

MUTANTS OF THE BACTERIOPHAGE LAMBDA INTEGRASE

Technical Field

The present invention relates to mutants of
5 bacteriophage lambda integrases and to nucleic acid
molecules comprising a nucleotide sequence encoding such
mutants.

Background

Phage integrases are enzymes that mediate
10 unidirectional site-specific recombination between two DNA
recognition sequences, the phage attachment site, *attP*,
and the bacterial attachment site, *attB*. Integrases may
be grouped into two major families, the tyrosine
recombinases and the serine recombinases, based on their
15 mode of catalysis.

Tyrosine family integrases, such as lambda integrase,
utilize a catalytic tyrosine to mediate strand cleavage,
tend to recognize longer *attP* sequences, and require other
proteins encoded by the phage or the host bacteria.

20 Phage integrases from the serine family are larger,
use a catalytic serine for strand cleavage, recognize
shorter *attP* sequences, and do not require host cofactors.
Phage integrases mediate efficient site-specific
recombination between two different sequences that are
25 relatively short, yet long enough to be specific on a
genomic scale.

These properties give phage integrases growing
importance for the genetic manipulation of living
eukaryotic cells, especially those with large genomes such
30 as mammals and most plants, for which there are few tools
for precise manipulation of the genome.

The use of lambda integrases has been subject to
extensive research for catalyzing site-specific DNA
recombination. For example, two mutant lambda integrases,

Int-h (E174K) and its derivative Int-h/218 (E174K/E218K) have been described and were shown to catalyze intermolecular recombination reactions at least as efficiently as the corresponding intramolecular recombination reactions in human cells. Although the presence of arm-site sequences have been shown to increase the recombination of core-sites by Int-h/218 in vivo, given the absence of an attB site in the human genome, recombination reactions occur in non-cognate sites in an essentially random manner.

This makes it difficult to engineer cell lines in a controlled, reproducible fashion.

Therefore, there remains a need to provide mutant integrases having greater efficiency and specificity in catalyzing site specific recombination reactions.

Summary

In one aspect, there is provided a lambda integrase comprising at least one amino acid mutation selected from the group consisting of I43F, E319G and D336V.

In another aspect, there is provided a lambda integrase comprising an amino acid mutation at position 336 of the lambda integrase as set forth in SEQ ID NO: 1, wherein the amino acid residue aspartate at sequence position 336 is replaced by an aliphatic amino acid.

In another aspect, there is provided a nucleic acid molecule comprising a nucleotide sequence encoding the lambda integrase as disclosed herein.

In another aspect, there is provided a nucleic acid molecule. The nucleic acid molecule includes a nucleotide sequence encoding a mutant as described herein.

In a further aspect, there is provided a host cell. The host cell includes a nucleic acid molecule as described herein.

In another aspect, there is provided a method of recombining a nucleic acid of interest into a target nucleic acid, the method comprising contacting a targeting nucleic acid comprising the nucleic acid of interest with the target nucleic acid in the presence of a lambda integrase as disclosed herein.

In yet another aspect, there is provided a method of recombining a nucleic acid of interest into a target nucleic acid. The method includes contacting a targeting nucleic acid comprising the nucleic acid of interest with

the target nucleic acid in the presence of a mutant as described herein.

In another aspect, there is provided a sequence specific recombination kit comprising:

- a. a targeting nucleic acid into which a nucleic acid of interest can be inserted, and
- b. a lambda integrase as disclosed herein or a nucleic acid as disclosed herein.

In yet a further aspect, there is provided a sequence specific recombination kit. The kit includes a targeting nucleic acid into which a nucleic acid of interest can be inserted, and a mutant as described herein.

Definitions

The following words and terms used herein shall have the meaning indicated:

The term "mutant" refers to a protein arising as a result of a mutation or a recombinant DNA procedure.

The term "Int" or "integrase" refers to the lambda phage integrase protein.

As used herein, "nucleic acid" refers to any nucleic acid in any possible configuration, such as linearized single stranded, double stranded or a combination thereof. Nucleic acids may include, but are not limited to DNA molecules (e.g., cDNA or genomic DNA), RNA molecules (e.g., mRNA) , analogues of the DNA or RNA generated using nucleotide analogues or using nucleic acid chemistry, cDNA synthetic DNA, a copolymer of DNA and RNA, oligonucleotides, and PNA (protein nucleic acids) . DNA or RNA may be of genomic or synthetic origin and may be single or double stranded. A respective nucleic acid may furthermore contain non-natural nucleotide analogues and/or be linked to an affinity tag or a label.

As used herein, nucleotides include nucleoside mono-, di-, and triphosphates. Nucleotides also include modified-nucleotides, such as, but not limited to, phosphorothioate nucleotides and deazapurine nucleotides and other nucleotide analogs.

An "attB/attP reaction" or a "B/P reaction" is a recombination reaction between an attB recognition site and an attP recognition site mediated by an Int.

5 An "attH/attPH reaction" or an "H/PH reaction" is a recombination reaction between an attH recognition site and an attPH recognition site mediated by an Int.

An "att site" is an attachment site on a DNA molecule for an integrase or integrase complex. As used herein, "att site" is generally used interchangeably with
10 "recognition site," described in greater detail below. Generally, "att site" is used to refer to a particular type of recognition site, such as, for example, an attB, an attP, an attL, or an attR site.

"Chromosomally-integrated" or "integrated" refers to
15 the integration of a foreign gene or nucleotide sequence into a host genome by covalent bonds that are formed with the host DNA.

"Deletion reaction" and "excision reaction" are used interchangeably and refer to a recombination reaction
20 between two recognition sites that are on the same DNA molecule and are in direct orientation with respect to one another. This reaction results in the removal of a nucleotide sequence that is positioned between the two recognition sites.

25 "Direct orientation" refers to an orientation of two or more recognition sites such that 15 base pair core regions of the recognition sites are oriented in the same 5' to 3' direction. "Direct repeat," as used herein, refers to two or more recognition sites that are in direct
30 orientation with respect to each other.

"Donor," "donor molecule," "donor sequence," and "donor DNA" are used interchangeably to refer to a nucleotide sequence that has been selected to undergo recombination with the target DNA sequence using site-

directed recombination. The donor nucleotide sequence can be any nucleotide sequence, such as, for example, a gene, an expression cassette, a promoter, a molecular marker, a selectable marker, a visible marker, a portion of any of these, or the like. The donor DNA sequence comprises at least one recombinase recognition site.

"Endogenous" as used herein means "of the same origin," i.e., derived from a host cell.

"Expression cassette" as used herein includes a nucleotide sequence that is capable of directing or driving the expression of another nucleotide sequence in an appropriate host cell. An expression cassette typically comprises a promoter operably linked to a nucleotide sequence, such as a nucleotide sequence of interest, for example, which is operably linked to a termination signal. The expression cassette also typically comprises sequences needed for proper translation of the nucleotide sequence. The nucleotide sequence of interest usually codes for a protein of interest but can also code for a functional RNA of interest, for example antisense RNA or a non-translated RNA that, in the sense or antisense direction, inhibits expression of a particular gene, e.g., antisense RNA. The expression cassette comprising the nucleotide sequence can be chimeric, meaning that at least one of its components is heterologous with respect to at least one of its other components. The expression cassette can include endogenous DNA that has been obtained in a recombinant form and is useful for heterologous expression. Typically, however, the expression cassette is heterologous with respect to the host; that is, the particular DNA sequence of the expression cassette does not occur naturally in the host cell and must be introduced into the host cell or an ancestor of the host cell through a transformation event. The expression of the nucleotide sequence in the

expression cassette can be under the control of any suitable promoter, such as for example, either a constitutive promoter or an inducible promoter that initiates transcription only when the host cell is exposed to some particular external stimulus. In the case of a multicellular organism, the promoter can also be specific to a particular tissue or organ or stage of development.

A "foreign" gene or DNA refers to a gene or a nucleotide sequence that is not normally found in the host organism but can be introduced by gene transfer. Foreign genes and DNA that are not integrated into the genome of the host cell are referred to as "extrachromosomal."

The term "gene" is used broadly to include any segment of a nucleotide sequence associated with a biological function. Thus, a gene can include a coding sequence either with or without the regulatory sequences required for its expression. Further, a gene can include both exon and intron sequences or can include only exon sequences. A gene can also include non-expressed DNA segments that, for example, form recognition sequences for other proteins. A "portion of a gene" or "an incomplete gene" as used herein means a part of a gene that is non-functional because it does not contain all of the sequence needed for functionality. The portion can be the 5' portion of a gene (i.e., the sequence at the 3' end of the gene is not present), or the portion can be the 3' portion of a gene (i.e., the sequence at the 5' end of the gene is not present). The 5' and 3' portions can be nonfunctional on their own, but when the 5' and 3' portions are operably linked, the gene is "functional" or "complete."

"Gene of interest," "sequence of interest," "nucleic acid of interest," and "DNA of interest" are used interchangeably and include any nucleotide sequence which, when transferred to a cell, confers upon the cell a

desired characteristic, such as virus resistance, insect resistance, antibiotic stress resistance, disease resistance, resistance to other pests, herbicide tolerance, improved nutritional value, improved performance in an industrial process, or altered reproductive capability, for example. The sequence of interest can also be one that is transferred to cell lines or mammals or plants for the production of commercially valuable enzymes or metabolites. In this context, the "target nucleic acid" as used herein refers to a nucleotide sequence containing at least one recognition site. The target nucleotide sequence can be a gene, an expression cassette, a promoter, a molecular marker, a portion of any of the above, or the like. The target nucleic acid can be stably transformed into a host cell to create a transformed cell line comprising the target sequence integrated into a chromosomal location in the genome. Accordingly, in some embodiments, the target nucleic acid can include genomic DNA. The genomic DNA can be comprised in a cell. In other embodiments, the target nucleic acid can include a sequence selected from the group consisting of an *attH* sequence (SEQ ID NO: 7) and an *attH4X* sequence (SEQ ID NO: 31).

"Genome" refers to the complete genetic material of an organism.

"Heterologous" as used herein means "of different natural origin," i.e., representing a non-natural state. For example, if a host cell is transformed with a gene derived from another organism, particularly from another species, that gene is heterologous with respect to both the host cell and descendants of the host cell that carry the gene. Similarly, "heterologous" refers to a nucleotide sequence which is derived from a natural or original cell type and is inserted into that same natural or original

cell type, but which is present in a non-natural state, such as, for example, in a different copy number, under the control of different regulatory elements, or the like.

To "identify" a recombination product means that the recombination product is detected and distinguished from both the target and donor sequences. There are many means for identifying a recombination product. For example, a selectable marker gene can be used, whereby site-specific integration results in the selectable marker becoming operatively linked with a promoter only in a recombinant product. Alternatively, a visible marker gene can be used, whereby a gain or loss of marker gene expression identifies a recombination product. Alternatively, a negative selectable marker gene can be used, whereby a loss or lack of expression of the marker gene identifies a recombination product. Additionally, molecular markers that are characteristic of the target sequence and/or donor sequence can be used, such that the molecular marker pattern is unique for the recombination product.

"Integrase" as used herein refers to a bacteriophage lambda-derived integrase, including wild-type integrase and any of a variety of mutant or modified integrases. "Integrase complex" as used herein refers to a complex comprising integrase and integration host factor (IHF). "Integrase complex" as used herein may also refer to a complex comprising integrase, integration host factor, and a bacteriophage lambda-derived excisionase (Xis). Further, as used herein, "Int" refers to both "integrase" and "integrase complex."

An "integrase-mediated recombination product" is a recombination product formed between target and donor sequences in the presence of an integrase or integrase complex. The integrase-mediated recombination results in strand exchange between at least one recombinase

recognition site on the target and at least one recombina-
se recognition site on the donor, whereby a
recombination product is formed. Consistent with the usage
defined above, "Int-mediated recombination" or "Int-
5 mediated recombination product" means a recombination or
recombination product that is mediated by either an
integrase or an integrase complex.

"Intramolecular recombination" refers to
recombination between recognition sites on a single
10 nucleic acid molecule. Recombination between recognition
sites on different molecules is termed "intermolecular
recombination."

"Intrachromosomal recombination" refers to
recombination between recognition sites on a single
15 chromosome. Recombination between recognition sites on
different chromosomes is termed "interchromosomal
recombination."

An "inversion reaction" refers to an intramolecular
recombination reaction between two *att* sites that are in
20 inverted orientation with respect to each other. For
example, an inversion reaction can be effected by an
intramolecular reaction between either an *attB* site and an
attP site in inverted orientation or an *attL* site and an
attR site in inverted orientation.

25 "Inverted orientation" refers to an orientation of
two recognition sites such that 15 base pair core regions
of the recognition sites are oriented in the opposite 5'
to 3' direction.

"Operably linked" or "operatively linked" refers to
30 the relationship between two or more nucleotide sequences
that interact physically or functionally. For example, a
promoter or regulatory nucleotide sequence is said to be
operably linked to a nucleotide sequence that codes for an
RNA or a protein if the two sequences are situated such

that the regulatory nucleotide sequence will affect the expression level of the coding or structural nucleotide sequence. A 5' portion of a gene is operatively or operably linked with a 3' portion of a gene if the two portions are situated to form a functional gene.

"Recognition site" or "recombination site" refers to a nucleotide sequence that can be recognized by a recombinase protein. The recognition site is the nucleotide sequence at which binding, cleavage, and strand exchange is performed by the recombinase and any associated accessory proteins. Integrase or integrase complex recognizes recognition sites comprising an *attB*, *attL*, *attR*, *attP*, and/or suitable mutations of such sites. The *attB* site can be approximately 25-30 bps and includes two 7 bp core sequences and a 7 bp overlap (or spacer) region, whereas the *attP* site can be approximately 240 bps and comprises binding sites for an integrase and one or more accessory proteins. The *attB* and *attP* sites can be recombined together by Int or, alternatively, the *attL* and *attR* sites can be recombined together by Int.

"Recombinase" refers to an enzyme that is capable of performing site-specific recombination of DNA. Recombinase enzymes possess endonuclease and ligase activities. A recombinase can function either as a single protein or as a part of a complex of proteins. As used herein integrase and integrase complex are recombinases.

Generally, if a recombinase-mediated recombination occurs between two recombinase recognition sites that are on the same molecule, the recombination reaction results in either the deletion or inversion of a sequence flanked by the two recognition sites. If a recombinase-mediated recombination occurs between two recombinase recognition sites that are on different molecules (e.g., between a recombinase recognition site on a target sequence and a

recombinase recognition site on a donor sequence), the recombination reaction results in the insertion of a sequence from one of the molecules into the other molecule (e.g., the insertion of a donor sequence into a target molecule). When particular recognition sites that are capable of recombining are present on both the target and the donor (e.g., an *attB* site on the target and an *attP* site on the donor or an *attL* site on the target and an *attR* site on the donor), the recombination product represents an exchange of nucleotide sequence between the two sites, resulting in two new sites. Each of these new sites contains a part of the original recognition sites from both the donor and target molecules. For example, when recombination occurs between an *attB* site on the target and an *attP* site on the donor, *attL* and *attR* sites are created in the recombination products. Additionally, the newly formed *attL* and *attR* sites are flanked on one side by sequence obtained from the donor molecule and on the other side by sequence obtained from the target molecule.

"Regulatory element" includes a nucleotide sequence that is involved in conferring upon a host cell the expression of another nucleotide sequence, such as, for example, a sequence of interest. A regulatory element can comprise a promoter that is operably linked to the nucleotide sequence of interest and to a termination signal. Regulatory elements also typically encompass sequences useful for proper translation of the nucleotide sequence of interest.

"Selectable marker" or "selectable marker gene" refers to a nucleotide sequence whose expression in a cell gives the cell a selective advantage under particular conditions. The selective advantage possessed by the cell transformed with the selectable marker gene can be an

improved ability to grow in the presence of a negative selective agent, such as an antibiotic or an herbicide, for example, as compared to the ability of non-transformed cells. Alternatively, the selective advantage possessed by the transformed cells can be an enhanced capacity, relative to non-transformed cells, to utilize a particular compound as a nutrient, growth factor, or energy source.

Alternatively, the selective advantage possessed by the transformed cell can be the loss of a previously possessed trait or characteristic, effecting what is termed "negative selection." In this last case, the host cell is exposed to or contacted by a compound that is toxic only to cells that have not lost the ability to express a specific trait or characteristic (such as a negative selectable marker gene, for example) that was present in the parent cell, which is typically a transgenic parent cell.

"Site-directed recombination" as used herein refers to recombination between two nucleotide sequences that each comprises at least one recognition site.

"Site-specific" means at a particular nucleotide sequence, which can be in a specific location in the genome of a host cell. The nucleotide sequence can be endogenous to the host cell, either in its natural location in the host genome or at some other location in the genome, or it can be a heterologous nucleotide sequence, which has been previously inserted into the genome of the host cell by any of a variety of known methods.

"Stably transformed" refers to a host cell that contains a nucleotide sequence of interest that has been stably integrated into the genome of the host cell.

"Target," "target molecule," "target sequence," and "target DNA" are used interchangeably to refer to a

nucleotide sequence containing at least one recombinase recognition site. The target nucleotide sequence can be a gene, an expression cassette, a promoter, a molecular marker, a portion of any of these, or the like. The target
5 sequence can be stably transformed into a cell to create a "target line" comprising the target sequence integrated into a chromosomal location in a genome.

A "targeted integration event" or "target event" refers to a recombination product formed between target
10 and donor sequences in the presence of an integrase or integrase complex. In particular, it refers to the integration of a donor sequence into a target sequence as a consequence of an Int-mediated recombination when the target sequence is stably transformed into a cell.

15 A "visible marker gene" refers to a gene or nucleotide sequence whose expression in a transformed cell may not confer an advantage to that cell but can be detected or made visible. Examples of visible markers include, but are not limited to, β -glucuronidase (GUS),
20 luciferase (LUC), and fluorescent proteins (such as green fluorescent protein (GFP) or cyan fluorescent protein (CFP), for example).

The word "substantially" does not exclude "completely" e.g. a composition which is "substantially
25 free" from Y may be completely free from Y. Where necessary, the word "substantially" may be omitted from the definition of the invention.

Unless specified otherwise, the terms "comprising" and "comprise", and grammatical variants thereof, are
30 intended to represent "open" or "inclusive" language such that they include recited elements but also permit inclusion of additional, non-recited elements.

As used herein, the term "about", in the context of concentrations of components of the formulations,

typically means +/- 5% of the stated value, more typically +/- 4% of the stated value, more typically +/- 3% of the stated value, more typically, +/- 2% of the stated value, even more typically +/- 1% of the stated value, and even
5 more typically +/- 0.5% of the stated value.

Throughout this disclosure, certain embodiments may be disclosed in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an
10 inflexible limitation on the scope of the disclosed ranges. Accordingly, the description of a range should be considered to have specifically disclosed all the possible sub-ranges as well as individual numerical values within that range. For example, description of a range such as
15 from 1 to 6 should be considered to have specifically disclosed sub-ranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual numbers within that range, for example, 1, 2, 3, 4, 5, and 6. This applies regardless of
20 the breadth of the range.

Certain embodiments may also be described broadly and generically herein. Each of the narrower species and subgeneric groupings falling within the generic disclosure also form part of the disclosure. This includes the
25 generic description of the embodiments with a proviso or negative limitation removing any subject matter from the genus, regardless of whether or not the excised material is specifically recited herein.

30 **Disclosure of Optional Embodiments**

Exemplary, non-limiting embodiments of a lambda integrase comprising at least one amino acid mutation at positions 43, 319 and 336, will now be disclosed.

In this context, mutations present in the lambda integrase described herein may comprise any mutations such as substitutions, deletions and also insertions of the natural amino acid sequence of the lambda integrase as long as the resulting polypeptide folds into a three-dimensionally stable structure and shows the desired (enhanced) recombination activity. The lambda integrase described herein may comprise conservative and/or non-conservative mutations. Examples of possible mutations are conservatively modified variations where the alteration is the substitution of an amino acid with a chemically similar amino acid. In addition to the above, the lambda integrase may comprise mutations, such as conservative mutations, outside of the regions as mentioned above. Such conservative substitutions are known to those of skill in the art and may include substitutions between: 1) alanine, serine, threonine; 2) aspartic acid and glutamic acid; 3) asparagine and glutamine; 4) arginine and lysine; 5) isoleucine, leucine, methionine, valine; and 6) phenylalanine, tyrosine, tyroptophan.

The "amino acid residue" as used herein refers to any amino acid and can either be in the D or L form or to an amino acid mimetic that can be incorporated into a polypeptide by an amide bond.

Accordingly, the positively charged amino acid residue can for example either be a naturally occurring amino acid residue that is positively charged under physiological conditions such as arginine or lysine or a non-natural mimetic such as a lysine residue the alpha-amino group of which is alkylated in order to yield a (quarternary) ammonium-salt having a permanent positive charge.

In one embodiment a lambda integrase comprising at least one amino acid substitution at positions 43, 319 and 336 of the lambda integrase as set forth in SEQ ID NO: 1.

5 In another embodiment, the lambda integrase as described herein comprises an amino acid substitution at positions 43, 319 and 336 of the lambda integrase as set forth in SEQ ID NO: 1.

10 In another embodiment, the amino acid residue isoleucine at sequence position 43 is replaced by an aromatic amino acid. The aromatic amino acid may be selected from the group consisting of phenylalanine, tyrosine and tryptophan. In one embodiment the aromatic amino acid is phenylalanine.

15 In another embodiment, the amino acid residue glutamate at sequence position 319 is replaced by glycine.

In another embodiment, the amino acid residue aspartate at sequence position 336 may be replaced by a hydrophobic amino acid. The hydrophobic amino acid may be an aliphatic amino acid. The aliphatic amino acid may be
20 selected from the group consisting of isoleucine, leucine and valine. In one embodiment the aliphatic amino acid is valine.

In a further embodiment, the lambda integrase as described herein may comprise the amino acid substitutions
25 I43F, E319G and D336V.

In an alternative embodiment, the lambda integrase as described herein may comprise an amino acid substitution at position 336 of the lambda integrase as set forth in SEQ ID NO: 1. The amino acid residue aspartate at sequence
30 position 336 may be replaced by a hydrophobic amino acid. The hydrophobic amino acid may be an aliphatic amino acid.

The aliphatic amino acid may be selected from the group consisting of isoleucine, leucine and valine. In one embodiment, the aliphatic amino acid is valine.

5 The mutations in the lambda integrase as described herein are generally important in directing recombinase specificity and efficiency.

10 The lambda integrase as described herein can be generated through various selection systems known to persons skilled in the art. For example, bacterial selection systems relying on identification of functional mutants through reporter gene activation or substrate-linked protein evolution (SLiPE) have been previously described. These selection systems are one of many different approaches for engineering altered site-specificities in recombinases. For example, a genetic selection system in yeast has also been described that yielded HIV-1 integrase variants displaying altered DNA binding affinities. As another example, *in vitro* compartmentalization (IVC) can be used as a selection system for generating and identifying variants such as the mutants of the invention as described herein.

25 The bacteriophage lambda integrase is the prototypical member of the large tyrosine-recombinase family. Generally, the bacteriophage lambda integrase comprises 3 distinct domains that collaborate within a higher-order tetrameric structure to form a dynamic recombinogenic complex. These 3 domains are the N-terminal DNA binding domain (amino acid residues 1-64); the core DNA-binding domain (amino acid residues 65-175); and the C-terminal catalytic domain (amino acid residues 176-356). The bacteriophage lambda integrase is central to the bacteriophage lifecycle, facilitating the controlled integration and excision of its genome into and

out of the host bacterial chromosome, respectively. In its natural function, the bacteriophage lambda integrase is able to catalyze site-specific recombination between a pair of target sequences, termed *att* sites, in the absence of high-energy cofactors. The target sequences (*attP* in the bacteriophage genome, *attB* in the bacterial genome) comprise a pair of 7bp inverted core-binding sites separated by a 7bp "overlap" region. The "overlap region" or "overlap sequence" as used herein defines the sequence of the recombination sequences where the DNA strand exchange, including strand cleavage and re-ligation, takes place and relates to the consensus DNA sequence 5'-TTTATAC-3' in wild-type *att* sites or said sequence having functional nucleotide substitutions. The bacteriophage lambda integrase DNA core-binding domain primarily recognizes the 7bp *attP* x *attB* core DNA sequence motifs. In the much longer *attP* site, the core sequence is flanked by binding sites for accessory DNA-bending factors such as integration host factor (IHF), factor for inversion stimulation (FIS) and excisionase (Xis). In addition to these accessory sites, several 'arm' binding sites for the N-terminal domain of the bacteriophage lambda integrase also flank the *attP* core site. Binding of the N-domain of the bacteriophage lambda integrase to 'arm' binding sites allosterically modulates the coupled core binding and catalytic domain to increase the affinity to core sites, which ultimately enables DNA strand cleavage and productive recombination of *attB* x *attP*. Therefore, these 'arm' regions are essential for activating efficient DNA cleavage by the C-terminal catalytic domain of bacteriophage lambda integrase, and thus contribute to the regulation of recombination directionality.

Generally, when a recombinase-mediated recombination occurs between two recognition sites, the recombination

reaction can either occur on two different molecules or within the same molecule (e.g., between a recognition site on a target sequence and a recognition site on a donor sequence). In this context, the lambda integrase as
5 described herein can catalyze either intermolecular or intramolecular recombination reactions or both intermolecular and intramolecular recombination reactions.

As used herein, "site-specific recombination" or "sequence-specific recombination" refers to recombination
10 between two nucleotide sequences that each comprises at least one recognition site or at least one non-cognate site. "Site-specific" means at a particular nucleotide sequence, which can be in a specific location in the genome of a host cell for example. The nucleotide sequence
15 can be endogenous to the host cell, either in its natural location in the host genome or at some other location in the genome, or it can be a heterologous nucleotide sequence, which has been previously inserted into the genome of the host cell by any of a variety of known
20 methods.

As described herein, "recognition sites" or "cognate sites" refer to a nucleotide sequence that can be recognized by a recombinase protein. The "recognition site" is the nucleotide sequence upon which binding,
25 cleavage and strand exchange is performed by the recombinase protein and any associated accessory proteins. The lambda integrase recognizes cognate sites comprising *attB*, *attP*, *attL*, *attR*, and/or suitable mutations of such sites. The *attB* site and *attP* sites can be recombined
30 together by the lambda integrase, or alternatively, the *attL* and *attR* sites can be recombined by the lambda integrase. In this context, the lambda integrase (Int mutants) described herein can facilitate recombination between, for example, the *attB* and *attP* sites.

Advantageously, the lambda integrase described herein is able to recombine into non-cognate sites (such as the *attH* site) with greater efficiency, as compared to the parental Int-h/218integrase.

5 In another embodiment there is provided a nucleic acid molecule comprising a nucleotide sequence encoding the lambda integrase as described herein.

 It will be appreciated that the degeneracy of the genetic code permits substitutions of certain codons by
10 other codons which specify the same amino acid and hence give rise to the same protein, the invention is not limited to a specific nucleic acid molecule but includes all nucleic acid molecules comprising a nucleotide sequence coding for the lambda integrase described herein.

15 In one embodiment, the nucleic acid molecule is operably linked to a regulatory sequence to permit expression of the nucleic acid molecule.

 It will be appreciated that the precise nature of the regulatory regions needed for gene sequence expression may
20 vary from organism to organism, but shall, in general include a promoter region which, in prokaryotes, contains only the promoter or both the promoter which directs the initiation of RNA transcription as well as the DNA sequences which, when transcribed into RNA will signal the
25 initiation of synthesis. Such regions will normally include non-coding regions which are located 5' and 3' to the nucleotide sequence to be expressed and which are involved with initiation of transcription and translation such as the TATA box, capping sequence and CAAT sequences.
30 These regions can for example, also contain enhancer sequences or translated signal and leader sequences for targeting the produced polypeptide to a specific compartment of a host cell, which is used for producing a

recombinant lambda integrase of the present invention. In one embodiment regulatory sequence comprises a promoter sequence.

In some embodiments, a nucleic acid of the invention
5 comprises a transcriptional initiating region functional in a cell and a transcriptional terminating region functional in a cell. Suitable promoter sequences that can be used are for example, the lac promoter, the tet-promoter or the T7 promoter in the case of bacterial
10 expression. An example of a promoter suitable for expression in eukaryotic systems is the SV 40 promoter.

In further embodiments, the nucleic acid molecule is comprised in a vector, particularly in an expression vector. Such an expression vector can comprise, besides
15 the above-mentioned regulatory sequences and a nucleic acid sequence which codes for a lambda integrase, a sequence coding for restriction cleavage site which adjoins the nucleic acid sequence coding for the lambda integrase in 5' and/or 3' direction. This vector also
20 permits the introduction of another nucleic acid sequence coding for a protein to be expressed. The expression vector may also contain replication sites and control sequences derived from a species compatible with the host that is to be used for expression. The expression vector
25 may be based on plasmids well known to person skilled in the art such as pBR322, pUC16, pBluescript (RTM) and the like.

In one embodiment there is also provided a host cell containing a nucleic acid molecule. The vector containing
30 the nucleic acid molecule can be transformed into host cells capable of expressing the genes. The transformation can be carried out in accordance with standard techniques. In this context, the transformed host cells can be cultured under conditions suitable for expression of the

nucleotide sequence encoding the lambda integrase. Host cells can be established, adapted and completely cultivated under serum free conditions, and optionally in media which are free of any protein/peptide of animal origin. Commercially available media such as RPMI-1640 (Sigma), Dulbecco's Modified Eagle's Medium (DMEM; Sigma), Minimal Essential Medium (MEM; Sigma), CHO-S-SFMII (Invitrogen), serum free-CHO Medium (Sigma), and protein-free CHO Medium (Sigma) are exemplary appropriate nutrient solutions. Any of the media may be supplemented as necessary with a variety of compounds, examples of which are hormones and/or other growth factors (such as insulin, transferrin, epidermal growth factor, insulin like growth factor), salts (such as sodium chloride, calcium, magnesium, phosphate), buffers (such as HEPES), nucleosides (such as adenosine, thymidine), glutamine, glucose or other equivalent energy sources, antibiotics, trace elements. Any other necessary supplements may also be included at appropriate concentrations that are known to those skilled in the art.

In yet another embodiment, there is provided a method of recombining a nucleic acid of interest into a target nucleic acid. The method comprises contacting a targeting nucleic acid comprising the nucleic acid of interest with the target nucleic acid in the presence of a lambda integrase as described herein.

In some embodiments, the method of recombining the nucleic acid of interest into the target nucleic acid is a sequence specific recombination. The sequence specific recombination can be performed in the presence of one or more cofactors. The cofactors can be selected from the group consisting of integration host factor (IHF), factor for inversion stimulation (FIS) and excisionase (Xis).

The "targeting nucleic acid" as used herein refers to a nucleotide sequence that contains at least one recognition site. The targeting nucleic acid can contact a target nucleic acid in the presence of a mutant of the invention, in order to recombine a nucleic acid of interest into the target nucleic acid. The targeting nucleotide sequence can be a gene, an expression cassette, a promoter, a molecular marker, a portion of any of the above, or the like. In some embodiments, the targeting nucleic acid can be a vector. In other embodiments, the targeting nucleic acid comprises a sequence selected from the group consisting of an *attP*H sequence (SEQ ID NO: 8) and an *attP*4X sequence (SEQ ID NO: 9). The term "nucleic acid of interest" as used herein refers to a polynucleotide sequence of any length that encodes a product of interest. The selected sequence can be a full length or a truncated gene, a fusion or tagged gene, and can be a cDNA, a genomic DNA, or a DNA fragment. It can also be the native sequence, i.e., naturally occurring form(s), or can be mutated or otherwise modified as desired. These modifications can include codon optimizations to optimize codon usage in the selected cell or host cell, humanization or tagging. The selected sequence can encode a secreted, cytoplasmic, nuclear, membrane bound or cell surface polypeptide. The "product of interest" can include, but are not limited to proteins, polypeptides, fragments thereof, peptides, antisense RNA, all of which can be produced in the selected host cell.

In one embodiment, the genomic DNA is comprised in a cell. The method described herein may be performed in all eukaryotic cells. Cells and cell lines may be present, for example in a cell culture and include but are not limited to eukaryotic cells, such as yeast, plant, insect or mammalian cells. For example, the cells may be

oocytes, embryonic stem cells, hematopoietic stem cells or any type of differentiated cells. In certain embodiments, the method of the invention can be performed in a mammalian cell. The mammalian cell lines can include, but
5 are not limited to a human, simian, murine, mice, rat, monkey, rabbit, rodent, hamster, goat, bovine, sheep or pig cell lines. Exemplary cell lines can include, but are not limited to Chinese hamster ovary (CHO) cells, murine myeloma cells such as NSO and Sp2/0 cells, COS cells, Hela
10 cells and human embryonic kidney (HEK-293) cells.

The target nucleic acid may comprise DNA. The DNA may be genomic DNA.

In one embodiment, the target nucleic acid comprises a sequence selected from the group consisting of an *attH* sequence (SEQ ID NO: 7) and an *attH4X* sequence (SEQ ID NO:
15 31). The targeting nucleic acid may be a vector. In one embodiment, the targeting nucleic acid comprises a sequence selected from the group consisting of an *attH* sequence (SEQ ID NO: 7) and an *attH4X* sequence (SEQ ID NO:
20 31).

In another embodiment sequence specific recombination may be performed in the presence of one or more cofactors. The cofactors may be selected from the group consisting of XIS, FIS and IHF.

25 In another embodiment, there is provided a sequence specific recombination kit comprising a targeting nucleic acid into which a nucleic acid of interest can be inserted, and lambda integrase or a nucleic acid as described herein.

30 The kit as described herein may comprise at least one reagent for inserting a nucleic acid of interest into the targeting nucleic acid. The reagent may be a restriction

enzyme or ligase. In another embodiment, the targeting nucleic acid may comprise a sequence selected from the group consisting of an *attH* sequence (SEQ ID NO: 7) and an *attH4X* sequence (SEQ ID NO: 31).

5 In one embodiment, the kit as described herein may further comprise buffer(s) and/or instructions for recombining the nucleic acid of interest with a given target nucleic acid.

In one embodiment, the kit as described herein may further
10 comprise at least one reagent for determining successful sequence specific recombination event. In one embodiment, the reagent component is a primer pair. The primer pair may be supplied in combination with the kit or supplied separately from the kit.

15

Brief Description Of Drawings

The accompanying drawings illustrate a disclosed embodiment and serves to explain the principles of the disclosed embodiment. It is to be understood, however,
20 that the drawings are designed for purposes of illustration only, and not as a definition of the limits of the invention.

Figure 1 shows the sequence alignment of the core
25 bacterial *attB*, and human *attH* and *attH4X* sequences. The 7 base pairs (bp) highlighted in grey represent the overlap sequence, which must be identical in both recombination partners, i.e. *attB* & *attP*, *attH* & *attPH*, *attH4X* and *attP4X*. The *attH* site differs from the bacterial *attB* site
30 at one position in the 7bp overlap sequence and three positions in the right arm core binding sequence. The non-cognate *attH4X* site occurs approximately 940 times in the human genome as part of human *Line1* (long interspersed

nuclear elements/remnants of retrotransposons/non-coding). The first three nucleotides of the attH4X sequence are degenerate.

Figure 2 shows the methodology for Rapid *E. coli* Chromosomal Integration using DNA minicircles and lambda integrase technology. The first step is to create a minicircle comprising lambda integrase under control of a suitable promoter (e.g. T7), and a minicircle comprising gene to be stably integrated, a selectable marker for antibiotic resistance (e.g. lactamase gene cassette) and the attP sequence. The second step is to transform both minicircles by electroporation or heat shock into *E. coli*, and then to plate and culture on selectable media (eg ampicillin plates). The third step is to confirm integration into attB site by PCR, sequencing and Southern blot.

Figure 3A shows in vivo recombination of exogenous DNA (lactamase gene cassette) into the attB site of *E. coli* chromosomal DNA mediated by integrase variant C3 as determined by PCR amplification from colonies growing on 100ug/mL (1X) and 70ug/mL (0.7X) ampicillin plates after being transformed with appropriate minicircles described in Figure 2. Chromosomal integration was verified using PCR primers flanking the endogenous attB site, ecoliatattBF (SEQ ID NO: 16) and ecoliatattBR (SEQ ID NO: 17). In absence of integration, the expected PCR product is ~200bp (as seen for colonies 1,5,6). Integration of the lactamase gene cassette results in a PCR product of ~1650bp (colonies 2-4, 7-10).

Figure 3B compares integration into attB of *E. coli* mediated by parental Int-h/218 or C3 integrase. When parental Int-h/218 was used, only 2 colonies were observed (both of which, i.e. 100%, had correctly inserted

lactamase cassette). In the case of C3, 27 colonies were observed. 10 of these were tested and 9 (90%) showed correctly inserted lactamase cassette. Therefore, of the 27 colonies, we can predict that ~90% (~24 colonies) contained the correctly inserted lactamase cassette. This corresponds to an improvement of ~12 fold (24 divided by 2).

Figure 3C shows the nucleotide sequence of an integrant *E. coli* colony generated using C3 harbouring the lactamase cassette. Bacterial chromosomal DNA flanking the cassette is in lower case. The *attL* and *attR* sites generated through recombination of *attB* and *attP* are underlined and in bold. The lactamase open reading frame is in bold.

Figure 4A shows the improved intramolecular recombination activity of parental Int-h/218 and indicated lambda integrase mutants expressed using an in vitro transcription/translation system for cognate (*attB/attP*) and non-cognate (*attH/attPH* and *attH4x/attP4x*) sites. Recombination is denoted relative to parental Int-h/218 efficiency with *attB/P* substrate (set to 1). Mutant lambda integrase proteins of the invention (C2 and C3) are more efficient at performing the respective recombination reactions. C2 denotes lambda integrase mutants with the D336V mutation. I43F C2 denotes the C2 lambda integrase mutants with an additional I43F mutation. C3 denotes lambda integrase mutants with the I43F, E319G, D336V mutations. N=2, bars indicate means +/- SD.

Figure 4B shows intramolecular recombination activity of parental Int-h/218 and C3 integrase on indicated substrates. Intramolecular recombination was carried out with 5µg of purified recombinant integrase protein incubated with 10ng plasmid substrate containing either

attB/attP sites, *attH/attPH* sites or *attH4x/attP4x* sites. The reaction volume was 25µL and was carried out for 1.5 hours at 37°C in recombination buffer (100mM Tris pH7.5, 500mM NaCl, 25mM DTT, 10mM EDTA, 5mg/mL bovine serum albumin). The reaction was diluted 1/10 before taking 2µL for real-time PCR quantification of recombination efficiency. Real-time PCR quantification was carried out with 250nM each of primers pLIR-F1 (SEQ ID NO: 27) and pLIR-R1 (SEQ ID NO: 28) in a final volume of 20µL with SsoAdvanced™ Universal SYBR® Green Supermix. The activities of the recombinant integrase proteins are presented relative to activity of Int-h/218 on *attB/attP* plasmid substrate (set as value of 1). Error bars indicate average +/- SD of 2 independent experiments.

Figure 5A shows the intra-molecular recombination efficiency of the mutant integrases C2 without or with a C-terminal nuclear localization sequence (C2-N) of an episomal plasmid substrate in the HT1080 cell line with *attB* and *attP* sites. Y-axis denotes percentage eGFP positive cells (indicative of recombination between *attB/attP* sites) and activities are presented relative to the transfection efficiency as determined by EGFP vector transfection (100%). The mutant C2 recombined *attB* x *attP* more efficiently than Int-h/218.

Figure 5B shows the intra-molecular recombination efficiency of the mutant integrases C2 without or with a C-terminal nuclear localization sequence (C2-N) of an episomal plasmid substrate in the HT1080 cell line with *attH* and *attPH* sites. Y-axis denotes percentage eGFP positive cells (indicative of recombination between *attH/PH* sites) and activities are presented relative to the transfection efficiency as determined by EGFP vector

transfection (100%). The mutant C2 recombined *attH* and *attPH* more efficiently than Int-h/218.

Figure 5C shows the intra-molecular recombination efficiency of the mutant integrases of an episomal plasmid substrate in the HT1080 cell line with *attB* and *attP* sites. Y-axis denotes percentage eGFP positive cells (indicative of recombination between *attB/attP* sites) and activities are presented relative to the transfection efficiency as determined by EGFP vector transfection (100%). The mutant C3 recombined *attB* x *attP* more efficiently than Int-h/218 or a codon-optimized Int-h/218 (opt Int).

Figure 5D shows the intra-molecular recombination efficiency of the mutant integrase C3 of an episomal plasmid substrate in the HT1080 cell line with *attH* and *attPH* sites. Y-axis denotes percentage eGFP positive cells (indicative of recombination between *attH/PH* sites) and activities are presented relative to the transfection efficiency as determined by EGFP vector transfection (100%). The mutant C3 recombined *attH* and *attPH* more efficiently than Int-h/218 or a codon-optimized Int-h/218 (opt Int).

Figure 6A shows the PCR results of screening for *attH4x* and *attP4x* recombination events at L1 loci in the HT1080 clones. PCR amplifications of the expected size (~1200bp; for the *attL* site generated through recombination between *attH4x* and *attP4x*) were detected for the HT1080 clones 3, 19 and 21. W, no DNA template control; HT, negative control (genomic DNA from the parental HT1080 cells); Ina, genomic DNA from puromycin resistant colonies obtained through co-transfection of pPGKssPuro-*attP4x* and pCMVssKZ-Inactivie Int (plasmid expressing integrase with an inactivating mutation wherein

the amino acid residue tyrosine at sequence position 342 is replaced by the amino acid alanine); +, positive control (genomic DNA from HT1080 clone having an *attH4x* X *attP4x* integration event in L1 element); M, 100 bp DNA ladder; 1 to 33, genomic DNA from puromycin resistant HT1080 colonies obtained through co-transfection of pPGKssPuro-*attP4x* and pCMVssKZ-IntC3-CNLS.

Figure 6B shows the nucleotide sequence of *attL* site generated through recombination between the *attH4x* and *attP4x* in the HT1080 clones 3 and 19. Human genomic DNA sequence flanking the *attL* sequence is in lower case. HOP' sequence is in italics, bold and underlined. The *attL* sequence is underlined and in bold. The PGK promoter sequence (part of the pPGKssPuro-*attP4x* targeting vector and driving the expression of Puromycin resistance gene) is in upper case. Genomic locus of the targeted *attH4x* sequence in each clone is specified.

Examples

Non-limiting examples of the invention, including the best mode, and a comparative example will be further described in greater detail by reference to specific Examples, which should not be construed as in any way limiting the scope of the invention.

Materials

SEQ ID NO: 1

Int-h/218

MGRRRSHERRDLPPNLYIRNNGYYCYRDPRTGKEFGLGRDRRIAITEAIQANIELFSG
HKHKPLTARINSDNSVTLHSWLDREYKILASRGIKQKTLINYMSKIKAIRGLPDAPL
EDITTKEIAAMLNGYIDEGKAASAKLIRSTLSDAFREAIAEGHITTNHVAATRAAKSK
VRRSRLTADEYLKIYQAAESSPCWLRRLAMELAVVTGQRVGDLCKMKWSDIVDGYLYVE
QSKTGVKIAIPTALHIDALGISMKETLCKEILGGETIIASTRREPLSSGTVSRYFM

RARKASGLSFEGDPPTFHELRSLSARLYEKQISDKFAQHLLGHKSDTMASQYRDDGRGR
EWDKIEIK

SEQ ID NO: 2

- 5 C2 integrase mutant:
MGRRRSHERRDLPPNLYIRNNGYYCYRDPRTGKEFGLGRDRRIAITEAIQANIELFSG
HKHKPLTARINSDNSVTLHSLDRYEKILASRGIKQKTLINYMSKIKAIRRGLPDAPL
EDITTKEIAAMLNGYIDEGKAASAKLIRSTLSDAFREAI AEGHITTNHVAATRAAKSK
VRRSRLTADEYLKIYQAAESSPCWLRRLAMELAVVTGQRVGDLCKMKWSDIVDGYLYVE
10 QSKTGVKIAIPTALHIDALGISMKETLDKCKEILGGETIIASTREPLSSGTVSRYFM
RARKASGLSFEGDPPTFHELRSLSARLYEKQISDKFAQHLLGHKSVTMASQYRDDGRGR
EWDKIEIK

SEQ ID NO: 3

- 15 C3 integrase mutant:
MGRRRSHERRDLPPNLYIRNNGYYCYRDPRTGKEFGLGRDRRFAITEAIQANIELFSG
HKHKPLTARINSDNSVTLHSLDRYEKILASRGIKQKTLINYMSKIKAIRRGLPDAPL
EDITTKEIAAMLNGYIDEGKAASAKLIRSTLSDAFREAI AEGHITTNHVAATRAAKSK
VRRSRLTADEYLKIYQAAESSPCWLRRLAMELAVVTGQRVGDLCKMKWSDIVDGYLYVE
20 QSKTGVKIAIPTALHIDALGISMKETLDKCKEILGGETIIASTREPLSSGTVSRYFM
RARKASGLSFEGDPPTFHELRSLSARLYGKQISDKFAQHLLGHKSVTMASQYRDDGRGR
EWDKIEIK

SEQ ID NO: 4

- 25 Lambda integrase:
ATGGGAAGAAGGCGAAGTCATGAGCGCCGGGATTTACCCCTAACCTTTATATAAGAA
ACAATGGATATTACTGCTACAGGGACCCAAGGACGGGTAAAGAGTTTGGATTAGGCAG
AGACAGGCGAATCGCAATCACTGAAGCTATACAGGCCAACATTGAGTTATTTTCAGGA
CACAAACACAAGCCTCTGACAGCGAGAATCAACAGTGATAATTCGGTTACGTTACATT
30 CATGGCTTGATCGCTACGAAAAAATCCTGGCCAGCAGAGGAATCAAGCAGAAGACACT
CATAAATTACATGAGCAAAATTAAAGCAATAAGGAGGGGTCTGCCTGATGCTCCACTT
GAAGACATCACCACAAAAGAAATTGCGGCAATGCTCAATGGATACATAGACGAGGGCA
AGGCGGCGTCAGCCAAGTTAATCAGATCAACACTGAGCGATGCATTCCGAGAGGCAAT
AGCTGAAGGCCATATAACAACAAACCATGTCGCTGCCACTCGCGCAGCAAAATCAAAG

GTAAGGAGATCAAGACTTACGGCTGACGAATACCTGAAAATTTATCAAGCAGCAGAAT
 CATCACCATGTTGGCTCAGACTTGCAATGGAAGTGGCTGTTGTTACCGGGCAACGAGT
 TGGTGATTTATGCAAAATGAAGTGGTCTGATATCGTAGATGGATATCTTTATGTCGAG
 CAAAGCAAAACAGGCGTAAAAATTGCCATCCCAACAGCATTGCATATTGATGCTCTCG
 5 GAATATCAATGAAGGAAACACTTGATAAATGCAAAGAGATTCTTGGCGGAGAAACCAT
 AATTGCATCTACTCGTCGCGAACCGCTTTCATCCGGCACAGTATCAAGGTATTTTATG
 CGCGCACGAAAAGCATCAGGTCTTTCCTTCGAAGGGGATCCGCCTACCTTTCACGAGT
 TGCGCAGTTTGTCTGCAAGACTCTATGAGAAGCAGATAAGCGATAAGTTTGCTCAACA
 TCTTCTCGGGCATAAGTCGGACACCATGGCATCACAGTATCGTGATGACAGAGGCAGG
 10 GAGTGGGACAAAATTGAAATCAAATAA

SEQ ID NO: 5

attB: CTGCTTTTTT ATACTAACTT G

15 SEQ ID NO: 6

attP: CAGCTTTTTT ATACTAAGTT G

SEQ ID NO: 7

attH: CTGCTTTCTT ATACCAAGTG G

20

SEQ ID NO: 8

attPH: CAGCTTTCTT ATACCAAGTT G

SEQ ID NO: 9

25 *attP4X*: CAGCTTTATT TCATTAAGTT G

SEQ ID NO: 10

petF2: CATCGGTGATGTCGGCGAT

30 SEQ ID NO: 11

petR: CGGATATAGTTCCTCCTTTCAGCA

SEQ ID NO: 12

attP-F: cacagaattcCGT CTG TTA CAG GTC ACT AAT ACC ATC T

SEQ ID NO: 13

attPSOE-R: ACA TTT CCC CGA AAA GTG CCA CCT GAA CAT CAC CGG
GAA ATC AAA TAA TGA T

5

SEQ ID NO: 14

TEM1prom-F: TTC AGG TGG CAC TTT TCG GGG AAA TGT

SEQ ID NO: 15

10 TEM1prom-R: TGT GGA ATT CCT ACA CTA GAA GGA CAG TAT TTG
GTA TCT GC

SEQ ID NO: 16

EcoliAttB-F: CTG AAA ATG TGT TCA CAG GTT GCT

15

SEQ ID NO: 17

EcoliattB-R: GCA ATG CCA TCT GGT ATC ACT

SEQ ID NO: 18

20 C2 gene sequence:

ATGGGAAGAAGGCGAAGTCATGAGCGCCGGGATTTACCCCCTAACCTTTATATAAGAA
ACAATGGATATTACTGCTACAGGGACCCAAGGACGGGTAAAGAGTTTGGATTAGGCAG
AGACAGGCGAATCGCAATCACTGAAGCTATACAGGCCAACATTGAGTTATTTTCAGGA
CACAAACACAAGCCTCTGACAGCGAGAATCAACAGTGATAATTCCGTTACGTTACATT
25 CATGGCTTGATCGCTACGAAAAATCCTGGCCAGCAGAGGAATCAAGCAGAAGACACT
CATAAATTACATGAGCAAAATTAAAGCAATAAGGAGGGGTCTGCCTGATGCTCCACTT
GAAGACATCACCACAAAAGAAATTGCGGCAATGCTCAATGGATACATAGACGAGGGCA
AGGCGGCGTCAGCCAAGTTAATCAGATCAACGCTGAGCGATGCATTCCGAGAGGCAAT
AGCTGAAGGCCATATAACAACAAACCATGTCGCTGCCACTCGCGCAGCAAAGTCAAAG
30 GTAAGGAGATCAAGACTTACGGCTGACGAATACCTGAAAATTTATCAAGCAGCAGAAT
CATCACCATGTTGGCTCAGACTTGCAATGGAAGTGGCTGTTGTTACCGGGCAACGAGT
TGGTGACTTGTGCAAAATGAAGTGGTCTGATATCGTAGATGGATATCTTTATGTGCGAG
CAAAGCAAAACAGGCGTAAAAATTGCCATCCCAACAGCATTGCATATTGATGCTCTCG
GAATATCAATGAAGGAAACACTTGATAAATGCAAAGAGATTCTTGCGGAGAAACCAT

AATTGCATCTACTCGTCGCGAACCGCTTTCATCCGGCACAGTATCAAGGTATTTTATG
 CGCGCACGAAAAGCATCAGGTCTTTCCTTCGAAGGGGATCCGCCTACCTTTCACGAGT
 TGCGCAGTTTGTCTGCAAGACTCTATGAGAAGCAGATAAGCGATAAGTTTGCTCAACA
 TCTTCTCGGGCATAAGTCGGTCACCATGGCATCACAGTATCGTGATGACAGAGGCAGG
 5 GAGTGGGACAAAATTGAAATCAAATAA

SEQ ID NO: 19

C3 gene sequence:

ATGGGAAGAAGGCGAAGTCATGAGCGCCGGGATTTACCCCTAACCTTTATATAAGAA
 10 ACAATGGATATTACTGCTACAGGGACCCAAGGACGGGTAAAGAGTTTGGATTAGGCAG
 AGACAGGCGATTTCGCAATCACTGAAGCTATACAGGCCAACATTGAGTTATTTTCAGGA
 CACAAACACAAGCCTCTGACAGCGAGAATCAACAGTGATAATTCCGTTACGTTACATT
 CATGGCTTGATCGCTACGAAAAAATCCTGGCCAGCAGAGGAATCAAGCAGAAGACACT
 CATAAATTACATGAGCAAAATTAAAGCAATAAGGAGGGGTCTGCCTGATGCTCCACTT
 15 GAAGACATCACCACAAAAGAAATTGCGGCAATGCTCAATGGATACATAGACGAGGGCA
 AGGCGGCGTCAGCCAAGTTAATCAGATCAACGCTGAGCGATGCATTCCGAGAGGCAAT
 AGCTGAAGGCCATATAACAACAAACCATGTCGCTGCCACTCGCGCGGCAAAGTCAAAG
 GTAAGGAGATCAAGACTTACGGCTGACGAATACCTGAAAATTTATCAAGCAGCAGAAT
 CATCACCATGTTGGCTCAGACTTGCAATGGAAGTGGCTGTTGTTACCGGGCAACGAGT
 20 TGGTGACTTGTGCAAAATGAAGTGGTCTGATATCGTAGATGGATATCTTTATGTTCGAG
 CAAAGCAAAACAGGCGTAAAAATTGCCATCCCAACAGCATTGCATATTGATGCTCTCG
 GAATATCAATGAAGGAAACACTTGATAAATGCAAAGAGATTCTTGCGGAGAAACCAT
 AATTGCATCTACTCGTCGCGAACCGCTCTCATCCGGCACAGTATCAAGGTATTTTATG
 CGCGCACGAAAAGCATCAGGTCTTTCCTTCGAAGGGGATCCGCCTACCTTTCACGAGT
 25 TGCGCAGTTTGTCTGCAAGACTCTATGGGAAGCAGATAAGCGATAAGTTTGCTCAACA
 TCTTCTCGGGCATAAGTCGGTCACCATGGCATCACAGTATCGTGATGACAGAGGCAGG
 GAGTGGGACAAAATTGAAATCAAATAA

SEQ ID NO: 20

30 C3 minicircle:

CATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCC
 GGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGAC
 TCACTATAGGGGAATTGTGAGCGGATAACAATCCCCTCTAGAAATAATTTTGTTTAA
 CTTTAAGAAGGAGATATACATATGGGAAGAAGGCGAAGTCATGAGCGCCGGGATTTAC

CCCCTAACCTTTATATAAGAAACAATGGATATTACTGCTACAGGGACCCAAGGACGGG
TAAAGAGTTTGGATTAGGCAGAGACAGGCGATTTCGCAATCACTGAAGCTATACAGGCC
AACATTGAGTTATTTTCAGGACACAAACACAAGCCTCTGACAGCGAGAATCAACAGTG
ATAATTCCGTTACGTTACATTCATGGCTTGATCGCTACGAAAAAATCCTGGCCAGCAG
5 AGGAATCAAGCAGAAGACACTCATAAATTACATGAGCAAAATTAAAGCAATAAGGAGG
GGTCTGCCTGATGCTCCACTTGAAGACATCACCACAAAAGAAATTGCGGCAATGCTCA
ATGGATACATAGACGAGGGCAAGGCGGCGTCAGCCAAGTTAATCAGATCAACGCTGAG
CGATGCATTCCGAGAGGCAATAGCTGAAGGCCATATAACAACAAACCATGTGCTGCC
ACTCGCGCGGCAAAGTCAAAGGTAAGGAGATCAAGACTTACGGCTGACGAATACCTGA
10 AAATTTATCAAGCAGCAGAATCATCACCATGTTGGCTCAGACTTGCAATGGAACCTGGC
TGTTGTTACCGGGCAACGAGTTGGTGACTTGTGCAAAATGAAGTGGTCTGATATCGTA
GATGGATATCTTTATGTGCGAGCAAAGCAAAACAGGCGTAAAAATTGCCATCCCAACAG
CATTGCATATTGATGCTCTCGGAATATCAATGAAGGAAACACTTGATAAATGCAAAGA
GATTCTTGCGCGGAGAAACCATAATTGCATCTACTCGTCGCGAACCCTCTCATCCGGC
15 ACAGTATCAAGGTATTTTATGCGCGCACGAAAAGCATCAGGTCTTTCCTTCGAAGGGG
ATCCGCCTACCTTTTACGAGTTGCGCAGTTTGTCTGCAAGACTCTATGGGAAGCAGAT
AAGCGATAAGTTTGCTCAACATCTTCTCGGGCATAAGTCGGTCACCATGGCATCACAG
TATCGTGATGACAGAGGCAGGGAGTGGGACAAAATTGAAATCAAACATCATCACCATC
ACCACTAATGAGAATTCgagctccgtcgacaagcttgcggccgcactcgagcaccacc
20 accaccaccactgagatccggctgctaacaagcccgaaaggaagctgagttggctgc
tgccaccgctgagcaataactagcataacccttggggcctctaaacgggtcttgagg
ggtttttttgcgtgaaaggaggaactatatccg

SEQ ID NO: 21

25 attP-TEM1:
cacagaattcCGtctgttacagggtcactaataccatctaagtagttgattcatagtga
ctgcatatatattgtgttttacagtattatgtagtctgttttttatgcaaaatctaattt
aatatatattgatatttatatcattttacgtttctcgttcagctttttataactaagttg
gcattataaaaaagcatttgcttatcaatttggttgcaacgaacagggtcactatcagtca
30 aaataaaatcattatttgATTTCCCGGTGATGttcagggtggcacttttcggggaaatg
tgcgcggaaccctatattgtttatttttctaaatacattcaaatatgtatccgctcat
gagacaataaccctgataaatgcttcaataatattgaaaaaggaagagtatgagtatt
caacatttccgtgtcgcccttattcccttttttgcggcattttgccttcctgtttttg
ctcaccagaaacgctggtgaaagtaaaagatgctgaagatcagttgggtgcacgagt

gggttacatcgaactggatctcaacagcggtaagatccttgagagttttcgccccgaa
 gaacgttttccaatgatgagcacttttaaaagtctgctatgtggcgcggtattatccc
 gtattgacgccgggcaagagcaactcggtcgccgcatacactattctcagaatgactt
 ggttgagtactcaccagtcacagaaaagcatcttacggatggcatgacagtaagagaa
 5 ttatgcagtgctgccataaccatgagtataacactgcggccaacttacttctgacaa
 cgatcggaggaccgaaggagctaaccgcttttttgcacaacatgggggatcatgtaac
 tcgccttgatcgttggaaccggagctgaatgaagccataccaaacgacgagcgtgac
 accacgatgcctgcagcaatggcaacaacgttgcgcaaactattaactggcgaactac
 ttactctagcttcccggcaacaattaatagactggatggaggcggataaagttgcagg
 10 accacttctgcgctcggcccttcgggctggctggtttattgctgataaatctggagcc
 ggtgagcgtgggtctcgcggtatcattgcagcactggggccagatggtaagccctccc
 gtatcgtagttatctacacgacggggagtcaggcaactatggatgaacgaaatagaca
 gatcgtgagataggtgcctcactgattaagcattggtaactgtcagaccaagtttac
 tcatatatacttttagattgattttaaacttcatttttaattttaaaggatctaggtga
 15 agatcctttttgataatctcatgacaaaaatcccttaacgtgagttttcgttccactg
 agcgtcagacccccgtagaaaagatcaaaggatcttcttgagatccttttttctgcgc
 gtaatctgctgcttgcaaacaaaaaaaccaccgctaccagcgggtggtttgtttgccgg
 atcaagagctaccaactctttttccgaaggtaactggcttcagcagagcgcagatacc
 aaatactgtccttctagtgtagccgtagttagg

20

SEQ ID NO: 22

HOP': ATGCTTTATTTTCATTAAGTTG

SEQ ID NO: 23

25 attL:GCATTATAAAAAAGCATTGCTTATCAATTTGTTGCAACGAACAGGTCCTATC
 AGTCAAAATACAATCATTATTTGATTTCAATTTTGTCCCACTCCCTCCCG

SEQ ID NO: 24

PGK promoter:

30 AATTCTACCGGGTAGGGGAGGCGCTTTTCCCAAGGCAGTCTGGAGCATGCGCTTTAGC
 AGCCCCGCTGGGCACTTGGCGCTACACAAGTGGCCTCTGGCCTCGCACACATTCCACA
 TCCACCGGTAGGCGCCAACCGGCTCCGTTCTTTGGTGGCCCCCTTCGCGCCACCTTCTA
 CTCCTCCCCTAGTCAGGAAGTTCCCCCCCCGCCCCGCAGCTCGCGTCGTGCAGGACGTG

ACAAATGGAAGTAGCACGTCTCACTAGTCTCGTGCAGATGGACAGCACCGCTGAGCAA
TGGAAGCGGGTAGGCCTTTGGGGCAGCGGCCAATAGCAGCT

SEQ ID NO: 25

5 HOP' attH4X_F1: GAGTGTTTTCCAACCTTGGTTCCATT

SEQ ID NO: 26

PuroRev24: CACCGTGGGCTTGTA CTCTCGGTC

10

SEQ ID NO: 27

pLIR-F1: CTGCATCGATT CAGCTAGCTG

SEQ ID NO: 28

15 pLIR-R1: CTGATAGTGACCTGTTCGTTGC

SEQ ID NO: 29

pPGKssPuro-attP4x (targeting vector):

20 gaattcctctgttacagggtcactaataccatctaagtagttgattcatagtgactgca
tatgttggtgttttacagtattatgtagtctgttttttatgcaaaatctaatttaatat
attgatatttatatcattttacgtttctcgttcagctttatttcattaagttggcatt
ataaaaaagcattgcttatcaatttggttgcaacgaacagggtcactatcagtcaaaata
aatcattatatttgatttcaattttgtccactccctcccgaattctaccgggtagggg
25 aggcgcttttcccaaggcagtcctggagcatgcgcttttagcagccccgctggcacttg
cgctacacaagtggcctctggcctcgcacacattccacatccaccggtagcgccaacc
ggctccgttctttggtggcccccttcgcgccacttctactcctcccctagtcaggaagt
ttcccccccgccccgcagctcgcgtcgtgcaggacgtgacaaatggaagtagcacgtc
tcactagtctcgtgcagatggacagcaccgctgagcaatggaagcgggtaggcctttg
30 gggcagcggccaatagcagctttgctccttcgctttctgggctcagaggctgggaagg
ggtgggtccgggggcgggctcaggggcgggctcaggggcggggcgggcgcccgaagg
cctccggaggcccggtattctgcacgcttcaaaagcgcacgtctgccgcgctgttctc
ctcttcctcatctccgggcctttcgaccaattcgctgtctgcgagggccagctgttgg
ggtgagtactccctctcaaaagcgggcatgacttctgcgctaagattgtcagtttcca

aaaacgaggaggatttgatattcacctggcccgcggtgatgcctttgaggggtggccgc
gtccatctggtcagaaaagacaatctttttgttgtcaagcttgagggtgtggcaggctt
gagatctggccatacacttgagtgacaatgacatccactttgcctttctctccacagg
tgtccactcccagggtccaactgcagatgaccgagtacaagcccacgggtgcgcctcgcc
5 acccgcgacgacgtcccccgggcggtacgcaccctcgccgcccgcgttcgcccactacc
ccgccacgcgccacaccgtcgacccggaccgccacatcgagcgggtcaccgagctgca
agaactcttcctcacgcgcgtcgggctcgacatcggcaagggtgtgggtcgcgagcagc
ggcgccgcggtggcggtctggaccacgcccggagagcgtcgaagcgggggcggtgttcg
ccgagatcgggcccgcgcatggccgagttgagcgggtcccggtggccgcgagcaaca
10 gatggaaggcctcctggcgccgcaccggcccaaggagcccgcgtgggtcctggccacc
gtcggcgctctcgcccaccaccagggcaagggtctgggcagcgccgtcgtgctccccg
gagtggaggcgccgagcgcgcgggggtgccgccttcctggagacctccgcgccccg
caacctccccttctacgagcggctcggcttcaccgtcaccgcccagctcgaggtgcc
gaaggaccgcgcacctgggtgcatgaccgcgaagcccgggtgcctgatctagagctcgct
15 gatcagcctcgactgtgccttctagtgtccagccatctgttggttgcccctccccgt
gccttccttgaccctggaagggtgccactcccactgtcctttcctaataaaatgaggaa
attgcatcgcattgtctgagtaggtgtcattctattctggggggtgggggtggggcagg
acagcaagggggaggattgggaagacaatagcaggcatgctggggatgcgggtgggctc
tatggcttctgaggcggaagaaccagctggggctcgagatccactagttctagcctc
20 gaggctagagcggccgccaccgcggtggagctccaattcgccctatagtgagtcgtat
tacgcgcgctcactggccgctcgttttacaacgtcgtgactgggaaaaccctggcgta
cccaacttaatcgccttgacgcacatcccccttcgccagctggcgtaatagcgaaga
ggcccgccaccgatcgcccttcccaacagttgcgcagcctgaatggcgaatgggacgcg
ccctgtagcggcgcatthaagcgcggcggtgtggtgttacgcgcagcgtgaccgcta
25 cacttgccagcgccttagcgcggcgtcctttcgctttcttcccttcctttctcgccac
gttcgccggctttccccgtcaagctctaaatcgggggctccctttagggttccgattt
agtgttttacggcacctcgacccccaaaaaacttgattaggggtgatgggtcacgtagtg
ggccatcgccctgatagacggtttttcgcccttgacgttgagtcacgttctttaa
tagtggactcttggttccaaactggaacaacactcaaccctatctcgggtctattctttt
30 gatttataagggattttgcccatttcggcctattgggttaaaaaatgagctgatttaac
aaaaatttaacgcgaatttttaacaaaatattaacgcttacaatttaggtggcactttt
cggggaaatgtgcgcggaaccctatttggttatttttctaaatacattcaaatatgt
atccgctcatgagacaataaccctgataaatgcttcaataatattgaaaaaggaagag
tatgagtattcaacatttccgtgtcgccttattcccttttttgcggcattttgcctt

cctgttttttgctcaccagaaacgctgggtgaaagtaaaagatgctgaagatcagttgg
gtgcacgagtggtttacatcgaactggatctcaacagcggtaagatccttgagagttt
tcgccccgaagaacgttttccaatgatgagcacttttaagttctgctatgtggcgcg
gtattatccccgtattgacgccgggcaagagcaactcggtcgccgcatacactattctc
5 agaatgacttgggttgagtactcaccagtcacagaaaagcatcttacggatggcatgac
agtaagagaattatgcagtgctgccataaccatgagtgataaactgcggccaactta
cttctgacaacgatcggaggaccgaaggagctaaccgcttttttgcacaacatggggg
atcatgtaactcgccttgatcggttggaaccggagctgaatgaagccataccaaacga
cgagcgtgacaccacgatgcctgtagcaatggcaacaacgttgcgcaaactattaact
10 ggcgaactacttactctagcttcccggcaacaattaatagactggatggaggcggata
aagttgcaggaccacttctgcgctcggcccttccggctggctggtttattgctgataa
atctggagccggtgagcgtgggtctcgcggtatcattgcagcactggggccagatggg
aagccctcccgtatcgtagttatctacacgacggggagtcaggcaactatggatgaac
gaaatagacagatcgctgagataggtgcctcactgattaagcattggtaactgtcaga
15 ccaagtttactcatatatacttttagattgattttaaacttcatttttaatttaaagg
atctaggtgaagatcctttttgataatctcatgaccaaatacccttaacgtgagtttt
cgttccactgagcgtcagaccccgtagaaaagatcaaaggatcttcttgagatccttt
ttttctgcgcgtaatctgctgcttgcaaacaaaaaaccaccgctaccagcgggtggtt
tgtttgccggatcaagagctaccaactctttttccgaaggtaactggcttcagcagag
20 cgcagataccaaatactgtccttctagtgtagccgtagttaggccaccacttcaagaa
ctctgtagcaccgcctacatacctcgtctgctaatacctgttaccagtggctgctgcc
agtggcgataagtcgtgtcttaccgggttgactcaagacgatagttaccgggataagg
cgcagcggctcgggctgaacggggggttcgtgcacacagcccagcttgagcgaacgac
ctacaccgaactgagatacctacagcgtgagctatgagaaagcgccacgcttcccga
25 gggagaaaggcggacaggtatccggtaagcggcagggtcggaacaggagagcgcacga
gggagcttccagggggaaacgcctgggtatctttatagtcctgtcgggtttcgccacct
ctgacttgagcgtcgatttttgtgatgctcgtcagggggcgaggcctatggaaaaac
gccagcaacgcggcctttttacggttcctggccttttgctggccttttgctcacatgt
tctttcctgcgttatccctgattctgtggataaccgtattaccgcctttgagtgagc
30 tgataccgctcgcgcagccgaacgaccgagcgcagcagtcagtgagcaggaagcg
gaagagcgcccaatacgcgaacgcctctccccgcgcgttgggcgattcattaatgca
gctggcacgacaggtttcccgactggaaagcgggcagtgagcgaacgcgaattaatgt
gagttagctcactcattaggcaccaccaggctttacactttatgcttccggctcgtatg
ttgtgtggaattgtgagcggataacaatttcacacaggaaacagctatgaccatgatt

acgccaaagcgcgcaattaaccctcactaaaggggaacaaaagctgggtaccggggccccc
cctcgagggtcgacggtatcgataagcttgatatac

SEQ ID NO: 30

5

pCMVssKZ-IntC3-CNLS (the integrase expression plasmid):

gaattcctcacatgttctttcctgcgttatcccctgattctgtggataaccgtattac
cgccatgcattagttattaatagtaatacaattacgggggtcattagttcatagccata
10 tatggagttccgcgttacataacttacggtaaatggcccgctggctgaccgcccac
gacccccgcccattgacgtcaataatgacgtatgttcccatagtaacgccaataggga
ctttccattgacgtcaatgggtggagtatttacggtaaaactgccacttggcagtaca
tcaagtgtatcatatgccaagtacgccccctattgacgtcaatgacggtaaatggccc
gcctggcattatgccagtacatgaccttatgggactttcctacttggcagtacatct
15 acgtattagtcacgctattaccatgggtgatgcgggttttggcagtacatcaatgggcg
tgatagcgggtttgactcacggggattttccaagtctccaccccattgacgtcaatggg
agtttggtttggcaccaaaatcaacgggactttccaaaatgtcgtaacaactccgccc
cattgacgcaaatgggcggtaggcgtgtacgggtgggaggtctatataagcagagctgg
tttagtgaaccgtcagatccgctagcaattcgctgtctgcgagggccagctgttgggg
20 tgagtactccctctcaaaagcgggcatgacttctgcgctaagattgtcagttttccaaa
aacgaggaggatttgatattcacctggcccgcggtgatgcctttgagggtggccgcgt
ccatctgggtcagaaaagacaatctttttgttgtaagcttgagggtgtggcaggcttga
gatctggccatacacttgagtgaacaatgacatccactttgcctttctctccacagggtg
tccactcccagggtccaactgcagctcgagggtccaccatgggaagaaggcgaagtcag
25 agcgcggggttttacccttaacctttatataagaacaatggatattactgctacag
ggaccaaggacgggtaaaagagtttggattaggcagagacaggcgattcgcaatcact
gaagctatacaggccaacattgagttattttcaggacacaaacacaagcctctgacag
cgagaatcaacagtgataattccgttacgttacattcatggcttgatcgctacgaaaa
aatcctggccagcagaggaatcaagcagaagacactcataaattacatgagcaaaaatt
30 aaagcaataaggaggggtctgcctgatgctccacttgaagacatcaccacaaaagaaa
ttgcggaatgctcaatggatacatagacgagggcaaggcggcgctcagccaagttaat
cagatcaacgctgagcgtatgcattccgagaggcaatagctgaaggccatataacaaca
aaccatgtcgctgccactcgcgcggaaggtcaaaggtaaggagatcaagacttacgg
ctgacgaatacctgaaaatttatcaagcagcagaatcatcaccatgttggctcagact
35 tgcaatggaactggctgttggtaccgggcaacgagttgggtgacttgtgcaaaatgaag
tggtctgatatacgtagatggatatctttatgtcgagcaaaagcaaaacaggcgtaaaaa

ttgccatcccaacagcattgcatattgatgctctcggaaatatcaatgaaggaaacact
tgataaatgcaaagagattcttggcggagaaaccataattgcatctactcgtcgcgaa
ccgctctcatccggcacagtatcaaggatTTTTATGCGCGCACGAAAAGCATCAGGTC
tttccttcgaaggggatccgcctacctttcacgagttgCGCAGTTTGTCTGCAAGACT
5 ctatgggaagcagataagcgataagtttgctcaacatcttctcgggcataagtcgggtc
accatggcatcacagtatcgtgatgacagaggcaggagtgggacaaaattgaaatca
aatccggaggcggccctaagaagaagagaaaaggatgataatctagagctcgtgatc
agcctcgactgtgccttctagttgccagccatctgttgtttgccctccccctgcct
tccttgaccctggaagggtgccactcccactgtcctttcctaataaaatgaggaaattg
10 catcgcattgtctgagtaggtgtcattctattctggggggtgggggtggggcaggacag
caagggggaggattgggaagacaatagcaggcatgctggggatgCGGTGGGCTCTATG
gcttctgaggcggaaagaaccagctggggctcgagatccactagttctagcctcgagg
ctagagcggccgccaccgCGGTGGAGCTCCAATTCGCCCTATAGTGAGTCGTATTACG
CGCGCTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCA
15 acttaatcgcttgcagcacatccccctttcgccagctggcgtaatagcgaagaggcc
cgcaccgatcgcccttcccaacagttgCGCAGCCTGAATGGCGAATGGGACGCGCCCT
gtagcggcgcattaagcgcggcgggtgtggtgggttacgCGCAGCGTGACCGCTACACT
tgccagcgccttagcgcCGCTCCTTTCGCTTTCTTCCCTTCTCTCGCCACGTTT
gccggctttccccgtcaagctctaaatcgggggctccctttagggttccgatttagtg
20 ctttacggcacctcgacccccaaaaacttgattagggtgatgggttcacgtagtgggcc
atcgccctgatagacgggtttttcgccctttgacggttgaggtccacgttctttaatagt
ggactcttggtccaaactggaacaacactcaaccctatctcgggtctattcttttgatt
tataagggtattttgccgatttcggcctattgggttaaaaaatgagctgatttaacaaaa
atttaacgcgaattttaacaaaatattaacgcttacaatttaggtggcacttttcggg
25 gaaatgtgcgcggaaccctatttgtttatttttctaaatacattcaaatatgtatcc
gctcatgagacaataaccctgataaatgcttcaataatattgaaaaaggaagagtatg
agtattcaacattttccgtgtcgccttattcccttttttgCGGCATTTTGCCTTCCTG
TTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGC
ACGAGTGGGTACATCGAAGTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGC
30 CCCGAAGAAGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTAT
TATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAA
TGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTA
AGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTC
TGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCA

tgtaactcgccttgatcggttggaaccggagctgaatgaagccataccaaacgacgag
 cgtgacaccacgatgcctgtagcaatggcaacaacgttgcgcaaactattaactggcg
 aactacttactctagcttcccggcaacaattaatagactggatggaggcggataaagt
 tgcaggaccacttctgcgctcggcccttccggctggctggtttattgctgataaatct
 5 ggagccggtgagcgtgggtctcgcggtatcattgcagcactggggccagatggtaagc
 cctcccgatcgtagttatctacacgacggggagtcaggcaactatggatgaacgaaa
 tagacagatcgctgagataggtgcctcactgattaagcattggtaactgtcagaccaa
 gtttactcatatatacttttagattgatttaaaacttcattttttaatttaaaaggatct
 aggtgaagatcctttttgataatctcatgacccaaaatcccttaacgtgagttttcggt
 10 ccactgagcgtcagaccccgtagaaaagatcaaaggatcttcttgagatccttttttt
 ctgcgcgtaatctgctgcttgcaaacaaaaaaaccaccgctaccagcgggtggtttggt
 tgccggatcaagagctaccaactctttttccgaaggtaactggcttcagcagagcgca
 gataccaataactgtccttctagtgtagccgtagttaggccaccacttcaagaactct
 gtagcaccgcctacatacctcgtctctgctaatacctgttaccagtggtgctgccagt
 15 gcgataagtcgtgtcttaccggggttgactcaagacgatagttaccggataaggcgca
 gcggtcgggctgaacgggggggttcgtgcacacagcccagcttggagcgaacgacctac
 accgaactgagatacctacagcgtgagctatgagaaagcgccacgcttcccgaaggga
 gaaaggcggacaggtatccggtaagcggcagggctcggaacaggagagcgcacgagggga
 gcttccaggggggaaacgcctgggtatctttatagtcctgtcgggtttcgccacctctga
 20 cttgagcgtcgatttttgtgatgctcgtcaggggggaggagcctatggaaaaacgcca
 gcaacgcggcctttttacggttcctggccttttgctggccttttgctcacatgttctt
 tcctgcgttatccccctgattctgtggataaccgtattaccgcctttgagtgagctgat
 accgctcggcgagccgaacgaccgagcgcagcgagtcagtgagcggaggaagcggaag
 agcgcccaatacgcgaaccgcctctccccgcgcggttgccgattcattaatgcagctg
 25 gcacgacaggtttcccgactggaaagcgggcagtgagcgcaacgcaattaatgtgagt
 tagctcactcattaggcaccccaggctttacactttatgcttccggctcgtatgttggt
 gtggaattgtgagcggataacaatttcacacaggaaacagctatgaccatgattacgc
 caagcgcgcaattaaccctcactaaaggggaacaaaagctgggtaccggggccccccctc
 gaggtcgacggtatcgataagcttgatatc

30

SEQ ID NO: 31

attH4X: acgctttattttcattaagttg

35

Example 1: Rapid E.coli chromosomal integration

The present example follows the methodology depicted in Figure 2. C3INT-HIS-PET22b(+) was amplified with petF2 (SEQ ID NO: 10) and petR (SEQ ID NO: 11) and the PCR product subsequently intramolecularly ligated to produce a C3INT-HIS minicircle. attP-PET22b(+) was amplified with attP-F (SEQ ID NO: 12) and attPSOE-R (SEQ ID NO: 13) while PET22b(+) was amplified with TEM1prom-F (SEQ ID NO: 14) and TEM1promR (SEQ ID NO: 15) which produced PCR products encoding *attP* and ampicillin-resistant gene respectively. Splice overlap extension PCR (SOE-PCR) was carried out with these two PCR products using attP-F (SEQ ID NO: 12) and TEM1prom-R (SEQ ID NO: 15). The PCR product was subsequently intramolecularly ligated to produce attP-TEM1 minicircle. 100ng of C3INT-HIS minicircle and 100ng attP-TEM1 minicircle were combined and electroporated to 25µL electrocompetent TG1 cells. The cells were allowed to recover for 1hr before being plated on varying concentrations of ampicillin-LB agar plates (0.01mg/mL, 0.02mg/mL, 0.05mg/mL, 0.07mg/mL and 0.1mg/mL). Incubation was carried out at 37°C for 12-14hrs to allow for expression of C3 integrase and chromosomal integration of the ampicillin-resistance cassette by C3 integrase. Colony PCR was carried out with EcoliAttB-F (SEQ ID NO: 16) and EcoliAttB-R (SEQ ID NO: 17), TEM1prom-F (SEQ ID NO: 14) and EcoliAttB-F (SEQ ID NO: 16), or TEM1prom-F (SEQ ID NO: 14) and EcoliAttB-R (SEQ ID NO: 17) to verify the presence of chromosomal integration of the ampicillin-resistance cassette. The PCR products were also sequenced with the same primers to confirm the results. The sequencing indicated a correct integration event into the chromosomal *attB* site (Figure 3C).

Example 2: Recombination activity of the parental and integrase variants C2 and C3

The present example demonstrates the recombination activity of the parental Int-h/218 and selected mutants (C2, C3 and indicated variants thereof). Figure 4A depicts results from an *in vitro* intramolecular recombination reaction using integrases produced by *in vitro* transcription/translation. Plasmids encoding the respective integrase (Int-h/218, C2, C3 or variant thereof) were amplified using primers IntrRBS-F and INTstop-R, and the PCR products re-amplified with primers Univeral and INTstop-R to get integrase amplicons with T7 promoter and ribosome binding site required for *in vitro* transcription-translation (IVT). 20ng of each integrase amplicon was expressed using *PURExpress® In Vitro Protein Synthesis Kit* in a total volume of 9µL at 30°C for 1 hour. Intramolecular recombination was then carried out by adding 10ng plasmid substrate containing either *attB/attP* sites, *attPH/attH* sites or *attH4x/attP4x* sites (Figure 1) to a total volume of 10µL. The mixture was allowed to incubate for 2 hours at 37°C. The reaction was subsequently diluted 1/10 before taking 1µL for real-time PCR quantification of recombination efficiency. Real-time PCR quantification was carried out with 250nM each of primers pLIR-F1 (SEQ ID NO: 27) and pLIR-R1 (SEQ ID NO: 28) in a final volume of 20µL with SsoAdvanced™ Universal SYBR® Green Supermix. The activities of the recombinant integrase proteins are presented relative to activity of WT Int-h/218 on *attB/attP* plasmid substrate (set as value of 1). Error bars indicate standard deviation of two independent experiments.

The results show significant increases in recombination efficiency for the C2 and C3 integrases

compared to parental Int-h/218. The data in Figure 4 show the strong contribution of the E319G mutation present in C3 on the efficiency of intramolecular recombination. Removing this mutation from C3 yields I43F C2 which shows
5 ~3 fold reduced activity on all substrate pairs tested. The contribution of the I43F mutation for intramolecular recombination is not readily apparent, as addition of this to C2 or removal from C3 does not lead to any significant change in recombination efficiency. However, it could
10 impact on other parameters such as intermolecular recombination *in vivo* and/or protein stability.

Figure 4B depicts results from *in vitro* intramolecular recombination reaction using integrases produced recombinantly in *E. coli*. Plasmids expressing
15 integrase Int-h/218 and C3 were transformed into *E. coli* BL21(DE3)pLysS (Invitrogen) competent cells. The bacterial cells were grown in LB medium at 37°C and induced at OD_{600nm} of ~0.6 with 0.5mM IPTG at 30°C for 6 hours. The cells were then harvested by centrifugation, resuspended in 50mM Tris
20 pH 8.0, 1M NaCl, 20mM Imidazole and lysed by sonication. The cell lysate was clarified by high-speed centrifugation and the supernatant was then applied to a 1mL HisTrapTM FF column (GE Healthcare) pre-equilibrated in binding buffer of 50mM Tris-HCl pH 8.0, 1M NaCl, 20mM Imidazole, 0.5mM
25 EDTA and 2mM DTT. The column was washed with binding buffer and the integrase proteins were eluted off the column with 50mM Tris-HCl pH 8.0, 1M NaCl, 500mM Imidazole, 0.5mM EDTA and 2mM DTT. Collected fractions were analyzed by SDS-PAGE gel and the appropriate fractions were
30 dialyzed and concentrated in 50mM Tris pH 8.0, 1M NaCl, 0.5mM EDTA and 2mM DTT using Amicon-Ultra (10 kDa MWCO) prior to storage at -80°C.

Intramolecular recombination was carried out with 5µg of purified recombinant integrase protein incubated with

10ng plasmid substrate containing either *attB/attP* sites, *attPH/attH* sites or *attH4x/attP4x* sites. The reaction volume was 25µL and was carried out for 1.5hours at 37°C in recombination buffer (100mM Tris pH7.5, 500mM NaCl, 25mM DTT, 10mM EDTA, 5mg/mL bovine serum albumin). The reaction was diluted 1/10 before taking 2µL for real-time PCR quantification of recombination efficiency. Real-time PCR quantification was carried out with 250nM each of primers pLIR-F1 (SEQ ID NO: 27) and pLIR-R1 (SEQ ID NO: 28) in a final volume of 20µL with SsoAdvanced™ Universal SYBR® Green Supermix. The activities of the recombinant integrase proteins are presented relative to activity of Int-h/218 (WT) on *attB/attP* plasmid substrate (set as value of 1). Error bars indicate average +/- SD of 2 independent experiments. The data again show increased recombination on all substrates tested for the C3 integrase compared to Int-h/218 parent.

The improved recombination activities of C2 and C3 observed in these experiments (Figures 4A and 4B) correlate with those seen in cell-based assays (Figure 5).

Example 3: Cell culture conditions, transfection procedure and selection of puromycin-resistant recombinants for endogenous *attH4x* targeting in HT1080 cells

For endogenous targeting in the HT1080 cell line, 3x10⁶ cells were seeded in Dulbecco's Modified Eagle Medium [DMEM (Life technologies) supplemented with 10% FBS, 1% L-glutamine and 100 Units/mL of Penicillin and Streptomycin each] per 10cm cell culture dish a day before transfection to obtain 70-90% confluence at the time of transfection. Transfections were done using Lipofectamine 2000 reagent (Life technologies). Plasmid DNA-Lipid complexes were prepared by mixing 5ng of the targeting vector (pPGKssPuro-*attP4x* (SEQ ID NO: 29)) and 100ng of the

integrase expression plasmid (pCMVssKZ-IntC3-CNLS (SEQ ID NO: 30)) diluted in 75µl of Opti-MEM medium with 2.5µl of Lipofectamine 2000 reagent diluted in 75µl of Opti-MEM medium (Life technologies) and incubating for 20 minutes at room temperature. The transfection mix was added onto the cells (under DMEM without antibiotics) and transfection was allowed to proceed for 4-6 hours following which the complexes were removed by replacing with fresh medium. 48 hours post-transfection, the cells were grown in growth medium containing 3µg Puromycin per ml to select for puromycin-resistant colonies. After 3 weeks of selection, puromycin-resistant colonies were picked and expanded. Genomic DNA was extracted using DNeasy Blood & Tissue Kit (Qiagen).

15

Example 4: Cell culture conditions, Transfection procedure and FACS analysis for episomal intra-molecular recombination assay

For the episomal intra-molecular recombination assays in HT1080 cell line, 3×10^5 cells were seeded in Dulbecco's Modified Eagle Medium [DMEM (Life technologies) supplemented with 10% FBS, 1% L-glutamine and 100 Units/mL of Penicillin and Streptomycin each] per well of 6 well plate a day before transfection to obtain 70-90% confluence at the time of transfection. Transfections were done using Lipofectamine 2000 reagent. For every transfection per well, plasmid DNA-Lipid complexes were prepared by mixing 1.5µg of pLIR and 1.5µg of the λ integrase expression plasmid diluted in 100µl of Opti-MEM medium with 6µl of Lipofectamine 2000 reagent diluted in 100µl of Opti-MEM medium and incubating for 20 minutes at room temperature. The transfection mix was added dropwise onto the cells (under DMEM without antibiotics) and transfection was allowed to proceed for 4-6 hours

following which the complexes were removed by replacing with fresh DMEM medium. 48-72 hours post-transfection, the cells were trypsinised and harvested with DMEM into eppendorf tubes, pelleted by centrifugation (at 1000x rcf for 5 minutes) and resuspended in 1ml fresh DMEM. GFP positive cell were quantified by FACS on a BD FACSCalibur™ machine (Becton-Dickinson).

Example 5: Identifying successful sequence specific recombination event

10

PCR was performed using GoTaq Flexi DNA polymerase (Promega) with primers HOP' attH4X_F1 (SEQ ID NO: 25) and PuroRev24 (SEQ ID NO: 26) and 200ng of genomic DNA as template per PCR reaction in 50µl volume. The following thermal cycling parameters were used for the PCR: an initial step of 95°C for 5 minutes, 35 cycles of 95°C for 1 minute, 57°C for 30 seconds and 72°C for 1 minute, and a final step of 72°C for 5 minutes. The PCR samples were analyzed by electrophoresis in 0.8% agarose gel in Tris-Boric acid-EDTA buffer.

20

Figure 6A depicts the PCR results of screening for attH4x and attP4x recombination events at the L1 loci in the HT1080 clones. PCR amplifications of the expected size (~1200bp; for the attL site generated through recombination between attH4x and attP4x) were detected for the HT1080 clones 3, 19 and 21. Figure 6B shows the nucleotide sequence of attL site generated through recombination between the attH4x and attP4x in the HT1080 clones 3 and 19.

25

30

Applications

The improved *in vitro* recombination by using the integrase variants of the present invention and the
5 *attH/attPH* and *attH4X/attP4X* substrate pairs indicates that the integrase variants described herein may be a useful reagent tool for biotechnology applications such as recombination-based cloning applications.

It will be apparent that various other modifications
10 and adaptations of the invention will be apparent to the person skilled in the art after reading the foregoing disclosure without departing from the spirit and scope of the invention and it is intended that all such modifications and adaptations come within the scope of the
15 appended claims.

Claims

1. A lambda integrase comprising an amino acid mutation at position 336 of the lambda integrase as set forth in SEQ ID NO: 1, wherein the amino acid residue aspartate at sequence position 336 is replaced by an aliphatic amino acid.
2. The lambda integrase according to claim 1, wherein the lambda integrase further comprises an amino acid mutation at position 319, 43 or both.
3. The lambda integrase according to claims 1 or 2, wherein the aliphatic amino acid at sequence position 336 is selected from the group consisting of isoleucine, leucine and valine.
4. The lambda integrase according to claim 3, wherein the lambda integrase comprises the amino acid mutation D336V in the lambda integrase as set forth in SEQ ID NO: 2.
5. The lambda integrase according to claim 4, wherein the amino acid residue glutamate at sequence position 319 is replaced by glycine; optionally wherein said lambda integrase comprises the amino acid mutations E319G and D336V in the lambda integrase as set forth in SEQ ID NO: 3.
6. The lambda integrase according to claims 1 to 5, wherein the lambda integrase comprises amino acid mutations at positions 43, 319 and 336 of the lambda integrase as set forth in SEQ ID NO: 1; optionally wherein the amino acid residue isoleucine at sequence position 43 is replaced by an aromatic amino acid; optionally wherein the aromatic amino acid is selected from the group consisting of phenylalanine, tyrosine and tryptophan.

7. The lambda integrase according to any one of claims 1 to 6, wherein the lambda integrase comprises amino acid mutations I43F, E319G and D336V in the lambda integrase as set forth in SEQ ID NO: 3.
8. A nucleic acid molecule comprising a nucleotide sequence encoding the lambda integrase according to any of claims 1 to 7.
9. The nucleic acid molecule according to claim 8, wherein the nucleic acid molecule is operably linked to a regulatory sequence to permit expression of the nucleic acid molecule; optionally wherein the regulatory sequence comprises a promoter sequence.
10. The nucleic acid molecule according to claims 8 or 9, wherein the nucleic acid molecule is located in a vector.
11. A host cell containing a nucleic acid molecule according to any one of claims 8 to 10.
12. A method of recombining a nucleic acid of interest into a target nucleic acid, the method comprising contacting a targeting nucleic acid comprising the nucleic acid of interest with the target nucleic acid in the presence of a lambda integrase as defined in any one of claims 1 to 7.
13. The method according to claim 12, wherein the target nucleic acid comprises DNA; optionally wherein the target nucleic acid comprises genomic DNA; optionally wherein the target nucleic acid comprises a sequence selected from the group consisting of an attH sequence (SEQ ID NO: 7) and an attH4X sequence (SEQ ID NO: 31); optionally wherein the targeting nucleic acid is a vector; optionally wherein the targeting nucleic acid comprises a sequence selected from the group consisting of an attPH sequence (SEQ ID NO: 8) and an attP4X sequence (SEQ ID NO: 9).

14. The method according to claims 12 or 13, wherein the recombination is performed in the presence of one or more cofactors; optionally wherein the cofactors is selected from the group consisting of XIS, FIS and IHF.
15. The method according to any one of claims 13 or 14, wherein the genomic DNA is comprised in a cell.
16. A sequence specific recombination kit comprising:
 - a. a targeting nucleic acid into which a nucleic acid of interest can be inserted, and
 - b. a lambda integrase according to any one of claims 1 to 7 or a nucleic acid as claimed in any one of claims 8 to 10.
17. The kit according to claim 16, further comprising at least one reagent for inserting said nucleic of interest into said targeting nucleic acid; optionally wherein said targeting nucleic acid comprises a sequence selected from the group consisting of an attPH sequence (SEQ ID NO: 8) and an attP4X sequence (SEQ ID NO: 9); optionally wherein said targeting nucleic acid comprises a sequence selected from the group consisting of an attH sequence (SEQ ID NO: 7) and an attH4X sequence (SEQ ID NO: 31).
18. The kit according to claims 16 or 17 further comprising buffer (s) and/or instructions for recombining said nucleic acid of interest with a given target nucleic acid.
19. The kit according to any one of claims 16 to 18 further comprising at least one reagent for determining a successful sequence specific recombination event; optionally wherein said reagent is a primer pair.