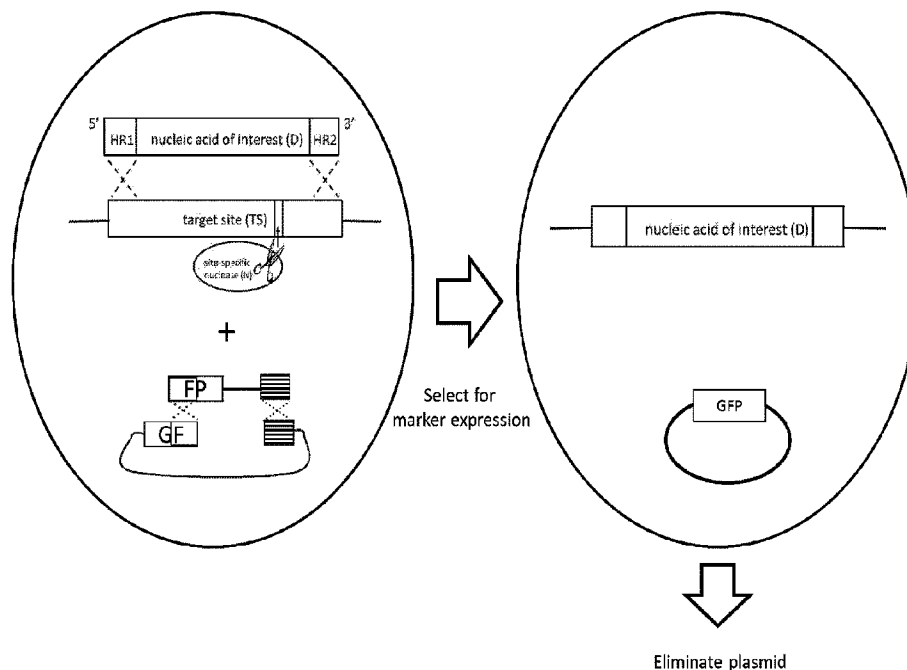




(86) Date de dépôt PCT/PCT Filing Date: 2014/12/19
(87) Date publication PCT/PCT Publication Date: 2015/06/25
(45) Date de délivrance/Issue Date: 2022/08/30
(85) Entrée phase nationale/National Entry: 2016/06/14
(86) N° demande PCT/PCT Application No.: US 2014/071693
(87) N° publication PCT/PCT Publication No.: 2015/095804
(30) Priorités/Priorities: 2013/12/19 (US61/918,625);
2014/02/07 (US61/937,444)

(51) Cl.Int./Int.Cl. *C12N 15/11* (2006.01),
C12N 15/90 (2006.01)
(72) Inventeurs/Inventors:
HORWITZ, ANDREW, US;
HAWKINS, KRISTY MICHELLE, US;
SCHUBERT, MAX, US;
SZETO, WAYNE, US
(73) Propriétaire/Owner:
AMYRIS, INC., US
(74) Agent: OSLER, HOSKIN & HARCOURT LLP

(54) Titre : PROCEDES D'INTEGRATION GENOMIQUE
(54) Title: METHODS FOR GENOMIC INTEGRATION



(57) **Abrégé/Abstract:**

Provided herein are methods of integrating one or more exogenous nucleic acids into one or more selected target sites of a host cell genome. In certain embodiments, the methods comprise contacting the host cell genome with one or more integration polynucleotides comprising an exogenous nucleic acid to be integrated into a genomic target site, a nuclease capable of causing a break at the genomic target site, and a linear nucleic acid capable of homologous recombination with itself or with one or more additional linear nucleic acids contacted with the population of cells, whereupon said homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker. In some embodiments, the methods further comprise selecting a host cell that expresses the selectable marker.

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

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
25 June 2015 (25.06.2015)



(10) International Publication Number
WO 2015/095804 A1

(51) International Patent Classification:

C12N 15/11 (2006.01) C12N 15/90 (2006.01)
C12Q 1/68 (2006.01)

(21) International Application Number:

PCT/US2014/071693

(22) International Filing Date:

19 December 2014 (19.12.2014)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

61/918,625 19 December 2013 (19.12.2013) US
61/937,444 7 February 2014 (07.02.2014) US

(71) Applicant: AMYRIS, INC. [US/US]; 5885 Hollis Street, Suite 100, Emeryville, California 94608 (US).

(72) Inventors: HORWITZ, Andrew; 5885 Hollis Street, Suite 100, Emeryville, California 94608 (US). HAWKINS, Kristy Michelle; 5885 Hollis Street, Suite 100, Emeryville, California 94608 (US). SCHUBERT, Max; 5885 Hollis Street, Suite 100, Emeryville, California 94608 (US). SZETO, Wayne; 5885 Hollis Street, Suite 100, Emeryville, California 94608 (US).

(74) Agents: PATHAK, Rahul et al.; Squire Patton Boggs (US) LLP, 275 Battery Street, Suite 2600, San Francisco, California 94111 (US).

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

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

Declarations under Rule 4.17:

- as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))
- as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))

[Continued on next page]

(54) Title: METHODS FOR GENOMIC INTEGRATION

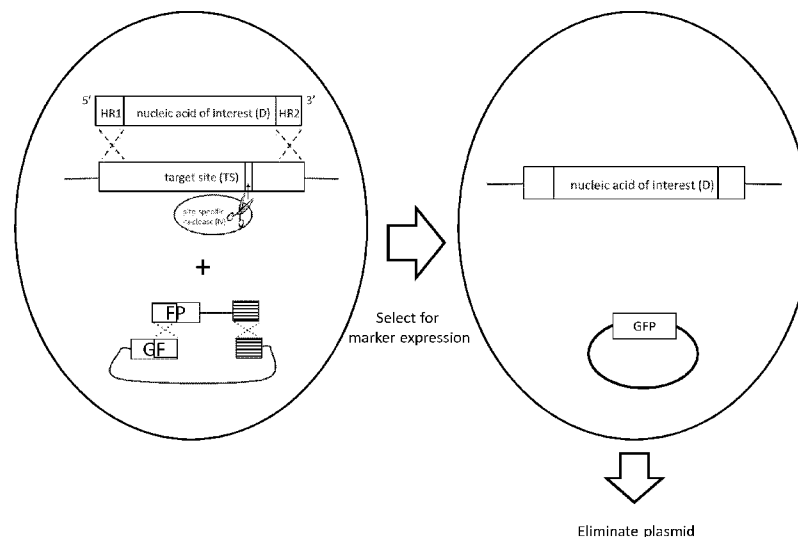


FIGURE 1

(57) Abstract: Provided herein are methods of integrating one or more exogenous nucleic acids into one or more selected target sites of a host cell genome. In certain embodiments, the methods comprise contacting the host cell genome with one or more integration polynucleotides comprising an exogenous nucleic acid to be integrated into a genomic target site, a nuclease capable of causing a break at the genomic target site, and a linear nucleic acid capable of homologous recombination with itself or with one or more additional linear nucleic acids contacted with the population of cells, whereupon said homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker. In some embodiments, the methods further comprise selecting a host cell that expresses the selectable marker.

WO 2015/095804 A1



Published:

- *with international search report (Art. 21(3))*
- *with sequence listing part of description (Rule 5.2(a))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

METHODS FOR GENOMIC INTEGRATION

This application claims the benefit of U.S. Provisional Application No. 61/918,625, filed December 19, 2013 and U.S. Provisional Application No. 61/937,444, filed February 7, 2014.

1. FIELD OF THE INVENTION

[0001] The methods and compositions provided herein generally relate to the fields of molecular biology and genetic engineering.

2. BACKGROUND

[0002] Genetic engineering techniques to introduce targeted modifications into a host cell genome find use in a variety of fields. Fundamentally, the determination of how genotype influences phenotype relies on the ability to introduce targeted insertions or deletions to impair or abolish native gene function. In the field of synthetic biology, the fabrication of genetically modified microbes capable of producing compounds of interest requires the insertion of customized DNA sequences into a chromosome of the host cell; industrial scale production generally requires the introduction of dozens of genes, *e.g.*, whole biosynthetic pathways, into a single host genome. In a therapeutic context, the ability to introduce precise genome modifications has enormous potential to address diseases resulting from single-gene defects, *e.g.*, X-linked severe combined immune deficiency (SCID), hemophilia B, beta-thalassemia, cystic fibrosis, muscular dystrophy and sickle-cell disease.

[0003] Recent advances in genome engineering have enabled the manipulation and/or introduction of virtually any gene across a diverse range of cell types and organisms. In particular, the advent of site-specific designer nucleases has enabled site-specific genetic modifications by introducing targeted breaks into a host cell genome, *i.e.*, genome editing. These nucleases include zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and clustered regulatory interspaced short palindromic repeats CRISPR/Cas (CRISPR-associated)-based RNA-guided endonucleases. ZFNs have been utilized, *inter alia*, to modify target loci in crops (Wright *et al.*, *Plant J* 44:693-705 (2005)), to improve mammalian cell culture lines expressing therapeutic antibodies (Malphettes *et al.*, *Biotechnol Bioeng* 106(5):774-783 (2010)), and to edit the human genome to evoke resistance to HIV (Urnov *et al.*, *Nat Rev Genet* 11(9):636-646 (2010)). Similarly, TALENs have been

utilized to modify a variety of genomes, including those of crop plants (Li, *et al.*, *Nat. Biotechnol.* 30: 390–392 (2012)), human, cattle, and mouse (Xu *et al.*, *Molecular Therapy—Nucleic Acids* 2, e112 (2013)). More recently, CRISPRs have been successfully utilized to edit the genomes of bacteria (*e.g.*, Jiang *et al.*, *Nature Biotechnology* 31(3):233-239 (2013); Qi *et al.*, *Cell*, 5, 1173–1183 (2013); yeast (*e.g.*, DiCarlo *et al.*, *Nucleic Acids Res.*, 7, 4336–4343 (2013)); zebrafish (*e.g.* Hwang *et al.*, *Nat. Biotechnol.*, 3, 227–229(2013)); fruit flies (*e.g.*, Gratz *et al.*, *Genetics*, 194, 1029–1035 (2013)); human cells (*e.g.*, Cong *et al.*, *Science* 6121, 819–823, (2013); Mali *et al.*, *Science*, 6121, 823–826 (2013); Cho *et al.*, *Nat. Biotechnol.*, 3, 230–232 (2013)); and plants (*e.g.*, Jiang *et al.*, *Nucleic Acids Research* 41(20):e188 (2013)); Belhaj *et al.*, *Plant Methods* 9(39) (2013)).

[0004] Site-specific nucleases induce breaks in chromosomal DNA that stimulate the host cell's cellular DNA repair mechanisms, including non-homologous end joining (NHEJ), single-strand annealing (SSA), and homology-directed repair (HDR). NHEJ-mediated repair of a nuclease-induced double-strand break (DSB) leads to the introduction of small deletions or insertions at the targeted site, leading to impairment or abolishment of gene function, *e.g.*, *via* frameshift mutations. The broken ends of the same molecule are rejoined by a multi-step enzymatic process that does not involve another DNA molecule. NHEJ is error prone and imprecise, producing mutant alleles with different and unpredictable insertions and deletions of variable size at the break-site during the repair. Similarly, SSA occurs when complementary strands from sequence repeats flanking the DSB anneal to each other, resulting in repair of the DSB but deletion of the intervening sequence. In contrast, HDR typically leads to an accurately restored molecule, as it relies on a separate undamaged molecule with homologous sequence to help repair the break. There are two major sources of homologous donor sequence native to the cell: the homologous chromosome, available throughout the cell cycle, and the sister chromatid of the broken molecule (which is only available after the DNA is replicated). However, genome engineering techniques routinely introduce exogenous donor DNAs that comprise regions homologous with the target site of the DSB, and can recombine with the target site. By including desired modifications to the target sequence within the exogenous donor, these modifications can be integrated into and replace the original target sequence via HDR.

[0005] Upon nuclease-induced breakage of DNA, the host cell's choice of repair pathways depends on a number of factors, and the outcome can dictate the precision of a desired genomic modification. Such factors include the DNA damage signaling pathways of the host

cell, the nature of the break, chromatin remodeling, transcription of specific repair proteins, and cyclin-dependent kinase activities present in later phases of the cell cycle. *See, e.g.,* Beucher *et al.*, *EMBO J* 28:3413-27(2009); Sørensen *et al.*, *Nat Cell Biol* 7:195-201 (2005); Jazayeri *et al.*, *Nat Cell Biol* 8:37-45 (2006); Huertas *et al.*, *Nature* 455:689-92 (2008); Moyal *et al.*, *Mol Cell* 41:529-42 (2011); and Chernikova *et al.*, *Radiat Res* 174:558-65 (2010). If a donor DNA with strong homology to the cleaved DNA is present, the chances of integration of the donor by homologous recombination increase significantly. *See, e.g.,* Moehle *et al.*, *Proc. Natl Acad. Sci. USA*, 9:3055–3060 (2007); Chen *et al.*, *Nat. Methods*, 9, 753–755 (2011). However, the overall frequency at which a homologous donor DNA is integrated via HDR into a cleaved target site, as opposed to non-integrative repair of the target site via NHEJ, can still be quite low. Recent studies suggest that HDR-mediated editing is generally a low efficiency event, and the less precise NHEJ can predominate as the mechanism of repair for DSBs.

[0006] For example, Mali *et al.* (*Science* 339:823-826 (2013)) attempted gene modification in human K562 cells using CRISPR (guide RNA and Cas9 endonuclease) and a concurrently supplied single-stranded donor DNA, and observed an HDR-mediated gene modification at the AAVS1 locus at a frequency of 2.0%, whereas NHEJ-mediated targeted mutagenesis at the same locus was observed at a frequency of 38%. Li *et al.* (*Nat Biotechnol.* (8):688-91 (2013)) attempted gene replacement in the plant *Nicotiana benthamiana* using CRISPR (guide RNA and Cas9 endonuclease) and a concurrently supplied double-stranded donor DNA, and observed an HDR-mediated gene replacement at a frequency of 9.0%, whereas NHEJ-mediated targeted mutagenesis was observed at a frequency of 14.2%. Kass *et al.* (*Proc Natl Acad Sci U S A.* 110(14): 5564–5569 (2013)) studied HDR in primary normal somatic cell types derived from diverse lineages, and observed that mouse embryonic and adult fibroblasts as well as cells derived from mammary epithelium, ovary, and neonatal brain underwent HDR at I-SceI endonuclease-induced DSBs at frequencies of approximately 1% (0.65-1.7%). Kass and others have reported higher HDR activity when cells are in S and G2 phases of the cell cycle. Li *et al.* (*Nat Biotechnol.* (8):688-91 (2013)) tested the possibility of enhancing HDR in *Nicotiana benthamiana* by triggering ectopic cell division, via co-expression of *Arabidopsis* CYCD3 (Cyclin D-Type 3), a master activator of the cell cycle; however, this hardly promoted the rate of HDR (up to 11.1% from 9% minus CYCD3). Strategies to improve HDR rates have also included knocking out the antagonistic NHEJ repair mechanism. For example, Qi *et al.* (*Genome Res* 23:547–554 (2013)) reported an increase of 5–16 fold in HDR-mediated gene targeting in *Arabidopsis* for the *ku70* mutant and 3–4 fold for the *lig4* mutant. However, the

overall rates were observed to be no higher than ~5%, with most less than 1%. Furthermore, once the desired gene-targeting event was produced, the *ku70* or *lig4* mutations had to be crossed out of the mutant plants.

[0007] Given the relatively low rate of HDR-mediated integration in most cell types, insertion of exogenous DNA into the chromosome typically requires the concomitant integration of a selectable marker, which enables enrichment for transformed cells that have undergone the desired integration event. However, this introduces extraneous sequences into the genome which may not be compatible with downstream applications, and prolonged expression of the marker may also have deleterious effects. For example, integration of the neomycin resistance gene into human cell genomes, followed by extended culturing times in G418, has been reported to cause changes to the cell's characteristics, and expression of enhanced green fluorescent protein (EGFP) and other fluorescent proteins has been reported to cause immunogenicity and toxicity. See, e.g., Barese *et al.*, *Human Gene Therapy* 22:659-668 (2011); Morris *et al.*, *Blood* 103:492-499 (2004); and Hanazono *et al.*, *Human Gene Therapy* 8:1313-1319 (1997). Additionally, the integration of selectable-marker genes in genetically modified (GM) plants has raised concerns of horizontal transfer to other organisms; in the case of antibiotic resistance markers, there is particular concern that these markers could lead to an increase in antibiotic resistant bacterial strains. A similar concern relates to the integration of herbicide-resistance markers and the possible creation of new aggressive weeds. At a minimum, removal of integrated marker sequences at later stages is time and labor intensive. This is particularly problematic where only a limited cache of selectable markers are available in a given host, and markers must be recycled to enable additional engineering steps. Thus, certain applications warrant introducing only the minimum exogenous sequences needed to effect a desired phenotype, e.g., for safety and/or regulatory compliance, and may ultimately require the avoidance of marker integration altogether.

[0008] Thus, there exists a need for methods and compositions that improve the efficiency and/or selection of HDR-mediated integration of one or more exogenous nucleic acids into a host cell genome. Moreover, there exists a need for genome engineering strategies that do not require co-integration of coding sequences for selectable markers. These and other needs are met by the compositions and methods provided herein.

3. SUMMARY

[0009] The methods and compositions provided herein relate to methods for selecting a homologous recombination (HR)-competent host cell. Without being bound by theory of operation, it is believed that HR-competence among a cell population can be selected for by selecting for a host cell that can homologously recombine one or more linear fragment(s), introduced into the host cell, to form a circular vector expressing a selectable marker. Here, this feature is exploited to enhance the identification of host cells that have site-specifically integrated, via HR, of one or more exogenous nucleic acids into the host cell's genome. In some embodiments, site-specific integration is enhanced by contacting the host cell genome with a site-specific nuclease that is capable of creating a break at the intended site of integration. Thus, by introducing to a host cell:

- (i) one or more exogenous nucleic acids having homologous regions to one or more target sites of the host cell genome;
 - (ii) one or more nucleases capable of selectively creating a break at the intended target site(s); and
 - (iii) a linear nucleic acid that can homologously recombine by itself, or with one or more additional linear DNA fragments introduced into the host cell, to form a circular, functional expression vector from which the selectable marker is expressed,
- and selecting for expression of the selectable marker, co-selection of cells that have integrated the one or more exogenous nucleic acids into their respective target site(s) is also achieved. The increased frequency of recovering host cells that have performed the desired integrations provided by the methods and compositions provided herein enables genetic engineering of otherwise difficult to engineer or intractable host cells, and improves the efficiency of higher order engineering designs, such as multiplex integrations.

[0010] Thus, in one aspect, provided herein is a method for integrating one or more exogenous nucleic acids into one or more target sites of a host cell genome, the method comprising contacting one or more host cells with one or more exogenous nucleic acids (ES) capable of recombining, via homologous recombination, at one or more target sites (TS) of the host cell genome; and one or more nucleases (N) capable of generating a break at each TS; and selecting a host cell competent for homologous recombination. In some embodiments, the selecting comprises selecting a host cell in which an exogenous nucleic acid has homologously recombined. In some embodiments, the nucleic acid that has homologously recombined is a

linear nucleic acid capable of homologous recombination with itself or with one or more additional linear nucleic acids. In some embodiments, the linear nucleic acid(s) form a circular nucleic acid upon homologous recombination. In some embodiments, the nucleic acid that has homologously recombined encodes a selectable marker. In some embodiments homologous recombination of the linear nucleic acid to form a circular nucleic acid forms a coding sequence for the selectable marker.

[0011] In another aspect, provided herein a method of selecting a host cell competent for homologous recombination, comprising:

(a) contacting one or more host cells with a linear nucleic acid capable of homologous recombination with itself or with one or more additional linear nucleic acids contacted with the population of cells, whereupon said homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker; and

(b) selecting a host cell that expresses the selectable marker.

[0012] In another aspect, provided herein is a method for integrating an exogenous nucleic acid into a target site of a host cell genome, the method comprising:

(a) contacting one or more host cells with:

(i) an exogenous nucleic acid (ES) capable of recombining, via homologous recombination, at the target site (TS) of the host cell genome;

(ii) a nuclease (N) capable of generating a break at TS; and

(iii) a linear nucleic acid capable of homologous recombination with itself or with one or more additional linear nucleic acids contacted with the host cell, whereupon said homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker;

and

(b) selecting a host cell that expresses the selectable marker.

[0013] In some embodiments, the linear nucleic acid comprises two internal homology regions that are capable of homologously recombining with each other, whereupon homologous recombination of the internal homology regions results in formation of the circular extrachromosomal nucleic acid expressing the selectable marker. In some

embodiments, the linear nucleic acid comprises a homology region that is capable of recombining with a homology region of an additional linear nucleic acid contacted with the host cell, whereupon homologous recombination of the two linear nucleic acids results in formation of the circular extrachromosomal nucleic acid expressing the selectable marker. In some embodiments, the linear nucleic acid comprises a partial, interrupted and/or non-contiguous coding sequence for the selectable marker, wherein the selectable marker cannot be expressed from the linear nucleic acid, whereupon said formation of the circular extrachromosomal nucleic acid results in formation of a complete coding sequence of the selectable marker, wherein the selectable marker can be expressed from the circular extrachromosomal nucleic acid.

[0014] In some embodiments, the contacted host cell(s) are cultured for a period of at least about 12, 24, 36, 48, 72 or more than 72 hours prior to said selecting. In some embodiments, the contacted cells are cultured under culturing conditions that select against the survival of cells not expressing the selectable marker. In some embodiments, said selecting of step (b) comprises detecting the expression of the selectable marker via visual, colorimetric or fluorescent detection methods. In some embodiments, the method further comprises the step of recovering a host cell wherein ES has homologously recombined at TS. In some embodiments, said recovering does not require integration of a selectable marker into the host cell genome. In some embodiments, said recovering occurs at a frequency of at least about one every 10, 9, 8, 7, 6, 5, 4, 3, or 2 contacted host cells, or clonal populations thereof, screened. In some embodiments, the method further comprises the step of eliminating the circular extrachromosomal nucleic acid from the selected host cell.

[0015] In some embodiments, the method comprises integrating a plurality of (n) exogenous nucleic acids into a plurality of (n) target sites of the host cell genome, wherein n is at least two, wherein step (a) comprises contacting the host cell with said plurality of exogenous nucleic acids, wherein x is an integer that varies from 1 to n, and for each integer x, each exogenous nucleic acid (ES)_x is capable of recombining, via homologous recombination, at a target site (TS)_x selected from said plurality of (n) target sites of said host cell genome; and for each said target site (TS)_x, the cell is also contacted with a nuclease (N)_x capable of generating a break at (TS)_x. In some embodiments, a single nuclease is capable of cleaving each (TS)_x. In some embodiments, n= 3, 4, 5, 6, 7, 8, 9 or 10. In some embodiments, (ES)_x comprises a first homology region (HR1)_x and a second homology region (HR2)_x, wherein (HR1)_x and (HR2)_x are capable of recombining, via homologous recombination, with a third

homology region (HR3)_x and a fourth homology region (HR4)_x, respectively, wherein (HR3)_x and (HR4)_x are each at TS. In some embodiments, (N)_x is capable of generating a single stranded break or a double stranded break at (TS)_x. In some embodiments, (ES)_x further comprises a nucleic acid of interest (D)_x. In some embodiments, (D)_x is selected from the group consisting of a selectable marker, a promoter, a nucleic acid sequence encoding an epitope tag, a gene of interest, a reporter gene, and a nucleic acid sequence encoding a termination codon. In some embodiments, (ES)_x is linear.

[0016] In some embodiments, the circular extrachromosomal nucleic acid further comprises a coding sequence for the nuclease. In some embodiments, the nuclease is an RNA-guided DNA endonuclease. In some embodiments, the RNA-guided DNA endonuclease is a Cas9 endonuclease. In some embodiments, the circular extrachromosomal nucleic acid further comprises a sequence that encodes a crRNA activity and a tracrRNA activity that enables site-specific recognition and cleavage of TS by the RNA-guided DNA endonuclease. In some embodiments, the crRNA activity and the tracrRNA activity are expressed as a single contiguous RNA molecule.

[0017] In some embodiments, the nuclease is selected from the group consisting of an endonuclease, a zinc finger nuclease, a TAL-effector DNA binding domain-nuclease fusion protein (TALEN), a transposase, and a site-specific recombinase. In some embodiments, the zinc finger nuclease is a fusion protein comprising the cleavage domain of a TypeIIS restriction endonuclease fused to an engineered zinc finger binding domain. In some embodiments, the TypeIIS restriction endonuclease is selected from the group consisting of HO endonuclease and Fok I endonuclease. In some embodiments, the zinc finger binding domain comprises 3, 5 or 6 zinc fingers.

[0018] In some embodiments, the endonuclease is a homing endonuclease selected from the group consisting of: an LAGLIDADG (SEQ ID NO:1) homing endonuclease, an HNH homing endonuclease, a His-Cys box homing endonuclease, a GIY-YIG (SEQ ID NO:2) homing endonuclease, and a cyanobacterial homing endonuclease. In some embodiments, the endonuclease is selected from the group consisting of: H-DreI, I-SceI, I-SceII, I-SceIII, I-SceIV, I-SceV, I-SceVI, I-SceVII, I-CeuI, I-CeuAIIIP, I-CreI, I-CrepsbIP, I-CrepsbIIP, I-CrepsbIIIP, I-CrepsbIIVP, I-TliI, I-PpoI, Pi-PspI, F-SceI, F-SceII, F-SuvI, F-CphI, F-TevI, F-TevII, I-AmaI, I-AniI, I-ChuI, I-CmoI, I-CpaI, I-CpaII, I-CsmI, I-CvuI, I-CvuAIP, I-DdiI, I-DdiII, I-DirI, I-DmoI, I-HmuI, I-HmuII, I-HsNIP, I-LlaI, I-MsoI, I-NaaI, I-NanI, I-NclIP, I-NgrIP, I-NitI, I-NjaI, I-Nsp236IP, I-PakI, I-PboIP, I-PcuIP, I-PcuAI, I-PcuVI, I-PgrIP, I-

PobIP, I-PorI, I-PorIIP, I-PbpIP, I-SpBetaIP, I-ScaI, I-SexIP, I-SneIP, I-SpomI, I-SpomCP, I-SpomIP, I-SpomIIP, I-SquIP, I-Ssp68031, I-SthPhiJP, I-SthPhiST3P, I-SthPhiSTe3bP, I-TdeIP, I-TevI, I-TevII, I-TevIII, i-UarAP, i-UarHGPAIP, I-UarHGPA13P, I-VinIP, I-ZbiIP, PI-MgaI, PI-MtuI, PI-MtuHIP PI-MtuHIIP, PI-PfuI, PI-PfuII, PI-PkoI, PI-PkoII, PI-Rma43812IP, PI-SpBetaIP, PI-SceI, PI-TfuI, PI-TfuII, PI-ThyI, PI-TliI, or PI-TliII. In some embodiments, the endonuclease is modified to specifically bind an endogenous genomic sequence, wherein the modified endonuclease no longer binds to its wild type endonuclease recognition sequence. In some embodiments, the modified endonuclease is derived from a homing endonuclease selected from the group consisting of: an LAGLIDADG (SEQ ID NO:1) homing endonuclease, an HNH homing endonuclease, a His-Cys box homing endonuclease, a GIY-YIG (SEQ ID NO:2) homing endonuclease, and a cyanobacterial homing endonuclease. In some embodiments, the modified endonuclease is derived from an endonuclease selected from the group consisting of: H-DreI, I-SceI, I-SceII, I-SceIII, I-SceIV, I-SceV, I-SceVI, I-SceVII, I-CeuI, I-CeuAIIP, I-CreI, I-CrepsbIP, I-CrepsbIIP, I-CrepsbIIIP, I-CrepsbIVP, I-TliI, I-PpoI, Pi-PspI, F-SceI, F-SceII, F-SuvI, F-CphI, F-TevI, F-TevII, I-AmaI, I-AniI, I-ChuI, I-Cmoel, I-CpaI, I-CpaII, I-Csml, I-CvuI, I-CvuAIP, I-DdiI, I-DdiII, I-DirI, I-DmoI, I-HmuI, I-HmuII, I-HsNIP, I-LlaI, I-MsoI, I-NaaI, I-NanI, I-NclIP, I-NgrIP, I-NitI, I-NjaI, I-Nsp236IP, I-PakI, I-PboIP, I-PcuIP, I-PcuAI, I-PcuVI, I-PgrIP, I-PobIP, I-PorI, I-PorIIP, I-PbpIP, I-SpBetaIP, I-ScaI, I-SexIP, I-SneIP, I-SpomI, I-SpomCP, I-SpomIP, I-SpomIIP, I-SquIP, I-Ssp68031, I-SthPhiJP, I-SthPhiST3P, I-SthPhiSTe3bP, I-TdeIP, I-TevI, I-TevII, I-TevIII, i-UarAP, i-UarHGPAIP, I-UarHGPA13P, I-VinIP, I-ZbiIP, PI-MgaI, PI-MtuI, PI-MtuHIP PI-MtuHIIP, PI-PfuI, PI-PfuII, PI-PkoI, PI-PkoII, PI-Rma43812IP, PI-SpBetaIP, PI-SceI, PI-TfuI, PI-TfuII, PI-ThyI, PI-TliI, or PI-TliII.

[0019] In some embodiments, the host cell is a prokaryotic cell. In some embodiments, the host cell is a eukaryotic cell. In some embodiments, the host cell is selected from the group consisting of a fungal cell, a bacterial cell, a plant cell, an insect cell, an avian cell, a fish cell and a mammalian cell. In some embodiments, the host cell is a mammalian cell selected from the group consisting of a rodent cell, a primate cell and a human cell. In some embodiments, the host cell is a yeast cell. In some embodiments, the yeast is *Saccharomyces cerevisiae*.

[0020] In another aspect, provided herein is a host cell comprising: an exogenous nucleic acid (ES) capable of recombining, via homologous recombination, at a target site (TS) of the host cell genome; a nuclease (N) capable of generating a break at TS; and a linear nucleic acid capable of homologous recombination with itself or with one or more additional linear nucleic

acid within the host cell, whereupon said homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker. In some embodiments, the linear nucleic acid comprises two internal homology regions that are capable of homologously recombining with each other, whereupon homologous recombination of the internal homology regions results in formation of the circular extrachromosomal nucleic acid expressing the selectable marker. In some embodiments, the linear nucleic acid comprises a homology region that is capable of recombining with a homology region of an additional linear nucleic acid within the host cell, whereupon homologous recombination of the two linear nucleic acids results in formation of the circular extrachromosomal nucleic acid expressing the selectable marker. In some embodiments, the linear nucleic acid comprises a partial, interrupted and/or non-contiguous coding sequence for the selectable marker, wherein the selectable marker cannot be expressed from the linear nucleic acid, whereupon said formation of the circular extrachromosomal nucleic acid results in formation of a complete coding sequence of the selectable marker, wherein the selectable marker can be expressed from the circular extrachromosomal nucleic acid.

[0021] In another aspect, provided herein is a composition comprising: a site-specific nuclease, or a nucleic acid comprising a coding sequence for a site-specific nuclease; and a linear nucleic acid comprising two internal homology regions that are capable of homologously recombining with each other in a host cell, whereupon homologous recombination of the internal homology regions results in formation of a circular nucleic acid comprising a coding sequence for a selectable marker. In some embodiments, the linear nucleic acid comprises a partial, interrupted and/or non-contiguous coding sequence for the selectable marker, wherein the selectable marker cannot be expressed from the linear nucleic acid in a host cell, whereupon said formation of the circular nucleic acid results in formation of a complete coding sequence of the selectable marker, wherein the selectable marker can be expressed from the circular nucleic acid in a host cell.

[0022] In another aspect, provided herein is a composition comprising a site-specific nuclease, or a nucleic acid comprising a coding sequence for a site-specific nuclease; and a first linear nucleic acid and one or more additional linear nucleic acids, wherein the first and second linear nucleic acids are capable of homologously recombining with each other in a host cell, whereupon said homologous recombination results in formation of a circular nucleic acid comprising a coding sequence for a selectable marker. In some embodiments, each linear nucleic acid comprises a partial, interrupted and/or non-contiguous coding sequence for the

selectable marker, wherein the selectable marker cannot be expressed from each linear nucleic acid in a host cell, whereupon said formation of the circular nucleic acid results in formation of a complete coding sequence of the selectable marker, wherein the selectable marker can be expressed from the circular nucleic acid in a host cell. In some embodiments, the circular nucleic acid further comprises a coding sequence for a site-specific nuclease. In some embodiments, the site-specific nuclease is an RNA-guided DNA endonuclease. In some embodiments, the RNA-guided DNA endonuclease is a Cas9 endonuclease. In some embodiments, the compositions further comprise a ribonucleic acid comprising a crRNA activity and a ribonucleic acid comprising a tracrRNA activity; or a deoxyribonucleic acid that encodes a ribonucleic acid comprising a crRNA activity and a deoxyribonucleic acid that encodes a ribonucleic acid comprising a tracrRNA activity. In some embodiments, the circular nucleic acid further comprises a deoxyribonucleic acid that encodes a ribonucleic acid comprising a crRNA activity and a deoxyribonucleic acid that encodes a ribonucleic acid comprising a tracrRNA activity. In some embodiments, the deoxyribonucleic acid that encodes the crRNA activity and the tracrRNA activity encodes said activities on a single contiguous RNA molecule. In other embodiments, the site-specific nuclease is selected from the group consisting of an endonuclease, a zinc finger nuclease, a TAL-effector DNA binding domain-nuclease fusion protein (TALEN), a transposase, and a site-specific recombinase. Also provided herein is a host cell comprising any of the aforementioned compositions. Also provided herein is a cell culture composition comprising a cell culture medium and any of the host cells described herein. In some embodiments, the cell culture composition further comprises a compound that selects for expression of the selectable marker.

[0023] In another aspect, also provided herein is a linear nucleic acid comprising a first homology region (HR1) and a second homology region (HR2), wherein HR1 and HR2 are capable of recombining with each other via homologous recombination, whereupon homologous recombination of HR1 with HR2 results in formation of a circular nucleic acid comprising a coding sequence for a selectable marker. In some embodiments, HR1 comprises a first incomplete coding sequence of the selectable marker and HR2 comprises a second incomplete coding sequence of the selectable marker, and homologous recombination of HR1 with HR2 results in reconstitution of a complete coding sequence for the selectable marker. In some embodiments, the linear nucleic acid further comprises a coding sequence for a site-specific nuclease described herein.

[0024] Also provided herein are methods and compositions for genomic integration of one or more donor DNAs into a host cell genome mediated by site-specific RNA guided endonucleases (RGEN), for example, CRISPR/Cas9. In one aspect, provided herein is a method for integrating one or more exogenous nucleic acids into one or more target sites of a host cell genome, the method comprising:

- (a) contacting one or more host cells with:
 - (i) one or more exogenous donor nucleic acids (ES) capable of recombining, via homologous recombination, at one or more target sites (TS) of the host cell genome;
 - (ii) an RNA-guided endonuclease (RGEN);
 - (iii) one or more ribonucleic acids that enable site-specific recognition and cleavage of the one or more TS by the RGEN; and
 - (iv) a linear pre-recombination nucleic acid capable of homologous recombination with itself or with one or more additional linear pre-recombination nucleic acids contacted with the host cell, whereupon said homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker;
- and
- (b) selecting a host cell that expresses the selectable marker, thereby selecting for a cell that has integrated the one or more exogenous nucleic acids into the one or more target sites of a host cell genome.

[0025] In some embodiments, the homologous recombination results in formation of a complete coding sequence of the selectable marker within the circular extrachromosomal nucleic acid. In some embodiments, at least one linear pre-recombination nucleic acid comprises a sequence that encodes the one or more ribonucleic acids that enables site-specific recognition and cleavage of TS by the RNA-guided DNA endonuclease. In some embodiments, the one or more ribonucleic acids comprise a crRNA activity and a tracrRNA on a single contiguous guide RNA (gRNA) molecule. In some embodiments, at least one linear pre-recombination nucleic acid comprises a sequence that encodes the RNA-guided DNA endonuclease. In some embodiments, the RNA-guided DNA endonuclease is Cas9.

[0026] In some embodiments, the formation of the circular extrachromosomal nucleic acid results from homologous recombination of two or three linear pre-recombination nucleic acids. In some embodiments, the one or more linear pre-recombination nucleic acids are generated *in vivo* by RGEN cleavage of one or more circular nucleic acids comprising the one or more pre-recombination nucleic acids. In some embodiments, a plurality of (n) exogenous nucleic acids is integrated into a plurality of (n) target sites of the host cell genome, wherein n is at least two, wherein step (a) comprises contacting the host cell with:

(i) said plurality of exogenous nucleic acids, wherein x is an integer that varies from 1 to n, and for each integer x, each exogenous nucleic acid (ES)_x is capable of recombining, via homologous recombination, at a target site (TS)_x selected from said plurality of (n) target sites of said host cell genome;

(ii) for each said target site (TS)_x, a guide RNA (gRNA)_x that enables site-specific recognition and cleavage of (TS)_x by the RGEN.

[0027] In some embodiments, the selectable marker is a drug resistance marker, a fluorescent protein or a protein detectable by colorimetric or fluorescent detection methods. In some embodiments, ES further comprises a nucleic acid of interest D. In some embodiments, D is selected from the group consisting of a selectable marker, a promoter, a nucleic acid sequence encoding an epitope tag, a gene of interest, a reporter gene, and a nucleic acid sequence encoding a termination codon. In some embodiments, the host cell is selected from the group consisting of a fungal cell, a bacterial cell, a plant cell, an insect cell, an avian cell, a fish cell and a mammalian cell. In some embodiments, the contacted host cell(s) are cultured for a period of at least about 12, 24, 36, 48, 72 or more than 72 hours prior to said selecting. In some embodiments, the contacted cells are cultured under culturing conditions that select against the survival of cells not expressing the selectable marker. In some embodiments, the selecting of step (b) comprises detecting the expression of the selectable marker via visual, colorimetric or fluorescent detection methods.

[0028] In another aspect, provided herein is a composition for integrating one or more exogenous nucleic acids into one or more target sites of a host cell genome, the composition comprising:

(a) one or more exogenous donor nucleic acids (ES) capable of recombining, via homologous recombination, at one or more target sites (TS) of a host cell genome;

- (b) an RNA-guided endonuclease (RGEN), or a nucleic acid encoding said RGEN;
- (c) one or more ribonucleic acids that enable site-specific recognition and cleavage of the one or more TS by the RGEN, or one or more nucleic acids encoding said one or more ribonucleic acids; and
- (d) a linear pre-recombination nucleic acid capable of *in vivo* homologous recombination with itself or with one or more additional linear pre-recombination nucleic acids in the composition, whereupon said *in vivo* homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker.

[0029] In some embodiments, said homologous recombination results in formation of a complete coding sequence of the selectable marker within the circular extrachromosomal nucleic acid. In some embodiments, at least one linear pre-recombination nucleic acid comprises a sequence that encodes the one or more ribonucleic acids that enables site-specific recognition and cleavage of TS by the RNA-guided DNA endonuclease. In some embodiments, the one or more ribonucleic acid molecules comprise a crRNA activity and a tracrRNA activity on a single contiguous guide RNA (gRNA) molecule. In some embodiments, at least one linear pre-recombination nucleic acid comprises a sequence that encodes the RNA-guided DNA endonuclease. In some embodiments, the RNA-guided DNA endonuclease is Cas9. In some embodiments, the composition comprises two or three linear pre-recombination nucleic acids capable of homologously recombining to form the circular extrachromosomal nucleic acid. In some embodiments, the one or more linear pre-recombination nucleic acids are generated *in vivo* by RGEN cleavage of one or more circular nucleic acids comprising the one or more pre-recombination nucleic acids.

[0030] In some embodiments, the composition comprises:

- (a) a plurality of (n) exogenous nucleic acids capable of integrating into a plurality of (n) target sites of the host cell genome, wherein n is at least two, wherein x is an integer that varies from 1 to n, and for each integer x, each exogenous nucleic acid (ES)_x is capable of recombining, via homologous recombination, at a target site (TS)_x selected from said plurality of (n) target sites of said host cell genome; and
- (b) for each said target site (TS)_x, a guide RNA (gRNA)_x that enables site-specific recognition and cleavage of (TS)_x by the RGEN.

[0031] In some embodiments, the selectable marker is a drug resistance marker, a fluorescent protein or a protein detectable by colorimetric or fluorescent detection methods. In some embodiments, ES further comprises a nucleic acid of interest D. In some embodiments, D is selected from the group consisting of a selectable marker, a promoter, a nucleic acid sequence encoding an epitope tag, a gene of interest, a reporter gene, and a nucleic acid sequence encoding a termination codon.

[0032] In another aspect, provided herein is a host cell comprising any of the compositions for the RGEN-mediated integration of one or more exogenous nucleic acids into one or more target sites of a host cell genome described herein. In some embodiments, the host cell is selected from the group consisting of a fungal cell, a bacterial cell, a plant cell, an insect cell, an avian cell, a fish cell and a mammalian cell.

4. BRIEF DESCRIPTION OF THE FIGURES

[0033] FIG. 1 provides an exemplary embodiment of genomic integration in a host cell of an exogenous nucleic acid (D) using a site-specific nuclease (N) and two pre-recombination molecules capable of homologous recombination (HR) with each other to form a circular plasmid comprising a coding sequence for a selectable marker (GFP). In this example, the coding sequence for green fluorescent protein (GFP) is split among the two pre-recombination molecules, and the coding sequence is reconstituted *in vivo* upon HR of overlapping homology regions between the two molecules. Selection for expression of the selectable marker also selects for cells which have integrated the exogenous nucleic acid into its target site, and following selection, the plasmid comprising the selectable marker can be eliminated. HR1 – upstream homology region; HR2 – downstream homology region; TS – target site; N – site-specific nuclease; D – nucleic acid of interest.

[0034] FIG. 2 provides an exemplary embodiment of simultaneous genomic integration in a host cell of a plurality of exogenous nucleic acids using a plurality of site-specific nucleases and two pre-recombination molecules capable of homologous recombination with each other to form a circular plasmid comprising a coding sequence for a selectable marker (GFP). In this example, two pre-recombination molecules are simultaneously introduced with three exogenous donor DNAs, each having homology regions specific to a unique target site in the host cell genome, and one or more nucleases capable of cleaving at the three target sites. Selection for expression of the selectable marker also selects for cells which have integrated each exogenous nucleic acid into its respective target site. HR1 – upstream homology region;

HR2 – downstream homology region; TS – target site; N – site-specific nuclease; D – nucleic acid of interest.

[0035] **FIG. 3** provides two exemplary embodiments for selecting cells capable of HR-mediated assembly of pre-recombination molecules and targeted genomic integration of exogenous donor DNA. **FIG. 3A** depicts a selection strategy based on HR-mediated formation of a plasmid comprising a fluorescence-based selectable marker. Host cells are transformed with one or more exogenous donor DNAs, one or more pre-recombination molecules, which upon *in vivo* HR-mediated assembly, forms a circular plasmid comprising a fluorescence-based selectable marker (e.g., GFP), and optionally, one or more site-specific nucleases capable of cleaving one or more target sites of the host cell genome. HR competent cells are marked by fluorescence, and can be isolated from the host cell population using standard techniques such as flow cytometry. **FIG. 3B** depicts a selection strategy based on HR-mediated formation of a drug-based selectable marker. In this embodiment, HR-competent cells are marked by drug resistance and survival when cultured in media containing the appropriate selective agent, whereas non-HR competent drug-sensitive cells are eliminated. Cells or clonal cell populations isolated under either selection scheme can be expanded and confirmed for harboring the targeted integration of one or more exogenous donor DNAs, for example, by PCR and/or sequencing of the genomic target regions.

[0036] **FIG. 4** provides exemplary pre-recombination compositions useful in the methods of genomic integration provided herein. For any of the pre-recombination compositions described herein, the compositions can be transformed directly into a host cell as linear nucleic acid molecules, or alternatively, parental circular molecules comprising the pre-recombination molecules can be introduced into the host cell and cleaved *in vivo* by one or more nucleases to liberate the pre-recombination molecules. In an HR-competent host cell, the linear pre-recombination molecule(s) homologously recombine to form a circular vector comprising a selectable marker. **FIG. 4A** depicts two exemplary embodiments of a pre-recombination molecule for 1-piece *in vivo* assembly of the marker plasmid. (L) A single linear pre-recombination molecule can comprise two overlapping homology regions (represented by vertical striped boxes) outside of, i.e., non-inclusive of an intact coding sequence of the selectable marker (represented by GFP). (C) Alternatively, the single linear pre-recombination molecule can comprise two overlapping homology regions which each comprise a partial coding sequence of a selectable marker (GF and FP, respectively; overlap shaded in gray). (R) For both embodiments, *in vivo* homologous recombination of the single linear pre-

recombination molecule with itself results in the formation of a circular plasmid comprising the complete coding sequence of the selectable marker. In some embodiments, the single linear pre-recombination molecule can further comprise a coding sequence for a site-specific nuclease (not shown).

[0037] FIG. 4B depicts two exemplary embodiments of pre-recombination molecule compositions for 2-piece *in vivo* assembly of the marker plasmid. Two linear pre-recombination molecules can each comprise two non-overlapping homology regions (represented by vertically and horizontally striped boxes, respectively), with each homology region being homologous to a homology region of the other pre-recombination molecule. (L) One of the two linear pre-recombination molecules can comprise an intact coding sequence of a selectable marker (represented by GFP) separate from the two non-overlapping homology regions. (C) Alternatively, each of the two pre-recombination molecules can comprise a partial coding sequence of the selectable marker having homology to a partial marker coding sequence on the other pre-recombination molecule (GF and FP, respectively; overlap shaded in gray). (R) For both embodiments, *in vivo* homologous recombination of the two linear pre-recombination molecules with each other results in formation of a circular plasmid comprising the complete coding sequence of the selectable marker. In some embodiments, one of the linear pre-recombination molecules can further comprise a complete coding sequence for a site-specific nuclease, or alternatively, each of the two pre-recombination molecules can comprise a partial nuclease coding sequence having homology to a partial nuclease coding sequence on the other pre-recombination molecule (not shown). Such an embodiment may be useful where nuclease expression is desired only in HR-competent cells, for example, to reduce the incidence of nuclease-mediated NHEJ in non-HR-competent cells.

[0038] FIG. 4C depicts two exemplary embodiments of pre-recombination molecule compositions for 3-piece *in vivo* assembly of the marker plasmid. The three linear pre-recombination molecules can each comprise two non-overlapping homology regions (represented by vertically and horizontally striped boxes; a vertically striped box and a diamond-filled box; and a horizontally striped box and a diamond-filled box, respectively), with each homology region being homologous to a homology region on one of the other pre-recombination molecules. (L) One of the three linear pre-recombination molecules can comprise an intact coding sequence of a selectable marker (represented by GFP) separate from the two non-overlapping homology regions. (C) Alternatively, each of at least two pre-recombination molecules can comprise a partial coding sequence of the selectable marker

having homology to a partial marker coding sequence on one other pre-recombination molecule (GF and FP, respectively). (R) For both embodiments, homologous recombination of the three linear pre-recombination molecules with each other results in the formation of a circular plasmid comprising the complete coding sequence of the selectable marker. In some embodiments, as in 2-piece assembly, one of the linear pre-recombination molecules can further comprise a complete coding sequence for a site-specific nuclease, or alternatively, each of at least two pre-recombination molecules can comprise a partial nuclease coding sequence having homology to a partial nuclease coding sequence on one other pre-recombination molecule (not shown).

[0039] **FIG. 5** provides exemplary pre-recombination compositions useful in RNA-guided DNA endonuclease (RGEN) specific embodiments of the methods of genomic integration provided herein. **FIG. 5A:** In some embodiments of a 1-piece *in vivo* assembly depicted in **FIG. 4A**, the single pre-recombination molecule can further comprise one or more sequences that encode a crRNA activity and a tracrRNA activity (e.g. a guide RNA (gRNA) sequence) that enables site-specific recognition and cleavage of a genomic target site by an RGEN (e.g., CRISPR/Cas9). In some embodiments, the pre-recombination molecule can further comprise a coding sequence for the RGEN (e.g., Cas9; not shown). **FIG. 5B:** In some embodiments of a 2-piece *in vivo* assembly depicted in **FIG. 4B**, one of the two pre-recombination molecules can further comprise a gRNA sequence. In other embodiments, one of the two pre-recombination molecules can further comprise a complete coding sequence of an RGEN, or alternatively, one of the two pre-recombination molecules can comprise a partial nuclease coding sequence having homology to a partial nuclease coding sequence on the other pre-recombination molecule (not shown). **FIG. 5C:** In some embodiments of a 3-piece *in vivo* assembly, one of the three pre-recombination molecules can further comprise a gRNA sequence. In other embodiments, one of the three pre-recombination molecules can further comprise a complete coding sequence of an RGEN, or alternatively, each of at least two pre-recombination molecules can comprise a partial nuclease coding sequence having homology to a partial nuclease coding sequence on one other pre-recombination molecule (not shown).

[0040] **FIG. 6** provides exemplary linear pre-recombination compositions useful in RGEN-mediated multiplex genomic integration. **FIG. 6A:** In some embodiments of a 2-piece *in vivo* assembly where one of the two pre-recombination molecules participating in the assembly comprises a gRNA sequence, several of these molecules can be provided at once (e.g., 3: gRNA-1, gRNA-2, gRNA-3), each comprising a unique gRNA sequence that targets a

different genomic target site. In this embodiment, the other pre-recombination molecule represents a common vector backbone that may comprise a complete coding sequence for a selectable marker (L), or a partial marker coding sequence that is homologous with a partial coding sequence common to each of the gRNA containing fragments (C). (R) HR-competent cells are able to recombine each unique gRNA containing fragment with the common vector backbone to reconstitute three different marker plasmids each comprising a unique gRNA sequence. **FIG. 6B:** In other embodiments, one of the two linear pre-recombination molecules can further comprise a complete coding sequence of an RGEN (e.g. Cas9). Alternatively, each of the two pre-recombination molecules can comprise a partial nuclease coding sequence having homology to a partial nuclease coding sequence on the other pre-recombination molecule (not shown). Multiplex genomic integrations can be performed with 2-piece, 3-piece, or higher order pre-recombination compositions, in combination with a plurality of unique gRNA cassettes positioned within one or more of the pre-recombination molecules of the composition.

[0041] **FIG. 7** depicts compositions used in determining optimal modes of gRNA delivery for CRISPR/Cas-9 mediated multiplex donor DNA integrations as described in Example 1. (L) Unique gRNA cassettes are depicted as crescents, and unique drug selectable markers as depicted as rectangles. (R) gRNA cassettes were introduced to host cells as: (1) circular vectors, wherein each of three unique gRNA cassettes was cloned into a plasmid comprising a unique selectable marker; (2) circular vectors, wherein each of three unique gRNA cassettes was cloned into a plasmid comprising the same selectable marker; (3) linear expression cassettes, wherein the three linear gRNA cassettes were co-transformed with a circular plasmid comprising a selectable marker; and (4) linear expression cassettes, each having ends that are homologous with the ends of a co-transformed linear plasmid comprising a selectable marker, thus allowing for HR-mediated *in vivo* assembly of circular plasmids comprising each gRNA and a common selectable marker.

[0042] **FIG. 8** provides the results of an experiment to determine optimal modes of gRNA delivery as described in Example 1. Cas9-expressing host cells (*S. cerevisiae*) were transformed with donor DNAs for simultaneous, marker-less integration/deletion of RHR2, HO and ADH5 open reading frames and gRNA constructs targeting each locus. Modes of gRNA delivery were 1) three plasmids with three different selectable markers, 2) three plasmids with the same marker, 3) a single marker plasmid, with three linear gRNA cassettes, and 4) a single linearized marker plasmid with flanking sequences for gap repair of three linear

gRNA cassettes. Colonies were assayed by cPCR using an upstream forward primer outside of the deletion construct, and a reverse primer binding to a short linker sequence integrated in place of each open reading frame. 11 colonies were assayed for each delivery mode, as well as a parent colony that serves as a negative control ("N").

[0043] **FIG. 9** provides results of an experiment (described in Example 2) to determine the benefit of gap repair of a marker vector, uncoupled from the benefit of selecting for gRNA expression, towards CRISPR/Cas-9 mediated single integration of a donor DNA into the RHR2, HO and ADH5 locus, respectively. Cas9-expressing host cells (*S. cerevisiae*) were transformed with a donor DNA for marker-less deletion of RHR2, HO or ADH5 open reading frames and gRNA constructs targeting each locus. In addition to the appropriate donor DNA, linear gRNA cassettes were co-transformed with 1) a closed marker vector (A), or 2) the same vector, but linearized and truncated such that gap repair of an additional supplied fragment is required to close the vector and reconstitute the marker cassette (B). Colonies were assayed by PCR using an upstream forward primer outside of the deletion construct, and a reverse primer binding to a short linker sequence integrated in place of each open reading frame. 23 colonies were assayed for each delivery mode, as well as a parent colony that serves as a negative control ("N"). The experiment was repeated 3 times; results from a single experiment are shown.

[0044] **FIG. 10** provides a summation of three gap repair experiments results demonstrated in **FIG. 9**. For each of three experiments, 23 colonies were assayed.

[0045] **FIG. 11** provides a box plot summation of three gap repair experiments demonstrated in **FIG. 9**.

[0046] **FIG. 12:** provides the fold-increase via gap-repair assembly of the marker plasmid; summation of three gap repair experiments demonstrated in **FIG. 9**.

[0047] **FIG. 13** provides the results of an experiment to determine the benefit of 2-piece *in vivo* assembly versus 3-piece *in vivo* assembly of a marker/gRNA vector towards CRISPR/Cas-9 mediated simultaneous integration of three donor DNAs into the Gal80, HO and ADH5 locus, respectively. Cas9-expressing host cells (haploid *S. cerevisiae*) were transformed with donor DNAs for simultaneous, marker-less deletion of Gal80, HO and ADH5 open reading frames, gRNA constructs targeting each locus, and pre-recombination molecules for either 2 or 3 piece marker/gRNA vector assembly. Colonies were assayed by cPCR using an upstream forward primer outside of the deletion construct, and a reverse primer binding to a

short linker sequence integrated in place of each open reading frame. 11 colonies were assayed for each delivery mode, as well as a parent colony that serves as a negative control (“N”).

[0048] FIG. 14 provides the results of an experiment to determine the benefit of 2-piece *in vivo* assembly versus 3-piece *in vivo* assembly of a marker/gRNA vector towards CRISPR/Cas-9 mediated simultaneous integration of three donor DNAs into the Gal80, HO and ADH5 locus, respectively. Cas9-expressing cells of the Diploid yeast strain CAT-1 (*S. cerevisiae*) were transformed with donor DNAs for simultaneous, pan-allelic, marker-less deletion of Gal80, HO and ADH5 open reading frames, gRNA constructs targeting each locus, and pre-recombination molecules for either 2 or 3 piece marker/gRNA vector assembly. Colonies were assayed by PCR using an upstream forward primer outside of the deletion construct, and a reverse primer binding to a short linker sequence integrated in place of each open reading frame. The experiment used a selection scheme in which cells must process transformed DNA reagents using 2 or 3 homologous recombination events to create a selective plasmid. The rate of simultaneous, pan-allelic triple integration was nearly ten-fold higher when 3 events were required. The number of colonies recovered from the experiment was also roughly ten-fold fewer when 3 events were required (not shown), indicating that the selection scheme was responsible for the increased rate of triple integration.

[0049] FIG. 15 provides a schematic for introduction of a point mutation in the context of a “heterology block.” A targeted amino acid is boxed, and an adjacent cleavage site is annotated with cleavage site and PAM sequence (Top panel). A donor DNA containing the desired point mutation in the context of a heterology block of silent codon changes and flanking homology can be generated synthetically by annealing and extending 60-mer oligos (Middle panel) or with larger cloned constructs. Integration of the donor DNA yields the desired point mutation (Lower panel).

[0050] FIG. 16 provides results of an experiment to introduce single point mutations encoded in donor DNA using CRISPR in combination with 2-piece *in vivo* assembly of a marker/gRNA vector. Candidate colonies (1-11) and parent negative control (c) were assayed by colony PCR against the heterology block and flanking sequence (Left panel, and table). Selected positive colonies were confirmed by sequencing a larger PCR product spanning the integration locus.

[0051] FIG. 17 provides the results of an experiment to introduce in multiplex fashion point mutations encoded in donor DNA using CRISPR in combination with 2-piece *in vivo*

assembly of a marker/gRNA vector. The ECM38, PGD1, and ADH2 loci were targeted for the simultaneous introduction of three point mutations. Donor DNAs were cloned, with 500 bp of upstream and downstream homology flanking each target site. Candidate colonies were identified by colony PCR against the heterology block and flanking sequence (Left panel, and table). 10/11 colonies (90.9%) were positive for integration of all three heterology blocks.

[0052] FIG. 18 provides the results of an experiment to introduce in multiplex fashion point mutations encoded in donor DNA using CRISPR in combination with 2-piece *in vivo* assembly of a marker/gRNA vector. The ADH2, PGD1, ECM38, SIN4 and CYS4 loci were targeted for the simultaneous introduction of five point mutations. Donor DNAs were cloned in this case, with 500 bp of upstream and downstream homology flanking each target site. Candidate colonies were identified by colony PCR against the heterology block and flanking sequence (Left panel, and table). 2/11 colonies (18.2%) were positive for integration of all five heterology blocks (clone #'s 4 and 9).

[0053] FIG. 19 provides the results of an experiment demonstrating integration of a short linker sequence at the GAL80 locus in haploid CENPK2 (A), and pan-allelic integration of the same construct of the GAL80 locus in diploid industrial strain CAT-1 (B) and diploid industrial strain PE-2 (C). Each colony was assayed for integration of the short linker sequence (odd numbered lanes) as well as for the presence of the wild type allele (even numbered lanes). The final two lanes on each gel are a parental (negative) control.

[0054] FIG. 20 provides the results of an experiment to introduce in multiplex fashion a 12-gene biosynthetic pathway (totaling ~30kb) for the production of the isoprenoid farnesene, using CRISPR in combination with 2-piece *in vivo* assembly of a marker/gRNA vector. The Gal80, HO and BUD9 loci were targeted for the simultaneous introduction of 3 donor DNAs comprising coding sequences for the farnesene pathway components (donor 1: the transcriptional regulator GAL4; farnesene synthase (2 copies) from *Artemisia annua*; ERG10, encoding acetyl-CoA thiolase; and ERG13, encoding HMG-CoA synthase; donor 2: tHMG1 (2 copies) encoding HMG-CoA reductase; and donor 3: ERG12, encoding mevalonate kinase; ERG8, encoding phosphomevalonate kinase; ERG19, encoding mevalonate pyrophosphate decarboxylase; IDI1, encoding isopentenyl pyrophosphate isomerase; and ERG20, encoding farnesyl pyrophosphate synthetase). Donor DNAs were cloned with 500 bp of upstream and downstream homology flanking each target site. Candidate colonies were identified by colony PCR against an internal linker sequence and sequence flanking the integration target sites. 11/47 colonies (23.4%) were positive for integration of the entire pathway.

[0055] **FIG. 21** provides the results of an experiment to confirm farnesene production in a batch sucrose plate model assay for the 11 clones identified by cPCR as having fully integrated the farnesene pathway. Each cPCR positive clone produced farnesene in amounts ranging from ~0.1 to 1.5 g/L farnesene.

[0056] **FIG. 22** provides results of allele swap cPCRs which demonstrate high rates of single and multiplexed allele swaps.

[0057] **FIGS. 23 (A) – (F)** provide results of experiments that demonstrate that multiplexed allele swaps produced using CRISPR display synergistic phenotypes. **(A)** Truncation of ACE2 results in incomplete cell division and clumping. **(B)** Secretory and cell cycle mutants do not grow at 37°C. **(C)** Cell cycle mutants arrest in G1 at non-permissive temperature. **(D)** SEC3-GFP is localized correctly to the bud at permissive temperature (23°C), but mislocalized at elevated temperature in secretory mutants. **(E)** Two alleles individually increase heat tolerance, and together produce an even more heat tolerant strain. **(F)** Several mutations impart ethanol resistance, but all alleles together synergize for even further increased ethanol tolerance. All five changes were made simultaneously using CRISPR.

[0058] **FIG. 24** provides results demonstrating integration of the entire muconic acid biosynthesis pathway into a naive yeast strain in a single transformation. **(A)** Schematic of the muconic biosynthesis pathway. **(B)** The muconic acid pathway was introduced into three separate loci via six pieces of donor DNA totaling 28kb. Each piece recombined into the genome through a region of homology upstream (US) or downstream (DS) of the targeted locus (ends) as well as with another piece of donor DNA with overlapping homology (center). **(C)** One-step integration of the pathway permitted fast diagnosis of the pathway bottleneck: AroY. Strains with the integrated pathway produce ~3g/L PCA (second line from bottom). When fed catechol (first line from bottom), these strains fully convert all available catechol to muconic acid (third, fourth, and fifth lines from bottom). **(D)** The muconic acid pathway was also introduced into three separate loci in *K. lactis* in a single step (10kb). **(E)** *K. lactis* strains with the integrated pathway produce ~1g/L PCA (first line from bottom), exhibiting the same pathway bottleneck as *S. cerevisiae*. When fed catechol, these strains also fully convert all available catechol to muconic acid (second line from bottom).

[0059] **FIG. 25** provides results of an RFLP assay on amplicons of a targeted genomic locus in 293T cells following transfection with CRISPR reagents and donor DNA. Cells were transfected as follows: (2) Closed “no gRNA” plasmid + linear donor; (3) Open “no gRNA”

plasmid; (4) Open “no gRNA” plasmid + CD4 gap fragment; (6) Closed gRNA plasmid + linear donor; (8) Open gRNA plasmid + full gap + linear donor.

5. DETAILED DESCRIPTION OF THE EMBODIMENTS

5.1 Definitions

[0060] As used herein, the terms “cleaves,” “cleavage” and/or “cleaving” with respect to a nuclease, *e.g.* a homing endonuclease, zinc-finger nuclease, TAL-effector nuclease, or RNA-Guided DNA endonuclease (*e.g.*, CRISPR/Cas9) refer to the act of creating a break in a particular nucleic acid. The break can leave a blunt end or sticky end (*i.e.*, 5' or 3' overhang), as understood by those of skill in the art. The terms also encompass single strand DNA breaks (“nicks”) and double strand DNA breaks.

[0061] As used herein, the term “engineered host cell” refers to a host cell that is generated by genetically modifying a parent cell using genetic engineering techniques (*i.e.*, recombinant technology). The engineered host cell may comprise additions, deletions, and/or modifications of nucleotide sequences to the genome of the parent cell.

[0062] As used herein, the term “heterologous” refers to what is not normally found in nature. The term “heterologous nucleotide sequence” refers to a nucleotide sequence not normally found in a given cell in nature. As such, a heterologous nucleotide sequence may be: (a) foreign to its host cell (*i.e.*, is “exogenous” to the cell); (b) naturally found in the host cell (*i.e.*, “endogenous”) but present at an unnatural quantity in the cell (*i.e.*, greater or lesser quantity than naturally found in the host cell); or (c) be naturally found in the host cell but positioned outside of its natural locus.

[0063] As used herein, the term “homology” refers to the identity between two or more nucleic acid sequences, or two or more amino acid sequences. Sequence identity can be measured in terms of percentage identity (or similarity or homology); the higher the percentage, the more near to identical the sequences are to each other. Homologs or orthologs of nucleic acid or amino acid sequences possess a relatively high degree of sequence identity when aligned using standard methods. Methods of alignment of sequences for comparison are well known in the art. Various programs and alignment algorithms are described in: Smith & Waterman, *Adv. Appl. Math.* 2:482, 1981; Needleman & Wunsch, *J. Mol. Biol.* 48:443, 1970; Pearson & Lipman, *Proc. Natl. Acad. Sci. USA* 85:2444, 1988; Higgins & Sharp, *Gene*, 73:237-44, 1988; Higgins & Sharp, *CABIOS* 5:151-3, 1989; Corpet *et al.*, *Nuc. Acids Res.* 16:10881-90, 1988; Huang *et al. Computer Appls. Biosc.* 8, 155-65, 1992; and Pearson *et al.*,

Meth. Mol. Bio. 24:307-31, 1994. Altschul *et al.*, *J. Mol. Biol.* 215:403-10, 1990, presents a detailed consideration of sequence alignment methods and homology calculations. The NCBI Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, *J. Mol. Biol.* 215:403-10, 1990) is available from several sources, including the National Center for Biological Information (NCBI, National Library of Medicine, Building 38A, Room 8N805, Bethesda, Md. 20894) and on the Internet, for use in connection with the sequence analysis programs blastp, blastn, blastx, tblastn and tblastx. Additional information can be found at the NCBI web site.

[0064] As used herein, the term “markerless” refers to integration of a donor DNA into a target site within a host cell genome without accompanying integration of a selectable marker. In some embodiments, the term also refers to the recovery of such a host cell without utilizing a selection scheme that relies on integration of selectable marker into the host cell genome. For example, in certain embodiments, a selection marker that is episomal or extrachromosomal may be utilized to select for cells comprising a plasmid encoding a nuclease capable of cleaving a genomic target site. Such use would be considered “markerless” so long as the selectable marker is not integrated into the host cell genome.

[0065] As used herein, the term “operably linked” refers to a functional linkage between nucleic acid sequences such that the sequences encode a desired function. For example, a coding sequence for a gene of interest, *e.g.*, a selectable marker, is in operable linkage with its promoter and/or regulatory sequences when the linked promoter and/or regulatory region functionally controls expression of the coding sequence. It also refers to the linkage between coding sequences such that they may be controlled by the same linked promoter and/or regulatory region; such linkage between coding sequences may also be referred to as being linked in frame or in the same coding frame. “Operably linked” also refers to a linkage of functional but non-coding sequences, such as an autonomous propagation sequence or origin of replication. Such sequences are in operable linkage when they are able to perform their normal function, *e.g.*, enabling the replication, propagation, and/or segregation of a vector bearing the sequence in host cell.

[0066] As used herein, the term “selecting a host cell expressing a selectable marker” also encompasses enriching for host cells expressing a selectable marker from a population of transformed cells.

[0067] As used herein, the term “selectable marker” refers to a gene which functions as guidance for selecting a host cell comprising a marker vector as described herein. The

selectable markers may include, but are not limited to: fluorescent markers, luminescent markers and drug selectable markers, and the like. The fluorescent markers may include, but are not limited to, genes encoding fluorescence proteins such as green fluorescent protein (GFP), cyan fluorescent protein (CFP), yellow fluorescent protein (YFP), red fluorescent protein (dsRFP) and the like. The luminescent markers may include, but are not limited to, genes encoding luminescent proteins such as luciferases. Drug selectable markers suitable for use with the methods and compositions provided herein include, but are not limited to, resistance genes to antibiotics, such as ampicillin, streptomycin, gentamicin, kanamycin, hygromycin, tetracycline, chloramphenicol, and neomycin. In some embodiments, the selection may be positive selection; that is, the cells expressing the marker are isolated from a population, e.g. to create an enriched population of cells comprising the selectable marker. In other instances, the selection may be negative selection; that is, the population is isolated away from the cells, e.g. to create an enriched population of cells that do not comprise the selectable marker. Separation may be by any convenient separation technique appropriate for the selectable marker used. For example, if a fluorescent marker has been utilized, cells may be separated by fluorescence activated cell sorting, whereas if a cell surface marker has been inserted, cells may be separated from the heterogeneous population by affinity separation techniques, e.g. magnetic separation, affinity chromatography, “panning” with an affinity reagent attached to a solid matrix, or other convenient technique.

[0068] As used herein, the term “simultaneous,” when used with respect to multiple integration, encompasses a period of time beginning at the point at which a host cell is co-transformed with a nuclease, e.g. a plasmid encoding a nuclease, and more than one donor DNA to be integrated into the host cell genome, and ending at the point at which the transformed host cell, or clonal populations thereof, is screened for successful integration of the donor DNAs at their respective target loci. In some embodiments, the period of time encompassed by “simultaneous” is at least the amount of time required for the nuclease to bind and cleave its target sequence within the host cell’s chromosome(s). In some embodiments, the period of time encompassed by “simultaneous” is at least 6, 12, 24, 36, 48, 60, 72, 96 or more than 96 hours, beginning at the point at which the a host cell is co-transformed with a nuclease, e.g. a plasmid encoding a nuclease, and more than one donor DNA.

5.2 Methods of Integrating Exogenous Nucleic Acids

[0069] Provided herein are methods of integrating one or more exogenous nucleic acids into one or more selected target sites of a host cell genome. In certain embodiments, the

methods comprise contacting the host cell with one or more integration polynucleotides, *i.e.*, donor DNAs, comprising an exogenous nucleic acid to be integrated into the genomic target site; one or more nucleases capable of causing a double-strand break near or within the genomic target site; and a linear nucleic acid capable of homologous recombination with itself or with one or more additional linear nucleic acids contacted with the host cell, whereupon said homologous recombination of the linear nucleic acid in the host cell results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker. In some embodiments, the contacted host cell is then grown under selective conditions. Without being bound by theory of operation, it is believed that forcing the host cell to circularize the expression vector via HR, in order to be selected in accordance with the methods described herein, increases the likelihood that the selected cell has also successfully performed the one or more intended HR-mediated genomic integrations of exogenous DNA.

[0070] In a particular aspect, provided herein is a method for markerless integration of an exogenous nucleic acid into a target site of a host cell genome, the method comprising:

- (a) contacting a host cell with:
 - (i) an exogenous nucleic acid (ES) comprising a first homology region (HR1) and a second homology region (HR2), wherein (HR1) and (HR2) are capable of initiating host cell mediated homologous recombination at said target site (TS);
 - (ii) a nuclease (N) capable of cleaving at (TS), whereupon said cleaving results in homologous recombination of (ES) at (TS); and
 - (iii) a linear nucleic acid capable of homologous recombination with itself or with one or more additional linear nucleic acids contacted with the host cell, whereupon said homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker;
- and
- (b) selecting a host cell that expresses the selectable marker.

In some embodiments, the method comprises recovering a host cell having (ES) integrated at (TS), wherein said recovering does not require integration of a selectable marker.

[0071] **FIG. 1** provides an exemplary embodiment of genomic integration of an exogenous nucleic acid using a site-specific nuclease and a pre-recombination composition capable of assembling *in vivo* via host cell mediated HR to form a circular marker expression vector. A

donor polynucleotide is introduced to a host cell, wherein the polynucleotide comprises a nucleic acid of interest (D) flanked by a first homology region (HR1) and a second homology region (HR2). HR1 and HR2 share homology with 5' and 3' regions, respectively, of a genomic target site (TS). A site-specific nuclease (N) is also introduced to the host cell, wherein the nuclease is capable of recognizing and cleaving a unique sequence within the target site. Also introduced to the cell is a pre-recombination composition, which in this example comprises two linear pre-recombination molecules each comprising two homology regions capable of homologously recombining with each other. In this example, the homology regions are positioned at the 5' and 3' termini of each pre-recombination molecule. One homology region of each pre-recombination molecule comprises a partial coding sequence for a selectable marker (GF and FP, respectively), such that upon HR between the two homology regions, a complete and operable coding sequence of the selectable marker (GFP) is reconstituted on a circularized marker expression vector. In general, such a circularization is selected for by culturing the cells under conditions that select for expression of the selectable marker, for example, by supplementing the culture medium with a selective agent (*e.g.*, an antibiotic) where the selectable marker is a drug resistance marker, or sorting for cells which express a marker detectable by colorimetric or fluorescent detection methods. Concomitantly, in cells that are competent for HR, induction of a double-stranded break within the target site by the site-specific nuclease facilitates the HR-mediated integration of the donor nucleic acid of interest at the cleaved target site. By making it a requirement that the host cell circularize the expression vector via HR in order to be selected, the recovery of cells that have also performed HR-mediated integration of the exogenous donor DNA is also increased. This increased frequency of recovery obviates the need to co-integrate a selectable marker in order to select transformants having undergone a recombination event. By eliminating the need for selectable markers, for example, during construction of an engineered microbe, the time needed to engineer a host cell genome is greatly reduced. In addition, engineering strategies are no longer limited by the need to recycle selectable markers due to there being a limited cache of markers available for a given host organism.

[0072] In some embodiments, markerless recovery of a transformed cell comprising a successfully integrated exogenous nucleic acid occurs within a frequency of about one every 1000, 900, 800, 700, 600, 500, 400, 300, 200 or 100 contacted host cells, or clonal populations thereof, screened. In particular embodiments, markerless recovery of a transformed cell comprising a successfully integrated exogenous nucleic acid occurs within a frequency of

about one every 90, 80, 70, 60, 50, 40, 30, 20, or 10 contacted host cells, or clonal populations thereof, screened. In more particular embodiments, markerless recovery of a transformed cell comprising a successfully integrated exogenous nucleic acid occurs within a frequency of about one every 9, 8, 7, 6, 5, 4, 3, or 2 contacted host cells, or clonal populations thereof, screened.

[0073] A variety of methods are available to identify those cells having an altered genome at or near the target site without the use of a selectable marker. In some embodiments, such methods seek to detect any change in the target site, and include but are not limited to PCR methods, sequencing methods, nuclease digestion, *e.g.*, restriction mapping, Southern blots, and any combination thereof. Phenotypic readouts, for example, a predicted gain or loss of function, can also be used as a proxy for effecting the intended genomic modification(s).

[0074] In another aspect, provided herein is a method for integrating a plurality of exogenous nucleic acids into a host cell genome, the method comprising:

(a) contacting a host cell with:

(i) a plurality of exogenous nucleic acids, wherein each exogenous nucleic acid (ES)_x comprises a first homology region (HR1)_x and a second homology region (HR2)_x, wherein (HR1)_x and (HR2)_x are capable of initiating host cell mediated homologous recombination of (ES)_x at a target site (TS)_x of said host cell genome;

(ii) for each said target site (TS)_x, a nuclease (N)_x capable of cleaving at (TS)_x, whereupon said cleaving results in homologous recombination of (ES)_x at (TS)_x; and

(iii) a linear nucleic acid capable of homologous recombination with itself or with one or more additional linear nucleic acids contacted with the host cell, whereupon said homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker;

and

(b) selecting a host cell that expresses the selectable marker.

In some embodiments, the method further comprises recovering a host cell wherein each selected exogenous nucleic acid (ES)_x has integrated at each selected target sequence (TS)_x,

wherein x is any integer from 1 to n wherein n is at least 2.

[0075] FIG. 2 provides an exemplary embodiment of simultaneous genomic integration of a plurality of exogenous nucleic acids using a plurality of site-specific nucleases. In this example, three different donor polynucleotides are introduced to a host cell, wherein each polynucleotide comprises an exogenous nucleic acid (ES)_x comprising a nucleic acid of interest (D)_x, wherein x=1, 2 or 3. Each (D)_x is flanked by a first homology region (HR1)_x and a second homology region (HR2)_x. (HR1)_x and (HR2)_x share homology with 5' and 3' regions, respectively, of a selected target site (TS)_x, of three total unique target sites in the genome. One or more site-specific nucleases (N)_x (for example, one or more (e.g. "x" number of) endonucleases having a unique recognition site; or an RNA-guided endonuclease together with one or more (e.g. "x" number of) guide RNAs) are also introduced to the host cell, wherein each nuclease (N)_x is capable of recognizing and cleaving a unique sequence within its corresponding target site, (TS)_x. Also introduced to the cell is a pre-recombination composition, which in this example comprises two linear pre-recombination molecules each comprising two homology regions capable of homologously recombining with each other. In this example, the homology regions are positioned at the 5' and 3' termini of each pre-recombination molecule. One homology region of each pre-recombination molecule comprises a partial coding sequence for a selectable marker (GF and FP, respectively), such that upon HR between the two homology regions, a complete and operable coding sequence of the selectable marker (GFP) is reconstituted on a circularized marker expression vector. Such a circularization is selected for by culturing the cells under conditions that select for expression of the selectable marker. Concomitantly, in cells that are competent for HR, cleavage of a target site (TS)_x by its corresponding site-specific nuclease (N)_x facilitates integration of the corresponding nucleic acid interest (D)_x at (TS)_x by the host cell's endogenous homologous recombination machinery. By making it a requirement that the host cell circularize the expression vector via HR in order to be selected, the recovery of cells that have also performed HR-mediated integration of the exogenous donor DNAs is also increased.

[0076] In particular embodiments, each exogenous nucleic acid (ES)_x, optionally comprising a nucleic acid of interest (D)_x, is integrated into its respective genomic target site (TS)_x simultaneously, *i.e.*, with a single transformation of the host cell with the plurality of integration polynucleotides and plurality of nucleases. In some embodiments, the methods are useful to simultaneously integrate any plurality of exogenous nucleic acids (ES)_x, that is, where x is any integer from 1 to n wherein n is at least 2, in accordance with the variables recited for the above described method. In some embodiments, the method of simultaneous integration

provided herein is useful to simultaneously integrate up to 10 exogenous nucleic acids (ES)_x into 10 selected target sites (TS)_x, that is, where x is any integer from 1 to n wherein n= 2, 3, 4, 5, 6, 7, 8, 9, or 10. In some embodiments, the method of simultaneous integration provided herein is useful to simultaneously integrate up to 20 exogenous nucleic acids (ES)_x into 20 selected target sites (TS)_x, that is, where x is any integer from 1 to n wherein n= 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20. In some embodiments, n=2. In some embodiments, n= 3. In some embodiments, n= 4. In some embodiments, n= 5. In some embodiments, n= 6. In some embodiments, n= 7. In some embodiments, n= 8. In some embodiments, n= 9. In some embodiments, n= 10. In some embodiments, n= 11. In some embodiments, n= 12. In some embodiments, n= 13. In some embodiments, n= 14. In some embodiments, n= 15. In some embodiments, n= 16. In some embodiments, n= 17. In some embodiments, n= 18. In some embodiments, n= 19. In some embodiments, n= 20. In some embodiments, the method of simultaneous integration provided herein is useful to simultaneously integrate more than 20 exogenous nucleic acids.

[0077] As with integration of a single exogenous nucleic acid at a single target site, the recovery of a host cell that has successfully integrated each exogenous nucleic acid at its respective target site occurs at a substantially higher frequency as compared to not contacting the host cell with one or more linear pre-recombination molecules described herein, and selecting for expression of the selectable marker. In some embodiments, this increased frequency of integration obviates the requirement for co-integration of one or more selectable markers for the identification of the plurality of recombination events. In some embodiments, markerless recovery of a transformed cell comprising a plurality of successfully integrated exogenous nucleic acid occurs within a frequency of about one every 1000, 900, 800, 700, 600, 500, 400, 300, 200 or 100 contacted host cells, or clonal populations thereof, screened. In particular embodiments, markerless recovery occurs within a frequency of about one every 90, 80, 70, 60, 50, 40, 30, 20, or 10 contacted host cells, or clonal populations thereof, screened. In more particular embodiments, markerless recovery occurs within a frequency of about one every 9, 8, 7, 6, 5, 4, 3, or 2 contacted host cells, or clonal populations thereof, screened.

5.2.1 HR-mediated *in vivo* assembly of circular marker expression vectors

[0078] The methods provided herein comprise host cell mediated assembly of a circular expression vector via gap repair. Gap repair is a fast and efficient method for assembling recombinant DNA molecules *in vivo*. The technique has been described to be effective for assembling and/or repairing plasmids in a number of organisms, including bacteria

(*Escherichia coli*; see, e.g., Datta *et al.*, *Gene* 379:109-115 (2006)), yeast (*Saccharomyces cerevisiae*; see e.g., Bessa *et al.*, *Yeast* 29:419-423 (2012)), insects (*Drosophila melanogaster*; see, e.g., Carreira-Rosario *et al.*, *J Vis Exp* 77:e50346 (2013)) and mammalian cells (human cells; see e.g., Adar *et al.*, *Nucleic Acids Research* 37(17):5737-5748 (2009)). Gap repair can produce a circular DNA molecule by homologous recombination between two homologous regions of a single linear DNA, or between two or more separate linear DNA fragments. Typically, the assembled circularized DNA acts as a vector carrying replicative sequences and a selective marker. See, e.g., Orr-Weaver *et al.*, *Methods Enzymol* 101:228-245 (1983). The technique, outlined in **FIG. 4**, typically starts with co-transformation of a linear “gapped” vector and a linear DNA fragment (insert) (Orr-Weaver *et al.*, 1983). In cells competent for homologous recombination, recombination occurs between two pairs of flanking stretches of homologous sequences between vector and insert, resulting in a larger circular vector wherein the gap has been repaired. A simple way to provide flanking homology of the insert is by polymerase chain reaction (PCR) where tailed primers provide the homology regions.

[0079] In one aspect of the methods and compositions provided herein, the host cell is contacted with a single contiguous linear (gapped) nucleic acid that serves as a pre-recombination vector intermediate. As used herein, the phrase “single nucleic acid” includes the embodiment of multiple copies of the same nucleic acid molecule. In some embodiments, the pre-recombination vector is self-circularizing, and comprises two sequence-specific recombination regions capable of homologous recombination with each other, such that introduction into a recombination-competent host cell results in formation of a circular expression vector. In some embodiments, the recombination regions are positioned at or near the termini of the linear pre-recombination vector (e.g., one recombination region is positioned at each termini of the linear vector, with additional sequences intervening the two regions), internal to the termini (e.g., each recombination region is flanked on both ends by additional sequences), or a combination thereof (e.g., one recombination is at one termini of the linear vector and the other is internal thereto and flanked on both sides by additional sequences). In some embodiments, the first and second recombination regions can comprise any nucleotide sequence of sufficient length and share any sequence identity that allows for homologous recombination with each other. In some embodiments, “sufficient sequence identity” refers to sequences with at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95%, at least 99%, or 100%, identity between recombination regions, over a length of, for example, at least 15 base pairs, at least 20 base pairs, at least 50 base pairs, at least 100 base pairs, at

least 250 base pairs, at least 500 base pairs, or more than 500 base pairs. The extent of sequence identity may be determined using any computer program and associated parameters, including those described herein, such as BLAST 2.2.2 or FASTA version 3.0t78, with the default parameters. For a discussion of effective lengths of homology between recombination regions, see Hasty *et al.*, *Mol Cell Biol* 11:5586-91 (1991).

[0080] In some embodiments, the first and second recombination regions share at least 25% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 30% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 35% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 40% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 45% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 50% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 60% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 65% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 70% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 75% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 80% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 85% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 90% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 95% nucleotide sequence identity. In some embodiments, the first and second recombination regions share at least 99% nucleotide sequence identity. In some embodiments, the first and second recombination regions share 100% nucleotide sequence identity.

[0081] In certain embodiments, each of the first and second recombination regions consists of about 50 to 5,000 nucleotides. In certain embodiments, each of the first and second recombination regions comprises about 50 to 5,000 nucleotides. In certain embodiments, each of the first and second recombination regions consists of about 100 to 2,500 nucleotides. In certain embodiments, each of the first and second recombination regions consists of about 100 to 1,000 nucleotides. In certain embodiments, each of first and second recombination regions consists of about 250 to 750 nucleotides. In certain embodiments, each of the first and second

recombination regions consists of about 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3100, 3200, 3300, 3400, 3500, 3600, 3700, 3800, 3900, 4000, 4100, 4200, 4300, 4400, 4500, 4600, 4700, 4800, 4900 or 5,000 nucleotides. In some embodiments, each of the first and second recombination regions consists of about 500 nucleotides. In some embodiments, each of first and second recombination regions comprises at least 18 nucleotide base pairs. In some embodiments, each of first and second recombination regions consists of 15 to 500 nucleotide base pairs. In some embodiments, each of first and second recombination regions consists of 15 to 500, 15 to 495, 15 to 490, 15 to 485, 15 to 480, 15 to 475, 15 to 470, 15 to 465, 15 to 460, 15 to 455, 15 to 450, 15 to 445, 15 to 440, 15 to 435, 15 to 430, 15 to 425, 15 to 420, 15 to 415, 15 to 410, 15 to 405, 15 to 400, 15 to 395, 15 to 390, 15 to 385, 15 to 380, 15 to 375, 15 to 370, 15 to 365, 15 to 360, 15 to 355, 15 to 350, 15 to 345, 15 to 340, 15 to 335, 15 to 330, 15 to 325, 15 to 320, 15 to 315, 15 to 310, 15 to 305, 15 to 300, 15 to 295, 15 to 290, 15 to 285, 15 to 280, 15 to 275, 15 to 270, 15 to 265, 15 to 260, 15 to 255, 15 to 250, 15 to 245, 15 to 240, 15 to 235, 15 to 230, 15 to 225, 15 to 220, 15 to 215, 15 to 210, 15 to 205, 15 to 200, 15 to 195, 15 to 190, 15 to 185, 15 to 180, 15 to 175, 15 to 170, 15 to 165, 15 to 160, 15 to 155, 15 to 150, 15 to 145, 15 to 140, 15 to 135, 15 to 130, 15 to 125, 15 to 120, 15 to 115, 15 to 110, 15 to 105, 15 to 100, 15 to 95, 15 to 90, 15 to 85, 15 to 80, 15 to 75, 15 to 70, 15 to 65, 15 to 60, 15 to 55, 15 to 50, 15 to 45, 15 to 40, 15 to 35, 15 to 30, 15 to 25, or 15 to 20 nucleotide base pairs. In some embodiments, each of first and second recombination regions consists of 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, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199 or 200 nucleotide base pairs.

[0082] In preferable embodiments of the methods and compositions provided herein, each homology region of a homology region pair comprises nucleotide sequences of sufficient length and sequence identity that allows for homologous recombination with each other, but

not with other regions of the pre-recombination molecule(s) participating in the assembly, nor with any genomic regions of the host cell.

[0083] While in some embodiments, the circularized expression vector is formed from a single pre-recombination molecule that is self-circularizing, in other embodiments, the circularized expression vector is formed by the HR-mediated assembly of two or more linear pre-recombination molecules. For example, a circularized vector may be assembled from two linear pre-recombination molecules, wherein the first molecule is a gapped vector and the second molecule is an insert comprising two separate homologous regions capable of recombining with two homology regions on the gapped vector. For each of the gapped linear vector and the linear insert, the recombination regions can be positioned at or near the termini of the linear pre-recombination vector (*e.g.*, one recombination region is positioned at each termini, with additional sequences intervening the two regions), internal to the termini (*e.g.*, each recombination region is flanked on both ends by additional sequences), or a combination thereof (*e.g.*, one recombination is at one termini and the other is internal thereto and flanked on both sides by additional sequences). In still other embodiments, the insert which repairs the gapped vector can itself be assembled from at least two linear nucleic acids comprising homologous regions to each other. For example, the circularized vector may be formed from three distinct linear pre-recombination fragments, wherein the first linear molecule comprises homology regions A₁ and B₁, the second linear molecule comprises B₂ and C₂, and the third linear molecule comprises C₃ and A₃, such that recombination between homologous regions of each fragment (*i.e.*, A₁ with A₃, B₁ with B₂, and C₂ with C₃) in an HR-competent host cell results in formation of a circularized expression vector comprising regions A → B → C.

[0084] In still other embodiments, the circularized vector is assembled in a HR-competent host cell from at least 4, 5, 6, 7, 8, 9 or 10 distinct linear pre-recombination fragments in a similar fashion. Without being bound by theory of operation, it is believed that requiring the circularized expression vector to be assembled from more than two linear pre-recombination molecules selects for host cells that are particularly adept at homologous recombination. Thus, assembly of the circular expression vector from multiple pre-recombination molecules may be preferred when higher order integration events are desired, *e.g.*, multiplex genomic integration (for example, of 2 or more donor exogenous DNAs), or when performing genomic integration into a cell type known or suspected to have very low rates of HR. In one example, for a multiplex (*i.e.*, simultaneous) integration of three exogenous donor nucleic acids into three respective genomic target sites of a host cell, the host cell is “forced” to assemble at least three

linear pre-recombination fragments to form the circular expression vector. Only cells that can successfully recombine the three fragments to form the circular vector that expresses a selectable marker can survive the selection, *i.e.* be selected for, and these cells will be more likely to have successfully integrated each of the three exogenous donor nucleic acids into their respective genomic target sites. In some embodiments, when multiplex integration of at least 2, 3, 4, 5, 6, 7, 8, 9, 10 or greater than 10 exogenous donor nucleic acids into at least 2, 3, 4, 5, 6, 7, 8, 9, 10 or greater than 10 respective genomic target sites is desired, the host cell is forced to assemble, *i.e.*, recombine at least 2, 3, 4, 5, 6, 7, 8, 9, 10 or greater than 10 pre-recombination fragments to form the circular expression vector, from which a selectable marker is expressed in the host cell. In some embodiments, the *in vivo* assembly of at least 2, 3, 4, 5, 6, 7, 8, 9, 10 or greater than 10 pre-recombination fragments forms a circular expression vector comprising coding sequences for more than one selectable marker (*e.g.* two, three, or more than three different selectable markers).

[0085] In preferred embodiments, the circularized expression vector, once assembled, comprises a coding sequence for a selectable marker, and suitable regulatory sequences such as a promoter and/or a terminator that enables expression of the marker in the host cell. Selectable markers include those which both enable selection of host cells containing a selectable marker, while other embodiments enable selecting cells which do not contain the selectable marker, which find use in embodiments wherein elimination of the circularized expression vector after selection is desired (as discussed below). In some embodiments, prior to assembly of the circular expression vector, the linear pre-recombination molecule, or at least one of the pre-recombination fragments of a multi-fragment assembly, comprises an intact coding sequence for the selectable marker, in operable linkage to its regulatory sequences, separate and apart from the homology regions involved in the HR-mediated assembly. Thus, assembly of the circular expression vector does not alter the coding sequence of the selectable marker nor any of its regulatory sequences needed for expression. In some such embodiments, the circularization event merely enables the propagation of, and sustained expression from, the coding sequence of the marker, whereas non-circularized linear vector cannot be propagated and/or maintained in the host cell. In preferred embodiments, host cells which do not comprise a circularized expression vector do not survive the selection step and/or are not selected for, in the methods described herein. Without being bound by theory of operation, it is believed that by requiring the host cell to circularize the expression vector via HR, in order to be selected in accordance with the methods described herein, increases the likelihood that the selected cell

has also successfully performed the one or more intended HR-mediated genomic integrations of exogenous DNA.

[0086] In other embodiments, the sequence encoding the selectable marker, and/or its regulatory sequences required for expression, is not intact, *i.e.*, is not in operable linkage, on any single pre-recombination molecule. In some such embodiments, only when the expression vector is circularized is the marker coding sequence, along with its regulatory elements, brought into operable linkage. Thus, the sequence encoding the marker, and or its necessary regulatory sequences may be divided into any number of overlapping homologous sequences distributed among any number fragments participating in the assembly, so long as HR between the component pre-recombination fragments results in reconstitution of the coding sequence of the selectable marker in operable linkage with its regulatory sequences. These embodiments are particularly useful to avoid selecting host cells in which formation of a circularized expression vector results from joining of the pre-recombination fragments via non-HR mechanisms, for example, non-homologous end-joining or single strand annealing; such cells surviving the selection would represent false positives. The frequency of these unwanted events can be lowered by removing the 5'-phosphate groups on the pre-recombination fragment(s) using phosphatase, which is the standard method used for *in vitro* ligation. Vector religation may also be avoided by treatment of the pre-recombination fragment(s) with Taq DNA polymerase and dATP; this has been reported to be particularly effective at preventing vector re-circularization *in vivo*, facilitating the screening for true recombinant clones. See Bessa *et al.*, *Yeast* 29:419-423 (2012). In addition, false positives caused by erroneous introduction of pre-circularized DNA may be avoided by prepping the pre-recombination fragment(s) by PCR rather than linearizing a circular vector by endonuclease digestion then isolating the fragment, which may carry over non-digested circular template. Nevertheless, without being bound by theory of operation, it is believed that requiring the host cell to reconstitute the marker coding sequence via HR, *e.g.*, from at least two partial sequences, in order to survive the selection process, increases the likelihood that a cell selected in accordance with the methods described herein has also successfully performed the one or more intended HR-mediated genomic integrations of exogenous DNA.

[0087] In some embodiments of the methods provided herein, the circularized expression vector, once assembled, further comprises a coding sequence for a site-specific nuclease described herein, and suitable regulatory sequences such as a promoter and/or a terminator that enables expression of the nuclease in the host cell. In some embodiments, the nuclease is

selected from the group consisting of CRISPR/Cas-associated RNA-guided endonuclease, a meganuclease, a zinc finger nuclease, a TAL-effector DNA binding domain-nuclease fusion protein (TALEN), a transposase, and a site-specific recombinase. In some embodiments, the nuclease is a CRISPR-associated RNA-guided endonuclease. In some such embodiments, the circular expression vector further comprises a sequence or sequences that encode obligate guide sequences for a RNA-guided endonuclease, for example, a crRNA activity and a tracrRNA activity, which enables site-specific recognition and cleavage of the genomic target DNA by the RNA-guided DNA endonuclease. In some embodiments, the crRNA activity and the tracrRNA activity are expressed from the circular expression vector as a single contiguous RNA molecule, *i.e.*, a chimeric guide RNA (gRNA) molecule. In some embodiments, the circular expression vector comprises one or more sequences encoding a guide sequence(s) (*e.g.*, a gRNA) for an RNA-guided endonuclease, without also comprising the coding sequence for the nuclease. In some such embodiments, one or more sequences encoding the RNA-guided nuclease may be supplied on a separate vector, integrated into the genome, or the nuclease may be introduced to the cell as a protein, *e.g.*, expressed and purified *in vitro*.

[0088] In any of the aforementioned embodiments in which nuclease coding sequences, and/or additional sequences required for expression and operability of the nuclease in the cell is included in the circularized expression vector, these sequences may be intact (*i.e.*, in operable linkage) in any of the one or more pre-recombination linear fragment(s) participating in the assembly reaction, or alternatively, divided into any number of overlapping homologous sequences distributed among any number fragments participating in the assembly, so long as HR between the component pre-recombination fragments results in reconstitution of the coding sequence of the nuclease in operable linkage with its regulatory sequences. Advantageously, coupling the coding sequence of the nuclease (and/or sequences encoding guide RNA sequences where the nuclease is an RNA-guided nuclease) to the circularized expression vector ensures that expression of these sequences is maintained at a level and duration sufficient to assist in the HR-mediated integration event. In accordance with the methods described herein, the efficiency of gene targeting can be improved when combined with a targeted genomic double-stranded break (DSB) introduced near the intended site of integration. *See e.g.*, Jasin, M., *Trends Genet* 12(6):224-228 (1996); and Urnov *et al.*, *Nature* 435(7042):646-651 (2005). Moreover, coupling the coding sequence of the nuclease and/or associated guide sequence(s) to the circularized expression vector eliminates the need for introducing multiple vectors to the host cell in order to effect expression of these sequences. Additionally, such coupling allows

for simultaneous elimination of the nuclease and the marked plasmid following selection of a host cell having performed the desired integrations. Thus, needlessly prolonged expression of the nuclease is avoided, and consequently, any toxicity associated therewith (*see e.g.*, Cho *et al.*, *Genome Res*, “Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases,” (2013); Sander *et al.*, “*In silico* abstraction of zinc finger nuclease cleavage profiles reveals an expanded landscape of off-target sites,” *Nucleic Acids Research* 41(19):e181 (2013)).

[0089] In some embodiments of the methods provided herein, the circularized expression vector, once assembled, further comprises one or more exogenous donor nucleic acids, described in Section 5.2.2 below. In some such embodiments, the exogenous donor nucleic acids may be released from the circularized expression vector by flanking the exogenous donor nucleic acids with recognition sequences for a nuclease also introduced into the host cell, for example a nuclease also encoded by the circularized expression vector.

[0090] As will be clear to those in the art, the circularized expression vector will preferably contain an autonomous propagation sequence which enables the expression vector to be replicated, propagated, and segregated during multiple rounds of host cell division. The autonomous propagation sequence can be either prokaryotic or eukaryotic, and includes an origin of replication. Replication origins are unique polynucleotides that comprise multiple short repeated sequences that are recognized by multimeric origin-binding proteins and that play a key role in assembling DNA replication enzymes at the origin site. Suitable origins of replication for use in the entry and assembly vectors provided herein include but are not limited to *E. coli* oriC, *colE1* plasmid origin, 2 μ and ARS (both useful in yeast systems), *sfl*, SV40 EBV oriP (useful in mammalian systems), or those found in pSC101. Particular embodiments of an expression vector include both prokaryotic and eukaryotic autonomous propagation sequences. This sequence may be intact (*i.e.*, in operable linkage) in any of the one or more pre-recombination linear fragment(s) participating in the assembly reaction, or alternatively, divided into any number of overlapping homologous sequences distributed among any number fragments participating in the assembly, so long as HR between the component pre-recombination fragments results in reconstitution of the autonomous propagation sequence. In particular embodiments, the autonomous propagation sequence is not intact, *i.e.*, is not in operable linkage, on any single pre-recombination molecule. In some such embodiments, only when the expression vector is circularized by recombination of the pre-recombination molecule(s) is the autonomous propagation sequence brought into operable

linkage. These embodiments are particularly useful to avoid selecting host cells in which formation of a circularized expression vector results from non-HR joining of the pre-recombination fragments comprising an intact autonomous propagation sequence, for example, by non-homologous end-joining or single strand annealing; such cells surviving the selection would represent false positives.

[0091] As will also be clear to those of skill in the art, the *in vivo* recombination between one or more linear pre-recombination DNA fragments described above, which results in formation of a circular expression vector in the host cell, can also be achieved with *circular* pre-recombination nucleic acids comprising the appropriate (*e.g.*, same) homology regions as a starting point. For any of the pre-recombination compositions described herein, the compositions can be transformed directly into a host cell as linear nucleic acid molecules, or alternatively, parental circular molecules comprising the pre-recombination molecules can be introduced into the host cell and cleaved *in vivo* by one or more nucleases to liberate the pre-recombination molecules. In some embodiments, the one or more linear pre-recombination molecules participating in the marker vector assembly can be liberated from a parental circular plasmid via *in vivo* cleavage in the host cell by the one or more nucleases targeting the one or more genomic target sites for cleavage.

[0092] In another aspect, provided herein are methods of making pre-recombination expression vector intermediates useful in the practice of the integration methods provided herein. In some embodiments, a base vector comprising an autonomous propagation sequence, a first primer binding sequence, and a second primer binding sequence is amplified using at least a first primer and a second primer. The first primer typically comprises of 5' portion having a first sequence-specific recombination sequence and a 3' portion having a priming portion substantially complementary (*i.e.*, having sufficient complementarity to enable amplification of the desired nucleic acids but not other, undesired molecules) to the first primer binding sequence of the base vector. Similarly, the second primer comprises a 5' portion having a second sequence-specific recombination sequence and a 3' portion having a priming portion substantially complementary to the second primer binding sequence of the base vector. Amplification of the base vector (which can be either linear or circular prior to initiation of the amplification process) results in the production of a linear expression vector intermediate having a first terminus comprising a first sequence-specific recombination region and a second terminus comprising a second sequence-specific recombination region. In certain embodiments, the base vector is a plasmid, particularly a plasmid such as are known in the art and which are based on various bacterial- or yeast-derived extra-chromosomal elements. In certain other embodiments, the base vector further comprises one or more selectable markers, transcription initiation sequences, and/or transcription termination

sequences. As those in the art will appreciate, elements intended to regulate expression of genes carried in the target nucleic acid should be positioned in the expression vector so as to be functionally or operably associated with the gene(s) to be expressed, once the circular expression vector is assembled *in vivo*. The particular positioning of such elements depends upon those elements employed, the host cell, the gene(s) to be expressed, and other factors, including the number of desired integrations in the host cell, as described above. As a result, the final design of a particular expression vector made in accordance with the instant teachings is a matter of choice and depends upon the specific application.

[0093] Yet other aspects concern expression vector intermediates made in accordance with the foregoing methods, and host cells containing the same. Still another aspect relates to methods of making multiple distinct expression vector intermediates useful in the practice of the present integration methods. In such methods, a base vector is amplified to generate two or more expression vector intermediates each having unique sequence-specific recombination regions which allow for homologous recombination with different insert nucleic acids. Such amplification reactions are preferably carried in separate reaction mixtures to produce distinct expression vector intermediates. In particularly preferred embodiments of such a high throughput approach, the requisite manipulations are performed in an automated fashion wherein one or more steps are performed by a computer-controlled device.

[0094] In some embodiments, any vector may be used to construct a pre-recombination molecule as provided herein. In particular, vectors known in the art and those commercially available (and variants or derivatives thereof) may be engineered to include recombination regions as described above. Such vectors may be obtained from, for example, Vector Laboratories Inc., InVitrogen, Promega, Novagen, NEB, Clontech, Boehringer Mannheim, Pharmacia, EpiCenter, OriGenes Technologies Inc., Stratagene, Perkin Elmer, Pharmingen, Life Technologies, Inc., and Research Genetics. General classes of vectors of particular interest include prokaryotic and/or eukaryotic cloning vectors, expression vectors, fusion vectors, two-hybrid or reverse two-hybrid vectors, shuttle vectors for use in different hosts, mutagenesis vectors, transcription vectors, vectors for receiving large inserts, and the like. Other vectors of interest include viral origin vectors (M13 vectors, bacterial phage λ vectors, adenovirus vectors, adeno-associated virus vectors (AAV) and retrovirus vectors), high, low and adjustable copy number vectors, vectors that have compatible replicons for use in combination in a single host (PACYC184 and pBR322) and eukaryotic episomal replication vectors (pCDM8). In other embodiments, a pre-recombination molecule may be obtained by standard procedures known in the art from cloned DNA (*e.g.*, a DNA “library”), by chemical synthesis, by cDNA cloning, or by the cloning of genomic DNA, or fragments thereof, purified from the desired cell, or by PCR amplification and cloning. *See, for example, Sambrook et al., Molecular Cloning, A Laboratory Manual*, 3d. ed., Cold

Spring Harbor Laboratory Press, Cold Spring Harbor, New York (2001); Glover, D.M. (ed.), *DNA Cloning: A Practical Approach*, 2d. ed., MRL Press, Ltd., Oxford, U.K. (1995).

5.2.1.1 Selectable Markers

[0095] In preferred embodiments, the circularized expression vector, once assembled, comprises a coding sequence for a selectable marker, and suitable regulatory sequences such as a promoter and/or a terminator that enables expression of the marker in the host cell. Useful selectable markers include those which function in both positive and negative selection systems.

[0096] In some embodiments, selection of the desired cells is based on selecting for drug resistance encoded by the selectable marker (**FIG. 3B**). Positive selection systems are those that promote the growth of transformed cells. They may be divided into conditional-positive or non-conditional-positive selection systems. A conditional-positive selection system consists of a gene coding for a protein, usually an enzyme, that confers resistance to a specific substrate that is toxic to untransformed cells or that encourages growth and/or differentiation of the transformed cells. In conditional-positive selection systems the substrate may act in one of several ways. It may be an antibiotic, an herbicide, a drug or metabolite analogue, or a carbon supply precursor. In each case, the gene codes for an enzyme with specificity to a substrate to encourage the selective growth and proliferation of the transformed cells. The substrate may be toxic or non-toxic to the untransformed cells. The *nptII* gene, which confers kanamycin resistance by inhibiting protein synthesis, is a classic example of a system that is toxic to untransformed cells. The *manA* gene, which codes for phosphomannose isomerase, is an example of a conditional-positive selection system where the selection substrate is not toxic. In this system, the substrate mannose is unable to act as a carbon source for untransformed cells but it will promote the growth of cells transformed with *manA*. Non-conditional-positive selection systems do not require external substrates yet promote the selective growth and differentiation of transformed cells. An example in plants is the *ipt* gene that enhances shoot development by modifying the plant hormone levels endogenously.

[0097] Negative selection systems result in the death of transformed cells. These are dominant selectable marker systems that may be described as conditional and non-conditional selection systems. When the selection system is not substrate dependent, it is a non-conditional-negative selection system. An example is the expression of a toxic protein, such as a ribonuclease to ablate specific cell types. When the action of the toxic gene requires a substrate to express toxicity, the system is a conditional negative selection system. These

include the bacterial *codA* gene, which codes for cytosine deaminase, the bacterial cytochrome P450 mono-oxygenase gene, the bacterial haloalkane dehalogenase gene, or the *Arabidopsis* alcohol dehydrogenase gene. Each of these converts non-toxic agents to toxic agents resulting in the death of the transformed cells. The *codA* gene has also been shown to be an effective dominant negative selection marker for chloroplast transformation. The *Agrobacterium aux2* and *tms2* genes are interesting in that they can also be used in positive selection systems. Combinations of positive-negative selection systems are particularly useful for the integration methods provided herein, as positive selection can be utilized to enrich for cells that have successfully recombined the circular expression vector (and presumptively, have performed one or more intended HR-mediated genomic integrations), and negative selection can be used to eliminate (“cure”) the expression vector from the same population once the desired genomic integrations have been confirmed.

[0098] A wide variety of selectable markers are known in the art (*see*, for example, Kaufman, *Meth. Enzymol.*, 185:487 (1990); Kaufman, *Meth. Enzymol.*, 185:537 (1990); Srivastava and Schlessinger, *Gene*, 103:53 (1991); Romanos *et al.*, in *DNA Cloning 2: Expression Systems*, 2nd Edition, pages 123-167 (IRL Press 1995); Markie, *Methods Mol. Biol.*, 54:359 (1996); Pfeifer *et al.*, *Gene*, 188:183 (1997); Tucker and Burke, *Gene*, 199:25 (1997); Hashida-Okado *et al.*, *FEBS Letters*, 425:117 (1998)). In some embodiments, the selectable marker is a drug resistant marker. A drug resistant marker enables cells to detoxify an exogenous drug that would otherwise kill the cell. Illustrative examples of drug resistant markers include but are not limited to those which confer resistance to antibiotics such as ampicillin, tetracycline, kanamycin, bleomycin, streptomycin, hygromycin, neomycin, Zeocin™, and the like. Other selectable markers include a bleomycin-resistance gene, a metallothionein gene, a hygromycin B-phosphotransferase gene, the AURI gene, an adenosine deaminase gene, an aminoglycoside phosphotransferase gene, a dihydrofolate reductase gene, a thymidine kinase gene, a xanthine-guanine phosphoribosyltransferase gene, and the like.

[0099] pBR and pUC-derived plasmids contain as a selectable marker the bacterial drug resistance marker *AMP^r* or *BLA* gene (*See*, Sutcliffe, J. G., *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 75:3737 (1978)). The *BLA* gene encodes the enzyme Tem-1, which functions as a beta-lactamase and is responsible for bacterial resistance to beta-lactam antibiotics, such as narrow-spectrum cephalosporins, cephamycins, and carbapenems (ertapenem), cefamandole, and cefoperazone, and all the anti-gram-negative-bacterium penicillins except temocillin.

[00100] Other useful selectable markers include but are not limited to: *NAT1*, *PAT*, *AUR1-C*, *PDR4*, *SMR1*, *CAT*, mouse dhfr, *HPH*, *DSDA*, *KAN^R*, and *SH BLE* genes. The *NAT1* gene of *S. noursei* encodes nourseothricin N-acetyltransferase and confers resistance to nourseothricin. The *PAT* gene from *S. viridochromogenes* Tu94 encodes phosphinothricin N-acetyltransferase and confers resistance to bialaphos. The *AUR1-C* gene from *S. cerevisiae* confers resistance to Auerobasidin A (AbA), an antifungal antibiotic produced by *Auerobasidium pullulans* that is toxic to budding yeast *S. cerevisiae*. The *PDR4* gene confers resistance to cerulenin. The *SMR1* gene confers resistance to sulfometuron methyl. The *CAT* coding sequence from Tn9 transposon confers resistance to chloramphenicol. The mouse dhfr gene confers resistance to methotrexate. The *HPH* gene of *Klebsiella pneumonia* encodes hygromycin B phosphotransferase and confers resistance to Hygromycin B. The *DSDA* gene of *E. coli* encodes D-serine deaminase and allows yeast to grow on plates with D-serine as the sole nitrogen source. The *KAN^R* gene of the Tn903 transposon encodes aminoglycoside phosphotransferase and confers resistance to G418. The *SH BLE* gene from *Streptoalloteichus hindustanus* encodes a Zeocin binding protein and confers resistance to Zeocin (bleomycin).

[00101] In other embodiments, the selectable marker is an auxotrophic marker. An auxotrophic marker allows cells to synthesize an essential component (usually an amino acid) while grown in media that lacks that essential component. Selectable auxotrophic gene sequences include, for example, *hisD*, which allows growth in histidine free media in the presence of histidinol. In some embodiments, the selectable marker rescues a nutritional auxotrophy in the host strain. In such embodiments, the host strain comprises a functional disruption in one or more genes of the amino acid biosynthetic pathways of the host that cause an auxotrophic phenotype, such as, for example, *HIS3*, *LEU2*, *LYS1*, *MET15*, and *TRP1*, or a functional disruption in one or more genes of the nucleotide biosynthetic pathways of the host that cause an auxotrophic phenotype, such as, for example, *ADE2* and *URA3*. In particular embodiments, the host cell comprises a functional disruption in the *URA3* gene. The functional disruption in the host cell that causes an auxotrophic phenotype can be a point mutation, a partial or complete gene deletion, or an addition or substitution of nucleotides. Functional disruptions within the amino acid or nucleotide biosynthetic pathways cause the host strains to become auxotrophic mutants which, in contrast to the prototrophic wild-type cells, are incapable of optimum growth in media without supplementation with one or more nutrients. The functionally disrupted biosynthesis genes in the host strain can then serve as

auxotrophic gene markers which can later be rescued, for example, upon introducing one or more plasmids comprising a functional copy of the disrupted biosynthesis gene.

[00102] In yeast, utilization of the *URA3*, *TRP1*, and *LYS2* genes as selectable markers has a marked advantage because both positive and negative selections are possible. Positive selection is carried out by auxotrophic complementation of the *URA3*, *TRP1*, and *LYS2* mutations whereas negative selection is based on the specific inhibitors 5-fluoro-orotic acid (FOA), 5-fluoroanthranilic acid, and α -aminoadipic acid (aAA), respectively, that prevent growth of the prototrophic strains but allow growth of the *URA3*, *TRP1*, and *LYS2* mutants, respectively. The *URA3* gene encodes orotidine-5'phosphate decarboxylase, an enzyme that is required for the biosynthesis of uracil. Ura3- (or ura5-) cells can be selected on media containing FOA, which kills all *URA3*+ cells but not *ura3*- cells because FOA appears to be converted to the toxic compound 5-fluorouracil by the action of decarboxylase. The negative selection on FOA media is highly discriminating, and usually less than 10^{-2} FOA-resistant colonies are Ura+. The FOA selection procedure can be used to produce *ura3* markers in haploid strains by mutation, and, more importantly, for selecting those cells that do not have the *URA3*-containing plasmids. The *TRP1* gene encodes a phosphoribosylanthranilate isomerase that catalyzes the third step in tryptophan biosynthesis. Counterselection using 5-fluoroanthranilic acid involves antimetabolism by the strains that lack enzymes required for the conversion of anthranilic acid to tryptophan and thus are resistant to 5-fluoroanthranilic acid. The *LYS2* gene encodes an aminoadipate reductase, an enzyme that is required for the biosynthesis of lysine. *Lys2*- and *lys5*- mutants, but not normal strains, grow on a medium lacking the normal nitrogen source but containing lysine and aAA. Apparently, *lys2* and *lys5* mutations cause the accumulation of a toxic intermediate of lysine biosynthesis that is formed by high levels of aAA, but these mutants still can use aAA as a nitrogen source. Similar with the FOA selection procedure, *LYS2*-containing plasmids can be conveniently expelled from *lys2* hosts. In other embodiments, the selectable marker is a marker other than one which rescues an auxotrophic mutation.

[00103] For any of the methods and compositions described herein, reporter genes, such as the *lac Z* reporter gene for facilitating blue/white selection of transformed colonies, or fluorescent proteins such as green, red and yellow fluorescent proteins, can be used as selectable marker genes to facilitate selection of HR-competent host cells that are able to successfully assemble the circular expression vector from one or more pre-recombination fragments (see **FIG. 3A**). In these embodiments, rather than growing the transformed cells in

media containing selective compound, *e.g.*, antibiotic, the cells are grown under conditions sufficient to allow expression of the reporter, and selection can be performed via visual, colorimetric or fluorescent detection of the reporter. Drug-free and selective pressure-free cell maintenance of the host cells can provide a number of advantages. For example, selective drugs and other selective pressure factors are often mutagenic or otherwise interfere with the physiology of the cells, leading to skewed results in cell-based assays. For example, selective drugs may decrease susceptibility to apoptosis (Robinson et al., *Biochemistry*, 36(37):11169-11178 (1997)), increase DNA repair and drug metabolism (Deffie et al., *Cancer Res.* 48(13):3595-3602 (1988)), increase cellular pH (Thiebaut et al., *J Histochem Cytochem.* 38(5):685-690 (1990); Roepe et al., *Biochemistry*. 32(41):11042-11056 (1993); Simon et al., *Proc Natl Acad Sci USA*. 91(3):1128-1132 (1994)), decrease lysosomal and endosomal pH (Schindler et al., *Biochemistry*. 35(9):2811-2817 (1996); Altan et al., *J Exp Med.* 187(10):1583-1598 (1998)), decrease plasma membrane potential (Roepe et al., *Biochemistry*. 32(41):11042-11056 (1993)), increase plasma membrane conductance to chloride (Gill et al., *Cell*. 71(1):23-32 (1992)) and ATP (Abraham et al., *Proc Natl Acad Sci USA*. 90(1):312-316 (1993)), and increase rates of vesicle transport (Altan et al., *Proc Natl Acad Sci USA*. 96(8):4432-4437 (1999)). Thus, the methods provided herein can be practiced with drug-free selection that allows for screening that is free from the artifacts caused by selective pressure.

[00104] A flow cytometric cell sorter can be used to isolate cells positive for expression of fluorescent markers or proteins (*e.g.*, antibodies) coupled to fluorophores and having affinity for the marker protein. In some embodiments, multiple rounds of sorting may be carried out. In one embodiment, the flow cytometric cell sorter is a FACS machine. Other fluorescence plate readers, including those that are compatible with high-throughput screening can also be used. MACS (magnetic cell sorting) can also be used, for example, to select for host cells with proteins coupled to magnetic beads and having affinity for the marker protein. This is especially useful where the selectable marker encodes, for example, a membrane protein, transmembrane protein, membrane anchored protein, cell surface antigen or cell surface receptor (*e.g.*, cytokine receptor, immunoglobulin receptor family member, ligand-gated ion channel, protein kinase receptor, G-protein coupled receptor (GPCR), nuclear hormone receptor and other receptors; CD14 (monocytes), CD56 (natural killer cells), CD335 (NKp46, natural killer cells), CD4 (T helper cells), CD8 (cytotoxic T cells), CD1c (BDCA-1, blood dendritic cell subset), CD303 (BDCA-2), CD304 (BDCA-4, blood dendritic cell subset), NKp80 (natural killer cells, gamma/delta T cells, effector/memory T cells), “6B11”

(Va24/Vb11; invariant natural killer T cells), CD137 (activated T cells), CD25 (regulatory T cells) or depleted for CD138 (plasma cells), CD4, CD8, CD19, CD25, CD45RA, CD45RO). Thus, in some embodiments, the selectable marker comprises a protein displayed on the host cell surface, which can be readily detected with an antibody, for example, coupled to a fluorophore or to a colorimetric or other visual readout.

5.2.1.2 Cell Culture

[00105] In some embodiments of the methods described herein, host cells transformed with one or more pre-recombination fragments are cultured for a period of time sufficient for expression of the selectable marker from the circularized expression vector.

[00106] In some embodiments where the selectable marker is a drug resistance marker, the culturing is carried out for a period of time sufficient to produce an amount of the marker protein that can support the survival of cells expressing the marker in selectable media. In preferable embodiments, these conditions also select against the survival of cells not expressing the selectable marker. Selective pressure can be applied to cells using a variety of compounds or treatments that would be known to one of skill in the art. Without being limited by theory, selective pressure can be applied by exposing host cells to conditions that are suboptimal for or deleterious to growth, progression of the cell cycle or viability, such that cells that are tolerant or resistant to these conditions are selected for compared to cells that are not tolerant or resistant to these conditions. Conditions that can be used to exert or apply selective pressure include but are not limited to antibiotics, drugs, mutagens, compounds that slow or halt cell growth or the synthesis of biological building blocks, compounds that disrupt RNA, DNA or protein synthesis, deprivation or limitation of nutrients, amino acids, carbohydrates or compounds required for cell growth and viability from cell growth or culture media, treatments such as growth or maintenance of cells under conditions that are suboptimal for cell growth, for instance at suboptimal temperatures, atmospheric conditions (e.g., % carbon dioxide, oxygen or nitrogen or humidity) or in deprived media conditions. The level of selective pressure that is used can be determined by one of skill in the art. This can be done, for example, by performing a kill curve experiment, where control cells and cells that comprise resistance markers or genes are tested with increasing levels, doses, concentrations or treatments of the selective pressure and the ranges that selected against the negative cells only or preferentially over a desired range of time (e.g., from 1 to 24 hours, 1 to 3 days, 3 to 5 days, 4 to 7 days, 5 to 14 days, 1 to 3 weeks, 2 to 6 weeks). The exact levels, concentrations, doses, or treatments of selective pressure that can be used depends on the cells that are used, the

desired properties themselves, the markers, factors or genes that confer resistance or tolerance to the selective pressure as well as the levels of the desired properties that are desired in the cells that are selected and one of skill in the art would readily appreciate how to determine appropriate ranges based on these considerations.

[00107] The culturing may be performed in a suitable culture medium in a suitable container, including but not limited to a cell culture plate, a flask, or a fermentor. In some embodiments, the culture medium is an aqueous medium comprising assimilable carbon, nitrogen and phosphate sources. Such a medium can also include appropriate salts, minerals, metals and other nutrients. In some embodiments, in addition to the selection agent, the suitable medium is supplemented with one or more additional agents, such as, for example, an inducer (*e.g.*, when one or more nucleotide sequences encoding a gene product are under the control of an inducible promoter), a repressor (*e.g.*, when one or more nucleotide sequences encoding a gene product are under the control of a repressible promoter). Materials and methods for the maintenance and growth of cell cultures are well known to those skilled in the art of microbiology or fermentation science (*see*, for example, Bailey *et al.*, Biochemical Engineering Fundamentals, second edition, McGraw Hill, New York, 1986). Consideration must be given to appropriate culture medium, pH, temperature, and requirements for aerobic, microaerobic, or anaerobic conditions, depending on the specific requirements of the host cell, the fermentation, and the process. In some embodiments, the culturing is carried out for a period of time sufficient for the transformed population to undergo a plurality of doublings until a desired cell density is reached. In some embodiments, the culturing is carried out for a period of time sufficient for the host cell population to reach a cell density (OD_{600}) of between 0.01 and 400 in the fermentation vessel or container in which the culturing is being carried out. In some embodiments, the culturing is carried out until an OD_{600} of at least 0.01 is reached. In some embodiments, the culturing is carried out until an OD_{600} of at least 0.1 is reached. In some embodiments, the culturing is carried out until an OD_{600} of at least 1.0 is reached. In some embodiments, the culturing is carried out until an OD_{600} of at least 10 is reached. In some embodiments, the culturing is carried out until an OD_{600} of at least 100 is reached. In some embodiments, the culturing is carried out until an OD_{600} of between 0.01 and 100 is reached. In some embodiments, the culturing is carried out until an OD_{600} of between 0.1 and 10 is reached. In some embodiments, the culturing is carried out until an OD_{600} of between 1 and 100 is reached. In other embodiments, the culturing is carried for a period of at least 12, 24, 36, 48, 60, 72, 84, 96 or more than 96 hours. In some embodiments, the culturing is carried

out for a period of between 3 and 20 days. In some embodiments, the culturing is carried out for a period of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more than 20 days.

[00108] In some embodiments of the methods described herein, the methods further comprise the step of eliminating the circularized expression vector, *i.e.*, plasmid, from the host cell, for example, once a selected host cell has been identified as comprising the desired genomic integration(s). Plasmid-based systems generally require selective pressure on the plasmids to maintain the foreign DNA in the cell. For example, most plasmids in yeast are relatively unstable, as a yeast cell typically loses 10% of plasmids contained in the cell after each mitotic division. Thus, in some embodiments, elimination of a plasmid encoding the selective marker from a selected cell can be achieved by allowing the selected cells to undergo sufficient mitotic divisions such that the plasmid is effectively diluted from the population. Alternatively, plasmid-free cells can be selected by selecting for the absence of the plasmid, *e.g.*, by selecting against a counter-selectable marker (such as, for example, URA3) or by plating identical colonies on both selective media and non-selective media and then selecting a colony that does not grow on the selective media but does grow on the non-selective media.

5.2.2. Exogenous Donor Nucleic Acids

[00109] Advantageously, an integration polynucleotide, *i.e.*, donor DNA, facilitates integration of one or more exogenous nucleic acid constructs into a selected target site of a host cell genome. In preferred embodiments, an integration polynucleotide comprises an exogenous nucleic acid (ES)_x comprising a first homology region (HR1)_x and a second homology region (HR2)_x, and optionally a nucleic acid of interest positioned between (HR1)_x and (HR2)_x. In some embodiments, the integration polynucleotide is a linear DNA molecule. In other embodiments, the integration polynucleotide is a circular DNA molecule. In some embodiments, the integration polynucleotide is a single-stranded DNA molecule, *i.e.*, an oligonucleotide. In other embodiments, the integration polynucleotide is a double-stranded DNA molecule.

[00110] The integration polynucleotide can be generated by any technique apparent to one skilled in the art. In certain embodiments, the integration polynucleotide is generated using polymerase chain reaction (PCR) and molecular cloning techniques well known in the art. *See, e.g., PCR Technology: Principles and Applications for DNA Amplification*, ed. HA Erlich, Stockton Press, New York, N.Y. (1989); Sambrook *et al.*, 2001, *Molecular Cloning – A*

Laboratory Manual, 3rd edition, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; *PCR Technology: Principles and Applications for DNA Amplification*, ed. HA Erlich, Stockton Press, New York, N.Y. (1989); U.S. Patent No. 8,110,360.

5.2.2.1 Genomic Integration Sequences

[00111] In preferred embodiments, an integration polynucleotide comprises an exogenous nucleic acid (ES)_x comprising a first homology region (HR1)_x and a second homology region (HR2)_x, wherein (HR1)_x and (HR2)_x are capable of initiating host cell mediated homologous recombination at a selected target site (TS)_x within the host cell genome. To integrate an exogenous nucleic acid into the genome by homologous recombination, the integration polynucleotide preferably comprises (HR1)_x at one terminus and (HR2)_x at the other terminus. In some embodiments, (HR1)_x is homologous to a 5' region of the selected genomic target site (TS)_x, and (HR2)_x is homologous to a 3' region of the selected target site (TS)_x. In some embodiments, (HR1)_x is about 70%, 75%, 80%, 85%, 90%, 95% or 100% homologous to a 5' region of the selected genomic target site (TS)_x. In some embodiments, (HR2)_x is about 70%, 75%, 80%, 85%, 90%, 95% or 100% homologous to a 3' region of the selected target site (TS)_x.

[00112] In certain embodiments, (HR1)_x is positioned 5' to a nucleic acid of interest (D)_x. In some embodiments, (HR1)_x is positioned immediately adjacent to the 5' end of (D)_x. In some embodiments, (HR1)_x is positioned upstream to the 5' of (D)_x. In certain embodiments, (HR2)_x is positioned 3' to a nucleic acid of interest (D)_x. In some embodiments, (HR2)_x is positioned immediately adjacent to the 3' end of (D)_x. In some embodiments, (HR2)_x is positioned downstream to the 3' of (D)_x.

[00113] Properties that may affect the integration of an integration polynucleotide at a particular genomic locus include but are not limited to: the lengths of the genomic integration sequences, the overall length of the excisable nucleic acid construct, and the nucleotide sequence or location of the genomic integration locus. For instance, effective heteroduplex formation between one strand of a genomic integration sequence and one strand of a particular locus in a host cell genome may depend on the length of the genomic integration sequence. An effective range for the length of a genomic integration sequence is 50 to 5,000 nucleotides. For a discussion of effective lengths of homology between genomic integration sequences and genomic loci. *See, Hasty et al., Mol Cell Biol* 11:5586-91 (1991).

[00114] In some embodiments, (HR1)_x and (HR2)_x can comprise any nucleotide sequence of sufficient length and sequence identity that allows for genomic integration of the exogenous nucleic acid (ES)_x at any yeast genomic locus. In certain embodiments, each of (HR1)_x and (HR2)_x independently consists of about 50 to 5,000 nucleotides. In certain embodiments, each of (HR1)_x and (HR2)_x independently consists of about 100 to 2,500 nucleotides. In certain embodiments, each of (HR1)_x and (HR2)_x independently consists of about 100 to 1,000 nucleotides. In certain embodiments, each of (HR1)_x and (HR2)_x independently consists of about 250 to 750 nucleotides. In certain embodiments, each of (HR1)_x and (HR2)_x independently consists of about 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3100, 3200, 3300, 3400, 3500, 3600, 3700, 3800, 3900, 4000, 4100, 4200, 4300, 4400, 4500, 4600, 4700, 4800, 4900 or 5,000 nucleotides. In some embodiments, each of (HR1)_x and (HR2)_x independently consists of about 500 nucleotides.

5.2.2.2 Nucleic Acids of Interest

[00115] In some embodiments, the integration polynucleotide further comprises a nucleic acid of interest (D)_x. The nucleic acid of interest can be any DNA segment deemed useful by one of skill in the art. For example, the DNA segment may comprise a gene of interest that can be “knocked in” to a host genome. In other embodiments, the DNA segment functions as a “knockout” construct that is capable of specifically disrupting a target gene upon integration of the construct into the target site of the host cell genome, thereby rendering the disrupted gene non-functional. Useful examples of a nucleic acid of interest (D)_x include but are not limited to: a protein-coding sequence, reporter gene, fluorescent marker coding sequence, promoter, enhancer, terminator, transcriptional activator, transcriptional repressor, transcriptional activator binding site, transcriptional repressor binding site, intron, exon, poly-A tail, multiple cloning site, nuclear localization signal, mRNA stabilization signal, integration loci, epitope tag coding sequence, degradation signal, or any other naturally occurring or synthetic DNA molecule. In some embodiments, (D)_x can be of natural origin. Alternatively, (D)_x can be completely of synthetic origin, produced *in vitro*. Furthermore, (D)_x can comprise any combination of isolated naturally occurring DNA molecules, or any combination of an isolated naturally occurring DNA molecule and a synthetic DNA molecule. For example, (D)_x may comprise a heterologous promoter operably linked to a protein coding sequence, a protein coding sequence linked to a poly-A tail, a protein coding sequence linked in-frame with a epitope tag coding sequence, and the like. The nucleic acid of interest (D)_x may be obtained by

standard procedures known in the art from cloned DNA (*e.g.*, a DNA “library”), by chemical synthesis, by cDNA cloning, or by the cloning of genomic DNA, or fragments thereof, purified from the desired cell, or by PCR amplification and cloning. *See*, for example, Sambrook *et al.*, *Molecular Cloning, A Laboratory Manual*, 3d. ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (2001); Glover, D.M. (ed.), *DNA Cloning: A Practical Approach*, 2d. ed., MRL Press, Ltd., Oxford, U.K. (1995).

[00116] In particular embodiments, the nucleic acid of interest (D)_x does not comprise nucleic acid encoding a selectable marker. In these embodiments, the high efficiency of integration provided by the methods described herein allows for the screening and identification of integration events without the requirement for growth of transformed cells on selection media. However, in other embodiments where growth on selective media is nonetheless desired, the nucleic acid of interest (D)_x can comprise a selectable marker that may be used to select for the integration of the exogenous nucleic acid into a host genome.

[00117] The nucleic acid of interest (D)_x can be of any size, including sizes ranging from about 300 nucleotides to up to about 1 million nucleotide base pairs. Such nucleic acids of interest may include one or more genes and/or their associated regulatory regions. These nucleic acids of interest can be derived from any source, for example, from genomic sources or from cDNA libraries, including tissue-specific, normalized, and subtractive cDNA libraries. Genomic sources include the genomes (or fragments thereof) of various organisms, including pathogenic organisms such as viruses (*e.g.*, HIV and hepatitis viruses) and cellular pathogens. Moreover, nucleic acids of interest can be obtained from any organism, including any plant or any animal, be they eukaryotic or prokaryotic. In certain embodiments, a nucleic acid of interest encodes a gene which is a disease-associated gene, *i.e.*, the presence, absence, expression, lack of expression, altered level of expression, or existence of an altered form of which correlates with or causes a disease.

[00118] In some embodiments, the nucleic acid of interest encodes a point mutation of a targeted allele of a host cell, which can be utilized for the introduction of a missense SNP (*i.e.* an “allele swap”) to the targeted allele. In some such embodiments, the selection of nuclease (*e.g.*, CRISPR/Cas9 and gRNA) target sites for an allele swap is considerably more constrained than for deletion or integration into an ORF. In preferred embodiments, the nuclease cleavage site should be unique in the genome, and should be as close to the targeted nucleotide as possible, such that recombination will incorporate the mutant sequence, rather than just the flanking sequence of the donor DNA. This is because recombination to repair the cut site does

not require incorporation of the desired SNP, and the likelihood of its inclusion is expected to decrease with distance from the cut site. Additionally, for optimal efficiency, the donor DNA should be designed such that it is not also a target for the nuclease (*e.g.*, CRISPR/Cas9 and gRNA). Thus, to make the donor DNA immune to cutting, and simultaneously improve the chances that recombination events include the desired SNP, a heterology block approach can be utilized whereby silent mutations are made in the codons between the target site and the point mutation, reducing the potential for recombination events that would omit the desired SNP. Donor DNAs can be designed with flanking homology surrounding a central “heterology block”. The heterology block introduces silent mutations to the sequence surrounding the nuclease target site, and serves several purposes. First, it removes bases critical for nuclease (*e.g.*, CRISPR-Cas9) recognition, such that the donor DNA will not be cut. Additionally, integration of the heterology block provides a novel primer binding site to identify candidate clones by PCR. **FIG. 15** provides a schematic for CRISPR/Cas9-mediated introduction of a point mutation in the context of a “heterology block.” A targeted amino acid is boxed, and an adjacent cleavage site is annotated with cleavage site and PAM sequence (Top panel). A donor DNA containing the desired point mutation in the context of a heterology block of silent codon changes and flanking homology can be generated synthetically by annealing and extending 60-mer oligos (Middle panel) or with larger cloned constructs. Integration of the donor DNA yields the desired point mutation (Lower panel).

5.2.3. Nucleases

[00119] In some embodiments of the methods described herein, a host cell genome is contacted with one or more nucleases capable of cleaving, *i.e.*, causing a break at a designated region within a selected target site. In some embodiments, the break is a single-stranded break, that is, one but not both DNA strands of the target site are cleaved (*i.e.*, “nicked”). In some embodiments, the break is a double-stranded break. In some embodiments, a break inducing agent is any agent that recognizes and/or binds to a specific polynucleotide recognition sequence to produce a break at or near the recognition sequence. Examples of break inducing agents include, but are not limited to, endonucleases, site-specific recombinases, transposases, topoisomerases, and zinc finger nucleases, and include modified derivatives, variants, and fragments thereof.

[00120] In some embodiments, each of the one or more nucleases is capable of causing a break at a designated region within a selected target site (TS)_x. In some embodiments, the nuclease is capable of causing a break at a region positioned between the 5' and 3' regions of

(TS)_x with which (HR1)_x and (HR2)_x share homology, respectively. In other embodiments, the nuclease is capable of causing a break at a region positioned upstream or downstream of the 5' and 3' regions of (TS)_x.

[00121] A recognition sequence is any polynucleotide sequence that is specifically recognized and/or bound by a break inducing agent. The length of the recognition site sequence can vary, and includes, for example, sequences that are at least 10, 12, 14, 16, 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 or more nucleotides in length.

[00122] In some embodiments, the recognition sequence is palindromic, that is, the sequence on one strand reads the same in the opposite direction on the complementary strand. In some embodiments, the nick/cleavage site is within the recognition sequence. In other embodiments, the nick/cleavage site is outside of the recognition sequence. In some embodiments, cleavage produces blunt end termini. In other embodiments, cleavage produces single-stranded overhangs, *i.e.*, “sticky ends,” which can be either 5' overhangs, or 3' overhangs.

[00123] In some embodiments, the recognition sequence within the selected target site can be endogenous or exogenous to the host cell genome. When the recognition site is an endogenous sequence, it may be a recognition sequence recognized by a naturally-occurring, or native break inducing agent. Alternatively, an endogenous recognition site could be recognized and/or bound by a modified or engineered break inducing agent designed or selected to specifically recognize the endogenous recognition sequence to produce a break. In some embodiments, the modified break inducing agent is derived from a native, naturally-occurring break inducing agent. In other embodiments, the modified break inducing agent is artificially created or synthesized. Methods for selecting such modified or engineered break inducing agents are known in the art. For example, amino acid sequence variants of the protein(s) can be prepared by mutations in the DNA. Methods for mutagenesis and nucleotide sequence alterations include, for example, Kunkel, (1985) *Proc Natl Acad Sci USA* 82:488-92; Kunkel, *et al.*, (1987) *Meth Enzymol* 154:367-82; U.S. Pat. No. 4,873,192; Walker and Gaastra, eds. (1983) *Techniques in Molecular Biology* (MacMillan Publishing Company, New York) and the references cited therein. Guidance regarding amino acid substitutions not likely to affect biological activity of the protein is found, for example, in the model of Dayhoff, *et al.*, (1978) *Atlas of Protein Sequence and Structure* (Natl Biomed Res Found, Washington, D.C.).

Conservative substitutions, such as exchanging one amino acid with another having similar properties, may be preferable. Conservative deletions, insertions, and amino acid substitutions are not expected to produce radical changes in the characteristics of the protein, and the effect of any substitution, deletion, insertion, or combination thereof can be evaluated by routine screening assays. Assays for double strand break inducing activity are known and generally measure the overall activity and specificity of the agent on DNA substrates containing recognition sites.

5.2.3.1 **Clustered Regulatory Interspaced Short Palindromic Repeats (CRISPR)**

[00124] In some embodiments of the methods provided herein, the nuclease is a CRISPR/Cas-derived RNA-guided endonuclease. CRISPR is a genome editing tool based on the type II prokaryotic CRISPR (clustered regularly interspaced short palindromic repeats) adaptive immune system. CRISPR systems in eubacteria and archaea use small RNAs and CRISPR-associated (Cas) endonucleases to target and cleave invading foreign DNAs. *See, e.g.,* Bhaya et al., *Annu Rev Genet* 45:273-297 (2011); Terns et al., *Curr Opin Microbiol* 14(3):321-327 (2011); and Wiedenheft et al., *Nature* 482 (7385):331-338. In bacteria, CRISPR loci are composed of a series of repeats separated by segments of exogenous DNA (of ~30bp in length) called spacers. The repeat-spacer array is transcribed as a long precursor and processed within repeat sequences to generate small crRNAs that specify the target sequences (also known as protospacers) cleaved by the CRISPR nuclease. CRISPR spacers are then used to recognize and silence exogenous genetic elements at the RNA or DNA level. Essential for cleavage is a sequence motif immediately downstream on the 3' end of the target region, known as the protospacer-adjacent motif (PAM). The PAM is present in the target DNA, but not the crRNA that targets it.

[00125] One of the simplest CRISPR systems is the type II CRISPR system from *Streptococcus pyogenes*. The CRISPR-associated Cas9 endonuclease and two small RNAs, a target-complimentary CRISPR RNA (crRNA); and a transacting crRNA (tracrRNA), are sufficient for RNA-guided cleavage of foreign DNAs. The Cas9 protein, a hallmark protein of the type II CRISPR-Cas system, is a large monomeric DNA nuclease containing two nuclease domains homologous to RuvC and HNH nucleases. Cas9 is guided to a DNA target sequence adjacent to the PAM (protospacer adjacent motif) sequence motif by a crRNA:tracrRNA complex. Mature crRNA base-pairs to tracrRNA to form a two-RNA structure that directs Cas9 to the target DNA. At sites complementary to the crRNA-guide sequence, the Cas9 HNH

nuclease domain cleaves the complementary strand, whereas the Cas9 RuvC-like domain cleaves the noncomplementary strand, resulting in a double strand break in the target DNA. *See, e.g., Deltcheva et al., Nature 47(7340):602-607 (2011).*

[00126] Recent studies show that a single guide RNA (gRNA) chimera that mimics the crRNA:tracrRNA complex can be utilized with Cas9 as a genome editing tool to guide Cas9 to introduce site specific DNA double-stranded breaks *in vitro*. Specificity of the cleavage within the target genome is determined by the spacer-like moiety of a chimeric guide RNA molecule (gRNA), which mimics the native crRNA:tracrRNA complex. Thus, the minimum number of components in a functional CRISPR/Cas system is two: Cas9 and sgRNA. The sgRNA guide sequence located at its 5' end confers DNA target specificity. Therefore, by modifying the guide sequence, it is possible to create sgRNAs with different target specificities. The canonical length of the guide sequence is 20 bp. Consequently, a DNA target is also 20 bp followed by a PAM sequence that follows the consensus NGG. Use of this modified CRISPR system has been demonstrated *in vitro* (*see, e.g., Jinek et al., Science 337(6096):816-821 (2012)*), in mammalian cell lines (*see, e.g., Mali et al., Science 339(6121):823-826 (2013)*, Jinek *et al., Elife 2:e00417 (2013)*; Cong *et al., Science 339(6121):819-823 (2013)*; and Cho *et al., Nat Biotechnol 31(3):230-232 (2013)*), in bacteria (*see, e.g., Jiang et al., Nat Biotechnol 31(3):233-239 (2013)*; and Gasiunas *et al., Proc Natl Acad Sci USA 109(39):E2579-E2586. (2012)*), yeast (*see, e.g., DiCarlo et al., Nucleic Acid Res 41(7):4336-4343 (2013)*), zebrafish (*see, e.g., Hwang et al., Nat Biotechnol 31(3):227-229 (2013)*; and Chang *et al., Cell Res 23(4):465-472 (2013)*), mice (*see, e.g., Wang et al., Cell 153(4):910-918 (2013)*), and plants (*see e.g., Belhaj et al., Plant Methods 9:39 (2013)*).

[00127] The Cas9 nuclease may be modified by: (1) codon optimization for increased expression within a heterologous host; (2) fusion to a nuclear localization signal (NLS) for proper compartmentalization; and (3) site directed mutagenesis of either the HNH or RuvC domain to convert the nuclease into a strand-specific nickase. Site-directed mutagenesis of Cas9 in either the RuvC- or HNH-motif showed strand cleavage specificity, thereby providing two strand-specific nickases, in addition to the wild-type endonuclease and enabling targeted single-strand breaks of DNA. *See, e.g., Jinek et al., Science 337(6096):816-821 (2012)*, and Gasiunas *et al., Proc Natl Acad Sci USA 109(39):E2579-E2586. (2012)*. As has been reported for zinc finger nucleases and TALENs, modifying the nuclease to function as a nickase that breaks only one strand reduces toxicity from off-target cutting, and may also lower rates of

break repair via non-HR mechanisms, *e.g.*, NHEJ. *See, e.g.*, Jinek *et al.*, *Science* 337(6096):816-821 (2012).

[00128] Any CRISPR/Cas system known in the art finds use as a nuclease in the methods and compositions provided herein. The highly diverse CRISPR-Cas systems are categorized into three major types, which are further subdivided into ten subtypes, based on core element content and sequences (*see, e.g.*, Makarova *et al.*, *Nat Rev Microbiol* 9:467-77 (2011)). The structural organization and function of nucleoprotein complexes involved in crRNA-mediated silencing of foreign nucleic acids differ between distinct CRISPR/Cas types (*see* Wiedenheft *et al.*, *Nature* 482:331-338 (2012)). In the Type I-E system, as exemplified by *Escherichia coli*, crRNAs are incorporated into a multisubunit effector complex called Cascade (CRISPR-associated complex for antiviral defence) (Brouns *et al.*, *Science* 321 : 960-4 (2008)), which binds to the target DNA and triggers degradation by the signature Cas3 protein (Sinkunas *et al.*, *EMBO J* 30:1335-2 (2011); Beloglazova *et al.*, *EMBO J* 30:616-27 (2011)). In Type III CRISPR/Cas systems of *Sulfolobus solfataricus* and *Pyrococcus furiosus*, Cas RAMP module (Cmr) and crRNA complex recognize and cleave synthetic RNA *in vitro* (Hale *et al.*, *Mol Cell* 45:292-302 (2012); Zhang *et al.*, *Mol Cell*, 45:303-13 (2012)), while the CRISPR/Cas system of *Staphylococcus epidermidis* targets DNA *in vivo* (Marraffini & Sontheimer, *Science*. 322:1843-5 (2008)). RNP complexes involved in DNA silencing by Type II CRISPR/Cas systems, more specifically in the CRISPR3/Cas system of *Streptococcus thermophilus* DGCC7710 (Horvath & Barrangou, *Science* 327:167-70 (2010)), consists of four cas genes: *cas9*, *cas1*, *cas2*, and *csn2*, that are located upstream of 12 repeat-spacer units. *Cas9* (formerly named *cas5* or *csn1*) is the signature gene for Type II systems (Makarova *et al.*, *Nat Rev Microbiol* 9:467-77 (2011)).

[00129] CRISPR systems that find use in the methods and compositions provided herein also include those described in International Publication Numbers WO 2013/142578 A1 and WO 2013/098244 A1.

5.2.3.2 Transcription Activator-like Effector Nucleases (TALENs)

[00130] In some embodiments of the methods provided herein, one or more of the nucleases is a TAL-effector DNA binding domain-nuclease fusion protein (TALEN). TAL effectors of plant pathogenic bacteria in the genus *Xanthomonas* play important roles in disease, or trigger defense, by binding host DNA and activating effector-specific host genes. *see, e.g.*, Gu *et al.*

(2005) *Nature* 435:1122-5; Yang *et al.*, (2006) *Proc. Natl. Acad. Sci. USA* 103:10503-8; Kay *et al.*, (2007) *Science* 318:648-51; Sugio *et al.*, (2007) *Proc. Natl. Acad. Sci. USA* 104:10720-5; Romer *et al.*, (2007) *Science* 318:645-8; Boch *et al.*, (2009) *Science* 326(5959):1509-12; and Moscou and Bogdanove, (2009) 326(5959):1501. A TAL effector comprises a DNA binding domain that interacts with DNA in a sequence-specific manner through one or more tandem repeat domains. The repeated sequence typically comprises 34 amino acids, and the repeats are typically 91-100% homologous with each other. Polymorphism of the repeats is usually located at positions 12 and 13, and there appears to be a one-to-one correspondence between the identity of repeat variable-diresidues at positions 12 and 13 with the identity of the contiguous nucleotides in the TAL-effector's target sequence.

[00131] The TAL-effector DNA binding domain may be engineered to bind to a desired target sequence, and fused to a nuclease domain, *e.g.*, from a type II restriction endonuclease, typically a nonspecific cleavage domain from a type II restriction endonuclease such as FokI (*see e.g.*, Kim *et al.* (1996) *Proc. Natl. Acad. Sci. USA* 93:1156-1160). Other useful endonucleases may include, for example, HhaI, HindIII, Nod, BbvCI, EcoRI, BglII, and AlwI. Thus, in preferred embodiments, the TALEN comprises a TAL effector domain comprising a plurality of TAL effector repeat sequences that, in combination, bind to a specific nucleotide sequence in the target DNA sequence, such that the TALEN cleaves the target DNA within or adjacent to the specific nucleotide sequence. TALENS useful for the methods provided herein include those described in WO10/079430 and U.S. Patent Application Publication No. 2011/0145940.

[00132] In some embodiments, the TAL effector domain that binds to a specific nucleotide sequence within the target DNA can comprise 10 or more DNA binding repeats, and preferably 15 or more DNA binding repeats. In some embodiments, each DNA binding repeat comprises a repeat variable-diresidue (RVD) that determines recognition of a base pair in the target DNA sequence, wherein each DNA binding repeat is responsible for recognizing one base pair in the target DNA sequence, and wherein the RVD comprises one or more of: HD for recognizing C; NG for recognizing T; NI for recognizing A; NN for recognizing G or A; NS for recognizing A or C or G or T; N* for recognizing C or T, where * represents a gap in the second position of the RVD; HG for recognizing T; H* for recognizing T, where * represents a gap in the second position of the RVD; IG for recognizing T; NK for recognizing G; HA for recognizing C; ND for recognizing C; HI for recognizing C; HN for recognizing G; NA for recognizing G; SN for recognizing G or A; and YG for recognizing T.

[00133] In some embodiments of the methods provided herein, one or more of the nucleases is a site-specific recombinase. A site-specific recombinase, also referred to as a recombinase, is a polypeptide that catalyzes conservative site-specific recombination between its compatible recombination sites, and includes native polypeptides as well as derivatives, variants and/or fragments that retain activity, and native polynucleotides, derivatives, variants, and/or fragments that encode a recombinase that retains activity. For reviews of site-specific recombinases and their recognition sites, see, Sauer (1994) *Curr Op Biotechnol* 5:521-7; and Sadowski, (1993) *FASEB* 7:760-7. In some embodiments, the recombinase is a serine recombinase or a tyrosine recombinase. In some embodiments, the recombinase is from the Integrase or Resolvase families. In some embodiments, the recombinase is an integrase selected from the group consisting of FLP, Cre, lambda integrase, and R. For other members of the Integrase family, see for example, Esposito, *et al.*, (1997) *Nucleic Acids Res* 25:3605-14 and Abremski, *et al.*, (1992) *Protein Eng* 5:87-91. Methods for modifying the kinetics, cofactor interaction and requirements, expression, optimal conditions, and/or recognition site specificity, and screening for activity of recombinases and variants are known, see for example Miller, *et al.*, (1980) *Cell* 20:721-9; Lange-Gustafson and Nash, (1984) *J Biol Chem* 259:12724-32; Christ, *et al.*, (1998) *J Mol Biol* 288:825-36; Lorbach, *et al.*, (2000) *J Mol Biol* 296:1175-81; Vergunst, *et al.*, (2000) *Science* 290:979-82; Dorgai, *et al.*, (1995) *J Mol Biol* 252:178-88; Dorgai, *et al.*, (1998) *J Mol Biol* 277:1059-70; Yagu, *et al.*, (1995) *J Mol Biol* 252:163-7; Scimiente, *et al.*, (2001) *Nucleic Acids Res* 29:5044-51; Santoro and Schultze, (2002) *Proc Natl Acad Sci USA* 99:4185-90; Buchholz and Stewart, (2001) *Nat Biotechnol* 19:1047-52; Voziyanov, *et al.*, (2002) *Nucleic Acids Res* 30:1656-63; Voziyanov, *et al.*, (2003) *J Mol Biol* 326:65-76; Klippel, *et al.*, (1988) *EMBO J* 7:3983-9; Arnold, *et al.*, (1999) *EMBO J* 18:1407-14; WO03/08045; WO99/25840; and WO99/25841. The recognition sites range from about 30 nucleotide minimal sites to a few hundred nucleotides. Any recognition site for a recombinase can be used, including naturally occurring sites, and variants. Variant recognition sites are known, see for example Hoess, *et al.*, (1986) *Nucleic Acids Res* 14:2287-300; Albert, *et al.*, (1995) *Plant J* 7:649-59; Thomson, *et al.*, (2003) *Genesis* 36:162-7; Huang, *et al.*, (1991) *Nucleic Acids Res* 19:443-8; Siebler and Bode, (1997) *Biochemistry* 36:1740-7; Schlake and Bode, (1994) *Biochemistry* 33:12746-51; Thygarajan, *et al.*, (2001) *Mol Cell Biol* 21:3926-34; Umlauf and Cox, (1988) *EMBO J* 7:1845-52; Lee and Saito, (1998) *Gene* 216:55-65; WO01/23545; WO99/25821; WO99/25851; WO01/11058; WO01/07572 and U.S. Pat. No. 5,888,732.

[00134] In some embodiments of the methods provided herein, one or more of the nucleases is a transposase. Transposases are polypeptides that mediate transposition of a transposon from one location in the genome to another. Transposases typically induce double strand breaks to excise the transposon, recognize subterminal repeats, and bring together the ends of the excised transposon, in some systems other proteins are also required to bring together the ends during transposition. Examples of transposons and transposases include, but are not limited to, the Ac/Ds, Dt/rdt, Mu-M1/Mn, and Spm(En)/dSpm elements from maize, the Tam elements from snapdragon, the Mu transposon from bacteriophage, bacterial transposons (Tn) and insertion sequences (IS), Ty elements of yeast (retrotransposon), Ta1 elements from *Arabidopsis* (retrotransposon), the P element transposon from *Drosophila* (Gloor, *et al.*, (1991) *Science* 253:1110-1117), the Copia, Mariner and Minos elements from *Drosophila*, the Hermes elements from the housefly, the PiggyBack elements from *Trichplusia ni*, Tc1 elements from *C. elegans*, and IAP elements from mice (retrotransposon).

5.2.3.3 Zinc Finger Nucleases (ZFNs)

[00135] In some embodiments of the methods provided herein, one or more of the nucleases is a zinc-finger nuclease (ZFN). ZFNs are engineered break inducing agents comprised of a zinc finger DNA binding domain and a break inducing agent domain. Engineered ZFNs consist of two zinc finger arrays (ZFAs), each of which is fused to a single subunit of a non-specific endonuclease, such as the nuclease domain from the *FokI* enzyme, which becomes active upon dimerization. Typically, a single ZFA consists of 3 or 4 zinc finger domains, each of which is designed to recognize a specific nucleotide triplet (GGC, GAT, *etc.*). Thus, ZFNs composed of two "3-finger" ZFAs are capable of recognizing an 18 base pair target site; an 18 base pair recognition sequence is generally unique, even within large genomes such as those of humans and plants. By directing the co-localization and dimerization of two *FokI* nuclease monomers, ZFNs generate a functional site-specific endonuclease that creates a break in DNA at the targeted locus.

[00136] Useful zinc-finger nucleases include those that are known and those that are engineered to have specificity for one or more target sites (TS) described herein. Zinc finger domains are amenable for designing polypeptides which specifically bind a selected polynucleotide recognition sequence, for example, within the target site of the host cell genome. ZFNs consist of an engineered DNA-binding zinc finger domain linked to a non-specific endonuclease domain, for example nuclease domain from a Type II endonuclease such as HO or *FokI*. Alternatively, engineered zinc finger DNA binding domains can be fused

to other break inducing agents or derivatives thereof that retain DNA nicking/cleaving activity. For example, this type of fusion can be used to direct the break inducing agent to a different target site, to alter the location of the nick or cleavage site, to direct the inducing agent to a shorter target site, or to direct the inducing agent to a longer target site. In some examples a zinc finger DNA binding domain is fused to a site-specific recombinase, transposase, or a derivative thereof that retains DNA nicking and/or cleaving activity. Additional functionalities can be fused to the zinc-finger binding domain, including transcriptional activator domains, transcription repressor domains, and methylases. In some embodiments, dimerization of nuclease domain is required for cleavage activity.

[00137] Each zinc finger recognizes three consecutive base pairs in the target DNA. For example, a 3 finger domain recognized a sequence of 9 contiguous nucleotides, with a dimerization requirement of the nuclease, two sets of zinc finger triplets are used to bind a 18 nucleotide recognition sequence. Useful designer zinc finger modules include those that recognize various GNN and ANN triplets (Dreier, *et al.*, (2001) *J Biol Chem* 276:29466-78; Dreier, *et al.*, (2000) *J Mol Biol* 303:489-502; Liu, *et al.*, (2002) *J Biol Chem* 277:3850-6), as well as those that recognize various CNN or TNN triplets (Dreier, *et al.*, (2005) *J Biol Chem* 280:35588-97; Jamieson, *et al.*, (2003) *Nature Rev Drug Discov* 2:361-8). See also, Durai, *et al.*, (2005) *Nucleic Acids Res* 33:5978-90; Segal, (2002) *Methods* 26:76-83; Porteus and Carroll, (2005) *Nat Biotechnol* 23:967-73; Pabo, *et al.*, (2001) *Ann Rev Biochem* 70:313-40; Wolfe, *et al.*, (2000) *Ann Rev Biophys Biomol Struct* 29:183-212; Segal and Barbas, (2001) *Curr Opin Biotechnol* 12:632-7; Segal, *et al.*, (2003) *Biochemistry* 42:2137-48; Beerli and Barbas, (2002) *Nat Biotechnol* 20:135-41; Carroll, *et al.*, (2006) *Nature Protocols* 1:1329; Ordiz, *et al.*, (2002) *Proc Natl Acad Sci USA* 99:13290-5; Guan, *et al.*, (2002) *Proc Natl Acad Sci USA* 99:13296-301; WO2002099084; WO00/42219; WO02/42459; WO2003062455; US20030059767; US Patent Application Publication Number 2003/0108880; U.S. Pat. Nos. 6,140,466, 6,511,808 and 6,453,242. Useful zinc-finger nucleases also include those described in WO03/080809; WO05/014791; WO05/084190; WO08/021207; WO09/042186; WO09/054985; and WO10/065123.

5.2.3.4 Endonucleases

[00138] In some embodiments of the methods provided herein, one or more of the nucleases is an endonuclease. Endonucleases are enzymes that cleave the phosphodiester bond within a polynucleotide chain, and include restriction endonucleases that cleave DNA at specific sites without damaging the bases. Restriction endonucleases include Type I, Type II, Type III, and

Type IV endonucleases, which further include subtypes. Restriction endonucleases are further described and classified, for example in the REBASE database (webpage at rebase.neb.com; Roberts, *et al.*, (2003) *Nucleic Acids Res* 31:418-20), Roberts, *et al.*, (2003) *Nucleic Acids Res* 31:1805-12, and Belfort, *et al.*, (2002) in *Mobile DNA II*, pp. 761-783, Eds. Craigie, *et al.*, ASM Press, Washington, D.C.

[00139] As used herein, endonucleases also include homing endonucleases, which like restriction endonucleases, bind and cut at a specific recognition sequence. However the recognition sites for homing endonucleases are typically longer, for example, about 18 bp or more. Homing endonucleases, also known as meganucleases, have been classified into the following families based on conserved sequence motifs: an LAGLIDADG (SEQ ID NO:1) homing endonuclease, an HNH homing endonuclease, a His-Cys box homing endonuclease, a GIY-YIG (SEQ ID NO:2) homing endonuclease, and a cyanobacterial homing endonuclease. *See, e.g.*, Stoddard, *Quarterly Review of Biophysics* 38(1): 49-95 (2006). These families differ greatly in their conserved nuclease active-site core motifs and catalytic mechanisms, biological and genomic distributions, and wider relationship to non-homing nuclease systems. *See, for example*, Guhan and Muniyappa (2003) *Crit Rev Biochem Mol Biol* 38:199-248; Lucas, *et al.*, (2001) *Nucleic Acids Res* 29:960-9; Jurica and Stoddard, (1999) *Cell Mol Life Sci* 55:1304-26; Stoddard, (2006) *Q Rev Biophys* 38:49-95; and Moure, *et al.*, (2002) *Nat Struct Biol* 9:764. Examples of useful specific homing endonucleases from these families include, but are not limited to: I-CreI (*see*, Rochaix *et al.*, *Nucleic Acids Res.* 13: 975-984 (1985), I-MsoI (*see*, Lucas *et al.*, *Nucleic Acids Res.* 29: 960-969 (2001), I-SceI (*see*, Foury *et al.*, *FEBS Lett.* 440: 325-331 (1998), I-SceIV (*see*, Moran *et al.*, *Nucleic Acids Res.* 20: 4069-4076 (1992), H-DreI (*see*, Chevalier *et al.*, *Mol. Cell* 10: 895-905 (2002), I-HmuI (*see*, Goodrich-Blair *et al.*, *Cell* 63: 417-424 (1990); Goodrich-Blair *et al.*, *Cell* 84: 211-221 (1996), I-PpoI (*see*, Muscarella *et al.*, *Mol. Cell. Biol.* 10: 3386-3396 (1990), I-DirI (*see*, Johansen *et al.*, *Cell* 76: 725-734 (1994); Johansen, *Nucleic Acids Res.* 21: 4405 (1993), I-NjaI (*see*, Elde *et al.*, *Eur. J. Biochem.* 259: 281-288 (1999); De Jonckheere *et al.*, *J. Eukaryot. Microbiol.* 41: 457-463 (1994), I-NanI (*see*, Elde *et al.*, *S. Eur. J. Biochem.* 259: 281-288 (1999); De Jonckheere *et al.*, *J. Eukaryot. Microbiol.* 41: 457-463 (1994)), I-NitI (*see*, De Jonckheere *et al.*, *J. Eukaryot. Microbiol.* 41: 457-463 (1994); Elde *et al.*, *Eur. J. Biochem.* 259: 281-288 (1999), I-TevI (*see*, Chu *et al.*, *Cell* 45: 157-166 (1986), I-TevII (*see*, Tomaschewski *et al.*, *Nucleic Acids Res.* 15: 3632-3633 (1987), I-TevIII (*see*, Eddy *et al.*, *Genes Dev.* 5: 1032-1041 (1991), F-TevI (*see*, Fujisawa *et al.*, *Nucleic Acids Res.* 13: 7473-7481 (1985), F-TevII (*see*, Kadyrov *et al.*, *Dokl. Biochem.*

339: 145-147 (1994); Kaliman, *Nucleic Acids Res.* 18: 4277 (1990), F-CphI (*see, Zeng et al., Curr. Biol.* 19: 218-222 (2009)), PI-MgaI (*see, Saves et al., Nucleic Acids Res.* 29:4310-4318 (2001)), I-CsmI (*see, Colleaux et al., Mol. Gen. Genet.* 223:288-296 (1990)), I-CeuI (*see, Turmel et al., J. Mol. Biol.* 218: 293-311 (1991)) and PI-SceI (*see, Hirata et al., J. Biol. Chem.* 265: 6726-6733 (1990)).

[00140] In some embodiments of the methods described herein, a naturally occurring variant, and/or engineered derivative of a homing endonuclease is used. Methods for modifying the kinetics, cofactor interactions, expression, optimal conditions, and/or recognition site specificity, and screening for activity are known. *See, for example, Epinat, et al., (2003) Nucleic Acids Res* 31:2952-62; Chevalier, *et al., (2002) Mol Cell* 10:895-905; Gimble, *et al., (2003) Mol Biol* 334:993-1008; Seligman, *et al., (2002) Nucleic Acids Res* 30:3870-9; Sussman, *et al., (2004) J Mol Biol* 342:31-41; Rosen, *et al., (2006) Nucleic Acids Res* 34:4791-800; Chames, *et al., (2005) Nucleic Acids Res* 33:e178; Smith, *et al., (2006) Nucleic Acids Res* 34:e149; Gruen, *et al., (2002) Nucleic Acids Res* 30:e29; Chen and Zhao, (2005) *Nucleic Acids Res* 33:e154; WO2005105989; WO2003078619; WO2006097854; WO2006097853; WO2006097784; and WO2004031346. Useful homing endonucleases also include those described in WO04/067736; WO04/067753; WO06/097784; WO06/097853; WO06/097854; WO07/034262; WO07/049095; WO07/049156; WO07/057781; WO07/060495; WO08/152524; WO09/001159; WO09/095742; WO09/095793; WO10/001189; WO10/015899; and WO10/046786.

[00141] Any homing endonuclease can be used as a double-strand break inducing agent including, but not limited to: H-DreI, I-SceI, I-SceII, I-SceIII, I-SceIV, I-SceV, I-SceVI, I-SceVII, I-CeuI, I-CeuAIP, I-CreI, I-CrepsbIP, I-CrepsbIIP, I-CrepsbIIIP, I-CrepsbIVP, I-TliI, I-PpoI, Pi-PspI, F-SceI, F-SceII, F-SuvI, F-CphI, F-TevI, F-TevII, I-AmaI, I-AniI, I-ChuI, I-Cmoel, I-CpaI, I-CpaII, I-CsmI, I-CvuI, I-CvuAIP, I-DdiI, I-DdiII, I-Dirl, I-DmoI, I-HmuI, I-Hmull, I-HsNIP, I-LlaI, I-MsoI, I-NaaI, I-NanI, I-NclIP, I-NgrIP, I-NitI, I-Njal, I-Nsp236IP, I-PakI, I-PboIP, I-PcuIP, I-PcuAI, I-PcuVI, I-PgrIP, I-PobIP, I-PorI, I-PorIIP, I-PbpIP, I-SpBetaIP, I-ScaI, I-SexIP, I-SneIP, I-SpomI, I-SpomCP, I-SpomIP, I-SpomIIP, I-SquIP, I-Ssp68031, I-SthPhiJP, I-SthPhiST3P, I-SthPhiSTe3bP, I-TdeIP, I-TevI, I-TevII, I-TevIII, I-UarAP, I-UarHGPAIP, I-UarHGPA13P, I-VinIP, I-ZbiIP, PI-MgaI, PI-MtuI, PI-MtuHIIP, PI-MtuHIIIP, PI-PfuI, PI-PfuII, PI-PkoI, PI-PkoII, PI-Rma43812IP, PI-SpBetaIP, PI-SceI, PI-TfuI, PI-TfuII, PI-ThyI, PI-TliI, or PI-TliII, or any variant or derivative thereof.

[00142] In some embodiments, the endonuclease binds a native or endogenous recognition sequence. In other embodiments, the endonuclease is a modified endonuclease that binds a non-native or exogenous recognition sequence and does not bind a native or endogenous recognition sequence.

5.2.3.5 Genomic Target Sites

[00143] In the methods provided herein, a nuclease is introduced to the host cell that is capable of causing a double-strand break near or within a genomic target site, which greatly increases the frequency of homologous recombination at or near the cleavage site. In preferred embodiments, the recognition sequence for the nuclease is present in the host cell genome only at the target site, thereby minimizing any off-target genomic binding and cleavage by the nuclease.

[00144] In some embodiments, the genomic target site is endogenous to the host cell, such as a native locus. In some embodiments, the native genomic target site is selected according to the type of nuclease to be utilized in the methods of integration provided herein.

[00145] If the nuclease to be utilized is a CRISPR/Cas-derived RNA-guided endonuclease, optimal target sites may be selected in accordance with the requirements for target recognition of the particular CRISPR-Cas endonuclease being used. For example Cas9 target recognition occurs upon detection of complementarity between a “protospacer” sequence in the target DNA and the remaining spacer sequence in the crRNA. Cas9 cuts the DNA only if a correct protospacer-adjacent motif (PAM) is also present at the 3' end. Different Type II systems have differing PAM requirements. The *S. pyogenes* system requires an NGG sequence, where N can be any nucleotide. *S. thermophilus* Type II systems require NGGNG and NNAGAAW, respectively, while different *S. mutans* systems tolerate NGG or NAAR. Bioinformatic analyses have generated extensive databases of CRISPR loci in a variety of bacteria that may serve to identify new PAMs and expand the set of CRISPR-targetable sequences. See, e.g., Rho *et al.*, *PLoS Genet.* 8, e1002441 (2012); and D. T. Pride *et al.*, *Genome Res.* 21, 126 (2011). In *S. thermophilus*, Cas9 generates a blunt-ended double-stranded break 3bp upstream of the protospacer, a process mediated by two catalytic domains in the Cas9 protein: an HNH domain that cleaves the complementary strand of the DNA and a RuvC-like domain that cleaves the non-complementary strand.

[00146] If the nuclease to be utilized is a zinc finger nuclease, optimal target sites may be selected using a number of publicly available online resources. See, e.g., Reyon *et al.*, *BMC*

Genomics 12:83 (2011). For example, Oligomerized Pool Engineering (OPEN) is a highly robust and publicly available protocol for engineering zinc finger arrays with high specificity and *in vivo* functionality, and has been successfully used to generate ZFNs that function efficiently in plants, zebrafish, and human somatic and pluripotent stem cells. OPEN is a selection-based method in which a pre-constructed randomized pool of candidate ZFAs is screened to identify those with high affinity and specificity for a desired target sequence. ZFNGenome is a GBrowse-based tool for identifying and visualizing potential target sites for OPEN-generated ZFNs. ZFNGenome provides a compendium of potential ZFN target sites in sequenced and annotated genomes of model organisms. ZFNGenome currently includes a total of more than 11.6 million potential ZFN target sites, mapped within the fully sequenced genomes of seven model organisms; *S. cerevisiae*, *C. reinhardtii*, *A. thaliana*, *D. melanogaster*, *D. rerio*, *C. elegans*, and *H. sapiens*. Additional model organisms, including three plant species; *Glycine max* (soybean), *Oryza sativa* (rice), *Zea mays* (maize), and three animal species *Tribolium castaneum* (red flour beetle), *Mus musculus* (mouse), *Rattus norvegicus* (brown rat) will be added in the near future. ZFNGenome provides information about each potential ZFN target site, including its chromosomal location and position relative to transcription initiation site(s). Users can query ZFNGenome using several different criteria (*e.g.*, gene ID, transcript ID, target site sequence).

[00147] If the nuclease to be utilized is a TAL-effector nuclease, in some embodiments, optimal target sites may be selected in accordance with the methods described by Sanjana *et al.*, *Nature Protocols*, 7:171-192 (2012). In brief, TALENs function as dimers, and a pair of TALENs, referred to as the left and right TALENs, target sequences on opposite strands of DNA. TALENs are engineered as a fusion of the TALE DNA-binding domain and a monomeric FokI catalytic domain. To facilitate FokI dimerization, the left and right TALEN target sites are chosen with a spacing of approximately 14–20 bases. Therefore, for a pair of TALENs, each targeting 20-bp sequences, an optimal target site should have the form 5'-TN¹⁹N¹⁴⁻²⁰N¹⁹A-3', where the left TALEN targets 5'-TN¹⁹-3' and the right TALEN targets the antisense strand of 5'-N¹⁹A-3' (N = A, G, T or C).

[00148] In other embodiments of the methods provided herein, the genomic target site is exogenous to the host cell. For example, one or more genomic target sites can be engineered into the host cell genome using traditional methods, *e.g.*, gene targeting, prior to performing the methods of integration described herein. In some embodiments, multiple copies of the

same target sequence are engineered into the host cell genome at different loci, thereby facilitating simultaneous multiple integration events with the use of only a single nuclease that specifically recognizes the target sequence. In other embodiments, a plurality of different target sequences is engineered into the host cell genome at different loci. In some embodiments, the engineered target site comprises a target sequence that is not otherwise represented in the native genome of the host cell. For example, homing endonucleases target large recognition sites (12-40 bp) that are usually embedded in introns or inteins, and as such, their recognition sites are extremely rare, with none or only a few of these sites present in a mammalian-sized genome. Thus, in some embodiments, the exogenous genomic target site is a recognition sequence for a homing endonuclease. In some embodiments, the homing nuclease is selected from the group consisting of: H-DreI, I-SceI, I-SceII, I-SceIII, I-SceIV, I-SceV, I-SceVI, I-SceVII, I-CeuI, I-CeuAIIIP, I-CreI, I-CrepsbIP, I-CrepsbIIP, I-CrepsbIIIP, I-CrepsbIVP, I-TliI, I-PpoI, Pi-PspI, F-SceI, F-SceII, F-SuvI, F-CphI, F-TevI, F-TevII, I-Amal, I-AniI, I-ChuI, I-Cmoel, I-CpaI, I-CpaII, I-CsmI, I-CvuI, I-CvuAIP, I-DdiI, I-DdiII, I-DirI, I-DmoI, I-HmuI, I-HmuII, I-HsNIP, I-LlaI, I-MsoI, I-NaaI, I-NanI, I-NclIP, I-NgrIP, I-NitI, I-NjaI, I-Nsp236IP, I-PakI, I-PboIP, I-PcuIP, I-PcuAI, I-PcuVI, I-PgrIP, I-PobIP, I-PorI, I-PorIIP, I-PbpIP, I-SpBetaIP, I-ScaI, I-SexIP, I-SnelP, I-SpomI, I-SpomCP, I-SpomIP, I-SpomIIP, I-SquIP, I-Ssp68031, I-SthPhiJP, I-SthPhiST3P, I-SthPhiSTe3bP, I-TdeIP, I-TevI, I-TevII, I-TevIII, I-UarAP, I-UarHGPAIP, I-UarHGPA13P, I-VinIP, I-ZbiIP, PI-MgaI, PI-MtuI, PI-MtuHIP, PI-MtuHIIP, PI-PfuI, PI-PfuII, PI-PkoI, PI-PkoII, PI-Rma43812IP, PI-SpBetaIP, PI-SceI, PI-TfuI, PI-TfuII, PI-ThyI, PI-TliI, or PI-TliII, or any variant or derivative thereof. In particular embodiments, the exogenous genomic target site is the recognition sequence for I-SceI, VDE (PI-SceI), F-CphI, PI-MgaI or PI-MtuII, each of which are provided below.

Table 1: Recognition and cleavage sites for select homing endonucleases.

Nuclease	Recognition sequence
I-SceI	TAGGGATAACAGGGTAAT (SEQ ID NO:52)
VDE (PI-SceI)	TATGTCGGGTGCGGAGAAAGAGGTAATGAAA (SEQ ID NO:53)
F-CphI	GATGCACGAGCGCAACGCTCACAA (SEQ ID NO:54)
PI-MgaI	GCGTAGCTGCCCAGTATGAGTCAG (SEQ ID NO:55)
PI-MtuII	ACGTGCACTACGTAGAGGGTCGCACCGCACCGATCTACAA (SEQ ID NO:56)

5.2.3.6 Delivery

[00149] In some embodiments, the one or more nucleases useful for the methods described herein are provided, *e.g.*, delivered into the host cell as a purified protein. In other embodiments, the one or more nucleases are provided via polynucleotide(s) comprising a nucleic acid encoding the nuclease. In other embodiments, the one or more nucleases are introduced into the host cell as purified RNA which can be directly translated in the host cell nucleus.

[00150] In certain embodiments, an integration polynucleotide, a polynucleotide encoding a nuclease, or a purified nuclease protein as described above, or any combination thereof, may be introduced into a host cell using any conventional technique to introduce exogenous protein and/or nucleic acids into a cell known in the art. Such methods include, but are not limited to, direct uptake of the molecule by a cell from solution, or facilitated uptake through lipofection using, *e.g.*, liposomes or immunoliposomes; particle-mediated transfection; *etc.* See, *e.g.*, U.S. Patent No. 5,272,065; Goeddel *et al.*, eds, 1990, *Methods in Enzymology*, vol. 185, Academic Press, Inc., CA; Krieger, 1990, *Gene Transfer and Expression -- A Laboratory Manual*, Stockton Press, NY; Sambrook *et al.*, 1989, *Molecular Cloning -- A Laboratory Manual*, Cold Spring Harbor Laboratory, NY; and Ausubel *et al.*, eds., Current Edition, *Current Protocols in Molecular Biology*, Greene Publishing Associates and Wiley Interscience, NY. Particular methods for transforming cells are well known in the art. See Hinnen *et al.*, *Proc. Natl. Acad. Sci. USA* 75:1292-3 (1978); Cregg *et al.*, *Mol. Cell. Biol.* 5:3376-3385 (1985). Exemplary techniques include but are not limited to, spheroplasting, electroporation, PEG 1000 mediated transformation, and lithium acetate or lithium chloride mediated transformation.

[00151] In some embodiments, biolistics are utilized to introduce an integration polynucleotide, a polynucleotide encoding a nuclease, a purified nuclease protein, or any combination thereof into the host cell, in particular, host cells that are otherwise difficult to transform/transfect using conventional techniques, such as plants. Biolistics work by binding the transformation reaction to microscopic gold particles, and then propelling the particles using compressed gas at the target cells.

[00152] In some embodiments, the polynucleotide comprising nucleic acid encoding the nuclease is an expression vector that allows for the expression of a nuclease within a host cell. Suitable expression vectors include but are not limited to those known for use in expressing

genes in *Escherichia coli*, yeast, or mammalian cells. Examples of *Escherichia coli* expression vectors include but are not limited to pSCM525, pDIC73, pSCM351, and pSCM353.

Examples of yeast expression vectors include but are not limited to pPEX7 and pPEX408.

Other examples of suitable expression vectors include the yeast-*Escherichia coli* pRS series of shuttle vectors comprising CEN.ARS sequences and yeast selectable markers; and 2 μ plasmids. In some embodiments, a polynucleotide encoding a nuclease can be modified to substitute codons having a higher frequency of usage in the host cell, as compared to the naturally occurring polynucleotide sequence. For example the polynucleotide encoding the nuclease can be modified to substitute codons having a higher frequency of usage in *S. cerevisiae*, as compared to the naturally occurring polynucleotide sequence.

[00153] In some embodiments where the nuclease functions as a heterodimer requiring the separate expression of each monomer, as is the case for zinc finger nucleases and TAL-effector nucleases, each monomer of the heterodimer may be expressed from the same expression plasmid, or from different plasmids. In embodiments where multiple nucleases are introduced to the cell to effect double-strand breaks at different target sites, the nucleases may be encoded on a single plasmid or on separate plasmids.

[00154] In certain embodiments, the nuclease expression vector further comprises a selectable marker that allows for selection of host cells comprising the expression vector. Such selection can be helpful to retain the vector in the host cell for a period of time necessary for expression of sufficient amounts of nuclease to occur, for example, for a period of 12, 24, 36, 48, 60, 72, 84, 96, or more than 96 hours, after which the host cells may be grown under conditions under which the expression vector is no longer retained. In certain embodiments, the selectable marker is selected from the group consisting of: URA3, hygromycin B phosphotransferase, aminoglycoside phosphotransferase, zeocin resistance, and phosphinothricin N-acetyltransferase. In some embodiments, the nuclease expression vector may comprise a counter-selectable marker that allows for selection of host cells that do not contain the expression vector subsequent to integration of the one or more donor nucleic acid molecules. The nuclease expression vector used may also be a transient vector that has no selection marker, or is one that is not selected for. In particular embodiments, the progeny of a host cell comprising a transient nuclease expression vector loses the vector over time.

[00155] In certain embodiments, the expression vector further comprises a transcription termination sequence and a promoter operatively linked to the nucleotide sequence encoding the nuclease. In some embodiments, the promoter is a constitutive promoter. In some

embodiments, the promoter is an inducible promoter. Illustrative examples of promoters suitable for use in yeast cells include, but are not limited to the promoter of the TEF1 gene of *K. lactis*, the promoter of the PGK1 gene of *Saccharomyces cerevisiae*, the promoter of the TDH3 gene of *Saccharomyces cerevisiae*, repressible promoters, *e.g.*, the promoter of the CTR3 gene of *Saccharomyces cerevisiae*, and inducible promoters, *e.g.*, galactose inducible promoters of *Saccharomyces cerevisiae* (*e.g.*, promoters of the GAL1, GAL7, and GAL10 genes).

[00156] In some embodiments, an additional nucleotide sequence comprising a nuclear localization sequence (NLS) is linked to the 5' of the nucleotide sequence encoding the nuclease. The NLS can facilitate nuclear localization of larger nucleases (>25 kD). In some embodiments, the nuclear localization sequence is an SV40 nuclear localization sequence. In some embodiments, the nuclear localization sequence is a yeast nuclear localization sequence.

[00157] A nuclease expression vector can be made by any technique apparent to one skilled in the art. In certain embodiments, the vector is made using polymerase chain reaction (PCR) and molecular cloning techniques well known in the art. *See, e.g., PCR Technology: Principles and Applications for DNA Amplification*, ed. HA Erlich, Stockton Press, New York, N.Y. (1989); Sambrook *et al.*, 2001, *Molecular Cloning – A Laboratory Manual*, 3rd edition, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

5.3 Host Cells

[00158] In another aspect, provided herein is a modified host cell generated by any of the methods of genomically integrating one or more exogenous nucleic acids described herein. Suitable host cells include any cell in which integration of a nucleic acid or “donor DNA” of interest into a chromosomal or episomal locus is desired. In some embodiments, the cell is a cell of an organism having the ability to perform homologous recombination. Although several of the illustrative embodiments are demonstrated in yeast (*S. cerevisiae*), it is believed that the methods of genomic modification provided herein can be practiced on all biological organisms having a functional recombination system, even where the recombination system is not as proficient as in yeast. Other cells or cell types that have a functional homologous recombination systems include bacteria such as *Bacillus subtilis* and *E. coli* (which is RecE RecT recombination proficient; Muylers *et al.*, *EMBO rep.* 1: 239-243, 2000); protozoa (*e.g.*, *Plasmodium*, *Toxoplasma*); other yeast (*e.g.*, *Schizosaccharomyces pombe*); filamentous fungi (*e.g.*, *Ashbya gossypii*); plants, for instance the moss *Physcomitrella patens* (Schaefer and

Zryd, *Plant J.* 11: 1195-1206, 1997); and animal cells, such as mammalian cells and chicken DT40 cells (Dieken *et al.*, *Nat. Genet.* 12:174-182, 1996).

[00159] In some embodiments, the host cell is a prokaryotic cell. In some embodiments, the host cell is a eukaryotic cell. In some embodiments, the host cell is selected from the group consisting of a fungal cell, a bacterial cell, a plant cell, an insect cell, an avian cell, a fish cell and a mammalian cell. In some embodiments, the mammalian cell selected from the group consisting a rodent cell, a primate cell and a human cell. In some embodiments, the cell is a fungal cell (for instance, a yeast cell), a bacteria cell, a plant cell, or an animal cell (for instance, a chicken cell). In some embodiments, the host cell is a Chinese hamster ovary (CHO) cell, a COS-7 cell, a mouse fibroblast cell, a mouse embryonic carcinoma cell, or a mouse embryonic stem cell. In some embodiments, the host cell is an insect cell. In some embodiments, the host cell is a S2 cell, a Schneider cell, a S12 cell, a 5B1-4 cell, a Tn5 cell, or a Sf9 cell. In some embodiments, the host cell is a unicellular eukaryotic organism cell.

[00160] In particular embodiments, the host cell is a yeast cell. Useful yeast host cells include yeast cells that have been deposited with microorganism depositories (*e.g.* IFO, ATCC, *etc.*) and belong to the genera *Aciculoconidium*, *Ambrosiozyma*, *Arthroascus*, *Arxiozyma*, *Ashbya*, *Babjevia*, *Bensingtonia*, *Botryoascus*, *Botryozyma*, *Brettanomyces*, *Bullera*, *Bulleromyces*, *Candida*, *Citeromyces*, *Clavispora*, *Cryptococcus*, *Cystofilobasidium*, *Debaryomyces*, *Dekkara*, *Dipodascopsis*, *Dipodascus*, *Eeniella*, *Endomycopsella*, *Eremascus*, *Eremothecium*, *Erythrobasidium*, *Fellomyces*, *Filobasidium*, *Galactomyces*, *Geotrichum*, *Guilliermondella*, *Hanseniaspora*, *Hansenula*, *Hasegawaea*, *Holtermannia*, *Hormoascus*, *Hyphopichia*, *Issatchenkia*, *Kloeckera*, *Kloeckeraspora*, *Kluyveromyces*, *Kondoa*, *Kuraishia*, *Kurtzmanomyces*, *Leucosporidium*, *Lipomyces*, *Lodderomyces*, *Malassezia*, *Metschnikowia*, *Mrakia*, *Myxozyma*, *Nadsonia*, *Nakazawaea*, *Nematospora*, *Ogataea*, *Oosporidium*, *Pachysolen*, *Phachytichospora*, *Phaffia*, *Pichia*, *Rhodosporidium*, *Rhodotorula*, *Saccharomyces*, *Saccharomycodes*, *Saccharomycopsis*, *Saitoella*, *Sakaguchia*, *Saturnospora*, *Schizoblastosporion*, *Schizosaccharomyces*, *Schwanniomyces*, *Sporidiobolus*, *Sporobolomyces*, *Sporopachydermia*, *Stephanoascus*, *Sterigmatomyces*, *Sterigmatosporidium*, *Symbiotaphrina*, *Sympodiomyces*, *Sympodiomycopsis*, *Torulaspora*, *Trichosporiella*, *Trichosporon*, *Trigonopsis*, *Tsuchiyaea*, *Udeniomyces*, *Waltomyces*, *Wickerhamia*, *Wickerhamiella*, *Williopsis*, *Yamadazyma*, *Yarrowia*, *Zygoascus*, *Zygosaccharomyces*, *Zygowilliopsis*, and *Zygozyma*, among others.

[00161] In some embodiments, the yeast host cell is a *Saccharomyces cerevisiae* cell, a *Pichia pastoris* cell, a *Schizosaccharomyces pombe* cell, a *Dekkera bruxellensis* cell, a *Kluyveromyces lactis* cell, a *Arxula adeninivorans* cell, or a *Hansenula polymorpha* (now known as *Pichia angusta*) cell. In a particular embodiment, the yeast host cell is a *Saccharomyces cerevisiae* cell. In some embodiments, the yeast host cell is a *Saccharomyces fragilis* cell or a *Kluyveromyces lactis* (previously called *Saccharomyces lactis*) cell. In some embodiments, the yeast host cell is a cell belonging to the genus *Candida*, such as *Candida lipolytica*, *Candida guilliermondii*, *Candida krusei*, *Candida pseudotropicalis*, or *Candida utilis*. In another particular embodiment, the yeast host cell is a *Kluyveromyces marxianus* cell.

[00162] In particular embodiments, the yeast host cell is a *Saccharomyces cerevisiae* cell selected from the group consisting of a Baker's yeast cell, a CBS 7959 cell, a CBS 7960 cell, a CBS 7961 cell, a CBS 7962 cell, a CBS 7963 cell, a CBS 7964 cell, a IZ-1904 cell, a TA cell, a BG-1 cell, a CR-1 cell, a SA-1 cell, a M-26 cell, a Y-904 cell, a PE-2 cell, a PE-5 cell, a VR-1 cell, a BR-1 cell, a BR-2 cell, a ME-2 cell, a VR-2 cell, a MA-3 cell, a MA-4 cell, a CAT-1 cell, a CB-1 cell, a NR-1 cell, a BT-1 cell, and a AL-1 cell. In some embodiments, the host cell is a *Saccharomyces cerevisiae* cell selected from the group consisting of a PE-2 cell, a CAT-1 cell, a VR-1 cell, a BG-1 cell, a CR-1 cell, and a SA-1 cell. In a particular embodiment, the *Saccharomyces cerevisiae* host cell is a PE-2 cell. In another particular embodiment, the *Saccharomyces cerevisiae* host cell is a CAT-1 cell. In another particular embodiment, the *Saccharomyces cerevisiae* host cell is a BG-1 cell.

[00163] In some embodiments, the yeast host cell is a cell that is suitable for industrial fermentation, *e.g.*, bioethanol fermentation. In particular embodiments, the cell is conditioned to subsist under high solvent concentration, high temperature, expanded substrate utilization, nutrient limitation, osmotic stress due, acidity, sulfite and bacterial contamination, or combinations thereof, which are recognized stress conditions of the industrial fermentation environment.

5.4 Applications

5.4.1. Gene and Cell Therapy

[00164] The methods and compositions described herein provide advantages in therapeutic applications which seek to correct genetic defects *e.g.*, *ex vivo* in a cell population derived from a subject. For example, Schwank *et al.* (Cell Stem Cell 13:653-658 (2013)) recently reported utilization of the CRISPR/Cas9 genome editing system to correct the CFTR locus by

homologous recombination in cultured intestinal stem cells of human CF patients. The corrected allele was expressed and fully functional as measured in clonally expanded organoids, and thus this report provides proof of concept for gene correction by homologous recombination in primary adult stem cells derived from patients with a single-gene hereditary defect. However, correction of the CFTR locus in the cultured stem cells required genomic integration of a puromycin resistance cassette along with the donor DNA, followed by selection in puromycin. It has been reported that integration of the neomycin resistance gene into human cell genomes, followed by extended culturing times in G418, causes changes to the cell's characteristics, while expression of enhanced green fluorescent protein (EGFP) and other fluorescent proteins has been reported to cause immunogenicity and toxicity. *See, e.g.,* Barese *et al.*, *Human Gene Therapy* 22:659-668 (2011); Morris *et al.*, *Blood* 103:492-499 (2004); and Hanazono *et al.*, *Human Gene Therapy* 8:1313-1319 (1997). Thus, the methods and compositions provided herein can be utilized to perform gene correction by homologous recombination in primary adult stem cells without the need for integration of a selectable marker.

[00165] In another report of HR-mediated correction of a genetic disease, Wu *et al.* (*Cell Stem Cell* 13:659-662 (2013)) demonstrated that mice with a dominant mutation in the *Crygc* gene that causes cataracts could be rescued by coinjection into zygotes of Cas9 mRNA and a single-guide RNA (sgRNA) targeting the mutant allele. Correction occurred via homology-directed repair (HDR) based on an exogenously supplied oligonucleotide or the endogenous WT allele, and the resulting mice were fertile and able to transmit the corrected allele to their progeny. However, the rate of HDR-mediated repair was much lower than the incidence of repair or non-repair by NHEJ (*see* Wu *et al.* at Table 1). Thus, the methods and compositions provided herein can be utilized to improve the efficiency of gene correction by providing a useful selection mechanism that selects for HR-mediated gene modifications.

5.4.2. Methods for Metabolic Pathway Engineering

[00166] The methods and compositions described herein provide particular advantages for constructing recombinant organisms comprising optimized biosynthetic pathways, for example, towards the conversion of biomass into biofuels, pharmaceuticals or biomaterials. Functional non-native biological pathways have been successfully constructed in microbial hosts for the production of precursors to the antimalarial drug artemisinin (*see, e.g.,* Martin *et al.*, *Nat Biotechnol* 21:796-802 (2003)); fatty acid derivatives fuels and chemicals (*e.g.,* fatty esters, fatty alcohols and waxes; *see, e.g.,* Steen *et al.*, *Nature* 463:559-562 (2010)); methyl halide-

derived fuels and chemicals (*see, e.g., Bayer et al., J Am Chem Soc* 131:6508-6515 (2009); polyketide synthases that make cholesterol lowering drugs (*see, e.g., Ma et al., Science* 326:589-592 (2009); and polyketides (*see, e.g., Kodumal, Proc Natl Acad Sci USA* 101:15573-15578 (2004).

[00167] Traditionally, metabolic engineering, and in particular, the construction of biosynthetic pathways, has proceeded in a one-at-a-time serial fashion whereby pathway components have been introduced, *i.e.*, integrated into the host cell genome at a single loci at a time. The methods of integration provided herein can be utilized to reduce the time typically required to engineer a host cell, for example, a microbial cell, to comprise one or more heterologous nucleotide sequences encoding enzymes of a new metabolic pathway, *i.e.*, a metabolic pathway that produces a metabolite that is not endogenously produced by the host cell. In other particular embodiments, the methods of integration provided herein can be used to efficiently engineer a host cell to comprise one or more heterologous nucleotide sequences encoding enzymes of a metabolic pathway that is endogenous to the host cell, *i.e.*, a metabolic pathway that produces a metabolite that is endogenously produced by the host cell. In one example, a design strategy may seek to replace three native genes of a host cell with a complementary exogenous pathway. Modifying these three endogenous loci using the current state of the art requires three separate transformations. By contrast, the methods of simultaneous multiple integration provided herein enables all three integrations to be performed in a single transformation, thus reducing the rounds of engineering needed by three-fold. Moreover, the methods enable the porting of DNA assemblies, comprising optimized pathway components integrated at multiple sites in one host cell chassis, to analogous sites in a second host cell chassis. By reducing the number of rounds needed to engineer a desired genotype, the pace of construction of metabolic pathways is substantially increased.

5.4.2.1 Isoprenoid Pathway Engineering

[00168] In some embodiments, the methods provided herein can be utilized to simultaneously introduce or replace one or more components of a biosynthetic pathway to modify the product profile of an engineered host cell. In some embodiments, the biosynthetic pathway is the isoprenoid pathway.

[00169] Terpenes are a large class of hydrocarbons that are produced in many organisms. When terpenes are chemically modified (*e.g.*, via oxidation or rearrangement of the carbon skeleton) the resulting compounds are generally referred to as terpenoids, which are also

known as isoprenoids. Isoprenoids play many important biological roles, for example, as quinones in electron transport chains, as components of membranes, in subcellular targeting and regulation via protein prenylation, as photosynthetic pigments including carotenoids, chlorophyll, as hormones and cofactors, and as plant defense compounds with various monoterpenes, sesquiterpenes, and diterpenes. They are industrially useful as antibiotics, hormones, anticancer drugs, insecticides, and chemicals.

[00170] Terpenes are derived by linking units of isoprene (C_5H_8), and are classified by the number of isoprene units present. Hemiterpenes consist of a single isoprene unit. Isoprene itself is considered the only hemiterpene. Monoterpenes are made of two isoprene units, and have the molecular formula $C_{10}H_{16}$. Examples of monoterpenes are geraniol, limonene, and terpineol. Sesquiterpenes are composed of three isoprene units, and have the molecular formula $C_{15}H_{24}$. Examples of sesquiterpenes are farnesenes and farnesol. Diterpenes are made of four isoprene units, and have the molecular formula $C_{20}H_{32}$. Examples of diterpenes are cafestol, kahweol, cembrene, and taxadiene. Sesterterpenes are made of five isoprene units, and have the molecular formula $C_{25}H_{40}$. An example of a sesterterpenes is geranylfarnesol. Triterpenes consist of six isoprene units, and have the molecular formula $C_{30}H_{48}$. Tetraterpenes contain eight isoprene units, and have the molecular formula $C_{40}H_{64}$. Biologically important tetraterpenes include the acyclic lycopene, the monocyclic gamma-carotene, and the bicyclic alpha- and beta-carotenes. Polyterpenes consist of long chains of many isoprene units. Natural rubber consists of polyisoprene in which the double bonds are cis.

[00171] Terpenes are biosynthesized through condensations of isopentenyl pyrophosphate (isopentenyl diphosphate or IPP) and its isomer dimethylallyl pyrophosphate (dimethylallyl diphosphate or DMAPP). Two pathways are known to generate IPP and DMAPP, namely the mevalonate-dependent (MEV) pathway of eukaryotes (FIG. 3), and the mevalonate-independent or deoxyxylulose-5-phosphate (DXP) pathway of prokaryotes. Plants use both the MEV pathway and the DXP pathway. IPP and DMAPP in turn are condensed to polyprenyl diphosphates (*e.g.*, geranyl diphosphate or GPP, farnesyl diphosphate or FPP, and geranylgeranyl diphosphate or GGPP) through the action of prenyl diphosphate synthases (*e.g.*, GPP synthase, FPP synthase, and GGPP synthase, respectively). The polyprenyl diphosphate intermediates are converted to more complex isoprenoid structures by terpene synthases.

[00172] Terpene synthases are organized into large gene families that form multiple products. Examples of terpene synthases include monoterpene synthases, which convert GPP

into monoterpenes; diterpene synthases, which convert GGPP into diterpenes; and sesquiterpene synthases, which convert FPP into sesquiterpenes. An example of a sesquiterpene synthase is farnesene synthase, which converts FPP to farnesene. Terpene synthases are important in the regulation of pathway flux to an isoprenoid because they operate at metabolic branch points and dictate the type of isoprenoid produced by the cell. Moreover, the terpene synthases hold the key to high yield production of such terpenes. As such, one strategy to improve pathway flux in hosts engineered for heterologous isoprenoid production is to introduce multiple copies of nucleic acids encoding terpene synthases. For example, in engineered microbes comprising the MEV pathway where the production of sesquiterpenes such as farnesene is desired, a sesquiterpene synthase, *e.g.*, a farnesene synthase is utilized as the terminal enzyme of the pathway, and multiple copies of farnesene synthase genes may be introduced into the host cell towards the generation of a strain optimized for farnesene production.

[00173] Because the biosynthesis of any isoprenoid relies on the same pathway components upstream of the prenyl diphosphate synthase and terpene synthase, these pathway components, once engineered into a host “platform” strain, can be utilized towards the production of any sesquiterpene, and the identity of the sesquiterpene can be dictated by the particular sesquiterpene synthase introduced into the host cell. Moreover, where production of terpenes having different isoprene units is desired, for example a monoterpene instead of a sesquiterpene, both the prenyl diphosphate synthase and the terpene synthase can be replaced to produce the different terpene while still utilizing the upstream components of the pathway.

[00174] Accordingly, the methods and compositions provided herein can be utilized to efficiently modify a host cell comprising an isoprenoid producing pathway, *e.g.*, the MEV pathway to produce a desired isoprenoid. In some embodiments, the host cell comprises the MEV pathway, and the methods of simultaneous multiple integration provided herein can be utilized to simultaneously introduce multiple copies of a prenyl diphosphate synthase and /or a terpene synthase to define the terpene product profile of the host cell. In some embodiments, the prenyl diphosphate synthase is GPP synthase and the terpene synthase is a monoterpene synthase. In some embodiments, the prenyl diphosphate synthase is FPP synthase and the terpene synthase is a sesquiterpene synthase. In some embodiments, the prenyl diphosphate synthase is GGPP synthase and the terpene synthase is a diterpene synthase. In other embodiments, the host cell comprises the MEV pathway and a prenyl diphosphate synthase and/or a terpene synthase for the production of a first type of terpene, for example, farnesene,

and the methods of simultaneous multiple integration provided herein can be utilized to simultaneously replace one or more copies of the prenyl diphosphate synthase and/or a terpene synthase to produce a second type of terpene, for example, amorphadiene. These embodiments are exemplified in Examples 3 and 4 below. The methods provided herein can be similarly utilized towards the construction and/or modification of any biosynthetic pathway which utilizes multiple copies of pathway components, and are particularly useful for engineering host cells whose product profile can be readily modified with the addition or exchange of multiple copies of a single pathway component.

6. EXAMPLES

6.1 Example 1: A comparison of multiple modes of gRNA delivery for simultaneous deletion, via integration of deletion constructs, of 3 genes using CRISPR

[00175] This Example provides results which demonstrate the use of CRISPR for simultaneous deletion of the RHR2, HO and ADH5 ORFs (with integration of a short linker sequence) in *S. cerevisiae*. In brief, chimeric gRNAs were generated targeting unique sequences contained in the ORF of RHR2, HO and ADH5, and were transformed in various configurations into host cells expressing the Cas9 protein from the type II bacterial CRISPR system of *Streptococcus pyogenes*. Transformed colonies were screened by colony PCR (cPCR) for the replacement of one, two or three ORFs with a short linker sequence.

6.1.1. Constitutive Expression of Cas9p

[00176] A wild-type *Saccharomyces cerevisiae* strain, (CEN.PK2, Mat a, *ura3*-, TRP1+, *leu2*-, MAL2-8C, SUC2) was used as a host for the constitutive expression of Cas9p from *Streptococcus pyogenes*. Genomic integration of a construct containing a yeast codon optimized coding sequence for Cas9 under the control of the medium strength FBA1 promoter (SEQ ID NO:3) was targeted to the GRE3 locus.

6.1.2. Selection of CRISPR Target Sites in RHR2, HO and ADH5 ORFs

[00177] Candidate CRISPR targets inside the targeted ORFs were identified based on the presence of a PAM sequence N₍₁₉₎NGG. The NGG sequence is referred to as a PAM sequence and the 8 base pairs of DNA preceding the PAM sequence are especially important for enforcing specificity (Fu *et al.*, *Nat Biotechnol* 31(9):822-826 (2013); Ran *et al.*, *Nat Protoc* 8(11):2281-2308 (2013)). Candidate sites were then ranked based on uniqueness of the target sequence in the genome, and the site with the lowest similarity to other sites in the genome was chosen to minimize the risk of off target cutting. Target sites are shown in **Table 2**.

[00178] **Table 2: CRISPR Target sites**

ORF	Target Sequence (NGG omitted)
RHR2	ACCTCTGGTACCCGTGACA (SEQ ID NO:4)
HO	CCGGCTTGATCGACTCAGA (SEQ ID NO:5)
ADH5	GGGTCATTGGTATCGATGG (SEQ ID NO:6)

6.1.3. gRNA Delivery Modes

[00179] Cas9p is targeted to cut sites by association with a generic structural RNA and a specific targeting RNA. The now standard “chimeric” configuration was adopted in this work, in which the targeting and structural RNAs are fused to create a single guide RNA, or gRNA (Ran, Hsu et al. 2013). Expression of the gRNA construct(s) was driven by the *SNR52* polymerase III promoter, with a *SUP4* terminator (DiCarlo, Norville et al. 2013). Sequences for gRNA constructs targeting the RHR2, HO and ADH5 locus, respectively, are provided herein as SEQ ID NOs: 7, 8 and 9. Several modes of gRNA delivery were used, as described in section 5.2.1 above and depicted in **FIG. 7**. Expression of the gRNA cassette from a pRS4XX-series 2μ vectors (Sikorski and Hieter 1989) was achieved either by: 1) standard cloning methods to generate finished circular plasmids (**FIG. 7.1 and 7.2**) prior to transformation into a Cas9-expressing yeast strain, or 2) by assembling the gRNA cassette into a circularized plasmid by gap repair *in vivo*, by transforming the cassette directly into a Cas9-expressing yeast strain, along with a linearized vector backbone (Orr-Weaver, Szostak et al. 1983) (**FIG. 7.4**). Regions of homology (~500 bp) between the termini of the gRNA cassette and the linear vector backbone (SEQ ID NO:10) facilitate assembly of a circular gRNA plasmid *in vivo*. Alternatively, 3) the gRNA was expressed directly from the linear cassette, co-transformed with a closed plasmid bearing a NatA (Nourseothricin acetyltransferase from *Streptomyces noursei*) selectable marker (SEQ ID NO:11) to select for transformed cells (**FIG. 7.3**). Finally, in a variation of the third method, 4) the linear cassette was co-transformed with vector linearized by PCR inside of the NatA marker (SEQ ID NO:12) such that a central coding sequence for the NatA marker is missing; a complementary overlapping NatA ORF fragment (SEQ ID NO:13) that can recombine via gap repair to re-create the closed plasmid bearing a complete NatA expression cassette was also co-transformed (**FIG. 9B**; discussed in Example 2 below).

[00180] To create circular gRNA plasmids (delivery mode 1), annealed oligonucleotides containing the CRISPR seed sequence and 20 bp of upstream/downstream homology to the cassette were gap-repaired into a linearized backbone in *E. coli* (Mandecki 1986), correct

clones were identified, and the finished plasmid transformed into a host strain. To prepare full length linear gRNA cassettes with ~500 bp flanking homology to the linearized vector (delivery modes 2, 3 and 4), a PCR assembly method was employed. Using a generic gRNA cassette as template, half cassettes were amplified using primers to create a central overlap of 22 base pairs comprising the unique CRISPR seed sequence. The two half cassettes were then assembled in a second PCR reaction to generate a full length gRNA expression cassette. 10 µl of unpurified PCR assembly (typically 20-60 ng/µl concentration as determined by comparison to DNA marker ladder) and 150 ng of linearized 2µ vector were used per transformation.

6.1.4. Generation of Linear Donor DNA

[00181] Linear donor DNAs comprise 500 bp upstream and downstream homology regions targeting each ORF, flanking a central linker (CGCTCGTCCAACGCCGGCGGACCT), and were generated by the methods of polynucleotide assembly described in U.S. Patent No. 8,221,982. Donor DNA sequences for integration into the RHR2, HO and ADH5 locus, respectively, are provided herein as SEQ ID NOs: 14, 15 and 16.

6.1.5. Simultaneous deletion of ORF and integration of a short linker sequence using CRISPR

[00182] Donor DNA (~1 µg) and the appropriate gRNA reagents were co-transformed into each Cas9 expressing strain using optimized LiAc methods (Gietz and Woods 2002) with the addition of a 30 minute incubation of cells at 30 degrees C prior to heat shock at 42 degrees C. Cells were recovered overnight in non-selective YPD (Yeast extract peptone dextrose) media before plating to selective, antibiotic-containing media (nourseothricin, 100 µg/ml) to maintain the gRNA or marker plasmid. Marker-less integrations were scored as positive if colony PCR (cPCR) using primers binding upstream of the 5' integration flank and to the integrated linker sequence (Table 3) produced the correct amplicon, a result indicative of a targeted integration event.

[00183] **Table 3. Primer sequences for cPCR verification of linker integration at RHR2, HO and ADH5 loci**

Primer Name	Description	Sequence	SEQ ID NO
RHR2-US-F	RHR2 locus US FOR	gggtgcgaagtaccaccacgtttcttttcatctct	SEQ ID NO: 17
HO-US-F	HO locus US FOR	acgtgtgtgtctcatggaaattgatgcagttgaagaca	SEQ ID NO: 18
ADH5-US-F	ADH5 locus US FOR	ggcgttatatccaaacatttcagacagaagatt	SEQ ID NO: 19

R5	Linker REV	AGGTCCGCCGGCGTTGGACGAGC G	SEQ ID NO:20
----	------------	------------------------------	--------------

6.1.6. Results and Discussion

[00184] Gap Repair Delivery Modes: Four modes of gRNA delivery were assessed for efficiency of simultaneous deletion of the RHR2, HO and AHD5 open reading frames, with integration of a short linker sequence. Colony PCR results for assessing triple integration are shown in **FIG. 8** and rates are summarized in **Table 4**.

[00185] **Table 4.** Rates of triple integrations with varying delivery of gRNA constructs

Description	Triple rate
Triple selection, plasmid gRNAs	0.91
Single selection, plasmid gRNAs	0
marker plasmid, linear gRNAs	0
gap repair linear gRNAs	0.64

[00186] In the first mode of gRNA delivery, each of three gRNAs (targeting RHR2, HO and ADH5, respectively) was supplied on a plasmid bearing unique markers (NatA, URA3 cassette and HIS3 cassette; see **FIG. 7.1**), and cells were transformed with all three plasmids and triply selected for the expression of each marker. Very high efficiencies (91%) of triple deletion (via integration) were observed (**FIG. 8, panel 1**). These high frequencies likely result from sustained expression of all three gRNAs by triple selection for the plasmids bearing their cassettes. By contrast, the second mode, where the gRNAs were supplied on three plasmids bearing the same marker (NatA; see **FIG. 7.2**), failed to generate any triple deletions (**FIG. 8, panel 2**). Instead, single deletions dominated, which is consistent with the selection requirement to maintain only one of the three plasmids. In the third mode, the gRNAs were supplied as linear cassettes, with a NatA marked plasmid included to select for transformed cells (see **FIG. 7.3**). No triple deletions were observed, and very low rates of any deletion event were observed (**FIG. 8, panel 3**). This mode of delivery is expected to result in transient expression of the gRNA constructs, and this seems to be inferior to sustained expression. The fourth delivery mode that was explored requires gap repair of the three gRNA cassettes into a linearized vector bearing the NatA marker (see **FIG. 7.4**). We observed 64% of colonies were triply deleted (**FIG. 8, panel 4**). This is a surprising result as this mode of delivery does not enforce sustained expression of all three gRNAs as the first mode does. Indeed, results from the second mode indicate clearly that co-transformation of three like-marked gRNA plasmids is ineffective, and the results from the third mode indicate that transient expression of gRNAs from linear cassettes is also non-functional. Thus, there is an unexpected advantageous benefit

towards CRISPR/Cas-9 mediated genomic integration events associated with gap repair as a mode of delivery for gRNA cassettes.

6.2 Example 2: Selection of HR Competent Cells Via Gap Repair

[00187] This example demonstrates the benefit of gap repair, independent of the benefit of selecting for gRNA expression, for improving the efficiency of a nuclease-mediated integration event.

[00188] One mechanism by which gap repair might improve the recovery of clones engineered by CRISPR (or any site-specific nuclease) is by enforcing an additional selection for cells that are proficient for homologous recombination (HR). HR proficiency can vary widely in an asynchronous cell population (*see e.g.*, Brnzei and Foiani, *Nat Rev Mol Cell Bio* 9(4):297-308 (2008)), and selection for cells that can accomplish gap repair of a plasmid bearing a selectable marker may select a population that is particularly HR proficient. This could explain the surprising success of the fourth mode of gRNA delivery (**FIGS. 7.4 and 8.4**), discussed in Example 1. To uncouple the effects of gap repair as a selection mechanism from that of sustained expression of at least one of the gRNAs, we assessed rates of single deletion of the RHR2, HO and ADH5 locus, respectively, by co-transformation of the appropriate donor DNA and linear gRNA cassette (described in Example 1, above) and one of two marker vectors. The first marker vector was closed, i.e. no gap repair is required for the expression of a NatA marker in transformed cells (**FIG. 9A**). The second marker was linearized such that a central portion of the NatA marker was missing, but could be complemented by gap repair of an overlapping fragment to produce a closed, functional marker vector bearing a complete NatA expression cassette (**FIG. 9B**). In both cases, expression of the gRNA is transient only. Over three independent experiments, we noted improvements up to 8-fold (~3-5-fold range average) in the rate of single locus deletion (integration) when the marker plasmid required gap repair (**FIG. 9**, with rates summarized in **FIGS. 10-12**).

[00189] These results support the hypothesis that gap repair can act as an additional selection for cells proficient in HR, and thus most capable of successful nuclease-assisted engineering, and in particular, targeted integrations of donor DNA. We note that *S. cerevisiae* are especially adept at HR, and in cells that favor NHEJ and other imperfect repair methods (e.g. mammalian cells), we propose that gap repair may be particularly effective at increasing the recovery of cells bearing one or more targeted integrations.

6.3 Example 3: Enhanced Selection of HR-Competent Cells Via Multi-Piece Gap Repair

[00190] This example demonstrates that by increasing the complexity of gap repair of a marker vector by further fragmenting the vector, it is possible to further increase the efficiency of nuclease-assisted engineering.

[00191] Cas9-expressing haploid yeast cells (*S. cerevisiae*) were transformed with donor DNAs for simultaneous, marker-less deletion of Gal80, HO and ADH5 open reading frames, gRNA constructs targeting each locus, and linearized vector backbone, as described in Example 1, with the addition of a transformation whereby the vector backbone was fragmented into two pieces, with each piece comprising overlapping homologous regions to each other (47 bp) as well as to the gRNA cassette (~500 bp). This allowed for a 3-piece *in vivo* assembly via gap repair of a circular NatA marker plasmid incorporating the gRNA cassette. The NatA marker ORF was comprised on one of the two backbone fragments while the promoter driving NatA expression (*K. lactis* Tef1 promoter) was comprised on the other, and thus, NatA expression is possible only upon HR-mediated assembly of the fragments. Sequences of donor DNA targeting each of the Gal80, HO and ADH5 locus, respectively; gRNA constructs targeting each locus; and marker plasmid fragments for two-piece and three-piece *in vivo* assembly are provided herein as SEQ ID NOs:21 to 29.

[00192] Target sites for the Gal80, HO and ADH5 locus, respectively, are shown in **Table 5**.

[00193] **Table 5: CRISPR Target sites**

ORF	Target Sequence (NGG omitted)	SEQ ID NO
Gal80	TAAGGCTGCTGCTGAACGT	SEQ ID NO:30
HO	CCGGCTTGATCGACTCAGA	SEQ ID NO:31
ADH5	GGGTCATTGGTATCGATGG	SEQ ID NO:32

[00194] Cells were transformed and cultured as described in Example 1, and colonies appearing on selection were assayed for integration of the deletion construct at each locus by cPCR using an upstream forward primer that binds outside of the deletion construct, and a reverse primer that binds to a short linker sequence integrated in place of each open reading frame. 11 colonies were assayed for each delivery mode, as well as a parent colony THAT serves as a negative control ("N").

[00195] **Table 6. Primer sequences for cPCR verification of linker integration at Gal80, HO and ADH5 loci**

Primer Name	Description	Sequence	SEQ ID NO
Gal80-US-F	Gal80 locus US FOR	CAAACGGCCGCCTCTGCCATGGC AAAGAATGCTTTCCA	SEQ ID NO: 33
HO-US-F	HO locus US FOR	ACGTGTGTGTCTCATGGAAATTG ATGCAGTTGAAGACA	SEQ ID NO: 34
ADH5-US-F	ADH5 locus US FOR	GGCGTTATATCCAAACATTTTCAG ACAGAAGATT	SEQ ID NO: 35
R_	Linker REV	AGGTCCGCCGGCGTTGGACGAGC G	SEQ ID NO:36

[00196] As shown in **FIG. 13**, the rate of simultaneous triple integration at all three loci was substantially higher when three HR events were required to assemble the marker vector, compared to when only two HR events were required. In particular, with a 2-piece *in vivo* assembly of the marker vector, 6/11 colonies had an integration at the Gal80 locus, 10/11 had an integration at the HO locus, 7/11 colonies had an integration at the ADH5 locus, and 5/11 colonies (45.4%) had an integration at all three loci. By comparison, with a 3-piece *in vivo* assembly of the marker vector, 9/11 colonies had an integration at the Gal80 locus, 10/11 had an integration at the HO locus, 10/11 colonies had an integration at the ADH5 locus, and 9/11 colonies (81.8%) had an integration at all three loci. Thus, requiring a 3-piece gap repair of the marker vector instead of a 2-piece gap repair improved the rate of triple integration by nearly two-fold.

[00197] To determine if this improvement in multiplex integration rate could also be seen in diploid strains of *S. cerevisiae*, Cas9-expressing cells of the diploid yeast strain CAT-1 were similarly transformed with donor DNAs for simultaneous, pan-allelic, marker-less deletion of Gal80, HO and ADH5 open reading frames, gRNA constructs targeting each locus, and selective DNA fragmented into either 2 or 3 overlapping pieces. Colonies were assayed by cPCR using an upstream forward primer outside of the deletion construct, and a reverse primer binding to a short linker sequence integrated in place of each open reading frame (**Table 6**).

[00198] As shown in **FIG. 14**, the rate of simultaneous triple integration at all three loci of the diploid strain was also substantially higher when three HR events were required to assemble the marker vector. In particular, with a 2-piece *in vivo* assembly of the marker vector, 3/24 colonies had an integration at the Gal80 locus, 7/24 had an integration at the HO locus, 2/24 colonies had an integration at the ADH5 locus, and 1/24 colonies (4.2%) had an integration at all three loci. By comparison, with a 3-piece *in vivo* assembly of the marker vector, 3/8 colonies had an integration at the Gal80 locus, 5/8 had an integration at the HO locus, 3/8 colonies had an integration at the ADH5 locus, and 3/8 colonies (37.5%) had an

integration at all three loci. Thus, requiring a 3-piece gap repair of the marker vector instead of a 2-piece gap repair improved the rate of triple integration in the diploid strain by nearly ten-fold. The number of colonies recovered from the experiment was also roughly ten-fold fewer when 3 events were required (data not shown), suggesting that requiring higher order assembly of the marker vector selects for only the most HR competent cells of the population.

6.4 Example 4: Introduction of Single and Multiplex Point Mutations Using CRISPR in Combination with Gap Repair

[00199] This example demonstrates the application of the optimized protocol as described in Example 1(mode 4: *in vivo* HR-mediated incorporation of gRNA cassette(s) into a marker vector backbone) for introducing precise point mutations or corrections to point mutations.

[00200] Currently, introduction of a point mutation at a single locus is a tedious process. The *Delitto Perfetto* method allows marker-less introduction of point mutations, but requires integration of a marked cassette containing an inducible meganuclease in close proximity to the targeted site (Storici *et al.*, *Proc Natl Acad Sci USA* 100(25):14994-9 (2003)).

Alternatively, a complex integration cassette bearing a URA3 marker can be designed and integrated at the target site, such that subsequent loop out of the URA3 by 5-FOA counter selection reconstitutes the genetic locus with the point mutation included. Both of these methods are problematic for essential genes, require at least two rounds of genetic engineering, and are not amenable to multiplexing.

[00201] There are several considerations for the introduction of a point mutation (i.e., a missense SNP) by CRISPR. First, in addition to being unique in the genome, the site targeted for cutting should be as close as possible to the site of the desired SNP (**FIG. 15**). This is because recombination to repair the cut site does not require incorporation of the desired SNP, and the likelihood of its inclusion is expected to decrease with distance from the cut site. Secondly, for optimal efficiency, the donor DNA should be designed such that it is not also a target for CRISPR. Indeed, this issue was cited by DiCarlo *et al.* to explain the low rates of SNP integration by CRISPR observed in their experiments (DiCarlo *et al.*, *Nucleic Acids Res.*, 7,4336-4343 (2013))). To escape cutting, the desired SNP would need to disrupt the CRISPR target site in the donor DNA, an impossible requirement to satisfy at most loci. To make the donor DNA immune to cutting, and simultaneously improve the chances that recombination events include the desired SNP, a heterology block approach was adopted whereby silent mutations were made in the codons between the target site and the point mutation, reducing the potential for recombination events that would omit the desired SNP (**FIG. 15**). Additionally,

integration of the heterology block provides a novel primer binding site to identify candidate clones by PCR.

[00202] As a proof of principle, mutant alleles of yeast cells (*S. cerevisiae*) that had undergone mutagenesis were targeted for replacement with corresponding wild type alleles using the above-described approach. The mutagenized strain was made to constitutively express Cas9 under the control of the medium strength FBA1 promoter as described in Example 1. The Cas9p-expressing strain was then transformed with donor DNAs (one at a time, for single integration events) targeting each of the TRS31, CUE5, ECM38, PGD1, SMC6, NTO1 and DGA1 open reading frames, and gRNA constructs targeting each locus, each comprising overlapping homologous with a linear NatA marker plasmid backbone, allowing for *in vivo* assembly of a circular plasmid via gap repair. Cells were transformed and cultured as described in Example 1, then assessed for introduction of a point mutation (reversion allele) at each of the 7 loci. Candidate colonies and parent negative control (c) were assayed by colony PCR against the heterology block and flanking sequence, and selected positive colonies were confirmed by sequencing a larger PCR product spanning the integration locus.

[00203] Table 7. Primer sequences for cPCR verification of allele swaps

Primer Name	Description	Sequence	SEQ ID NO
TRS31-US-F	TRS31 locus US FOR	GTGCATTTGGCTCGAGTTGCTG	SEQ ID NO: 37
TRS31-DS-R	TRS31 locus DS REV	GGGAAGTTATCTACTATCATATA TTCATTGTCACG	SEQ ID NO: 38
TRS31- het-R	Heterology block primer	GAAAAGTAGAGATTCAGAATAG ATCCTTGAC	SEQ ID NO: 39
CUE5-US-F	CUE5 locus US FOR	GGAAGGTATCAAGGATTCTTCTC TCC	SEQ ID NO: 40
CUE5- DS-R	CUE5 locus DS REV	GAGGTGGCACATCTTCATCATCT TC	SEQ ID NO: 41
CUE5- het-R	Heterology block primer	CCAATAACTCATCCTGCTCCAAT TGT	SEQ ID NO: 42
ECM38-US-F	ECM38 locus US FOR	CAGACGCTGCAGTAACACAAGC	SEQ ID NO: 43
ECM38-DS-R	ECM38 locus DS REV	CTGAAGTGGGCAGTTCCATGC	SEQ ID NO: 44
ECM38- het-R	Heterology block primer	CAGTGATCTGGATCGTAGAAGGG C	SEQ ID NO: 45
PGD1-US-F	PGD-1 locus US FOR	CCAAGAGCATGCCACGGTTG	SEQ ID NO: 46
PGD1-DS-R	PGD-1 locus DS REV	GAGTTCCCATAGTACTACCGC	SEQ ID NO: 47

PGD1- het-R	Heterology block primer	GCAGACCTTATCTCTTGTCTCG	SEQ ID NO: 48
SMC6-US-F	SMC6 locus US FOR	GAGCTACTTTCACTGACTGCGC	SEQ ID NO: 49
SMC6-DS-R	SMC6 locus DS REV	GCGCTTCAATAGTAGTACCATCA GATG	SEQ ID NO: 50
SMC6- het-R	Heterology block primer	GCCGTTCTCTGATCTCAAAGAGA AT	SEQ ID NO: 51
NT01-US-F	NT01 locus US FOR	CTCAGTATGACATGGATGAACAG GATG	SEQ ID NO: 52
NT01-DS-R	NT01 locus DS REV	GGTACCTCCTGTAAGCTCCCTTTT C	SEQ ID NO: 53
NT01- het-R	Heterology block primer	GACTGAGACGTTCTGGACTCCTT C	SEQ ID NO: 54
DGA1-US-F	DGA1 locus US FOR	CTTAACCAAGCACGACAGTGGTC	SEQ ID NO: 55
DGA1-DS-R	DGA1 locus DS REV	GATTCCCTAGCGCCACCAAC	SEQ ID NO: 56
DGA1- het-R	Heterology block primer	CCTCTCCGGTGGCTGGTGATCTG	SEQ ID NO: 57
ADH2-US-F	ADH2 locus US FOR	CGAGACTGATCTCCTCTGCCGGA AC	SEQ ID NO: 58
ADH2-DS-R	ADH2 locus DS REV	GAATACTTCACCACCGAGCGAG	SEQ ID NO: 59
ADH2- het-R	Heterology block primer	GCATGTAAGTCTGTATGACATAC TCCTG	SEQ ID NO: 60
SIN4-US-F	SIN4 locus US FOR	CAAACGTCCTAAATGACCCATCG TTG	SEQ ID NO: 61
SIN4-DS-R	SIN4 locus DS REV	CAACTTCGGGTTTTGTTGTTGGTT AG	SEQ ID NO: 62
SIN4- het-R	Heterology block primer	CAATGGCAATTTACCGTAGTTGA AACCG	SEQ ID NO: 63
CYS4-US-F	CYS4 locus US FOR	CTCCAGAATCACATATTGGTGTT GC	SEQ ID NO: 64
CYS4-DS-R	CYS4 locus DS REV	CCATCTTAGTAACGATATGGATT GGTTTC	SEQ ID NO: 65
CYS4-het-R	Heterology block primer	CTGATGGAGTCAGGAAAGATGGC	SEQ ID NO: 66

[00204] Sequences of donor DNA targeting each of the TRS31, CUE5, ECM38, PGD1, SMC6, NT01, DGA1, ADH2, SIN4 and CYS4 locus, respectively; the target sequence of each locus, and gRNA constructs targeting each locus are provided herein as SEQ ID NOs:67-96.

[00205] As shown in **FIG. 16**, a high rate of heterology block integration was observed at each locus (ranging from 36.4% to 90.9%), and subsequent sequencing of PCR fragments spanning the desired mutations revealed a majority of clones contained the desired allele.

[00206] To determine the feasibility of multiplex introduction of point mutations, three loci (ECM38, PGD1 and ADH2) were targeted simultaneously for heterology block integration (allele swapping) using the optimized delivery mode for multiple gRNAs. As shown in **FIG. 17**, high rates of triple heterology block integration were observed by PCR assay (90.9 to 100%). To determine if even higher order multiplex integrations were feasible, five loci (ADH2, PGD1, ECM38, SIN4 and CYS4) were simultaneously targeted in a similar fashion. As shown in **FIG. 18**, simultaneous quintuple heterology block integration was confirmed by cPCR assay in 2/11 colonies (18.2%).

[00207] As a second proof of principle, eleven different mutant alleles were introduced into naïve *S. cerevisiae* strains and the resulting strains were tested for phenotypes conferred by these SNPs. Among the alleles examined was one allele relevant to industrial fermentation, conferring faster sedimentation (ACE2 S372*) (Oud, Guadalupe-Medina et al., *Proc Natl Acad Sci U S A* 110(45): E4223-4231, 2013), a series of temperature sensitive alleles in genes essential for cell division and the secretory pathway (SEC1 G443E, SEC6 L633P, MYO2 E511K, CDC28 R283Q) (Lorincz and Reed, *Mol Cell Biol* 6(11): 4099-4103, 1986; Roumanie, Wu et al. *J Cell Biol* 170(4): 583-594, 2005) and another pair related to improved high temperature growth (NCS2 H71L and END3 S258N) (Sinha, David et al., *Genetics* 180(3): 1661-1670, 2008; Yang, Foulquie-Moreno et al., *PLoS Genet* 9(8): e1003693, 2013). In addition, a series of five alleles associated with resistance to elevated ethanol concentrations were tested (SPT15 F177S, SPT15 Y195H, SPT15 K218R, PRO1 D154N and PUT1 deletion) (Takagi, Takaoka et al., *Appl Environ Microbiol* 71(12): 8656-8662, 2005; Alper, Moxley et al., *Science* 314(5805): 1565-1568, 2006).

[00208] High rates of heterology block integration (>90%) were observed for the introduction of most individual alleles (**FIG. 22**), and subsequent sequencing of PCR fragments spanning the desired mutations confirmed these changes. The temperature sensitive mutants were assayed at permissive and restrictive temperatures to confirm their intended phenotypes. Incomplete separation of cells during division caused by truncation of ACE2 was confirmed by bright-field microscopy, with dramatic clumping of cells (**FIG. 23 A**). Temperature-sensitive alleles of CDC28, MYO2 and SEC1 failed to grow at 37°C as expected (**FIG. 23 B**), and the CDC28 allele strain arrested in the G1 phase of growth (**FIG. 23 C**). To demonstrate secretory defects at the restrictive temperature in the SEC1 and SEC6 mutants, the

exocyst complex component SEC3 was carboxy-terminally GFP-tagged at its endogenous locus (also using CRISPR) to function as a reporter of secretory activity. SEC3 is normally localized to the bud in wild-type cells but its localization is clearly disrupted in both secretory mutants (**FIG. 23 D**).

[00209] In many cases, a phenotype results from the synergy of multiple alleles, but engineering such strains is even more time consuming, and is often not attempted. The optimized multiplex method was applied to this problem. Naïve CENPK2 bears one allele for high temperature growth (MKT1 D30G) (Yang, Foulquie-Moreno et al. 2013), and two additional mutations were introduced in NCS2 and END3. When grown overnight at a range of temperatures, neither of the individual alleles had an effect. However, strains containing both additional alleles integrated in a single step survived temperatures up to 42.7°C (**FIG. 23 E**).

[00210] Ethanol resistance alleles also conferred a synergistic effect. Wild-type CENPK2 tolerated up to 17.5% ethanol in this experiment (**FIG. 23F**). Mutations in SPT15 increased resistance up to 20% ethanol (**FIG. 23F**). To examine the interaction of these alleles, five targeted changes over three loci were simultaneously introduced into a naïve strain (three mutations in SPT15, PRO1 D154N and the deletion of PUT1). The three mutations in SPT15 were introduced on a single donor DNA by using two gRNAs to excise ~150 bp of the gene containing the three alleles. 27% of the resulting clones contained all five modifications as assessed by colony PCR and confirmed by sequencing (**FIG. 22**). The resulting strain had the highest ethanol tolerance, up to 22.5% ethanol (**FIG. 23F**). As these results demonstrate, multiplexed CRISPR allows rapid evaluation of hypotheses about combinations of causal alleles.

[00211] These results demonstrate that precise point mutations or reversions can be achieved at a high efficiency and at high multiplexity using the optimized methods and compositions for CRISPR-mediated genomic integration provided herein.

6.5 Example 5: Bi-allelic Engineering of Diploid Cells

[00212] This example demonstrates application of the optimized protocol as described in Example 1 (mode 4: *in vivo* HR-mediated incorporation of gRNA cassette(s) into a marker vector backbone) for simultaneous bi-allelic integration in diploid yeast strains.

[00213] Diploid industrial SC strains are highly heterozygous, with many unmapped but advantageous traits for fermentation. However, these strains are difficult to engineer by standard methods, requiring two sequential integration steps and distinct markers to delete a

gene or to introduce bi-allelic engineering. Thus, the efficacy of CRISPR-mediated bi-allelic deletion of the GAL80 locus in the CAT-1 and PE-2 diploid industrial strains of *S. cerevisiae* was tested using the optimized protocol of Example 1. Donor and gRNA sequences targeting the Gal80 locus are described in Example 3.

[00214] As shown in **FIG. 19**, cPCR of Cas9-expressing strains transformed with donor DNA and a linear gRNA cassette targeting the Gal80 locus and having homologous ends to a co-transformed linear NatA marker vector backbone revealed bi-allelic donor integration rates of 100% in CAT-1 diploid cells (**FIG. 19B**) and 90% in PE-2 diploid cells (**FIG. 19C**). These rates are comparable to the rate at which the same deletion in a haploid CENPK2 strain was obtained (100%; **FIG. 19A**). These results demonstrate the efficacy of the optimized methods and compositions for CRISPR-mediated genomic integration provided herein for engineering diploid host cells.

6.6 Example 6: Multiplex Integration of a Complete Biosynthetic Pathway

[00215] This example demonstrates application of the optimized protocol as described in Example 1 (mode 4: *in vivo* HR-mediated incorporation of gRNA cassette(s) into a marker vector backbone) for simultaneous integration into a naïve yeast strain of an entire biosynthetic pathway.

[00216] Typically, engineering metabolic pathways, even in tractable hosts such as *S. cerevisiae*, is time consuming. This timetable would be greatly improved if integrations of genetic cassettes could be conducted in parallel, and without requiring any integration of drug selectable markers. Therefore, the optimized protocol for CRISPR-mediated integration was applied towards the simultaneous integration of 12 gene cassettes totaling approximately 30 kb of DNA, encoding a functional pathway for production of farnesene (see U.S. Patent No. 8,415,136), into *S. cerevisiae*. Gene cassettes were designed and cloned using methods of polynucleotide assembly described in U.S. Patent Nos. 8,221,982 and 8,332,160. The pathway was divided into three donor constructs for integration of 12 genes: ERG10, encoding acetyl-CoA thiolase; ERG13, encoding HMG-CoA synthase; tHMG1 (2 copies), encoding HMG-CoA reductase; ERG12, encoding mevalonate kinase; ERG8, encoding phosphomevalonate kinase; ERG19, encoding mevalonate pyrophosphate decarboxylase; IDI1, encoding isopentenyl pyrophosphate isomerase; farnesene synthase (2 copies) from *Artemisia annua*; ERG20, encoding farnesyl pyrophosphate synthetase; and the transcriptional regulator GAL4. The three constructs were targeted for integration into the Gal80, HO, and BUD9 loci in a naïve CENPK2 *S. cerevisiae*, wherein Cas9 under the control of the medium strength FBA1

promoter (SEQ ID NO:3) was targeted to the GRE3 locus. Simultaneous, marker-less integration of all three constructs was attempted using the optimized gap repair method described in Example 1, and clones were assayed by cPCR primer pairs that bind the 5' flank of the integration target locus and an internal linker sequence within each donor construct.

[00217] As shown in **FIG. 20**, out of 47 clones screened, 11 clones (23.4%) were positive for integration of the 30 kb constituting the entire farnesene pathway (clones 22, 24, 29-32, 41-43, 45 and 46). The 48th clone is a parent negative control. All triple positive candidates were subsequently assayed by cPCR at the 3' flank of each target locus as well and confirmed to be positive for integration at both flanks. Subsequently, the 11 cPCR positive clones were tested for farnesene production in a batch sucrose plate model assay. As shown in **FIG. 21**, all 11 clones produced farnesene in amounts ranging from 0.1 to near 1.5 g/L amounts. Taken together, these results demonstrate the efficacy of the optimized methods and compositions for CRISPR-mediated genomic integration provided herein for rapid, multiplex, metabolic engineering. The optimized protocol for multiplex engineering can be applied to drastically shorten the timeline for engineering of complex pathways. For example, the simultaneous introduction of three point mutations in an *S. cerevisiae* strain would require approximately 6 weeks of work (1 week to introduce each allele, 1 week to recycle each marker, 3 cycles total), versus 1 week when using the optimized methods provided herein. The timeline for introduction of a rudimentary farnesene pathway demonstrated here is likewise compressed 6-fold, as the amount of time saved scales with the number of loci targeted in parallel.

[00218] As an additional proof of concept, multiplex integration of 11 gene cassettes containing 24 kb of DNA distributed over three loci, encoding a novel route to muconic acid was attempted in haploid CENPK2 (**FIG. 24**). Muconic acid is a precursor molecule with great potential for the production of bioplastics including nylon-6,6, polyurethane, and polyethylene terephthalate (PET). Currently, muconic acid is obtained from petroleum derived feedstocks via organic synthesis, but a renewable source is desirable. Biosynthesis of muconic acid is achieved by overexpression of the aromatic amino acid (shikimate) biosynthetic pathway. Previously, high level production of muconic acid (36.8 g/L, fed batch fermentation) was achieved in *E. coli* (Niu, Draths et al., *Biotechnol Prog* 18(2): 201-211, 2002). However, lower pH fermentation with *S. cerevisiae* would facilitate downstream processing and industrialization of the process. In a proof of principle effort, titers up to 141 mg/L have been observed in shake flask experiments in an engineered *S. cerevisiae* strain (Curran, Leavitt et al., *Metab Eng* 15: 55-66, 2013). In contrast to this initial attempt in *S. cerevisiae*, this

experiment utilized the *E. coli* shikimate pathway genes AROF, AROB and AROD rather than the native ARO1 gene. The engineered pathway is shown in **FIG. 24A**.

[00219] For integration, the pathway was divided into three split constructs (with internal overlap for reconstitution by homologous recombination *in vivo*) targeting the GAL80, HO, and ARO1 loci in CENPK (**FIG. 24B**). HO was chosen as a neutral locus, while GAL80 was selected to remove glucose repression of the galactose operon, and the ARO1 locus was deleted to force flux through the engineered pathway. ARO1 deletion also makes the strains auxotrophic for aromatic amino acids, creating a simple switch mechanism between the biomass production phase (in rich media) and the production phase (in minimal media). Simultaneous, marker-less integration of all three constructs was attempted using the optimized gap repair method, and clones were assayed by PCR, revealing a 4.2% rate of triple integration (n = 48). It is notable that integration of these three constructs requires nine recombination events (two flanking and one internal event per locus). While the observed rate is lower than seen for multiplex deletions, introduction of a complete biosynthetic pathway is expected to confer a fitness defect, and this may limit recovery of properly integrated strains.

[00220] Production of muconic acid and intermediates were tested in a 96 well shake-plate assay, with analysis by HPLC. The one-step integrated strains showed high titers of PCA (~3 g/L), indicating a bottleneck at AroY, which converts PCA to catechol (**FIG. 24A**). To confirm functionality of the downstream pathway, up to 1 g/L catechol was supplied directly to the production media wells, and quantitative conversion of catechol to *cis-trans* muconic acid was observed in the engineered, but not parent strain, unambiguously identifying a single defect in the pathway design at AroY.

[00221] To test the efficacy of the optimized gap repair method method in a second industrially relevant yeast, an attempt was made to integrate a more compact version of the muconic acid pathway comprising six genes in *K. lactis*. The pathway was divided into three integration constructs targeting the DIT1, ADH1, and NDT80 loci. A naïve *K. lactis* strain (ATCC 8585) was prepared by integrating Cas9 at the GAL80 locus, and deleting YKU80 to minimize the effects of non-homologous end joining (Kooistra, Hooykaas et al., Yeast 21(9): 781-792, 2004; Wesolowski-Louvel, FEMS Yeast Res 11(6): 509-513, 2011). Marker-less integration of all three constructs was accomplished in one step using the same gap-repair method, but with a plasmid backbone containing the pKD1 stability element (Chen, Wesolowski-Louvel et al., J Basic Microbiol 28(4): 211-220, 1988). Triple integrations occurred at a rate of 2.1%, as assayed by PCR (n=48). In analogy to the CENPK results, high titers of PCA (1 g/L) were observed, but no muconic acid production (**FIG. 24E**). Catechol

feeding experiments confirmed the same defect in AROY function. It is notable that ARO1 was not deleted in this *K. lactis* strain, and this discrepancy may explain the lower titers of PCA that were observed. Nonetheless, these results demonstrate the ability to prototype muconic acid production in two industrially relevant yeast strains and identify a limiting enzyme in less than a month, a workflow that facilitates rapid design iterations and allowed sampling of two potential hosts.

6.7 Example 7: Improved Integration in Mammalian Cells

[00222] This example demonstrates application of the optimized protocol as described in Example 1 (mode 4: *in vivo* HR-mediated incorporation of gRNA cassette(s) into a marker vector backbone) to achieve improved integration rates in a mammalian host cell.

[00223] To test whether the gap repair delivery method for gRNAs described above for *S. cerevisiae* might also improve integration rates in mammalian cells, a series of reagents were generated for transfection experiments in HEK-293T cells. In broad overview, cells were co-transfected with a linearized plasmid backbone containing a Cas9 expression cassette fused via 2A-linker to the 5' portion of the CD4 epitope tag, with a fragment containing a gRNA cassette targeting the AAVS1 locus, and with a donor DNA fragment for repair of the locus by homologous recombination, comprising upstream and downstream homology flanking an EcoRI site for later diagnostic purposes (gap repair condition). This transfection was compared to a control reaction with a Cas9-2A-CD4 expression cassette and gRNA cassette contained in a closed plasmid (positive control). In addition, a plasmid containing no gRNA was used as a negative control to assess whether homologous integration of the donor DNA occurred at a measureable rate in the absence of CRISPR-Cas9. Following transfection, CD4+ cells (transfected cells) were isolated using antibody-coupled magnetic beads, and cells were eluted and used in genomic DNA preparations. PCR of a region encompassing the integration site was performed, and PCR products were digested using EcoRI to determine the fraction of cells that had integrated the donor DNA at the target site.

[00224] Materials and Methods

[00225] Expression of the Cas9 nuclease and associated gRNA. The LifeTech/GeneArt CRISPR Nuclease with CD4 enrichment kit (A21175) was used. Following manufacturer's instructions, a double-stranded oligonucleotide (prepared by annealing oligos CUT1216 and CUT1217) encoding a sequence inside the AAVS1 region (AAVS1, T2 gRNA from Mali et al)

was ligated into the provided linearized vector to create pAM3473 (SEQ ID NO:98). The plasmid was maxi-prepped (Qiagen).

[00226] Generation of a version of pAM3473 suitable for testing gap repair. The pAM3473 plasmid was digested with Bst1107I and NheI to remove the entire gRNA cassette a portion of the CD4 ORF. The backbone was CIP (alkaline phosphatase) treated and gel purified. A multiple cloning site (MCS) double stranded oligo containing unique ClaI and XmaI sites and with compatible overhangs to the linear vector was prepared by annealing CUT1214 (SEQ ID NO:103) and CUT1215 (SEQ ID NO:104) oligos. The double-stranded oligo was ligated into the purified backbone to create pAM3472 (SEQ ID NO:97). The plasmid was maxi-prepped (Qiagen). Prior to use, the plasmid was linearized by digestion with ClaI and NheI, and purified/concentrated by ethanol precipitation.

[00227] Generation of a control plasmid containing CD4 epitope only (and no gRNA). pAM3473 (SEQ ID NO:98) was digested with Bst1107I and PacI, and the backbone was CIP treated and gel purified. A double-stranded oligo designed to re-circularize the gRNA cassette-less backbone was prepared by annealing CUT1254 (SEQ ID NO:113) and CUT1255 (SEQ ID NO:114) oligos and ligated into the vector backbone to create pAM15068 (SEQ ID NO:102; formerly known as "A2").

[00228] Preparation of fragments for gap repair of the AAVS1 gRNA cassette into linearized pAM3472. The primers CUT 1220 (SEQ ID NO:107) and CUT1221 (SEQ ID NO:108) were used to amplify a 2850 bp fragment from pAM3473. The product was sub-cloned by gap repair (*E. coli*) into the RYSE09 acceptor vector, and the construct was verified by sequencing to make pAM3515 (SEQ ID NO:100). Prior to transfection, linear fragment was prepared by Phusion PCR amplification using flanking RYSE0 (SEQ ID NO:117) and RYSE19 (SEQ ID NO:118) primers, and the PCR product was purified using Ampure beads (Axygen). Prior to transfection, linear fragment was prepared by Phusion PCR amplification using flanking RYSE0 and RYSE19 primers, and the PCR product was purified using Ampure beads (Axygen).

[00229] Preparation of fragments for gap repair of the CD4-only control fragment into linearized pAM3472. Primers CUT1220 (SEQ ID NO:107) and CUT1252 (SEQ ID NO:111) were used in a Phusion PCR reaction to amplify an upstream fragment containing the 3' end of the CD4 ORF using pAM3473 as template, and primers CUT1253 (SEQ ID NO:112) and CUT1221 (SEQ ID NO:108) were used to amplify a downstream fragment containing flanking

homology downstream of the gRNA cassette. These two fragments were gel purified and used in a second fusion PCR reaction, with primers RYSE0 and RYSE19 used for amplification of the ~2 kb product. The product was sub-cloned by gap repair (*E. coli*) into the RYSE09 acceptor vector, and the construct was verified by sequencing to make pAM3516 (SEQ ID NO:101). Prior to transfection, linear fragment was prepared from this template by Phusion PCR amplification using flanking RYSE0 and RYSE19 primers, and the PCR product was purified using Ampure beads (Axygen).

[00230] Preparation of donor DNA for introduction of an EcoRI site at the AAVS1 target locus. Primers CUT1226 (SEQ ID NO:119) and CUT1223 (SEQ ID NO:120) were used to amplify a ~570 bp upstream fragment containing a synthetic EcoRI site at its 3' end from human genomic DNA (derived from HEK-293 cells) using Phusion polymerase. Primers CUT1224 (SEQ ID NO:109) and CUT1227 (SEQ ID NO:110) were used to amplify a ~540 bp downstream fragment containing the EcoRI site at its 5' end from the same human genomic template. The fragments were assembled by fusion PCR using Phusion polymerase with the flanking RYSE0 and RYSE19 primers, and the ~1100 bp fragment was sub-cloned into linearized RYSE09 vector by gap repair (*E. coli*). The construct was verified by sequencing to make pAM3514 (SEQ ID NO:99). Prior to transfection, linear fragment was prepared from this template by Phusion PCR amplification using flanking RYSE0 and RYSE19 primers, and the PCR product was purified using Ampure beads (Axygen).

[00231] Transfection experiments. 70% confluent adherent 293T cells were transfected with DNA using Lipofectamine 3000 according to manufacturer's instructions (with a 1.5 fold DNA to LF3000 ratio). Table 8 provides the DNA constructs and amount of DNA used for each transfection (performed in duplicate).

Table 8.

	Vectors		Linear Fragments				
	linear	closed	No gRNA	linear	full gap	CD4 gap	ug total DNA
Transfection	pAM3472	pAM3473	"A2"	pAM3514	pAM3515	pAM3516	
2			10	5			15
3	10						10
4	10					5	15
6		10		5			15
8	10			5	5		20

[00232] After 48 hrs, cells were harvested using TrypLE reagent (LifeTech), and CD4+ cells were purified using the Dynabeads CD4 Positive Isolation Kit (LifeTech). Bound cells were

eluted from beads per manufacturer's instructions, and genomic DNA was prepared using the PrepGem Tissue Kit (Zygem) according to manufacturer's instructions.

[00233] RFLP assay for integration of EcoRI site. An RFLP assay was performed on PCR fragments (920 bp) amplified using Phusion Polymerase with a primer set (CUT1297 (SEQ ID NO:116) and CUT1294 (SEQ ID NO:115)) encompassing the EcoRI integration site with an "outside" primer, such that only donor DNA in the context of the intended genomic integration would yield a product. Fragments were purified using Ampure beads (Axygen) and digested with EcoRI. PCR products with an EcoRI site integrated by homologous recombination yielded 348 bp and 572 bp fragments. The fraction of template with integrated EcoRI site was calculated by densitometry (Image J) using the formula: digest band density (348 bp + 572 bp densities)/Total density (348 bp + 572 bp + 920 bp densities).

[00234] Results

[00235] To test whether imposing a requirement for gap repair might increase rates of homologous integration, we compared rates of EcoRI site donor DNA insertion in HEK-293T cells transfected with several different combinations of plasmid and linear DNA (Table 8 and **FIG. 25**). To assess whether the EcoRI donor DNA might integrate at the AAVS1 locus at some measureable level in the absence of targeted cutting by CRISPR-Cas9, cells were transfected with plasmid pAM15068, which contains the Cas9-2A-CD4 ORF, but no gRNA cassette, and linear donor DNA (pAM3514 PCR product). Transfected cells were purified using the dynabeads, genomic template was prepared, and digestion of the PCR product spanning the integration site yielded no digestion products, indicating that the rate of EcoRI site integration is not measurable in the absence of CRISPR-Cas9 (**FIG. 25**, transfection 2). To confirm that the linearized pAM3472 plasmid lacking a complete CD4 ORF could not confer a CD4⁺ phenotype on its own, transfections were conducted with just this linear fragment. No PCR product was obtained from template prepared from these purified cells, indicating that there was insufficient association of the transfected cells with the dynabeads to act as template for a PCR reaction (**FIG. 25**, transfection 3). To confirm that gap repair could reconstitute the CD4 ORF, the linearized pAM3472 plasmid was co-transfected with the CD4 gap repair fragment (pAM3516 PCR product), and template from cells purified from these transfections yielded a 920 bp band but no digestion products, as no gRNA was present. (**FIG. 25**, transfection 4). Next, transfections including the AAVS1 gRNA were examined. To establish a baseline for functionality of the CRISPR-Cas9 system with the AAVS1 gRNA, we co-transfected a closed plasmid containing an expression cassette for Cas9, CD4 and the gRNA

(pAM3473) and the linear donor DNA construct (pAM3514 PCR product). EcoRI digestion of the PCR product showed dropout products of 572 and 348 bp, indicative of digestion of a fraction of the total PCR product (**FIG. 25**, transfection 6). Quantification of the band densities using Image J software revealed that 22.5% of the total template contained an EcoRI site. To assess whether gap repair might improve this rate, we substituted the linearized pAM3472 vector and the gap repair fragment containing the missing portion of CD4 and the gRNA cassette (pAM3515 PCR product) for the closed vector (**FIG. 25**, transfection 8). Repeating the densitometry process, we observed that 47.5% of total template contained an EcoRI site. This represents a 2.1% fold improvement over the rate observed for the closed plasmid (transfection 6), thus confirming in mammalian cells the efficacy of the improved gap repair method for genomic integration.

[00236] Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be readily apparent to those of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.

The embodiments of the present invention for which an exclusive property or privilege is claimed are defined as follows:

1. An *in vitro* method for integrating an exogenous nucleic acid into a target site of a host cell genome, the method comprising:

- (a) contacting one or more host cells with:
 - (i) an exogenous nucleic acid (ES) capable of recombining, via homologous recombination, at the target site (TS) of the host cell genome, wherein ES comprises a first homology region (HR1) and a second homology region (HR2), wherein HR1 and HR2 are capable of recombining, via homologous recombination, with a third homology region (HR3) and a fourth homology region (HR4), respectively, wherein HR3 and HR4 are each at TS;
 - (ii) a nuclease (N) capable of generating a break at TS; and
 - (iii) a linear nucleic acid comprising a homology region that is capable of homologous recombination with a homology region of one or more additional linear nucleic acids contacted with the host cell, whereupon homologous recombination of the linear nucleic acids results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker;and
- (b) selecting a host cell that expresses the selectable marker.

2. The method of claim 1,
wherein the linear nucleic acid comprises a partial, interrupted and/or non-contiguous coding sequence for the selectable marker, wherein the selectable marker cannot be expressed from the linear nucleic acid, whereupon said formation of the circular extrachromosomal nucleic acid results in formation of a complete coding sequence of the selectable marker, wherein the selectable marker can be expressed from the circular extrachromosomal nucleic acid.

3. The method of either claim 1 or claim 2, wherein said selecting of step (b) comprises detecting the expression of the selectable marker via visual, colorimetric or fluorescent detection methods.

4. The method of any one of claims 1 to 3, further comprising the step of recovering a host cell wherein ES has homologously recombined at TS, optionally further wherein:

- (a) said recovering does not require integration of a selectable marker into the host cell genome; or
- (b) said recovering occurs at a frequency of about one every 10, 9, 8, 7, 6, 5, 4, 3, or 2 contacted host cells, or clonal populations thereof, screened.

5. The method of any one of claims 1 to 4, wherein a plurality of (n) exogenous nucleic acids is integrated into a plurality of (n) target sites of the host cell genome, wherein n is at least two, wherein step (a) comprises contacting the host cell with:

- (i) said plurality of exogenous nucleic acids, wherein x is an integer that varies from 1 to n, and for each integer x, each exogenous nucleic acid (ES)_x is capable of recombining, via homologous recombination, at a target site (TS)_x selected from said plurality of (n) target sites of said host cell genome;
- (ii) for each said target site (TS)_x, a nuclease (N)_x capable of generating a break at (TS)_x;

optionally further wherein:

- (a) a single nuclease is capable of cleaving each (TS)_x; or
- (b) n= 3, 4, 5, 6, 7, 8, 9 or 10.

6. The method of any one of claims 1 to 5, wherein:

- (a) N is capable of generating a single stranded break or a double stranded break at TS; or
- (b) ES is linear.

7. The method of any one of claims 1 to 6, wherein ES further comprises a nucleic acid of interest D, optionally further wherein D is selected from the group consisting

of a selectable marker, a promoter, a nucleic acid sequence encoding an epitope tag, a gene of interest, a reporter gene, and a nucleic acid sequence encoding a termination codon.

8. The method of any one of claims 1 to 7, wherein the circular extrachromosomal nucleic acid further comprises a coding sequence for the nuclease, optionally further wherein the nuclease is an RNA-guided DNA endonuclease, optionally further wherein the nuclease is a Cas9 endonuclease.

9. The method of claim 8, wherein the circular extrachromosomal nucleic acid further comprises a sequence that encodes a crRNA activity and a tracrRNA activity that enables site-specific recognition and cleavage of TS by the RNA-guided DNA endonuclease, optionally further wherein the crRNA activity and the tracrRNA activity are expressed as a single contiguous RNA molecule.

10. The method of any one of claims 1 to 9, wherein the nuclease is selected from the group consisting of an endonuclease, a zinc finger nuclease, a TAL-effector DNA binding domain-nuclease fusion protein (TALEN), a transposase, and a site-specific recombinase, optionally further wherein:

- (a) the zinc finger nuclease is a fusion protein comprising the cleavage domain of a TypeII restriction endonuclease fused to an engineered zinc finger binding domain, wherein the TypeII restriction endonuclease is selected from the group consisting of HO endonuclease and Fok I endonuclease, and wherein the zinc finger binding domain comprises 3, 5 or 6 zinc fingers; or
- (b) the endonuclease is:
 - (i) a homing endonuclease selected from the group consisting of an LAGLIDADG homing endonuclease, an HNH homing endonuclease, a His-Cys box homing endonuclease, a GIY-YIG homing endonuclease, and a cyanobacterial homing endonuclease;
 - (ii) selected from the group consisting of: H-DreI, I-SceI, I-SceII, I-SceIII, I-SceIV, I-SceV, I-SceVI, I-SceVII, I-CeuI, I-CeuAII, I-CreI, I-CrepsbIP, I-CrepsbIIP, I-CrepsbIIIP, I-CrepsbIVP, I-TliI, I-PpoI, Pi-PspI, F-SceI, F-SceII, F-SuvI, F-CphI, F-TevI, F-TevII, I-Amal, I-AniI, I-ChuI, I-Cmoel, I-Cpal, I-CpaiI, I-CsmI, I-CvuI, I-CvuAIP, I-

DdiI, I-DdiII, I-DirI, I-DmoI, I-HmuI, I-HmuII, I-HsNIP, I-LlaI, I-MsoI, I-NaaI, I-NanI, I-NclIP, I-NgrIP, I-NitI, I-NjaI, I-Nsp236IP, I-PakI, I-PboIP, I-PcuIP, I-PcuAI, I-PcuVI, I-PgrIP, I-PobIP, I-PorI, I-PorIIP, I-PbpIP, I-SpBetaIP, I-ScaI, I-SexIP, I-SneIP, I-SpomI, I-SpomCP, I-SpomIP, I-SpomIIP, I-SquIP, I-Ssp68031, I-SthPhiJP, I-SthPhiST3P, I-SthPhiSTe3bP, I-TdeIP, I-TevI, I-TevII, I-TevIII, i-UarAP, i-UarHGPAIP, I-UarHGPA13P, I-VinIP, I-ZbiIP, PI-MgaI, PI-MtuI, PI-MtuHIP, PI-MtuHIIP, PI-PfuI, PI-PfuII, PI-PkoI, PI-PkoII, PI-Rma43812IP, PI-SpBetaIP, PI-SceI, PI-TfuI, PI-TfuII, PI-ThyI, PI-TliI, and PI-TliII; or

- (iii) modified to specifically bind an endogenous genomic sequence, wherein the modified endonuclease no longer binds to its wild type endonuclease recognition sequence and is derived from a wild type endonuclease of (i) or (ii).

11. The method of any one of claims 1 to 10, wherein the host cell is:
 - (a) a prokaryotic cell;
 - (b) a eukaryotic cell;
 - (c) selected from the group consisting of a fungal cell, a bacterial cell, a plant cell, an insect cell, an avian cell, a fish cell and a mammalian cell;
 - (d) a mammalian cell selected from the group consisting of a rodent cell, a primate cell and a human cell; or
 - (e) a yeast cell, which is optionally *Saccharomyces cerevisiae*.
12. An isolated host cell comprising:
 - (i) an exogenous nucleic acid (ES) comprising homologous regions that are capable of recombining, via homologous recombination, at a target site (TS) of the host cell genome, wherein ES comprises a first homology region (HR1) and a second homology region (HR2), wherein HR1 and HR2 are capable of recombining, via homologous recombination, with a third homology region (HR3) and a fourth homology region (HR4), respectively, wherein HR3 and HR4 are each at TS;
 - (ii) a nuclease (N) capable of generating a break at TS; and

(iii) at least two linear nucleic acids comprising homology regions that are capable of homologous recombination with each other, whereupon said homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker;

and optionally further wherein the linear nucleic acid comprises a partial, interrupted and/or non-contiguous coding sequence for the selectable marker, wherein the selectable marker cannot be expressed from the linear nucleic acid, whereupon said formation of the circular extrachromosomal nucleic acid results in formation of a complete coding sequence of the selectable marker, wherein the selectable marker can be expressed from the circular extrachromosomal nucleic acid.

13. The host cell of claim 12, wherein the host cell comprises:

- (i) a plurality of exogenous nucleic acids, wherein
x is an integer that varies from 1 to n, and for each integer x, each exogenous nucleic acid (ES)_x is capable of recombining, via homologous recombination, at a target site (TS)_x selected from said plurality of (n) target sites of said host cell genome;
- (ii) for each said target site (TS)_x, a nuclease (N)_x capable of generating a break at (TS)_x.

14. The host cell of claim 12 or claim 13, wherein:

- (a) N is capable of generating a single stranded break or a double stranded break at TS; or
- (b) ES is linear.

15. The host cell of any one of claims 12 to 14, wherein ES further comprises a nucleic acid of interest D, optionally further wherein D is selected from the group consisting of a selectable marker, a promoter, a nucleic acid sequence encoding an epitope tag, a gene of interest, a reporter gene, and a nucleic acid sequence encoding a termination codon.

16. The host cell of any one of claims 12 to 15, wherein the circular extrachromosomal nucleic acid further comprises a coding sequence for the nuclease,

optionally further wherein the nuclease is an RNA-guided DNA endonuclease, optionally further wherein the nuclease is a Cas9 endonuclease.

17. The host cell of claim 16, wherein the circular extrachromosomal nucleic acid further comprises a sequence that encodes a crRNA activity and a tracrRNA activity that enables site-specific recognition and cleavage of TS by the RNA-guided DNA endonuclease, optionally further wherein the crRNA activity and the tracrRNA activity are expressed as a single contiguous RNA molecule.

18. The host cell of any one of claims 12 to 17, wherein the nuclease is selected from the group consisting of an endonuclease, a zinc finger nuclease, a TAL-effector DNA binding domain-nuclease fusion protein (TALEN), a transposase, and a site-specific recombinase, optionally further wherein:

- (a) the zinc finger nuclease is a fusion protein comprising the cleavage domain of a TypeII restriction endonuclease fused to an engineered zinc finger binding domain, wherein the TypeII restriction endonuclease is selected from the group consisting of HO endonuclease and Fok I endonuclease, and wherein the zinc finger binding domain comprises 3, 5 or 6 zinc fingers; or
- (b) the endonuclease is:
 - (i) a homing endonuclease selected from the group consisting of an LAGLIDADG homing endonuclease, an HNH homing endonuclease, a His-Cys box homing endonuclease, a GIY-YIG homing endonuclease, and a cyanobacterial homing endonuclease;
 - (ii) selected from the group consisting of: H-DreI, I-SceI, I-SceII, I-SceIII, I-SceIV, I-SceV, I-SceVI, I-SceVII, I-CeuI, I-CeuAIIIP, I-CreI, I-CrepsbIP, I-CrepsbIIP, I-CrepsbIIIP, I-CrepsbIVP, I-TliI, I-PpoI, Pi-PspI, F-SceI, F-SceII, F-SuvI, F-CphI, F-TevI, F-TevII, I-Amal, I-AniI, I-ChuI, I-Cmoel, I-CpaI, I-CpaII, I-CsmI, I-CvuI, I-CvuAIP, I-DdiI, I-DdiII, I-DriI, I-DmoI, I-HmuI, I-HmuII, I-HsNIP, I-LlaI, I-MsoI, I-Naai, I-NanI, I-NclIP, I-NgrIP, I-NitI, I-NjaI, I-Nsp236IP, I-PakI, I-PboIP, I-PcuIP, I-PcuAI, I-PcuVI, I-PgrIP, I-PobIP, I-PorI, I-PorIIP, I-PbpIP, I-SpBetaIP, I-ScaI, I-SexIP, I-SneIP, I-SpomI, I-SpomCP, I-SpomIP, I-SpomIIP, I-SquIP, I-Ssp6803I, I-SthPhiJP, I-

SthPhiST3P, I-SthPhiSTe3bP, I-TdeIP, I-TevI, I-TevII, I-TevIII, i-UarAP, i-UarHGPAIP, I-UarHGPA13P, I-VinIP, I-ZbiIP, PI-Mgal, PI-MtuI, PI-MtuHIP, PI-MtuHIIP, PI-PfuI, PI-PfuII, PI-PkoI, PI-PkoII, PI-Rma43812IP, PI-SpBetaIP, PI-SceI, PI-TfuI, PI-TfuII, PI-ThyI, PI-TliI, and PI-TliII; or

- (iii) modified to specifically bind an endogenous genomic sequence, wherein the modified endonuclease no longer binds to its wild type endonuclease recognition sequence and is derived from a wild type endonuclease of (i) or (ii).

19. A composition comprising:

- (a) a site-specific nuclease, or a nucleic acid comprising a coding sequence for a site-specific nuclease;
- (b) a first linear nucleic acid and one or more additional linear nucleic acid, wherein the first and second linear nucleic acids are capable of homologously recombining with each other after introduction into a host cell, whereupon said homologous recombination results in formation of a circular nucleic acid comprising a coding sequence for a selectable marker; and
- (c) an exogenous nucleic acid comprising homologous regions that are capable of recombining, via homologous recombination, at a target site of a host cell genome.

20. A cell culture composition comprising a cell culture medium and the host cell of any one of claims 12 to 18.

21. An *in vitro* method for integrating an exogenous nucleic acid into a target site of a host cell genome, the method comprising:

- (a) contacting one or more host cells with:
 - (i) an exogenous nucleic acid (ES) capable of recombining, via homologous recombination, at the target site (TS) of the host cell genome; and
 - (ii) a linear nucleic acid capable of homologous recombination with one or more additional linear nucleic acids contacted with the host cell,

whereupon said homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker and a coding sequence for a nuclease (N) capable of generating a break at TS; and

- (b) selecting a host cell that expresses the selectable marker.

22. A method for integrating an exogenous nucleic acid into a target site of a host cell genome, the method comprising:

- (a) obtaining one or more isolated host cells;
- (b) contacting one or more host cells with:
 - (i) an exogenous nucleic acid (ES) capable of recombining, via homologous recombination, at the target site (TS) of the host cell genome, wherein ES comprises a first homology region (HR1) and a second homology region (HR2), wherein HR1 and HR2 are capable of recombining, via homologous recombination, with a third homology region (HR3) and a fourth homology region (HR4), respectively, wherein HR3 and HR4 are each at TS;
 - (ii) a nuclease (N) capable of generating a break at TS; and
 - (iii) a linear nucleic acid comprising a homology region that is capable of homologous recombination with a homology region of one or more additional linear nucleic acids contacted with the host cell, whereupon homologous recombination of the linear nucleic acids results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker;

and

- (c) selecting a host cell that expresses the selectable marker.

23. A method for integrating an exogenous nucleic acid into a target site of a host cell genome, the method comprising:

- (a) obtaining one or more isolated host cells;
- (b) contacting one or more host cells with:

- (i) an exogenous nucleic acid (ES) capable of recombining, via homologous recombination, at the target site (TS) of the host cell genome; and
 - (ii) a linear nucleic acid capable of homologous recombination with one or more additional linear nucleic acids contacted with the host cell, whereupon said homologous recombination results in formation of a circular extrachromosomal nucleic acid comprising a coding sequence for a selectable marker and a coding sequence for a nuclease (N) capable of generating a break at TS; and
- (c) selecting a host cell that expresses the selectable marker.

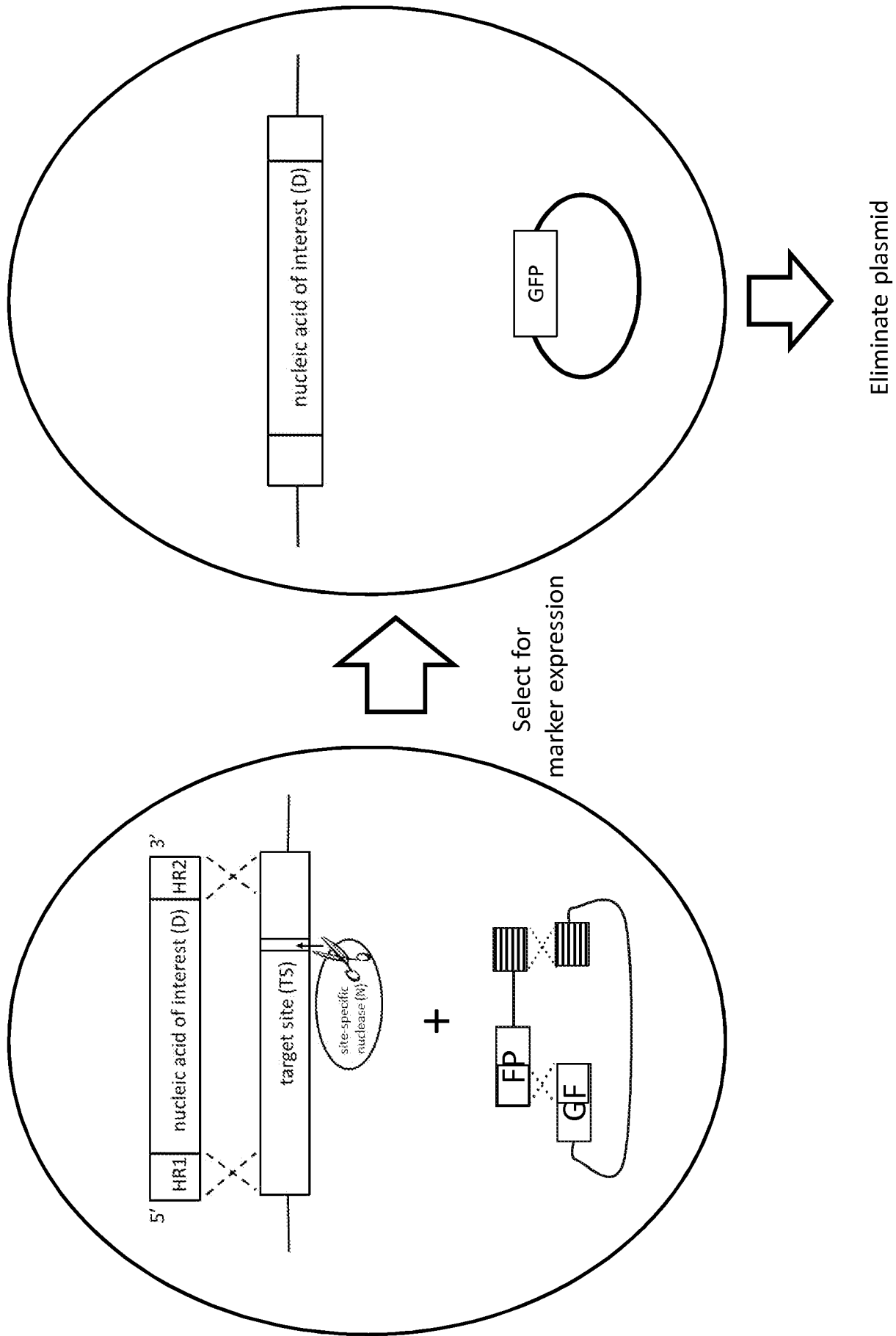


FIGURE 1

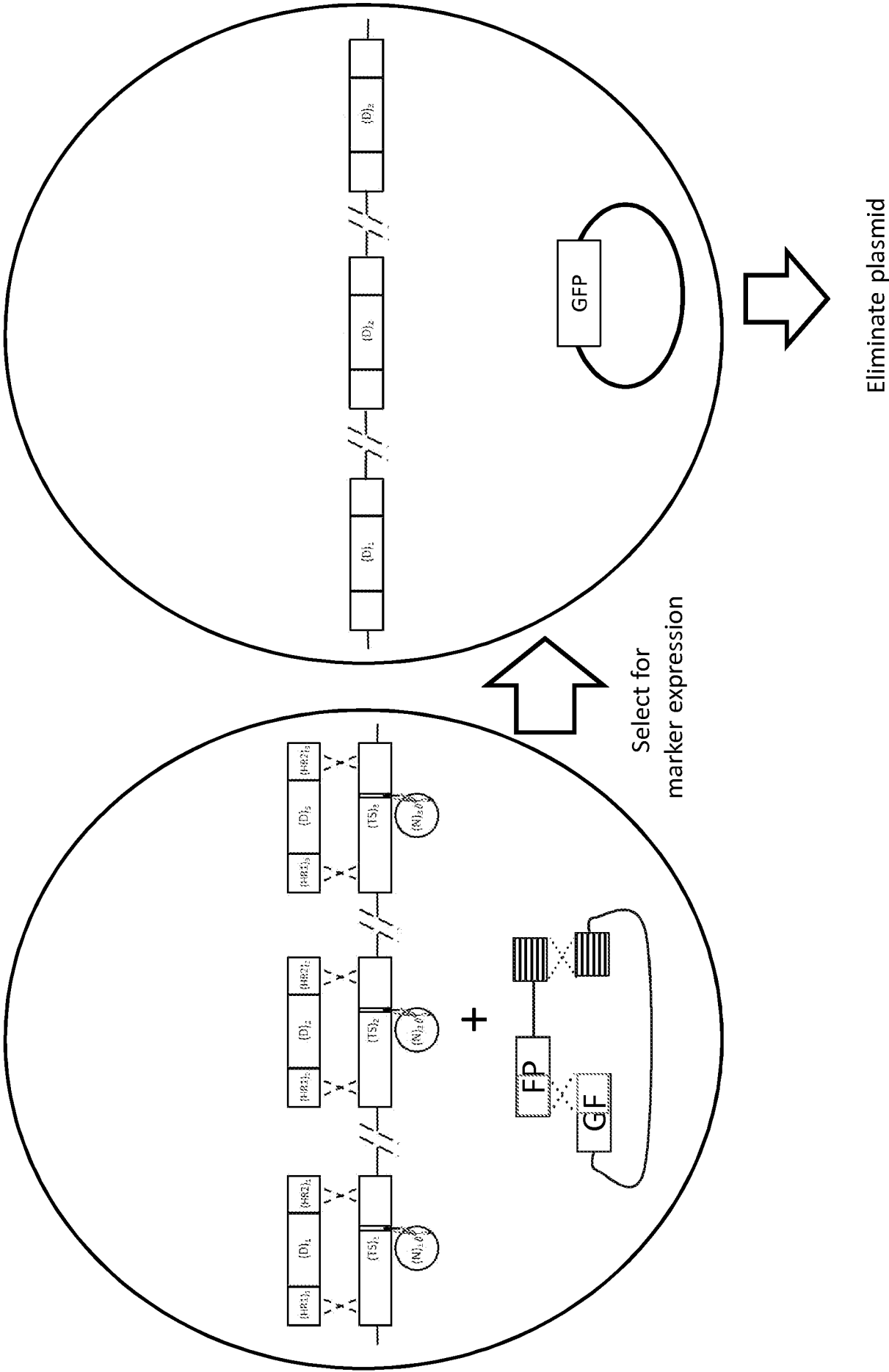


FIGURE 2

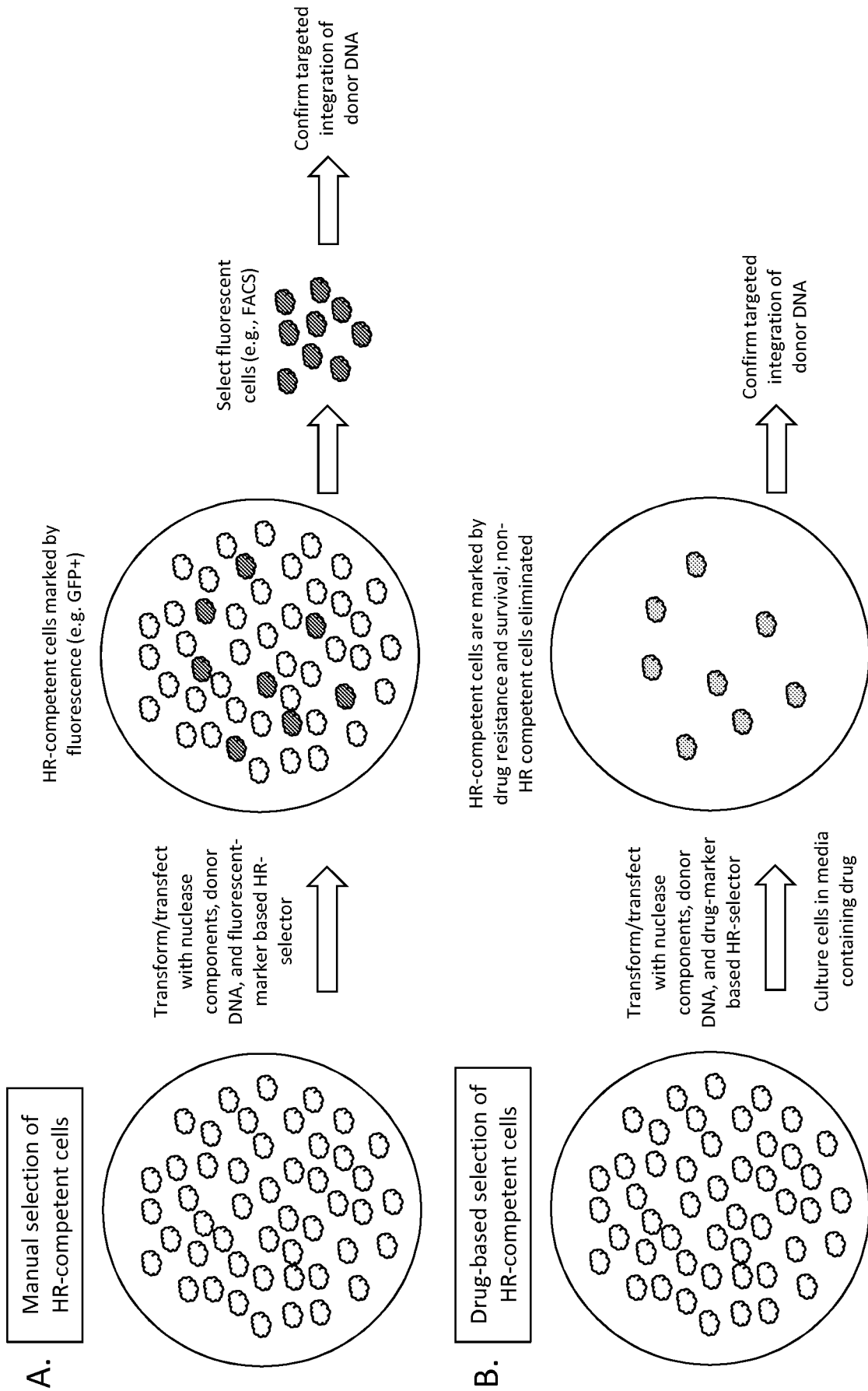
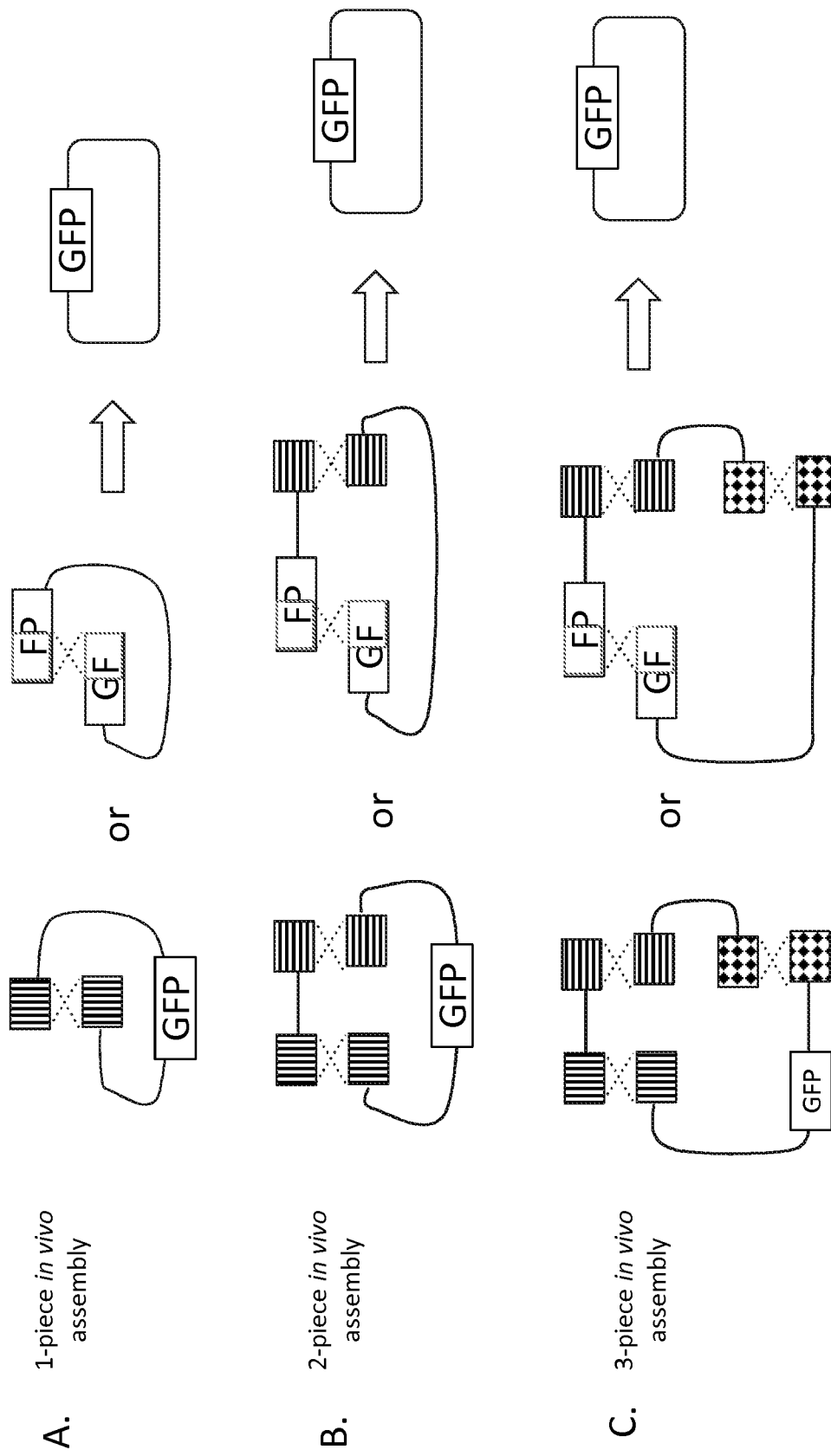


FIGURE 3



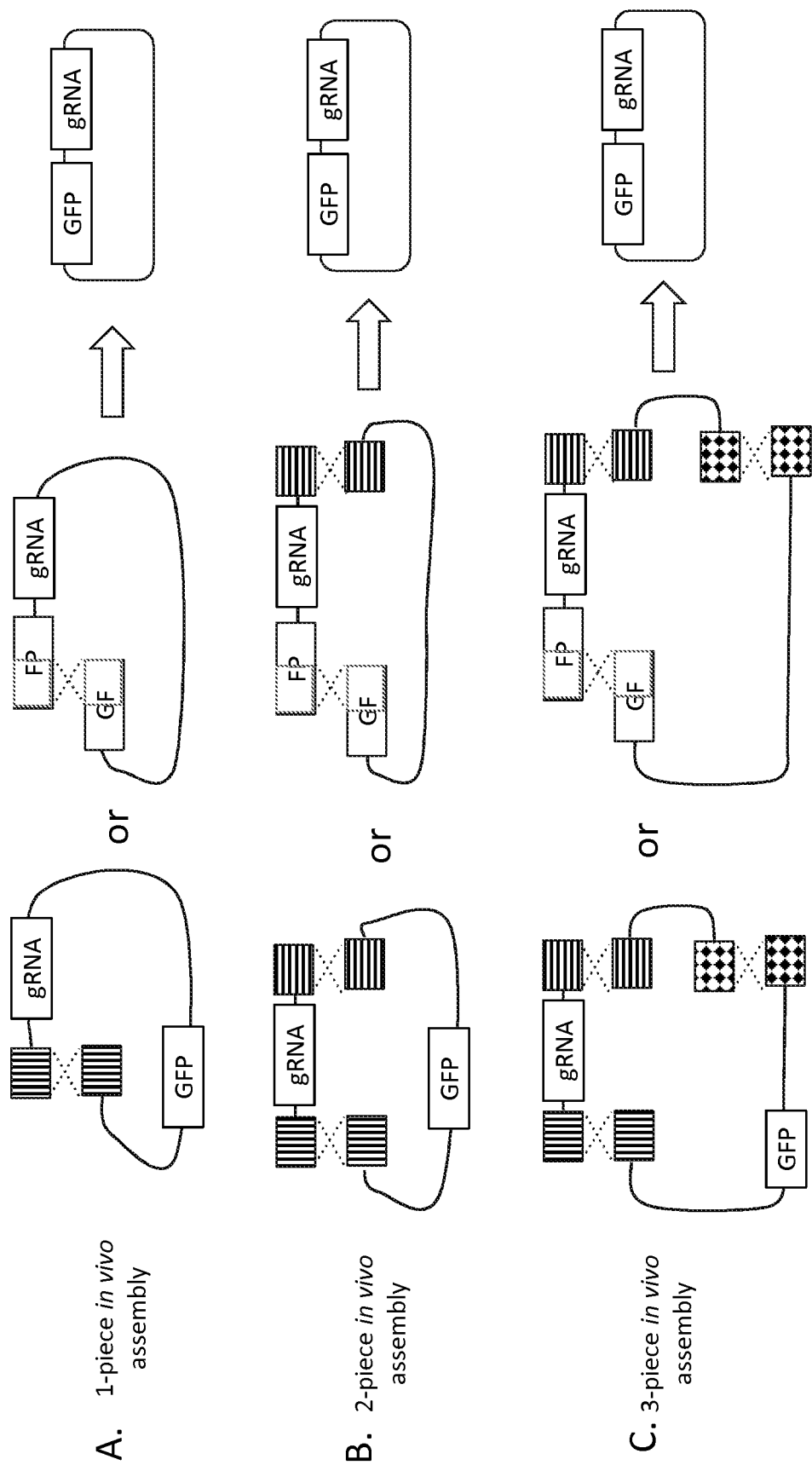
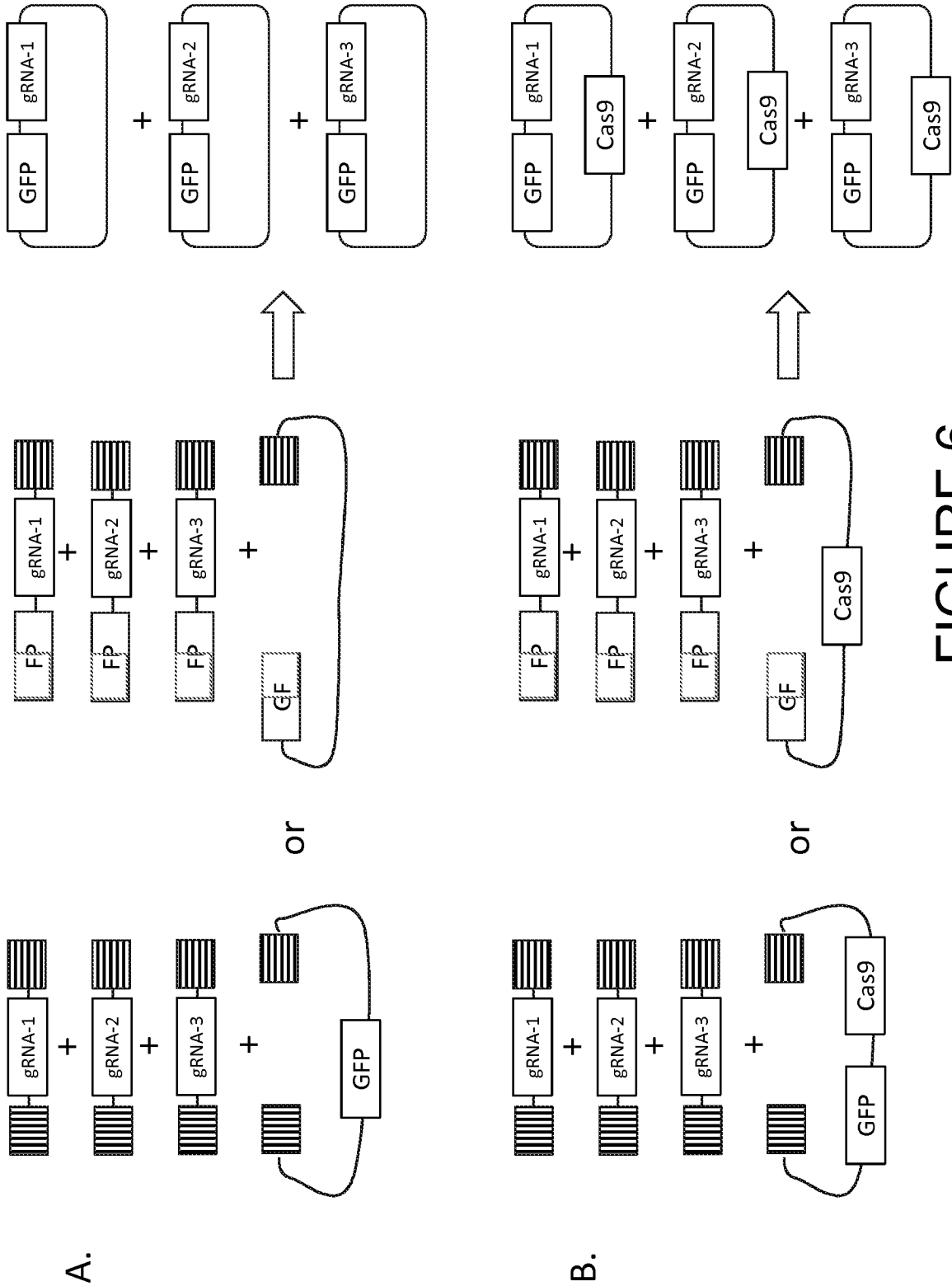
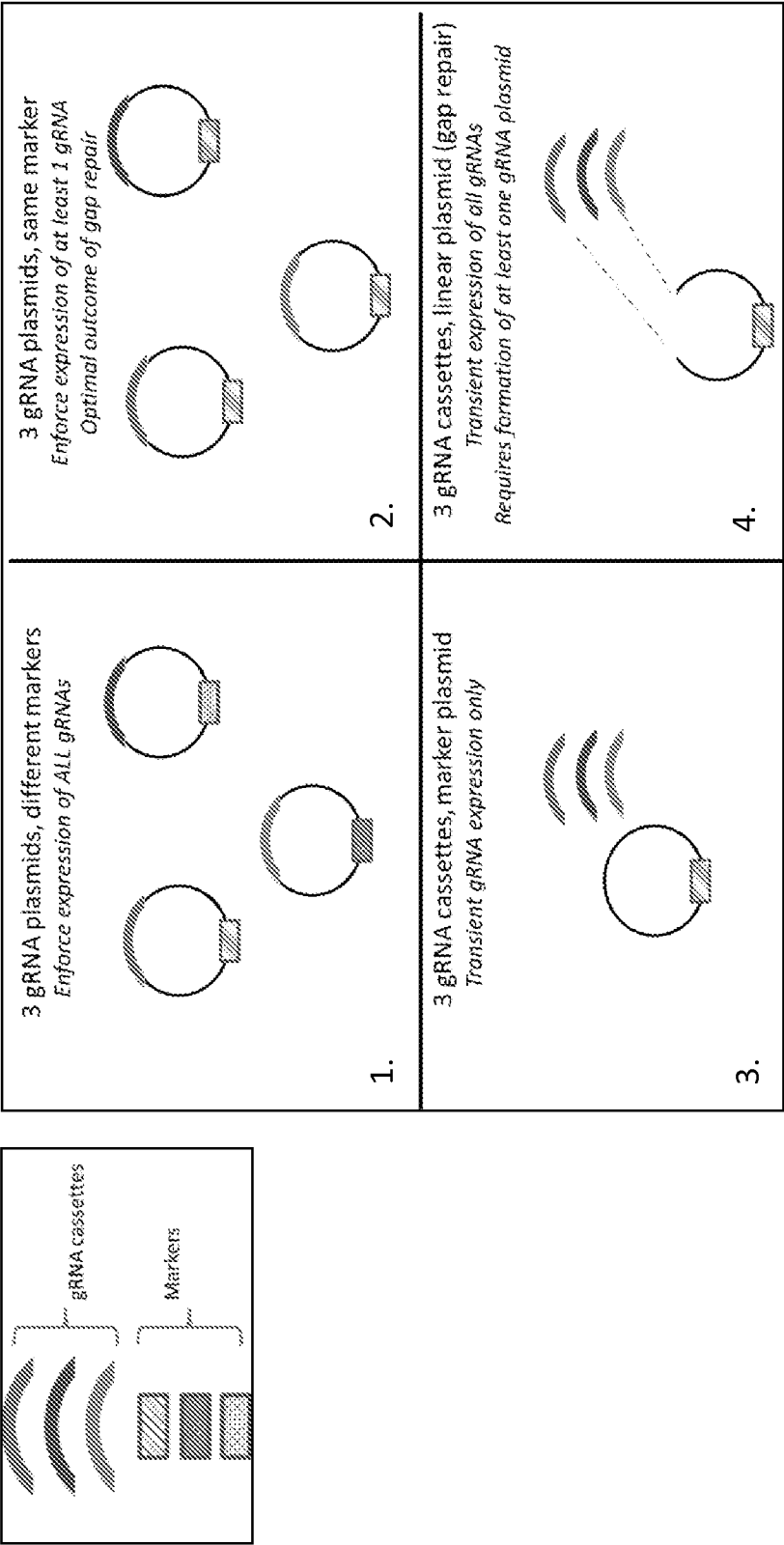


FIGURE 5

2-piece *in vivo* assembly for CRISPR-mediated multiplex integrations





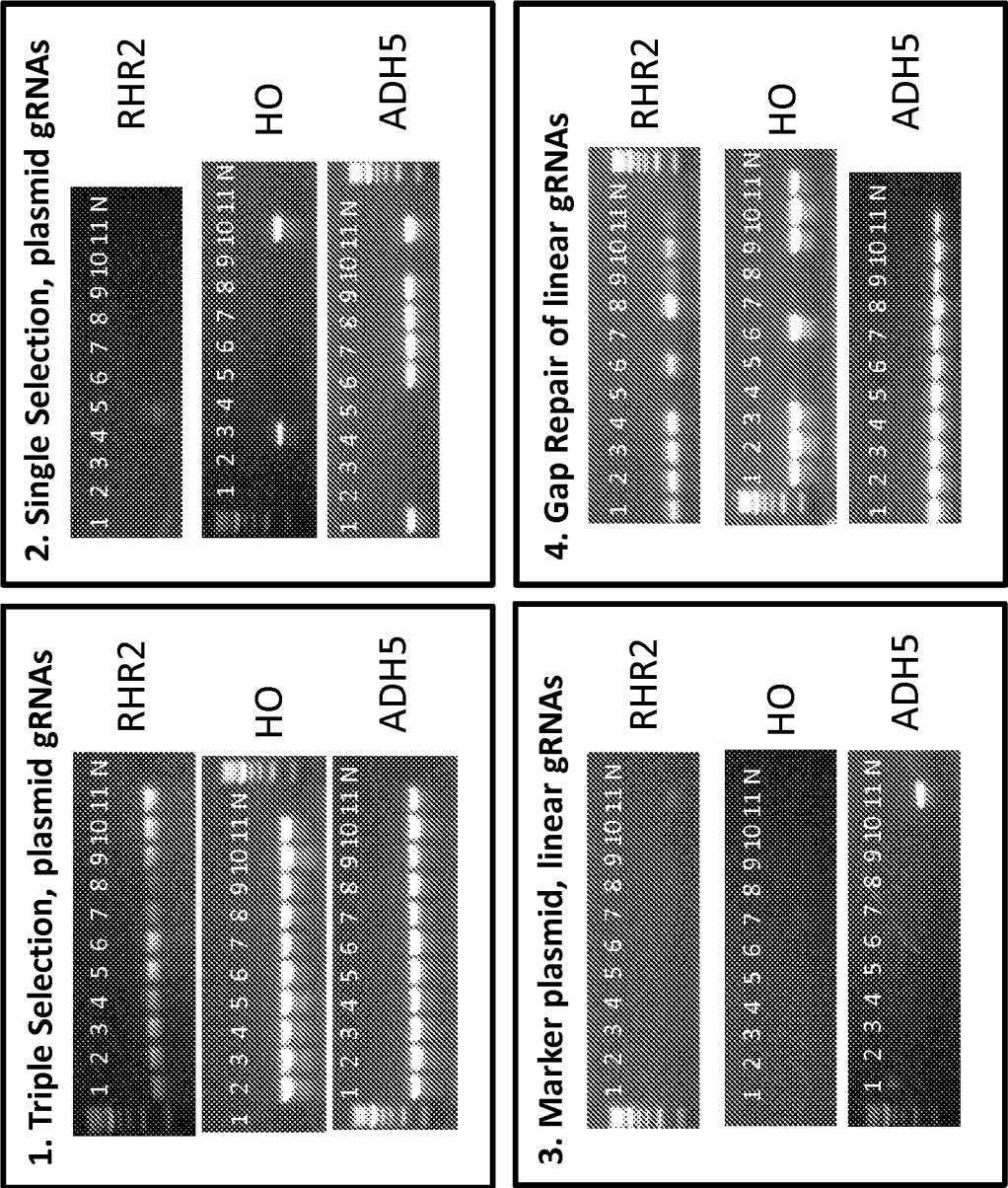


FIGURE 8

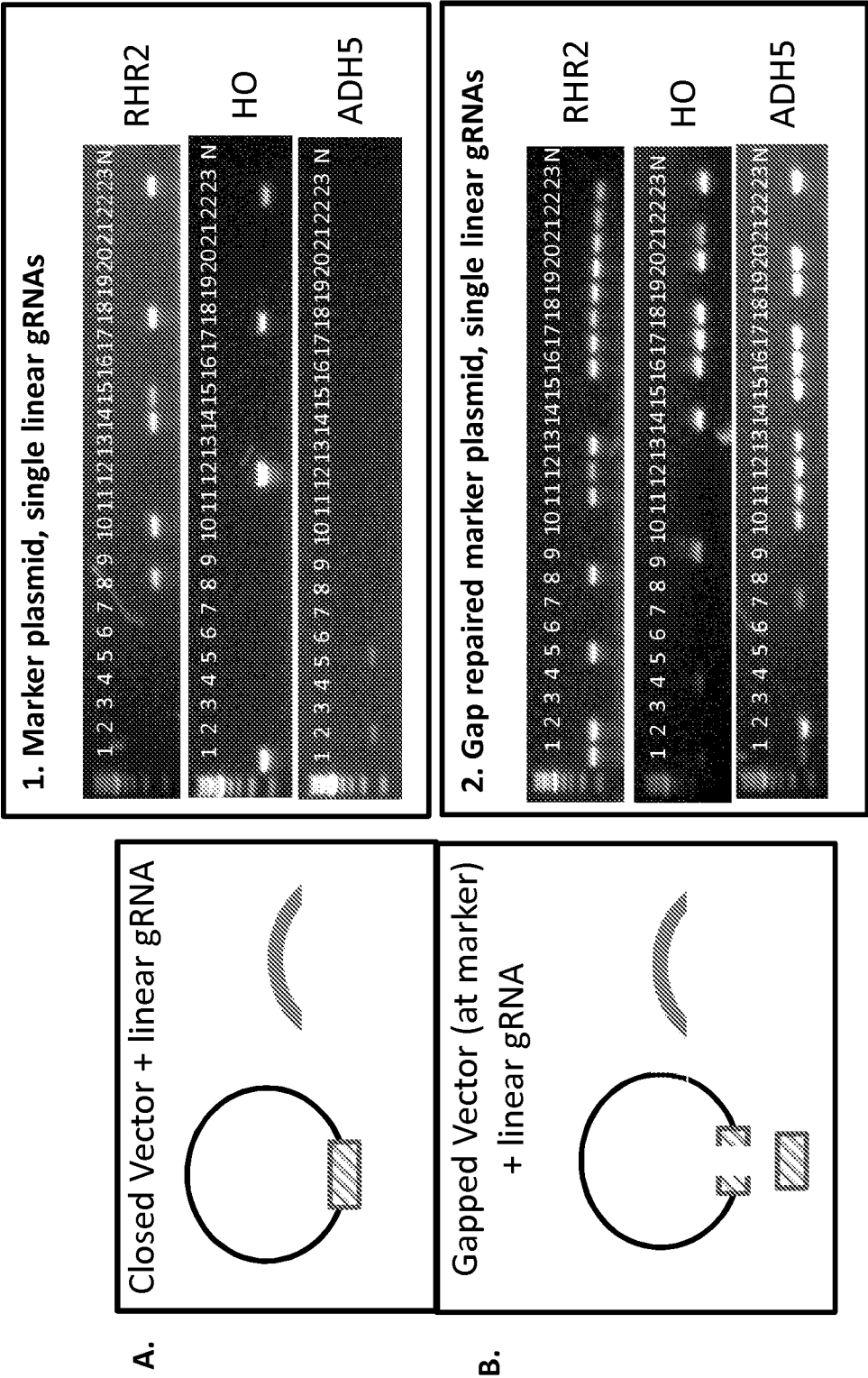


FIGURE 9

		1st Expt		2nd Expt		3rd Expt	
		#positive / #screened	fold increase	#positive / #screened	fold increase	#positive / #screened	fold increase
gRNA	marker vector						avg. fold increase
RHR2	Closed	34.80%		4.30%		26.10%	
	Gap repair vector						
RHR2		34.80%	0	30.40%	7.07	65.20%	2.50
HO	Closed	13.00%		4.30%		17.40%	
	Gap repair vector						
HO		43.50%	3.35	34.80%	8.09	43.50%	2.50
ADH5	Closed	34.80%		17.40%		8.70%	
	Gap repair vector						
ADH5		60.90%	1.75	47.80%	2.75	56.50%	6.49
							3.66

FIGURE 10

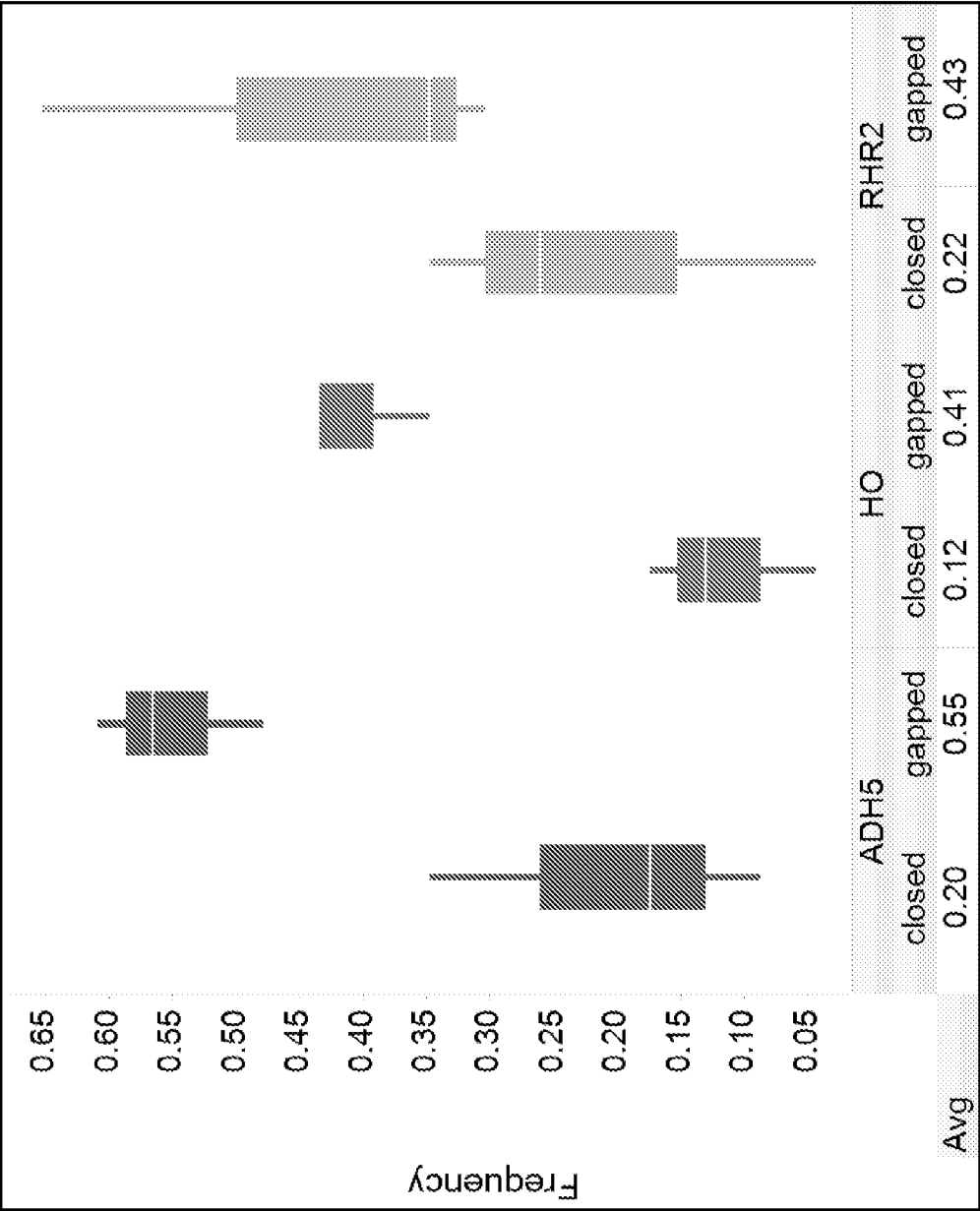


FIGURE 11

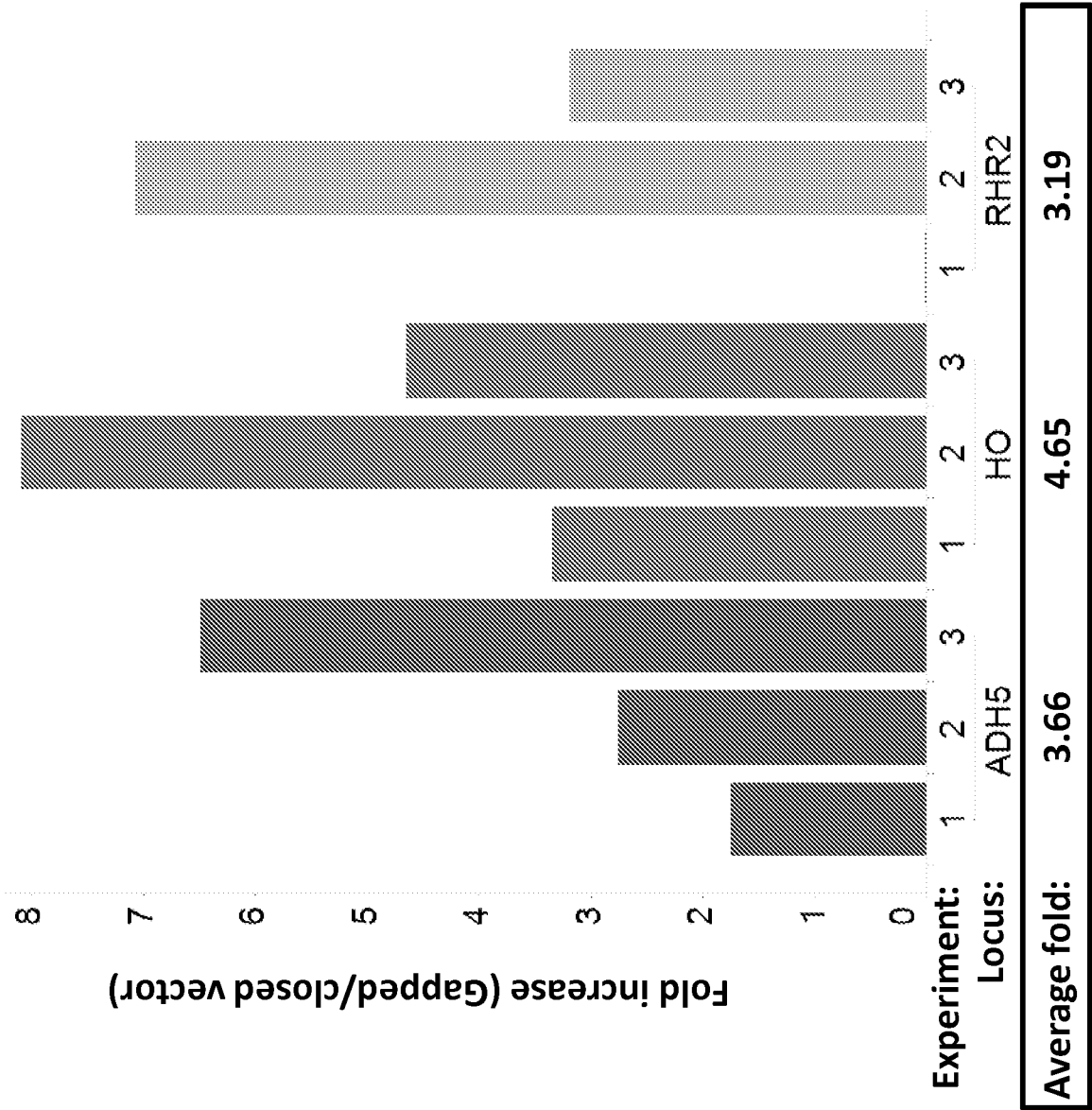


FIGURE 12

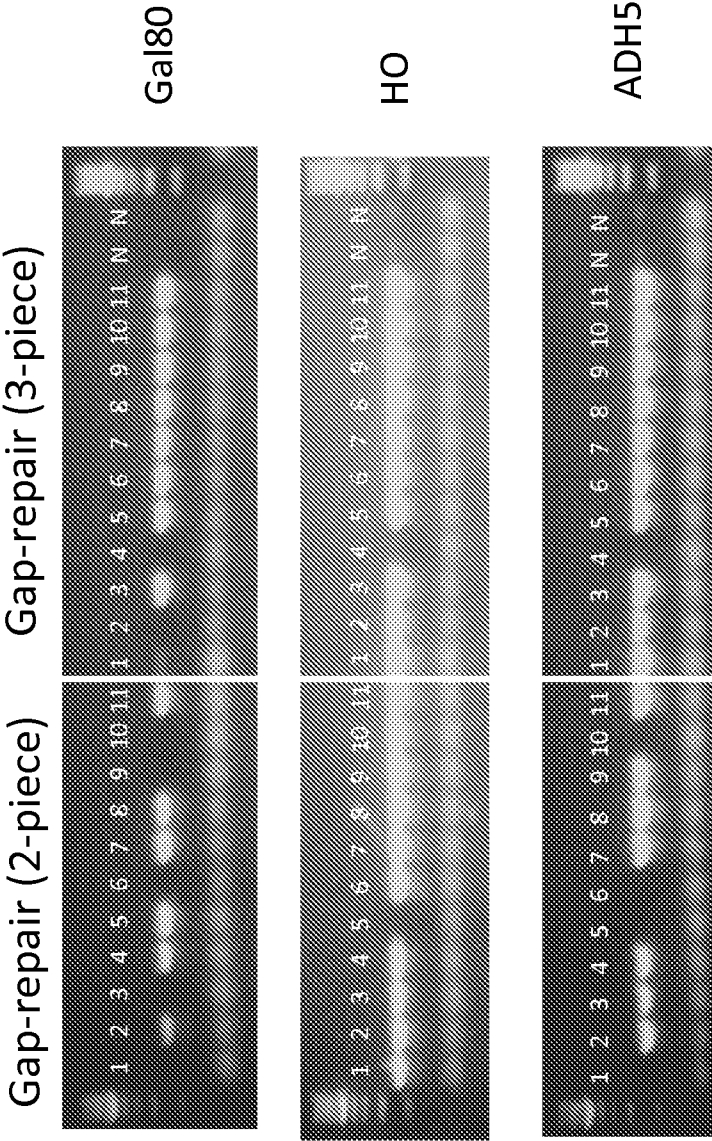
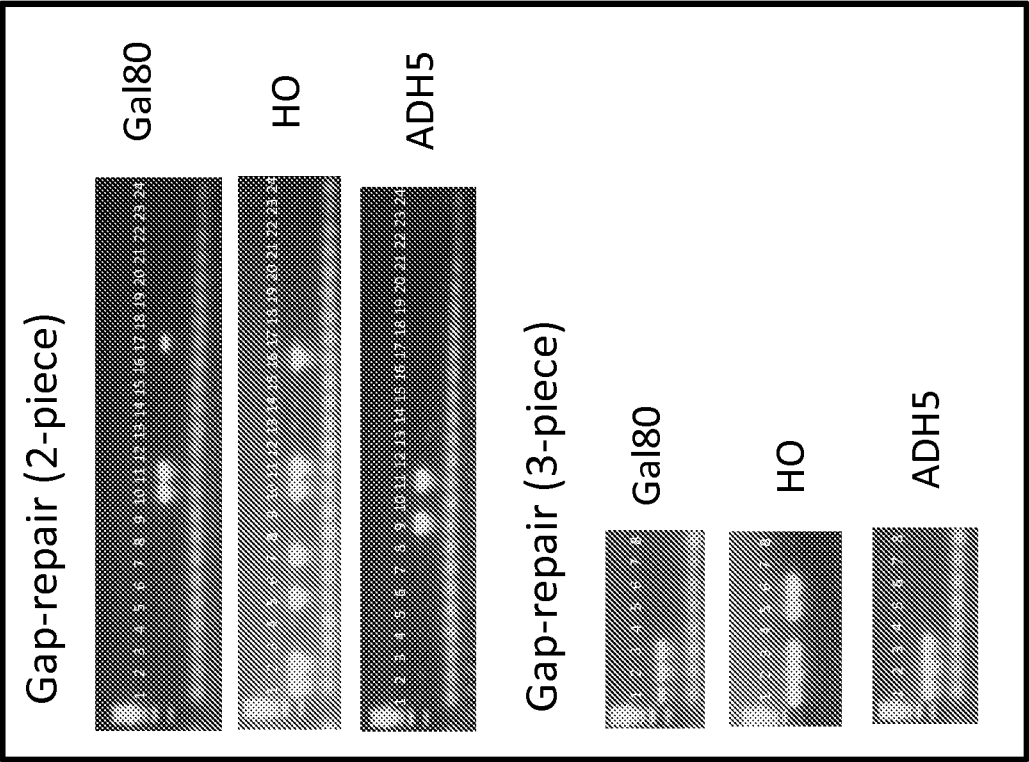


FIGURE 13



Description	Triple Integration Rate	N
2-piece gap repair	4%	24
3-piece gap repair	37.50%	8

FIGURE 14

FIGURE 15

FIGURE 15

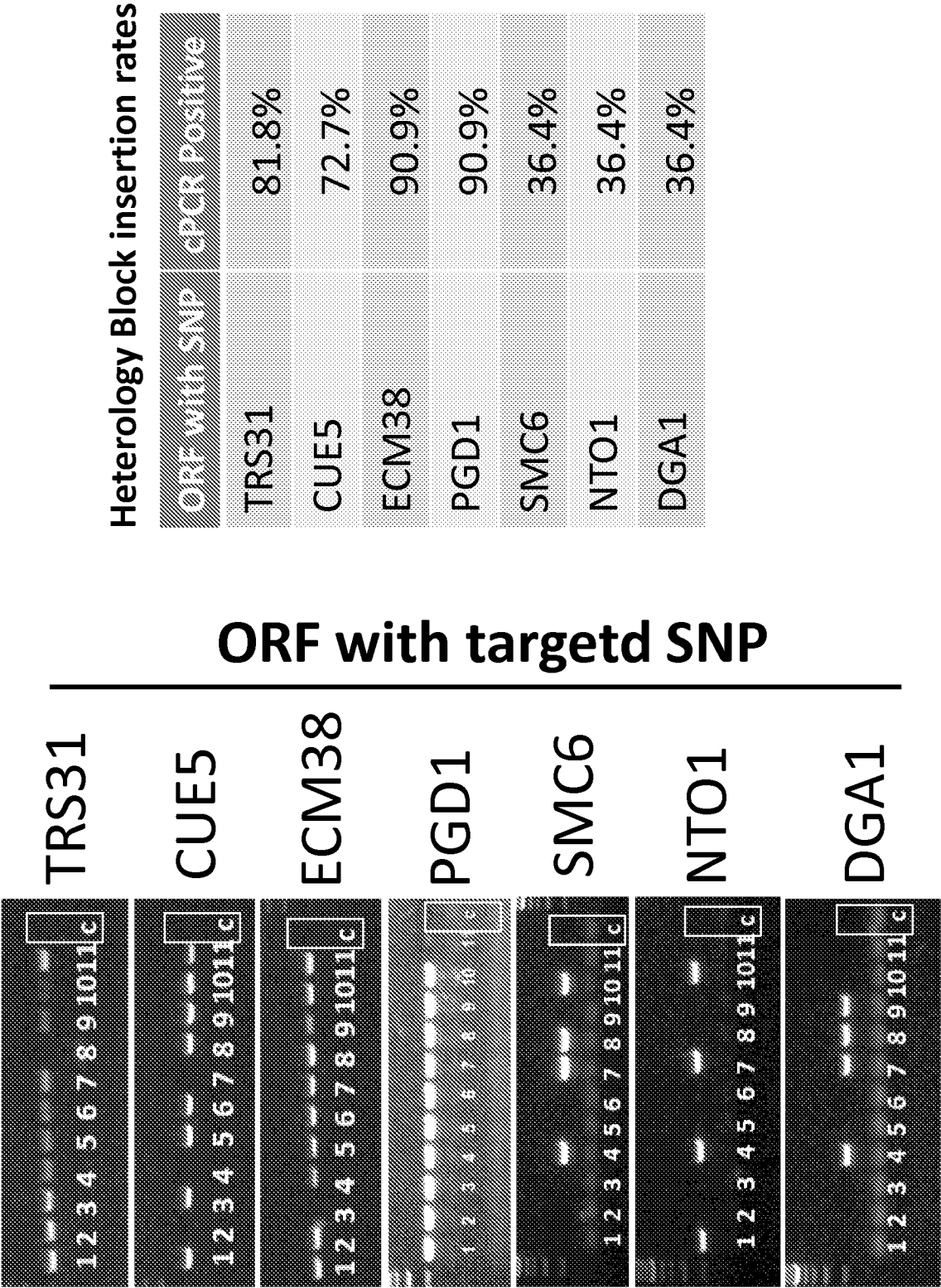
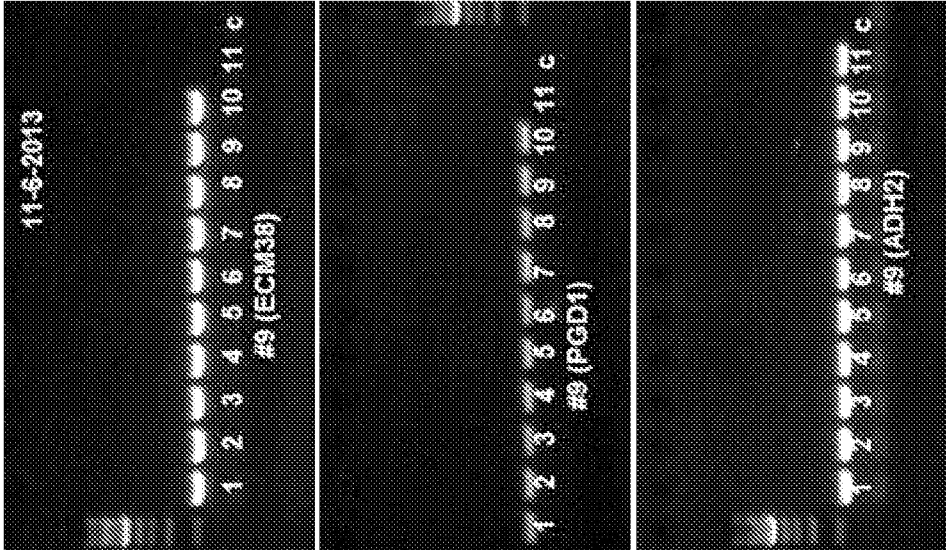


FIGURE 16



Heterology Block insertion rates

ORF with SNP	cPCR Positive
ECM38	90.9%
PGD1	90.9%
ADH2	100%
Triple	90.9%

FIGURE 17

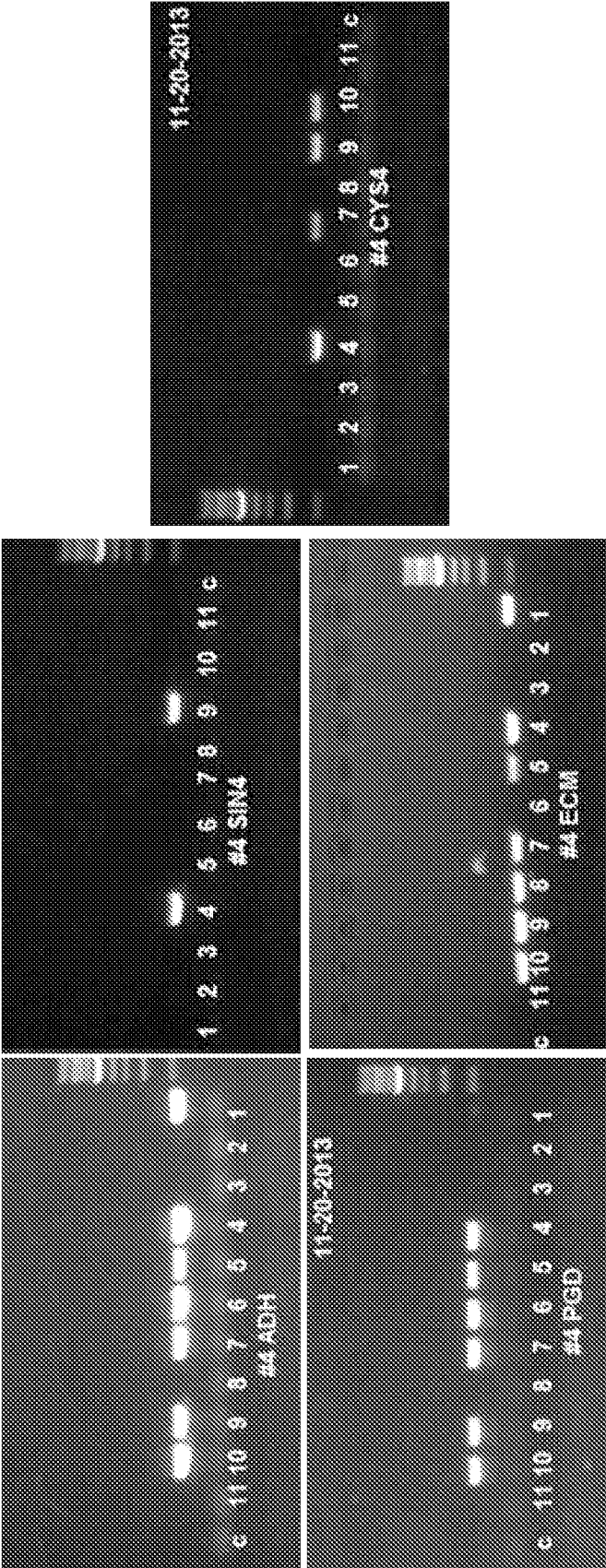


FIGURE 18

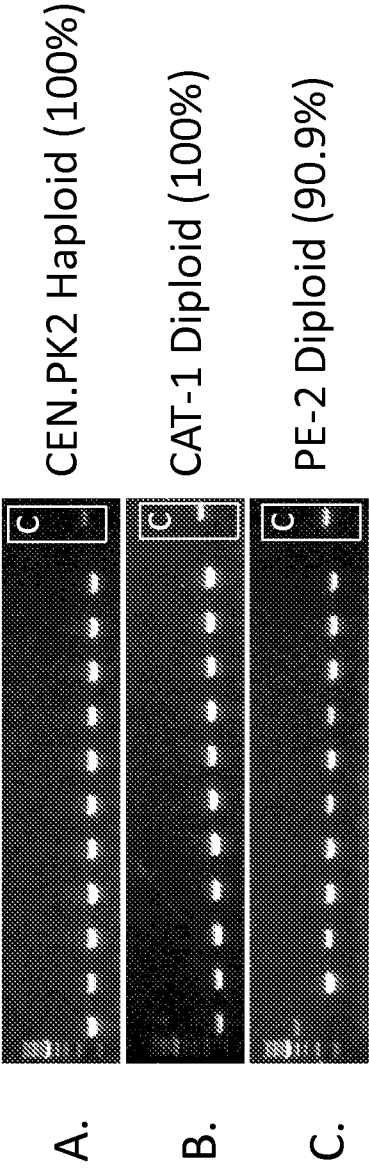


FIGURE 19

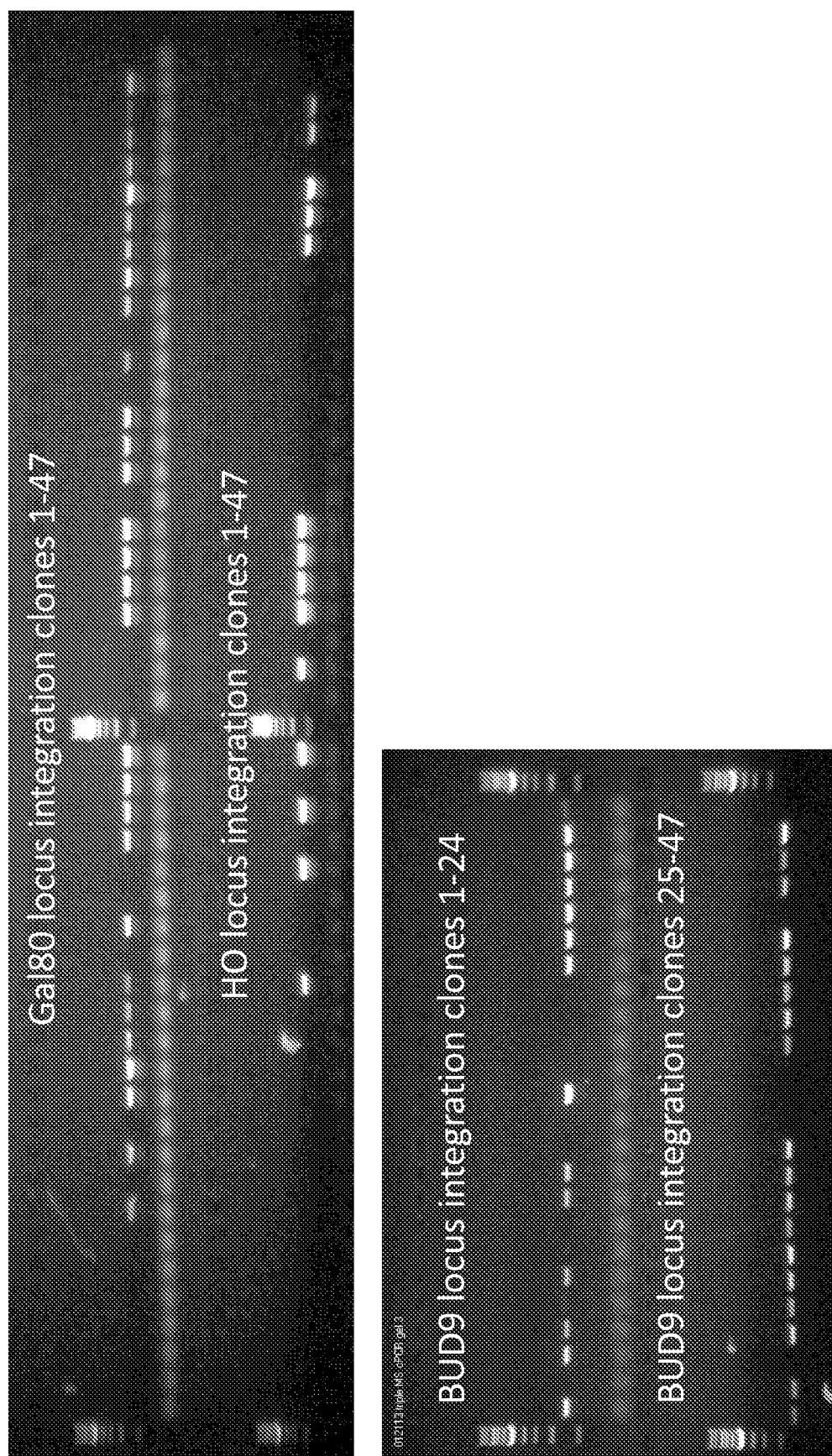


FIGURE 20

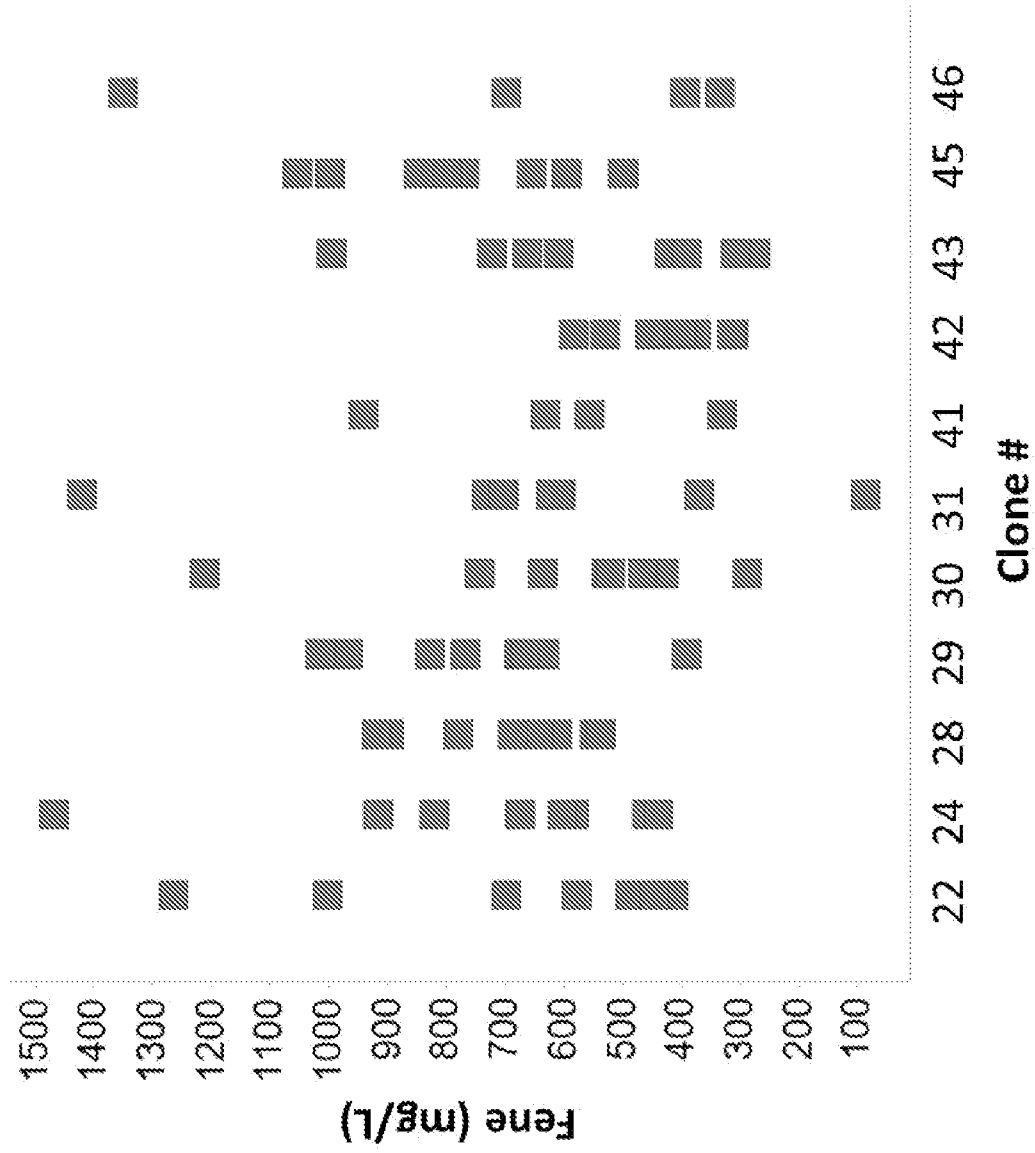


FIGURE 21

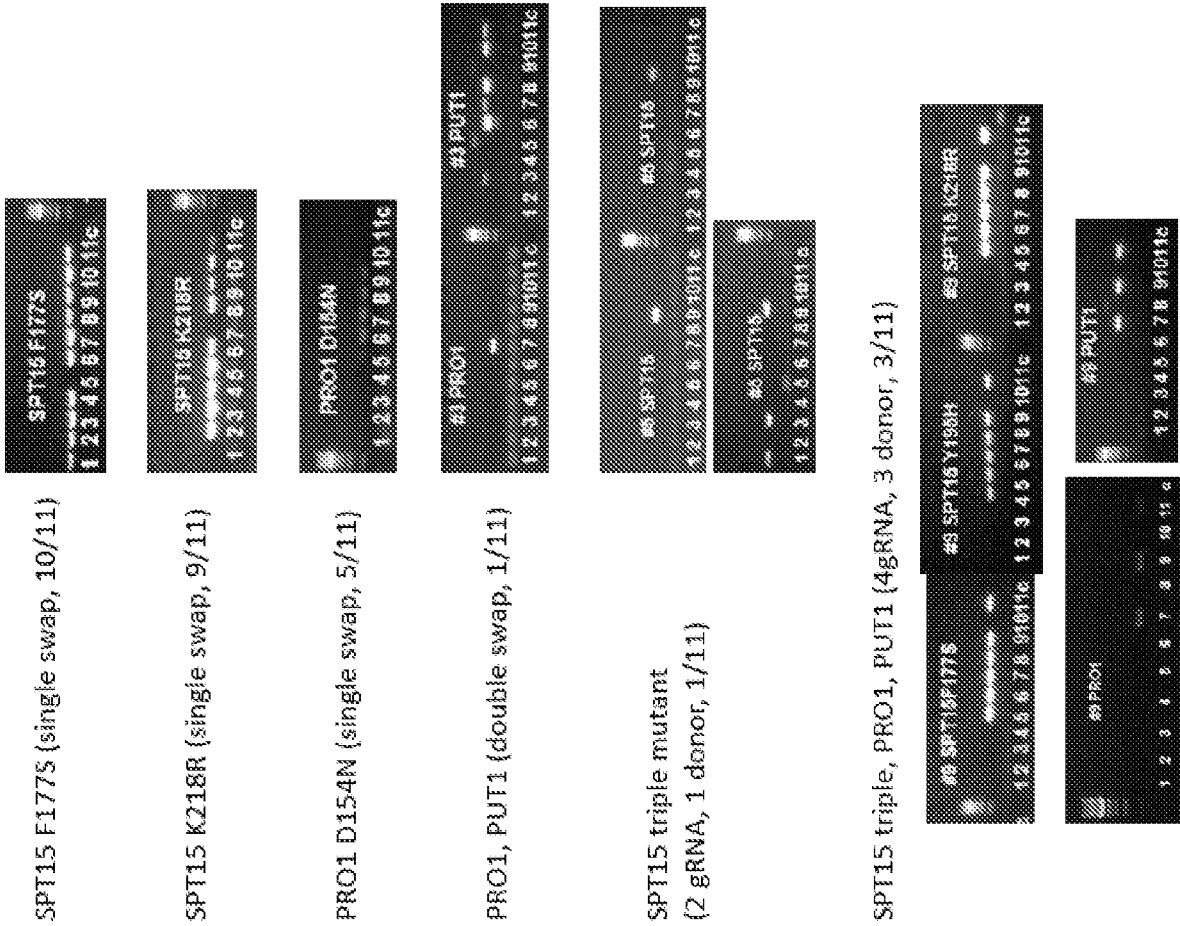


FIGURE 22

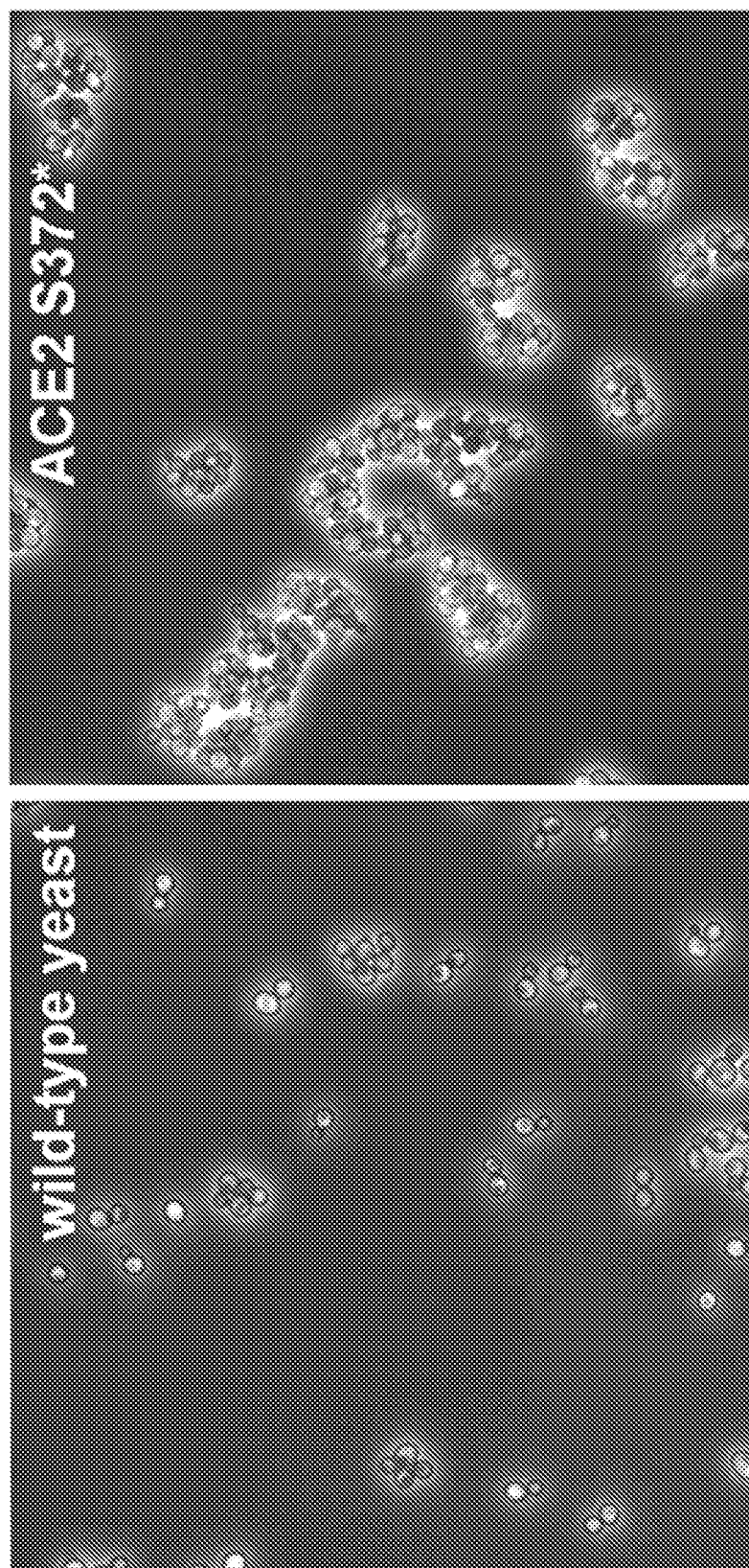


FIGURE 23A

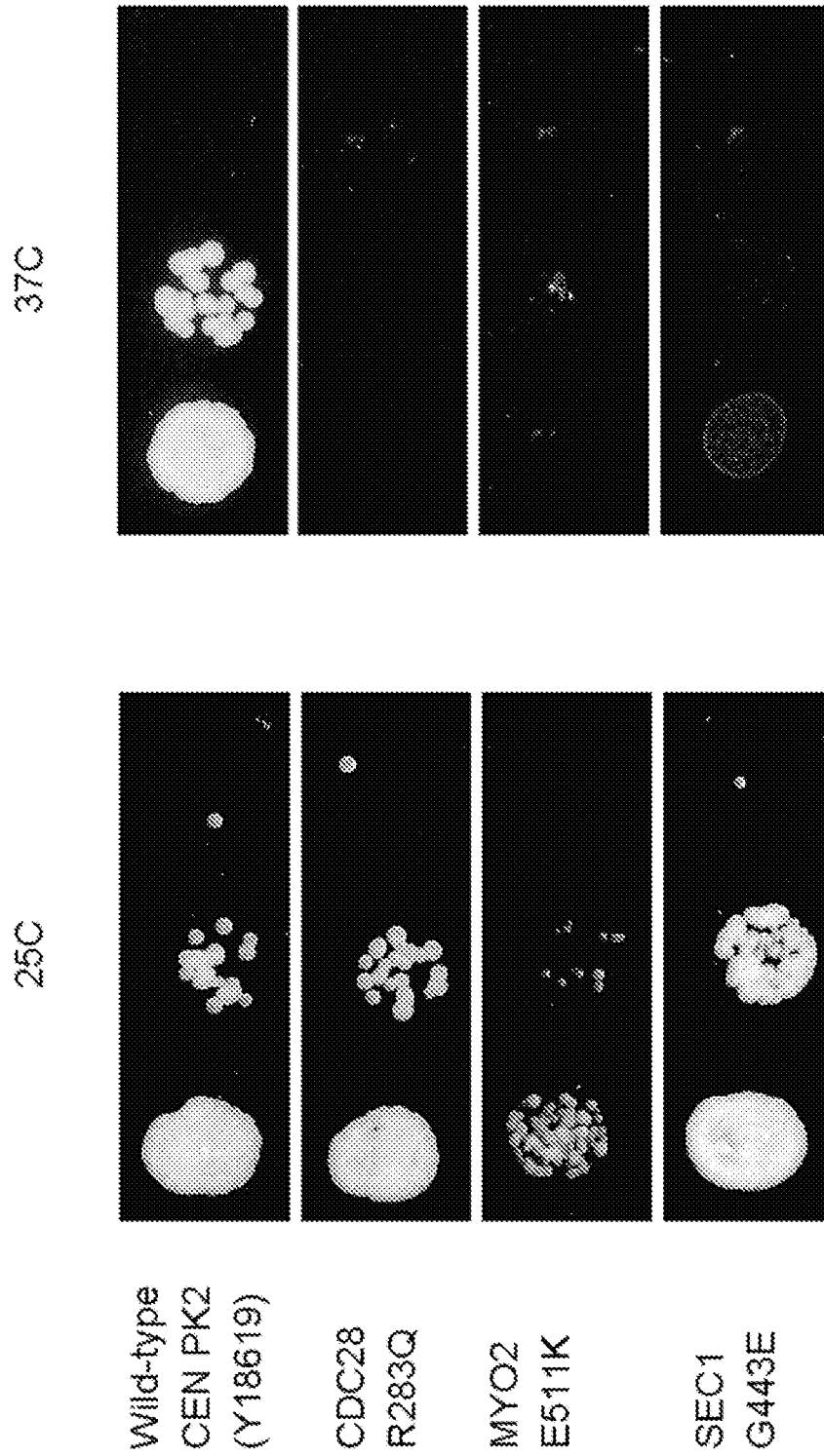


FIGURE 23B

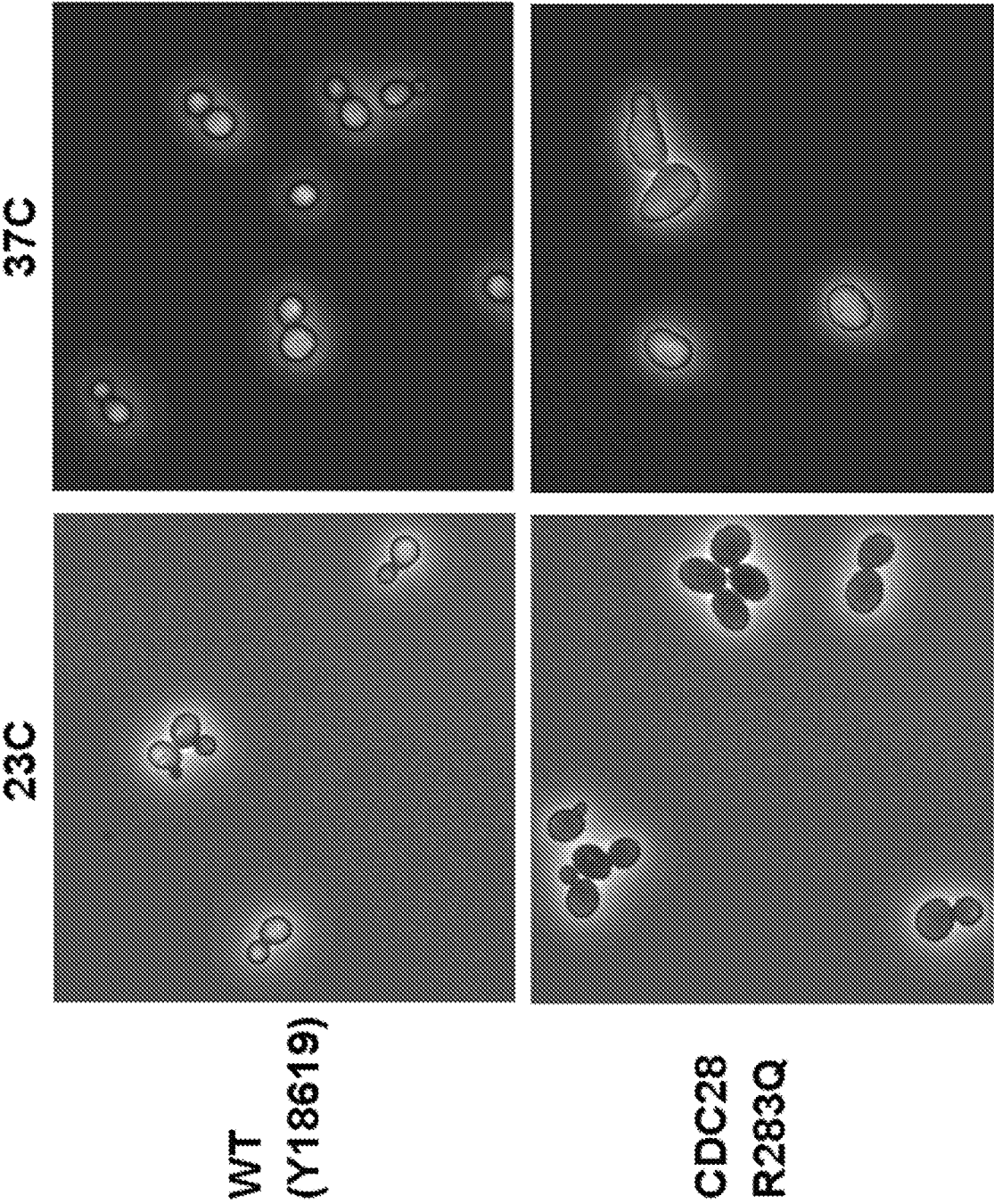


FIGURE 23C

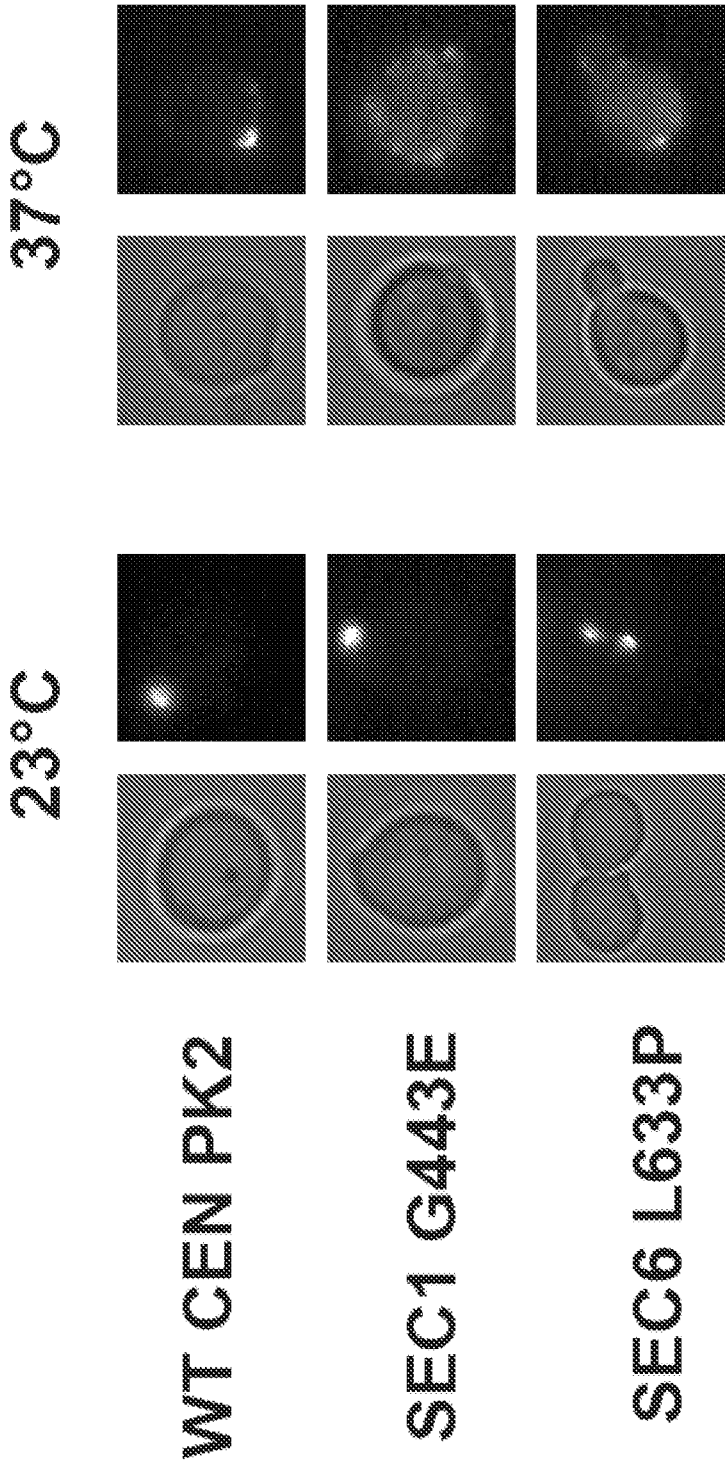


FIGURE 23D

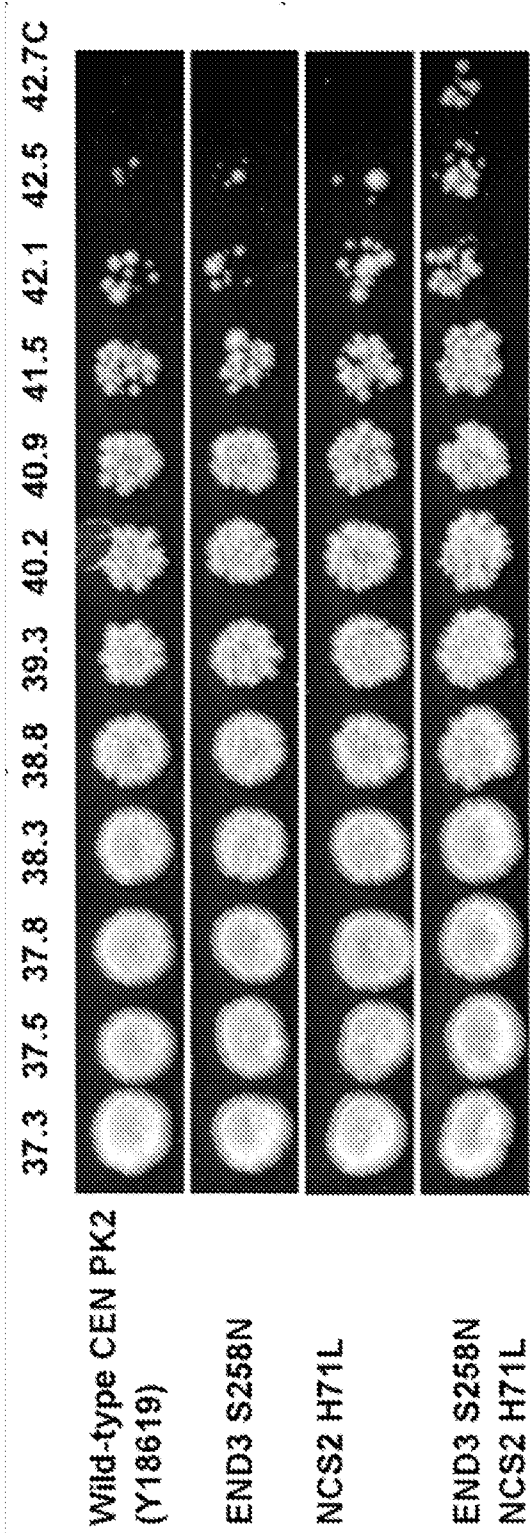


FIGURE 23E

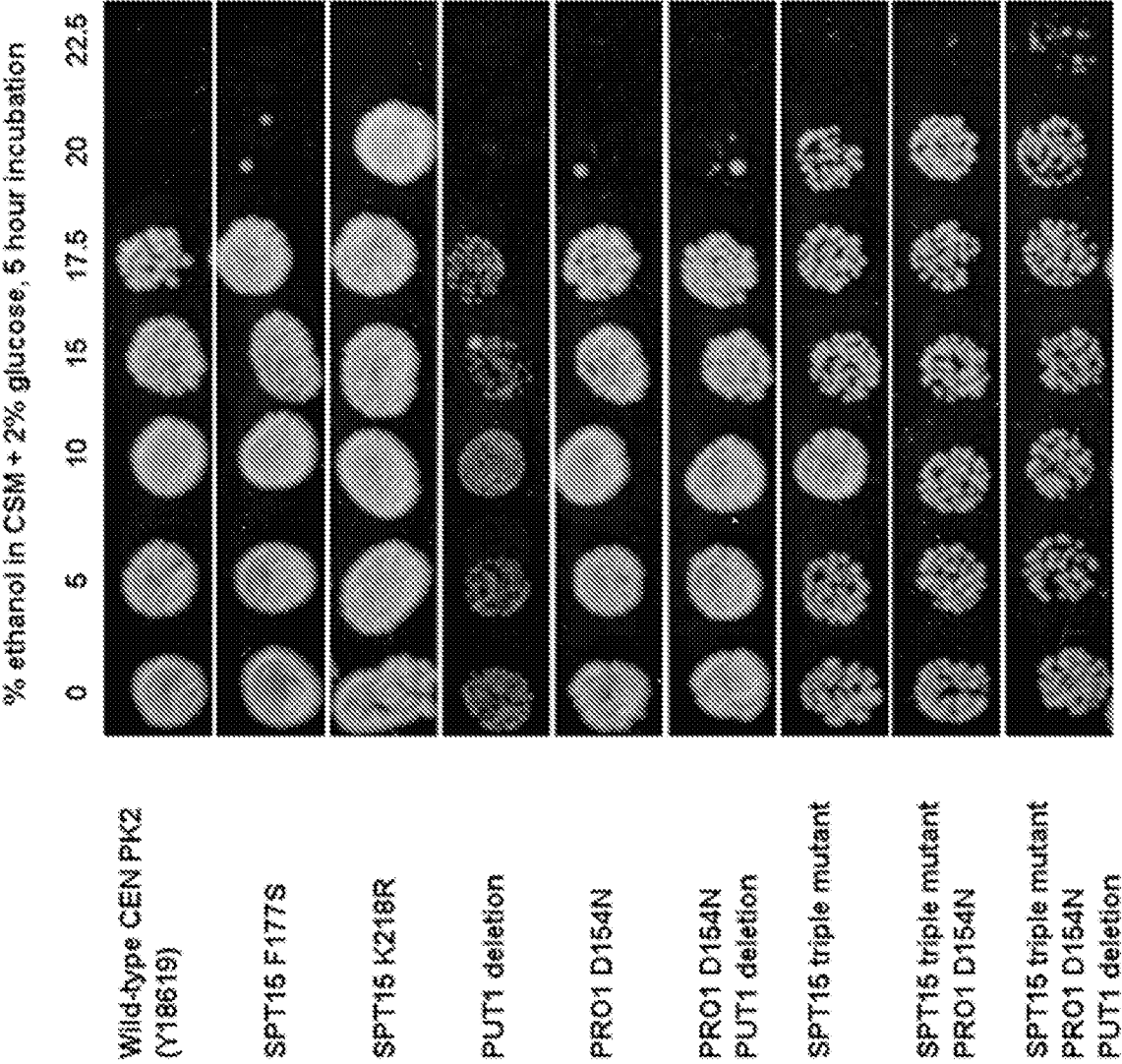


FIGURE 23F

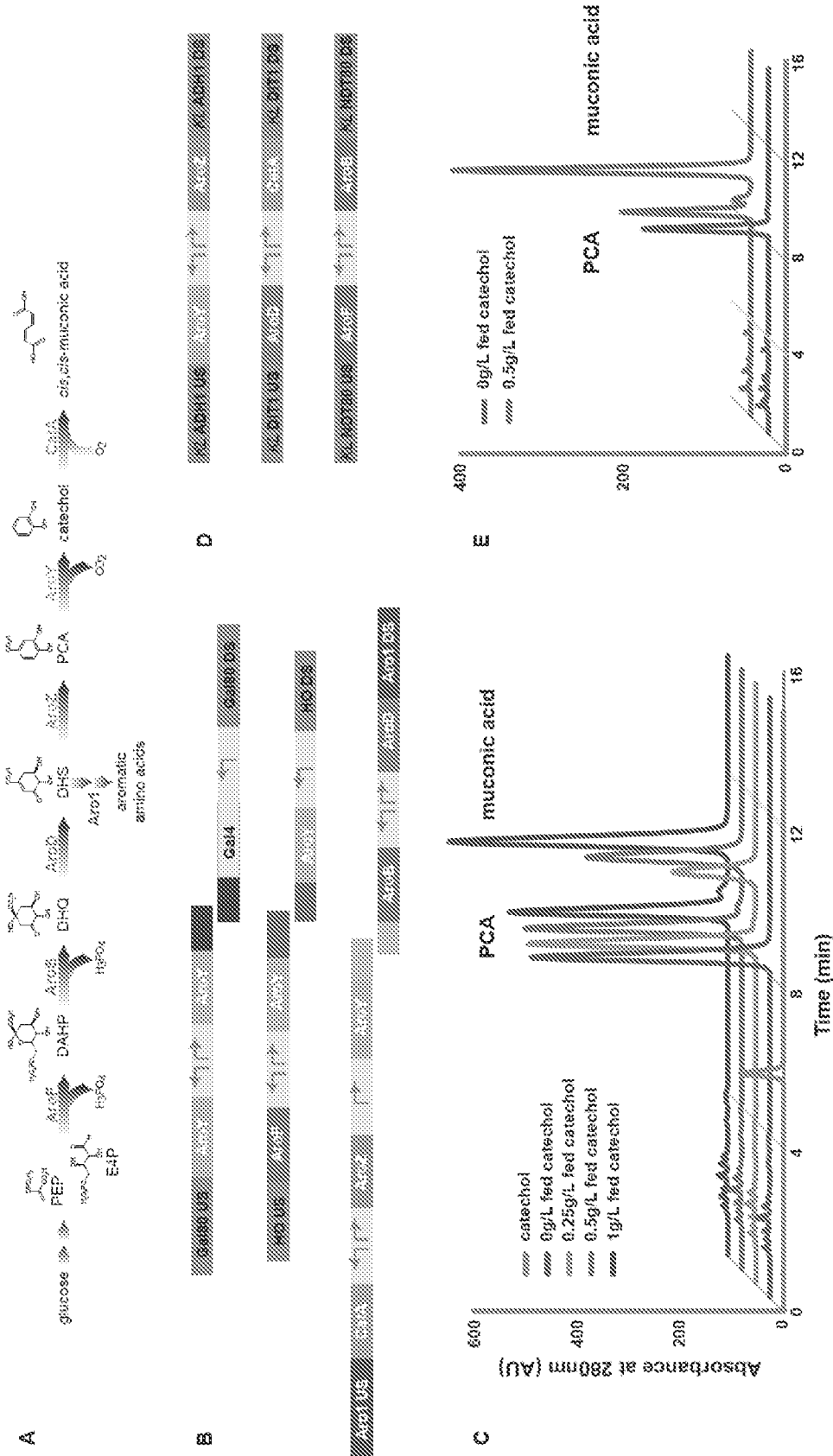


FIGURE 24

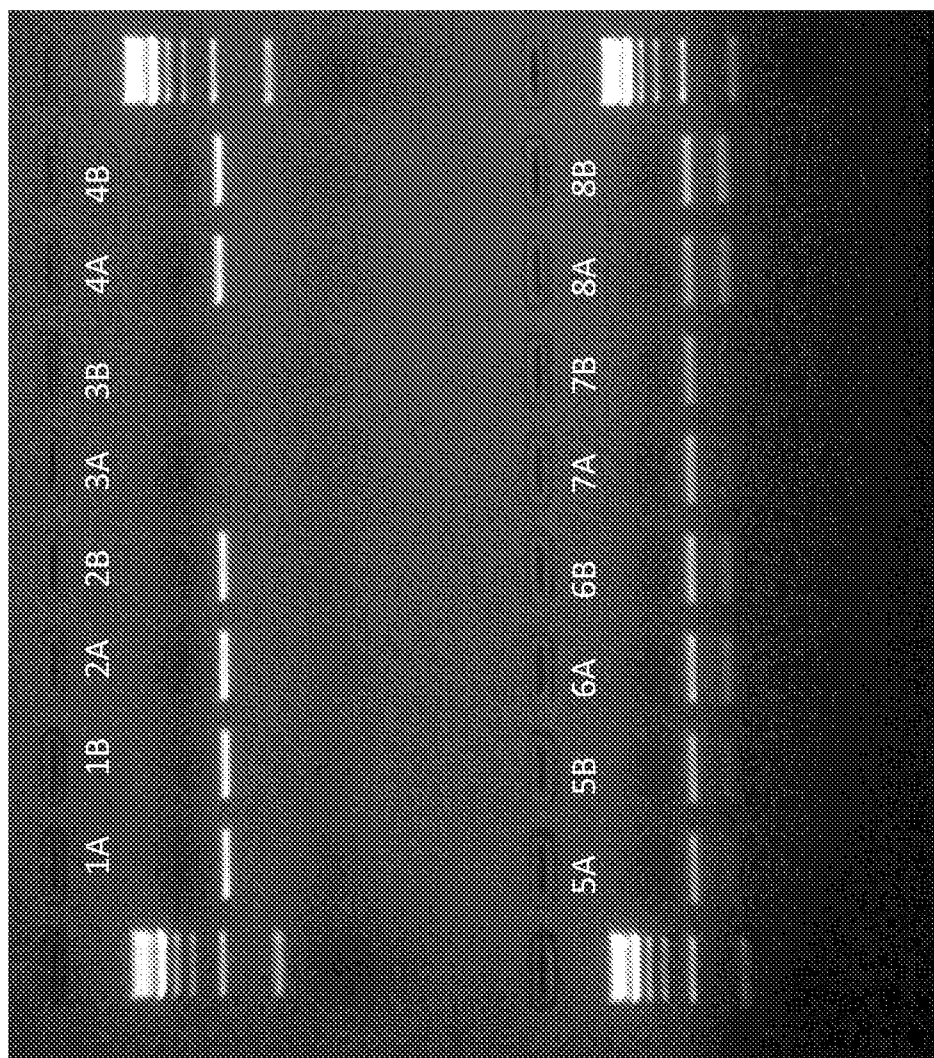
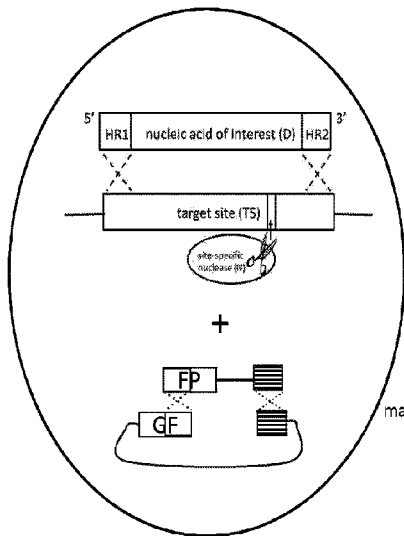
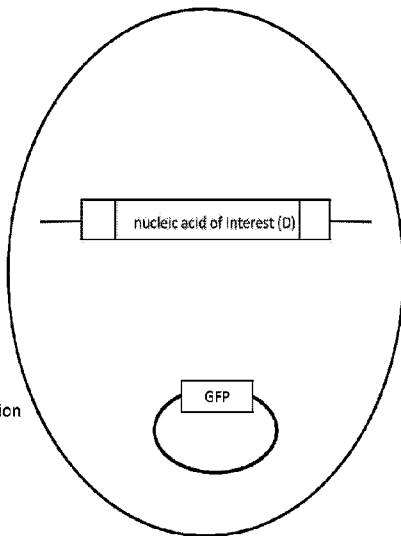


FIGURE 25



Select for
marker expression



Eliminate plasmid