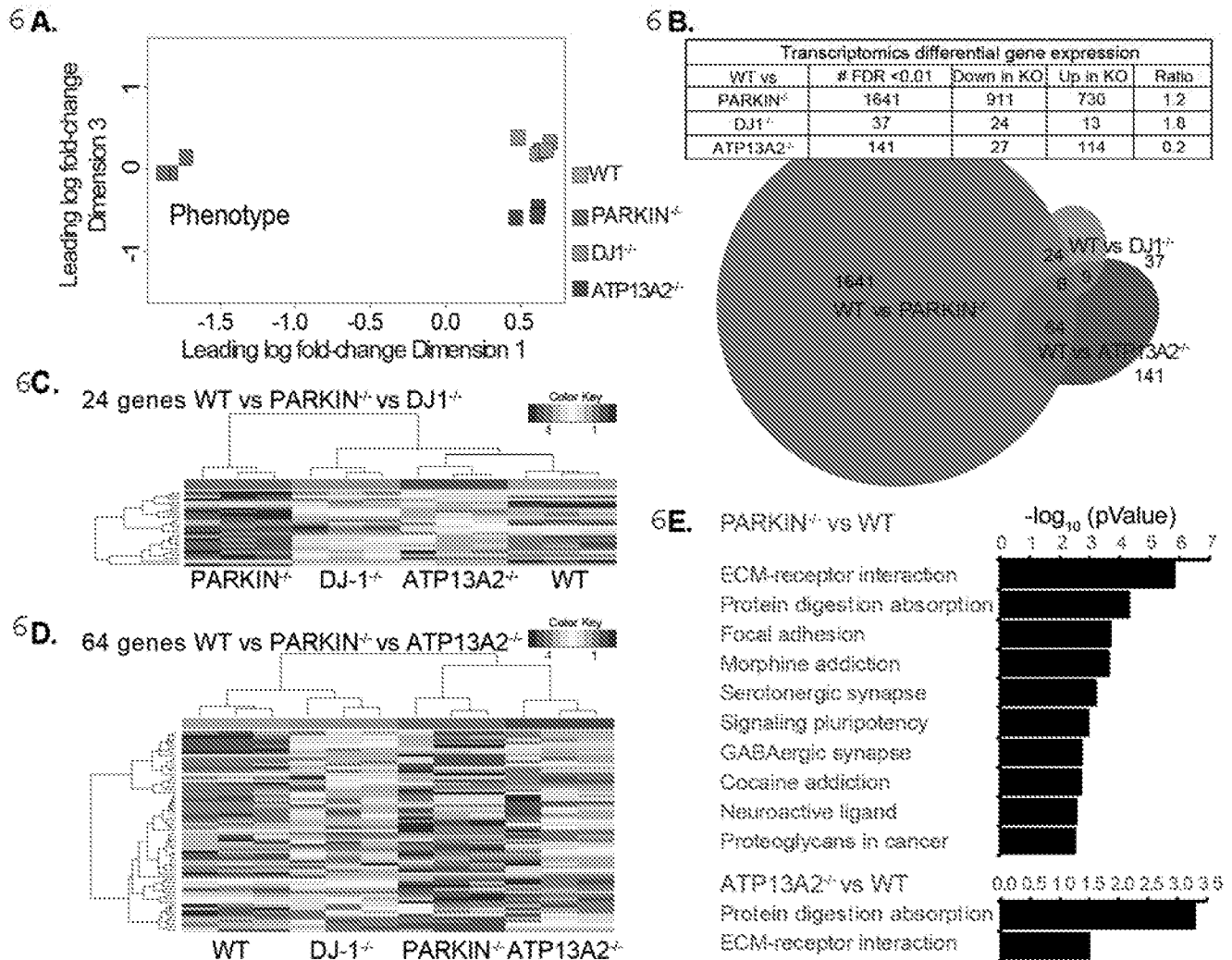
FIG. 3



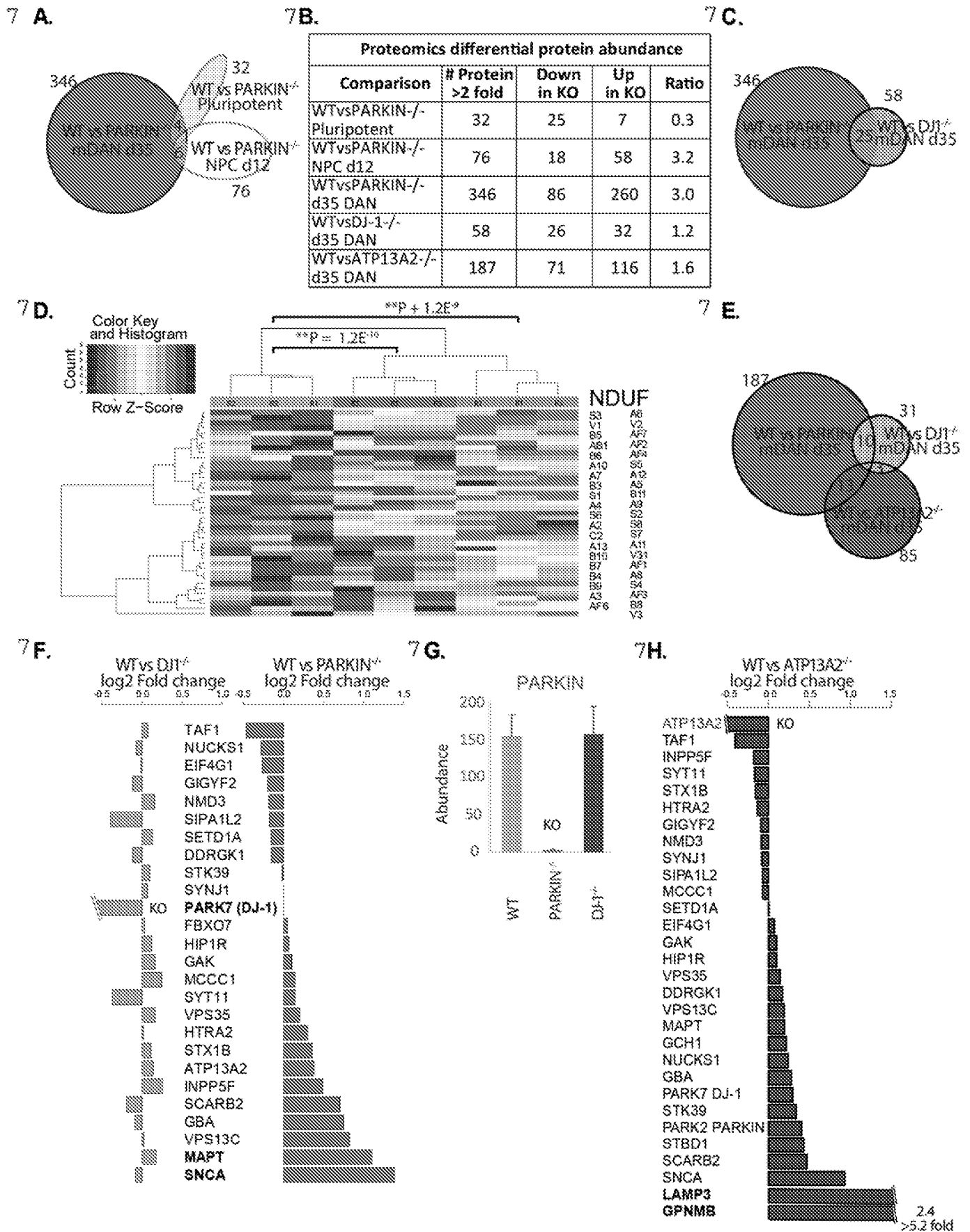




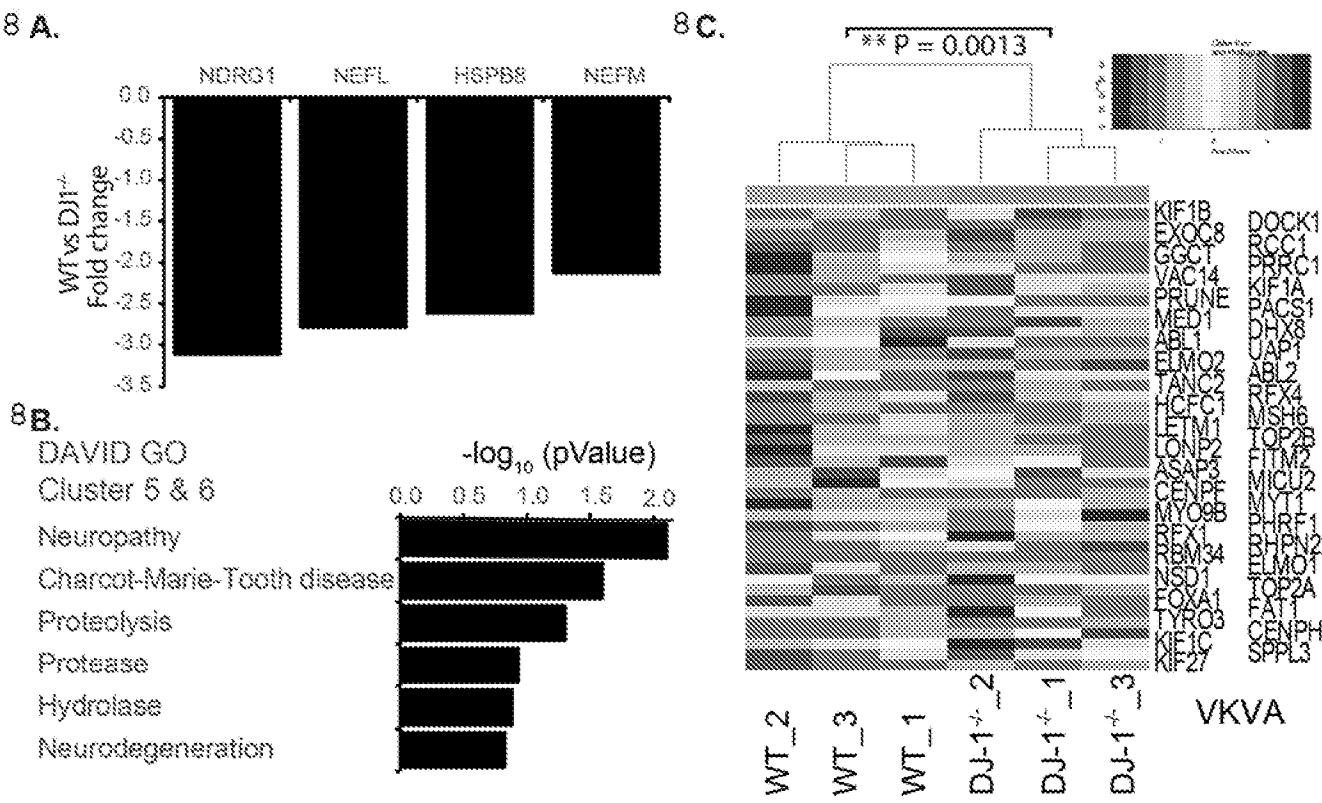
FIGS. 6A-6E



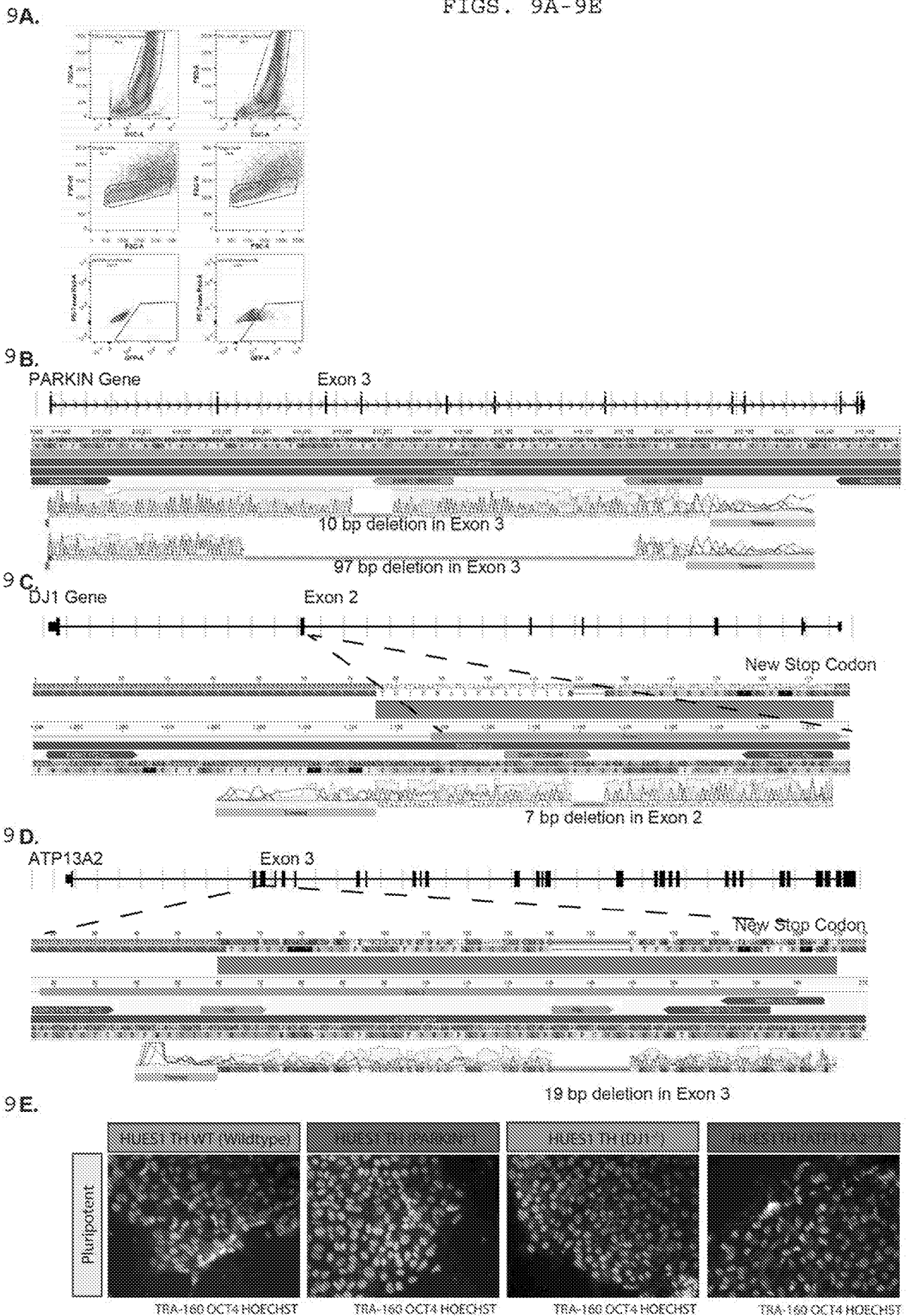
FIGS. 7A-7H



FIGS. 8A-8C

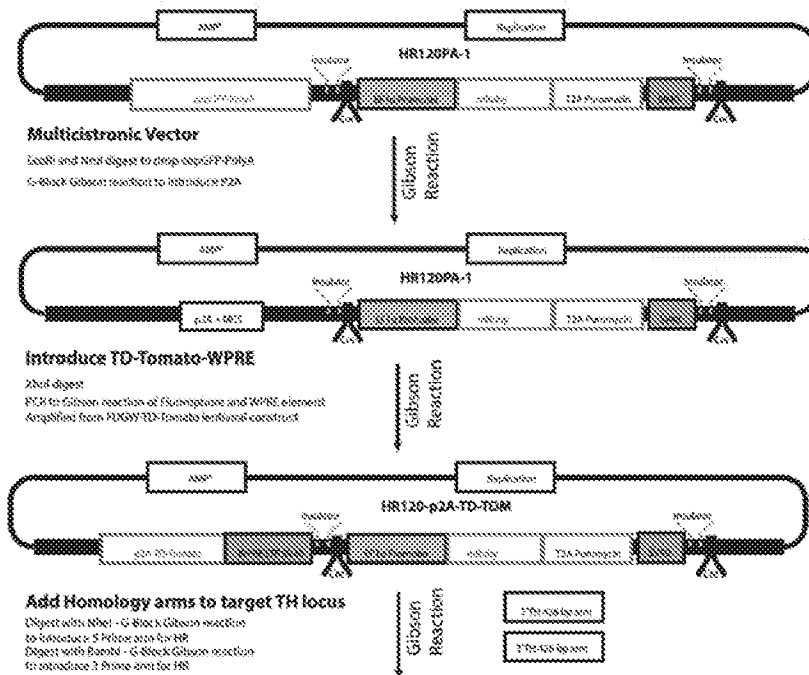


FIGS. 9A-9E

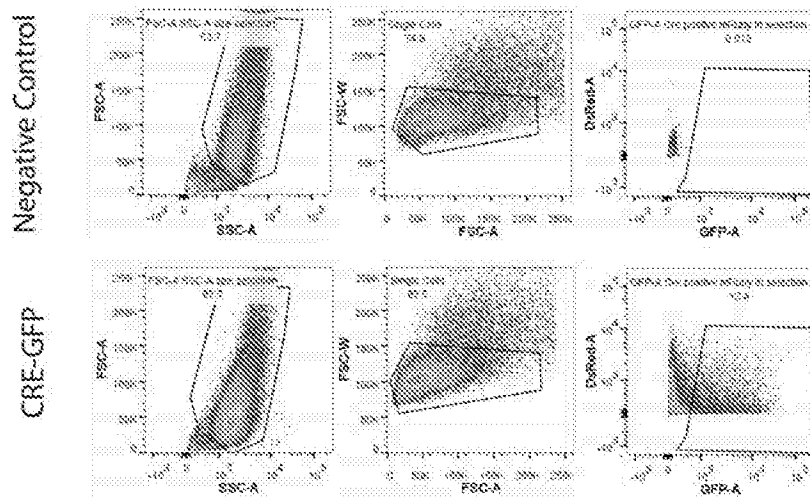


FIGS. 10A-10D

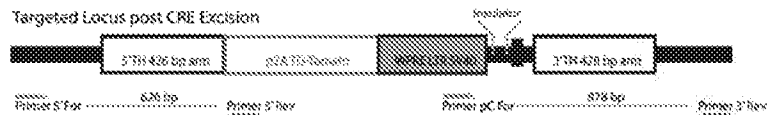
10 A.



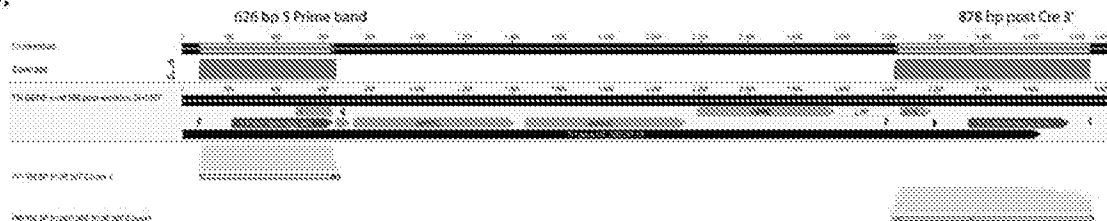
10 B.



10 C.

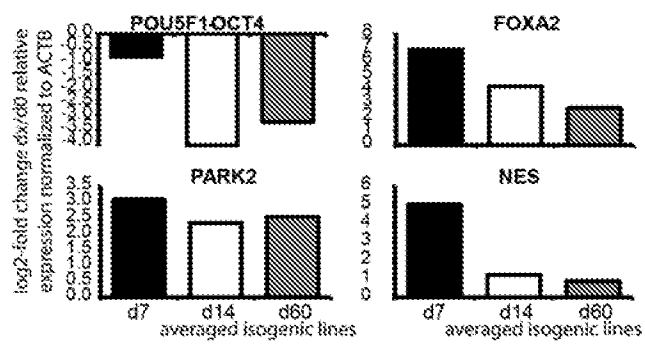


10 D.

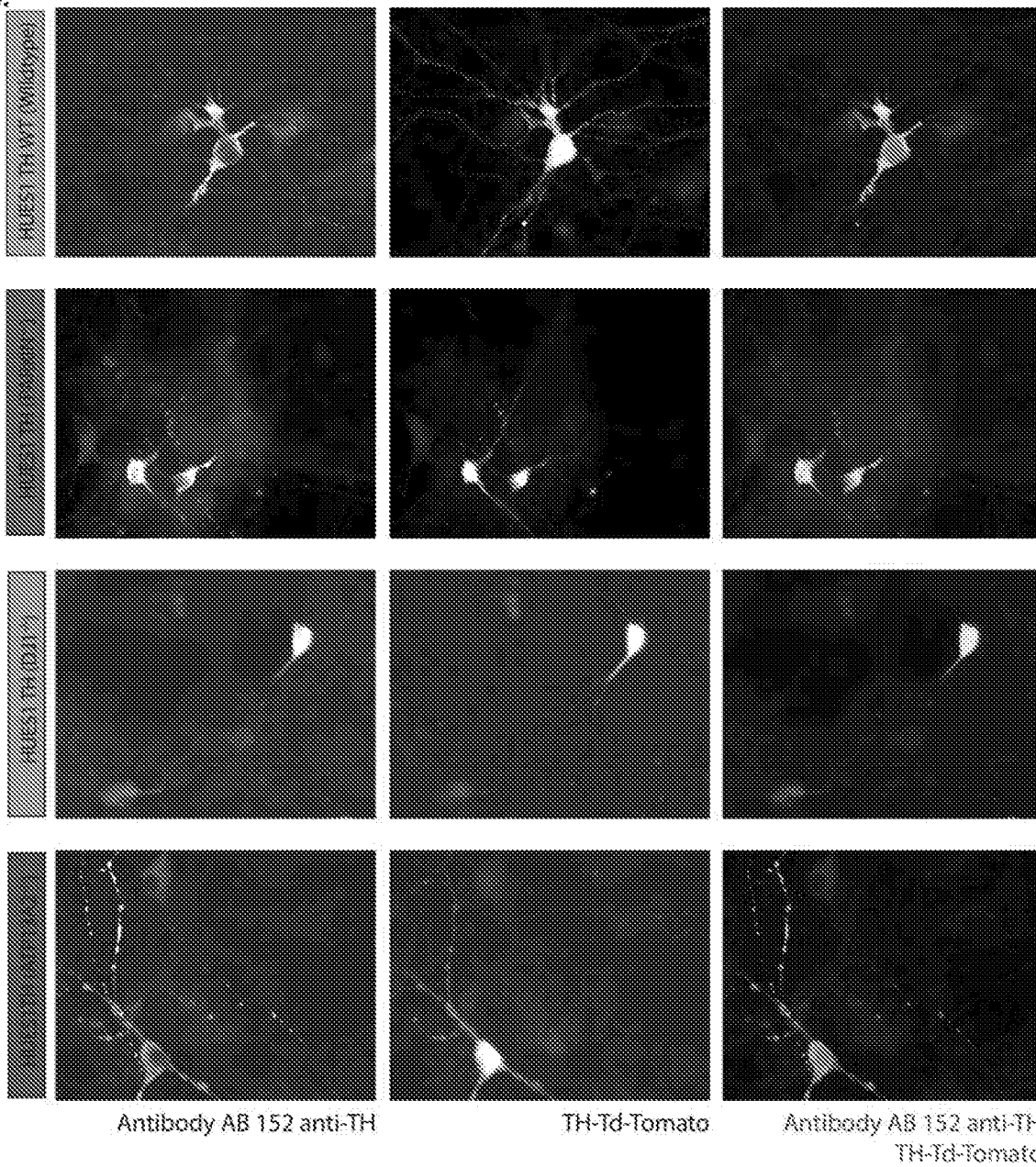


FIGS. 11A-11B

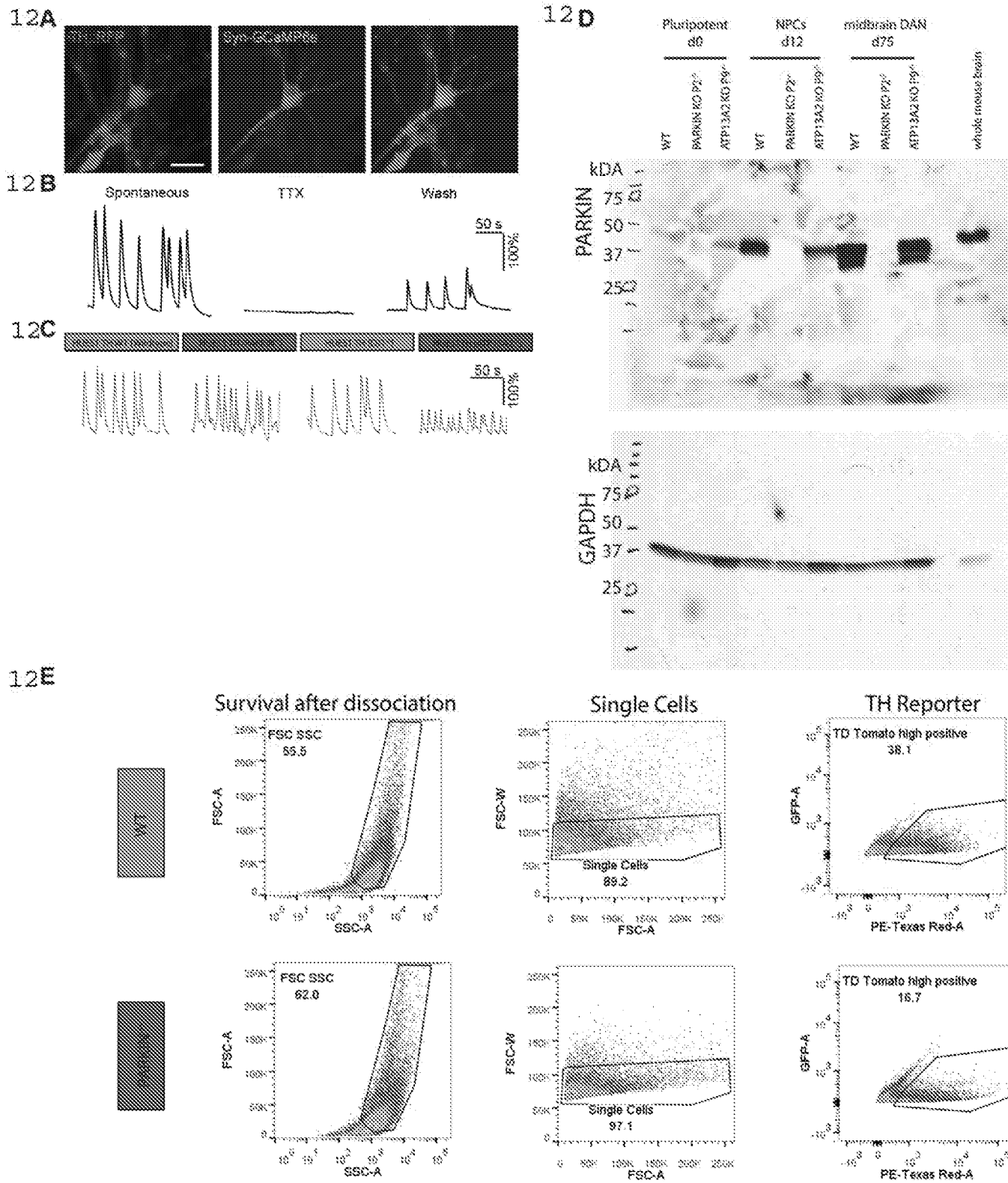
11A.



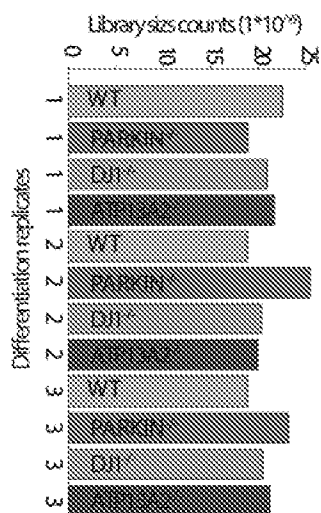
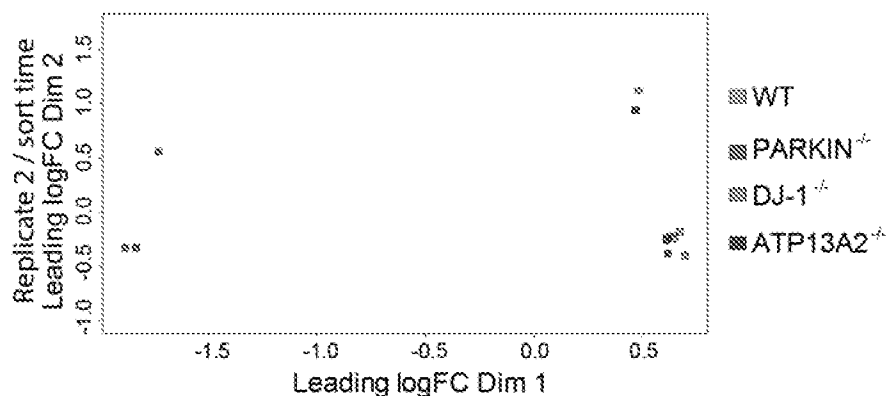
11B.



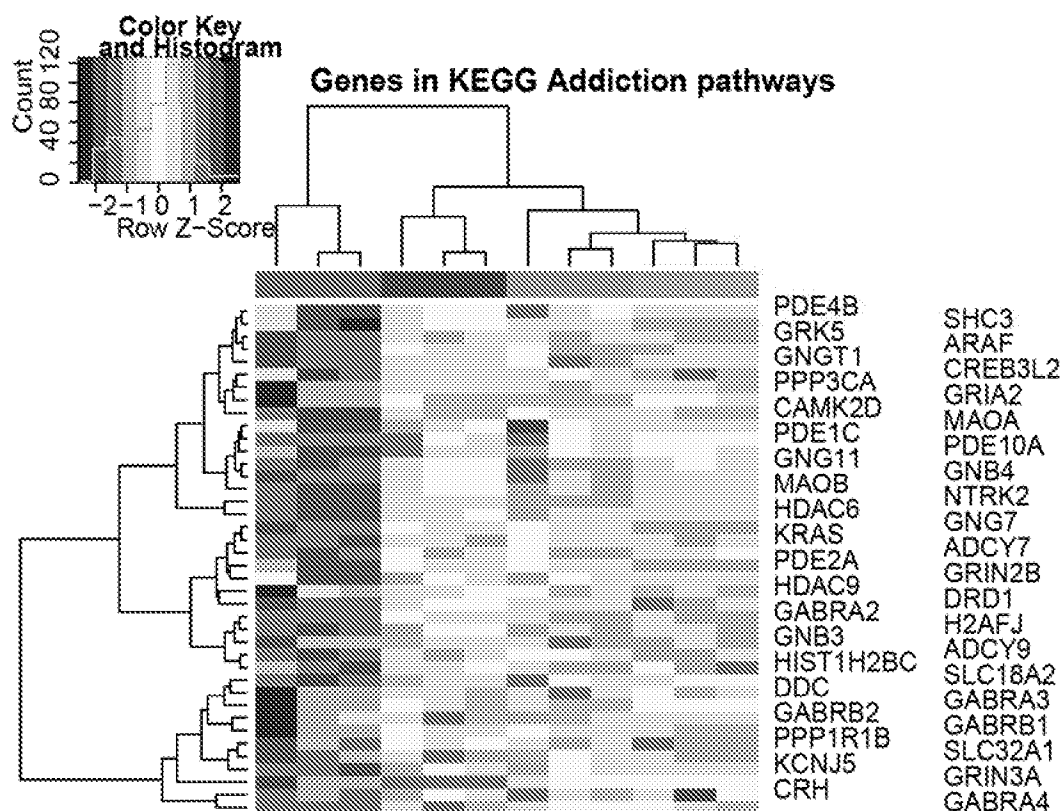
FIGS. 12A-12E



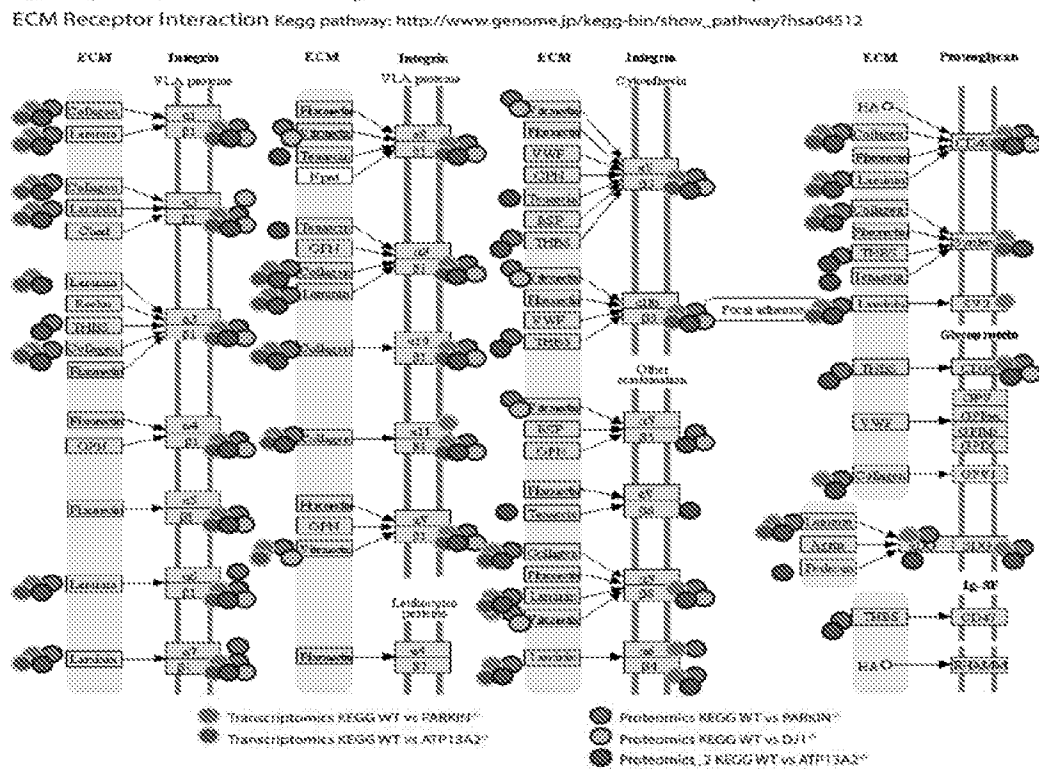
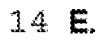
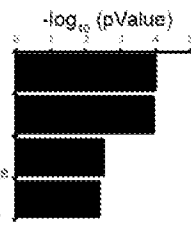
13A.

Replicate 2 / sort time
Leading logFC Dim 2

13C.

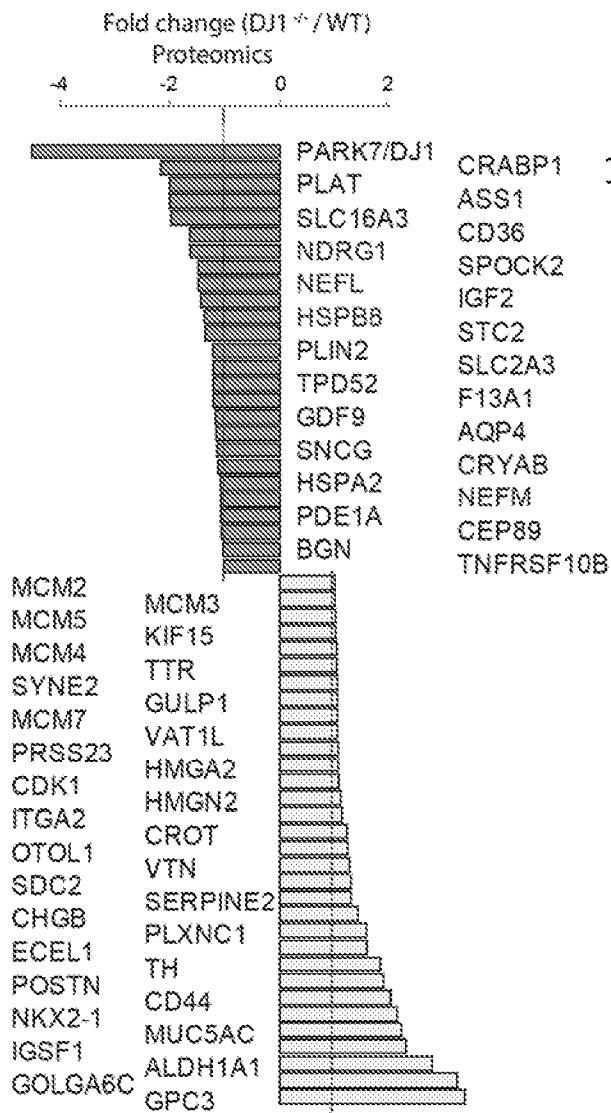


14A.



FIGS. 15A-15C

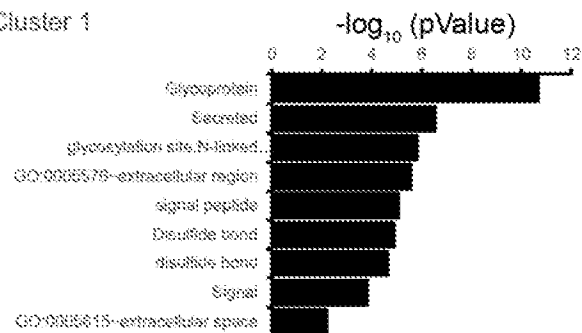
15A.



15B.

Charcot-Marie-Tooth	Gene Symbol	Description
OMT4D	NDRG1	Protein NDRG1
OMT1F/OMT2E	NEFL	Neurofilament light polypeptide
OMT2	HSPB8	Heat shock protein beta-8
	NEFM	Neurofilament medium polypeptide

15C

DJ1 vs WT 2-FOLD GO Functional Annotation
Cluster 1

Cluster 2

