TRANSPOSITION OF NUCLEIC ACID CONSTRUCTS INTO EUKARYOTIC GENOMES WITH A TRANSPOSASE FROM AMYELOIS

CROSS REFERENCE TO RELATED APPLICATIONS

[001] The present application claims priority to U.S. Provisional Application No. 62/831,097 filed April 8, 2019; U.S. Provisional Application No. 62/873,342 filed July 12, 2019; and U.S. Provisional Application No. 62/990,568 filed March 17, 2020, each incorporated by reference in its entirety for all purposes.

REFERENCE TO A SEQUENCE LISTING

[002] The application refers to sequences disclosed in a txt file named 546917SEQLST.TXT, of 2,392,133 bytes, created April 6, 2020, incorporated by reference.

1. FIELD OF THE INVENTION

[003] The field of the present invention relates to configurations of DNA vectors for making stable modifications of the genomes of target cells, and the use of non-natural transposons and transposases.

2. BACKGROUND OF THE INVENTION

[004] The expression levels of genes encoded on a polynucleotide integrated into the genome of a cell depend on the configuration of sequence elements within the polynucleotide. The efficiency of integration and thus the number of copies of the polynucleotide that are integrated into each genome, and the genomic loci where integration occurs also influence the expression levels of genes encoded on the polynucleotide. The efficiency with which a polynucleotide may be integrated into the genome of a target cell can often be increased by placing the polynucleotide into a transposon.

[005] Transposons comprise two ends that are recognized by a transposase. The transposase acts on the transposon to remove it from one DNA molecule and integrate it into another. The DNA between the two transposon ends is transposed by the transposase along with the transposon ends. Heterologous DNA flanked by a pair of transposon ends, such that it is recognized and transposed by a transposase is referred to herein as a synthetic transposon. Introduction of a synthetic transposon and a corresponding transposase into the nucleus of a eukaryotic cell may result in transposition of the transposon into the genome of the cell. These outcomes are useful because they increase transformation efficiencies and

because they can increase expression levels from integrated heterologous DNA. There is thus a need in the art for hyperactive transposases and transposons.

[006] Transposition by a piggyBac-like transposase is perfectly reversible. The transposon is initially integrated at an integration target sequence in a recipient DNA molecule, during which the target sequence becomes duplicated at each end of the transposon inverted terminal repeats (ITRs). Subsequent transposition removes the transposon and restores the recipient DNA to its former sequence, with the target sequence duplication and the transposon removed. However, this is not sufficient to remove a transposon from a genome into which it has been integrated, as it is highly likely that the transposon will be excised from the first integration target sequence but transposed into a second integration target sequence in the genome. Transposases that are deficient for the integration (or transposition) function, on the other hand, can excise the transposon from the first target sequence, but will be unable to integrate into a second target sequence. Integration-deficient transposases are thus useful for reversing the genomic integration of a transposon.

[007] One application for transposases is for the engineering of eukaryotic genomes. Such engineering may require the integration of more than one different polynucleotide into the genome. These integrations may be simultaneous or sequential. When transposition into a genome of a first transposon comprising a first heterologous polynucleotide by a first transposase is followed by transposition into the same genome of a second transposon comprising a second heterologous polynucleotide by a second transposase, it is advantageous that the second transposase not recognize and transpose the first transposon. This is because the location of a polynucleotide sequence within the genome influences the expressibility of genes encoded on said polynucleotide, so transposition of the first transposon to a different chromosomal location by the second transposase could change the expression properties of any genes encoded on the first heterologous polynucleotide. There is therefore a need for a set of transposons and their corresponding transposases in which the transposases within the set recognize and transpose only their corresponding transposons, but not any other transposons in the set.

[008] Since its discovery in 1983, the piggyBac transposon and transposase from the looper moth *Trichoplusia ni* has been widely used for inserting heterologous DNA into the genomes of target cells from many different organisms. The piggyBac system is a particularly valuable transposase system because of: "its activity in a wide range of organisms, its ability to integrate multiple large transgenes with high efficiency, the ability to add domains to the transposase without loss of activity, and excision from the genome

without leaving a footprint mutation" (Doherty et al., Hum. Gene Ther. 23, 311-320 (2012), at p. 312, LHC, ¶ 2).

[009] The value and versatility of the piggyBac system has inspired significant efforts to identify other active piggyBac-like transposons (commonly referred to as piggyBac-like elements, or PLEs) but these have been largely unsuccessful. "Since piggyBac is one of the most popular transposons used for transgenesis, searching for new active PLEs has attracted lots of attention. However, only a few active PLEs have been reported to date." (Luo et al., BMC Molecular Biology 15, 28 (2014) http://www.biomedcentral.com/1471-2199/15/28. p.4 of 12, RHC, ¶ 1 "Discussion").

[0010] Although there are large numbers of homologs of piggyBac transposons and transposases in sequence databases, few active ones have been identified because the vast majority are inactivated by their hosts to avoid activity deleterious to the hosts as illustrated by the following excerpts: "Related piggyBac transposable elements have been found in plants, fungi and animals, including humans [125], although they are probably inactive due to mutation." (Munoz-Lopez & Garcia-Perez, Current Genomics 11, 115-128 (2010) at p. 120, RHC, ¶ 1). "It is believed that transposons invade a genome and subsequently spread throughout it during evolution. The ''selfish'' mobility of transposons is harmful to the host; hence, they are eliminated or inactivated by the host through natural selection. Even harmless transposons lose the activity eventually because of the absence of conservative selection for them. Thus, in general, transposons have a short life span in a host and they subsequently become fossils in the genome." (Hikosaka et al., Mol. Biol. Evol. 24, 2648-3656 (2007) at p. 2648, LHC, ¶ 1 "Introduction"). "Frequent movement of transposable elements in a genome is harmful (Belancio et al., 2008; Deininger & Batzer, 1999; Le Rouzic & Capy, 2006; Oliver & Greene, 2009). As a result, most transposable elements are inactivated shortly after they invade a new host." (Luo et al., Insect Science 18, 652-662 (2011) at p. 660, LHC, ¶1).

[0011] Three classes of piggyBac-like elements have been found: (1) those that are very similar to the original piggyBac from the looper moth (typically >95% identical at the nucleotide level), (2) those that are moderately related (typically 30-50% identical at the amino acid level), and (3) those that are very distantly related (Wu et al., Insect Science 15, 521-528 (2008) at p. 521, RHC. ¶ 2).

[0012] PiggyBac-like transposases highly related to the looper moth transposase have been described by several groups. They are extremely highly conserved. Very similar transposase sequences to the original piggyBac (95-98% nucleotide identity) have been reported in three different strains of the fruit fly *Bactrocera dorsalis* (Handler & McCombs,

Insect Molecular Biology 9, 605-612, (2000)). Comparably conserved piggyBac sequences have been found in other *Bactrocera* species (Bonizzoni et al., Insect Molecular Biology 16, 645-650 (2007)). Two species of noctuid moth (*Helicoverpa zea* and *Helicoverpa armigera*) and other strains of the looper moth *Trichoplusia ni* had genomic copies of the piggyBac transposase with 93-100% nucleotide identity to the original piggyBac sequence (Zimowska & Handler, Insect Biochemistry and Molecular Biology, 36, 421-428 (2006)). Zimowska & Handler also found multiple copies of much more significantly mutated (and truncated) versions of the piggyBac transposase in both *Helicoverpa* species, as well as a homolog in the armyworm *Spodptera frugiperda*. None of these groups attempted to measure any activity for these transposases. Wu et. al (2008), supra, reported isolating a transposase from *Macdunnoughia crassisigna* with 99.5% sequence identity with the looper moth piggyBac. They also demonstrated that this transposon and transposase are active, by showing that they could measure both excision and transposition. Their Discussion summarized previous results as follows: "Other reportedly closely related IFP2 class sequences were in various Bactrocera species, T. ni genome, Heliocoverpa armigera, and H. zea (Handler & McCombs, 2000; Zimowska & Handler, 2006; Bonizzoni et al., 2007). These sequences were partial fragments of piggyBac-like elements, and most of them were truncated or inactivated by accumulating random mutations." (Wu et. al., Insect Science 15, 521-528 (2008) at p. 526, LHC, ¶ 3.)

[0013] It has proved very difficult to identify active piggyBac-like transposases that are moderately related to the looper moth enzyme simply by looking at sequence. The presence of features that are known to be necessary: a full-length open reading frame, catalytic aspartate residues and intact ITRs, has not proven to be predictive of activity. "A large diversity of PLEs in eukaryotes has been documented in a computational analysis of genomic sequence data [citations omitted]. However, few elements were isolated with an intact structure consistent with function, and only the original IFP2 piggyBac has been developed into a vector for routine transgenesis." (Wu et al., Genetica 139, 149-154 (2011), at p. 152, RHC, ¶ 2.). Wu et al.'s group from Nanjing University (the "Nanjing group") published several papers over a 6-year period, each identifying moderately related piggyBac homologs. Although the Nanjing group showed in 2008 that they could measure both excision and transposition of the *Macdunnoughia crassisigna* transposon by its corresponding transposase, and in each subsequent paper they express the desire to identify novel active piggyBac-like transposases, they only show excision activity and that only for one transposase from *Aphis gossypii*. They conclude that the usefulness of this transposase "remains to be explored with further experiments" (Luo et. al. 2011, p. 660, LHC ¶ 2

"Discussion"). However, none of the other papers published by the Nanjing group in which piggyBac-like sequences were identified from a variety of other insects, show that any activity was found. Three papers identifying other putative active piggyBac-like transposases were published by a group at Kansas State University. None of these papers reports any activity data. Wang et al., Insect Molecular Biology 15, 435-443 (2006) found multiple copies of piggyBac-like sequences in the genome of the tobacco budworm *Heliothis virescens*. Many of these had obvious mutations or deletions that led the authors not to consider them to be candidate active transposases. Wang et. al., Insect Biochemistry and Molecular Biology 38, 490-498 (2008) reported more than 30 piggyBac-like sequences in the genome of the red flour beetle *Tribolium castaneum*. They concluded "All the TcPLEs identified here, except TcPLE1, were apparently defective due to the presence of multiple stop codons and/or indels in the putative transposase encoding regions." Even for TcPLE1 there was "no evidence supporting recent or current mobilization events" (p. 492, section 3.1, ¶ ¶ 2&3). Wang et al. (2010) used PCR to identify piggyBac-like sequences from the pink bollworm *Pectinophora gossypiella*. Again, they found many obviously defective copies, as well as one transposase with characteristics the authors believe to be consistent with activity (page 179, RHC, ¶ 2). But no follow up report indicating transposase activity can be found. Other groups have also attempted to identify active piggyBac-like transposases. These reports conclude with statements that the piggyBac-like elements identified are undergoing testing for activity, but there are no subsequent reports of success. For example, Sarkar et. al. (2003) conclude their Discussion by re-stating the value of novel active piggyBac-like transposons, and describing their ongoing efforts to identify one: "The mobility of the original T. ni piggyBac element in various insects suggests that piggyBac family transposons might prove to be useful genetic tools in organisms other than insects. We are currently isolating an intact piggyBac element from An. gambiae (AgaPB1) to test its mobility in various organisms." (Mol. Gen. Genomics 270, 173-180 at p. 179, LHC, ¶1). There appear to be no further published reports of this putative active transposase. Xu et al. analyzed the silkworm genome looking for piggyBac-like sequences (Xu et al., Mol Gen Genomics 276, 31-40 (2006)). They found 98 piggyBac-like sequences and performed various computational analyses of putative transposase sequence and ITR sequences. They conclude: "We have isolated several intact piggyBac-like elements from B. mori and are currently testing their activity and the feasibility of using them as transformation vectors." (p 38, RHC, ¶3). There appear to be no further published reports of these putative active transposases.

[0014] Four published papers discussing the third class of distantly related piggyBac-like transposases. The first three of these demonstrate only the excision part of the reaction and acknowledge that this is different from full transposition. Hikosaka et. al., Mol Biol Evol 24, 2648-2656 (2007) reported that " In the present study, we demonstrated that the Xtr-Uribo2 Tpase has excision activity toward the target transposon, although there is no evidence for the integration of the excised target into the genome thus far." (page 2654, RHC, ¶2). Luo et. al., Insect Science 18, 652-662 (2011) reported "These results demonstrated the activity of the Ago-PLE1.1 transposase in mediating the first step of the cut and- paste movement of the element" (page 658, LHC, ¶1). Daimon et. al., Genome 53, 585-593 (2010) discussed the transposase systems yabusabe-1 and yabusabe-W. Although Daimon et al. reported detecting an excision event by PCR, they also report screening approximately 100,000 recovered plasmids for the excision of yabusame-1 and yabusame-W without identifying a single recovered plasmid from which the elements had excised. By contrast Daimon reports the transposition frequency of wildtype piggyBac enzyme as around 0.3-1.4. Thus, it appears from Daimon et al. that the excision frequency of yabusabe-1 or -W is less than 0.001% (1:100,000). This is at least 2-3 orders of magnitude less than can be achieved with a wild-type piggyBac enzyme and even less than available genetically engineered variants of piggyBac transposase, which achieve ten- fold higher transposition than wildtype. The implied transposition frequency for yabasume-1 from Daimon et al. is also two orders of magnitude lower than random integration frequency in mammalian cells (which is of the order of 0.1%). Thus, Daimon et al. show that yabusame-1 was essentially inactive and would not be useful as a genetic engineering tool. Such a view likely underlies Daimon et al.'s own conclusion: "Although we could detect the excision event in the highly sensitive PCR-based assay, our data indicate that both elements have lost their excision activity almost entirely." This also suggests that the PCR-based excision assay used to show activity of Uribo2 and Ago-PLE1.1 is not predictive of transposition activity that will be useful for inserting heterologous DNA into the genome of a target cell. The only report of a fully active piggyBac-like transposase (competent for both excision and integration) of the third category of distantly related transposases to the original piggyBac transposase from *Trichoplusia Ni* is one from the bat *Myotis lucifugus* (Mitra et. al., Proc. Natl. Acad. Sci. 110, 234-239 (2013)). These authors used a yeast system to demonstrate both excision and transposition activities for the bat transposase. All of the work described here shows that it has been extremely difficult to identify fully active piggyBac-like transposases, even though there are a large

number of candidate sequences. There is therefore a need for new piggyBac-like transposons and their corresponding transposases.

3. SUMMARY OF THE INVENTION

[0015] Heterologous gene expression from polynucleotide constructs that stably integrate into a target cell genome can be improved by placing the expression polynucleotide between a pair of transposon ends: sequence elements that are recognized and transposed by transposases. DNA sequences inserted between a pair of transposon ends can be excised by a transposase from one DNA molecule and inserted into a second DNA molecule. A novel piggyBac-like transposon-transposase system is disclosed that is not derived from the looper moth *Trichoplusia ni*. It is derived from the naval orangeworm moth *Amyelois transitella* (the *Amyelois* transposase). The *Amyelois* transposon comprises sequences that function as transposon ends and that can be used in conjunction with a corresponding *Amyelois* transposase that recognizes and acts on those transposon ends, as a gene transfer system for stably introducing nucleic acids into the DNA of a cell. The gene transfer systems of the invention can be used in methods including but not limited to genomic engineering of eukaryotic cells, heterologous gene expression, gene therapy, cell therapy, insertional mutagenesis, or gene discovery.

[0016] Transposition may be effected using a polynucleotide comprising an open reading frame encoding an *Amyelois* transposase, the amino acid sequence of which is at least 90% identical to SEQ ID NO: 18, operably linked to a heterologous promoter. The heterologous promoter may be active in a eukaryotic cell. The heterologous promoter may be active in a mammalian cell. mRNA may be prepared from a polynucleotide comprising an open reading frame encoding an *Amyelois* transposase, the amino acid sequence of which is at least 90% identical to SEQ ID NO: 18, operably linked to a heterologous promoter that is active in an *in vitro* transcription reaction. The transposase may comprises a mutation as shown in columns C and D in Table 1, relative to the sequence of SEQ ID NO: 18. The transposase may comprise a mutation at an amino acid position selected from 79, 95, 115, 116, 121, 139, 166, 179, 187, 198, 203, 211, 238, 273, 304, 329, 345, 362, 366, 408, 416, 435, 458, 475, 483, 491, 529, 540, 560 and 563, relative to the sequence of SEQ ID NO: 18. The transposase may comprise a mutation selected from D79N, R95S, L115D, E116P, H121Q, K139E, V166F, G179N, W187F, P198R, L203R, N211R, E238D, L273M, L273I, D304R, D304K, Q329G, T345L, K362R, T366R, L408M, S416E, S435G, L458M, V475I, N483K, I491M, A529P, K540R, S560K and S563K, relative to the sequence of SEQ ID NO:

18, the transposase optionally including at least 2, 3, 4, or 5 selected from the group. The amino acid sequence of the transposase may be selected from SEQ ID NO: 96-170. The transposase can excise or transpose a transposon from SEQ ID NO: 192. The excision activity or transposition activity of the transposase is at least 2-fold higher than the activity of SEQ ID NO: 18, optionally 2 to 10 fold higher. Codons of the open reading frame of the transposase may be selected for mammalian cell expression. An isolated mRNA may encode a polypeptide, the amino acid sequence of which is at least 90% identical with SEQ ID NO: 18, and wherein the mRNA sequence comprises at least 10 synonymous codon differences relative to SEQ ID NO: 929, and wherein the mRNA sequence comprises at least 10 synonymous codon differences relative to SEQ ID NO: 929 at corresponding positions between the mRNA and SEQ ID NO:929, optionally wherein codons in the mRNA at the corresponding positions are selected for mammalian cell expression. The open reading frame encoding the transposase may further encode a heterologous nuclear localization sequence fused to the transposase. The open reading frame encoding the transposase may further encode a heterologous DNA binding domain (for example derived from a Crispr Cas system, or a zinc finger protein, or a TALE protein) fused to the transposase. A non-naturally occurring polynucleotide may encode a polypeptide, the sequence of which is at least 90% identical to SEQ ID NO: 18.

[0017] An *Amyelois* transposon comprises SEQ ID NO: 9 and SEQ ID NO: 10 flanking a heterologous polynucleotide. The transposon may further comprise a sequence at least 90% identical to SEQ ID NO: 13 on one side of the heterologous polynucleotide and a sequence at least 90% identical to SEQ ID NO: 16 on the other. The heterologous polynucleotide may comprise a heterologous promoter that is active in eukaryotic cells. The promoter may be operably linked to at least one or more of: i) an open reading frame; ii) a nucleic acid encoding a selectable marker; iii) a nucleic acid encoding a counter-selectable marker; iii) a nucleic acid encoding a regulatory protein; iv) a nucleic acid encoding an inhibitory RNA. The heterologous promoter may comprise a sequence selected from SEQ ID NOs: 474-558. The heterologous polynucleotide may comprise a heterologous enhancer that is active in eukaryotic cells. The heterologous enhancer may be selected from SEQ ID NOs: 453-473. The heterologous polynucleotide may comprise a heterologous intron that is spliceable in eukaryotic cells. The nucleotide sequence of the heterologous intron may be selected from SEQ ID NO: 561-621. The heterologous polynucleotide may comprise an insulator sequence. The nucleic acid sequence of the insulator may be selected from SEQ ID NO: 435-441. The heterologous polynucleotide may comprise two open reading frames, each operably

linked to a separate promoter. The heterologous polynucleotide may comprise a sequence selected from SEQ ID NOs: 745-928. The heterologous polynucleotide may comprise or encode a selectable marker. The selectable marker may be selected from a glutamine synthetase enzyme, a dihydrofolate reductase enzyme, a puromycin acetyltransferase enzyme, a blasticidin acetyltransferase enzyme, a hygromycin B phosphotransferase enzyme, an aminoglycoside 3'-phosphotransferase enzyme and a fluorescent protein. A eukaryotic cell whose genome comprises SEQ ID NO: 9 and SEQ ID NO: 10 flanking a heterologous polynucleotide is an embodiment of the invention. The cell may be an animal cell, a mammalian cell, a rodent cell or a human cell.

[0018] A transposon may be integrated into the genome of a eukaryotic cell by (a) introducing into the cell a transposon comprising SEQ ID NO: 9 and SEQ ID NO: 10 flanking a heterologous polynucleotide, (b) introducing into the cell a transposase, the sequence of which is at least 90% identical with SEQ ID NO: 18 wherein the transposase transposes the transposon to produce a genome comprising SEQ ID NO: 9 and SEQ ID NO: 10 flanking the heterologous polynucleotide. The transposase may be introduced as a polynucleotide encoding the transposase, the polynucleotide may be an mRNA molecule or a DNA molecule. The transposase may be introduced as a protein. The heterologous polynucleotide may also encode a selectable marker, and the method may further comprise selecting a cell comprising the selectable marker. The cell may be an animal cell, a mammalian cell, a rodent cell or a human cell. The human cell may be a human immune cell, for example a B-cell or a T-cell. The heterologous polynucleotide may encode a chimeric antigen receptor. A polypeptide may be expressed from the transposon integrated into the genome of the eukaryotic cell. The polypeptide may be purified. The purified polypeptide may be incorporated into a pharmaceutical composition.

4. BRIEF DESCRIPTION OF FIGURES

Figure 1. Structure of an *Amyelois* transposon.

[0019] An *Amyelois* transposon comprises a left transposon end and a right transposon end flanking a heterologous polynucleotide. The left transposon end comprises (i) a left target sequence, which is often 5'-TTAA-3', although a number of other target sequences are used at lower frequency (Li *et al.*, 2013. Proc. Natl. Acad. Sci vol. 110, no. 6, E478-487); (ii) a left ITR (e.g. SEQ ID NO: 9) and (iii) (optionally) additional left transposon end sequences (e.g. SEQ ID NO: 13). The right transposon end comprises (i) (optionally) additional right transposon end sequences (e.g. SEQ ID NO: 16); (ii) a right ITR (e.g. SEQ ID NO: 10) which is a perfect or imperfect repeat of the left ITR, but in inverted orientation

relative to the left ITR and (iii) a right target sequence which is typically the same as the left target sequence.

5. DETAILED DESCRIPTION OF THE INVENTION

5.1 DEFINITIONS

[0020] Use of the singular forms "a," "an," and "the" include plural references unless the context clearly dictates otherwise. Thus, for example, reference to "a polynucleotide" includes a plurality of polynucleotides, reference to "a substrate" includes a plurality of such substrates, reference to "a variant" includes a plurality of variants, and the like.

[0021] Terms such as "connected," "attached," "linked," and "conjugated" are used interchangeably herein and encompass direct as well as indirect connection, attachment, linkage or conjugation unless the context clearly dictates otherwise. Where a range of values is recited, it is to be understood that each intervening integer value, and each fraction thereof, between the recited upper and lower limits of that range is also specifically disclosed, along with each subrange between such values. The upper and lower limits of any range can independently be included in or excluded from the range, and each range where either, neither or both limits are included is also encompassed within the invention. Where a value being discussed has inherent limits, for example where a component can be present at a concentration of from 0 to 100%, or where the pH of an aqueous solution can range from 1 to 14, those inherent limits are specifically disclosed. Where a value is explicitly recited, it is to be understood that values which are about the same quantity or amount as the recited value are also within the scope of the invention. Where a combination is disclosed, each sub combination of the elements of that combination is also specifically disclosed and is within the scope of the invention. Conversely, where different elements or groups of elements are individually disclosed, combinations thereof are also disclosed. Where any element of an invention is disclosed as having a plurality of alternatives, examples of that invention in which each alternative is excluded singly or in any combination with the other alternatives are also hereby disclosed; more than one element of an invention can have such exclusions, and all combinations of elements having such exclusions are hereby disclosed.

[0022] Unless defined otherwise herein, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Singleton, et al., *Dictionary of Microbiology and Molecular Biology*, 2nd Ed., John Wiley and Sons, New York (1994), and Hale & Marham, *The Harper Collins Dictionary of Biology*, Harper Perennial, NY, 1991, provide one of skill with a general dictionary of many of the terms used in this invention. Although any methods and materials

similar or equivalent to those described herein can be used in the practice or testing of the present invention, the preferred methods and materials are described. Unless otherwise indicated, nucleic acids are written left to right in 5' to 3' orientation; amino acid sequences are written left to right in amino to carboxy orientation, respectively. The terms defined immediately below are more fully defined by reference to the specification as a whole.

[0023] The "configuration" of a polynucleotide means the functional sequence elements within the polynucleotide, and the order and direction of those elements.

[0024] The terms "corresponding transposon" and "corresponding transposase" are used to indicate an activity relationship between a transposase and a transposon. A transposase transposes its corresponding transposon. Many transposases may correspond with a single transposon, and many transposons may correspond with a single transposase.

[0025] The term "counter-selectable marker" means a polynucleotide sequence that confers a selective disadvantage on a host cell. Examples of counter-selectable markers include *sacB, rpsL, tetAR, pheS, thyA, gata-1, ccdB, kid* and *barnase* (Bernard, 1995, Journal/Gene, 162: 159-160; Bernard et al., 1994. Journal/Gene, 148: 71-74; Gabant et al., 1997, Journal/Biotechniques, 23: 938-941; Gababt et al., 1998, Journal/Gene, 207: 87-92; Gababt et al., 2000, Journal/ Biotechniques, 28: 784-788; Galvao and de Lorenzo, 2005, Journal/Appl Environ Microbiol, 71: 883-892; Hartzog et al., 2005, Journal/Yeat, 22:789-798; Knipfer et al., 1997, Journal/Plasmid, 37: 129-140; Reyrat et al., 1998, Journal/Infect Immun, 66: 4011-4017; Soderholm et al., 2001, Journal/Biotechniques, 31: 306-310, 312; Tamura et al., 2005, Journal/ Appl Environ Microbiol, 71: 587-590; Yazynin et al., 1999, Journal/FEBS Lett, 452: 351-354). Counter-selectable markers often confer their selective disadvantage in specific contexts. For example, they may confer sensitivity to compounds that can be added to the environment of the host cell, or they may kill a host with one genotype but not kill a host with a different genotype. Conditions which do not confer a selective disadvantage on a cell carrying a counter-selectable marker are described as "permissive". Conditions which do confer a selective disadvantage on a cell carrying a counter-selectable marker are described as "restrictive".

[0026] The term "coupling element" or "translational coupling element" means a DNA sequence that allows the expression of a first polypeptide to be linked to the expression of a second polypeptide. Internal ribosome entry site elements (IRES elements) and cis-acting hydrolase elements (CHYSEL elements) are examples of coupling elements.

[0027] The terms "DNA sequence", "RNA sequence" or "polynucleotide sequence" mean a contiguous nucleic acid sequence. The sequence can be an oligonucleotide of 2 to 20

nucleotides in length to a full length genomic sequence of thousands or hundreds of thousands of base pairs.

[0028] The term "expression construct" means any polynucleotide designed to transcribe an RNA. For example, a construct that contains at least one promoter which is or may be operably linked to a downstream gene, coding region, or polynucleotide sequence (for example, a cDNA or genomic DNA fragment that encodes a polypeptide or protein, or an RNA effector molecule, for example, an antisense RNA, triplex-forming RNA, ribozyme, an artificially selected high affinity RNA ligand (aptamer), a double-stranded RNA, for example, an RNA molecule comprising a stem-loop or hairpin dsRNA, or a bi-finger or multi-finger dsRNA or a microRNA, or any RNA). An "expression vector" is a polynucleotide comprising a promoter which can be operably linked to a second polynucleotide. Transfection or transformation of the expression construct into a recipient cell allows the cell to express an RNA effector molecule, polypeptide, or protein encoded by the expression construct. An expression construct may be a genetically engineered plasmid, virus, recombinant virus, or an artificial chromosome derived from, for example, a bacteriophage, adenovirus, adeno-associated virus, retrovirus, lentivirus, poxvirus, or herpesvirus. Such expression vectors can include sequences from bacteria, viruses or phages. Such vectors include chromosomal, episomal and virus-derived vectors, for example, vectors derived from bacterial plasmids, bacteriophages, yeast episomes, yeast chromosomal elements, and viruses, vectors derived from combinations thereof, such as those derived from plasmid and bacteriophage genetic elements, cosmids and phagemids. An expression construct can be replicated in a living cell, or it can be made synthetically. For purposes of this application, the terms "expression construct", "expression vector", "vector", and "plasmid" are used interchangeably to demonstrate the application of the invention in a general, illustrative sense, and are not intended to limit the invention to a particular type of expression construct.

[0029] The term "expression polypeptide" means a polypeptide encoded by a gene on an expression construct.

[0030] The term "expression system" means any *in vivo* or *in vitro* biological system that is used to produce one or more gene product encoded by a polynucleotide.

[0031] A “gene” refers to a transcriptional unit including a promoter and sequence to be expressed from it as an RNA or protein. The sequence to be expressed can be genomic or cDNA among other possibilities. Other elements, such as introns, and other regulatory sequences may or may not be present.

[0032] A "gene transfer system" comprises a vector or gene transfer vector, or a polynucleotide comprising the gene to be transferred which is cloned into a vector (a "gene transfer polynucleotide" or "gene transfer construct"). A gene transfer system may also comprise other features to facilitate the process of gene transfer. For example, a gene transfer system may comprise a vector and a lipid or viral packaging mix for enabling a first polynucleotide to enter a cell, or it may comprise a polynucleotide that includes a transposon and a second polynucleotide sequence encoding a corresponding transposase to enhance productive genomic integration of the transposon. The transposases and transposons of a gene transfer system may be on the same nucleic acid molecule or on different nucleic acid molecules. The transposase of a gene transfer system may be provided as a polynucleotide or as a polypeptide.

[0033] Two elements are "heterologous" to one another if not naturally associated. For example, a nucleic acid sequence encoding a protein linked to a heterologous promoter means a promoter other than that which naturally drives expression of the protein. A heterologous nucleic acid flanked by transposon ends or ITRs means a heterologous nucleic acid not naturally flanked by those transposon ends or ITRs, such as a nucleic acid encoding a polypeptide other than a transposase, including an antibody heavy or light chain. A nucleic acid is heterologous to a cell if not naturally found in the cell or if naturally found in the cell but in a different location (e.g., episomal or different genomic location) than the location described.

[0034] The term "host" means any prokaryotic or eukaryotic organism that can be a recipient of a nucleic acid. A "host," as the term is used herein, includes prokaryotic or eukaryotic organisms that can be genetically engineered. For examples of such hosts, see Maniatis et al., Molecular Cloning. A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y. (1982). As used herein, the terms "host," "host cell," "host system" and "expression host" can be used interchangeably.

[0035] A "hyperactive" transposase is a transposase that is more active than the naturally occurring transposase from which it is derived. "Hyperactive" transposases are thus not naturally occurring sequences.

[0036] 'Integration defective' or "transposition defective" means a transposase that can excise its corresponding transposon, but that integrates the excised transposon at a lower frequency into the host genome than a corresponding naturally occurring transposase.

[0037] An "IRES" or "internal ribosome entry site" means a specialized sequence that directly promotes ribosome binding, independent of a cap structure.

[0038] An 'isolated' polypeptide or polynucleotide means a polypeptide or polynucleotide that has been either removed from its natural environment, produced using recombinant techniques, or chemically or enzymatically synthesized. Polypeptides or polynucleotides of this invention may be purified, that is, essentially free from any other polypeptide or polynucleotide and associated cellular products or other impurities.

[0039] The terms "nucleoside" and "nucleotide" include those moieties which contain not only the known purine and pyrimidine bases, but also other heterocyclic bases which have been modified. Such modifications include methylated purines or pyrimidines, acylated purines or pyrimidines, or other heterocycles. Modified nucleosides or nucleotides can also include modifications on the sugar moiety, for example, where one or more of the hydroxyl groups are replaced with halogen, aliphatic groups, or is functionalized as ethers, amines, or the like. The term "nucleotidic unit" is intended to encompass nucleosides and nucleotides.

[0040] An "Open Reading Frame" or "ORF" means a portion of a polynucleotide that, when translated into amino acids, contains no stop codons. The genetic code reads DNA sequences in groups of three base pairs, which means that a double-stranded DNA molecule can read in any of six possible reading frames-three in the forward direction and three in the reverse. An ORF typically also includes an initiation codon at which translation may start.

[0041] The term "operably linked" refers to functional linkage between two sequences such that one sequence modifies the behavior of the other. For example, a first polynucleotide comprising a nucleic acid expression control sequence (such as a promoter, IRES sequence, enhancer or array of transcription factor binding sites) and a second polynucleotide are operably linked if the first polynucleotide affects transcription and/or translation of the second polynucleotide. Similarly, a first amino acid sequence comprising a secretion signal or a subcellular localization signal and a second amino acid sequence are operably linked if the first amino acid sequence causes the second amino acid sequence to be secreted or localized to a subcellular location.

[0042] The term "orthogonal" refers to a lack of interaction between two systems. A first transposon and its corresponding first transposase and a second transposon and its corresponding second transposase are orthogonal if the first transposase does not excise or transpose the second transposon and the second transposase does not excise or transpose the first transposon.

[0043] The term "overhang" or "DNA overhang" means the single-stranded portion at the end of a double-stranded DNA molecule. Complementary overhangs are those which will base-pair with each other.

[0044] A "piggyBac-like transposase" means a transposase with at least 20% sequence identity as identified using the TBLASTN algorithm to the piggyBac transposase from *Trichoplusia ni* (SEQ ID NO: 17), and as more fully described in Sakar, A. *et. al.*, (2003). Mol. Gen. Genomics 270: 173-180. "Molecular evolutionary analysis of the widespread piggyBac transposon family and related 'domesticated' species", and further characterized by a DDE-like DDD motif, with aspartate residues at positions corresponding to D268, D346, and D447 of *Trichoplusia ni* piggyBac transposase on maximal alignment. PiggyBac-like transposases are also characterized by their ability to excise their transposons precisely with a high frequency. A "piggyBac-like transposon" means a transposon having transposon ends which are the same or at least 80% and preferably at least 90, 95, 96, 97, 98 or 99% or 100% identical to the transposon ends of a naturally occurring transposon that encodes a piggyBac-like transposase. A piggyBac-like transposon includes an inverted terminal repeat (ITR) sequence of approximately 12-16 bases at each end, and is flanked on each side by a 4 base sequence corresponding to the integration target sequence which is duplicated on transposon integration (the Target Site Duplication or Target Sequence Duplication or TSD). PiggyBac-like transposons and transposases occur naturally in a wide range of organisms including *Argyrogramma agnate* (GU477713), *Anopheles gambiae* (XP_312615; XP_320414; XP_310729), *Aphis gossypii* (GU329918), *Acyrthosiphon pisum* (XP_001948139), *Agrotis ypsilon* (GU477714), *Bombyx mori* (BAD11135), *Ciona intestinalis* (XP_002123602), *Chilo suppressalis* (JX294476), *Drosophila melanogaster* (AAL39784), *Daphnia pulicaria* (AAM76342), *Helicoverpa armigera* (ABS18391), *Homo sapiens* (NP_689808), *Heliothis virescens* (ABD76335), *Macdunnoughia crassisigna* (EU287451), *Macaca fascicularis* (AB179012), *Mus musculus* (NP_741958), *Pectinophora gossypiella* (GU270322), *Rattus norvegicus* (XP_220453), *Tribolium castaneum* (XP_001814566) and *Trichoplusia ni* (AAA87375) and *Xenopus tropicalis* (BAF82026), although transposition activity has been described for almost none of these.

[0045] The terms "polynucleotide," "oligonucleotide," "nucleic acid" and "nucleic acid molecule" are used interchangeably to refer to a polymeric form of nucleotides of any length, and may comprise ribonucleotides, deoxyribonucleotides, analogs thereof, or mixtures thereof. This term refers only to the primary structure of the molecule. Thus, the term includes triple-, double- and single-stranded deoxyribonucleic acid ("DNA"), as well as triple-, double- and single-stranded ribonucleic acid ("RNA"). It also includes modified, for example by alkylation, and/or by capping, and unmodified forms of the polynucleotide. More particularly, the terms "polynucleotide," "oligonucleotide," "nucleic acid" and "nucleic

acid molecule" include polydeoxyribonucleotides (containing 2-deoxy-D-ribose), polyribonucleotides (containing D-ribose), including tRNA, rRNA, hRNA, siRNA and mRNA, whether spliced or unspliced, any other type of polynucleotide which is an N- or C-glycoside of a purine or pyrimidine base, and other polymers containing non-nucleotidic backbones, for example, polyamide (for example, peptide nucleic acids ("PNAs")) and polymorpholino (commercially available from the Anti-Virals, Inc., Corvallis, Oreg., as Neugene) polymers, and other synthetic sequence-specific nucleic acid polymers providing that the polymers contain nucleobases in a configuration which allows for base pairing and base stacking, such as is found in DNA and RNA. There is no intended distinction in length between the terms "polynucleotide," "oligonucleotide," "nucleic acid" and "nucleic acid molecule," and these terms are used interchangeably herein. These terms refer only to the primary structure of the molecule. Thus, these terms include, for example, 3'-deoxy-2', 5'-DNA, oligodeoxyribonucleotide N3' P5' phosphoramidates, 2'-O-alkyl-substituted RNA, double- and single-stranded DNA, as well as double- and single-stranded RNA, and hybrids thereof including for example hybrids between DNA and RNA or between PNAs and DNA or RNA, and also include known types of modifications, for example, labels, alkylation, "caps," substitution of one or more of the nucleotides with an analog, internucleotide modifications such as, for example, those with uncharged linkages (for example, methyl phosphonates, phosphotriesters, phosphoramidates, carbamates, or the like) with negatively charged linkages (for example, phosphorothioates, phosphorodithioates, or the like), and with positively charged linkages (for example, aminoalkylphosphoramidates, aminoalkylphosphotriesters), those containing pendant moieties, such as, for example, proteins (including enzymes (for example, nucleases), toxins, antibodies, signal peptides, poly-L-lysine, or the like), those with intercalators (for example, acridine, psoralen, or the like), those containing chelates (of, for example, metals, radioactive metals, boron, oxidative metals, or the like), those containing alkylators, those with modified linkages (for example, alpha anomeric nucleic acids, or the like), as well as unmodified forms of the polynucleotide or oligonucleotide.

[0046] A "promoter" means a nucleic acid sequence sufficient to direct transcription of an operably linked nucleic acid molecule. A promoter can be used with or without other transcription control elements (for example, enhancers) that are sufficient to render promoter-dependent gene expression controllable in a cell type-specific, tissue-specific, or temporal-specific manner, or that are inducible by external signals or agents; such elements, may be within the 3' region of a gene or within an intron. Desirably, a promoter is operably linked to

a nucleic acid sequence, for example, a cDNA or a gene sequence, or an effector RNA coding sequence, in such a way as to enable expression of the nucleic acid sequence, or a promoter is provided in an expression cassette into which a selected nucleic acid sequence to be transcribed can be conveniently inserted. A regulatory element such as promoter active in a mammalian cells means a regulatory element configurable to result in a level of expression of at least 1 transcript per cell in a mammalian cell into which the regulatory element has been introduced.

[0047] The term "selectable marker" means a polynucleotide segment or expression product thereof that allows one to select for or against a molecule or a cell that contains it, often under particular conditions. These markers can encode an activity, such as, but not limited to, production of RNA, peptide, or protein, or can provide a binding site for RNA, peptides, proteins, inorganic and organic compounds or compositions. Examples of selectable markers include but are not limited to: (1) DNA segments that encode products which provide resistance against otherwise toxic compounds (e.g., antibiotics); (2) DNA segments that encode products which are otherwise lacking in the recipient cell (e.g., tRNA genes, auxotrophic markers); (3) DNA segments that encode products which suppress the activity of a gene product; (4) DNA segments that encode products which can be readily identified (e.g., phenotypic markers such as beta-galactosidase, green fluorescent protein (GFP), and cell surface proteins); (5) DNA segments that bind products which are otherwise detrimental to cell survival and/or function; (6) DNA segments that otherwise inhibit the activity of any of the DNA segments described in Nos. 1-5 above (e.g., antisense oligonucleotides); (7) DNA segments that bind products that modify a substrate (e.g. restriction endonucleases); (8) DNA segments that can be used to isolate a desired molecule (e.g. specific protein binding sites); (9) DNA segments that encode a specific nucleotide sequence which can be otherwise non-functional (e.g., for PCR amplification of subpopulations of molecules); and/or (10) DNA segments, which when absent, directly or indirectly confer sensitivity to particular compounds.

[0048] Sequence identity can be determined by aligning sequences using algorithms, such as BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, Wis.), using default gap parameters, or by inspection, and the best alignment (i.e., resulting in the highest percentage of sequence similarity over a comparison window). Percentage of sequence identity is calculated by comparing two optimally aligned sequences over a window of comparison, determining the number of positions at which the identical residues occurs in both sequences

to yield the number of matched positions, dividing the number of matched positions by the total number of matched and mismatched positions not counting gaps in the window of comparison (i.e., the window size), and multiplying the result by 100 to yield the percentage of sequence identity. Unless otherwise indicated the window of comparison between two sequences is defined by the entire length of the shorter of the two sequences.

[0049] A "target nucleic acid" is a nucleic acid into which a transposon is to be inserted. Such a target can be part of a chromosome, episome or vector.

[0050] An "integration target sequence" or "target sequence" or "target site" for a transposase is a site or sequence in a target DNA molecule into which a transposon can be inserted by a transposase. The piggyBac transposase from *Trichoplusia ni* inserts its transposon predominantly into the target sequence 5'-TTAA-3'. Other useable target sequences for piggyBac transposons are 5'-CTAA-3', 5'-TTAG-3', 5'-ATAA-3', 5'-TCAA-3', 5'-AGTT-3', 5'-ATTA-3', 5'-GTTA-3', 5'-TTGA-3', 5'-TTTA-3', 5'-TTAC-3', 5'-ACTA-3', 5'-AGGG-3', 5'-CTAG-3', 5'-GTAA-3', 5'-AGGT-3', 5'-ATCA -3', , 5'- CTCC-3', 5'- TAAA-3', 5'-TCTC -3', 5'-TGAA -3', 5'- AAAT-3', 5'- AATC-3', 5'-ACAA -3', 5'- ACAT-3', 5'-ACTC -3', 5'-AGTG -3', 5'-ATAG -3', 5'- CAAA-3', 5'-CACA -3', 5'-CATA -3', 5'-CCAG -3', 5'-CCCA -3', 5'-CGTA -3', 5'-CTGA -3', 5'- GTCC-3', 5'- TAAG-3', 5'-TCTA -3', 5'-TGAG -3', 5'-TGTT -3', 5'-TTCA -3', 5'- TTCT-3' and 5'-TTTT -3' (Li *et al.*, 2013. Proc. Natl. Acad. Sci vol. 110, no. 6, E478-487). PiggyBac-like transposases transpose their transposons using a cut-and-paste mechanism, which results in duplication of their 4 base pair target sequence on insertion into a DNA molecule. The target sequence is thus found on each side of an integrated piggyBac-like transposon.

[0051] The term "translation" refers to the process by which a polypeptide is synthesized by a ribosome 'reading' the sequence of a polynucleotide.

[0052] A 'transposase' is a polypeptide that catalyzes the excision of a corresponding transposon from a donor polynucleotide, for example a vector, and (providing the transposase is not integration-deficient) the subsequent integration of the transposon into a target nucleic acid. An "*Amyelois* transposase" means a transposase with at least 80%, 90, 95, 96, 7, 98, 99 or 100% sequence identity to SEQ ID NO: 18, including hyperactive variants of SEQ ID NO: 18, that are able to transposase a corresponding transposon. A hyperactive transposase is a transposase that is more active than the naturally occurring transposase from which it is derived, for excision activity or transposition activity or both. A hyperactive transposase is preferably at least 1.5-fold more active, or at least 2-fold more active, or at least 5-fold more active, or at least 10-fold more active than the naturally occurring transposase from which it

is derived, e.g., 2-5 fold or 2-10 fold. A transposase may or more not be fused to one or more additional domains such as a nuclear localization sequence or DNA binding protein.

[0053] The term "transposition" is used herein to mean the action of a transposase in excising a transposon from one polynucleotide and then integrating it, either into a different site in the same polynucleotide, or into a second polynucleotide.

[0054] The term "transposon" means a polynucleotide that can be excised from a first polynucleotide, for instance, a vector, and be integrated into a second position in the same polynucleotide, or into a second polynucleotide, for instance, the genomic or extrachromosomal DNA of a cell, by the action of a corresponding trans-acting transposase. A transposon comprises a first transposon end and a second transposon end, which are polynucleotide sequences recognized by and transposed by a transposase. A transposon usually further comprises a first polynucleotide sequence between the two transposon ends, such that the first polynucleotide sequence is transposed along with the two transposon ends by the action of the transposase. This first polynucleotide in natural transposons frequently comprises an open reading frame encoding a corresponding transposase that recognizes and transposes the transposon. Transposons of the present invention are "synthetic transposons" comprising a heterologous polynucleotide sequence which is transposable by virtue of its juxtaposition between two transposon ends. Synthetic transposons may or may not further comprise flanking polynucleotide sequence(s) outside the transposon ends, such as a sequence encoding a transposase, a vector sequence or sequence encoding a selectable marker.

[0055] The term "transposon end" means the cis-acting nucleotide sequences that are sufficient for recognition by and transposition by a corresponding transposase. Transposon ends of piggyBac-like transposons comprise perfect or imperfect repeats such that the respective repeats in the two transposon ends are reverse complements of each other. These are referred to as inverted terminal repeats (ITR) or terminal inverted repeats (TIR). A transposon end may or may not include additional sequence proximal to the ITR that promotes or augments transposition.

[0056] The term "vector" or "DNA vector" or "gene transfer vector" refers to a polynucleotide that is used to perform a "carrying" function for another polynucleotide. For example, vectors are often used to allow a polynucleotide to be propagated within a living cell, or to allow a polynucleotide to be packaged for delivery into a cell, or to allow a polynucleotide to be integrated into the genomic DNA of a cell. A vector may further comprise additional functional elements, for example it may comprise a transposon.

5.2 DESCRIPTION

5.2.1 GENOMIC INTEGRATION

[0057] Expression of a gene from a heterologous polynucleotide in a eukaryotic host cell can be improved if the heterologous polynucleotide is integrated into the genome of the host cell. Integration of a polynucleotide into the genome of a host cell also generally makes it stably heritable, by subjecting it to the same mechanisms that ensure the replication and division of genomic DNA. Such stable heritability is desirable for achieving good and consistent expression over long growth periods. This is particularly important for cell therapies in which cells are genetically modified and then placed into the body. It is also important for the manufacturing of biomolecules, particularly for therapeutic applications where the stability of the host and consistency of expression levels is also important for regulatory purposes. Cells with gene transfer vectors, including transposon-based gene transfer vectors, integrated into their genomes are thus an important embodiment of the invention.

[0058] Heterologous polynucleotides may be more efficiently integrated into a target genome if they are part of a transposon (i.e., positioned between transposon ITRs), for example so that they may be integrated by a transposase A particular benefit of a transposon is that the entire polynucleotide between the transposon ITRs is integrated. A transposon comprising target sites flanking ITRs flanking a heterologous polynucleotide integrates at a target site in a genome to result in the genome containing the heterologous polynucleotide flanked by the ITRs, flanked by target sites. This is in contrast to random integration, where a polynucleotide introduced into a eukaryotic cell is often fragmented at random in the cell, and only parts of the polynucleotide become incorporated into the target genome, usually at a low frequency. The piggyBac transposon from the looper moth *Trichoplusia ni* has been shown to be transposed by its transposase in cells from many organisms (see e.g. Keith et al (2008) BMC Molecular Biology **9**:72 "Analysis of the piggyBac transposase reveals a functional nuclear targeting signal in the 94 c-terminal residues"). Heterologous polynucleotides incorporated into piggyBac-like transposons may be integrated into eukaryotic cells including animal cells, fungal cells or plant cells. Preferred animal cells can be vertebrate or invertebrate. Preferred vertebrate cells include cells from mammals including rodents such as rats, mice, and hamsters; ungulates, such as cows, goats or sheep; and swine. Preferred vertebrate cells also include cells from human tissues and human stem cells. Target cells types include hepatocytes, neural cells, muscle cells, blood cells,

embryonic stem cells, somatic stem cells, hematopoietic cells, embryos, zygotes, sperm cells (some of which are open to be manipulated in an *in vitro* setting) and immune cells including lymphocytes such as T cells, B cells and natural killer cells, T-helper cells, antigen-presenting cells, dendritic cells, neutrophils and macrophages. Preferred cells can be pluripotent cells (cells whose descendants can differentiate into several restricted cell types, such as hematopoietic stem cells or other stem cells) or totipotent cells (i.e., a cell whose descendants can become any cell type in an organism, e.g., embryonic stem cells). Preferred culture cells are Chinese hamster ovary (CHO) cells or Human embryonic kidney (HEK293) cells. Preferred fungal cells are yeast cells including *Saccharomyces cerevisiae* and *Pichia pastoris*. Preferred plant cells are algae, for example Chlorella, tobacco, maize and rice (Nishizawa-Yokoi et al (2014) Plant J. 77:454-63 "Precise marker excision system using an animal derived piggyBac transposon in plants").

[0059] Preferred gene transfer systems comprise a transposon in combination with a corresponding transposase protein that transposases the transposon, or a nucleic acid that encodes the corresponding transposase protein and is expressible in the target cell. A preferred gene transfer system comprises a synthetic *Amyelois* transposon and a corresponding *Amyelois* transposase.

[0060] A transposase protein can be introduced into a cell as a protein or as a nucleic acid encoding the transposase, for example as a ribonucleic acid, including mRNA or any polynucleotide recognized by the translational machinery of a cell; as DNA, e.g. as extrachromosomal DNA including episomal DNA; as plasmid DNA, or as viral nucleic acid. Furthermore, the nucleic acid encoding the transposase protein can be transfected into a cell as a nucleic acid vector such as a plasmid, or as a gene expression vector, including a viral vector. The nucleic acid can be circular or linear. mRNA encoding the transposase may be prepared using DNA in which a gene encoding the transposase is operably linked to a heterologous promoter, such as the bacterial T7 promoter, which is active *in vitro*. DNA encoding the transposase protein can be stably inserted into the genome of the cell or into a vector for constitutive or inducible expression. Where the transposase protein is transfected into the cell or inserted into the vector as DNA, the transposase encoding sequence is preferably operably linked to a heterologous promoter. There are a variety of promoters that could be used including constitutive promoters, cell-type specific promoters, organism-specific promoters, tissue-specific promoters, inducible promoters, and the like. Where DNA encoding the transposase is operably linked to a promoter and transfected into a target cell, the promoter should be operable in the target cell. For example if the target cell is a

mammalian cell, the promoter should be operable in a mammalian cell; if the target cell is a yeast cell, the promoter should be operable in a yeast cell; if the target cell is an insect cell, the promoter should be operable in an insect cell; if the target cell is a human cell, the promoter should be operable in a human cell; if the target cell is a human immune cell, the promoter should be operable in a human immune cell. All DNA or RNA sequences encoding piggyBac-like transposase proteins are expressly contemplated. Alternatively, the transposase may be introduced into the cell directly as protein, for example using cell-penetrating peptides (e.g. as described in Ramsey and Flynn (2015) Pharmacol. Ther. 154: 78-86 "Cell-penetrating peptides transport therapeutics into cells"); using small molecules including salt plus propanebetaine (e.g. as described in Astolfo et al (2015) Cell 161: 674-690); or electroporation (e.g. as described in Morgan and Day (1995) Methods in Molecular Biology 48: 63-71 "The introduction of proteins into mammalian cells by electroporation").

[0061] It is possible to insert the transposon into DNA of a cell through non-homologous recombination through a variety of reproducible mechanisms, and even without the activity of a transposase. The transposons described herein can be used for gene transfer regardless of the mechanisms by which the genes are transferred.

5.2.5 GENE TRANSFER SYSTEMS

[0062] Gene transfer systems comprise a polynucleotide to be transferred to a host cell. Preferably the polynucleotide comprises an *Amyelois* transposon and wherein the polynucleotide is to be integrated into the genome of a target cell.

[0063] When there are multiple components of a gene transfer system, for example the one or more polynucleotides comprising genes for expression in the target cell and optionally comprising transposon ends, and a transposase (which may be provided either as a protein or encoded by a nucleic acid), these components can be transfected into a cell at the same time, or sequentially. For example, a transposase protein or its encoding nucleic acid may be transfected into a cell prior to, simultaneously with or subsequent to transfection of a corresponding transposon. Additionally, administration of either component of the gene transfer system may occur repeatedly, for example, by administering at least two doses of this component.

[0064] Any of the transposase proteins described herein may be encoded by polynucleotides including RNA or DNA. Similarly, the nucleic acid encoding the transposase protein or the transposon of this invention can be transfected into the cell as a linear fragment or as a circularized fragment, either as a plasmid or as recombinant viral DNA.

[0065] An *Amyelois* transposase may be provided as a DNA molecule expressible in the target cell. The sequence encoding the *Amyelois* transposase should be operably linked to heterologous sequences that enable expression of the transposase in the target cell. A sequence encoding the *Amyelois* transposase may be operably linked to a heterologous promoter that is active in the target cell. For example, if the target cell is a mammalian cell, then the promoter should be active in a mammalian cell. If the target is a vertebrate cell, the promoter should be active in a vertebrate cell. If the target cell is a plant cell, the promoter should be active in a plant cell. If the promoter is an insect cell, the promoter should be active in an insect cell. The sequence encoding the *Amyelois* transposase may also be operably linked to other sequence elements required for expression in the target cell, for example polyadenylation sequences, terminator sequences etc.

[0066] An *Amyelois* transposase may be provided as an mRNA expressible in the target cell. mRNA is preferably prepared in an *in vitro* transcription reaction. For *in vitro* transcription, a sequence encoding the *Amyelois* transposase is operably linked to a promoter that is active in an *in vitro* transcription reaction. Exemplary promoters active in an *in vitro* transcription reaction include a T7 promoter (5'-TAATACGACTCACTATAG-3') which enables transcription by T7 RNA polymerase, a T3 promoter (5'-AATTAACCCTCACTAAAG-3') which enables transcription by T3 RNA polymerase and an SP6 promoter (5'-ATTTAGGTGACACTATAG-3') which enables transcription by SP6 RNA polymerase. Variants of these promoters and other promoters that can be used for in vitro transcription may also be operably linked to a sequence encoding an *Amyelois* transposase.

[0067] If the *Amyelois* transposase is provided as a polynucleotide (either DNA or mRNA) encoding the transposase, then it is advantageous to improve the expressibility of the transposase in the target cell. It is therefore advantageous to use a sequence other than a naturally occurring sequence to encode the transposase, in other words, to use codon-preferences of the cell type in which expression is to be performed. For example, if the target cell is a mammalian cell, then the codons should be biased toward the preferences seen in a mammalian cell. If the target is a vertebrate cell, then the codons should be biased toward the preferences seen in the particular vertebrate cell. If the target cell is a plant cell, then the codons should be biased toward the preferences seen in a in a plant cell. If the promoter is an insect cell, then the codons should be biased toward the preferences seen in an insect cell.

[0068] Preferable RNA molecules include those with appropriate cap structures to enhance translation in a eukaryotic cell, polyadenylic acid and other 3' sequences that enhance mRNA stability in a eukaryotic cell and optionally substitutions to reduce toxicity

effects on the cell, for example substitution of uridine with pseudouridine, and substitution of cytosine with 5-methyl cytosine. mRNA encoding the *Amyelois* transposase may be prepared such that it has a 5'-cap structure to improve expression in a target cell. Exemplary cap structures are a cap analog (G(5')ppp(5')G), an anti-reverse cap analog (3'-O-Me-m^7G(5')ppp(5')G, a clean cap (m7G(5')ppp(5')(2'OMeA)pG), an mCap (m7G(5')ppp(5')G). mRNA encoding the *Amyelois* transposase may be prepared such that some bases are partially or fully substituted, for example uridine may be substituted with pseudo-uridine, cytosine may be substituted with 5-methyl-cytosine. Any combinations of these caps and substitutions may be made.

[0069] The components of the gene transfer system may be transfected into one or more cells by techniques such as particle bombardment, electroporation, microinjection, combining the components with lipid-containing vesicles, such as cationic lipid vesicles, DNA condensing reagents (example, calcium phosphate, polylysine or polyethyleneimine), and inserting the components (that is the nucleic acids thereof into a viral vector and contacting the viral vector with the cell. Where a viral vector is used, the viral vector can include any of a variety of viral vectors known in the art including viral vectors selected from the group consisting of a retroviral vector, an adenovirus vector or an adeno- associated viral vector. The gene transfer system may be formulated in a suitable manner as known in the art, or as a pharmaceutical composition or kit.

5.2.3 SEQUENCE ELEMENTS IN GENE TRANSFER SYSTEMS

[0070] Expression of genes from a gene transfer polynucleotide such as a piggyBac-like transposon, including an *Amyelois* transposon, integrated into a host cell genome is often strongly influenced by the chromatin environment into which it integrates. Polynucleotides that are integrated into euchromatin have higher levels of expression than those that are either integrated into heterochromatin, or which become silenced following their integration. Silencing of a heterologous polynucleotide may be reduced if it comprises a chromatin control element. It is thus advantageous for gene transfer polynucleotides (including any of the transposons described herein) to comprise chromatin control elements such as sequences that prevent the spread of heterochromatin (insulators). Advantageous gene transfer polynucleotides including an *Amyelois* transposon comprise an insulator sequence that is at least 95% identical to a sequence selected from one of SEQ ID NOS: 435-441, they may also comprise ubiquitously acting chromatin opening elements (UCOEs) or stabilizing and anti-repressor elements (STARs), to increase long-term stable expression from the integrated gene

transfer polynucleotide. Advantageous gene transfer polynucleotides may further comprise a matrix attachment region for example a sequence that is at least 95% identical to a sequence selected from one of SEQ ID NOS: 442-452.

[0071] In some cases, it is advantageous for a gene transfer polynucleotide to comprise two insulators, one on each side of the heterologous polynucleotide that contains the sequence(s) to be expressed, and within the transposon ITRs. The insulators may be the same, or they may be different. Particularly advantageous gene transfer polynucleotides comprise an insulator sequence that is at least 95% identical to a sequence selected from one of SEQ ID NO: 440 or SEQ ID NO: 441 and an insulator sequence that is at least 95% identical to a sequence selected from one of SEQ ID NOS: 435-439. Insulators also shield expression control elements from one another. For example, when a gene transfer polynucleotide comprises genes encoding two open reading frames, each operably linked to a different promoter, one promoter may reduce expression from the other in a phenomenon known as transcriptional interference. Interposing an insulator sequence that is at least 95% identical to a sequence selected from one of SEQ ID NOS: 435-441 between the two transcriptional units can reduce this interference, increasing expression from one or both promoters.

[0072] Preferred gene transfer vectors comprise expression elements capable of driving high levels of gene expression. In eukaryotic cells, gene expression is regulated by several different classes of elements, including enhancers, promoters, introns, RNA export elements, polyadenylation sequences and transcriptional terminators.

[0073] Advantageous gene transfer polynucleotides for the transfer of genes for expression into eukaryotic cells comprise an enhancer operably linked to a heterologous gene. Advantageous gene transfer polynucleotides for the transfer of genes for expression into mammalian cells comprise an enhancer from immediate early genes 1, 2 or 3 of cytomegalovirus (CMV) from either human, primate or rodent cells (for example sequences at least 95% identical to any of SEQ ID NOS: 453-471), an enhancer from the adenoviral major late protein enhancer (for example sequences at least 95% identical to SEQ ID NO: 472), or an enhancer from SV40 (for example sequences at least 95% identical to SEQ ID NO: 473), operably linked to a heterologous gene.

[0074] Advantageous gene transfer polynucleotides for the transfer of genes for expression into eukaryotic cells comprise a promoter operably linked to a heterologous gene. Advantageous gene transfer polynucleotides for the transfer of genes for expression into mammalian cells comprise an EF1a promoter from any mammalian or avian species

including human, rat, mice, chicken and Chinese hamster, (for example any of SEQ ID NOS: 474-495); a promoter from the immediate early genes 1, 2 or 3 of cytomegalovirus (CMV) from either human, primate or rodent cells (for example any of SEQ ID NOS: 496-506); a promoter for eukaryotic elongation factor 2 (EEF2) from any mammalian or avian species including human, rat, mice, chicken and Chinese hamster, (for example any of SEQ ID NOS: 507-517); a GAPDH promoter from any mammalian or yeast species (for example any of SEQ ID NOS: 528-544), an actin promoter from any mammalian or avian species including human, rat, mice, chicken and Chinese hamster (for example any of SEQ ID NOS: 518-527); a PGK promoter from any mammalian or avian species including human, rat, mice, chicken and Chinese hamster (for example any of SEQ ID NOS: 545-551), or a ubiquitin promoter (for example SEQ ID NO: 552), operably linked to a heterologous gene. The promoter may be operably linked to i) a heterologous open reading frame; ii) a nucleic acid encoding a selectable marker; iii) a nucleic acid encoding a counter-selectable marker; iii) a nucleic acid encoding a regulatory protein; iv) a nucleic acid encoding an inhibitory RNA.

[0075] Advantageous gene transfer polynucleotides for the transfer of genes for expression into eukaryotic cells comprise an intron within a heterologous polynucleotide spliceable in a target cell. Advantageous gene transfer polynucleotides for the transfer of genes for expression into mammalian cells comprise an intron from immediate early genes 1, 2 or 3 of cytomegalovirus (CMV) from either human, primate or rodent cells (for example sequences at least 95% identical to any of SEQ ID NOS: 561-571), an intron from EF1a from any mammalian or avian species including human, rat, mice, chicken and Chinese hamster, (for example sequences at least 95% identical to any of SEQ ID NOS: 581-593), an intron from EEF2 from any mammalian or avian species including human, rat, mice, chicken and Chinese hamster, (for example sequences at least 95% identical to any of SEQ ID NOS: 613-620); an intron from actin from any mammalian or avian species including human, rat, mice, chicken and Chinese hamster (for example sequences at least 95% identical to any of SEQ ID NOS: 594-607), a GAPDH intron from any mammalian or avian species including human, rat, mice, chicken and Chinese hamster (for example sequences at least 95% identical to any of SEQ ID NOS: 608-610); an intron comprising the adenoviral major late protein enhancer for example sequences at least 95% identical to SEQ ID NO: 611-612) or a hybrid / synthetic intron (for example sequences at least 95% identical to any of SEQ ID NOS: 572-580) within a heterologous polynucleotide.

[0076] Advantageous gene transfer polynucleotides for the transfer of genes for expression into eukaryotic cells comprise an enhancer and promoter, operably linked to a

heterologous coding sequence. Such gene transfer polynucleotides may comprise combinations of enhancers and promoters in which an enhancer from one gene is combined with a promoter from a different gene, that is the enhancer is heterologous to the promoter. For example, for the transfer of genes for expression into mammalian cells, an immediate early CMV enhancer from rodent or human or primate (such as a sequence selected from SEQ ID NOS: 453-471) is advantageously followed by a promoter from an EF1a gene (such as a sequence selected from SEQ ID NOS: 474-495), or a promoter from a heterologous CMV gene (such as a sequence selected from SEQ ID NOS: 496-506), or a promoter from an EEF2 gene (such as a sequence selected from SEQ ID NOS: 507-517), or a promoter from an actin gene (such as a sequence selected from SEQ ID NOS: 518-527) , or a promoter from a GAPDH gene (such as a sequence selected from SEQ ID NOS: 528-544) operably linked to a heterologous sequence.

[0077] Advantageous gene transfer polynucleotides for the transfer of genes for expression into eukaryotic cells comprise an operably linked promoter and an intron, operably linked to a heterologous open reading frame. Such gene transfer polynucleotides may comprise combinations of promoters and introns in which a promoter from one gene is combined with an intron from a different gene, that is the intron is heterologous to the promoter. For example, for the transfer of genes for expression into mammalian cells, an immediate early CMV promoter from rodent or human or primate (such as a sequence selected from SEQ ID NOS: 496-506) is advantageously followed by an intron from an EF1a gene (such as a sequence that is at least 95% identical to a sequence selected from SEQ ID NOS: 581-593) or an intron from an EEF2 gene (such as a sequence that is at least 95% identical to a sequence selected from SEQ ID NOS: 613-620), or an intron from an actin gene (such as a sequence that is at least 95% identical to a sequence selected from SEQ ID NOS: 594-607) operably linked to a heterologous sequence.

[0078] Advantageous gene transfer polynucleotides for the transfer of genes for expression into eukaryotic cells, comprise composite transcriptional initiation regulatory elements comprising promoters that are operably linked to enhancers and / or introns, and the composite transcriptional initiation regulatory element is operably linked to a heterologous sequence. Examples of advantageous composite transcriptional initiation regulatory elements that may be operably linked to a heterologous sequence in gene transfer polynucleotides for the transfer of genes for expression into mammalian cells are sequences selected from SEQ ID NO: 622-714.

[0079] Expression of two open reading frames from a single polynucleotide can be accomplished by operably linking the expression of each open reading frame to a separate promoter, each of which may optionally be operably linked to enhancers and introns as described above. This is particularly useful when expressing two polypeptides that need to interact at specific molar ratios, such as chains of an antibody or chains of a bispecific antibody, or a receptor and its ligand. It is often advantageous to prevent transcriptional promoter interference by placing a genetic insulator between the two open reading frames, for example to the 3' of the polyadenylation sequence operably linked to the first open reading frame and to the 5' of the promoter operably linked to the second open reading frame encoding the second polypeptide. Transcriptional promoter interference may also be prevented by effectively terminating transcription of the first gene. In many eukaryotic cells the use of strong polyA signal sequences between two open reading frames will reduce transcriptional promote interference. Examples of polyA signal sequences that can be used to effectively terminate transcription are given as SEQ ID NOs: 715-744. Advantageous gene transfer polynucleotides comprise a sequence that is at least 95% identical to a sequence selected from SEQ ID NOs: 715-744 operably linked to a heterologous open reading frame. Advantageous composite regulatory elements for the termination of transcription of a first gene and the initiation of transcription of a second gene include sequences given as SEQ ID NOs: 745-928. Particularly advantageous gene transfer polynucleotides for the transfer of a first and a second open reading frame for co-expression into mammalian cells comprise a sequence at least 90% identical or at least 95% identical or at least 99% identical to or 100% identical to a sequence selected from SEQ ID NOS: 745-928, separating two heterologous open reading frames.

5.2.4 SELECTION OF TARGET CELLS COMPRISING GENE TRANSFER POLYNUCLEOTIDES

[0080] A target cell whose genome comprises a stably integrated transfer polynucleotide may be identified, if the gene transfer polynucleotide comprises an open reading frame encoding a selectable marker, by exposing the target cells to conditions that favor cells expressing the selectable marker ("selection conditions"). It is advantageous for a gene transfer polynucleotide to comprise an open reading frame encoding a selectable marker such as an enzyme that confers resistance to antibiotics such as neomycin (resistance conferred by an aminoglycoside 3'-phosphotransferase e.g. a sequence selected from SEQ ID NOs: 262-265), puromycin (resistance conferred by puromycin acetyltransferase e.g. a

sequence selected from SEQ ID NOs: 268-270), blasticidin (resistance conferred by a blasticidin acetyltransferase and a blasticidin deaminase e.g. SEQ ID NO: 272), hygromycin B (resistance conferred by hygromycin B phosphotransferase e.g. a sequence selected from SEQ ID NOs: 266-267) and zeocin (resistance conferred by a binding protein encoded by the *ble* gene, for example SEQ ID NO: 259). Other selectable markers include those that are fluorescent (such as open reading frames encoding GFP, RFP etc.) and can therefore be selected for example using flow cytometry. Other selectable markers include open reading frames encoding transmembrane proteins that are able to bind to a second molecule (protein or small molecule) that can be fluorescently labelled so that the presence of the transmembrane protein can be selected for example using flow cytometry.

[0081] A gene transfer polynucleotide may comprise a selectable marker open reading frame encoding glutamine synthetase (GS, for example a sequence selected from SEQ ID NOS: 274-278) which allows selection via glutamine metabolism. Glutamine synthase is the enzyme responsible for the biosynthesis of glutamine from glutamate and ammonia, it is a crucial component of the only pathway for glutamine formation in a mammalian cell. In the absence of glutamine in the growth medium, the GS enzyme is essential for the survival of mammalian cells in culture. Some cell lines, for example mouse myeloma cells do not express sufficient GS enzyme to survive without added glutamine. In these cells a transfected GS open reading frame can function as a selectable marker by permitting growth in a glutamine-free medium. In other cell lines, for example Chinese hamster ovary (CHO) cells express sufficient GS enzyme to survive without exogenously added glutamine. These cell lines can be manipulated by genome editing techniques including CRISPR/Cas9 to reduce or eliminate the activity of the GS enzyme. In all of these cases, GS inhibitors such as methionine sulphoximine (MSX) can be used to inhibit a cell's endogenous GS activity. Selection protocols include introducing a gene transfer polynucleotide comprising sequences encoding a first polypeptide and a glutamine synthase selectable marker, and then treating the cell with inhibitors of glutamine synthase such as methionine sulphoximine. The higher the levels of methionine sulphoximine that are used, the higher the level of glutamine synthase expression is required to allow the cell to synthesize sufficient glutamine to survive. Some of these cells will also show an increased expression of the first polypeptide.

[0082] Preferably the GS open reading frame is operably linked to a weak promoter or other sequence elements that attenuate expression as described herein, such that high levels of expression can only occur if many copies of the gene transfer polynucleotide are present, or if they are integrated in a position in the genome where high levels of expression occur. In

such cases it may be unnecessary to use the inhibitor methionine sulphoximine: simply synthesizing sufficient glutamine for cell survival may provide a sufficiently stringent selection if expression of the glutamine synthetase is attenuated.

[0083] A gene transfer polynucleotide may comprise a selectable marker open reading frame encoding dihydrofolate reductase (DHFR, for example a sequence selected from SEQ ID NO: 260-261) which is required for catalyzing the reduction of 5,6-dihydrofolate (DHF) to 5,6,7,8-tetrahydrofolate (THF). Some cell lines do not express sufficient DHFR to survive without added hypoxanthine and thymidine (HT). In these cells a transfected DHFR open reading frame can function as a selectable marker by permitting growth in a hypoxanthine and thymidine-free medium. DHFR-deficient cell lines, for example Chinese hamster ovary (CHO) cells can be produced by genome editing techniques including CRISPR/Cas9 to reduce or eliminate the activity of the endogenous DHRF enzyme. DHFR confers resistance to methotrexate (MTX). DHFR can be inhibited by higher levels of methotrexate. Selection protocols include introducing a construct comprising sequences encoding a first polypeptide and a DHFR selectable marker into a cell with or without a functional endogenous DHFR gene, and then treating the cell with inhibitors of DHFR such as methotrexate. The higher the levels of methotrexate that are used, the higher the level of DHFR expression is required to allow the cell to synthesize sufficient DHFR to survive. Some of these cells will also show an increased expression of the first polypeptide. Preferably the DHFR open reading frame is operably linked to a weak promoter or other sequence elements that attenuate expression as described above, such that high levels of expression can only occur if many copies of the gene transfer polynucleotide are present, or if they are integrated in a position in the genome where high levels of expression occur.

[0084] High levels of expression may be obtained from genes encoded on gene transfer polynucleotides that are integrated at regions of the genome that are highly transcriptionally active, or that are integrated into the genome in multiple copies, or that are present extrachromosomally in multiple copies. It is often advantageous to operably link the open reading frame encoding the selectable marker to expression control elements that result in low levels of expression of the selectable polypeptide from the gene transfer polynucleotide and / or to use conditions that provide more stringent selection. Under these conditions, for the expression cell to produce sufficient levels of the selectable polypeptide encoded on the gene transfer polynucleotide to survive the selection conditions, the gene transfer polynucleotide can either be present in a favorable location in the cell's genome for high levels of expression, or a sufficiently high number of copies of the gene transfer

polynucleotide can be present, such that these factors compensate for the low levels of expression achievable because of the expression control elements.

[0085] Genomic integration of transposons in which a selectable marker is operably linked to regulatory elements that only weakly express the marker usually requires that the transposon be inserted into the target genome by a transposase, see for example Section 6.1.3. By operably linking the selectable marker to elements that result in weak expression, cells are selected which either incorporate multiple copies of the transposon, or in which the transposon is integrated at a favorable genomic location for high expression. Using a gene transfer system that comprises a transposon and a corresponding transposase increases the likelihood that cells will be produced with multiple copies of the transposon, or in which the transposon is integrated at a favorable genomic location for high expression. Gene transfer systems comprising a transposon and a corresponding transposase are thus particularly advantageous when the transposon comprises a selectable marker operably linked to a weak promoter.

[0086] A nucleic acid to be expressed as an RNA or protein and a selectable marker may be included on the same gene transfer polynucleotide, but operably linked to different promoters. In this case low expression levels of the selectable marker may be achieved by using a weakly active constitutive promoter such as the phosphoglycerokinase (PGK) promoter (such as a promoter selected from SEQ ID NOS: 545-551), the Herpes Simplex Virus thymidine kinase (HSV-TK) promoter (e.g. a sequence selected from SEQ ID NOs: 553-554), the MC1 promoter (for example SEQ ID NO: 555), a ubiquitin promoter (for example a sequence selected from SEQ ID NO: 552). Other weakly active promoters maybe deliberately constructed, for example a promoter attenuated by truncation, such as a truncated SV40 promoter (for example a sequence selected from SEQ ID NO: 556-557), a truncated HSV-TK promoter (for example SEQ ID NO: 553), or a promoter attenuated by insertion of a 5'UTR unfavorable for expression (for example a sequence selected from SEQ ID NOS: 559-560) between a promoter and the open reading frame encoding the selectable polypeptide. Particularly advantageous gene transfer polynucleotides comprise a promoter sequence selected from SEQ ID NOS: 545-558, operably linked to an open reading frame encoding a selectable marker.

[0087] Expression levels of a selectable marker may also be advantageously reduced by other mechanisms such as the insertion of the SV40 small t antigen intron after the open reading frame for the selectable marker. The SV40 small t intron accepts aberrant 5' splice sites, which can lead to deletions within the preceding open reading frame in a fraction of the

spliced mRNAs, thereby reducing expression of the selectable marker. Particularly advantageous gene transfer polynucleotides comprise intron SEQ ID NO: 621, operably linked to an open reading frame encoding a selectable marker. For this mechanism of attenuation to be effective, it is preferable for the open reading frame encoding the selectable marker to comprise an intron donor within its coding region. DNA sequences SEQ ID NOs: 279-282 are exemplary nucleic acid sequences that encode glutamine synthetase sequences with SEQ ID NO: 274-277 respectively. Each of these nucleic acid sequences comprises an intron donor, and which may be operably linked to the SV40 small t antigen intron by placing the intron into the 3' UTR of the glutamine synthetase open reading frame. Sequence SEQ ID NO: 271 is an exemplary nucleic acid sequence encoding puromycin acetyl transferase SEQ ID NO: 270, which comprises an intron donor, and which may be operably linked to the SV40 small t antigen intron by placing the intron into the 3' UTR of the puromycin open reading frame. Advantageous gene transfer polynucleotides comprise a sequence at least 90% identical or at least 95% identical or at least 99% identical to, or 100% identical to a sequence selected from one of SEQ ID NO: 279-282 or 271, operably linked to SEQ ID NO: 621.

[0088] Expression levels of a selectable marker may also be advantageously reduced by other mechanisms such as insertion of an inhibitory 5'-UTR within the transcript, for example SEQ ID NO: 559-560. Particularly advantageous gene transfer polynucleotides comprise a promoter operably linked to an open reading frame encoding a selectable marker, wherein a sequence that is at least 90% identical or at least 95% identical or at least 99% identical to, or 100% identical to SEQ ID NO: 559-560 is interposed between the promoter and the selectable marker.

[0089] Exemplary nucleic acid sequences comprising the glutamine synthetase coding sequence operably linked to regulatory sequences expressible in mammalian cells include SEQ ID NOs: 300-370 or 432-434. A gene transfer polynucleotide comprising a sequence selected from SEQ ID NOs: 300-370 or 432-434, upon integration into the genome of a target cell, expresses glutamine synthetase, thereby helping a cell to grow in the absence of added glutamine or in the presence of MSX. Regulatory elements in these sequences have been balanced to produce low levels of expression of glutamine synthetase, providing a selective advantage for target cells whose genome comprises either multiple copies of the gene transfer polynucleotide, or for target calls whose genome comprises copies of the gene transfer polynucleotide in regions of the genome that are favorable for expression of encoded genes. Advantageous gene transfer polynucleotides comprise a sequence selected from SEQ ID

NOs: 300-370 or 432-434, they may further comprise a left transposon end and a right transposon end.

[0090] Exemplary nucleic acid sequences comprising the blasticidin-S-transferase coding sequence operably linked to regulatory sequences expressible in mammalian cells include SEQ ID NOs: 371-377. A gene transfer polynucleotide comprising a sequence selected from SEQ ID NOs: 371-377, upon integration into the genome of a target cell, expresses blasticidin-S-transferase, thereby helping a cell to grow in the presence of added blasticidin. Regulatory elements in these sequences have been balanced to produce low levels of expression of blasticidin-S-transferase, providing a selective advantage for target cells whose genome comprises either multiple copies of the gene transfer polynucleotide, or for target calls whose genome comprises copies of the gene transfer polynucleotide in regions of the genome that are favorable for expression of encoded genes. Advantageous gene transfer polynucleotides comprise a sequence selected from SEQ ID NOs: 371-377, they may further comprise a left transposon end and a right transposon end.

[0091] Exemplary nucleic acid sequences comprising the hygromycin B phosphotransferase coding sequence operably linked to regulatory sequences expressible in mammalian cells include SEQ ID NOs: 378-379. A gene transfer polynucleotide comprising a sequence selected from SEQ ID NOs: 378-379, upon integration into the genome of a target cell, expresses hygromycin B phosphotransferase, thereby helping a cell to grow in the presence of added hygromycin. Regulatory elements in these sequences have been balanced to produce low levels of expression of hygromycin B phosphotransferase, providing a selective advantage for target cells whose genome comprises either multiple copies of the gene transfer polynucleotide, or for target calls whose genome comprises copies of the gene transfer polynucleotide in regions of the genome that are favorable for expression of encoded genes. Advantageous gene transfer polynucleotides comprise a sequence selected from SEQ ID NOs: 378-379, they may further comprise a left transposon end and a right transposon end.

[0092] Exemplary nucleic acid sequences comprising the aminoglycoside 3'-phosphotransferase coding sequence operably linked to regulatory sequences expressible in mammalian cells include SEQ ID NOs: 380-382 or 408-409. A gene transfer polynucleotide comprising a sequence selected from SEQ ID NOs: 380-382 or 408-409, upon integration into the genome of a target cell, expresses aminoglycoside 3'-phosphotransferase, thereby helping a cell to grow in the presence of added neomycin. Regulatory elements in these sequences have been balanced to produce low levels of expression of aminoglycoside 3'-

phosphotransferase, providing a selective advantage for target cells whose genome comprises either multiple copies of the gene transfer polynucleotide, or for target calls whose genome comprises copies of the gene transfer polynucleotide in regions of the genome that are favorable for expression of encoded genes. Advantageous gene transfer polynucleotides comprise a sequence selected from SEQ ID NOs: 380-382 or 408-409, they may further comprise a left transposon end and a right transposon end.

[0093] Exemplary nucleic acid sequences comprising the puromycin acetyltransferase coding sequence operably linked to regulatory sequences expressible in mammalian cells include SEQ ID NOs: 383-402 or 410-434. A gene transfer polynucleotide comprising a sequence selected from SEQ ID NOs: 383-402 or 410-434, upon integration into the genome of a target cell, expresses puromycin acetyltransferase, thereby helping a cell to grow in the presence of added puromycin. Regulatory elements in these sequences have been balanced to produce low levels of expression of puromycin acetyltransferase, providing a selective advantage for target cells whose genome comprises either multiple copies of the gene transfer polynucleotide, or for target calls whose genome comprises copies of the gene transfer polynucleotide in regions of the genome that are favorable for expression of encoded genes. Advantageous gene transfer polynucleotides comprise a sequence selected from SEQ ID NOs: 383-402 or 410-434, they may further comprise a left transposon end and a right transposon end.

[0094] Exemplary nucleic acid sequences comprising the *ble* gene coding sequence operably linked to regulatory sequences expressible in mammalian cells include SEQ ID NOs: 403-407. A gene transfer polynucleotide comprising a sequence selected from SEQ ID NOs: 403-407, upon integration into the genome of a target cell, expresses the *ble* gene, thereby helping a cell to grow in the presence of added zeocin. Regulatory elements in these sequences have been balanced to produce low levels of expression of *ble* gene product, providing a selective advantage for target cells whose genome comprises either multiple copies of the gene transfer polynucleotide, or for target calls whose genome comprises copies of the gene transfer polynucleotide in regions of the genome that are favorable for expression of encoded genes. Advantageous gene transfer polynucleotides comprise a sequence selected from SEQ ID NOs: 403-407, they may further comprise a left transposon end and a right transposon end.

[0095] Exemplary nucleic acid sequences comprising the dihydrofolate reductase coding sequence operably linked to regulatory sequences expressible in mammalian cells include SEQ ID NOs: 283-299 or 408-431. A gene transfer polynucleotide comprising a

sequence selected from SEQ ID NOs: 283-299 or 408-431, upon integration into the genome of a target cell, expresses dihydrofolate reductase, thereby helping a cell to grow in the absence of added hypoxanthine and thymidine or in the presence of MTX. Regulatory elements in these sequences have been balanced to produce low levels of expression of dihydrofolate reductase, providing a selective advantage for target cells whose genome comprises either multiple copies of the gene transfer polynucleotide, or for target calls whose genome comprises copies of the gene transfer polynucleotide in regions of the genome that are favorable for expression of encoded genes. Advantageous gene transfer polynucleotides comprise a sequence selected from SEQ ID NOs: 283-299 or 408-431, they may further comprise a left transposon end and a right transposon end.

[0096] The use of transposons and transposases in conjunction with weakly expressed selectable markers has several advantages over non-transposon constructs. One is that linkage between expression of the first polypeptide and the selectable marker is better for transposons, because a transposase integrates the entire sequence that lies between the two transposon ends into the genome. In contrast when heterologous DNA is introduced into the nucleus of a eukaryotic cell, for example a mammalian cell, it is gradually broken into random fragments which may either be integrated into the cell's genome, or degraded. Thus, if a gene transfer polynucleotide comprising sequences that encode a first polypeptide and a selectable marker is introduced into a population of cells, some cells will integrate the sequences encoding the selectable marker but not those encoding the first polypeptide, and vice versa. Selection of cells expressing high levels of selectable marker is thus only somewhat correlated with cells that also express high levels of the first polypeptide. In contrast, because the transposase integrates all of the sequences between the transposon ends, cells expressing high levels of selectable marker are highly likely to also express high levels of the first polypeptide.

[0097] A second advantage of transposons and transposases is that they are much more efficient at integrating DNA sequences into the genome. A much higher fraction of the cell population is therefore likely to integrate one or more copies of the gene transfer polynucleotide into their genomes, so there will be a correspondingly higher likelihood of good stable expression of both the selectable marker and the first polypeptide.

[0098] A third advantage of piggyBac-like transposons and transposases is that piggyBac-like transposases are biased toward inserting their corresponding transposons into transcriptionally active chromatin. Each cell is therefore likely to integrate the gene transfer polynucleotide into a region of the genome from which genes are well expressed, so there

will be a correspondingly higher likelihood of good stable expression of both the selectable marker and the first polypeptide.

5.2.5 A NOVEL PIGGYBAC-LIKE TRANSPOSASE

[0099] Natural DNA transposons undergo a 'cut and paste' system of replication in which the transposon is excised from a first DNA molecule and inserted into a second DNA molecule. DNA transposons are characterized by inverted terminal repeats (ITRs) and are mobilized by an element-encoded transposase. The piggyBac transposon / transposase system is particularly useful because of the precision with which the transposon is integrated and excised (see for example "Fraser, M. J. (2001) The TTAA-Specific Family of Transposable Elements: Identification, Functional Characterization, and Utility for Transformation of Insects. Insect Transgenesis: Methods and Applications. A. M. Handler and A. A. James. Boca Raton, Fla., CRC Press: 249-268"; and "US 20070204356 A1: PiggyBac constructs in vertebrates" and references therein).

[00100] Many sequences with sequence similarity to the piggyBac transposase from *Trichoplusia ni* have been found in the genomes of phylogenetically distinct species from fungi to mammals, but very few have been shown to possess transposase activity (see for example Wu M, et al (2011) Genetica 139:149-54. "Cloning and characterization of piggyBac-like elements in lepidopteran insects", and references therein).

[00101] Two properties of transposases that are of particular interest for genomic modifications are their ability to integrate a polynucleotide into a target genome, and their ability to precisely excise a polynucleotide from a target genome. Both of these properties can be measured with a suitable system.

[00102] A system for measuring the first step of transposition, which is excision of a transposon from a first polynucleotide, comprises the following components: (i) A first polynucleotide encoding a first selectable marker operably linked to sequences that cause it to be expressed in a selection host and (ii) A transposon comprising transposon ends recognized by a transposase. The transposon is present in, and interrupts the coding sequence of, the first selectable marker, such that the first selectable marker is not active. The transposon is placed in the first selectable marker such that precise excision of the first transposon causes the first selectable marker to be reconstituted. If an active transposase that can excise the first transposon is introduced into a host cell which comprises the first polynucleotide, the host cell will express the active first selectable marker. The activity of the transposase in excising

the transposon can be measured as the frequency with which the host cells become able to grow under conditions that require the first selectable marker to be active.

[00103] If the transposon comprises a second selectable marker, operably linked to sequences that make the second selectable marker expressible in the selection host, transposition of the second selectable marker into the genome of the host cell will yield a genome comprising active first and second selectable markers. The activity of the transposase in transposing the transposon into a second genomic location can be measured as the frequency with which the host cells become able to grow under conditions that require the first and second selectable markers to be active. In contrast, if the first selectable marker is present, but the second is not, then this indicates that the transposon was excised from the first polynucleotide but was not subsequently transposed into a second polynucleotide. The selectable markers may, for example, be open reading frames encoding an antibiotic resistance protein, or an auxotrophic marker, or any other selectable marker.

[00104] We used such a system to test putative transposase / transposon combinations for activity, as described in Section 6.1. We used computational methods to search publicly available sequenced genomes for open reading frames with homology to known active piggyBac-like transposases. We selected transposase sequences that appeared to possess the DDDE motif characteristic of active piggyBac-like transposases and searched the DNA sequences flanking these putative transposases for inverted repeat sequences adjacent to a 5'-TTAA-3' target sequence. Amongst those that we identified were putative transposons with intact transposases from: *Spodoptera litura* (Genbank accession number MTZO01002002.1, protein accession number XP_022823959) with an open reading frame encoding a putative transposase with SEQ ID NO: 172 flanked by a putative left end with SEQ ID NO: 216 and a putative right end with SEQ ID NO: 217; *Pieris rapae* (NCBI genomic reference sequence NW_ 019093607.1, Genbank protein accession number XP_022123753.1) with an open reading frame encoding a putative transposase with SEQ ID NO: 173 flanked by a putative left end with SEQ ID NO: 218 and a putative right end with SEQ ID NO: 219; *Myzus persicae* (NCBI genomic reference sequence NW_019100532.1, protein accession number XP_022166603) with an open reading frame encoding a putative transposase with SEQ ID NO: 174 flanked by a putative left end with SEQ ID NO: 220 and a putative right end with SEQ ID NO: 221; *Onthophagus taurus* (NCBI genomic reference sequence NW_019280463, protein accession number XP_022900752) with an open reading frame encoding a putative transposase with SEQ ID NO: 175 flanked by a putative left end with SEQ ID NO: 222 and a putative right end with SEQ ID NO: 223; *Temnothorax curvispinosus* (NCBI genomic

reference sequence NW_020220783.1, protein accession number XP_024881886) with an open reading frame encoding a putative transposase with SEQ ID NO: 176 flanked by a putative left end with SEQ ID NO: 224 and a putative right end with SEQ ID NO: 225; *Agrlius planipenn* (NCBI genomic reference sequence NW_020442437.1, protein accession number XP_025836109) with an open reading frame encoding a putative transposase with SEQ ID NO: 177 flanked by a putative left end with SEQ ID NO: 226 and a putative right end with SEQ ID NO: 227; *Parasteatoda tepidariorum* (NCBI genomic reference sequence NW_018371884.1, protein accession number XP_015905033) with an open reading frame encoding a putative transposase with SEQ ID NO: 178 flanked by a putative left end with SEQ ID NO: 228 and a putative right end with SEQ ID NO: 229; *Pectinophora gossypiella* (Genbank accession number GU270322.1, protein ID ADB45159.1, also described in Wang et al, 2010. Insect Mol. Biol. 19, 177-184. "piggyBac-like elements in the pink bollworm, Pectinophora gossypiella") with an open reading frame encoding a putative transposase with SEQ ID NO: 179 flanked by a putative left end with SEQ ID NO: 230 and a putative right end with SEQ ID NO: 231; *Ctenoplusia agnata* (NCBI accession number GU477713.1, protein accession number ADV17598.1, also described by Wu M, et al (2011) Genetica 139:149-54. "Cloning and characterization of piggyBac-like elements in lepidopteran insects") with an open reading frame encoding a putative transposase with SEQ ID NO: 180 flanked by a putative left end with SEQ ID NO: 232 and a putative right end with SEQ ID NO: 233; *Macrostomum lignano* (NCBI genomic reference sequence NIVC01003029.1, protein accession number PAA53757) with an open reading frame encoding a putative transposase with SEQ ID NO: 181 flanked by a putative left end with SEQ ID NO: 234 and a putative right end with SEQ ID NO: 235; *Orussus abietinus* (NCBI accession number XM_012421754, protein accession number XP_012277177) with an open reading frame encoding a putative transposase with SEQ ID NO: 182 flanked by a putative left end with SEQ ID NO: 236 and a putative right end with SEQ ID NO: 237; *Eufriesea mexicana* (NCBI genomic reference sequence NIVC01003029.1, protein accession number XP_017759329) with an open reading frame encoding a putative transposase with SEQ ID NO: 183 flanked by a putative left end with SEQ ID NO: 238 and a putative right end with SEQ ID NO: 239; *Spodoptera litura* (NCBI genomic reference sequence NC_036206.1, protein accession number XP_022824855) with an open reading frame encoding a putative transposase with SEQ ID NO: 184 flanked by a putative left end with SEQ ID NO: 240 and a putative right end with SEQ ID NO: 241; *Vanessa tameamea* (NCBI genomic reference sequence NW_020663261.1, protein accession number XP_026490968) with an open reading frame

encoding a putative transposase with SEQ ID NO: 185 flanked by a putative left end with SEQ ID NO: 242 and a putative right end with SEQ ID NO: 243; *Blattella germanica* (NCBI genomic reference sequence PYGN01002011.1, protein accession number PSN31819) with an open reading frame encoding a putative transposase with SEQ ID NO: 186 flanked by a putative left end with SEQ ID NO: 244 and a putative right end with SEQ ID NO: 245; *Onthophagus taurus* (NCBI genomic reference sequence NW_019281532.1, protein accession number XP_022910826) with an open reading frame encoding a putative transposase with SEQ ID NO: 187 flanked by a putative left end with SEQ ID NO: 246 and a putative right end with SEQ ID NO: 247; *Onthophagus taurus* (NCBI genomic reference sequence NW_019281689.1, protein accession number XP_022911139) with an open reading frame encoding a putative transposase with SEQ ID NO: 188 flanked by a putative left end with SEQ ID NO: 248 and a putative right end with SEQ ID NO: 249; *Onthophagus taurus* (NCBI genomic reference sequence NW_019286114.1, protein accession number XP_022913435) with an open reading frame encoding a putative transposase with SEQ ID NO: 189 flanked by a putative left end with SEQ ID NO: 250 and a putative right end with SEQ ID NO: 251; *Megachile rotundata* (NCBI genomic reference sequence NW_003797295, protein accession number XP_012145925) with an open reading frame encoding a putative transposase with SEQ ID NO: 190 flanked by a putative left end with SEQ ID NO: 252 and a putative right end with SEQ ID NO: 253; *Xiphophorus maculatus* (NCBI genomic reference sequence NC_036460.1, protein accession number XP_023207869) with an open reading frame encoding a putative transposase with SEQ ID NO: 191 flanked by a putative left end with SEQ ID NO: 254 and a putative right end with SEQ ID NO: 255 and *Amyelois transitella* (NCBI accession number XM_013335311, protein accession number XP_013190765.1) with an open reading frame encoding a putative transposase with SEQ ID NO: 18 flanked by a putative left end with SEQ ID NO: 1 and a putative right end with SEQ ID NO: 3.

5.2.5.1 The *Amyelois* transposase and its corresponding transposon

[00105] One active transposase and its corresponding transposon identified by transposition activity in yeast was an *Amyelois* transposase, as described in Section 6.1.2. An *Amyelois* transposase comprises a polypeptide sequence that is at least 80%, 90%, 93%, 95%, 96%, 97%, 98%, 99% or 100% identical to the sequence given by SEQ ID NO: 18, and which is capable of transposing the transposon from transposase reporter construct SEQ ID

NO: 192, as described in Section 6.1.2. Exemplary *Amyelois* transposases include sequences given as SEQ ID NOs: 18-170.

[00106] An *Amyelois* transposase may be provided as a part of a gene transfer system as a protein, or as a polynucleotide encoding the *Amyelois* transposase, wherein the polynucleotide is expressible in the target cell. When provided as a polynucleotide, the *Amyelois* transposase may be provided as DNA or mRNA. If provided as DNA, the open reading frame encoding the *Amyelois* transposase is preferably operably linked to heterologous regulatory elements including a promoter that is active in the target cell such that the transposase is expressible in the target cell, for example a promoter that is active in a eukaryotic cell or a vertebrate cell or a mammalian cell. If provided as mRNA, the mRNA may be prepared *in vitro* from a DNA molecule in which the open reading frame encoding the *Amyelois* transposase is preferably operably linked to a heterologous promoter active in the invitro transcription system used to prepare the mRNA, for example a T7 promoter.

[00107] An *Amyelois* transposon comprises a heterologous polynucleotide flanked by a left transposon end comprising a left ITR with sequence given by SEQ ID NO: 9 and a right transposon end comprising a right ITR with sequence given by SEQ ID NO: 10, and wherein the distal end of each ITR is immediately adjacent to a target sequence. Here and elsewhere when inverted repeats are defined by a sequence including a nucleotide defined by an ambiguity code, the identity of that nucleotide can be selected independently in the two repeats. A preferred target sequence is 5'-TTAA-3', although other useable target sequences may be used; preferably the target sequence on one side of the transposon is a direct repeat of the target sequence on the other side of the transposon. The left transposon end may further comprise additional sequences proximal to the ITR, for example a sequence at least 90% identical to, or 100% identical to SEQ ID NO: 7, or SEQ ID NO: 13. The right transposon end may further comprise additional sequences proximal to the ITR, for example a sequence at least 90% identical to, or 100% identical to SEQ ID NO: 8, or SEQ ID NO: 16. The structure of a representative *Amyelois* transposon is shown in Figure 1. An *Amyelois* transposon can be transposed by a transposase with a polypeptide sequence given by SEQ ID NO: 18, for example as encoded by a polynucleotide with sequence given by SEQ ID NO: 171 operably linked to a Gal1 promoter.

[00108] Transposon ends, including ITRs and target sequences may be added to the ends of a heterologous polynucleotide sequence to create a synthetic *Amyelois* transposon which may be efficiently transposed into a target eukaryotic genome by an *Amyelois* transposase. For example, SEQ ID NO: 1 and SEQ ID NO: 2 each comprise a left 5'-TTAA-

3' target sequence followed by a left transposon ITR followed by additional end sequences that may be added to one side of a heterologous polynucleotide, with the target sequence distal relative to the heterologous polynucleotide, to generate a synthetic *Amyelois* transposon. SEQ ID NO: 3, and SEQ ID NO: 4 each comprise additional end sequences followed by a right transposon ITR sequence followed by a right 5'-TTAA-3' target sequence that may be added to the other side of a heterologous polynucleotide, with the target sequence distal relative to the heterologous polynucleotide, to generate a synthetic *Amyelois* transposon. The preceding transposon end sequences comprise 5'-TTAA-3' as the target sequence, but SEQ ID NO: 5 comprises a left transposon ITR followed by additional end sequences that may be added to one side of a heterologous polynucleotide, with the ITR sequence distal relative to the heterologous polynucleotide, and SEQ ID NO: 6 comprises additional end sequences followed by a right transposon ITR sequence that may be added to the other side of a heterologous polynucleotide, with the target sequence distal relative to the heterologous polynucleotide, to generate a synthetic *Amyelois* transposon which may then be flanked by alternative target sequences.

[00109] *Amyelois* transposases recognize synthetic *Amyelois* transposons. They excise the transposon from a first DNA molecule, by cutting the DNA at the target sequence at the left end of one transposon end and the target sequence at the right end of the second transposon end, re-join the cut ends of the first DNA molecule to leave a single copy of the target sequence. The excised transposon sequence, including any heterologous DNA that is between the transposon ends, is integrated by the transposase into a target sequence of a second DNA molecule, such as the genome of a target cell. A cell whose genome comprises a synthetic *Amyelois* transposon is an embodiment of the invention.

5.2.5.2 The *Amyelois* transposase is active in mammalian cells

[00110] The looper moth piggyBac transposase has been shown to be active in a very wide variety of eukaryotic cells. In Section 6.1.2 we show that the *Amyelois* transposase can transpose its corresponding transposon into the genome of the yeast *Saccharomyces cerevisiae*. In Section 6.1.3 we show that the *Amyelois* transposase can transpose its corresponding transposon into the genome of a mammalian CHO cell. These results provide evidence that, like the other known active piggyBac-like transposases, the *Amyelois* transposase is also active in transposing its corresponding transposon into the genomes of most eukaryotic cells. Although the *Amyelois* transposase is active in a wide range of

eukaryotic cells, the naturally occurring open reading frame encoding the *Amyelois* transposase (given by SEQ ID NO: 929) is unlikely to express well in a similarly wide range of cells, as optimal codon usage differs significantly between cell types. It is therefore advantageous to use a sequence other than a naturally occurring sequence to encode the transposase, in other words, to use codon-preferences of the cell type in which expression is to be performed. Likewise, the promoter and other regulatory sequences are selected so as to be active in the cell type in which expression is to be performed. An advantageous polynucleotide for expression of an *Amyelois* transposase comprises at least 2, 5, 10, 20, 30, 40 or 50 synonymous codon differences relative to SEQ ID NO: 929 at corresponding positions between the polynucleotide and SEQ ID NO:929, optionally wherein codons in the polynucleotide at the corresponding positions are selected for mammalian cell expression. Exemplary polynucleotide sequences for *Amyelois* transposases with polypeptide sequences given by SEQ ID NOs: 18, 96 and 170 wherein codons have been selected for expression in mammalian cells are given as SEQ ID NOs: 171, 930 and 931 respectively. The polynucleotide may be DNA or mRNA.

5.2.6 HYPERACTIVE *AMYELOIS* TRANSPOSASES

[00111] Individual favorable mutations may be combined in a variety of different ways, for example by "DNA shuffling" or by methods described in US Patent 8,635,029 B2 or by methods described in US Patent 8,635,029 B2 and Liao et al (2007, BMC Biotechnology 2007, 7:16 doi:10.1186/1472-6750-7-16 “Engineering proteinase K using machine learning and synthetic genes”). A transposase with modified activity, either for activity on a new target sequence, or increased activity on an existing target sequence may be obtained by using variations of the selection scheme described herein (for example Section 6.1.6) with an appropriate corresponding transposon.

[00112] An alignment of known active piggyBac-like transposases may be used to identify amino acid changes likely to result in enhanced activity. Transposases are often deleterious to their hosts, so tend to accumulate mutations that inactivate them. However the mutations that accumulate in different transposases are different, as each occurs by random chance. A consensus sequence can be obtained from an alignment of sequences, and this can be used to improve activity (Ivics et al, 1997. Cell 91: 501-510. “Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells.”). We aligned known active piggyBac-like transposases using the CLUSTAL algorithm, and enumerated the amino acids found at each position. This diversity is shown in Table 1

relative to an *Amyelois* transposase (relative to SEQ ID NO: 18), the amino acids shown in column C are found in known active piggyBac-like transposases at the equivalent position in an alignment, and are thus likely to be acceptable changes in an *Amyelois* transposase. Column D shows amino acid changes found in known active piggyBac-like transposases other than the *Amyelois* transposase at positions where there is good conservation within the rest of the transposase set, but the amino acid in the *Amyelois* transposase sequence is an outlier. Mutation of the position shown in column A to an amino acid shown in column D is particularly likely to result in enhanced transposase activity.

[00113] We selected 59 amino acid substitutions to make in *Amyelois* transposase SEQ ID NO: 18 from column D in Table 1. The substitutions were P65E, D79N, R95S, V100I, V106P, L115D, E116P, H121Q, V131E, K139E, T159N, V166F, G174A, G179N, W187F, P198R, L203R, I209L, N211R, A225R, P232A, E238D, V261L, L273M, L273I, D304R, D304K, Y310G, I323L, Q329G, Q329R, M336C, Y343I, T345L, K362R, T366R, C367E, T380S, L408M, E413S, S416E, I426M, K432Q, S435G, K442Q, N452K, L458M, N466D, T467M, A472S, V475I, N483K, I491M, A529P, K540R, D551R, S560K, T562K and S563K. Genes encoding *Amyelois* transposase variants comprising combinations of these substitutions were synthesized and tested for transposase activity as described in Section 6.1.6. We identified more than 70 *Amyelois* transposase variants with increased excision or transposition activity compared with the naturally occurring sequence SEQ ID NO: 18. Exemplary sequences of hyperactive *Amyelois* transposase variants are provided as SEQ ID NOs: 96-170. Particularly active *Amyelois* transposase variants are provided as SEQ ID NOs: 114-170.

[00114] *Amyelois* transposases can thus be created that are not naturally occurring sequences, but that are at least 99%, 98%, 97%, 96%, 95%, 90% or at least 80% identical to SEQ ID NO 18. Such variants can retain partial activity of the transposase of SEQ ID NO: 18 (as determined by either or both of transposition and/or excision activity), can be functionally equivalent of the transposase of SEQ ID NO: 18 in either or both of transposition and excision, or can have enhanced activity relative to the transposase of SEQ ID NO: 18 in transposition, excision activity or both. Such variants can include mutations shown herein to increase transposition and/or excision, mutations shown herein to be neutral as to transposition and/or excision, and mutations detrimental to transposition and/or integration. Preferred variants include mutations shown to be neutral or to enhance transposition/and or excision. Some such variants lack mutations shown to be detrimental to transposition and/or excision. Some such variants include only mutations shown to enhance transposition, only

mutations shown to enhance excision, or mutations shown to enhance both transposition and excision.

[00115] Enhanced activity means activity (e.g., transposition or excision activity) that is greater beyond experimental error than that of a reference transposase from which a variant was derived. The activity can be greater by a factor of e.g., 1.5, 2, 5, 10, 15, 20, 50 or 100 fold of the reference transposase. The enhanced activity can lie within a range of for example 2-100 fold, 2-50 fold, 5-50 fold or 2-10 fold of the reference transposase. Here and elsewhere activities can be measured as demonstrated in the examples.

[00116] Functional equivalence means a variant transposase can mediate transposition and/or excision of the same transposon with a comparable efficiency (within experimental error) to a reference transposase. Seventeen representative sequences of variant *Amyelois* transposases with transposition frequencies comparable to naturally occurring *Amyelois* transposase SEQ ID NO: 18 are SEQ ID NOs: 19-35.

[00117] Furthermore, variant sequences of SEQ ID NO:18 can be created by combining two, three, four, or five or more substitutions selected from Table 1 column D. Combining beneficial substitutions, for example those shown in column D of Table 1 can result in hyperactive variants of SEQ ID NO: 18. We identified more than 70 *Amyelois* transposases (comprising sequences provided as SEQ ID NOs: 96-170) with transposition or excision frequencies more than about 2-fold greater than that of naturally occurring *Amyelois* transposase SEQ ID NO: 18. These preferred hyperactive *Amyelois* transposases comprised one or more of the following substitutions (relative to SEQ ID NO: 18): D79N, R95S, L115D, E116P, H121Q, K139E, V166F, G179N, W187F, P198R, L203R, N211R, E238D, L273M, L273I, D304R, D304K, Q329G, T345L, K362R, T366R, L408M, S416E, S435G, L458M, V475I, N483K, I491M, A529P, K540R, S560K and S563K. Some hyperactive *Amyelois* transposases may further comprise a heterologous nuclear localization sequence.

[00118] We used machine learning methods as described in Liao et al (2007, BMC Biotechnology 2007, 7:16 doi:10.1186/1472-6750-7-16 "Engineering proteinase K using machine learning and synthetic genes") to determine the effect of various amino acid substitutions on the excision and transposition activities of an *Amyelois* transposase. Each substitution was empirically tested in a minimum of 5 different sequence contexts (i.e. in the presence of different other amino acid substitutions). As described in Liao et. al., the mean value for the regression weight for a substitution is a measure of the average effect of that substitution within multiple different *Amyelois* transposases. A substitution with a positive mean regression weight is one that on average has a positive effect on the transposition

activity of *Amyelois* transposases. Addition of a substitution with a positive mean regression weight to an active *Amyelois* transposase that does not already comprise such a substitution is thus expected to improve the transposition activity of that *Amyelois* transposase. Table 8 identifies 32 substitutions within an *Amyelois* transposase with positive mean regression weights for transposition: D79N, R95S, L115D, E116P, H121Q, K139E, V166F, G179N, W187F, P198R, L203R, N211R, E238D, L273M, L273I, D304R, D304K, Q329G, T345L, K362R, T366R, L408M, S416E, S435G, L458M, V475I, N483K, I491M, A529P, K540R, S560K and S563K.

[00119] SEQ ID NOs: 18-35 or 96-170 and comprise a substitution at a position selected from amino acid 79, 95, 115, 116, 121, 139, 166, 179, 187, 198, 203, 211, 238, 273, 304, 329, 345, 362, 366, 408, 416, 435, 458, 475, 483, 491, 529, 540, 560 and 563, relative to SEQ ID NO: 18. Preferably the substitution is one shown in Table 1 columns C or D. Preferably the hyperactive *Amyelois* transposase comprises an amino acid substitution, relative to the sequence of SQ ID NO: 18, selected from D79N, R95S, L115D, E116P, H121Q, K139E, V166F, G179N, W187F, P198R, L203R, N211R, E238D, L273M, L273I, D304R, D304K, Q329G, T345L, K362R, T366R, L408M, S416E, S435G, L458M, V475I, N483K, I491M, A529P, K540R, S560K and S563K, or any combination of substitutions thereof including at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or all of these mutations. The hyperactive *Amyelois* transposase may also comprise substitutions at many other positions that are not associated with enhanced transposition, for example conservative substitutions that have a neutral effect on transposition.

[00120] Preferred hyperactive *Amyelois* transposases comprise an amino acid sequence, other than a naturally occurring protein (e.g., not a transposase whose amino acid sequence comprises SEQ ID NO: 18), that is at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% or 100% identical to the amino acid sequence of any of SEQ ID NOs: 18-35 or 96-170 and comprise a substitution at a position selected from amino acid 79, 95, 115, 116, 121, 139, 166, 179, 187, 198, 203, 211, 238, 273, 304, 329, 345, 362, 366, 408, 416, 435, 458, 475, 483, 491, 529, 540, 560 and 563, relative to SEQ ID NO: 18. Preferably the substitution is one shown in Table 1 columns C or D. Preferably the hyperactive *Amyelois* transposase comprises an amino acid substitution, relative to the sequence of SQ ID NO: 18, selected from D79N, R95S, L115D, E116P, H121Q, K139E, V166F, G179N, W187F, P198R, L203R, N211R, E238D, L273M, L273I, D304R, D304K, Q329G, T345L, K362R, T366R, L408M, S416E, S435G, L458M, V475I, N483K, I491M, A529P, K540R, S560K and S563K, or any

combination of substitutions thereof including at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or all of these mutations.

[00121] Methods of creating transgenic cells using naturally occurring or hyperactive *Amyelois* transposases are an aspect of the invention. A method of creating a transgenic cell comprises (i) introducing into a eukaryotic cell a naturally occurring or hyperactive *Amyelois* transposase (as a protein or as a polynucleotide encoding the transposase) and a corresponding *Amyelois* transposon. Creating the transgenic cell may further comprise (ii) identifying a cell in which an *Amyelois* transposon is incorporated into the genome of the eukaryotic cell. Identifying the cell in which an *Amyelois* transposon is incorporated into the genome of the eukaryotic cell may comprise selecting the eukaryotic cell for a selectable marker encoded on the *Amyelois* transposon. The selectable marker may be any selectable polypeptide, including any described herein.

[00122] Activity of transposases may also be increased by fusion of nuclear localization signal (NLS) at the N -terminus, C-terminus, both at the N- and C-termini or internal regions of the transposase protein, as long as transposase activity is retained. A nuclear localization signal or sequence (NLS) is an amino acid sequence that 'tags' or facilitates interaction of a protein, either directly or indirectly with nuclear transport proteins for import into the cell nucleus. Nuclear localization signals (NLS) used can include consensus NLS sequences, viral NLS sequences, cellular NLS sequences, and combinations thereof.

[00123] Transposases may also be fused to other protein functional domains. Such protein functional domains can include DNA binding domains, flexible hinge regions that can facilitate one or more domain fusions, and combinations thereof. Fusions can be made either to the N-terminus, C-terminus, or internal regions of the transposase protein so long as transposase activity is retained. Fusions to DNA binding domains can be used to direct the *Amyelois* transposase to a specific genomic locus or loci. DNA binding domains may include a helix-turn-helix domain, a zinc-finger domain, a leucine zipper domain, a TALE (transcription activator-like effector) domain, a CRISPR-Cas protein or a helix-loop-helix domain. Specific DNA binding domains used can include a Gal4 DNA binding domain, a LexA DNA binding domain, or a Zif268 DNA binding domain. Flexible hinge regions used can include glycine/serine linkers and variants thereof.

5.3 KITS

[00124] The present invention also features kits comprising an *Amyelois* transposase as a protein or encoded by a nucleic acid, and/or an *Amyelois* transposon; or a gene transfer system as described herein comprising an *Amyelois* transposase as a protein or encoded by a nucleic acid as described herein, in combination with an *Amyelois* transposon; optionally together with a pharmaceutically acceptable carrier, adjuvant or vehicle, and optionally with instructions for use. Any of the components of the inventive kit may be administered and/or transfected into cells in a subsequent order or in parallel, e.g. an *Amyelois* transposase protein or its encoding nucleic acid may be administered and/or transfected into a cell as defined above prior to, simultaneously with or subsequent to administration and/or transfection of an *Amyelois* transposon. Alternatively, an *Amyelois* transposon may be transfected into a cell as defined above prior to, simultaneously with or subsequent to transfection of an *Amyelois* transposase protein or its encoding nucleic acid. If transfected in parallel, preferably both components are provided in a separated formulation and/or mixed with each other directly prior to administration to avoid transposition prior to transfection. Additionally, administration and/or transfection of at least one component of the kit may occur in a time staggered mode, e.g. by administering multiple doses of this component.

6. EXAMPLES

[00125] The following examples illustrate the methods, compositions and kits disclosed herein and should not be construed as limiting in any way. Various equivalents will be apparent from the following examples; such equivalents are also contemplated to be part of the invention disclosed herein.

6.1 A NEW TRANSPOSASE

6.1.1 MEASURING TRANSPOSASE ACTIVITY

[00126] As described in Section 5.2.5, transposition frequencies for active transposases may be measured using a system in which a transposon interrupts a selectable marker. Transposase reporter polynucleotides were constructed in which the open reading frame of the yeast *Saccharomyces cerevisiae* URA3 open reading frame was interrupted by a yeast TRP 1 open reading frame operably linked to a promoter and terminator such that it was expressible in the yeast *Saccharomyces cerevisiae*. The TRP1 gene was flanked by putative transposon ends with 5'-TTAA-3' target sites, such that excision of the putative transposon would leave a single copy of the 5'-TTAA-3' target site and exactly reconstitute the URA3

open reading frame. A yeast transposase reporter strain was constructed by integrating the transposase reporter polynucleotide into the URA3 gene of a haploid yeast strain auxotrophic for LEU2 and TRP1, such that the strain became LEU2-, URA3- and TRP1+.

[00127] Transposases were tested for their ability to transposase the TRP1 gene-containing transposons from within the URA3 open reading frame. Each open reading frame encoding a putative transposase was cloned into a *Saccharomyces cerevisiae* expression vector comprising a 2 micron origin of replication and a LEU2 gene expressible in *Saccharomyces*. Each transposase open reading frame was operably linked to a Gal1 promoter. Each cloned transposase open reading frame was transformed into a yeast transposase reporter strain and plated on minimal media lacking leucine. After 2 days, all LEU+ colonies were harvested by scraping the plates. The Gal promoter was induced by growing in galactose for 4 hours, and cells were then plated onto 3 different plates: plates lacking only leucine, plates lacking leucine and uracil, and plates lacking leucine, uracil and tryptophan. These plates were incubated for 2-4 days, and the colonies on each plate were counted, measuring the number of live cells, the number of transposon excision events and the number of transposon excision and re-integration (i.e. transposition events) respectively.

6.1.2 IDENTIFICATION OF AN ACTIVE *AMYELOIS* PIGGYBAC-LIKE TRANSPOSASE

[00128] As described in Section 5.2.5, twenty-one putative piggyBac-like transposases were identified from Genbank as being at least 20% identical to the piggyBac transposase from *Trichoplusia ni*. These putative transposases appeared to comprise the DDDE motif characteristic of active piggyBac-like transposases. The flanking DNA sequences were analyzed for the presence of inverted repeat sequences immediately adjacent to the 5'-TTAA-3' target sequence characteristic of piggyBac transposition. Putative left and right transposon end sequences comprising the sequence between the 5'-TTAA-3' target sequence and the open reading frame encoding the putative transposase were taken from these flanking sequences. These transposon ends were incorporated into transposase reporter constructs configured as described in Section 6.1.1 and integrated into the genome of *Saccharomyces cerevisiae* thereby generating transposase reporter strains. The corresponding transposase sequence for each reporter strain was back-translated, synthesized, cloned into a *Saccharomyces cerevisiae* expression vector and transformed into the reporter strain. Transposase activities were measured as described in Section 6.1.1.

[00129] The following twenty combinations showed no excision or transposition: reporter construct SEQ ID NO: 196 (comprising putative left transposon end SEQ ID NO: 216, and putative right transposon end SEQ ID NO: 217) with transposase SEQ ID NO: 172, reporter construct SEQ ID NO: 197 (comprising putative left transposon end SEQ ID NO: 218, and putative right transposon end SEQ ID NO: 219) with transposase SEQ ID NO: 173, reporter construct SEQ ID NO: 198 (comprising putative left transposon end SEQ ID NO: 220, and putative right transposon end SEQ ID NO: 221) with transposase SEQ ID NO: 174, reporter construct SEQ ID NO: 199 (comprising putative left transposon end SEQ ID NO: 222, and putative right transposon end SEQ ID NO: 223) with transposase SEQ ID NO: 175, reporter construct SEQ ID NO: 200 (comprising putative left transposon end SEQ ID NO: 224, and putative right transposon end SEQ ID NO: 225) with transposase SEQ ID NO: 176, reporter construct SEQ ID NO: 201 (comprising putative left transposon end SEQ ID NO: 226, and putative right transposon end SEQ ID NO: 227) with transposase SEQ ID NO: 177, reporter construct SEQ ID NO: 202 (comprising putative left transposon end SEQ ID NO: 228, and putative right transposon end SEQ ID NO: 229) with transposase SEQ ID NO: 178, reporter construct SEQ ID NO: 203 (comprising putative left transposon end SEQ ID NO: 230, and putative right transposon end SEQ ID NO: 231) with transposase SEQ ID NO: 179, reporter construct SEQ ID NO: 204 (comprising putative left transposon end SEQ ID NO: 232, and putative right transposon end SEQ ID NO: 233) with transposase SEQ ID NO: 180, reporter construct SEQ ID NO: 205 (comprising putative left transposon end SEQ ID NO: 234, and putative right transposon end SEQ ID NO: 235) with transposase SEQ ID NO: 181, reporter construct SEQ ID NO: 206 (comprising putative left transposon end SEQ ID NO: 236 and putative right transposon end SEQ ID NO: 237) with transposase SEQ ID NO: 182, reporter construct SEQ ID NO: 207 (comprising putative left transposon end SEQ ID NO: 238, and putative right transposon end SEQ ID NO: 239) with transposase SEQ ID NO: 183, reporter construct SEQ ID NO: 208 (comprising putative left transposon end SEQ ID NO: 240, and putative right transposon end SEQ ID NO: 241) with transposase SEQ ID NO: 184, reporter construct SEQ ID NO: 209 (comprising putative left transposon end SEQ ID NO: 242, and putative right transposon end SEQ ID NO: 243) with transposase SEQ ID NO: 185, reporter construct SEQ ID NO: 210 (comprising putative left transposon end SEQ ID NO: 244, and putative right transposon end SEQ ID NO: 245) with transposase SEQ ID NO: 186, reporter construct SEQ ID NO: 211 (comprising putative left transposon end SEQ ID NO: 246, and putative right transposon end SEQ ID NO: 247) with transposase SEQ ID NO: 187, reporter construct SEQ ID NO: 212 (comprising putative left transposon end SEQ ID NO:

248, and putative right transposon end SEQ ID NO: 249) with transposase SEQ ID NO: 188, reporter construct SEQ ID NO: 213 (comprising putative left transposon end SEQ ID NO: 250, and putative right transposon end SEQ ID NO: 251) with transposase SEQ ID NO: 189, reporter construct SEQ ID NO: 214 (comprising putative left transposon end SEQ ID NO: 252, and putative right transposon end SEQ ID NO: 253) with transposase SEQ ID NO: 190, reporter construct SEQ ID NO: 215 (comprising putative left transposon end SEQ ID NO: 254 and putative right transposon end SEQ ID NO: 255) with transposase SEQ ID NO: 191. This is consistent with reports in the literature that while computational recognition of sequences that are homologous to the piggyBac transposase from *Trichoplusia ni* is straightforward, most of these sequences are transpositionally inactive, even when they appear to have intact terminal repeats and the transposases appear to comprise the DDDE motif found in active piggyBac-like transposases. It is therefore necessary to measure excision and transposition activity, in order to identify novel active piggyBac-like transposases and transposons.

[00130] One transposase that showed good activity in excising its corresponding transposon from the reporter construct (shown by the appearance of URA+ colonies) and transposing the TRP gene in the transposon into another genomic location in the *Saccharomyces cerevisiae* reporter strain was transposase SEQ ID NO: 18. Transposase SEQ ID NO: 18 was able to transpose the transposon from reporter construct SEQ ID NO: 192. This is shown in Table 2: the number of excision events, measured by the appearance of URA+ colonies, is shown in column G; the number of full transposition events, measured by the appearance of URA+ TRP+ colonies, is shown in column H.

6.1.3 THE *AMYELOIS* TRANSPOSASE IS ACTIVE IN MAMMALIAN CELLS

[00131] PiggyBac-like transposases can transpose their corresponding transposons into the genomes of eukaryotic cells including yeast cells such as *Pichia pastoris* and *Saccharomyces cerevisiae*, and mammalian cells such as human embryonic kidney (HEK) and Chinese hamster ovary (CHO) cells. To determine the activity of piggyBac-like transposases in mammalian cells, we constructed gene transfer polynucleotides comprising transposon ends, and further comprising a selectable marker encoding glutamine synthetase with a polypeptide sequence given by SEQ ID NO: 277, encoded by DNA sequence given by SEQ ID NO: 282 and operably linked to regulatory elements that give weak glutamine synthetase expression, the sequence of the glutamine synthetase and its associated regulatory elements given by SEQ ID NO: 320. The gene transfer polynucleotides further comprised

open reading frames encoding the heavy and light chains of an antibody, each operably linked to a promoter and polyadenylation signal sequence. The gene transfer polynucleotide (with SEQ ID NO: 256) comprised a left transposon end comprising a 5'-TTAA-3' target integration sequence immediately followed by an *Amyelois* left transposon end with ITR sequence given by SEQ ID NO: 11, which is an embodiment of SEQ ID NO: 9. The gene transfer polynucleotide further comprised an *Amyelois* right transposon end with ITR sequence given by SEQ ID NO: 12 (which is an embodiment of SEQ ID NO: 10) immediately followed by a 5'-TTAA-3' target integration sequence. The two *Amyelois* transposon ends were placed on either side of the heterologous polynucleotide comprising the glutamine synthetase selectable marker and the open reading frames encoding the heavy and light chains of the antibody. The left transposon end further comprised a sequence given by SEQ ID NO: 7 immediately adjacent to the left ITR and proximal to the heterologous polynucleotide. The right transposon end further comprised a sequence given by SEQ ID NO: 8 immediately adjacent to the right ITR and proximal to the heterologous polynucleotide.

[00132] Gene transfer polynucleotides were transfected into CHO cells which lacked a functional glutamine synthetase gene. Cells were transfected by electroporation with 25 µg of gene transfer polynucleotide DNA, either with or without a co-transfection with 3 µg of DNA comprising a gene encoding a transposase operably linked to a human CMV promoter and a polyadenylation signal sequence. The cells were incubated in media containing 4 mM glutamine for 48 hours following electroporation, and subsequently diluted to 300,000 cells per ml in media lacking glutamine. Cells were exchanged into fresh glutamine-free media every 5 days. The viability of the cells from each transfection were measured at various times following transfection using a Beckman-Coulter Vi-Cell. The total number of viable cells were also measured with the same instrument. The results are shown in Table 3.

[00133] As shown in Table 3, the viability of cells transfected with the gene transfer polynucleotide, but no transposase fell to about 27% by 12 days post-transfection (column B). The total number of live cells fell to fewer than 50,000 per ml within 7 days (column C). At or below this density of live cells, viability measurements become inaccurate. The culture never recovered. In contrast when gene transfer polynucleotide with SEQ ID NO: 256 was co-transfected with *Amyelois* transposase SEQ ID NO: 18, cells recovered to greater than 90% viability within 17 days (Table 3 column D), by which time the density of live cells exceeded 2 million per ml (Table 3 column E). This shows that a gene transfer polynucleotide comprising a left and right *Amyelois* transposon end can be efficiently

transposed into the genome of a mammalian target cell by a corresponding *Amyelois* transposase.

[00134] The recovered pools of CHO cells comprising piggyBac-like transposons integrated into their genomes were grown in a 14-day fed-batch using Sigma Advanced Fed Batch media. Antibody titers were measured in culture supernatant using an Octet. Table 4 shows the titers measured at 7, 10, 12 and 14 days of the fed batch culture. The titer of antibody from cells comprising gene transfer polynucleotide with SEQ ID NO: 256, that had been integrated by co-transfection with the *Amyelois* transposase SEQ ID NO: 18 reached approximately 2 g/L after 14 days. This shows that the *Amyelois* transposon and its corresponding transposase, as described in Section 5.2.5, is a novel, piggyBac-like transposon/transposase system that is active in mammalian cells and useful for developing protein expressing cell lines and engineering the genomes of mammalian cells.

6.1.4 MESSENGER RNA ENCODING THE *AMYELOIS* TRANSPOSASE IS ACTIVE IN MAMMALIAN CELLS

[00135] We further tested gene transfer polynucleotide with SEQ ID NO: 256, whose configuration is described in Section 6.1.3, to determine whether the synthetic *Amyelois* transposon could be integrated into the genome of a mammalian cell if the corresponding transposase was provided in the form of mRNA. Gene transfer polynucleotide 354498 with SEQ ID NO: 256 comprised a selectable marker encoding glutamine synthetase with a polypeptide sequence given by SEQ ID NO: 277, encoded by DNA sequence given by SEQ ID NO: 282 and operably linked to regulatory elements that give weak glutamine synthetase expression, the sequence of the glutamine synthetase and its associated regulatory elements given by SEQ ID NO: 320. Gene transfer polynucleotide SEQ ID NO: 256 further comprised open reading frames encoding the heavy and light chains of an antibody, each operably linked to a promoter and polyadenylation signal sequence. Gene transfer polynucleotide SEQ ID NO: 256 further comprised an *Amyelois* left transposon end with sequence given by SEQ ID NO: 1 and an *Amyelois* right transposon end with sequence given by SEQ ID NO: 3.

[00136] mRNA encoding *Amyelois* transposase was prepared by *in vitro* transcription using T7 RNA polymerase. The mRNA comprised a 5' sequence SEQ ID NO: 257 preceding the open reading frame, an open reading frame encoding an Amyelois transposase (amino acid sequence SEQ ID NO: 18, nucleotide sequence SEQ ID NO: 171), and a 3' sequence SEQ ID NO: 258 following the stop codon at the end of the open reading frame. The mRNA had an anti-reverse cap analog (3'-O-Me-m^{7}G(5')ppp(5')G. DNA molecules comprising a

sequence encoding a transposase operably linked to a heterologous promoter that is active *in vitro* are useful for the preparation of transposase mRNA. Isolated mRNA molecules comprising a sequence encoding a transposase are useful for transposition of a corresponding transposon into a target genome.

[00137] Gene transfer polynucleotide SEQ ID NO: 256 was transfected into CHO cells which lacked a functional glutamine synthetase gene. Cells were transfected by electroporation: 25 µg of gene transfer polynucleotide DNA was co-transfected with 3 µg of mRNA comprising an open reading frame encoding a corresponding transposase (amino acid sequence SEQ ID NO: 18, nucleotide sequence SEQ ID NO: 171). The cells were incubated in media containing 4 mM glutamine for 48 hours following electroporation, and subsequently diluted to 300,000 cells per ml in media lacking glutamine. Cells were exchanged into fresh glutamine-free media every 5 days. The viability of the cells from each transfection were measured at various times following transfection using a Beckman-Coulter Vi-Cell. The total number of viable cells were also measured with the same instrument. The results are shown in Table 5.

[00138] When gene transfer polynucleotide with SEQ ID NO: 256 was co-transfected with mRNA encoding *Amyelois* transposase SEQ ID NO: 18, viability fell to around 26% by 9 days post-transfection (Table 5 column B), by which time the density of live cells was around 50,000 per ml (Table 5 column C). Cell viability and the density of live cells then increased until by 26 days post-transfection viability was above 92% and there were over 1 million live cells per ml. This shows that a gene transfer polynucleotide comprising a left and right *Amyelois* transposon end can be efficiently transposed into the genome of a mammalian target cell when co-transfected with mRNA encoding a corresponding *Amyelois* transposase.

6.1.5 *AMYELOIS* TRANSPOSON END SEQUENCES ACTIVE IN MAMMALIAN CELLS

[00139] When we originally tested the *Amyelois* transposon, we used the entire sequence between the 5'-TTAA-3' target sequences and the transposase open reading frame as transposon ends. We have found that for other piggyBac-like sequences this full sequence is generally not required for transposition activity. We therefore constructed synthetic *Amyelois* transposons with truncated ends to determine whether these were transposable by an *Amyelois* transposase. A heterologous polynucleotide with SEQ ID NO: 370 encoded glutamine synthetase with a polypeptide sequence given by SEQ ID NO: 278, operably linked to regulatory elements that give weak glutamine synthetase expression as a selectable

marker. On one side of the heterologous polynucleotide was a left *Amyelois* transposon end comprising a 5'-TTAA-3' integration target sequence immediately followed by a transposon ITR sequence with SEQ ID NO: 11, which is an embodiment of SEQ ID NO: 9. On the other side of the heterologous polynucleotide was a right *Amyelois* transposon end comprising a transposon ITR sequence with SEQ ID NO: 12 (which is an embodiment of SEQ ID NO: 10) immediately followed by a 5'-TTAA-3' integration target sequence. The transposon further comprised an additional sequence selected from SEQ ID NOs: 7 or 13 immediately adjacent to (following) the left transposon ITR sequence. The transposon further comprised an additional sequence selected from SEQ ID NOs: 8 or 16 immediately adjacent to (preceding) the right transposon ITR sequence. Transposons were transfected into CHO cells which lacked a functional glutamine synthetase gene. Cells were transfected by electroporation: 25 μg of gene transfer polynucleotide DNA were transfected, optionally the cells were co-transfected with 3 μg of mRNA comprising an open reading frame encoding a corresponding transposase (amino acid sequence SEQ ID NO: 18, nucleotide sequence SEQ ID NO: 171). The cells were incubated in media containing 4 mM glutamine for 48 hours following electroporation, and subsequently diluted to 300,000 cells per ml in media lacking glutamine. Cells were exchanged into fresh glutamine-free media every 5 days. The viability of the cells from each transfection were measured at various times following transfection using a Beckman-Coulter Vi-Cell. The total number of viable cells were also measured with the same instrument. The results are shown in Table 6.

[00140] Table 6 columns B and C show the reduction in cell viability and viable cell density when cells were transfected with a transposon comprising full length transposon ends in the absence of transposase. Cell viability and viable cell density can both be seen to fall throughout the experiment. In contrast when any the same transposon was co-transfected with mRNA encoding an *Amyelois* transposase, the cell viability and viable cell density fell initially, but had begun to recover by day 18 and was fully recovered between day 25 and 29 (Table 6 columns C and D). A comparable result was obtained when the left transposon end was truncated from the sequence given by SEQ ID NO: 7, to the sequence given by SEQ ID NO: 13 (compare Table 6 columns E and F with columns G and H respectively). A comparable result was also obtained when the right transposon end was truncated from the sequence given by SEQ ID NO: 8, to the sequence given by SEQ ID NO: 16 (compare Table 6 columns I and J with columns K and L respectively). This shows that in addition to a 5'-TTAA-3' integration target sequence immediately adjacent to a transposon ITR sequence with SEQ ID NO: 9, an *Amyelois* synthetic transposon left transposon end may further

comprise an additional sequence selected from SEQ ID NOs: 7 and 13 immediately adjacent to the left transposon ITR sequence; and an *Amyelois* synthetic transposon right transposon end may comprise an additional sequence selected from SEQ ID NOs: 8 and 16 immediately adjacent to the right transposon ITR sequence.

6.1.6 ENGINEERING HYPERACTIVE *AMYELOIS* TRANSPOSASES

[00141] To identify *Amyelois* transposase mutations that led to either increased transposition activity, or increased excision activity, relative to the naturally occurring *Amyelois* transposase sequence given by SEQ ID NO: 18, we analyzed a CLUSTAL alignment of active piggyBac-like transposases. Table 1 column C shows the amino acids found in active piggyBac-like transposases relative to each position in the *Amyelois* transposase (position shown in Table 1 column A). The amino acid present in *Amyelois* transposase given by SEQ ID NO: 18 is shown in column B of Table 1. Because transposases are often deleterious to their hosts, they tend to accumulate mutations that inactivate them. The mutations that accumulate in different transposases are different, as each occurs by random chance. A consensus sequence can therefore be used to approximate an ancestral sequence that pre-dates the accumulation of deleterious mutations. It is difficult to accurately calculate an ancestral sequence from a small number of extant sequences, so we chose to focus on positions where active transposases were more highly conserved, and where the consensus amino acid(s) differed from the one in the *Amyelois* transposase. We considered that mutating these amino acids to the consensus amino acids found in other active transposases would be likely to increase the activity of the *Amyelois* transposase. These candidate beneficial amino acid substitutions are shown in Table 1 column D.

6.1.6.1 First set of Amyelois Transposase Variants

[00142] A set of 95 genes encoding variant *Amyelois* transposases comprised one or more substitutions selected from P65E, D79N, R95S, V100I, V106P, L115D, E116P, H121Q, V131E, K139E, T159N, V166F, G174A, G179N, W187F, P198R, L203R, I209L, N211R, A225R, P232A, E238D, V261L, L273M, L273I, D304R, D304K, Y310G, I323L, Q329G, Q329R, M336C, Y343I, T345L, K362R, T366R, C367E, T380S, L408M, E413S, S416E, I426M, K432Q, S435G, K442Q, N452K, L458M, N466D, T467M, A472S, V475I, N483K, I491M, A529P, K540R, D551R, S560K, T562K and S563K. Each substitution was represented at least 5 times within the set of 96 variants, and the number of different pairwise combinations of substitutions was maximized so that each substitution was tested in as many

different sequence contexts as possible. Each variant gene was cloned into a vector comprising a leucine selectable marker; each gene encoding a transposase variant was operably linked to the *Saccharomyces cerevisiae* Gal-1 promoter. Each of these variants was then individually transformed into a *Saccharomyces cerevisiae* strain comprising a chromosomally integrated copy of SEQ ID NO: 192, as described above. After 48 hours cells were scraped from the plate into minimal media lacking leucine and with galactose as the carbon source. The A600 for each culture was adjusted to 2. Cultures were grown for 4 hours in galactose to induce expression of the transposases, then a 100x-diluted aliquot was plated on media lacking leucine, uracil and tryptophan (to count transposition), a 100x-diluted aliquot was plated on media lacking leucine and uracil (to count excision) and a 25,000x-diluted aliquot was plated on media lacking leucine (to count total live cells). Two days later, colonies were counted to determine transposition (= number of cells on –leu-ura-trp media divided by (250 x number of cells on –leu media)) and excision (= number of cells on –leu-ura media divided by (250 x number of cells on –leu media)) frequencies. The results are shown in Table 7. We identified 18 *Amyelois* transposase variants (with sequences given by SEQ ID NO: 96-113) with excision or transposition activities that were between 1.2 and 5-fold higher than the activities measured for the naturally occurring *Amyelois* transposase. We also identified 17 *Amyelois* transposase variants (with sequences given by SEQ ID NO: 19-35) with excision or transposition activities that were comparable to the naturally occurring *Amyelois* transposase. We also identified 45 *Amyelois* transposase variants (with sequences given by SEQ ID NO: 51-95) with excision or transposition activities that were less than the naturally occurring *Amyelois* transposase. Only 15 of the 95 variants (with sequences given by SEQ ID NO: 36-50) possessed activities that were so low that they were essentially inactive.

[00143] The effects of sequence changes on transposition frequencies were modelled using two different methods as described in Liao et al (2007, BMC Biotechnology 2007, 7:16 doi:10.1186/1472-6750-7-16 "Engineering proteinase K using machine learning and synthetic genes") and US patent 8,635,029. Mean values and standard deviations for the regression weights were calculated for each substitution, these are shown in Table 8. The effect of an individual substitution upon transposase activity may vary depending on the context (ie the other substitutions present). A positive mean regression weight indicates that on average, considering all of the different sequence contexts in which it has been tested, the substitution has a positive influence on the measured property. Incorporation of substitutions with positive mean regression weights into a sequence generally results in variants with

improved activity (Liao et. al., ibid). A further measure of this context-dependent variability is the standard deviation of the regression weight. If the mean regression weight plus the standard deviation of the regression weight for a substitution is zero or greater, then there are contexts within which the substitution has a positive effect. If the mean regression weight for a substitution is zero or greater, then the average effect of the substitution in all measured contexts is positive (Liao et al, ibid). If the mean regression weight minus the standard deviation of the regression weight for a substitution is zero or greater, then the substitution has a positive effect in the majority of sequence contexts (Liao et al, ibid). Thirty-three of the fifty nine substitutions we selected by looking for changes toward the consensus in other active piggyBac-like transposases had a mean regression weight of zero or greater as determined by one of the methods used: D79N, R95S, L115D, E116P, H121Q, K139E, V166F, G179N, W187F, P198R, L203R, N211R, E238D, L273M, L273I, D304R, D304K, Q329G, T345L, K362R, T366R, L408M, S416E, S435G, L458M, V475I, N483K, I491M, A529P, K540R, S560K, S563K. In addition to identifying specific substitutions with a beneficial effect, this also provides an indication of positions at which analogous substitutions may be beneficial. For example, replacement of the acidic aspartate at position 304 with either the basic residue lysine or the basic residue arginine has a positive effect. This provides evidence that multiple different analogous substitutions at a position can be beneficial. Analogous substitutions are those in which properties of the amino acids are conserved. For example: glycine and alanine are in the "small" amino acid group; valine, leucine, isoleucine and methionine are in the "hydrophobic" amino acid group; phenylalanine, tyrosine and tryptophan are in the "aromatic" amino acid group; aspartate and glutamate are in the "acidic" amino acid group; asparagine and glutamine are in the "amide" amino acid group; histidine, lysine and arginine are in the "basic" amino acid group; cysteine, serine and threonine are in the "nucleophilic" amino acid group. If a substitution at an amino acid position within the *Amyelois* transposase is beneficial for excision or transposition activity, other substitutions at the same position drawn from the same amino acid group are likely to be beneficial. For example, since replacing the basic residue lysine at position 139 with the acidic residue glutamate (K139E) is beneficial, replacing with the acidic residue aspartate (i.e. K139D) is likely also to be beneficial. Similarly, since replacing the hydrophobic residue valine at position 166 with the aromatic residue phenylalanine is beneficial, replacing with the aromatic residues tyrosine or tryptophan (i.e. V166Y or V166W) are likely also to be beneficial. An advantageous hyperactive *Amyelois* transposase comprises an amino acid substitution at one or more positions selected from amino acid 79,

95, 115, 116, 121, 139, 166, 179, 187, 198, 203, 211, 238, 273, 273, 304, 304, 329, 345, 362, 366, 408, 416, 435, 458, 475, 483, 491, 529, 540, 560 and 563 for example one or more substitutions selected from D79N, R95S, L115D, E116P, H121Q, K139E, V166F, G179N, W187F, P198R, L203R, N211R, E238D, L273M, L273I, D304R, D304K, Q329G, T345L, K362R, T366R, L408M, S416E, S435G, L458M, V475I, N483K, I491M, A529P, K540R, S560K and S563K, or analogous changes at the same positions.

[00144] Of the 59 amino acid substitutions selected, only one (G174A) was found exclusively in essentially inactive transposases. This indicates that amino acid changes that make the *Amyelois* transposase sequence closer to a consensus sequence for active piggyBac-like transposases are substitutions that can be incorporated the *Amyelois* transposase to create variants that retain transposition and excision activity. These are the changes listed in Table 1 column D. The positive results from the initial set of substitutions provide evidence that other changes in column D of Table 1 can also be incorporated into an *Amyelois* transposase to improve its activity.

6.1.6.2 Second set of *Amyelois* Transposase Variants

[00145] As described in Liao et al (2007, BMC Biotechnology 2007, 7:16 doi:10.1186/1472-6750-7-16 "Engineering proteinase K using machine learning and synthetic genes"), and US patent 8,635,029, Sections 5.4.2 and 5.4.3, substitutions that have been tested several times in the contexts of different combinations of other substitutions and that have "a positive regression coefficient, weight or other value describing its relative or absolute contribution to one or more activity" of a protein are usefully incorporated into a protein to obtain a protein that is "improved for one or more property, activity or function of interest". Based on the substitution weights shown in Table 8, we designed a set of open reading frames encoding 57 new variants (with sequences given by SEQ ID NOs: 114-170) combining some of the most positive substitutions (R95S, K139E, V166F, G179N, W187F, P198R, L203R, N211R, D304K, D304R, T345L, T366R, K540R). Each substitution was represented at least 10 times within the set of 57 variants, and the number of different pairwise combinations of substitutions was maximized so that each substitution was tested in as many different sequence contexts as possible. Each variant open reading frame was cloned into a vector comprising a leucine selectable marker; each open reading frame encoding a transposase variant was operably linked to the *Saccharomyces cerevisiae* Gal-1 promoter. Each of these variants was then individually transformed into a *Saccharomyces*

cerevisiae strain comprising a chromosomally integrated copy of SEQ ID NO: 192, as described in Section 6.1.6.1. After 48 hours cells were scraped from the plate into minimal media lacking leucine and with galactose as the carbon source. The A600 for each culture was adjusted to 2. Cultures were grown for 4 hours in galactose to induce expression of the transposases, then a 5,000x-diluted aliquot was plated on media lacking leucine, uracil and tryptophan (to count transposition) and a 25,000x-diluted aliquot was plated on media lacking leucine (to count total live cells). Two days later, colonies were counted to determine transposition (= number of cells on –leu -ura -trp media divided by (5 x number of cells on –leu media)) frequencies. The results are shown in Table 9.

[00146] In addition to the activities of the 57 new *Amyelois* transposase variants, Table 9 also shows the activities of 12 variants from the first set that were among the most active variants in that set. More than half of the new set of variants had greater transposition activity than the best variants from the first set, and all of the new variants had substantially (more than 50-fold) more transposition activity than the naturally occurring *Amyelois* transposase. A preferred hyperactive *Amyelois* transposase comprises an amino acid substitution selected from R95S, K139E, V166F, G179N, W187F, P198R, L203R, N211R, D304K, D304R, T345L, T366R, K540R, or analogous changes at the same positions.

BRIEF DESCRIPTION OF TABLES

Table 1. Amino acid changes likely to result in enhanced transposase activity.

[00147] Amino acid substitutions with the potential to improve transposase activity were identified as described in Section 5.2.6. Column A shows the position in an *Amyelois* transposase (relative to SEQ ID NO: 18), column B shows the amino acid in the native protein, column C shows the amino acids found in known active piggyBac-like transposases at the equivalent position in an alignment, column D shows amino acid changes found in known active piggyBac-like transposases other than the *Amyelois* transposase at positions where there is good conservation within the rest of the transposase set, but the amino acid in the *Amyelois* transposase sequence is an outlier. Mutation to these amino acids are particularly likely to result in enhanced transposase activity. More than one amino acid letter in column means that each of those individual amino acid substitutions are acceptable or beneficial, it is not intended to represent a peptide. For example, at position 2, amino acids T, A, R, D or N are all acceptable, so column C contains "TARDN" to indicate this.

Table 2. Excision and transposition of transposons in yeast.

[00148] Transposon and transposase sources are listed in column A. The left sequence with SEQ ID NO shown in column B and the right sequence with SEQ ID NO shown in column C were used to construct reporter plasmids as described in Section 6.1.2. The reporter plasmids have insert sequence given by the SEQ ID NO listed in column D. These reporter plasmids were integrated into the Ura3 gene of a Trp- strain of *Saccharomyces cerevisiae*. The amino acid sequence given by the SEQ ID NO shown in column E was backtranslated, synthesized and cloned into a plasmid comprising a Leu2 gene expressible in *Saccharomyces cerevisiae* and 2 micron origin of replication. The transposase gene was operably linked to a Gal1 promoter. The plasmid comprising the transposase was transformed into the reporter strain, expression was induced, and cells were plated as described in Section 6.1.1. Induced culture was diluted 25,000-fold prior to plating 100 µl on leu dropout plates, and 100-fold prior to plating 100 µl on leu ura or leu ura trp dropout plates. Column F shows the number of colonies on the leu dropout plates; column G shows the number of colonies on the leu ura dropout plates (indicating excision of the transposon from the middle of the ura gene in the reporter); column H shows the number of colonies on the leu ura trp dropout plates (indicating excision of the transposon from the middle of the ura gene in the reporter and transposition to another site in the genome).

Table 3. Transposition of transposons into the genome of CHO target cells.

[00149] Cells were transfected with transposons and transposases as described in Section 6.1.3. The transposon SEQ ID NO is shown in row 1, the co-transfected transposase SEQ ID NO is shown in row 2. For each transfection, viability (the percentage of cells that are viable) and the total viable cell density (in millions of cells per ml) are shown in adjacent columns, as indicated in row 3. Rows 4-18 show these measurements at various times post-transfection, the days elapsed are shown in column A.

Table 4. Antibody production from transposons integrated into the genome of CHO target cells.

[00150] Cells were transfected with a transposon (with SEQ ID NO: 256) and DNA encoding a transposase with polypeptide sequence given by SEQ ID NO: 18 as described in Section 6.1.3. Recovery is shown in Table 3. During a 14 day fed batch antibody production run, the culture supernatant contained the concentration of antibody shown: column A shows the titer on Day 7; column B shows the titer on Day 10; column C shows the titer on Day 12; column D shows the titer on Day 14.

Table 5. Transposition of transposons into the genome of CHO target cells by mRNA-encoded transposase.

[00151] Cells were transfected with a transposon with SEQ ID NO: 256 and mRNA-encoding transposase with SEQ ID NO: 18 as described in Section 6.1.4. The viability (the percentage of cells that are viable) and the total viable cell density (in millions of cells per ml) are shown in columns B and C respectively, as indicated in row 1. Rows 2-12 show these measurements at various times post-transfection, the days elapsed since transfection are shown in column A.

Table 6. Transposition of transposons with truncated end sequences into the genome of CHO target cells.

[00152] Cells were transfected with a transposon and mRNA-encoded transposase as described in Section 6.1.5. The transposon comprised a left transposon end comprising a 5'-TTAA-3' integration target sequence immediately adjacent to a transposon ITR sequence with SEQ ID NO: 11 which was immediately adjacent to a left end sequence with SEQ ID NO shown in row 2. The transposon further comprised a right transposon end comprising a right end sequence with SEQ ID NO shown in row 3 immediately adjacent to a transposon ITR sequence with SEQ ID NO: 12 which was immediately adjacent to a 5'-TTAA-3' integration target sequence. The two transposon ends were placed on either side of a glutamine synthetase reporter with SEQ ID NO: 370. Row 1 shows the SEQ ID NO of the transposon including the target sequences, the ITRs, the additional transposon end sequences and the selectable marker. Row 4 shows the SEQ ID NO of the transposase encoded by the transfected mRNA. The viability (the percentage of cells that are viable) is indicated in columns labelled "V" in row 5 and the total viable cell density (in millions of cells per ml) is indicated in columns labelled "VCD" in row 5. Rows 6-15 show these measurements at various times post-transfection, the days elapsed since transfection are shown in column M.

Table 7. Transposition and excision activities of *Amyelois* transposase variants (first set).

[00153] Genes encoding *Amyelois* transposase variants were designed, synthesized and cloned as described in Section 6.1.6.1. SEQ ID NOs of each variant are given in column A. Genes were transformed into a Saccharomyces cerevisiae strain whose genome comprised a single copy of transposase reporter SEQ ID NO: 192 and plated on media lacking leucine. After 48 hours cells were scraped from the plate into minimal media lacking leucine and with galactose as the carbon source. The A600 for each culture was adjusted to 2. Cultures were grown for 4 hours in galactose to induce expression of the transposases. Cultures were diluted 100-fold into minimal media lacking leucine; one 100 µl aliquot was plated onto minimal media agar plates lacking leucine and uracil (to measure transposon excision) another 100 µl

aliquot was plated onto minimal media agar plates lacking leucine, tryptophan and uracil (to measure transposon transposition). Each culture was also diluted 25,000-fold and a 100 µl aliquot was plated onto minimal media agar plates lacking leucine (to measure live cells). After 48 hours colonies on each plate were counted, the number of colonies on plates lacking leucine are shown in column B, the number of colonies on plates lacking leucine, uracil and tryptophan are shown in column C, the number of colonies on plates lacking leucine and uracil are shown in column D. Column E shows the transposition frequency (calculated as the number in column C, divided by the number in column B, and further divided by 250) Column F shows the excision frequency (calculated as the number in column D, divided by the number in column B, and further divided by 250). Asterisks in column G indicate variants where the colonies on plates lacking leucine and uracil or plates lacking leucine, tryptophan and uracil were too dense to be accurately counted. Re-measurements of these variants are shown in Table 9.

Table 8. Model weights for amino acid substitutions in *Amyelois* transposase variants.

[00154] The effects of sequence changes on *Amyelois* transposase excision and transposition activities were modelled as described in Liao et al (2007, BMC Biotechnology 2007, 7:16 doi:10.1186/1472-6750-7-16 "Engineering proteinase K using machine learning and synthetic genes") and US patent 8,635,029. The mean values and standard deviations for the regression weights were calculated for each substitution. The position (relative to SEQ ID NO: 18) is shown in column A, the amino acid found at this position in SEQ ID NO: 18 is shown in column B. The tested amino acid substitution is shown in column C. The regression weight for the substitution on transposition activity using a log data transformation is shown in column D, the standard deviation for this regression weight is shown in column E. The regression weight for the substitution on transposition activity using a log data transformation is shown in column F, the standard deviation for this regression weight is shown in column G.

Table 9. Transposition activities of *Amyelois* transposase variants (second set).

[00155] Genes encoding *Amyelois* transposase variants were designed, synthesized and cloned as described in Section 6.1.6.2. SEQ ID NOs of each variant are given in column A. Genes were transformed into a Saccharomyces cerevisiae strain whose genome comprised a single copy of transposase reporter SEQ ID NO: 192 and plated on media lacking leucine. After 48 hours cells were scraped from the plate into minimal media lacking leucine and with galactose as the carbon source. The A600 for each culture was adjusted to 2. Cultures were grown for 4 hours in galactose to induce expression of the transposases. Cultures were diluted

5,000-fold into minimal media lacking leucine and one 100 µl aliquot was plated onto minimal media agar plates lacking leucine, tryptophan and uracil (to measure transposon transposition). Each culture was also diluted 25,000-fold and a 100 µl aliquot was plated onto minimal media agar plates lacking leucine (to measure live cells). After 48 hours colonies on each plate were counted, the number of colonies on plates lacking leucine are shown in column B, the number of colonies on plates lacking leucine, uracil and tryptophan are shown in column C. Column D shows the transposition frequency (calculated as the number in column C, divided by the number in column B, and further divided by 5).

TABLES

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
1	M		
2	A	TARDN	
3	R	SRQFI	
4	G	SGFRLYE	
5	L	GLTDSRA	
6	T	RTANDQ	
7	D	KDEQH	
8	L	RLEDHN	
9	E	SEAIR	
10	I	ILA	
11	N	GNLASR	
12	Q	NQLTAH	
13	I	VIFLCM	
14	L	HLFM	
15	E	NEDQA	
16	L	QLSNE	
17	E	REDSV	
18	D	ADSLE	
19	V	AVEDST	
20	E	KEYLD	
21	N	NESVFY	
22	D	RDIPGS	
23	V	RVSLEGDY	
24	I	AIEDS	
25	F	VFISD	
26	D	VDES	
27	E	ED	
28	S	SLY	
29	G	PGSVE	
30	D	GDEP	
31	E	TEAKVP	
32	S	RSEAT	
33	D	DS	
34	H	FHNRCE	
35	V	GVDSC	
36	S	TSIVD	
37	I	TIDES	TDES
38	R	LRDSH	LDSH
39	V	SVDE	
40	E	WEFQN	
41	S	LSQY	
42	D	DFW	
43	T	NTSC	
44	E	EDS	
45	E	DESQ	
46	V	SVEA	SEA
47	E	SEMIFTRA	
48	I	GIDV	

TABLE 1 (contd.)

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
49	P	SPD	SD
50	T	ETYFASP	
51	L	VLD	
52	E	EHD	
53	P	DPSEVL	
54	Q	DQVNE	
55	Q	TQVNIME	
56	G	IGSLET	
57	S	SPADVY	
58	S	QSED	
59	D	SDQRE	
60	S	ESNDP	
61	E	SENVD	
62	N	ENLAPID	
63	D	EDGN	
64	Q	QMLV	
65	P	VPE	VE
66	L	ALQVGD	
67	S	DSTQ	
68	N	HNSELA	
69	L	VLSA	
70	A	TAGR	
71	R	ERSDQ	ESDQ
72	R	ERGHN	
73	S	HSTRAM	
74	F	SFTY	
75	Y	FYI	
76	K	YKCRTS	
77	G	GS	
78	K	K	
79	D	NDG	NG
80	N	RNKEG	
81	T	YTIHP	
82	I	KIVCA	
83	W	W	
84	N	ANSGY	
85	R	CRTPK	
86	A	QASTPN	
87	P	PKC	
88	P	LPHGSNQ	
89	N	SNQRTF	
90	P	RPST	
91	R	ARTN	
92	V	VGSI	
93	R	R	
94	T	VTL	
95	R	PRS	PS
96	S	QSAE	
97	E	HESIL	
98	N	NP	

TABLE 1 (contd.)

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
99	I	IP	
100	V	IVF	IF
101	T	QTKR	
102	G	RGMTSE	
103	T	TNQVR	
104	P	NPRA	
105	G	VGQL	
106	V	SVP	SP
107	K	NKT	
108	R	LRVNT	
109	Q	TQFMDIG	
110	A	EACT	
111	K	DKVRS	
112	N	DNT	
113	A	PAVI	
114	L	KLDVYFS	
115	L	DLET	DET
116	E	PEI	PI
117	L	FLQIYS	
118	D	SDLNEK	
119	C	ICAF	
120	F	WF	
121	H	NHQK	NQK
122	L	KLI	
123	F	LF	
124	V	MVIF	
125	N	DNST	DST
126	E	DEQS	
127	S	ESPAD	
128	I	IM	
129	L	LEI	
130	S	QSRHD	
131	V	EVID	ED
132	I	TIM	
133	L	LV	
134	E	KEDLT	
135	H	WHMY	
136	T	T	
137	N	N	
138	H	EHLVAS	
139	K	KEYS	EYS
140	I	IGMA	
141	R	IRSE	
142	S	QSRVLH	
143	E	YEVKRS	
144	R	RFLQ	
145	Q	SQRTV	
146	G	KGNQ	
147	K	FKLENTS	
148	N	SNIKPA	

TABLE 1 (contd.)

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
149	T	DTMLA	
150	S	KSITPE	
151	N	DNSR	
152	E	EAS	
153	Y	PYEAHFV	
154	A	EAKSTYH	
155	Y	LYWFMK	
156	S	RSKHQ	
157	E	NESDP	
158	T	LTI	
159	T	DTN	DN
160	L	MLQTEI	
161	T	VTMDACS	
162	E	ED	
163	L	LMI	
164	R	HRNWYK	
165	A	AR	
166	V	FVYL	FYL
167	I	IVF	
168	G	GA	
169	L	LI	
170	L	LT	
171	Y	LYITV	
172	L	FLMAI	
173	A	TAM	
174	G	AG	A
175	L	VL	
176	F	FYMRIT	
177	K	KR	
178	S	SDA	
179	G	NGK	NK
180	R	HRG	
181	Q	EQLMS	
182	N	NASL	
183	L	VLTE	
184	Q	NQSDK	
185	D	YDSE	
186	L	LW	
187	W	FWD	FD
188	A	ANDTR	
189	S	TSAR	
190	D	DETS	
191	G	GEFLV	
192	T	TNSL	
193	G	GS	
194	I	RIV	
195	E	EPTMD	
196	I	IRV	
197	F	FY	
198	P	RPVS	RVS

TABLE 1 (contd.)

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
199	M	CMAST	
200	T	VT	
201	M	M	
202	S	S	
203	L	KLR	KR
204	R	NREDQ	
205	R	RT	
206	F	FY	
207	A	LAHEDQY	
208	F	VFML	
209	I	IL	L
210	V	LVSIQ	
211	N	HNR	HR
212	C	CVFNS	
213	L	LIM	
214	R	RH	
215	F	FM	
216	D	DN	
217	D	ND	
218	S	PSRKT	
219	D	DTSA	
220	T	DTLIV	
221	R	RP	
222	E	EVPD	
223	E	EGTD	
224	R	RLQ	
225	A	RAPK	RPK
226	A	EASGQK	
227	I	SIDNHT	
228	D	D	
229	R	KRAVN	
230	L	ILFM	
231	A	AILTH	
232	P	APK	AK
233	I	IVLF	
234	R	SR	
235	Q	YQDKPS	
236	I	IVLM	
237	Y	FYWI	FWI
238	E	TED	TD
239	E	KEILSQ	
240	F	FWL	
241	V	VIS	
242	K	GKENHQ	
243	N	NIQRC	
244	C	CLF	
245	K	QKPRIA	
246	D	KDLQAN	
247	V	IVLNA	
248	Y	YH	

TABLE 1 (contd.)

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
249	T	NTVS	
250	P	VP	
251	Y	CYGS	CGS
252	E	EPGSAQ	
253	N	YNHF	
254	L	ALVI	
255	T	TC	
256	I	VI	
257	D	D	
258	E	E	
259	E	MERQS	
260	L	L	
261	V	VL	L
262	A	PAGLS	
263	F	F	
264	R	RK	
265	G	G	
266	R	R	
267	C	TCL	
268	K	HKPQL	
269	F	LF	
270	R	MR	
271	Q	IQMV	
272	Y	Y	
273	L	MLI	MI
274	P	P	
275	N	MNS	
276	K	K	
277	P	PR	
278	A	ADS	
279	K	KR	
280	Y	Y	
281	G	G	
282	I	LI	
283	K	KR	
284	I	LIF	
285	I	MIWPLYF	
286	A	CAMK	
287	L	LAM	
288	V	CV	
289	D	DAE	
290	A	AS	
291	Y	NYKAGS	
292	T	NTS	
293	Y	GYSKF	
294	Y	Y	
295	S	FSAMTV	
296	L	YLWISV	
297	N	NKDY	
298	M	CMAGFL	

TABLE 1 (contd.)

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
299	E	YEQIML	
300	I	IVP	
301	Y	Y	
302	A	TALE	
303	G	G	
304	D	RDK	RK
305	Q	GQSD	
306	P	SPT	
307	D	DGKQSL	
308	G	GTL	
309	P	APND	
310	Y	GYP	GP
311	K	LKEVPA	
312	V	TVKPG	
313	S	ESNC	
314	N	ENQP	
315	K	VKLRT	
316	P	PGSA	
317	H	THMEGF	
318	D	QDRFYKE	
319	V	SVYI	
320	V	V	
321	D	IDLKWE	
322	R	HRED	
323	I	LIM	LM
324	V	AVSTI	
325	Q	KQES	
326	P	PGT	
327	I	LIV	LV
328	S	FSQHLA	
329	Q	GQR	GR
330	T	STQ	
331	G	NGCH	
332	R	RHF	
333	N	NH	
334	V	IVL	
335	T	TY	
336	M	CMVF	CVF
337	D	D	
338	N	N	
339	W	WF	
340	F	FY	
341	T	TS	
342	S	SG	
343	Y	IY	I
344	P	EPRT	
345	T	LT	L
346	Y	IYGAFM	
347	A	EAKTL	
348	H	YHENA	

TABLE 1 (contd.)

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
349	L	LM	
350	L	KLQY	
351	K	KQCN	
352	N	KNRAEL	
353	H	HY	
354	K	GKNDR	
355	L	LT	
356	T	TP	
357	A	CAMIS	
358	V	VLCT	
359	G	G	
360	T	T	
361	M	MVI	
362	K	KRN	RN
363	S	KSR	
364	N	N	
365	K	KR	
366	T	RTPK	RPK
367	C	ECGQ	EGQ
368	I	ILM	
369	P	P	
370	P	KPSERD	
371	K	EKVAS	EVAS
372	F	FIL	
373	R	LRKIT	
374	E	PEKNDR	
375	R	SRIKT	
376	R	RN	
377	E	DEPRQ	
378	I	VIMGP	
379	N	GNHEA	
380	T	ST	S
381	S	SY	
382	L	LIMAV	
383	F	YFL	
384	G	GACR	
385	F	YFK	
386	Q	AQTDN	
387	D	GDEK	GEK
388	D	QDKPL	
389	F	NFALI	NALI
390	T	TA	
391	I	IVL	
392	V	LVK	
393	S	SF	
394	Y	HYF	
395	I	VICKDA	
396	P	P	
397	K	K	
398	R	KRP	

TABLE 1 (contd.)

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
399	N	NSAK	
400	K	KR	
401	N	ANMV	
402	V	V	
403	F	IFLYV	
404	M	LMVA	
405	L	LM	
406	S	ST	
407	S	ST	
408	L	MLCI	MCI
409	H	HD	
410	H	HTED	
411	D	ADNE	
412	S	ESAN	
413	E	AESV	ASV
414	I	VIL	
415	D	DSNR	
416	S	ESTQ	ETQ
417	E	TERSQ	
418	T	TDNR	
419	G	GDV	
420	E	EG	
421	Q	QK	
422	Q	QMN	
423	K	K	
424	P	P	
425	S	ESQDL	
426	I	IMC	MC
427	I	IVS	
428	T	GTLMK	
429	F	FDYE	
430	Y	Y	
431	N	NS	
432	K	KSQ	SQ
433	T	TY	
434	K	KM	
435	S	GSA	GA
436	G	G	
437	V	V	
438	D	D	
439	N	ENSTRV	
440	V	IVLFT	
441	D	D	
442	K	KQE	QE
443	L	KLVM	
444	I	CITQS	
445	R	ARKSH	
446	T	ITSVYN	
447	Y	YM	
448	D	TDSN	TSN

TABLE 1 (contd.)

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
449	V	SVCA	
450	S	SQNT	
451	R	R	
452	N	RNK	RK
453	S	TS	
454	R	RANK	
455	R	RA	
456	W	W	
457	P	PY	
458	L	MLK	M
459	T	VTAK	
460	I	VIL	
461	F	FLG	
462	F	YFI	
463	W	RWNGY	
464	I	MILV	
465	L	LVI	
466	N	DNQ	DQ
467	T	ITVM	IVM
468	A	SA	
469	G	TGAFCLS	
470	I	VIYR	
471	N	N	
472	A	SA	S
473	K	HKYFC	
474	I	LIV	
475	V	IVL	IL
476	Q	YQW	YW
477	M	DMSCKRQ	
478	L	ILEHAT	
479	N	HNIA	
480	S	HSNKV	
481	S	SQIPN	
482	D	DENSG	
483	N	KNVA	KVA
484	T	TLPV	
485	P	TPYIQSV	
486	T	ETRNSYK	
487	R	RY	
488	R	GRKT	
489	A	MALEKY	
490	F	FQ	
491	I	LIM	ML
492	K	KERQ	
493	K	QKENIS	
494	L	L	
495	G	AGSYP	
496	M	RMKTIAL	
497	S	TSADQL	
498	L	LM	

TABLE 1 (contd.)

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
499	I	VITF	
500	A	LATSGY	
501	P	PSGE	
502	H	QHKWFV	
503	Q	MQIEL	
504	A	KAQREH	
505	E	REKQS	
506	R	RT	
507	K	AKLVN	
508	T	LTRKQEP	
509	N	NPAEK	
510	S	ESAPMK	
511	K	RKTPN	
512	I	LISP	
513	P	PKS	
514	V	RVADTF	
515	S	ESATYNH	
516	L	LAVI	
517	R	RA	
518	K	LKVDRQ	
519	R	SRINL	
520	I	LIE	
521	G	AGKTSE	
522	S	RSINK	
523	H	VHKIQ	
524	L	LF	
525	G	GRKPI	
526	E	PETNKD	
527	S	DSVETP	
528	S	MSNVLT	
529	A	PAQ	P
530	S	VSATR	
531	P	PMSH	
532	A	DAEGV	
533	K	PKND	
534	I	QIVNSM	
535	P	EPDSTR	
536	N	VNTE	
537	V	FVPM	
538	G	KG	
539	V	TVKPR	
540	K	RKQY	RQY
541	K	RKSTV	
542	R	RYG	
543	C	C	
544	Y	HYQGTKR	
545	I	TIVFYDE	
546	C	C	
547	P	PSR	
548	V	LVSYKN	

TABLE 1 (contd.)

A	B	C	D
amyelois_position	amyelois	Acceptable	Beneficial
549	K	KR	
550	K	LKDI	
551	D	QDR	QR
552	R	RS	
553	K	KMD	
554	S	STA	
555	K	TKSNR	
556	Y	HYTAR	
557	I	TISQY	
558	C	CF	
559	I	YIVCKPN	
560	S	TSKA	TKA
561	C	C	
562	T	KTPA	KPA
563	S	KSNR	KNR
564	H	HFAVNP	
565	I	VIL	
566	C	C	
567	L	LRGFM	
568	E	QEK	
569	H	CHP	
570	A	ATNC	
571	N	KNVIF	
572	F	QFTDE	
573	V	VFMIL	
574	C	CY	
575	E	AEPQH	
576	N	DNST	
577	C	CQ	
578	R	VRGIFLA	

TABLE 1

A	B	C	D	E	F	G	H
Source	Tposon left end	Tposon right end	Tposon SEQ ID NO	Tpase SEQ ID	leu	leu ura	leu ura trp
Amyelois transitella	1	3	192	18	454	435	201
Spodoptera litura	216	217	196	172	>250	0	0
Pieris rapae	218	219	197	173	>250	0	0
Myzus persicae	220	221	198	174	>250	0	0
Onthophagus taurus	222	223	199	175	>250	0	0
Temnothorax curvispinosus	224	225	200	176	>250	0	0
Agrilus planipenn	226	227	201	177	>250	0	0
Parasteatoda tepidariorum	228	229	202	178	>250	0	0
Pectinophora gossypiella	230	231	203	179	>250	0	0
Ctenopusia agnata	232	233	204	180	>250	0	0
Macrostomum lignano	234	235	205	181	>250	0	0
Orussus abietinus	236	237	206	182	>250	0	0
Eufriesea mexicana	238	239	207	183	323	0	0
Spodoptera litura	240	241	208	184	400	0	0
Vanessa tameamea	242	243	209	185	389	0	0
Blattella germanica	244	245	210	186	248	0	0
Onthophagus taurus	246	247	211	187	>250	0	0
Onthophagus taurus	248	249	212	188	>250	0	0
Onthophagus taurus	250	251	213	189	>250	0	0
Megachile rotundata	252	253	214	190	>250	0	0
Xiphophorus maculatus	254	255	215	191	>250	0	0

TABLE 2

		A	B	C	D	E
1	Transposon SEQ ID NO		256	256	256	256
2	Transposase SEQ ID NO		none	none	18	18
3		Day	viability	viable cells	viability	viable cells
4		1	94.12	1.03	88.76	0.85
5		3	92.15	0.55	93.02	0.21
6		5	80.66	0.22	77.66	0.26
7		7	57.58	0.05	50.79	0.06
8		10	27.18	0.03	56.25	0.08
9		12	27.05	0.04	75.82	0.25
11		14	31.88	0.04	88.87	1.12
12		17	41.46	0.04	97.56	3.38
14		19	no live cells	no live cells	98.74	>3
15		21	no live cells	no live cells	99.09	>3
16		24	no live cells	no live cells	>99	>3
17		26	no live cells	no live cells	>99	>3
18		27	no live cells	no live cells	>99	>3

TABLE 3

A	B	C	D
Day 7	Day 10	Day 12	Day 14
1,130	1,823	1,932	1,974

TABLE 4

	A	B	C
1	Days post	viability	viable cells
2	1	94.81	0.78
3	2	92.41	0.20
4	5	80.89	0.23
5	7	44.26	0.06
6	9	26.54	0.05
7	14	40.70	0.07
8	16	39.77	0.11
9	19	57.91	0.26
10	21	68.62	0.43
11	23	74.54	0.66
12	26	92.75	1.23

TABLE 5

		A	B	C	D	E	F	G	H	I	J	K	L	M
1	Transposon SEQ ID NO	193	193	193	193	194	194	194	194	195	195	195	195	
2	Left end SEQ ID NO	7	7	7	7	13	13	13	13	7	7	7	7	
3	Right end SEQ ID NO	8	8	8	8	8	8	8	8	16	16	16	16	
4	Transposase SEQ ID NO	none	none	18	18	none	none	18	18	none	none	18	18	
5		V	VCD	V	VCD	V	VCD	V	VCD	V	VCD	V	VCD	Day
6		98.3	3.807	99.0	3.299	98.3	3.481	99.0	3.550	98.6	3.781	98.8	2.961	2
7		97.1	0.210	95.0	0.255	91.8	0.151	90.7	0.114	97.8	0.138	95.5	0.371	4
8		57.8	0.067	64.0	0.101	66.4	0.066	59.1	0.073	69.6	0.078	70.2	0.180	8
9		31.5	0.026	36.2	0.050	31.4	0.029	38.8	0.028	39.1	0.047	37.9	0.069	11
10		24.5	0.013	34.0	0.032	20.0	0.011	25.2	0.026	17.9	0.005	23.9	0.041	13
11		16.8	0.017	39.6	0.095	17.5	0.015	32.4	0.021	8.5	0.009	15.1	0.025	18
12		15.5	0.017	62.0	0.201	15.5	0.010	59.2	0.093	12.9	0.014	26.7	0.052	20
13		12.0	0.010	76.9	0.432	16.5	0.014	74.8	0.164	19.8	0.025	34.9	0.082	21
14		20.3	0.013	97.2	4.714	20.5	0.015	98.3	4.347	17.8	0.018	94.7	0.832	25
15		14.3	0.011	94.5	1.296	10.3	0.008	99.4	9.460	1.4	0.001	99.4	5.123	29

TABLE 6

A	B	C	D	E	F	G
SEQ ID NO	live	trans	ex	trans freq	ex freq	
36	436	0	2	0.0000	0.0000	
37	662	22	20	0.0001	0.0001	
38	379	1	5	0.0000	0.0001	
39	581	4	24	0.0000	0.0002	
40	411	3	21	0.0000	0.0002	
41	269	3	3	0.0000	0.0000	
42	338	10	6	0.0001	0.0001	
43	305	4	9	0.0001	0.0001	
44	439	9	5	0.0001	0.0000	
45	269	3	7	0.0000	0.0001	
46	158	2	0	0.0001	0.0000	
47	203	7	13	0.0001	0.0003	
48	273	1	1	0.0000	0.0000	
49	205	4	6	0.0001	0.0001	
50	212	4	7	0.0001	0.0001	
96	353	1200	1133	0.0136	0.0128	*
97	521	1200	1122	0.0092	0.0086	*
98	336	499	568	0.0059	0.0068	*
99	513	900	933	0.0070	0.0073	*
100	367	511	580	0.0056	0.0063	*
101	237	274	427	0.0046	0.0072	
102	189	336	285	0.0071	0.0060	
103	295	239	367	0.0032	0.0050	
104	321	561	879	0.0070	0.0110	*
105	181	169	193	0.0037	0.0043	
106	275	340	241	0.0049	0.0035	
107	221	610	693	0.0110	0.0125	*
108	204	487	583	0.0095	0.0114	*
109	156	134	176	0.0034	0.0045	
110	138	299	460	0.0087	0.0133	
111	240	220	281	0.0037	0.0047	
112	227	520	554	0.0092	0.0098	*
113	365	1220	934	0.0134	0.0102	*
19	464	224	329	0.0019	0.0028	
20	332	153	197	0.0018	0.0024	
21	408	180	102	0.0018	0.0010	
22	323	157	117	0.0019	0.0014	
23	181	135	218	0.0030	0.0048	
24	470	255	295	0.0022	0.0025	
25	355	185	291	0.0021	0.0033	
26	381	209	207	0.0022	0.0022	
27	281	205	183	0.0029	0.0026	
28	401	274	333	0.0027	0.0033	
29	332	201	277	0.0024	0.0033	
30	284	137	137	0.0019	0.0019	
31	222	101	110	0.0018	0.0020	
32	326	153	158	0.0019	0.0019	
33	298	189	272	0.0025	0.0037	

TABLE 7 (contd.)

A	B	C	D	E	F	G
SEQ ID NO	live	trans	ex	trans freq	ex freq	
34	219	111	164	0.0020	0.0030	
35	323	248	285	0.0031	0.0035	
51	387	33	39	0.0003	0.0004	
52	381	63	83	0.0007	0.0009	
53	371	16	32	0.0002	0.0003	
54	309	38	18	0.0005	0.0002	
55	364	142	162	0.0016	0.0018	
56	349	101	87	0.0012	0.0010	
57	433	28	16	0.0003	0.0001	
58	427	112	129	0.0010	0.0012	
59	419	19	16	0.0002	0.0002	
60	420	92	91	0.0009	0.0009	
61	629	193	181	0.0012	0.0012	
62	308	26	42	0.0003	0.0005	
63	225	22	28	0.0004	0.0005	
64	226	63	46	0.0011	0.0008	
65	159	28	45	0.0007	0.0011	
66	520	121	162	0.0009	0.0012	
67	351	69	131	0.0008	0.0015	
68	173	26	29	0.0006	0.0007	
69	484	149	223	0.0012	0.0018	
70	686	28	24	0.0002	0.0001	
71	219	35	43	0.0006	0.0008	
72	198	43	64	0.0009	0.0013	
73	358	43	33	0.0005	0.0004	
74	470	17	29	0.0001	0.0002	
75	306	41	82	0.0005	0.0011	
76	380	127	138	0.0013	0.0015	
77	289	37	33	0.0005	0.0005	
78	271	93	145	0.0014	0.0021	
79	445	187	202	0.0017	0.0018	
80	393	21	32	0.0002	0.0003	
81	179	17	15	0.0004	0.0003	
82	259	71	91	0.0011	0.0014	
83	287	68	74	0.0009	0.0010	
84	233	49	62	0.0008	0.0011	
85	255	61	83	0.0010	0.0013	
86	247	20	13	0.0003	0.0002	
87	233	25	28	0.0004	0.0005	
88	203	76	132	0.0015	0.0026	
89	219	61	75	0.0011	0.0014	
90	294	107	140	0.0015	0.0019	
91	219	72	99	0.0013	0.0018	
92	220	81	71	0.0015	0.0013	
93	270	10	29	0.0001	0.0004	
94	286	22	31	0.0003	0.0004	
95	255	67	134	0.0011	0.0021	
18	247	167	228	0.0027	0.0037	

TABLE 7

A	B	C	D	E	F	G
Position	**Amino acid**	**Substitution**	**Weight method 1**	**Weight Std method 1**	**Weight method 2**	**Weight Std method 2**
79	D	N	-0.201	0.094	0.016	0.023
95	R	S	-0.157	0.090	0.054	0.039
115	L	D	0.274	0.117	-0.063	0.030
116	E	P	0.082	0.098	-0.007	0.022
121	H	Q	0.046	0.106	0.118	0.049
139	K	E	0.450	0.125	0.347	0.055
166	V	F	0.388	0.152	0.230	0.078
179	G	N	0.418	0.111	0.349	0.046
187	W	F	0.321	0.102	0.225	0.051
198	P	R	-0.081	0.084	0.116	0.038
203	L	R	0.414	0.107	0.050	0.026
211	N	R	0.251	0.081	0.036	0.025
238	E	D	-0.107	0.110	0.039	0.052
273	L	M	0.047	0.090	0.022	0.021
273	L	I	-0.081	0.097	0.017	0.047
304	D	R	0.506	0.084	0.319	0.047
304	D	K	0.451	0.083	0.289	0.046
329	Q	G	0.064	0.089	-0.030	0.024
345	T	L	0.224	0.088	0.092	0.043
362	K	R	0.004	0.083	0.008	0.027
366	T	R	0.200	0.093	-0.026	0.028
408	L	M	-0.119	0.111	0.094	0.047
416	S	E	0.010	0.130	-0.057	0.027
435	S	G	0.012	0.095	0.102	0.028
458	L	M	-0.078	0.177	0.198	0.062
475	V	I	0.109	0.123	0.130	0.053
483	N	K	-0.113	0.086	0.075	0.029
491	I	M	-0.121	0.112	0.010	0.036
529	A	P	-0.015	0.101	0.007	0.033
540	K	R	0.229	0.077	0.117	0.034
560	S	K	-0.142	0.117	0.110	0.043
563	S	K	0.027	0.098	0.065	0.027
426	I	M	-0.112	0.086	-0.009	0.021
329	Q	R	-0.036	0.082	-0.033	0.031
100	V	I	-0.105	0.091	-0.034	0.038
413	E	S	-0.132	0.083	-0.037	0.018
452	N	K	-0.208	0.106	-0.043	0.035
65	P	E	-0.064	0.087	-0.046	0.026
562	T	K	-0.035	0.090	-0.048	0.021
209	I	L	-0.003	0.097	-0.054	0.026
159	T	N	-0.245	0.092	-0.063	0.028
551	D	R	-0.492	0.084	-0.065	0.037
131	V	E	-0.408	0.177	-0.072	0.064
472	A	S	-0.117	0.098	-0.074	0.025
380	T	S	-0.072	0.103	-0.091	0.023
225	A	R	-0.218	0.096	-0.093	0.034
336	M	C	-0.443	0.104	-0.096	0.021

TABLE 8 (contd.)

A	B	C	D	E	F	G
Position	**Amino acid**	**Substitution**	**Weight method 1**	**Weight Std method 1**	**Weight method 2**	**Weight Std method 2**
323	I	L	-0.028	0.106	-0.102	0.027
367	C	E	-0.180	0.079	-0.116	0.021
310	Y	G	-0.404	0.099	-0.118	0.019
432	K	Q	-0.098	0.091	-0.129	0.027
261	V	L	-0.303	0.107	-0.151	0.031
106	V	P	-0.398	0.069	-0.197	0.025
232	P	A	-0.400	0.118	-0.215	0.035
343	Y	I	-0.632	0.074	-0.252	0.029
467	T	M	-1.152	0.114	-0.293	0.033
442	K	Q	-0.929	0.103	-0.309	0.036
466	N	D	-1.027	0.133	-0.319	0.040
174	G	A	-1.102	0.204	-0.336	0.047

TABLE 8

A	B	C	D
SEQ ID NO	leu	leu ura trp	transposition frequency
170	123	393	0.639
156	178	449	0.504
159	135	334	0.495
155	216	498	0.461
168	118	264	0.447
143	183	408	0.446
160	151	331	0.438
154	458	1003	0.438
127	190	383	0.403
167	145	270	0.372
119	173	304	0.351
131	260	450	0.346
140	233	391	0.336
158	330	545	0.330
169	251	413	0.329
125	278	434	0.312
164	192	294	0.306
165	176	265	0.301
163	231	340	0.294
157	297	436	0.294
162	370	532	0.288
135	200	287	0.287
161	240	339	0.283
141	276	382	0.277
148	243	322	0.265
133	260	339	0.261
147	215	272	0.253
114	319	398	0.250
137	240	294	0.245
153	228	278	0.244
166	300	363	0.242
142	366	441	0.241
97	220	262	0.238
124	366	426	0.233
113	312	360	0.231
96	381	439	0.230
152	238	271	0.228
134	294	334	0.227
144	226	246	0.218
118	357	387	0.217
132	347	364	0.210
146	412	412	0.200
150	321	320	0.199
151	269	263	0.196
136	339	320	0.189
115	435	408	0.188
117	324	294	0.181
123	340	307	0.181

TABLE 9 (contd.)

A	B	C	D
SEQ ID NO	leu	leu ura trp	transposition frequency
116	345	302	0.175
145	280	239	0.171
126	483	405	0.168
139	377	314	0.167
120	331	272	0.164
138	405	328	0.162
121	436	342	0.157
149	302	230	0.152
104	477	338	0.142
129	366	254	0.139
128	321	211	0.131
130	509	282	0.111
99	312	169	0.108
122	556	301	0.108
112	179	82	0.092
107	271	107	0.079
108	198	59	0.060
98	278	60	0.043
100	318	45	0.028
101	319	34	0.021
102	156	16	0.021
28	179	18	0.020
18	223	2	0.002

TABLE 9

7. REFERENCES

[00156] All references cited herein are incorporated herein by reference in their entirety and for all purposes to the same extent as if each individual publication or patent or patent application was specifically and individually indicated to be incorporated by reference in its entirety for all purposes. To the extent the information associated with a citation may change with time, the version in effect at the effective filing date of this application is meant, the effective filing date being the filing date of the application or priority application in which the citation was first mentioned.

[00157] Many modifications and variations of this invention can be made without departing from its spirit and scope, as will be apparent to those skilled in the art. The specific embodiments described herein are offered by way of example only, and the invention is to be limited only by the terms of the appended claims, along with the full scope of equivalents to which such claims are entitled. Unless otherwise apparent from the context, any embodiment, aspect, element, feature or step can be used in combination with any other.

CLAIMS

What is claimed is

1. A polynucleotide comprising an open reading frame encoding a transposase, the amino acid sequence of which is at least 90% identical to SEQ ID NO: 18, operably linked to a heterologous promoter.
2. The polynucleotide of claim 1, wherein the transposase comprises a mutation as shown in columns C and D in Table 1, relative to the sequence of SEQ ID NO: 18.
3. The polynucleotide of claim 2, wherein the transposase comprises a mutation at an amino acid position selected from 79, 95, 115, 116, 121, 139, 166, 179, 187, 198, 203, 211, 238, 273, 304, 329, 345, 362, 366, 408, 416, 435, 458, 475, 483, 491, 529, 540, 560 and 563, relative to the sequence of SEQ ID NO: 18.
4. The polynucleotide of claim 3, wherein the transposase comprises a mutation selected from D79N, R95S, L115D, E116P, H121Q, K139E, V166F, G179N, W187F, P198R, L203R, N211R, E238D, L273M, L273I, D304R, D304K, Q329G, T345L, K362R, T366R, L408M, S416E, S435G, L458M, V475I, N483K, I491M, A529P, K540R, S560K and S563K, relative to the sequence of SEQ ID NO: 18, the transposase optionally including at least 2, 3, 4, or 5 selected from the group
5. The polynucleotide of claim 2, wherein the amino acid sequence of the transposase is selected from SEQ ID NO: 96-170.
6. The polynucleotide of any preceding claim, wherein the transposase can excise or transpose a transposon from SEQ ID NO: 192.
7. The polynucleotide of claim 6, wherein the excision activity or transposition activity of the transposase is at least 2-fold higher than the activity of SEQ ID NO: 39, optionally 2 to 50-fold higher.
8. The polynucleotide of any preceding claim, wherein the promoter is active in an *in vitro* transcription reaction.
9. The polynucleotide of any one of claims 1-7, wherein the promoter is active in a eukaryotic cell.

10. The polynucleotide of claim 9, wherein the eukaryotic cell is a mammalian cell, optionally, wherein codons of the open reading frame are selected for mammalian cell expression.
11. An isolated mRNA encoding a polypeptide, the amino acid sequence of which is at least 90% identical with SEQ ID NO: 18, and wherein the mRNA sequence comprises at least 10 synonymous codon differences relative to SEQ ID NO: 929 at corresponding positions between the mRNA and SEQ ID NO:781, optionally wherein codons in the mRNA at the corresponding positions are selected for mammalian cell expression.
12. The polynucleotide of any one of claims claim 1-11, wherein the open reading frame further encodes a nuclear localization sequence fused to the transposase.
13. The polynucleotide of any one of claims 1-12, wherein the open reading frame further encodes a heterologous DNA binding domain fused to the transposase.
14. The polynucleotide of claim 13, wherein the DNA binding domain is derived from a Crispr Cas system, or a zinc finger protein, or a TALE protein.
15. A non-naturally occurring polynucleotide encoding a polypeptide, the sequence of which is at least 90% identical to SEQ ID NO: 18, wherein the polynucleotide sequence comprises at least 10 synonymous codon differences relative to SEQ ID NO: 929 at corresponding positions between the polynucleotide and SEQ ID NO:929, optionally wherein codons in the polynucleotide at the corresponding positions are selected for mammalian cell expression.
16. A non-naturally occurring polypeptide encoded by the polynucleotide of any preceding claim.
17. A transposon comprising SEQ ID NO: 9 and SEQ ID NO: 10 flanking a heterologous polynucleotide.
18. The transposon of claim 17, further comprising a sequence at least 90% identical to SEQ ID NO: 13 on one side of the heterologous polynucleotide and a sequence at least 90% identical to SEQ ID NO: 16 on the other.
19. The transposon of claim 17 or 18, wherein the heterologous polynucleotide comprises a heterologous promoter that is active in eukaryotic cells.
20. The transposon of claim 19, wherein the promoter is operably linked to at least one or more of: i) an open reading frame; ii) a nucleic acid encoding a selectable marker; iii) a nucleic acid encoding a counter-selectable marker; iii)

a nucleic acid encoding a regulatory protein; iv) a nucleic acid encoding an inhibitory RNA.

21. The transposon of claim 19, wherein the heterologous promoter comprises a sequence selected from SEQ ID NOs: 474-558.
22. The transposon of any one of claims 17-21, wherein the heterologous polynucleotide comprises a heterologous enhancer that is active in eukaryotic cells.
23. The transposon of claim 22, wherein the heterologous enhancer is selected from SEQ ID NOs: 453-473.
24. The transposon of any one of claims 17-23, wherein the heterologous polynucleotide comprises a heterologous intron that is spliceable in eukaryotic cells.
25. The transposon of claim 24, wherein the nucleotide sequence of the heterologous intron is selected from SEQ ID NO: 561-621.
26. The transposon of any one of claims 17-25, wherein the heterologous polynucleotide comprises an insulator sequence.
27. The transposon of claim 26, wherein the nucleic acid sequence of the insulator is selected from SEQ ID NO: 435-441.
28. The transposon of any one of claims 17-27, wherein the heterologous polynucleotide comprises or encodes a selectable marker.
29. The transposon of claim 28, wherein the selectable marker is selected from a glutamine synthetase enzyme, a dihydrofolate reductase enzyme, a puromycin acetyltransferase enzyme, a blasticidin acetyltransferase enzyme, a hygromycin B phosphotransferase enzyme, an aminoglycoside 3'-phosphotransferase enzyme and a fluorescent protein.
30. A eukaryotic cell whose genome comprises the SEQ ID NO: 9 and SEQ ID NO: 10 flanking a heterologous polynucleotide.
31. The eukaryotic cell of claim 30, wherein the cell is an animal cell.
32. The animal cell of claim 31, wherein the cell is a mammalian cell.
33. The mammalian cell of claim 32, wherein the cell is a rodent cell.
34. The mammalian cell of claim 32, wherein the cell is a human cell.
35. The transposon of any one of claims 17-29, wherein the heterologous polynucleotide comprises two open reading frames, each operably liked to a separate promoter.

36. The transposon of claim 35, wherein the heterologous polynucleotide further comprises a sequence selected from SEQ ID NOs: 745-928.
37. A method of integrating a transposon into a eukaryotic cell, the method comprising
    a. introducing into the cell a transposon comprising SEQ ID NO: 9 and SEQ ID NO: 10 flanking a heterologous polynucleotide
    b. introducing into the cell a transposase, the sequence of which is at least 90% identical with SEQ ID NO: 18 wherein the transposase transposes the transposon to produce a genome comprising SEQ ID NO: 9 and SEQ ID NO: 10 flanking the heterologous polynucleotide.
38. The method of claim 37, wherein the transposase is introduced as a polynucleotide encoding the transposase
39. The method of claim 38, wherein the polynucleotide encoding the transposase is an mRNA molecule.
40. The method of claim 38, wherein the polynucleotide encoding the transposase is a DNA molecule.
41. The method of claim 37, wherein the transposase is introduced as a protein
42. The method of any one of claims 37-41, wherein the heterologous polynucleotide encodes a selectable marker, and the method further comprises
    c. selecting a cell comprising the selectable marker
43. The method of claim 37, wherein the cell is an animal cell.
44. The animal cell of claim 43, wherein the cell is a mammalian cell.
45. The mammalian cell of claim 44, wherein the cell is a rodent cell.
46. The mammalian cell of claim 44, wherein the cell is a human cell.
47. A method of expressing a polypeptide comprising culturing a eukaryotic cell having a genome comprising SEQ ID NO: 9 and SEQ ID NO: 10 flanking a heterologous polynucleotide, wherein the polynucleotide is expressed.
48. The method of claim 47, further comprising purifying the polypeptide from culture media.
49. The method of claim 47 or 48, further comprising incorporating the purified polypeptide into a pharmaceutical composition.