



(12) **DEMANDE DE BREVET CANADIEN**
CANADIAN PATENT APPLICATION

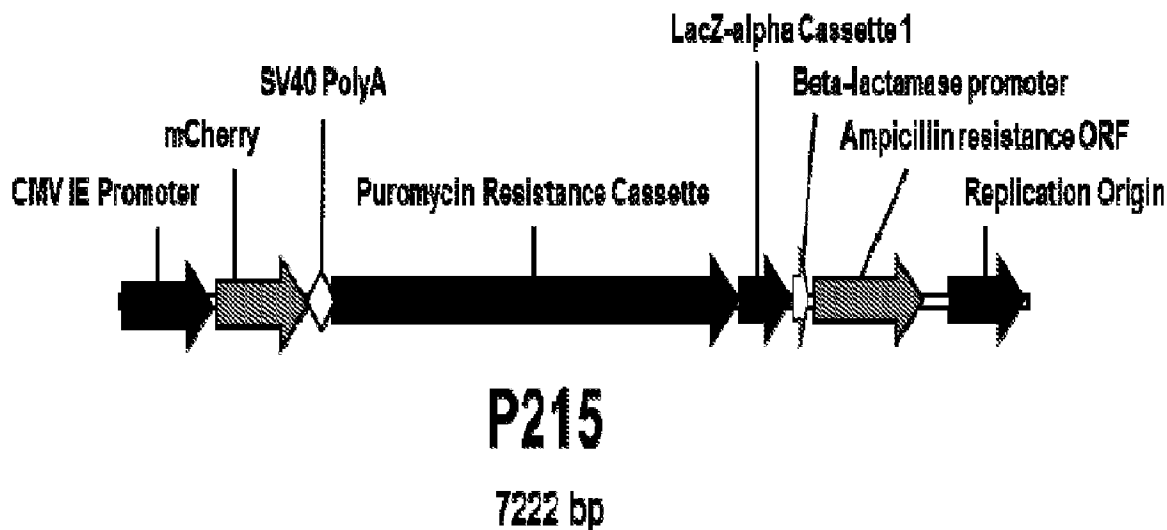
(13) **A1**

(86) Date de dépôt PCT/PCT Filing Date: 2020/01/14
(87) Date publication PCT/PCT Publication Date: 2020/07/23
(85) Entrée phase nationale/National Entry: 2021/07/16
(86) N° demande PCT/PCT Application No.: IB 2020/050267
(87) N° publication PCT/PCT Publication No.: 2020/148652
(30) Priorité/Priority: 2019/01/18 (US62/793,933)

(51) Cl.Int./Int.Cl. *C12N 15/63* (2006.01),
C12N 15/52 (2006.01), *C12N 15/72* (2006.01),
C12N 9/12 (2006.01), *C12N 9/38* (2006.01)
(71) Demandeur/Applicant:
JANSSEN BIOTECH, INC., US
(72) Inventeur/Inventor:
PERRY, WILLIAM, US
(74) Agent: NORTON ROSE FULBRIGHT CANADA
LLP/S.E.N.C.R.L., S.R.L.

(54) Titre : PEPTIDE ALPHA DE β -GALACTOSIDASE EN TANT QUE MARQUEUR DE SELECTION NON
ANTIBIOTIQUE ET UTILISATIONS CORRESPONDANTES
(54) Title: β -GALACTOSIDASE ALPHA PEPTIDE AS A NON-ANTIBIOTIC SELECTION MARKER AND USES THEREOF

FIG. 1



(57) Abrégé/Abstract:

Provided herein are methods of using a nucleic acid construct as a selectable marker. The nucleic acid construct comprises an isolated β -galactosidase expression cassette comprising a nucleic acid sequence encoding the amino-terminal fragment of β -galactosidase operably linked to a promoter. Also provided are isolated vectors comprising the β -galactosidase expression cassette, methods of generating the isolated vector, and kits comprising the isolated vector.



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

(19) World Intellectual Property

Organization

International Bureau

(43) International Publication Date

23 July 2020 (23.07.2020)



(10) International Publication Number

WO 2020/148652 A1

(51) International Patent Classification:

C12N 15/63 (2006.01) C12N 15/52 (2006.01)

C12N 9/12 (2006.01) C12N 15/72 (2006.01)

C12N 9/38 (2006.01)

(21) International Application Number:

PCT/IB2020/050267

(22) International Filing Date:

14 January 2020 (14.01.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/793,933 18 January 2019 (18.01.2019) US

(71) Applicant: JANSSEN BIOTECH, INC. [US/US];
800/850 Ridgeview Drive, Horsham, Pennsylvania 19044 (US).

(72) Inventor: PERRY, William; 1400 McKean Road, Spring House, Pennsylvania 19477 (US).

(74) Agent: SHIRTZ, Joseph F. et al.; Johnson & Johnson, One Johnson & Johnson Plaza, New Brunswick, New Jersey 08933 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

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

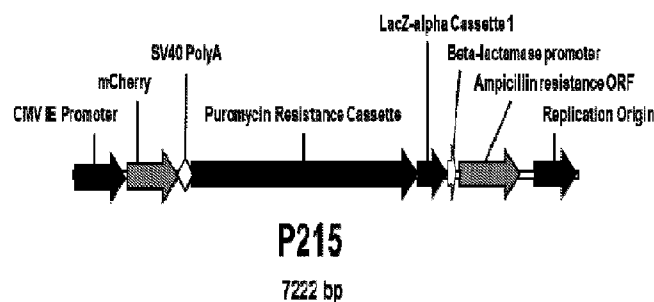
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, 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))

(54) Title: β -GALACTOSIDASE ALPHA PEPTIDE AS A NON-ANTIBIOTIC SELECTION MARKER AND USES THEREOF

FIG. 1



(57) Abstract: Provided herein are methods of using a nucleic acid construct as a selectable marker. The nucleic acid construct comprises an isolated β -galactosidase expression cassette comprising a nucleic acid sequence encoding the amino-terminal fragment of β -galactosidase operably linked to a promoter. Also provided are isolated vectors comprising the β -galactosidase expression cassette, methods of generating the isolated vector, and kits comprising the isolated vector.

[Continued on next page]



WO 2020/148652 A1

WO 2020/148652 A1

Published:

- *with international search report (Art. 21(3))*
- *with sequence listing part of description (Rule 5.2(a))*
- *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

β -GALACTOSIDASE ALPHA PEPTIDE AS A NON-ANTIBIOTIC SELECTION MARKER AND USES THEREOF

FIELD OF THE INVENTION

5 This invention relates to isolated β -galactosidase expression cassettes comprising a non-antibiotic selection marker. Specifically, the isolated β -galactosidase expression cassettes comprise the amino-terminal fragment of β -galactosidase operably linked to a promoter. Also provided are isolated vectors comprising the β -galactosidase expression cassettes, methods of producing the isolated vectors, and kits comprising the isolated
10 vectors.

REFERENCE TO SEQUENCE LISTING SUBMITTED ELECTRONICALLY

 This application contains a sequence listing, which is submitted electronically via EFS-Web as an ASCII formatted sequence listing with a file name
15 “JBI6031USPSP1Seq1” and a creation date of January 17, 2019 and having a size of 48 kb. The sequence listing submitted via EFS-Web is part of the specification and is herein incorporated by reference in its entirety.

BACKGROUND OF THE INVENTION

20 Plasmid vectors usually contain genes that are expressed in *E. coli* and provide a way to identify or select cells containing the plasmid from those which do not contain the plasmid when the plasmid is introduced into cells by transformation or electroporation. The most commonly used selectable markers are genes that confer resistance to antibiotics. However, there are several situations where antibiotic resistance genes are undesirable.
25 When plasmids are used to create manufacturing cell lines for biologics such as antibodies, the antibiotic resistance genes are usually removed or destroyed. For gene therapies, antibiotic resistance genes are also undesirable. While the kanamycin/neomycin resistance gene is often tolerated by the FDA, EU regulatory agencies are much stricter. The European Pharmacopei states “Unless otherwise justified and authorized, antibiotic
30 resistance genes used as selectable genetic markers, particularly for clinically useful

antibiotics, are not included in the vector construct. Other selection techniques for the recombinant plasmid are preferred” (“Gene transfer medical products for human use.” European Pharmacopei 7.0 (2011)). While destruction of the antibiotic selection marker may be possible when a small amount of the plasmid is needed for cell line development, these techniques are impractical for gene therapy applications where more of the plasmid needs to be manufactured.

Plasmid vectors where the replication origin and selection marker are a combined size of < 1 kb are needed for development of plasmid-based gene therapies to avoid gene silencing *in vivo*. Therapeutic transgenes were expressed longer and at higher levels in mice when the plasmid backbones were 1 kb or less compared to traditional plasmids with plasmid backbones 3 kb or more (Lu et al., Mol. Ther. 20(11):2111-9 (2012)). It was proposed that large blocks of DNA that were not expressed *in vivo* induced silencing. Thus, plasmids with smaller plasmid backbones might be much more efficacious.

Smaller plasmids are also needed for applications where transient transfection is used to manufacture therapeutics. One example is the production of Adeno-associated viral vectors where large-scale transfection of plasmids is used to generate clinical material. Smaller plasmids reduce the amount of DNA that must be transfected, reducing costs.

Thus, there is a need for generating smaller plasmids comprising a selectable marker that can be used for gene therapy applications.

BRIEF SUMMARY OF THE INVENTION

In one general aspect, provided are methods of using a nucleic acid construct as a selectable marker. The methods comprise (a) contacting a host cell comprising a deletion in a lac operon with the nucleic acid construct, wherein the nucleic acid construct comprises an isolated β -galactosidase expression cassette comprising a nucleic acid sequence encoding the amino-terminal fragment of β -galactosidase operably linked to a promoter; and (b) growing the host cell under conditions wherein the nucleic acid construct is maintained in the host cell.

In another general aspect, provided are isolated β -galactosidase expression cassettes. The isolated cassette comprises a nucleic acid sequence encoding the amino-terminal fragment of β -galactosidase operably linked to a promoter.

In certain embodiments, the amino-terminal fragment of β -galactosidase comprises
5 an amino acid sequence with at least 75% identity to SEQ ID NO:1. In certain
embodiments, the amino-terminal fragment of β -galactosidase comprises an amino acid
sequence of SEQ ID NO:1.

In certain embodiments, the nucleic acid sequence further comprises a replication
origin. The replication origin can, for example, be a high-copy replication origin. In
10 certain embodiments, the high-copy replication origin is the pUC57 replication origin. In
certain embodiments, the pUC57 replication origin comprises the nucleic acid sequence of
SEQ ID NO:19.

In certain embodiments, the isolated β -galactosidase expression cassette further
comprises a dimer resolution element. The dimer resolution element can, for example,
15 comprise a nucleic acid sequence comprising a site-specific recombinase recognition site.
The dimer resolution element can further comprise a nucleic acid sequence encoding a site
specific recombinase. In certain embodiments, the host cell comprises a nucleic acid
sequence encoding a site-specific recombinase. The dimer resolution element can, for
example, be a ColE1 dimer resolution element. In certain embodiments, the ColE1 dimer
20 resolution element comprises the nucleic acid sequence of SEQ ID NO:20.

Also provided are isolated vectors comprising the isolated β -galactosidase
expression cassettes of the invention. In certain embodiments, the isolated vector is less
than about 1.5 kilobases in size. In certain embodiments, the isolated vector comprises a
nucleic acid sequence selected from the group consisting of SEQ ID NOs:9-13, 17, and 18.

25 Also provided are methods of generating the isolated vectors of the invention. The
methods comprise (a) contacting a host cell with the isolated vector; (b) growing the host
cell under conditions to produce the vector; and (c) isolating the vector from the host cell.

In certain embodiments, the host cell is grown in minimal media. The minimal
media can comprise lactose as the sole carbon source. In certain embodiments, the

minimal media comprises about 1% to about 4% weight per volume (w/v) lactose. In certain embodiments, the minimal media comprises about 2% w/v lactose.

Also provided are kits comprising (a) an isolated β -galactosidase expression cassette of the invention; and (b) a host cell comprising a deletion in a lac operon. In certain embodiments, the kit further comprises minimal media comprising lactose as the sole carbon source. In certain embodiments, a vector comprises the isolated β -galactosidase expression cassette. In certain embodiments, the host cell comprises the LacZ Δ M15 deletion. In certain embodiments, the host cell is selected from the group consisting of an *E. coli* host cell and a yeast host cell.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing summary, as well as the following detailed description of preferred embodiments of the present application, will be better understood when read in conjunction with the appended drawings. It should be understood, however, that the application is not limited to the precise embodiments shown in the drawings.

FIG. 1 shows a schematic of the P215 plasmid.

FIG. 2 shows a schematic of the P216 plasmid.

FIG. 3 shows a schematic of the P217 plasmid.

FIG. 4 shows a schematic of the P218 plasmid.

FIG. 5 shows a schematic of the P219 plasmid.

FIG. 6 shows a schematic of the P469-2 plasmid.

DETAILED DESCRIPTION OF THE INVENTION

Various publications, articles and patents are cited or described in the background and throughout the specification; each of these references is herein incorporated by reference in its entirety. Discussion of documents, acts, materials, devices, articles or the like which has been included in the present specification is for the purpose of providing context for the invention. Such discussion is not an admission that any or all of these matters form part of the prior art with respect to any inventions disclosed or claimed.

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood to one of ordinary skill in the art to which this invention pertains. Otherwise, certain terms used herein have the meanings as set forth in the specification.

5 It must be noted that as used herein and in the appended claims, the singular forms “a,” “an,” and “the” include plural reference unless the context clearly dictates otherwise.

 Unless otherwise stated, any numerical values, such as a concentration or a concentration range described herein, are to be understood as being modified in all instances by the term “about.” Thus, a numerical value typically includes $\pm 10\%$ of the
10 recited value. For example, a concentration of 1 mg/mL includes 0.9 mg/mL to 1.1 mg/mL. Likewise, a concentration range of 1% to 10% (w/v) includes 0.9% (w/v) to 11% (w/v). As used herein, the use of a numerical range expressly includes all possible subranges, all individual numerical values within that range, including integers within such ranges and fractions of the values unless the context clearly indicates otherwise.

15 Unless otherwise indicated, the term “at least” preceding a series of elements is to be understood to refer to every element in the series. Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the invention.

20 As used herein, the terms “comprises,” “comprising,” “includes,” “including,” “has,” “having,” “contains” or “containing,” or any other variation thereof, will be understood to imply the inclusion of a stated integer or group of integers but not the exclusion of any other integer or group of integers and are intended to be non-exclusive or open-ended. For example, a composition, a mixture, a process, a method, an article, or an
25 apparatus that comprises a list of elements is not necessarily limited to only those elements but can include other elements not expressly listed or inherent to such composition, mixture, process, method, article, or apparatus. Further, unless expressly stated to the contrary, “or” refers to an inclusive or and not to an exclusive or. For example, a condition A or B is satisfied by any one of the following: A is true (or present) and B is false (or not
30 present), A is false (or not present) and B is true (or present), and both A and B are true (or present).

As used herein, the conjunctive term “and/or” between multiple recited elements is understood as encompassing both individual and combined options. For instance, where two elements are conjoined by “and/or,” a first option refers to the applicability of the first element without the second. A second option refers to the applicability of the second element without the first. A third option refers to the applicability of the first and second elements together. Any one of these options is understood to fall within the meaning, and therefore satisfy the requirement of the term “and/or” as used herein. Concurrent applicability of more than one of the options is also understood to fall within the meaning, and therefore satisfy the requirement of the term “and/or.”

As used herein, the term “consists of,” or variations such as “consist of” or “consisting of,” as used throughout the specification and claims, indicate the inclusion of any recited integer or group of integers, but that no additional integer or group of integers can be added to the specified method, structure, or composition.

As used herein, the term “consists essentially of,” or variations such as “consist essentially of” or “consisting essentially of,” as used throughout the specification and claims, indicate the inclusion of any recited integer or group of integers, and the optional inclusion of any recited integer or group of integers that do not materially change the basic or novel properties of the specified method, structure or composition. See M.P.E.P. § 2111.03.

It should also be understood that the terms “about,” “approximately,” “generally,” “substantially,” and like terms, used herein when referring to a dimension or characteristic of a component of the preferred invention, indicate that the described dimension/characteristic is not a strict boundary or parameter and does not exclude minor variations therefrom that are functionally the same or similar, as would be understood by one having ordinary skill in the art. At a minimum, such references that include a numerical parameter would include variations that, using mathematical and industrial principles accepted in the art (e.g., rounding, measurement or other systematic errors, manufacturing tolerances, etc.), would not vary the least significant digit.

The terms “identical” or percent “identity,” in the context of two or more nucleic acids or polypeptide sequences (e.g., amino-terminal β -galactosidase peptides and polynucleotides that encode them; nucleic acids of the isolated vectors described

herein), refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection.

5 For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test
10 sequence(s) relative to the reference sequence, based on the designated program parameters.

Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by
15 the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see generally, Current Protocols in Molecular Biology*, F.M. Ausubel *et al.*, eds., Current Protocols, a joint
20 venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (1995 Supplement) (Ausubel)).

Examples of algorithms that are suitable for determining percent sequence identity and sequence similarity are the BLAST and BLAST 2.0 algorithms, which are described in Altschul *et al.* (1990) *J. Mol. Biol.* 215: 403-410 and Altschul *et al.* (1997) *Nucleic
25 Acids Res.* 25: 3389-3402, respectively. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information. This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a
30 database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al., supra*). These initial neighborhood word hits act as seeds for initiating searches to find

longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased.

Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached.

The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff, *Proc. Natl. Acad. Sci. USA* 89:10915 (1989)).

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.,* Karlin & Altschul, *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001.

A further indication that two nucleic acid sequences or polypeptides are substantially identical is that the polypeptide encoded by the first nucleic acid is immunologically cross reactive with the polypeptide encoded by the second nucleic acid, as described below. Thus, a polypeptide is typically substantially identical to a second polypeptide, for example, where the two peptides differ only by conservative substitutions.

Another indication that two nucleic acid sequences are substantially identical is that the two molecules hybridize to each other under stringent conditions.

As used herein, the term “isolated” means a biological component (such as a nucleic acid, peptide, protein, or cell) has been substantially separated, produced apart from, or purified away from other biological components of the organism in which the component naturally occurs, i.e., other chromosomal and extrachromosomal DNA and RNA, proteins, cells, and tissues. Nucleic acids, peptides, proteins, and cells that have been “isolated” thus include nucleic acids, peptides, proteins, and cells purified by standard purification methods and purification methods described herein. “Isolated” nucleic acids, peptides, proteins, and cells can be part of a composition and still be isolated if the composition is not part of the native environment of the nucleic acid, peptide, protein, or cell. The term also embraces nucleic acids, peptides and proteins prepared by recombinant expression in a host cell as well as chemically synthesized nucleic acids.

As used herein, the term “polynucleotide,” synonymously referred to as “nucleic acid molecule,” “nucleotides” or “nucleic acids,” refers to any polyribonucleotide or polydeoxyribonucleotide, which can be unmodified RNA or DNA or modified RNA or DNA. “Polynucleotides” include, without limitation single- and double-stranded DNA, DNA that is a mixture of single- and double-stranded regions, single- and double-stranded RNA, and RNA that is mixture of single- and double-stranded regions, hybrid molecules comprising DNA and RNA that can be single-stranded or, more typically, double-stranded or a mixture of single- and double-stranded regions. In addition, “polynucleotide” refers to triple-stranded regions comprising RNA or DNA or both RNA and DNA. The term polynucleotide also includes DNAs or RNAs containing one or more modified bases and DNAs or RNAs with backbones modified for stability or for other reasons. “Modified” bases include, for example, tritylated bases and unusual bases such as inosine. A variety of modifications can be made to DNA and RNA; thus, “polynucleotide” embraces chemically, enzymatically or metabolically modified forms of polynucleotides as typically found in nature, as well as the chemical forms of DNA and RNA characteristic of viruses and cells. “Polynucleotide” also embraces relatively short nucleic acid chains, often referred to as oligonucleotides.

As used herein, the term “vector” is a replicon in which another nucleic acid segment can be operably inserted so as to bring about the replication or expression of the segment.

The term “expression” as used herein, refers to the biosynthesis of a gene product. The term encompasses the transcription of a gene into RNA. The term also encompasses translation of RNA into one or more polypeptides, and further encompasses all naturally occurring post-transcriptional and post-translational modifications. The expressed CAR
 5 can be within the cytoplasm of a host cell, into the extracellular milieu such as the growth medium of a cell culture, or anchored to the cell membrane.

The term “operatively linked” as used herein, refers to the linkage between nucleic acids (e.g., a promoter and a nucleic acid encoding a polypeptide) when it is placed into a structural or functional relationship. For example, one segment of a nucleic acid sequence
 10 can be operably linked to another segment of a nucleic acid sequence if they are positioned relative to one another on the same contiguous nucleic acid sequence and have a structural or functional relationship, such as a promoter or enhancer that is positioned relative to a coding sequence so as to facilitate transcription of the coding sequence; a ribosome binding site that is positioned relative to a coding sequence so as to facilitate translation; or
 15 a pre-sequence or secretory leader that is positioned relative to a coding sequence so as to facilitate expression of a pre-protein (e.g., a pre-protein that participates in the secretion of the encoded polypeptide). In other examples, the operably linked nucleic acid sequences are not contiguous, but are positioned in such a way that they have a functional relationship with each other as nucleic acids or as proteins that are expressed by them.
 20 Enhancers, for example, do not have to be contiguous. Linking can be accomplished by ligation at convenient restrictions sites or by using synthetic oligonucleotide adaptors or linkers.

The term “promoter” as used herein, refers to a nucleic acid sequence enabling the initiation of the transcription of a gene sequence in a messenger RNA, such transcription
 25 being initiated with the binding of an RNA polymerase on or nearby the promoter.

The term “replication origin” or “origin of replication” as used herein, refers to a nucleic acid sequence that is necessary for replication of a plasmid. Examples of replication origins include, but are not limited to, the pBR322 replication origin, the ColE1 replication origin, the pUC57 replication origin, a pMB1 replication origin, a pSC101
 30 replication origin, and a R6K gamma replication origin. Replication origins can be high- or low-copy. A high-copy replication origin, when present in a vector, can result in a high

number (e.g., 150 to 200) of copies of the vector per cell. A medium-copy replication origin, when present in a vector, can result in a medium number (e.g., 25 to 50) of copies of the vector per cell. A low-copy replication origin, when present in a vector, can result in a low number (e.g., 1 to 3) of copies of the vector per cell.

5 The term “dimer resolution element” as used herein, refers to a nucleic acid sequence that facilitates the *in vivo* conversion of multimers of the nucleic acid sequence (e.g., a vector or plasmid) to monomers in which said sequence is present. A dimer resolution element can comprise a nucleic acid sequence comprising a site-specific recombinase target site (e.g., a LoxP target site, a rfs target site, a FRT target site, a RP4
10 res target site, a RK2 res target site, and a res target site). A dimer resolution element can comprise a nucleic acid sequence encoding a site-specific recombinase (e.g., a Cre recombinase, a ResD recombinase, a Flp recombinase, a ParA recombinase, a Sin recombinase, a β recombinase, a $\gamma\delta$ recombinase, a tnpR recombinase, and a pSK41 resolvase). Dimers of isolated vectors/nucleic acids can be resolved by an enzyme acting
15 on the target DNA sequence comprised within the dimer resolution element. The enzyme recombines the target DNA sequence. By way of a non-limiting example, the enzymes XerC and XerD, expressed either by the host cell or the vector comprising the dimer resolution element, recognize the *cer* target site of the ColE1 dimer resolution element and work with several additional cofactors to ensure that a monomer of the vector/nucleic acid
20 is produced.

As used herein, the terms “peptide,” “polypeptide,” or “protein” can refer to a molecule comprised of amino acids and can be recognized as a protein by those of skill in the art. The conventional one-letter or three-letter code for amino acid residues is used herein. The terms “peptide,” “polypeptide,” and “protein” can be used interchangeably
25 herein to refer to polymers of amino acids of any length. The polymer can be linear or branched, it can comprise modified amino acids, and it can be interrupted by non-amino acids. The terms also encompass an amino acid polymer that has been modified naturally or by intervention; for example, disulfide bond formation, glycosylation, lipidation, acetylation, phosphorylation, or any other manipulation or modification, such as
30 conjugation with a labeling component. Also included within the definition are, for

example, polypeptides containing one or more analogs of an amino acid (including, for example, unnatural amino acids, etc.), as well as other modifications known in the art.

The peptide sequences described herein are written according to the usual convention whereby the N-terminal region of the peptide is on the left and the C-terminal
 5 region is on the right. Although isomeric forms of the amino acids are known, it is the L-form of the amino acid that is represented unless otherwise expressly indicated.

Polynucleotides, vectors, host cells, and methods of use

In one general aspect, provided are methods of using a nucleic acid construct as a selectable marker. The methods comprise (a) contacting a host cell comprising a deletion
 10 in a lac operon with the nucleic acid construct, wherein the nucleic acid construct comprises an isolated β -galactosidase expression cassette comprising a nucleic acid sequence encoding the amino-terminal fragment of β -galactosidase operably linked to a promoter; and (b) growing the host cell under conditions wherein the nucleic acid construct is maintained in the host cell.

15 In another general aspect, the invention relates to an isolated β -galactosidase expression cassette comprising a nucleic acid sequence encoding the amino-terminal fragment of β -galactosidase operably linked to a promoter.

In certain embodiments, the amino-terminal fragment of β -galactosidase comprises an amino acid sequence with at least 75% identity to SEQ ID NO:1. In certain
 20 embodiments, the amino-terminal fragment of β -galactosidase comprises an amino acid sequence with at least 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:1. The amino-terminal fragment of the β -galactosidase can comprise SEQ ID NO:1.

25 In certain embodiments, the nucleic acid sequence further comprises a replication origin. The replication origin can, for example, be a high-copy replication origin. In certain embodiments, the high-copy replication origin is the pUC57 replication origin. In certain embodiments, the pUC57 replication origin comprises a nucleic acid sequence with at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID

NO:19. In certain embodiments, the pUC57 replication origin comprises a nucleic acid sequence of SEQ ID NO:19.

In certain embodiments, the isolated β -galactosidase expression cassette can further comprise a dimer resolution element. The dimer resolution element can, for example, comprise a nucleic acid sequence comprising a site-specific recombinase recognition site. The site-specific recombinase recognition site can, for example, be selected from the group consisting of a LoxP site, a rfs site, a FRT site, a RP4 res site, a RK2 res site, and a res site. The dimer resolution element can further comprise a nucleic acid sequence encoding a site specific recombinase. In certain embodiments, the host cell comprises a nucleic acid sequence encoding a site-specific recombinase. The site-specific recombinase can, for example, be selected from the group consisting of a Cre recombinase, a ResD recombinase, a Flp recombinase, a ParA recombinase, a Sin recombinase, a β recombinase, a $\gamma\delta$ recombinase, a tnpR recombinase, and a pSK41 resolvase.

The dimer resolution element can, for example, be a ColE1 dimer resolution element. The ColE1 dimer resolution element can comprise a nucleic acid sequence with at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to SEQ ID NO:20. In certain embodiments, the ColE1 dimer resolution element comprises a nucleic acid sequence of SEQ ID NO:20.

In certain embodiments, an isolated vector comprises the isolated β -galactosidase expression cassettes of the invention. Any vector known to those skilled in the art in view of the present disclosure can be used, such as a plasmid, a cosmid, an artificial chromosome (e.g., a bacterial artificial chromosome (BAC), a yeast artificial chromosome (YAC), and/or a P1-derived artificial chromosome (PAC)), a transposon, a phage vector, or a viral vector. In some embodiments, the vector is a recombinant expression vector such as a plasmid. The vector can include any element to establish a conventional function of an expression vector, for example, a promoter, ribosome binding element, terminator, enhancer, selection marker, and origin of replication. The promoter can be a constitutive, inducible, or repressible promoter. A number of expression vectors capable of delivering nucleic acids to a cell are known in the art and can be used herein for the production of the amino-terminal fragment of the β -galactosidase peptide. Conventional cloning techniques

or artificial gene synthesis can be used to generate a recombinant expression vector according to embodiments of the invention.

In certain aspects, the isolated vector is less than about 1.5 kilobases in size. The isolated vector can, for example, be about 700 base pairs, about 800 base pairs, about 900
5 base pairs, about 1000 base pairs (about 1 kilobase), about 1100 base pairs (about 1.1 kilobases), about 1200 base pairs (about 1.2 kilobases), about 1300 base pairs (about 1.3 kilobases), about 1400 base pairs (about 1.4 kilobases), or about 1500 base pairs (about 1.5 kilobases) in length. In certain embodiments, the isolated vector is less than about 1 kilobase in size. In certain embodiments, the isolated vector is less than about 900 base
10 pairs in size. In certain embodiments, the isolated vector is less than about 800 base pairs in size.

In certain embodiments, the isolated vector comprises a nucleic acid sequence with at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to a nucleic acid selected from the group consisting of SEQ ID NOs:9-13, 17, and 18. In certain
15 embodiments, the isolated vector comprises a nucleic acid sequence selected from the group consisting of SEQ ID NOs:9-13, 17, and 18.

Also provided are methods of generating the isolated vector of the invention. The methods comprise (a) contacting a host cell with the isolated vector; (b) growing the host cell under conditions to produce the vector; and (c) isolating the vector from the host cell.

In certain embodiments, the host cell is grown in minimal media. The minimal media can comprise lactose as the sole carbon source. In certain embodiments, the minimal media comprises about 1% to about 4% weight per volume (w/v) lactose. In certain embodiments, the minimal media comprises about 1% to about 4% w/v, about 1%
20 to about 3% w/v, about 1% to about 2% w/v, about 1.5% to about 4% w/v, about 1.5% to about 3% w/v, about 1.5% to about 2% w/v, about 2% to about 4% w/v, about 2% to about 3% w/v, about 2.5% to about 4% w/v, about 2.5% to about 35% w/v, or about 3% to about 4% w/v lactose. In certain embodiments, the minimal media comprises about 2% w/v
25 lactose.

In certain embodiments, the invention relates to a host cell comprising an isolated
30 vector of the invention. Any host cell known to those skilled in the art in view of the present disclosure can be used for comprising an isolated vector of the invention. Suitable

host cells include cells with the LacZ Δ M15 deletion but with the rest of the lactose biosynthetic pathway intact. Strains that contain this mutation in the context of the bacteriophage Φ 80 integration (i.e., Φ 80lacZ Δ M15 marker) contain this mutation in the context of the complete lac operon, and, therefore, are suitable hosts. Other hosts with
5 different deletions in the amino-terminal (N-terminal) region of the LacZ gene, which produce significant levels of β -galactosidase when transformed with a LacZ- α complementation plasmid can also be suitable hosts. Suitable host cells of the invention can include an *E. coli* host cell or a yeast host cell.

Also provided are kits comprising (a) an isolated β -galactosidase expression
10 cassette of the invention; and (b) a host cell comprising a deletion in a lac operon. In certain embodiments, a vector comprises the isolated β -galactosidase expression cassette. In certain embodiments, the host cell comprises the LacZ Δ M15 deletion. In certain embodiments, the host cell can be selected from an *E. coli* host cell or a yeast host cell.

In certain embodiments, the kit further comprises minimal media comprising
15 lactose as the sole carbon source. In certain embodiments, the minimal media comprises about 1% to about 4% weight per volume (w/v) lactose. In certain embodiments, the minimal media comprises about 1% to about 4% w/v, about 1% to about 3% w/v, about 1% to about 2% w/v, about 1.5% to about 4% w/v, about 1.5% to about 3% w/v, about 1.5% to about 2% w/v, about 2% to about 4% w/v, about 2% to about 3% w/v, about 2.5%
20 to about 4% w/v, about 2.5% to about 35% w/v, or about 3% to about 4% w/v lactose. In certain embodiments, the minimal media comprises about 2% w/v lactose.

EMBODIMENTS

This invention provides the following non-limiting embodiments.

25 Embodiment 1 is a method of using a nucleic acid construct as a selectable marker, the method comprising:

- a. contacting a host cell comprising a deletion in a lac operon with the nucleic acid construct, wherein the nucleic acid construct comprises an isolated β -galactosidase expression cassette comprising a nucleic acid sequence encoding

the amino-terminal fragment of β -galactosidase operably linked to a promoter;
and

- b. growing the host cell under conditions wherein only the host cell containing the nucleic acid construct is maintained in the host cell.

5 Embodiment 2 is the method of embodiment 1, wherein the amino-terminal fragment of β -galactosidase comprises an amino acid sequence with at least 75% identity to SEQ ID NO:1.

Embodiment 3 is the method of embodiment 1 or 2, wherein the amino-terminal fragment of β -galactosidase comprises an amino acid sequence of SEQ ID NO:1.

10 Embodiment 4 is the method of any one of embodiments 1-3, wherein the nucleic acid sequence further comprises a replication origin.

Embodiment 5 is the method of embodiment 4, wherein the replication origin is a high-copy replication origin.

15 Embodiment 6 is the method of embodiment 5, wherein the high-copy replication origin is the pUC57 replication origin.

Embodiment 7 is the method of embodiment 6, wherein the pUC57 replication origin comprises the nucleic acid sequence of SEQ ID NO:19.

Embodiment 8 is the method of any one of embodiments 1-7, wherein the isolated β -galactosidase expression cassette further comprises a dimer resolution element.

20 Embodiment 9 is the method of embodiment 8, wherein the dimer resolution element comprises a nucleic acid sequence comprising a site-specific recombinase recognition site.

Embodiment 10 is the method of embodiment 8 or 9, wherein the dimer resolution element further comprises a nucleic acid sequence encoding a site-specific recombinase.

25 Embodiment 11 is the method of embodiment 8 or 9, wherein the host cell comprises a nucleic acid sequence encoding a site-specific recombinase.

Embodiment 12 is the method of any one of embodiments 8-11, wherein the dimer resolution element is a ColE1 dimer resolution element.

30 Embodiment 13 is the method of embodiment 12, wherein the ColE1 dimer resolution element comprises the nucleic acid sequence of SEQ ID NO:20.

Embodiment 14 is the method of any one of embodiments 1-13, wherein the host cell comprises a LacZΔM15 deletion.

Embodiment 15 is the method of any one of embodiments 1-14, wherein an isolated vector comprises the isolated β-galactosidase expression cassette.

5 Embodiment 16 is the method of embodiment 15, wherein the isolated vector is less than about 1.5 kilobases in size.

Embodiment 17 is the method of embodiment 15 or 16, wherein the isolated vector comprises a nucleic acid sequence selected from the group consisting of SEQ ID NOs:9-13, 17, and 18.

10 Embodiment 18 is a method of generating the isolated vector of any one of embodiments 15-17, wherein the method comprises:

- a. contacting a host cell with the isolated vector;
- b. growing the host cell under conditions to produce the vector;
- c. isolating the vector from the host cell.

15 Embodiment 19 is the method of embodiment 18, wherein the host cell is grown in minimal media.

Embodiment 20 is the method of embodiment 19, wherein the minimal media comprises lactose as the sole carbon source.

20 Embodiment 21 is the method of embodiment 20, wherein the minimal media comprises about 1% to about 4% weight per volume (w/v) lactose.

Embodiment 22 is the method of embodiment 21, wherein the minimal media comprises about 2% w/v lactose.

25 Embodiment 23 is an isolated β-galactosidase expression cassette comprising a nucleic acid sequence encoding the amino-terminal fragment of β-galactosidase operably linked to a promoter.

Embodiment 24 is the isolated β-galactosidase expression cassette of embodiment 23, wherein the amino-terminal fragment of β-galactosidase comprises an amino acid sequence with at least 75% identity to SEQ ID NO:1.

Embodiment 25 is the isolated β -galactosidase expression cassette of embodiment 23 or 24, wherein the amino-terminal fragment of β -galactosidase comprises an amino acid sequence of SEQ ID NO:1.

Embodiment 26 is the isolated β -galactosidase expression cassette of any one of
5 embodiments 23-25, wherein the nucleic acid sequence further comprises a replication origin.

Embodiment 27 is the isolated β -galactosidase expression cassette of embodiment 26, wherein the replication origin is a high-copy replication origin.

Embodiment 28 is the isolated β -galactosidase expression cassette of embodiment
10 27, wherein the high-copy replication origin is the pUC57 replication origin.

Embodiment 29 is the isolated β -galactosidase expression cassette of embodiment 28, wherein the pUC57 replication origin comprises the nucleic acid sequence of SEQ ID NO:19.

Embodiment 30 is the isolated β -galactosidase expression cassette of any one of
15 embodiments 23-29, wherein the isolated β -galactosidase expression cassette further comprises a dimer resolution element.

Embodiment 31 is the isolated β -galactosidase expression cassette of embodiment 30, wherein the dimer resolution element comprises a nucleic acid sequence comprising a site-specific recombinase recognition site.

Embodiment 32 is the isolated β -galactosidase expression cassette of embodiment
20 30 or 31, wherein the dimer resolution element further comprises a nucleic acid sequence encoding a site-specific recombinase.

Embodiment 33 is the isolated β -galactosidase expression cassette of any one of
25 embodiments 30-32, wherein the dimer resolution element is a ColE1 dimer resolution element.

Embodiment 34 is the isolated β -galactosidase expression cassette of embodiment 33, wherein the ColE1 dimer resolution element comprises the nucleic acid sequence of SEQ ID NO:20.

Embodiment 35 is an isolated vector comprising the isolated β -galactosidase
30 expression cassette of any one of embodiments 23-34.

Embodiment 36 is the isolated vector of embodiment 35, wherein the isolated vector is less than about 1.5 kilobases in size.

Embodiment 37 is the isolated vector of embodiment 35 or 36, wherein the isolated vector comprises a nucleic acid sequence selected from the group consisting of SEQ ID

5 NOs: 9-13, 17, and 18.

Embodiment 38 is a kit comprising:

- a. an isolated β -galactosidase expression cassette of any one of embodiments 23-37; and
- b. a host cell comprising a deletion in a lac operon.

10 Embodiment 39 is the kit of embodiment 38, further comprising minimal media comprising lactose as the sole carbon source.

Embodiment 40 is the kit of embodiment 38 or 39, wherein a vector comprises the isolated β -galactosidase expression cassette.

15 Embodiment 41 is the kit of any one of embodiments 38-40, wherein the host cell comprises the LacZ Δ M15 deletion.

Embodiment 42 is the kit of embodiment 41, wherein the host cell is selected from the group consisting of an *E. coli* host cell and a yeast host cell.

EXAMPLES

20 **Example 1: Plasmid selection via alpha-complementation of β -galactosidase instead of antibiotic selection in TOP10 cells**

Materials

Cells: One Shot Top10 competent cells (Thermo-Fisher; Waltham, MA, Catalog Number C404003). NEB 5-alpha (New England Biolabs, Ipswich, MA, Catalog Number (C2987).

25 GT115 (InvivoGen, San Diego, CA, Catalog Number GT115-21). NEB Stable (New England Biolabs, Catalog Number C3040H). Stellar (Takara Bio USA, Mountain View, CA, Catalog Number 636766). DH10B (Thermo-Fisher, Catalog Number 18297010). Stbl3 (Thermo-Fisher, Catalog Number C737303). Xli-blue (Agilent, Santa Clara, CA; Catalog Number 200236).

Plasmids: pUC19 (Thermo-Fisher Scientific; Catalog Number SD0061); pBluescript II. KS(-) (Agilent; Santa Clara, CA; Catalog Number 212208). Clones P215 (SEQ ID NO:9) and P216 (SEQ ID NO:10). GWIZ-Luciferase (Genlantis Corporation; San Diego, CA; P030200); P219 (SEQ ID NO:13; FIG. 5). P469-2 (SEQ ID NO:17; FIG. 6).

- 5 Media: M9 + Lactose Media (Teknova, Hollister CA; Catalog Number M1348-04 (plates)): 0.3% KH₂PO₄, 0.6% Na₂HPO₄, 0.5% (85mM) NaCl, 0.1% NH₄Cl, 2mM MgSO₄, 50 mg/liter L-leucine, 50 mg/L isoleucine; 1 mM thiamine, 2% lactose, and 1.5% agar. M9 + Glucose Media (Teknova Hollister CA; Catalog Number M1346-04 (plates)): 0.3% KH₂PO₄, 0.6% Na₂HPO₄, 0.5% (85mM) NaCl, 0.1% NH₄Cl, 2 mM MgSO₄, 50 mg/liter
- 10 L-leucine, 50 mg/liter isoleucine, 1 mM thiamine, 1% glucose, and 1.5% agar. LB-Carbenicillin(100) plates (Teknova, Hollister CA; Catalog number L1010). LB Plates (Teknova Hollister CA L1100). LB + 60µg/mL X-Gal, 0.1mM IPTG (Teknova Hollister CA L1920). SOC Media (Thermo-Fisher 15544034). LB Broth (Thermo-Fisher 10855021);
- 15 D-PBS, pH 7.1, no Mg²⁺ noCa²⁺ (ThermoFisher 14200-075)

Results

- Plasmids without antibiotic selection markers are desirable for gene therapy applications and cell line development for therapeutic products. It has also been reported that plasmid backbones 1 kb or smaller were useful in avoiding gene silencing when
- 20 delivered to animals *in vivo*. The purpose of these experiments was to explore a new strategy for developing a small metabolic selection marker for selection of plasmid-containing cells in *E. coli*.

- It was hypothesized that plasmids that express the alpha peptide of β -galactosidase could complement the LacZ Δ 15 allele in TOP10 cells, completing the lactose operon and
- 25 allowing cells to grow on minimal media with lactose as the sole carbon source. Plasmids pUC19 and pBluescript II both express β -galactosidase alpha peptide fusion proteins. Whether these plasmids were able to complement lac mutations in the Top10 host strain and allow growth on minimal media was tested.

- To test whether pUC19 and/or pBluescript II were capable of complementing the
- 30 LacZ Δ 15 mutations in TOP10 cells, these plasmids were transformed into the cells using the following procedure.

Two transformation mixtures were prepared in sterile microfuge tubes as follows:
 1) 1 μ l (100 pg) pBluescript II plasmid + 50 μ l One Shot TOP10 cells; 2) 1 μ l (10 pg)
 pUC19 plasmid + 50 μ l One Shot TOP10 cells. The transformation mixtures were
 incubated on ice 30 minutes, then heat shocked for 30 seconds at 42°C. After the heat
 5 shock, the transformation mixtures were incubated on ice for 1 minute. To the
 transformation mixtures, 450 μ l SOC media was added, and the cells were incubated
 shaking at 37°C for 1 hour. The transformation mixtures containing the cells were
 centrifuged, and the cells were resuspended in 500 μ l Sterile D-PBS buffer. The cells were
 centrifuged and resuspended twice more. Two 1:10 serial dilutions of the cells were made
 10 in D-PBS for each sample. 200 μ l of each dilution was spread onto M9 + Lactose plates.
 200 μ l of the first two dilutions were also spread onto LB-Carbenicillin (100) plates. The
 plates were incubated at 37°C overnight.

After overnight incubation there were many colonies from both transformations
 plated onto LB-Carbenicillin (100) plates; these plates were stored at 4°C. There were no
 15 visible colonies from either transformation plated onto M9 + Lactose plates; these plates
 were incubated for an additional 24 hours at 37°C. No colonies were visible on the M9-
 Lactose plates. Cells were cultured for an additional 48 hours at 30°C. No colonies were
 visible on these plates, even after extended incubation.

Neither of the cloning vectors expressing LacZ- α fusion peptides were able to
 20 complement the Lac mutation in the TOP10 host strain to allow growth in minimal media
 containing lactose as the sole carbon source.

It was possible that the expression of LacZ- α peptide fusion proteins by the pUC19
 and pBluescript II cloning vectors was not high enough to adequately complement the lac
 mutations in the host strains tested. Both vectors produce fusion proteins that transcribe
 25 through the multi-cloning region and such fusion proteins could be sub-optimal for
 complementing the LacZ Δ 15 mutation.

Example 2: LacZ expressing plasmids used as a metabolic selection marker in *E. coli*.

Two LacZ-alpha expression cassettes with medium and strong promoters (LacZYA
 30 and OmpF, respectively) were designed. The OmpF promoter sequence was based on the
 OmpF promoter used by Stavropoulos et al. (Stavropoulos and Strathdee, Genomics

72(1):99-104 (2001)). The LacZ α promoter was derived from the sequence in pBluescript along with the lac operator sequence bound by the lac repressor.

For the open reading frame (ORF) of the LacZ alpha peptide, Reddy (Reddy, Biotechniques 37(6):948-52 (2004) reported that the plasmid pUC19 produced about 10x more beta-galactosidase activity than pBluescript. These plasmids have the same promoter elements driving the lacZ alpha peptide. However, pBluescript has a much longer polylinker than pUC19 and pUC19 encodes non-lacZ C-terminal residues. It is unknown which of these differences result in higher pUC19 beta-galactosidase activity. Nishiyama et al found that the N-terminal alpha peptides of 60 amino acids had maximal β -galactosidase activity in their assay (Nishiyama et al., Protein Sci. 24(5):599-603 (2015)). The following wild type LacZ alpha region from strain MG1655 truncated at residue 60 was used: MTMITDSLAVVLQRRDWENPGVTQLNRLAAHPPFASWRNSEEARTDRPSQQLRSLNGEWR (SEQ ID NO:1).

The terminator sequence was derived from the rrnBT2 terminator described by Orosz et al. (Orosz et al., Eur. J. Biochem. 201(3):653-9 (1991)).

The P215 (SEQ ID NO:9) (FIG. 1) and P216 (SEQ ID NO:10) (FIG. 2) plasmids were constructed by gene synthesis at GeneWiz (South Plainfield, NJ). The plasmids contain an ampicillin resistance cassette and a 4.9 kb transgene.

Results

Plasmids without antibiotic selection markers are desirable for gene therapy applications and cell line development for therapeutic products. It has also been reported that plasmid backbones 1 kb or smaller were useful in avoiding gene silencing when delivered to animals *in vivo*. The purpose of these experiments was to explore a new strategy for developing a small metabolic selection marker for selection of plasmid-containing cells in *E. coli*.

It was hypothesized that plasmids that express the alpha peptide of β -galactosidase could complement the LacZ Δ 15 allele in Top10 cells, completing the lactose operon and allowing cells to grow on minimal media with lactose as the sole carbon source.

In Example 1, whether pUC19 and pBluescript vectors that express lacZ α fusion peptides could complement TOP10 cells and allow them to grow on minimal media with lactose was tested. These experiments were unsuccessful.

Based on the hypothesis that the lacZ α fusion proteins encoded by these vectors were suboptimal at complementing the LacZ Δ 15 mutation and were not expressed at high enough levels to enable growth on Lactose-containing minimal media, vectors were synthesized with new lacZ α expression cassettes. The ability of these vectors to

5 complement the LacZ Δ 15 mutation was tested. Ten nanograms (ng) of plasmids P215 and P216, and pBluescript II were transformed into 50 μ l OneShot Top10 cells. The cells were incubated with DNA on ice for 20 minutes, heat shocked at 42°C for 30 seconds, and returned to incubate on ice for 1 minute. 450 μ l of SOC was added to the cells, and the cells were incubated at 37°C for 1 hour while shaking. 250 μ l of cells were removed and

10 the remaining cells were returned to the incubator. The extracted cells were washed two times with 500 μ l of D-PBS and resuspended in 200 μ l of D-PBS after the last wash. 50 μ l of cells were plated on LB-carbenicillin (100), M9 + glucose, and M9 + lactose plates, and the plates were incubated at 37°C. After 4.5 hours post heat shock, the remaining cells from the incubator were washed, as described above, and plated onto M9 + glucose and

15 M9 + lactose plates. The plates were incubated at 37°C overnight.

Transformations plated on M9 + glucose made a lawn of cells, indicating that Top10 host cells can grow on these plates. Transformations plated on LB-carbenicillin (100) produced lots of colonies as well. The LB-carbenicillin plates were stored at 4°C. The M9 + lactose plates remained at 37°C to incubate for 24 more hours.

20 Transformations allowed to recover for either one hour or for four hours both produced a large number of colonies when plated on the M9 + lactose plates. There were no colonies on the pBluescript II transformations confirming the results from Example 1, indicating that pBluescript II was unable to produce enough β -galactosidase through complementation of the LacZ Δ 15 mutation to allow growth on lactose minimal media.

25 The plates were stored at 4°C.

Natural plasmids such as ColE1 are efficiently maintained in *E. coli* hosts in the absence of antibiotic selection while the pUC series of vectors can be lost from cells at a high rate in the absence of selection (Summers, Molecular Microbiology 29: 1137-1145 (1998)). However, given the much slower growth rate of P215 and P216-transformed cells

30 on minimal media versus rich LB media, it would be much faster and cheaper for plasmid

DNA purification to grow cell cultures in LB in the absence of selection if the frequency of plasmid loss was not too high. β -galactosidase alpha-complementation plasmid-containing cells are easily distinguished from plasmid-free cells grown on LB-IPTG-XGAL plates since the β -galactosidase hydrolyzes the XGAL (5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside) indicator turning the cells blue. This assay was used to investigate the frequency of plasmid loss when these cells are grown in the absence of antibiotics in LB media.

Pure populations of cells were obtained by streaking cells on LB-IPTG-XGAL plates, and colonies that contained plasmids turned blue. Most of the colonies streaked on the plates were blue, as expected.

After obtaining a pure population of cells, serial cultures of the cells were grown. A single blue colony was picked and grown in 2 mls of LB media in a 15 ml tube. The culture was incubated overnight at 37°C while shaking.

Cells from the cultures were streaked onto LB-IPTG-XGAL plates, and the plates were incubated overnight at 37°C. Colonies on the re-streaked plates were blue. A single colony was inoculated in 50 mls of LB in a 250 ml flask and incubated overnight at 37°C while shaking.

50 μ l of a 10^{-4} dilution of the overnight cultures were plated onto LB-IPTG-XGAL plates. The plates were incubated overnight at 37°C. 1 μ l of the 50 ml cultures was diluted to a new culture of 50 mls of LB (50,000-fold dilution). The cultures were grown overnight at 37°C.

After incubation overnight, all colonies on the plate were observed to be blue. 50 μ l of a 10^{-4} dilution of the 50 ml culture from the previous night were plated on LB-IPTG-XGAL plate. 1 μ l of the 50 ml cultures from the previous night was diluted to a new culture of 50 mls of LB (50,000-fold dilution). The cultures were grown overnight at 37°C.

After incubation overnight, there were about 1000 colonies observed on the plates with 50 μ l of the 10^{-4} dilution. All of the colonies of the P215 transformation were blue, and there were only 3 white colonies observed on the P216 transformation plate.

The results indicated that plasmids P215 and P216 were stable even in the absence of selection. These plasmids are 7.2 and 7.3 kb for P215 and P216, respectively. From a single colony to 50 mls and then diluted 1:50,000 and grown to confluence twice suggests that the cells could be grown to a volume of 1.25×10^8 liters without selection while still retaining the plasmid in most of the cells. The transformation efficiency was similar when cells were allowed to recover for one hour versus four hours in SOC media post-heat shock.

The alpha complementation plasmids constructed complemented the LacZ Δ 15 mutation in Top10 cells allowing growth on minimal media with lactose as the sole carbon source. These plasmids were also found to be stable in LB liquid cultures in the absence of selective pressure.

Example 3: Reducing the size of β -galactosidase- α complementation plasmids

In previous experiments, expression of the β -galactosidase alpha peptide from the P215 and P216 plasmids was demonstrated to be useful as selection marker on plasmids, replacing antibiotic resistance genes. Next it was sought to define which regions of the plasmids were essential for plasmid selection and replication in *E. coli* with the goal of defining the smallest possible replicon.

Results

Using standard cloning techniques, the mCherry and puromycin resistance genes were removed from plasmid P215 to create plasmid P217 (SEQ ID NO:11) (FIG. 3).

From plasmid P217, standard cloning techniques were used to remove the ampicillin resistance gene. Ligated DNA was transformed into 50 μ l of TOP10 cells, incubated on ice for 20 minutes, heat shocked for 30 seconds, and incubated on ice for an additional 3 minutes. After incubation, 450 μ l of SOC media was added to the cells, and the cells were incubated at 37°C for 1 hour while shaking. The cells were pelleted and washed 3 times with 1 ml of d-PBS. Cells were plated onto M9-lactose plates and incubated at 37°C for two days. Colonies from the transformation were picked and streaked onto an LB-IPTG-XGAL plate. The resulting colonies were blue for each clone. A single clone was picked (Clone P218 (SEQ ID NO:12; FIG. 4)), and DNA sequencing confirmed that the desired deletion had been created.

To further decrease the size of the β -galactosidase selection cassette, the rrnBT2 transcription terminator (SEQ ID NO:7) was deleted. In addition to the possibility that this sequence was not necessary to maintain transcript stability, it was reported that read-through transcription from promoters upstream of the pUC57/pMB1 origin can
5 increase copy number by increasing transcription through the replication primer region of the origin (Panayotatos, Nucleic Acid Res. 12(6):2641-8 (1984); Oka et al., Mol Gen Genet. 172(2):151-9 (1979)).

Using standard cloning techniques, colonies were obtained for the deletion construct P219 (SEQ ID NO:13; FIG. 5). The deletion was confirmed through DNA
10 sequencing.

The minimal β -galactosidase expression cassette/replication origin cassette that was elucidated by this work (SEQ ID NO:18) is 938 bp. It fulfills the goal of being smaller than 1 kb in order to avoid DNA silencing in mammalian cells associated with larger plasmid backbones (Lu et al., Mol. Ther. 20(11):2111-9 (2012)).

15 **Example 4: Creation of β -galactosidase- α complementation vector with Firefly Luciferase expression cassette**

In the examples provided above, plasmids that use alpha complementation of a β -galactosidase mutation as a selection marker instead of an antibiotic resistance gene were constructed. To determine whether DNA replication was still efficient when the plasmid
20 size increases, the minimal β -galactosidase expression cassette/replication origin sequence defined above (SEQ ID NO:18) was used to replace the antibiotic selection marker and replication origin of an existing plasmid using standard cloning techniques.

The CMV promoter-luciferase-polyA expression cassette from the GWIZ-
25 Luciferase plasmid (SEQ ID NO:16) was cloned into P219 using standard cloning techniques. Transformation into One Shot TOP10 cells, plating onto M9+Lactose plates, and incubation for 2 days at 37°C produced large colonies. Colonies were re-streaked onto LB-IPTG-XGAL plates and incubated overnight at 37°C.

Blue colonies of the transformation reaction were screened for inserts using primers
30 CNFOR (SEQ ID NO:14); and P455R2 (SEQ ID NO:15). Two PCR-positive colonies were picked and used to inoculate a 6 ml LB culture, which was grown at 37°C. DNA was

isolated from the cultures and the DNA yields were estimated by measuring their OD₂₆₀ with a Spectrophotometer (Table 1).

Table 1: DNA yields for selected clones

Sample name	A260 Concentration (ng/ul)
P469-1	132.69
P469-2	506.91

- 5 200 mls of LB in a 500 ml flask was inoculated with a single blue colony for clone P469-2 and grown for 18 hours at 37°C in a shaker incubator. DNA was purified from this culture using a Qiagen HiSpeed MaxiPrep kit and 440 µg of DNA was recovered. Plasmid P469-2 (SEQ ID NO:17) was sequenced confirmed at GeneWiz.

- 10 In this example, the kanamycin resistance gene and replication origin of GWIZ-Luciferase was successfully replaced by the minimal β-galactosidase/replication origin defined above. An acceptable plasmid yield was achieved when this clone was grown without selective pressure in LB media.

Example 5: Testing β-galactosidase-α complementation vector function in various E. coli strains

- 15 To identify additional *E. coli* strains where the β-galactosidase alpha peptide can be used as a selectable marker instead of an antibiotic resistance gene, one of the plasmids constructed above was tested by DNA transfection into 8 different strains.

Table 2: Bacterial Strains

Strain	Vendor	Genotype
Top10	Thermo-Fisher	F- <i>mcrA</i> Δ(<i>mrr-hsdRMS-mcrBC</i>) Φ80 <i>lacZ</i> Δ <i>M15</i> Δ <i>lacX74</i> <i>recA1</i> <i>araD139</i> Δ(<i>ara</i> <i>leu</i>)7697 <i>galU</i> <i>galK</i> <i>rpsL</i> (Str ^R) <i>endA1</i> <i>nupG</i>
NEB 5-alpha	New England Biolabs	<i>fhuA2</i> Δ(<i>argF-lacZ</i>)U169 <i>phoA</i> <i>glnV44</i> Φ80 Δ(<i>lacZ</i>) <i>M15</i> <i>gyrA96</i> <i>recA1</i> <i>relA1</i> <i>endA1</i> <i>thi-1</i> <i>hsdR17</i>
GT115	InVivogen	F- <i>mcrA</i> Δ(<i>mrr-hsdRMS-mcrBC</i>) φ80 <i>lacZ</i> Δ <i>M15</i> Δ <i>lacX74</i> <i>recA1</i> <i>rspL</i> (Str ^A) <i>endA1</i> Δ <i>dcm</i> <i>uidA</i> (Δ <i>MluI</i>): <i>pir-116</i> Δ <i>sbcC-sbcD</i>
NEB Stable	New England Biolabs	F' <i>proA+B+</i> <i>lacI^q</i> Δ(<i>lacZ</i>) <i>M15</i> <i>zzf::Tn10</i> (Tet ^R) Δ(<i>ara-leu</i>) 7697 <i>araD139</i> <i>fhuA</i> Δ <i>lacX74</i> <i>galK16</i> <i>galE15</i> <i>e14-</i>

		<i>Φ80dlacZΔM15 recA1 relA1 endA1 nupG rpsL (Str^R) rph spoT1 Δ(mrr-hsdRMS-mcrBC)</i>
Stellar	Takara Bio USA	F ⁻ , endA1, supE44, thi-1, recA1, relA1, gyrA96, phoA, Φ80d lacZΔ M15, Δ(lacZYA-argF) U169, Δ(mrr-hsdRMS-mcrBC), ΔmcrA, λ-
DH10B	Thermo-Fisher	F ⁻ mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ (ara, leu)7697 galU galK λ- rpsL nupG /pMON14272 / pMON7124
Stbl3	Thermo-Fisher	F ⁻ mcrB mrrhsdS20(r _B ⁻ , m _B ⁻) recA13 supE44 ara-14 galK2 lacY1 proA2 rpsL20(Str ^R) xyl-5 λ ⁻ leumtl-1
XL1-blue	Agilent	<i>recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac</i> [F' <i>proAB lacI^qZΔM15 Tn10</i> (Tetr)].

Results

50 μl of the E. coli strains in Table 2 were incubated with 1 ng of plasmid P469-2 on ice in a sterile microfuge tube for 30 minutes. The cells were heat shocked for 30 seconds at 42°C and incubated on ice for 1 minute. 450 μl SOC media was added to all cells except NEB-Stable cells. 450 μl of NEB-Stable outgrowth medium (supplied by the manufacturer) was added to the transformed NEB-Stable cells. The cells were incubated at 37°C for 1 hour while shaking. The cells were pelleted and washed 3 times with 1 ml of D-PBS. Cells were plated onto M9-lactose plates and incubated at 37°C for three days.

As expected, no colonies were detected on plates from the Stbl3-transformed cells that were included as a negative control. Five of the strains (Top10, GT115, NEB-Stable, Stellar, and DH10B) had normal-sized colonies. Two strains (NEB-Alpha and XL1-Blue) had small colonies. This was expected since a similar strain to NEB-alpha (DH5alpha) and XL1-Blue contain a mutation in the *purB* gene that results in slow growth on minimal media (Jung et al. Appl Environ. Micro. 76: 6307-6309 (2010)).

XL1-blue and NEB-Alpha plates were incubated for an additional day at 37°C. Pure colonies were obtained by streaking colonies from the M9-lactose plates onto LB-IPTG-XGAL plates and incubating at 37°C. Blue colonies (plasmid containing cells) were streaked a second time onto an LB-IPTG-XGAL plate and incubated at 37°C which produced mostly blue cells.

All of the tested strains that contained the *Φ80dlacZΔM15* marker could be transformed by the β-galactosidase alpha peptide expression plasmid P469-2 and selected

on M9 minimal media with lactose as the sole carbon source. Plasmid P469-2
transfectants of strain XL1-blue that contains the marker *lacI^qZΔM15* on the F episome
were also selectable on M9-Lactose plates. Hence, seven commercially available E. coli
strains have been demonstrated to be compatible with the β-galactosidase selectable
5 marker.

It will be appreciated by those skilled in the art that changes could be made to the
embodiments described above without departing from the broad inventive concept thereof.
It is understood, therefore, that this invention is not limited to the particular embodiments
10 disclosed, but it is intended to cover modifications within the spirit and scope of the
present invention as defined by the present description.

CLAIMS

It is claimed:

1. A method of using a nucleic acid construct as a selectable marker, the method comprising:
 - a. contacting a host cell comprising a deletion in a lac operon with the nucleic acid construct, wherein the nucleic acid construct comprises an isolated β -galactosidase expression cassette comprising a nucleic acid sequence encoding the amino-terminal fragment of β -galactosidase operably linked to a promoter; and
 - b. growing the host cell under conditions wherein the nucleic acid construct is maintained in the host cell.
2. The method of claim 1, wherein the amino-terminal fragment of β -galactosidase comprises an amino acid sequence with at least 75% identity to SEQ ID NO:1.
3. The method of claim 1 or 2, wherein the amino-terminal fragment of β -galactosidase comprises an amino acid sequence of SEQ ID NO:1.
4. The method of any one of claims 1-3, wherein the nucleic acid sequence further comprises a replication origin.
5. The method of claim 4, wherein the replication origin is a high-copy replication origin.
6. The method of claim 5, wherein the high-copy replication origin is the pUC57 replication origin.
7. The method of claim 6, wherein the pUC57 replication origin comprises the nucleic acid sequence of SEQ ID NO:19.
8. The method of any one of claims 1-7, wherein the isolated β -galactosidase expression cassette further comprises a dimer resolution element.
9. The method of claim 8, wherein the dimer resolution element comprises a nucleic acid sequence comprising a site-specific recombinase recognition site.
10. The method of claim 8 or 9, wherein the dimer resolution element further comprises a nucleic acid sequence encoding a site-specific recombinase.

11. The method of claim 8 or 9, wherein the host cell comprises a nucleic acid sequence encoding a site-specific recombinase.
12. The method of any one of claims 8-11, wherein the dimer resolution element is a ColE1 dimer resolution element.
13. The method of claim 12, wherein the ColE1 dimer resolution element comprises the nucleic acid sequence of SEQ ID NO:20.
14. The method of any one of claims 1-13, wherein the host cell comprises a LacZ Δ 15 deletion.
15. The method of any one of claims 1-14, wherein an isolated vector comprises the isolated β -galactosidase expression cassette.
16. The method of claim 15, wherein the isolated vector is less than about 1.5 kilobases in size.
17. The method of claim 15 or 16, wherein the isolated vector comprises a nucleic acid sequence selected from the group consisting of SEQ ID NOs:9-13, 17, and 18.
18. A method of generating the isolated vector of claims 15-17, wherein the method comprises:
 - a. contacting a host cell with the isolated vector;
 - b. growing the host cell under conditions to produce the vector;
 - c. isolating the vector from the host cell.
19. The method of claim 18, wherein the host cell is grown in minimal media.
20. The method of claim 19, wherein the minimal media comprises lactose as the sole carbon source.
21. The method of claim 20, wherein the minimal media comprises about 1% to about 4% weight per volume (w/v) lactose.
22. The method of claim 21, wherein the minimal media comprises about 2% w/v lactose.
23. A kit comprising:
 - a. an isolated β -galactosidase expression cassette of any one of claims 1-13; and
 - b. a host cell comprising a deletion in a lac operon.

24. The kit of claim 23, further comprising minimal media comprising lactose as the sole carbon source.
25. The kit of claim 23 or 24, wherein a vector comprises the isolated β -galactosidase expression cassette.
26. The kit of any one of claims 23-25, wherein the host cell comprises the LacZ Δ 15 deletion.
27. The kit of claim 26, wherein the host cell is selected from the group consisting of an *E. coli* host cell and a yeast host cell.

FIG. 1

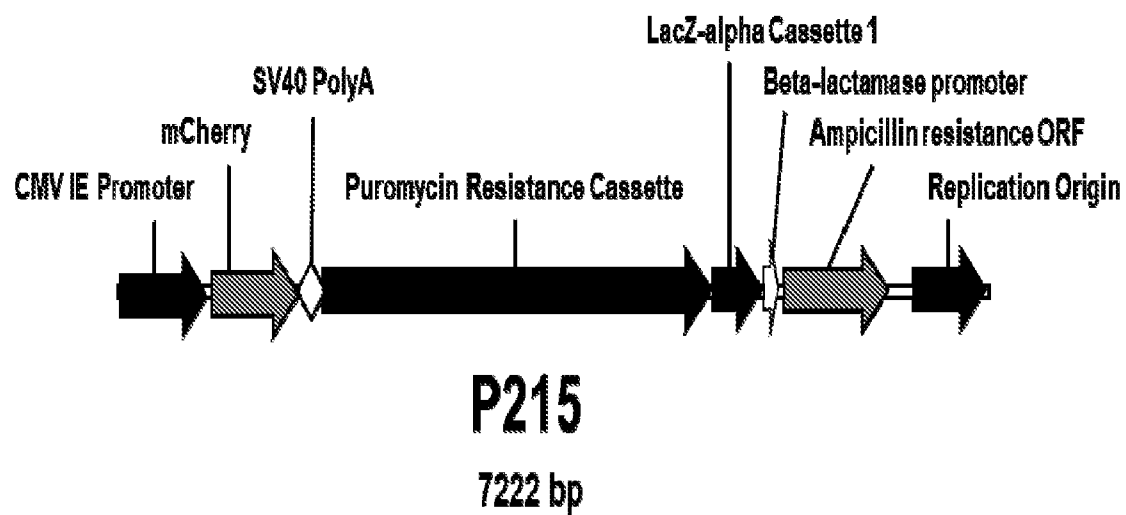


FIG. 2

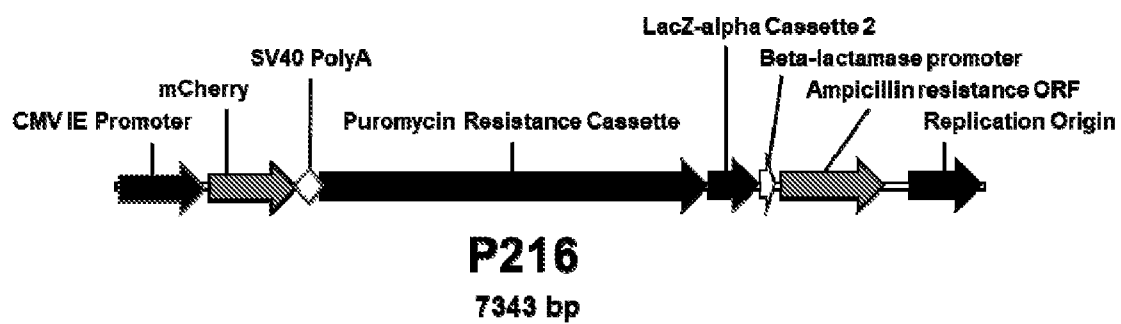


FIG. 3

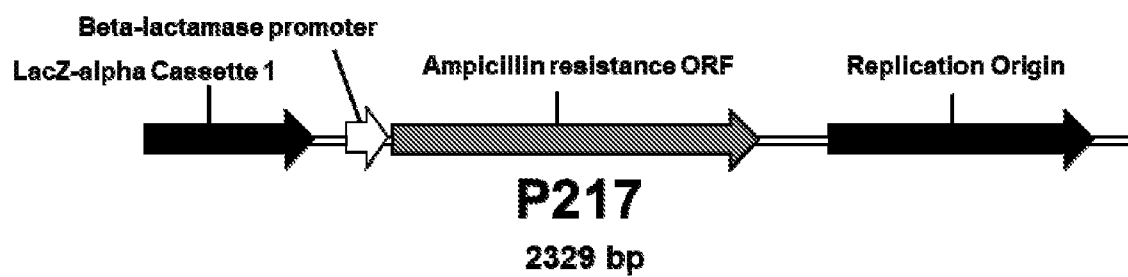


FIG. 4

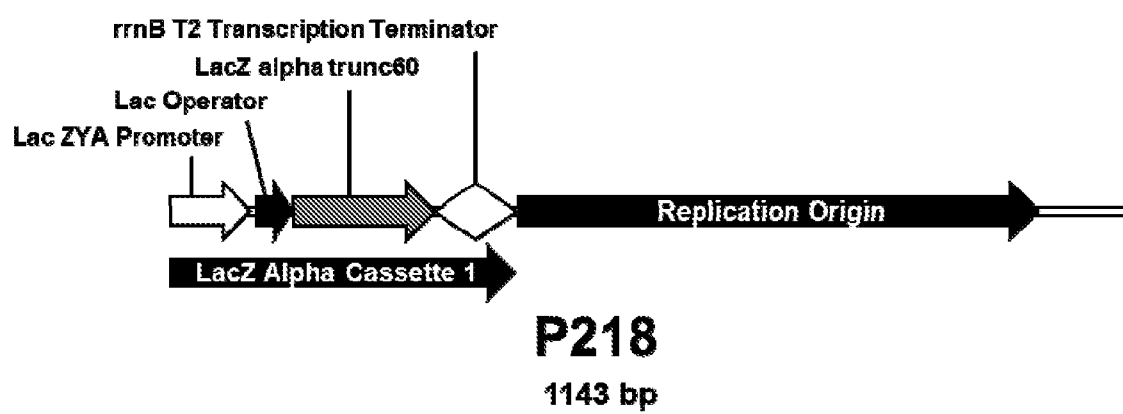


FIG. 5

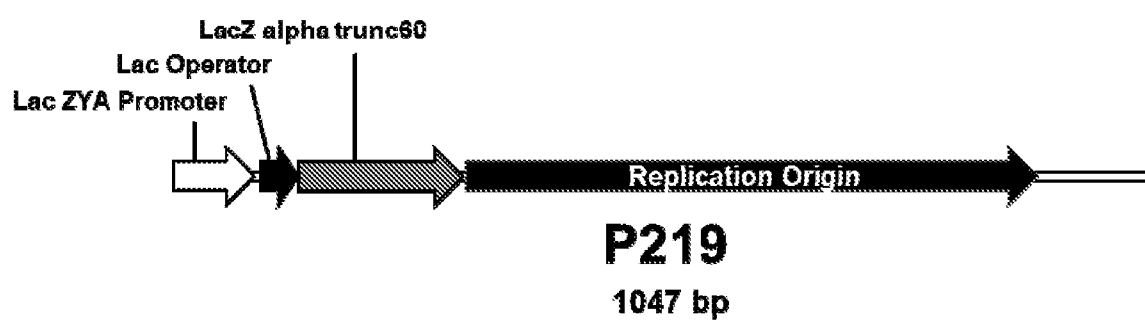


FIG. 6

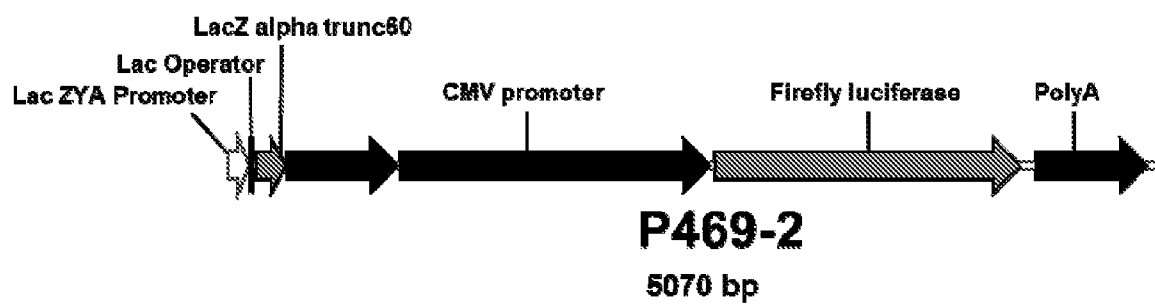


FIG. 1

