COMPOSITIONS AND METHODS FOR NON-TARGETED ACTIVATION OF ENDOGENOUS GENES

Inventors: John Joseph Harrington, Mentor, OH (US); Bruce A. Sherf, Spencer, OH (US); Stephen E. Rundlett, Coon Rapids, OH (US)

Correspondence Address:
LAWIVE & COCKFIELD, LLP
ONE POST OFFICE SQUARE
BOSTON, MA 02109-2127 (US)

Assignee: ATHERSYS, INC., Cleveland, OH

Filed: Sep. 6, 2005

Related U.S. Application Data

Continuation of application No. 09/513,997, filed on Feb. 26, 2000, now abandoned, which is a division of application No. 09/276,820, filed on Mar. 26, 1999, now Pat. No. 6,897,066, which is a continuation-in-part of application No. 09/263,814, filed on Mar. 8, 1999, now abandoned, which is a continuation-in-part of application No. 09/253,022, now abandoned, which is a continuation-in-part of application No. 09/159,643, filed on Sep. 24, 1998, now abandoned, which is a continuation-in-part of application No. 08/941,223, filed on Sep. 26, 1997, now abandoned.

Publication Classification

Publication Date: May 31, 2007

The present invention is directed generally to activating gene expression or causing over-expression of a gene by recombination methods in situ. The invention also is directed generally to methods for expressing an endogenous gene in a cell at levels higher than those normally found in the cell. In one embodiment of the invention, expression of an endogenous gene is activated or increased following integration into the cell, by non-homologous or illegitimate recombination, of a regulatory sequence that activates expression of the gene. The invention also provides methods for the identification, activation, isolation, and/or expression of genes undiscoverable by current methods since no target sequence is necessary for integration. Thus, by the present invention, endogenous genes, including those associated with human disease and development, may be activated and isolated without prior knowledge of the sequence, structure, function, or expression profile of the genes.
Random Activation of Gene Expression (RAGE)

FIG. 1.
FIG. 5A.
FIG. 5B.
FIG. 5C.
AGAGCGCAGGGAGCTCCAGGGGAAACGCCTGGTATCTTTATAGTCCTGTC
GGTTTCGCCACCTCTGACTTGGAGCGTCGATTTTTTGTGATGCTCGTCAGGG
GGCGGAGCCTATGGAAAAACGCCAGCAACCGCGCTTTTTACGGTTCCTGGCCTT
TGCTGGCCTTTTGCTCACATGGCT
CGAC3'

FIG. 5D.
FIG. 6A.
GAAAATTGTGATGCTATTTTGCTTTTATTTTGTAACCTTTTATAAGCTGCAATAAAA
CAAGTTGAAAAGAAACAAATCCAGGATCTTATTTTTATGTTTTCAGGTCAGGGGAGATGT
GGGAGGTTTTTAAAGCAGATAAAA
CCCTCTCAAAATGTGTAATATCCGCTATAAGGATCAGTTCCCGAGCCTGAATGGCGGA
ATGGGACCGCCCGCTGAGGCTGCTAGT
TAAGGCGGCGGCTGGTGTGCTTACCGCCGCTAGCGGTCTACACTTGCCAGGGC
CCTAGCAGGCGCTCTTCCGGTCTCTCT
TCCCTCTCTCTCTGCAAGGGGCTTCCCCGCTCAAGGCTTAAATATCGAGCTGTTGTCG
GCCCACTCGCCTGATAGCGGGTCTT
TCCGCTCTTTTGAATGGTGATCTCCGTTTTGTTTTCGGGATTTGCTGGGTTATTTT
GACTGTATTTAACAAAATTTAAAC
CGCAGAATTTTACAAATATATTAAGCTTAAATTTCGGCTGTGATACCTCTGTGAGGC
GGAAAGAACCAGCTGGGAAATGTTG
GTCAGTATTAGGGTGGAGAAGTTCCACGGTCTCCCATCGAGGCAGAAAGATATGCAAA
GCACATGGCTGAATTACATGGACCC
AGCTGTTGGAAGATCGCCCGCAGCTCCTCCGAGCGACAGACTGCAAGCATGCAT
CTCAATTAGTGCAACACATAGTCCC
GCCCTCAACTCGCGCATCCCAGCGCTAATCCGGCAGTTCCGGGACATTCTCG
CCCATGGGCTGACTATTTTTTTA
TTATAGGCAGGGCGAGCCGCCTCGGCTGCTGATATCAGGAAAGTAGTGAGG
AGGCTTTTTGAGGGCTAGGCTTTT
TGCAAAAAAGTTGATTCTTCTTGACACAACAGTCGAACTTAAGGCTAGACACAC
CATGATTAGAAAGATGATGGCAC
GCAGGTCTCTCCGGCCGCTTTGGAGAAGGGCTATTCCGCTATGACTGGGACAAC
AGAACATCTCGCTGCTCTGATGCCC
CGTGTTCCGGCTGTCCAGGCAAGGGGCACCGCGCCTTCTTTTGACTAGCGCCACCTG
TCCGGTCGGCTGATAGTAACTGCAAG
ACGAGGCGACGGGCAGCTATCGGGTGGCAGCAGCCGCGTCTTCGCGCAGCTG
TGCTGCAAGTTGCAGCTGCTGACGGGA
AGGGACTGCGTCGCTATTTGGGCGAAGTCCGGGCGCGGCAAGATCTCCTGTCATCTACC
TTGCTTCTCGCGGAAGAAATCTCCTAT
CATGCGTCTAGCAATCGGCCGCTGCTGATACGCTTGATCCGGCTACTCGGCCACAT
GACCCCAAGGCAAAACATCGATCG
AGCGAGACGTACTCGGTGAGAAGCCGCTTCTTTCGATCAAGGATGATCCTGGACG
AAGACATCAGGCTGCTCGGCCAGC
GAACTGTTCCGAGGCTCAAGGCCGCCGATGCGCCAGCGCCAGGAGATTCTCGTCG
ACCCTTAGCCTGGCTGTCTGGCGAAA
TTACATGCGTGGAAATGCGGCGCTTTTCGGATTTCTCGACTGTTGGCAGGGCTGTTG
GGCTGACGCTATCCAGACATAGCGGTCGGATCTCGGTGATATTGCGAGACG
TTGCCGCGGAAATGGCCGAGCGCTCTTCTCTGCTGTTTAAGCTTGATCGCGC
CCGGATTTCGACGCGCTCTCTCAGGGTTCTTTCAGAGGTCTCTTGAGGCCG
GACTCTGGGGTTCAGAAATTACGC
CAAGCCAGCCCAACTGCCAACACTGATGGGCGCAATAAAAATATCTATTTTTCA
TTACTCTGTGGTTTTTTTTG
GTGAAGAGTCGGCTATAGGGTTGACCTCAGAAATCTGGCTGTGATGGCGGATAGT
TAAGGCAAAGCCCAGACCCCGCAAAACTCGGTGACGCGCCCTGAGGGCTATAGT

FIG. 6B.
FIG. 6C.
Fig. 7A.
FIG. 7B.
AGTACAATCTGCTCTGTAGCCTGGCGATAGTTAAGCCAGCCCGCGACACCACCCCGCAA
CAGCCCGCAGCGCGCCCGTCGAGGCGGGCTTTGTCTCCTCGCCGCTCCTACAGACA
ACTGCTGAGGCCTCTCCGGAGGCTG
AGTGTTCAGAGGTGTTTCCACCCTCATCACCCGAAACGCGGAGACGAAAGGGCGCTG
TGATAAGCTCTATTTTATATGGAT
TGTCATGATAAATAATGGTTCTTACACGCTCAGGGCTTTCTTCCGGGAAATGTG
CGCGGAACCCCTATTGTATTTTT
TCTAATACATTTCAATATGATCCGTCTCATGAGACAAATAAACCCCTGATAAAATGT
TCAATAATATTGAAAAGAAAGAT
ATGAGATATTCAACATTTCGGCTGTCGCCCTTATTCCCTTTTTGCGGCATTTTGGCT
TCTCTTTTGGTCTCACCCAGAGCA
GCTGTGGAAGAAGAAGATGCTGAGACGACGTTGGGTGTCGAGAGTTGGTATC
CGAAGTGGATCTCAAGCCGGTAAGA
TCTTTAGAGTTTTCGCCCGCGAAACGTTTTTCAATGAGTGAACACTTTTAAGAT
TCTGGCTATGTTGGCGCGGTATTATAC
CGTATTGGACGGCGCGCGGAAAGCAGCGAATCCTGGCGCGCGCTACATCATTTCATGAA
ACTGGTGGTACGTCACTCACGCTAC
AGAAAGGACTTTACCTAGGAACTGGAAGATGTAAGAGAATTAGCAGTGGTCATCGATA
ACCATGAGTGTATAAACACTGCAGGCA
ACTTACTCTGTAACAGATCGGGAGGACCGGCAGCTAAACCCGGTTTGGCACAAG
CATGGGGAGATCGTCACTCGCTT
GATCGTGTGGGAAACCGGAGCTGAATGAAAGCTCACAAACGACGAGCGTGACACC
ACGAATGCGTGGTACCAACGGCGAAC
GTTCGCAGAAAATCATTAAACTTCGCGCAACTTTACTTCTACTTCTGCTTCCGGCCGAAACATTA
ATAGACTGTTAGGGACGGGATAAAAG
TTGCAGGACCACTTCTGGCGCTCGGGCCTTCCGGCTGCTGGTATTATGCTGAAAA
ACTGGAGCCGGCTGGTTGCTCCT
CGCGGTATCTTGACGACTGCGGCGAGATGGTAGACGGCCCTCCTCGTACTCGTAGTTA
TCTACAGCGACGGGGAGTGCAAGGCAC
TATGATGAAACGAAATAGACAGATCGCTGAGATAGTTGGCTCTACTCGATTAGCAT
TGGATACCTGACGACCAAGTTATTAC
CATATACTACCTTTAAGATTTTAAAACCTACATTTTTAATTTAAAGGATCTAGGTTG
AAGAATCTCCTTTTTGATAAATCTATG
ACCACAAATCTCTTAAAGTGGATTCTTTCGTTTCCACTGAGCTGACGCTGAGGAGAA
AGATCAAAGGATCTCTTCTGAGATCC
TTTTTTCTCGCGGTAATCTGCTGTCTGGGCAACAAACAAAAACACCAGCTACCAGC
GTTGGTGGTTTGGCGGGAATTGTGAGATCTCCAGACTGGGGTTTACACCGGATTAC
TACCCCATCTCTTTTCCGAAGGTAATCGGCTTCAGACAGCGCAAGATAACAATAC
TGCTTTCTTAGGCTAGCGTATG
GGGCCACACTTCCAGGACACTCCTGTAACCGGCTTACATACTCGCTCCTGCAATCC
GTATACAGTGGCGCTGGCAGTGG
CGATAGTTGTGTCTTACCCGGGTGAGACTCAAGAGTGGTTACCAGGATAAGGCG
CAGCCGGTTGGCGGTAAACGGGGGGTT
CGTGCACACAGCCACGCTGAGGCGAAGACGACTCACCCGAGACTGAGACCTAC
AGCCTGATCTAAGGACCGCAGCTTTCCGGAAGGAAAGGCGAGCAGT
ATCCGGTAGAGCCGGCGATTGGTGAAGAAGGAGACGCGAAGGAGCTCCGGAG
GGGAAGACCGCTGATCTTTTATATGTCCTGCTGGGTTTCCGACCTCTCGAAGT
CGTCGATTGGTGTATGCGTCAG
GGGGGGGAGCCTATCGAAAAACGCAGAGACCGGCGGCTTTTTTACGTTTCTG
CCTTTTCGCTTGGCTTCTGACATGGCTGCAC3'

FIG. 7C.
FIG. 12A.

FIG. 12B.

FIG. 12C.

FIG. 12D.

FIG. 12E.

FIG. 12F.

FIG. 12G.
FIG. 13.
FIG. 14B.
FIG. 15A
FIG. 15B.
FIG. 16A.
FIG. 16B.
FIG. 17A.

FIG. 17B.

FIG. 17C.

FIG. 17D.

FIG. 17E.

FIG. 17F.

FIG. 17G.
FIG. 19.
FIG. 23A
5' UTR: ACCAGGTTGAG
Exon A and Flanking Intron

FIG. 23B
5' UTR: ACCATGCCAGTGC
Exon B and Flanking Intron

FIG. 23C
5' UTR: ACCATGCGCAAGTTAG
Exon C and Flanking Intron

FIG. 23D
5' UTR: ACCATGGCGCAAGTTGAG
Exon D and Flanking Intron
FIG. 25A.

DNA Binding Domain  S/D  Neo

FIG. 25B.

DNA Binding Domain  S/D  Neo

FIG. 25C.

Activation Domain  S/D  Neo

FIG. 25D.
FIG. 26.

Gene A

Gene B

Transcription and Splicing

Translation

Detection of Protein Interaction

Transcription

Reporter Gene
FIG. 27A.

FIG. 27B.

FIG. 27C.

FIG. 27D.

FIG. 27E.
FIG. 29A.
FIG. 29B.
FIG. 30B.
FIG. 30C
FIG. 31B.
FIG. 31C.
FIG. 32A.
FIG. 32B
FIG. 32C.
FIG. 33A.
FIG. 33C.
FIG. 33D.
FIG. 34A.
FIG. 34B.
FIG. 35A.
FIG. 35B.
FIG. 36
FIG. 37A.
FIG. 37B.
FIG. 37C.
COMPOSITIONS AND METHODS FOR NON-TARGETED ACTIVATION OF ENDOGENOUS GENES

CROSS REFERENCE TO RELATED APPLICATIONS


BACKGROUND OF THE INVENTION

[0002] 1. Field of the Invention

[0003] The present invention is in the fields of molecular biology and cellular biology. The invention is directed generally to activation of gene expression or causing over-expression of a gene by recombination methods in situ. More specifically, the invention is directed to activation of endogenous genes by non-targeted integration of specialized activation vectors, which are provided by the invention, into the genome of a host cell. The invention also is directed to methods for the identification, activation, and isolation of genes that were heretofore undiscoverable, and to host cells and vectors comprising such isolated genes. The invention also is directed to isolated genes, gene products, nucleic acid molecules, and compositions comprising such genes, gene products and nucleic acid molecules, that may be used in a variety of therapeutic and diagnostic applications. Thus, by the present invention, endogenous genes, including those associated with human disease and development, may be identified, activated, and isolated without prior knowledge of the sequence, structure, function, or expression profile of the genes.

[0004] 2. Related Art

[0005] Identification and over-expression of novel genes associated with human disease is an important step towards developing new therapeutic drugs. Current approaches to creating libraries of cells for protein over-expression are based on the production and cloning of cDNA. Thus, in order to identify a new gene using this approach, the gene must be expressed in the cells that were used to make the library. The gene also must be expressed at sufficient levels to be adequately represented in the library. This is problematic because many genes are expressed only in very low quantities, in a rare population of cells, or during short developmental periods.

[0006] Furthermore, because of the large size of some mRNAs, it is difficult or impossible to produce full length cDNA molecules capable of expressing the biologically active protein. Lack of full-length cDNA molecules has also been observed for small mRNAs and is thought to be related to sequences in the message that are difficult to produce by reverse transcription or that are unstable during propagation in bacteria. As a result, even the most complete cDNA libraries express only a fraction of the entire set of possible genes.

[0007] Finally, many cDNA libraries are produced in bacterial vectors. Use of these vectors to express biologically active mammalian proteins is severely limited since most mammalian proteins do not fold correctly and/or are improperly glycosylated in bacteria.

[0008] Therefore, a method for creating a more representative library for protein expression, capable of facilitating faithful expression of biologically active proteins, would be extremely valuable.

[0009] Current methods for over-expressing proteins involve cloning the gene of interest and placing it in a construct, next to a suitable promoter/enhancer, polyadenylation signal, and splice site, and introducing the construct into an appropriate host cell.

[0010] An alternative approach involves the use of homologous recombination to activate gene expression by targeting a strong promoter or other regulatory sequence to a previously identified gene.

[0011] WO 90/14092 describes in situ modification of genes, in mammalian cells, encoding proteins of interest. This application describes single-stranded oligonucleotides for site-directed modification of genes encoding proteins of interest. A marker may also be included. However, the methods are limited to providing an oligonucleotide sequence substantially homologous to a target site. Thus, the method requires knowledge of the site required for activation by site-directed modification and homologous recombination. Novel genes are not discoverable by such methods.

[0012] WO 91/06667 describes methods for expressing a mammalian gene in situ. With this method, an amplifiable gene is introduced next to a target gene by homologous recombination. When the cell is then grown in the appropriate medium, both the amplifiable gene and the target gene are amplified and there is enhanced expression of the target gene. As above, methods of introducing the amplifiable gene are limited to homologous recombination, and are not useful for activating novel genes whose sequence (or existence) is unknown.

[0013] WO 91/01410 describes the inactivation of endogenous genes by modification of cells by homologous recombination. By these methods, homologous recombination is used to modify and inactivate genes and to produce cells which can serve as donors in gene therapy.

[0014] WO 92/20808 describes methods for modifying genomic target sites in situ. The modifications are described as being small, for example, changing single bases in DNA. The method relies upon genomic modification using homologous DNA for targeting.

[0015] WO 92/19255 describes a method for enhancing the expression of a target gene, achieved by homologous recombination in which a DNA sequence is integrated into the genome or large genomic fragment. This modified
sequence can then be transferred to a secondary host for expression. An amplifiable gene can be integrated next to the target gene so that the target region can be amplified for enhanced expression. Homologous recombination is necessary to this targeted approach.

WO 93/09222 describes methods of making proteins by activating an endogenous gene encoding a desired product. A regulatory region is targeted by homologous recombination and replacing or disabling the region normally associated with the gene whose expression is desired. This disabling or replacement causes the gene to be expressed at levels higher than normal.

WO 94/12650 describes a method for activating expression of and amplifying an endogenous gene in situ in a cell, which gene is not expressed or is not expressed at desired levels in the cell. The cell is transfected with exogenous DNA sequences which repair, alter, delete, or replace a sequence present in the cell or which are regulatory sequences not normally functionally linked to the endogenous gene in the cell. In order to do this, DNA sequences homologous to genomic DNA sequences at a preselected site are used to target the endogenous gene. In addition, amplifiable DNA encoding a selectable marker can be included. By culturing the homologously recombinant cells under conditions that select for amplification, both the endogenous gene and the amplifiable marker are co-amplified and expression of the gene increased.

WO 95/31560 describes DNA constructs for homologous recombination. The constructs include a targeting sequence, a regulatory sequence, an exon, and an unpaired splice donor site. The targeting is achieved by homologous recombination of the construct with genomic sequences in the cell and allows the production of a protein in vitro or in vivo.

WO 96/29411 describes methods using an exogenous regulatory sequence, an exogenous exon, either coding or non-coding, and a splice donor site introduced into a preselected site in the genome by homologous recombination. In this application, the introduced DNA is positioned so that the transcripts under control of the exogenous regulatory region include both the exogenous exon and endogenous exons present in either the thrombopoietin, DNase I, or β-interferon genes, resulting in transcripts in which the exogenous and endogenous exons are operably linked. The novel transcription units are produced by homologous recombination.

WO 97/272,071 describes the transcriptional activation of transcriptionally silent genes in a cell by inserting a DNA regulatory element capable of promoting the expression of a gene normally expressed in that cell. The regulatory element is inserted so that it is operably linked to the normally silent gene. The insertion is accomplished by means of homologous recombination by creating a DNA construct with a segment of the normally silent gene (the target DNA) and the DNA regulatory element used to induce the desired transcription.

WO 97/272,071 describes the transcriptional activation of transcriptionally silent genes in a cell by inserting a DNA regulatory element capable of promoting the expression of a gene normally expressed in that cell. The regulatory element is inserted so that it is operably linked to the normally silent gene. The insertion is accomplished by means of homologous recombination by creating a DNA construct with a segment of the normally silent gene (the target DNA) and the DNA regulatory element used to induce the desired transcription.

US 5,784,611 discloses activating expression of mammalian target genes by homologous recombination. A DNA sequence is integrated into the genome or a large genomic fragment to enhance the expression of the target gene. The modified construct can then be transferred to a secondary host. An amplifiable gene can be integrated adjacent to the target gene so that the target region is amplified for enhanced expression.

[0022] Both of the above approaches (construction of an over-expressing construct by cloning or by homologous recombination in vivo) require the gene to be cloned and sequenced before it can be over-expressed. Furthermore, using homologous recombination, the genomic sequence and structure must also be known.

[0023] Unfortunately, many genes have not yet been identified and/or sequenced. Thus, a method for over-expressing a gene of interest, whether or not it has been previously cloned, and whether or not its sequence and structure are known, would be useful.

BRIEF SUMMARY OF THE INVENTION

The invention is, therefore, generally directed to methods for over-expressing an endogenous gene in a cell, comprising introducing a vector containing a transcriptional regulatory sequence into the cell, allowing the vector to integrate into the genome of the cell by non-homologous recombination, and allowing over-expression of the endogenous gene in the cell. The method does not require previous knowledge of the sequence of the endogenous gene or even of the existence of the gene. Hence, the invention is directed to non-targeted gene activation, which as used herein means the activation of endogenous genes by non-targeted or non-homologous (as opposed to targeted or homologous) integration of specialized activation vectors into the genome of a host cell.

The invention also encompasses novel vector constructs for activating gene expression or over-expressing a gene through non-homologous recombination. The novel construct lacks homologous targeting sequences. That is, it does not contain nucleotide sequences that target host cell DNA and promote homologous recombination at the target site, causing over-expressing of a cellular gene via the introduced transcriptional regulatory sequence.

[0026] Novel vector constructs include a vector containing a transcriptional regulatory sequence operably linked to an unpaired splice donor sequence and further contains one or more amplifiable markers.

[0027] Novel vector constructs include constructs with a transcriptional regulatory sequence operably linked to a translational start codon, a signal sequence, and an unpaired splice donor site; constructs with a transcriptional regulatory sequence, operably linked to a translation start codon, an epitope tag, and an unpaired splice donor site; constructs containing a transcriptional regulatory sequence operably linked to a translational start codon, a signal sequence and an epitope tag, and an unpaired splice donor site; constructs containing a transcriptional regulatory sequence operably linked to a translation start codon, a signal sequence, an epitope tag, and a sequence-specific protease site, and an unpaired splice donor site.

[0028] The vector construct can contain one or more selectable markers for recombinant host cell selection. Alternatively, selection can be effected by phenotypic selection for a trait provided by the activated endogenous gene product.
These vectors, and indeed any of the vectors disclosed herein, and variants of the vectors that will be readily recognized by one of ordinary skill in the art, can be used in any of the methods described herein to form any of the compositions producible by these methods.

The transcriptional regulatory sequence used in the vector constructs of the invention includes, but is not limited to, a promoter. In preferred embodiments, the promoter is a viral promoter. In highly preferred embodiments, the viral promoter is the cytomegalovirus immediate early promoter. In alternative embodiments, the promoter is a cellular, non-viral promoter or inducible promoter.

The transcriptional regulatory sequence used in the vector construct of the invention may also include, but is not limited to, an enhancer. In preferred embodiments, the enhancer is a viral enhancer. In highly preferred embodiments, the viral enhancer is the cytomegalovirus immediate early enhancer. In alternative embodiments, the enhancer is a cellular non-viral enhancer.

In preferred embodiments of the methods described herein, the vector construct be, or may contain, linear RNA or DNA.

The cell containing the vector may be screened for expression of the gene.

The cell over-expressing the gene can be cultured in vitro under conditions favoring the production, by the cell, of desired amounts of the gene product (also referred to interchangeably herein as the "expression product") of the endogenous gene that has been activated or whose expression has been increased. The expression product can then be isolated and purified to use, for example, in protein therapy or drug discovery.

Alternatively, the cell expressing the desired gene product can be allowed to express the gene product in vivo. In certain such aspects of the invention, the cell containing a vector construct of the invention integrated into its genome may be introduced into a eukaryote (such as a vertebrate, particularly a mammal, more particularly a human) under conditions favoring the overexpression or activation of the gene by the cell in vivo in the eukaryote. In related such aspects of the invention, the cell may be isolated and cloned prior to being introduced into the eukaryote.

The invention is also directed to methods for over-expressing an endogenous gene in a cell, comprising introducing a vector containing a transcriptional regulatory sequence and one or more amplifiable markers into the cell, allowing the vector to integrate into the genome of the cell by non-homologous recombination, and allowing over-expression of the endogenous gene in the cell.

The cell containing the vector may be screened for over-expression of the gene.

The cell over-expressing the gene is cultured such that amplification of the endogenous gene is obtained. The cell can then be cultured in vitro so as to produce desired amounts of the gene product of the amplified endogenous gene that has been activated or whose expression has been increased. The gene product can then be isolated and purified.

Alternatively, following amplification, the cell can be allowed to express the endogenous gene and produce desired amounts of the gene product in vivo.

It is to be understood, however, that any vector used in the methods described herein can include one or more amplifiable markers. Thereby, amplification of both the vector and the DNA of interest (i.e., containing the over-expressed gene) occurs in the cell, and further enhanced expression of the endogenous gene is obtained. Accordingly, methods can include a step in which the endogenous gene is amplified.

The invention is also directed to methods for over-expressing an endogenous gene in a cell comprising introducing a vector containing a transcriptional regulatory sequence and an unpaired splice donor sequence into the cell, allowing the vector to integrate into the genome of the cell by non-homologous recombination, and allowing over-expression of the endogenous gene in the cell.

The cell containing the vector may be screened for expression of the gene.

The cell over-expressing the gene can be cultured in vitro so as to produce desirable amounts of the gene product of the endogenous gene whose expression has been activated or increased. The gene product can then be isolated and purified.

Alternatively, the cell can be allowed to express the desired gene product in vivo.

The vector construct can consist essentially of the transcriptional regulatory sequence.

The vector construct can consist essentially of the transcriptional regulatory sequence and one or more amplifiable markers.

The vector construct can consist essentially of the transcriptional regulatory sequence and the splice donor sequence.

Any of the vector constructs of the invention can also include a secretion signal sequence. The secretion signal sequence is arranged in the construct so that it will be operably linked to the activated endogenous protein. Thereby, secretion of the protein of interest occurs in the cell, and purification of that protein is facilitated. Accordingly, methods can include a step in which the protein expression product is secreted from the cell.

The invention also encompasses cells made by any of the above methods. The invention encompasses cells containing the vector constructs, cells in which the vector constructs have integrated into the cellular genome, and cells which are over-expressing desired gene products from an endogenous gene, over-expression being driven by the introduced transcriptional regulatory sequence.

The cells can be isolated and cloned.

The methods can be carried out in any cell of eukaryotic origin, such as fungal, plant or animal. In preferred embodiments, the methods of the invention may be carried out in vertebrate cells, and particularly mammalian cells including but not limited to rat, mouse, bovine, porcine, sheep, goat and human cells, and more particularly in human cells.

A single cell made by the methods described above can over-express a single gene or more than one gene. More than one gene in a cell can be activated by the integration of
a single type of construct into multiple locations in the genome. Similarly, more than one gene in a cell can be activated by the integration of multiple constructs (i.e., more than one type of construct) into multiple locations in the genome. Therefore, a cell can contain only one type of vector construct or different types of constructs, each capable of activating an endogenous gene.

[0053] The invention is also directed to methods for making the cells described above by one or more of the following: introducing one or more of the vector constructs of the invention into a cell; allowing the introduced construct(s) to integrate into the genome of the cell by non-homologous recombination; allowing over-expression of one or more endogenous genes in the cell; and isolating and cloning the cell. The invention is also directed to cells produced by such methods, which may be isolated cells.

[0054] The invention also encompasses methods for using the cells described above to over-express a gene, such as an endogenous cellular gene, that has been characterized (for example, sequenced), uncharacterized (for example, a gene whose function is known but which has not been cloned or sequenced), or a gene whose existence was, prior to over-expression, unknown. The cells can be used to produce desired amounts of an expression product in vitro or in vivo. If desired, this expression product can then be isolated and purified, for example by cell lysis or by isolation from the growth medium (as when the vector contains a secretion signal sequence).

[0055] The invention also encompasses libraries of cells made by the above described methods. A library can encompass all of the clones from a single transfection experiment or a subset of clones from a single transfection experiment. The subset can over-express the same gene or more than one gene, for example, a class of genes. The transfection can have been done with a single construct or with more than one construct.

[0056] A library can also be formed by combining all of the recombinant cells from two or more transfection experiments, by combining one or more subsets of cells from a single transfection experiment or by combining subsets of cells from separate transfection experiments. The resulting library can express the same gene, or more than one gene, for example, a class of genes. Again, in each of these individual transfections, a unique construct or more than one construct can be used.

[0057] Libraries can be formed from the same cell type or different cell types.

[0058] The invention is also directed to methods for making libraries by selecting various subsets of cells from the same or different transfection experiments.

[0059] The invention is also directed to methods of using the above-described cells or libraries of cells to over-express or activate endogenous genes, or to obtain the gene expression products of such over-expressed or activated genes. According to this aspect of the invention, the cell or library may be screened for the expression of the gene and cells that express the desired gene product may be selected. The cell can then be used to isolate or purify the gene product for subsequent use. Expression in the cell can occur by culturing the cell in vitro, under conditions favoring the production of the expression product of the endogenous gene by the cell, or by allowing the cell to express the gene in vivo.

[0060] In preferred embodiments of the invention, the methods include a process wherein the expression product is isolated or purified. In highly preferred embodiments, the cells expressing the endogenous gene product are cultured under conditions favoring production of sufficient amounts of gene product for commercial application, and especially for diagnostic, therapeutic and drug discovery uses.

[0061] Any of the methods can further comprise introducing double-strand breaks into the genomic DNA in the cell prior to or simultaneously with vector integration.

[0062] The invention also is directed to vector constructs that are useful for activating expression of endogenous genes and for isolating the mRNA and cDNA corresponding to the activated genes.

[0063] In one such embodiment, the vector construct may comprise (a) a first transcriptional regulatory sequence operably linked to a first unpaired splice donor sequence; (b) a second transcriptional regulatory sequence operably linked to a second unpaired splice donor sequence; and (c) a linearization site, which may be located between the first and second transcripational regulatory sequences. According to the invention, when the vector construct is transformed into a host cell and then integrates into the genome of the host cell, the first transcriptional regulatory sequence is preferably in an inverted orientation relative to the orientation of the second transcriptional regulatory sequence. In certain preferred such embodiments, the vector may be rendered linear by cleavage at the linearization site.

[0064] In another embodiment, the invention provides a linear vector construct having a 3' end and a 5' end, comprising a transcriptional regulatory sequence operably linked to an unpaired spliced donor site, wherein the transcriptional regulatory sequence is oriented in the linear vector construct in an orientation that directs transcription towards the 3' end or the 5' end of the linear vector construct.

[0065] In another embodiment, the invention provides a vector construct comprising, in sequential order, (a) a transcriptional regulatory sequence, (b) an unpaired splice donor site, (c) a rare cutting restriction site, and (d) a linearization site.

[0066] In another embodiment, the invention provides a vector construct comprising (a) a first transcriptional regulatory sequence operably linked to a selectable marker lacking a polyadenylation signal; and (b) a second transcriptional regulatory sequence operably linked to an exon-splice donor site complex, wherein the first transcriptional regulatory sequence is in the same orientation in the vector construct as is the second transcriptional regulatory sequence, and wherein the first transcriptional regulatory sequence is upstream of the second transcriptional regulatory sequence in the vector construct.

[0067] In additional embodiments, the invention provides vector constructs comprising a transcriptional regulatory sequence operably linked to a selectable marker lacking a polyadenylation signal, and further comprising an unpaired splice donor site.

[0068] In another embodiment, the invention provides vector constructs comprising a first transcriptional regula-
ory sequence operably linked to a selectable marker lacking a polyadenylation signal, and further comprising a second transcriptional regulatory sequence operably linked to an unpaired splice donor site.

[0069] According to the invention, the transcriptional regulatory sequence (or first or second transcriptional regulatory sequence, in vector constructs having more than one transcriptional regulatory sequence) may be a promoter, an enhancer, or a repressor, and is preferably a promoter, including an animal cell promoter, a plant cell promoter, or a fungal cell promoter, most preferably a promoter selected from the group consisting of a CMV immediate early gene promoter, an SV40 T antigen promoter, and a β-actin promoter. Other promoters of animal, plant, or fungal cell origin that may be used in accordance with the invention are known in the art and will be familiar to one of ordinary skill in view of the teachings herein.

[0070] The selectable marker used in the vector constructs of the invention may be any marker or marker gene that, upon integration of a vector containing the selectable marker into the host cell genome, permits the selection of a cell containing or expressing the marker gene. Suitable such selectable markers include, but are not limited to, a neomycin gene, a hygromycin phosphoribosyl transferase gene, a puromycin gene, a dihydroorotase gene, a glutamine synthetase gene, a histidine D gene, a carbamyl phosphate synthase gene, a dihydrofolate reductase gene, a multidrug resistance 1 gene, an aspartate transcarbamylase gene, a xanthine-guanine phosphoribosyl transferase gene, an adenosine deaminase gene, and a thymidine kinase gene.

[0071] In related embodiments, the invention provides vector constructs comprising a positive selectable marker, a negative selectable marker, and an unpaired splice donor site, wherein the positive and negative selectable markers and the splice donor site are oriented in the vector construct in an orientation that results in expression of the positive selectable marker in active form, and either non-expression of said negative selectable marker or expression of the negative selectable marker in inactive form, when the vector construct is integrated into the genome of an eukaryotic host cell and activates an endogenous gene in the genome. In certain preferred such embodiments, either the positive selection marker, the negative selection marker, or both, may lack a polyadenylation signal. The positive selection marker used in these aspects of the invention may be any selection marker that, upon expression, produces a protein capable of facilitating the isolation of cells expressing the marker, including but not limited to a neomycin gene, a hygromycin phosphoribosyl transferase gene, a puromycin gene, a dihydroorotase gene, a glutamine synthetase gene, a histidine D gene, a carbamyl phosphate synthase gene, a dihydrofolate reductase gene, a multidrug resistance 1 gene, an aspartate transcarbamylase gene, a xanthine-guanine phosphoribosyl transferase gene, or an adenosine deaminase gene. Analogously, the negative selection marker used in these aspects of the invention may be any selection marker that, upon expression, produces a protein capable of facilitating removal of cells expressing the marker, including but not limited to a hygromycin phosphoribosyl transferase gene, a thymidine kinase gene, or a diphtheria toxin gene.

[0072] The invention also is directed to eukaryotic host cells, which may be isolated host cells, comprising one or more of the vector constructs of the invention. Preferred such eukaryotic host cells include, but are not limited to, animal cells (including, but not limited to, mammalian (particularly human) cells, insect cells, avian cells, amphibian cells, reptilian cells, and fish cells), plant cells, and fungal (particularly yeast) cells. In certain such host cells, the vector construct may be integrated into the genome of the host cell.

[0073] The invention also is directed to primer molecules comprising a PCR-amplifiable sequence and a degenerate 3' terminus. Primer molecules according to this aspect of the invention preferably have the general structure:

$$5'-(\text{AT})_X-N_N-TTTATTT-3'$$

wherein $X$ is a whole number from 1 to 100 (preferably from 10 to 30), $X$ is a PCR-amplifiable sequence consisting of a nucleic acid sequence of about 10-20 nucleotides in length, $N$ is any nucleotide, and $b$ is a whole number from 0 to 6. One preferred such primer has the nucleotide sequence 5'-TTTTTTTTTTTGTCAGCGCCG-CATCINNNNTTATT-3' (SEQ ID NO:10). In related embodiments, the primer molecules according to this aspect of the invention may be biotinylated.

[0074] The invention also is directed to methods for first strand cDNA synthesis comprising (a) annealing a first primer of the invention (such as the primer described above) to an RNA template molecule to form an first primer-RNA complex, and (b) treating this first primer-RNA complex with reverse transcriptase and one or more deoxynucleoside triphosphate molecules under conditions favoring the reverse transcription of the first primer-RNA complex to synthesize a first strand cDNA.

[0075] The invention also is directed to methods for isolating activated genes, particularly from a host cell genome. These methods of the invention exploit the structure of the mRNA molecules produced using the non-targeted gene activation vectors of the invention. One such method of the invention comprises, for example, (a) introducing a vector construct comprising a transcriptional regulatory sequence and an unpaired splice donor site into a host cell (preferably one of the eukaryotic host cells described above), (b) allowing the vector construct to integrate into the genome of the host cell by non-homologous recombination, under conditions such that the vector activates an endogenous gene comprising an exon in the genome, (c) isolating RNA from the host cell, (d) synthesizing first strand cDNA according to the method of the invention described above, (e) annealing a second primer specific for the vector-encoded exon to the first strand cDNA to create a second primer-first strand cDNA complex, and (f) contacting the second primer-first strand cDNA complex with a DNA polymerase under conditions favoring the production of a second strand cDNA substantially complementary to the first strand cDNA. Methods according to this aspect of the invention may comprise one or more additional steps, such as treating the second strand cDNA with a restriction enzyme that cleaves at a restriction site located on the vector downstream of the unpaired splice donor site, or amplifying the second strand cDNA using a third primer specific for the vector-encoded exon and a fourth primer specific for the second primer. The invention also is directed to isolated genes produced according to these methods, and to vectors (which may be expression vectors) and host cells comprising
these isolated genes. The invention also is directed to methods of producing a polypeptide, comprising cultivating a host cell comprising the isolated gene (or a vector, particularly an expression vector, comprising the isolated gene), and culturing the host cell under conditions favoring the expression by the host cell of a polypeptide encoded by the isolated gene. The invention also provides additional methods of producing a polypeptide, comprising introducing into a host cell a vector comprising a transcriptional regulatory sequence operably linked to an exonic region followed by an unspliced splice donor site, and culturing the host cell under conditions favoring the expression by said host cell of a polypeptide encoded by the exonic region, wherein the exon contains a translational start site positioned at any of the open reading frame positions relative to the 5'-most base of the unspliced splice donor site (e.g., the “A” in the ATG start codon may be at position -3 or at an increment of 3 bases upstream theretofrom (e.g., -6, -9, -12, -15, -18, etc.), at position -2 or at an increment of 3 bases upstream therefrom (e.g., -5, -8, -11, -14, -17, -20, etc.), or at position -1 or at an increment of 3 bases upstream therefrom (e.g., -4, -7, -10, -13, -16, -19, etc.), relative to the 5'-most base of the splice donor site). In related embodiments, the methods of the invention may further comprise isolating the polypeptide. The invention also is directed to polypeptides, which may or may not be isolated polypeptides, produced according to these methods.

Other preferred embodiments of the present invention will be apparent to one of ordinary skill in light of the following drawings and description of the invention, and of the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1. Schematic diagram of gene activation events described herein. The activation construct is transfected into cells and allowed to integrate into the host cell chromosomes at DNA breaks. If breakage occurs upstream of a gene of interest (e.g., Epo), and the appropriate activation construct integrates at the break such that its regulatory sequence becomes operably linked to the gene of interest, activation of the gene will occur. Transcription and splicing produce a chimeric RNA molecule containing exonic sequences from the activation construct and from the endogenous gene. Subsequent translation will result in the production of the protein of interest. Following isolation of the recombinant cell, gene expression can be further enhanced via gene amplification. The polyA tail is set forth in SEQ ID NO:33.

FIG. 2. Schematic diagram of non-translated activation constructs. The arrows denote promoter sequences. The exonic sequences are shown as open boxes and the splice donor sequence is indicated by S/D. Construct numbers corresponding to the description below are shown on the left. The selectable and amplifiable markers are not shown.

FIG. 3. Schematic diagram of translated activation constructs. The arrows denote promoter sequences. The exonic sequences are shown as open boxes and the splice donor sequence is indicated by S/D. The translated, signal peptide, epitope tag, and protease cleavage sequences are shown in the legend below the constructs. Construct numbers corresponding to the description below are shown on the left. The selectable and amplifiable markers are not shown.

FIG. 4. Schematic diagram of an activation construct capable of activating endogenous genes.

FIG. 5A-5D. Nucleotide sequence of pRG8R1-CD2 (SEQ ID NO:7).

FIG. 6A-6C. Nucleotide sequence of pRG8R2-CD2 (SEQ ID NO:8).

FIG. 7A-7C. Nucleotide sequence of pRG8R3-CD2 (SEQ ID NO:9).

Fig. 8A-8F. Examples of poly(A) trap vectors. Each vector is illustrated schematically in its linearized form. Each horizontal line represents a DNA molecule. The arrows denote promoter sequences located on the DNA molecule, and face in the direction of transcription. Transcribed regions include all sequences located downstream of a promoter. Untranslated regions are designated by hatched boxes and open reading frames are designated by open boxes. The following designations were used: splice donor site (S/D), signal secretion sequence (SP), epitope tag (ET), neomycin resistance gene (Neo). In the vectors depicted in FIG. 8B-8E, it is possible to omit the splice donor site immediately downstream of the Neo gene. In vectors lacking a splice donor site between the neo gene and the downstream promoter, the Neo transcript will utilize the splice donor site located 3' of the downstream promoter. In addition, as shown in the vectors depicted in FIG. 8B-8E, a downstream promoter may drive expression of an exon. It is recognized that this exon, when present, may encode codons in any reading frame. Using multiple vectors, codons in each of the 3 possible reading frames can be created.

FIG. 9A-9F. Examples of splice acceptor trap vectors containing a positive and a negative selectable marker driven from a single promoter. Each vector is illustrated schematically in its linearized form. Each horizontal line represents a DNA molecule. The arrows denote promoter sequences located on the DNA molecule, and face in the direction of transcription. Transcribed regions include all sequences located downstream of a promoter. Untranslated regions are designated by hatched boxes. Poly(A) signals are not present in these examples. As described in the specification, however, poly(A) signals may be placed on the vector 3' of either or both selectable markers. The following designations were used: splice donor site (S/D), signal secretion sequence (SP), epitope tag (ET), internal ribosome entry site (ires), hypoxanthine phosphoribosyl transferase (HPRT), and neomycin resistance gene (Neo). In these examples, Neo represents the positive selectable marker and HPRT represents the negative selectable maker. In the vectors shown in FIGS. 9C and 9F, the region designated exon contains a translation start codon. As described in the Detailed Description, the exon may encode a methionine residue, a partial signal sequence, a full signal secretion sequence, a portion of a protein, or an epitope tag. In addition, the codons may be present in any reading frame relative to the splice donor site. In other vector examples not shown, the region designated exon lacks a translation start codon.

FIG. 10A-10F. Examples of splice acceptor trap vectors containing a positive and a negative selectable marker driven from different promoters. Each vector is illustrated schematically in its linearized form. Each horizontal line represents a DNA molecule. The arrows denote promoter
sequences located on the DNA molecule, and face in the direction of transcription. Transcribed regions include all sequences located downstream of a promoter. Untranslated regions are designated by hatched boxes. Poly(A) signals are not present in these examples. As described in the specification, however, poly(A) signals may be placed on the vector 3' of either or both selectable markers. The following designations were used: splice donor site (S/D), internal ribosome entry site (ires), hypoxanthine phosphoribosyl transferase (HPRT), and neomycin resistance gene (Neo). In the vectors shown in FIGS. 10A-10F, Neo represents the positive selectable marker and HPRT represents the negative selectable marker. As shown, the vectors depicted in FIGS. 10A-10F do not contain a splice donor site 3' of the Neo gene; however, in other vectors not shown, a splice donor site may be located 3' of the Neo gene to facilitate splicing of the positive selection marker to an endogenous exon. In the vectors shown in FIG. 10C and 10F, the region designated exon contains a translation start codon. As described in the Detailed Description, the exon may encode a methionine residue, a partial signal sequence, a full signal secretion sequence, a portion of a protein, or an epitope tag. In addition, the codons may be present in any reading frame relative to the splice donor site. In other vector examples not shown, the region designated exon lacks a translation start codon.

[0087] FIG. 11A-11C. Schematic diagram of bidirectional activation vectors. The arrows denote promoter sequences. The exons are shown as checkered boxes and splice donor sites are indicated by S/D. The hatched boxes indicate exon sequences operably linked to the upstream promoter. It is understood that the exons on these vectors may be untranslated, or may contain a start codon and additional codons as described herein. As illustrated in the vectors depicted in FIG. 11B-11C, the vectors may contain a selectable marker. In these vectors, the neomycin resistance (Neo) gene is illustrated. In FIG. 11B, a polyadenylation signal (pA) is located downstream of the selectable marker. In FIG. 11C, polyadenylation signals are omitted from the vector.

[0088] FIG. 12A-12G. Examples of vectors useful for recovering exon I from activated endogenous genes. Each vector is illustrated schematically in its linearized form. Each horizontal line represents a DNA molecule. The arrows denote promoter sequences located on the DNA molecule, and face in the direction of transcription. Transcribed regions include all sequences located downstream of a promoter. Untranslated regions are designated by hatched boxes. Poly(A) signals are not present in the vectors depicted. As discussed in the Detailed Description, however, poly(A) signals may be placed on the vector 3' of either or both selectable markers. The following designations were used: splice donor site (S/D), internal ribosome entry site (ires), hypoxanthine phosphoribosyl transferase (HPRT), and neomycin resistance gene (Neo). In these examples, Neo represents the positive selectable marker and HPRT represents the negative selectable marker. It is also recognized that in these examples, the region designated exon, when present, lacks a translation start codon. In other examples not shown, the region designated exon contains a translation start codon. Furthermore, when the vector exon contains a translation start codon, the exon may encode a methionine residue, a partial signal sequence, a full signal secretion sequence, a portion of a protein, or an epitope tag. In addition, the codons may be present in each reading frame relative to the splice donor site.

[0089] FIG. 13. Illustration depicting two transcripts produced from the integrated vectors described in FIGS. 12A-12G. DNA strands are depicted as horizontal lines. Vector DNA is shown as a black line. Endogenous genomic DNA is shown as a grey line. Rectangles depict exons. Vector-encoded exons are shown as open rectangles, while endogenous exons are shown as shaded boxes. S/D denotes a splice donor site. Following integration, the vector encoded promoters activate transcription of the endogenous gene. Transcription resulting from the upstream promoter produces a spliced RNA molecule containing the vector encoded exon joined to the second and subsequent exons from an endogenous gene. Transcription from the downstream promoter, on the other hand, produces a transcript containing the sequences downstream of the integrated exon and the subsequent exons from an endogenous gene. The polyA tail is set forth in SEQ ID NO:33.

[0090] FIG. 14A-14B. Nucleotide sequence of pRIG1 (SEQ ID NO:18).

[0091] FIG. 15A-15B. Nucleotide sequence of pRIG21b (SEQ ID NO:19).

[0092] FIG. 16A-16B. Nucleotide sequence of pRIG22b (SEQ ID NO:20).

[0093] FIG. 17A-17G. Examples of poly(A) trap vectors. Each vector is illustrated schematically in its linearized form. Each horizontal line represents a DNA molecule. The arrows denote promoter sequences located on the DNA molecule, and face in the direction of transcription. Transcribed regions include all sequences located downstream of a promoter. Boxes indicate exons. Hatched boxes indicate untranslated regions. The following designations were used: splice donor site (S/D), signal secretion sequence (SP), epitope tag (ET), neomycin resistance gene (Neo), vector promoter #1 (VP #1), and vector promoter #2 (VP #2). As shown in the vectors depicted in FIG. 17C-17G, a promoter operably linked to an exxon and an unpaired splice donor site can be positioned upstream of the selectable marker. It is recognized that this exon, when present, may encode codons a start codon in any reading frame relative to the splice donor site. To activate protein expression from genes with different reading frames, three separate vectors can be used, each with a start codon in a different reading frame relative to the splice donor site.

[0094] FIG. 18. Illustration of the transcripts produced by the vector from FIG. 17C upon integration into a host cell genome upstream of a multi-exon endogenous gene. Each horizontal line represents a DNA molecule. Vertical lines running through the DNA strand mark the upstream and downstream vector/cellular genome boundaries. The arrows denote promoter sequences located on the DNA molecule, and face in the direction of transcription. Transcribed regions include all sequences located downstream of a promoter. Boxes indicate exons. Hatched boxes indicate untranslated regions. The endogenous exons are numbered using roman numerals. The following designations were used: splice donor site (S/D), neomycin resistance gene (Neo), vector promoter #1 (VP #1), vector promoter #2 (VP #2), endogenous promoter (EP) and polyadenylation signal (pA). Following integration, vector promoter #1 expresses a
chimeric transcript containing the Neo gene linked to the genomic sequences downstream of the integration site, including the processed (spliced) exons from the endogenous gene. Since transcript #1 contains a poly (A) signal from the endogenous gene, the Neo gene product will be efficiently produced, thereby conferring drug resistance on the cell. In addition to transcript #1, the integrated vector will generate a second transcript, designated transcript #2, originating from vector promoter #2. The structure of transcript #2 facilitates efficient translation of the protein encoded by the endogenous gene. As exemplified in FIG. 17, vectors containing alternative coding information in the vector encoded exon can be used to produce different chimeric proteins, containing, for example, signal sequences and/or epitope tags.

FIG. 19. Example of dual positive selectable marker vector. The vector is illustrated schematically in its linearized form. The horizontal line represents a DNA molecule. The arrows denote promoter sequences located on the DNA molecule, and face in the direction of transcription. Transcribed regions include all sequences located downstream of a promoter. Boxes indicate exons. Hatched boxes indicate untranslated regions. Poly(A) signals are not present in these examples. The following designations were used: splice donor site (S/D), hygromycin resistance gene (Hyg), neomycin resistance gene (Neo), vector promoter #1, and vector promoter #2.

FIG. 20A-20B. Examples of transcripts produced by a dual positive selectable marker vector integrated into a host cell genome adjacent to an endogenous gene. FIG. 20A illustrates the transcripts produced upon vector integration near a multi-exon gene. FIG. 20B illustrates the transcripts produced upon vector integration near a single exon gene. Each horizontal line represents a DNA molecule. Vertical lines running through the DNA strand mark the upstream and downstream vector/cellular genome boundaries. The arrows denote promoter sequences located on the DNA molecule, and face in the direction of transcription. Transcribed regions include all sequences located downstream of each promoter. Boxes indicate exons. Hatched boxes indicate untranslated regions. The endogenous exons are numbered using roman numerals. The following designations were used: splice donor site (S/D), hygromycin resistance gene (Hyg), neomycin resistance gene (Neo), vector promoter #1 (VP #1), vector promoter #2 (VP #2), endogenous promoter (EP), and polyadenylation signal (pA). Following integration, vector promoter #1 expresses a chimeric transcript containing the Hyg gene linked to the genomic sequences downstream of the integration site, including the processed (spliced) exons from the endogenous gene. Since transcript #1 contains a poly (A) signal from the endogenous gene, the Hyg gene product will be efficiently produced, thereby conferring drug resistance on the cell. In addition to transcript #1, the integrated vector will generate a second transcript, designated transcript #2, originating from vector promoter #2. In FIG. 20A, the neo gene is removed from transcript #2 upon splicing from the vector encoded splice donor site, and the first endogenous splice acceptor located downstream of the vector integration site (i.e. exon II in this example). Since multi-exon genes contain splice acceptor sites at the 5’ end of each exon (except exon I), the neo gene will be removed from transcript #2 in cells in which the vector has integrated near, and transcriptionally activated, a multi-exon gene. As a result, cells having activated multi-exon genes may be eliminated by selecting with G418 and hygromycin. In FIG. 20B, the neo gene is not removed from transcript #2 by splicing, since single exon genes do not contain any splice acceptor sequences. Thus, cells containing a vector integrated near single exon genes will survive double selection with G418 and hygromycin. These cells can be used to efficiently isolate the activated single exon genes using methods described herein.

FIG. 21A-21B. Examples of dual trap vectors containing a positive and a negative selectable marker. Each vector is illustrated schematically in its linearized form. Each horizontal line represents a DNA molecule. The arrows denote promoter sequences located on the DNA molecule, and face in the direction of transcription. Transcribed regions include all sequences located downstream of a promoter. Boxes indicate exons. Hatched boxes indicate untranslated regions. The following designations were used: splice donor site (S/D), hypoxanthine phosphoribosyl transferase (HPRT), neomycin resistance gene (Neo), vector promoter #1 (VP #1), vector promoter #2 (VP #2), and vector promoter #3 (VP #3). In the vectors shown in FIGS. 21A-21B, Neo represents the positive selectable marker and HPRT represents the negative selectable marker. In re 21B a third promoter is located upstream of the selectable markers. This upstream promoter is operably linked to an exon and unpaired splice donor site. Fig. The region designated exon contains a translation start codon in this example. As described herein, the exon may encode a methionine residue, a partial signal sequence, a full signal sequence, a portion of a protein, or an epitope tag. In addition, the codons may be present in any reading frame relative to the splice donor site. In other vector examples not shown, the region designated exon lacks a translation start codon.
transcriptionally activated, a multi-exon gene. As a result, cells containing activated multi-exon genes may be isolated by selecting with G418 and 8-Azaguanine 6-Thioguanine (AgTgh). Thus, cells containing a vector integrated near single exon genes will survive double selection with G418 and AgTgh. These cells can be used to efficiently isolate the activated multi-exon genes using methods described herein. In addition to transcripts #1 and #2, a third transcript, designated transcript #3 is produced from the integrated vector. Transcript #3, originating from vector promoter #3, contains an exonic sequence suitable for directing protein expression from the endogenous gene. This occurs following splicing from the first splice donor site downstream of promoter #3 to the first downstream splice acceptor site from the endogenous gene. In addition to directing protein expression, transcript #3, and/or transcripts #1 and/or #2, can be isolated for gene discovery purposes using the methods described herein.

[0099] FIG. 23A-23D. Example of a multi-Promoter/Activation Exon Vector. Each vector is illustrated schematically in its linearized form. Each horizontal line represents a DNA molecule. The arrows denote promoter sequences. Boxes indicate exons. Hatched boxes indicate untranslated regions. It is understood that the exons on these vectors may be untranslated, or may contain a start codon and additional codons as described herein. The following designations were used: splice donor site (S/D), vector promoter #1 (VP #1), vector promoter #2 (VP #2), vector promoter #3 (VP #3), and vector promoter #4 (VP #4). Individual vector activation exons are designated A, B, C, and D (SEQ ID Nos. 29-32, respectively). Each activation exon may contain a different structure. The structure of each activation exon and its flanking intron are shown below. It is understood, however, that any activation exon described herein, may be used on these vectors, in any combination and/or order, including exons that encode signal sequences, partial signal sequences, epitope tags, proteins, portions of proteins, and protein motifs. Any of the exons may lack a start codon. In addition, while not illustrated in these examples, these vectors may contain a selectable marker and/or an amplifiable marker. The selectable marker may contain a poly (A) signal or a splice donor site. When present, the splice donor site may be located upstream or downstream of the selectable marker. Alternatively, the selectable marker may not be operably linked to a poly (A) signal and/or a splice donor site.

[0100] FIG. 24. Examples of transcripts produced from a multi-Promoter/Activation Exon Vector upon integration into a host cell genome upstream of an endogenous gene. Each horizontal line represents a DNA molecule. Vertical lines running through the DNA strand mark the upstream and downstream vector/cellular genome boundaries. The arrows denote promoter sequences located on the DNA molecule, and face in the direction of transcription. Transcribed regions include all sequences located downstream of each promoter. Boxes indicate exons. Hatched boxes indicate untranslated regions. The endogenous exons are numbered using roman numerals. The following designations were used: splice donor site (S/D), binding domain (BD), activation domain (AD), recognition sequence (RS), and polyadenylation signal (pA). The binding domain vector is shown integrated into the genome of a host cell, upstream of an endogenous gene, designated gene A. The activation domain vector is shown integrated into the genome of the same host cell upstream of an endogenous gene, designated gene B. Both vectors are integrated into the genome of the same host cell. Following integration, each vector is capable of producing a fusion protein containing the binding domain (or activation domain, as the case may be) and the protein encoded by the downstream endogenous gene. If the binding domain fusion protein interacts with the activation domain fusion protein, a protein complex will be formed. This complex is capable of increasing expression of a reporter gene present in the cell.

[0103] FIG. 27. Examples of activation vectors useful for in vitro and in vivo transposition. Each vector is illustrated schematically in its linearized form. Each horizontal line represents a DNA molecule. The arrows denote promoter sequences. Boxes indicate exons. Hatched boxes indicate untranslated regions. The solid boxes indicate the transposon signals. It is recognized that there is directionality to the transposon signals, and that the signals are oriented in the configuration suitable for the type of transposition reaction (integration, inversion, or deletion). The following designations were used: splice donor site (S/D), neomycin resistance gene (Neo), dihydrofolate reductase (DHFR), puromycin resistance gene (Puro), poly (A) signal (pA), and the Epstein Barr Virus origin of replication (ori P). It is also recognized that activation exon may be encode amino acids...
in any reading frame (relative to the splice donor site), allowing activation of endogenous genes with different reading frames.

[0104] FIG. 28. Schematic illustration depicting integration of an activation vector into a cloned genomic DNA fragment by in vitro transposition. Each horizontal line represents a DNA molecule. The cloned genomic DNA is in a BAC vector. The single line represents the genomic DNA and the rectangle depicts the BAC vector sequences. The arrows denote promoter sequences located on the DNA molecule, and face in the direction of transcription. Transcribed regions include all sequences located downstream of each promoter. The vector activation exon is depicted as an open box. Exons from a gene encoded in the cloned genomic fragment are depicted as hatched boxes. The solid boxes indicate the transposon signals. It is recognized that there is directionality to the transposon signals, and that the signals are oriented in the configuration suitable for the type of transposition reaction (integration, inversion, or deletion). The following designations were used: splice donor site (S/D), and polyadenylation signal (pA). To integrate the vector into the genomic fragment, the activation vector is incubated with the cloned genomic DNA in the presence of transposase. Following integration of the activation vector into the genomic fragment, the plasmid may be transfected directly into an appropriate eukaryotic host cell to express the gene located downstream of the vector integration site. Alternatively, the BAC plasmid may be transformed into E. coli to produce larger quantities of plasmid for transfection into the appropriate eukaryotic host cell.

[0105] FIG. 29A-29B. Nucleotide sequence of pRIG14 (SEQ ID NO:21).

[0106] FIG. 30A-30C. Nucleotide sequence of pRIG19 (SEQ ID NO:22).

[0107] FIG. 31A-31C. Nucleotide sequence of pRIG20 (SEQ ID NO:23).

[0108] FIG. 32A-32C. Nucleotide sequence of pRIGaGd1 (SEQ ID NO:24).

[0109] FIG. 33A-33D. Nucleotide sequence of pRIGbd1 (SEQ ID NO:25).

[0110] FIG. 34A-34B. Nucleotide sequence of pUniBAC (SEQ ID NO:26).

[0111] FIG. 35A-35B. Nucleotide sequence of pRIG22 (SEQ ID NO:27).

[0112] FIG. 36. Schematic diagram of pRIG-TP. The vector is shown in its linearized form. The horizontal line represents a DNA molecule. The arrows denote promoters. Open boxes indicate exons. Filled boxes represent transposon recombination signals (from Tn5—compatible with the in vitro transposition kit available from Epicentre Technologies). The following designations were used: splice donor site (S/D), puromycin resistance gene (puro), dihydrofolate reductase gene (DHFR), Epstein Barr nuclear antigen 1-replication protein (EBNA-1), Epstein Barr virus origin of replication (ori P), poly (A) signal (pA), and activation exon (AE). It is understood that the activation exon can contain any sequence capable of directing protein synthesis, including a translation start codon in any reading frame, a partial secretion signal sequence, an entire secretion signal sequence, an epitope tag, a protein, a portion of a protein, or a protein motif. The activation exon may also lack a translation start codon.

[0113] FIG. 37A-37C. Nucleotide sequence of pRIG-T (SEQ ID NO:28).

DETAILED DESCRIPTION OF THE INVENTION

[0114] There are great advantages to gene activation by non-homologous recombination over other gene activation procedures. Unlike previous methods of protein over-expression, the methods described herein do not require that the gene of interest be cloned (isolated from the cell). Nor do they require any knowledge of the DNA sequence or structure of the gene to be over-expressed (i.e., the sequence of the ORF, introns, exons, or upstream and downstream regulatory elements) or knowledge of a gene’s expression patterns (i.e., tissue specificity, developmental regulation, etc.). Furthermore, the methods do not require any knowledge pertaining to the genomic organization of the gene of interest (i.e., the intron and exon structure).

[0115] The methods of the present invention thus involve vector constructs that do not contain target nucleotide sequences for homologous recombination. A target sequence allows homologous recombination of vector DNA with cellular DNA at a predetermined site on the cellular DNA, the site having homology for sequences in the vector, the homologous recombination at the predetermined site resulting in the introduction of the transcriptional regulatory sequence into the genome and the subsequent endogenous gene activation.

[0116] The method of the present invention does not involve integration of the vector at predetermined sites. Instead, the present methods involve integration of the vector constructs of the invention into cellular DNA (e.g., the cellular genome) by nonhomologous or “illegitimate” recombination, also called “non-targeted gene activation.” In related embodiments, the present invention also concerns non-targeted gene activation. Non-targeted gene activation has a number of important applications. First, by activating genes that are not normally expressed in a given cell type, it becomes possible to isolate a cDNA copy of genes independent of their normal expression pattern. This facilitates isolation of genes that are normally expressed in rare cells, during short developmental periods, and/or at very low levels. Second, by translationally activating genes, it is possible to produce protein expression libraries without the need for cloning the full-length cDNA. These libraries can be screened for new enzymes and proteins and/or for interesting phenotypes resulting from over-expression of an endogenous gene. Third, cell-lines over-expressing a specific protein can be created and used to produce commercial quantities of protein. Thus, activating endogenous genes provides a powerful approach to discovering and isolating new genes and proteins, and to producing large amounts of specific proteins for commercialization.

[0117] The vectors described herein do not contain target sequences. A target sequence is a sequence on the vector that has homology with a sequence or sequences within the gene to be activated or upstream of the gene to be activated, the upstream region being up to and including the first functional splice acceptor site on the same coding strand of the
gene of interest, and by means of which homology the transcriptional regulatory sequence that activates the gene of interest is integrated into the genome of the cell containing the gene to be activated. In the case of an enhancer integration vector for activating an endogenous gene, the vector does not contain homology to any sequence in the genome upstream or downstream of the gene of interest (or within the gene of interest) for a distance extending as far as enhancer function is operative.

[0118] The present methods, therefore, are capable of identifying new genes that have been or can be missed using conventional and currently available cloning techniques. By using the constructs and methodology described herein, unknown and/or uncharacterized genes can be rapidly identified and over-expressed to produce proteins. The proteins have use as, among other things, human therapeutics and diagnostics and as targets for drug discovery.

[0119] The methods are also capable of producing over-expression of known and/or characterized genes for in vitro or in vivo protein production.

[0120] A “known” gene is directed to the level of characterization of a gene. The invention allows expression of genes that have been characterized, as well as expression of genes that have not been characterized. Different levels of characterization are possible. These include detailed characterization, such as cloning, DNA, RNA, and/or protein sequencing, and relating the regulation and function of the gene to the cloned sequence (e.g., recognition of promoter and enhancer sequences, functions of the open reading frames, introns, and the like). Characterization can be less detailed, such as having mapped a gene and related function, or having a partial amino acid or nucleotide sequence, or having purified a protein and ascertained a function. Characterization may be minimal, as when a nucleotide or amino acid sequence is known or a protein has been isolated but the function is unknown. Alternatively, a function may be known but the associated protein or nucleotide sequence is not known or is known but has not been correlated to the function. Finally, there may be no characterization in that both the existence of the gene and its function are not known. The invention allows expression of any gene at any of these or other specific degrees of characterization.

[0121] Many different proteins (also referred to herein interchangeably as “gene products” or “expression products”) can be activated or over-expressed by a single activation construct and in a single set of transfections. Thus, a single cell or different cells in a set of transfectants (library) can over-express more than one protein following transfection with the same or different constructs. Previous activation methods require a unique construct to be created for each gene to be activated.

[0122] Further, many different integration sites adjacent to a single gene can be created and tested simultaneously using a single construct. This allows rapid determination of the optimal genomic location of the activation construct for protein expression.

[0123] Using previous methods, the 5' end of the gene of interest had to be extensively characterized with respect to sequence and structure. For each activation construct to be produced, an appropriate targeting sequence had to be isolated. Usually, this must be an isogenic sequence isolated from the same person or laboratory strain of animal as the cells to be activated. In some cases, this DNA may be 50 kb or more from the gene of interest. Thus, production of each targeting construct required an arduous amount of cloning and sequencing of the endogenous gene. However, since sequence and structure information is not required for the methods of the present invention, unknown genes and genes with uncharacterized upstream regions can be activated.

[0124] This is made possible using in situ gene activation using non-homologous recombination of exogenous DNA sequences with cellular DNA. Methods and compositions (e.g., vector constructs) required to accomplish such in situ gene activation using non-homologous recombination are provided by the present invention.

[0125] DNA molecules can recombine to redistribute their genetic content by several different and distinct mechanisms, including homologous recombination, site-specific recombination, and non-homologous/illegitimate recombination. Homologous recombination involves recombination between stretches of DNA that are highly similar in sequence. It has been demonstrated that homologous recombination involves pairing between the homologous sequences along their length prior to redistribution of the genetic material. The exact site of crossover can be at any point in the homologous segments. The efficiency of recombination is proportional to the length of homologous targeting sequence (Hope, Development 113:399 (1991); Reddy et al., J Virol. 65:1507 (1991)), the degree of sequence identity between the two recombining sequences (von Melehrn et al., Genes Dev. 6:919 (1992)), and the ratio of homologous to non-homologous DNA present on the construct (Letson, Genetics 117:759 (1987)).

[0126] Site-specific recombination, on the other hand, involves the exchange of genetic material at a predetermined site, designated by specific DNA sequences. In this reaction, a protein recombinase binds to the recombination signal sequences, creates a strand scission, and facilitates DNA strand exchange. Cre/Lox recombination is an example of site specific recombination.

[0127] Non-homologous/illegitimate recombination, such as that used advantageously by the methods of the present invention, involves the joining (exchange or redistribution) of genetic material that does not share significant sequence homology and does not occur at site-specific recombination sequences. Examples of non-homologous recombination include integration of exogenous DNA into chromosomes at non-homologous sites, chromosomal translocations and deletions, DNA end-joining, double strand break repair of chromosome ends, bridge-breakage fusion, and concatemerization of transfected sequences. In most cases, non-homologous recombination is thought to occur through the joining of “free DNA ends.” Free ends are DNA molecules that contain an end capable of being joined to a second DNA end either directly, or following repair or processing. The DNA end may consist of a 5' overhang, 3' overhang, or blunt end.

[0128] As used herein, retroviral insertion and other transposition reactions are loosely considered forms of non-homologous recombination. These reactions do not involve the use of homology between the recombining molecules. Furthermore, unlike site-specific recombination, these types of recombination reactions do not occur between discrete sites. Instead, a specific protein/DNA complex is required on
only one of the recombination partners (i.e., the retrovirus or transposon), with the second DNA partner (i.e., the cellular genome) usually being relatively non-specific. As a result, these “vectors” do not integrate into the cellular genome in a targeted fashion, and therefore they can be used to deliver the activation construct according to the present invention.

[0129] Vector constructs useful for the methods described herein ideally may contain a transcriptional regulatory sequence that undergoes non-homologous recombination with genomic sequences in a cell to over-express an endogenous gene in that cell. The vector constructs of the invention also lack homologous targeting sequences. That is, they do not contain DNA sequences that target host cell DNA and promote homologous recombination at the target site. Thus, integration of the vector constructs of the present invention into the cellular genome occurs by non-homologous recombination, and can lead to over-expression of a cellular gene via the introduced transcriptional regulatory sequence contained on the integrated vector construct.

[0130] The invention is generally directed to methods for over-expressing an endogenous gene in a cell, comprising introducing a vector containing a transcriptional regulatory sequence into the cell, allowing the vector to integrate into the genome of the cell by non-homologous recombination, and allowing over-expression of the endogenous gene in the cell. The method does not require previous knowledge of the sequence of the endogenous gene or even of the existence of the gene. Where the sequence of the gene to be activated is known, however, the constructs can be engineered to contain the proper configuration of vector elements (e.g., location of the start codon, addition of codons present in the first exon of the endogenous gene, and the proper reading frame) to achieve maximal overexpression and/or the appropriate protein sequence.

[0131] In certain embodiments of the invention, the cell containing the vector may be screened for expression of the gene.

[0132] The cell over-expressing the gene can be cultured in vitro under conditions favoring the production, by the cell, of desired amounts of the gene product of the endogenous gene that has been activated or whose expression has been increased. If desired, the gene product can then be isolated or purified to use, for example, in protein therapy or drug discovery.

[0133] Alternatively, the cell expressing the desired gene product can be allowed to express the gene product in vivo.

[0134] The vector construct can consist essentially of the transcriptional regulatory sequence.

[0135] Alternatively, the vector construct can consist essentially of the transcriptional regulatory sequence and one or more amplifiable markers.

[0136] The invention, therefore, is also directed to methods for over-expressing an endogenous gene in a cell, comprising introducing a vector containing a transcriptional regulatory sequence and an amplifiable marker into the cell, allowing the vector to integrate into the genome of the cell by non-homologous recombination, and allowing over-expression of the endogenous gene in the cell.

[0137] The cell containing the vector is screened for over-expression of the gene.

[0138] The cell over-expressing the gene is cultured such that amplification of the endogenous gene is obtained. The cell can then be cultured in vitro so as to produce desired amounts of the gene product of the amplified endogenous gene that has been activated or whose expression has been increased. The gene product can then be isolated and purified.

[0139] Alternatively, following amplification, the cell can be allowed to express the endogenous gene and produce desired amounts of the gene product in vivo.

[0140] The vector construct can consist essentially of the transcriptional regulatory sequence and the splice donor sequence.

[0141] The invention, therefore, is also directed to methods for over-expressing an endogenous gene in a cell comprising introducing a vector containing a transcriptional regulatory sequence and an unpaired splice donor sequence into the cell, allowing the vector to integrate into the genome of the cell by non-homologous recombination, and allowing over-expression of the endogenous gene in the cell.

[0142] The cell containing the vector is screened for expression of the gene.

[0143] The cell over-expressing the gene can be cultured in vitro so as to produce desirable amounts of the gene product of the endogenous gene whose expression has been activated or increased. The gene product can then be isolated and purified.

[0144] Alternatively, the cell can be allowed to express the desired gene product in vivo.

[0145] The vector construct can consist essentially of a transcriptional regulatory sequence operably linked to an unpaired splice donor sequence and also containing an amplifiable marker.

[0146] Other activation vectors include constructs with a transcriptional regulatory sequence and an exonic sequence containing a start codon; a transcriptional regulatory sequence and an exonic sequence containing a translation start codon and a secretion signal sequence; constructs with a transcriptional regulatory sequence and an exonic sequence containing a translation start codon, and an epitope tag; constructs containing a transcriptional regulatory sequence and an exonic sequence containing a translation start codon, a signal sequence and an epitope tag; constructs containing a transcriptional regulatory sequence and an exonic sequence containing a translation start codon, a signal sequence, an epitope tag, and a sequence-specific protease site. In each of the above constructs, the exon on the construct is located immediately upstream of an unpaired splice donor site.

[0147] The constructs can also contain a regulatory sequence, a selectable marker lacking a poly(A) signal, an internal ribosome entry site (ires), and an unpaired splice donor site (FIG. 4). A start codon, signal secretion sequence, epitope tag, and/or a protease cleavage site may optionally be included between the ires and the unpaired splice donor sequence. When this construct integrates upstream of a gene, the selectable marker will be efficiently expressed since a poly(A) site will be supplied by the endogenous gene. In addition the downstream gene will also be expressed since the ires will allow protein translation to initiate at the
downstream open reading frame (i.e. the endogenous gene). Thus, the message produced by this activation construct will be polycistronic. The advantage of this construct is that integration events that do not occur near genes and in the proper orientation will not produce a drug resistant colony. The reason for this is that without a poly(A) tail (supplied by the endogenous gene), the neomycin resistance gene will not express efficiently. By reducing the number of nonproductive integration events, the complexity of the library can be reduced without affecting its coverage (the number of genes activated), and this will facilitate the screening process.

[0148] In another embodiment of this construct, cre-lox recombination sequences can be included between the regulatory sequence and the neo start codon and between the ires and the unpaired splice donor site (between the ires and the start codon, if present). Following isolation of cells that have activated the gene of interest, the neo gene and ires can be removed by transfecting the cells with a plasmid encoding the cre recombinase. This would eliminate the production of the polycistronic message and allow the endogenous gene to be expressed directly from the regulatory sequence on the integrated activation construct. Use of Cre recombination to facilitate deletion of genetic elements from mammalian chromosomes has been described (Gu et al., Science 265:103 (1994); Sauer, Meth. Enzymology 225:890-900 (1993)).

[0149] Thus, constructs useful in the methods described herein include, but are not limited to, the following (See also FIGS. 1-4):

[0150] 1) Construct with a regulatory sequence and an exon lacking a translation start codon.

[0151] 2) Construct with a regulatory sequence and an exon lacking a translation start codon followed by a splice donor site.

[0152] 3) Construct with a regulatory sequence and an exon containing a translation start codon in reading frame 1 (relative to the splice donor site), followed by an unpaired splice donor site.

[0153] 4) Construct with a regulatory sequence and an exon containing a translation start codon in reading frame 2 (relative to the splice donor site), followed by an unpaired splice donor site.

[0154] 5) Construct with a regulatory sequence and an exon containing a translation start codon in reading frame 3 (relative to the splice donor site), followed by an unpaired splice donor site.

[0155] 6) Construct with a regulatory sequence and an exon containing a translation start codon and a signal secretion sequence in reading frame 1 (relative to the splice donor site), followed by an unpaired splice donor site.

[0156] 7) Construct with a regulatory sequence and an exon containing a translation start codon and a signal secretion sequence in reading frame 2 (relative to the splice donor site), followed by an unpaired splice donor site.

[0157] 8) Construct with a regulatory sequence and an exon containing a translation start codon and a signal secretion sequence in reading frame 3 (relative to the splice donor site), followed by an unpaired splice donor site.

[0158] 9) Construct with a regulatory sequence and an exon containing (from 5' to 3') a translation start codon and an epitope tag in reading frame 1 (relative to the splice donor site), followed by an unpaired splice donor site.

[0159] 10) Construct with a regulatory sequence and an exon containing (from 5' to 3') a translation start codon and an epitope tag in reading frame 2 (relative to the splice donor site), followed by an unpaired splice donor site.

[0160] 11) Construct with a regulatory sequence and an exon containing (from 5' to 3') a translation start codon and an epitope tag in reading frame 3 (relative to the splice donor site), followed by an unpaired splice donor site.

[0161] 12) Construct with a regulatory sequence and an exon containing (from 5' to 3') a translation start codon, a signal secretion sequence, and an epitope tag in reading frame 1 (relative to the splice donor site), followed by an unpaired splice donor site.

[0162] 13) Construct with a regulatory sequence and an exon containing (from 5' to 3') a translation start codon, a signal secretion sequence, and an epitope tag in reading frame 2 (relative to the splice donor site), followed by an unpaired splice donor site.

[0163] 14) Construct with a regulatory sequence and an exon containing (from 5' to 3') a translation start codon, a signal secretion sequence, and an epitope tag in reading frame 3 (relative to the splice donor site), followed by an unpaired splice donor site.

[0164] 15) Construct with a regulatory sequence and an exon containing (from 5' to 3') a translation start codon, a signal secretion sequence, an epitope tag, and a sequence specific protease site in reading frame 1 (relative to the splice donor site), followed by an unpaired splice donor site.

[0165] 16) Construct with a regulatory sequence and an exon containing (from 5' to 3') a translation start codon, a signal secretion sequence, an epitope tag, and a sequence specific protease site in reading frame 2 (relative to the splice donor site), followed by an unpaired splice donor site.

[0166] 17) Construct with a regulatory sequence and an exon containing (from 5' to 3') a translation start codon, a signal secretion sequence, an epitope tag, and a sequence specific protease site in reading frame 3 (relative to the splice donor site), followed by an unpaired splice donor site.

[0167] 18) Construct with a regulatory sequence linked to a selectable marker, followed by an internal ribosome entry site, and an unpaired splice donor site.

[0168] 19) Construct 18 in which a crelox recombination signal is located between a) the regulatory sequence and the open reading frame of the selectable marker and b) between the ires and the unpaired splice donor site.
Construct with a regulatory sequence operably linked to an exon containing green fluorescent protein lacking a stop codon, followed by an unpaired splice donor site.

It is to be understood, however, that any vector used in the methods described herein can include one or more (i.e., one, two, three, four, five, or more, and most preferably one or two) amplifiable markers. Accordingly, methods can include a step in which the endogenous gene is amplified. Placement of one or more amplifiable markers on the activation construct results in the juxtaposition of the gene of interest and the one or more amplifiable markers in the activated cell. Once the activated cell has been isolated, expression can be further increased by selecting for cells containing an increased copy number of the locus containing both the gene of interest and the activation construct. This can be accomplished by selection methods known in the art, for example by culturing cells in selective culture media containing one or more selection agents that are specific for the one or more amplifiable markers contained on the genetic construct or vector.

Following activation of an endogenous gene by nonhomologous integration of any of the vectors described above, the expression of the endogenous gene may be further increased by selecting for increased copies of the amplifiable marker(s) located on the integrated vector. While such an approach may be accomplished using one amplifiable marker on the integrated vector, in an alternative embodiment the invention provides such methods wherein two or more (i.e., two, three, four, five, or more, and most preferably two) amplifiable markers may be included on the vector to facilitate more efficient selection of cells that have amplified the vector and flanking gene of interest. This approach is particularly useful in cells that have a functional endogenous copy of one or more of the amplifiable marker(s) that are contained on the vector, since the selection procedure can result in isolation of cells that have incorrectly amplified the endogenous amplifiable marker(s) rather than the vector-encoded amplifiable marker(s). This approach is also useful to select against cells that develop resistance to the selective agent by mechanisms that do not involve gene amplification. The approach using two or more amplifiable markers is advantageous in these situations because the probability of a cell developing resistance to two or more selective agents (resistance to which is encoded by two or more amplifiable markers) without amplifying the integrated vector and flanking gene of interest is significantly lower than the probability of the cell developing resistance to any single selective agent. Thus, by selecting for two or more vector encoded amplifiable markers, either simultaneously or sequentially, a greater percentage of cells that are ultimately isolated will contain the amplified vector and gene of interest.

Thus, in another embodiment, the vectors of the invention may contain two or more (i.e., two, three, four, five, or more, and most preferably two) amplifiable markers. This approach allows more efficient amplification of the vector sequences and adjacent gene of interest following activation of expression.

Examples of amplifiable markers that may be used constructing the present vectors include, but are not limited to, dihydrofolate reductase, adenosine deaminase, aspartate transcarbamylase, dihydro-orotase, and carbamyl phosphate synthase.

It is also understood that any of the constructs described herein may contain a eukaryotic viral origin of replication, either in place of, or in conjunction with an amplifiable marker. The presence of the viral origin of replication allows the integrated vector and adjacent endogenous gene to be isolated as an episome and/or amplified to high copy number upon introduction of the appropriate viral replication protein. Examples of useful viral origins include, but are not limited to, SV40 ori and EBV ori P.

The invention also encompasses embodiments in which the constructs disclosed herein consist essentially of the components specifically described for these constructs. It is also understood that the above constructs are examples of constructs useful in the methods described herein, but that the invention encompasses functional equivalents of such constructs.

The term “vector” is understood to generally refer to the vehicle by which the nucleotide sequence is introduced into the cell. It is not intended to be limited to any specific sequence. The vector could itself be the nucleotide sequence that activates the endogenous gene or could contain the sequence that activates the endogenous gene. Thus, the vector could be simply a linear or circular polynucleotide containing essentially only those sequences necessary for activation, or could be those sequences in a larger polynucleotide or other construct such as a DNA or RNA viral genome, a whole virion, or other biological construct used to introduce the critical nucleotide sequences into a cell. It is also understood that the phrase “vector construct” or the term “construct” may be used interchangeably with the term “vector” herein.

The vector can contain DNA sequences that exist in nature or that have been created by genetic engineering or synthetic processes.

The construct, upon nonhomologous integration into the genome of a cell, can activate expression of an endogenous gene. Expression of the endogenous gene may result in production of full length protein, or in production of a truncated biologically active form of the endogenous protein, depending on the integration site (e.g., upstream region versus intron 2). The activated gene may be a known gene (e.g., previously cloned or characterized) or unknown gene (previously not cloned or characterized). The function of the gene may be known or unknown.

Examples of proteins with known activities include, but are not limited to, cytokines, growth factors, neurotransmitters, enzymes, structural proteins, cell surface receptors, intracellular receptors, hormones, antibodies, and transcription factors. Specific examples of known proteins that can be produced by this method include, but are not limited to, erythropoietin, insulin, growth hormone, glucocerebrosidase, tissue plasminogen activator, granulocyte-colony stimulating factor (G-CSF), granulocyte/macrophage colony stimulating factor (GM-CSF), macrophage colony-stimulating factor (M-CSF) interferon α, interferon β, interferon γ, interleukin-2, interleukin-3, interleukin-4, interleukin-6, interleukin-8, interleukin-10, interleukin-11, interleukin-12, interleukin-13, interleukin-14, TGF-β, blood proteins, etc.
clotting factor V, blood clotting factor VII, blood clotting factor VIII, blood clotting factor IX, blood clotting factor X, TSH-62, bone growth factor-2, bone growth factor-7, tumor necrosis factor, alpha-1 antitrypsin, anti-thrombin III, leukemia inhibitory factor, glucagon, Protein C, protein kinase C, stem cell factor, follicle stimulating hormone β, uridine, nerve growth factors, insulin-like growth factors, insulinotropin, parathyroid hormone, lactoferrin, complement inhibitors, platelet derived growth factor, keratinocyte growth factor, hepatocyte growth factor, endothelial cell growth factor, neurotropin-3, thrombopoietin, chorionic gonadotropin, thrombomodulin, alpha glucosidase, epidermal growth factor, and fibroblast growth factor. The invention also allows the activation of a variety of genes expressing transmembrane proteins, and production and isolation of such proteins, including but not limited to cell surface receptors for growth factors, hormones, neurotransmitters and cytokines such as those described above, transmembrane ion channels, cholesterol receptors, receptors for lipid-proteins (including LDLs and HDLs) and other lipid moieties, integrins and other extracellular matrix receptors, cytoskeletal anchoring proteins, immunoglobulin receptors, CD antigens (including CD2, CD3, CD4, CD8, and CD34 antigens), and other cell surface transmembrane structural and functional proteins that are known in the art. As one of ordinary skill will appreciate, other cellular proteins and receptors that are known in the art may also be produced by the methods of the invention.

[0180] One of the advantages of the method described herein is that virtually any gene can be activated. However, since genes have different genomic structures, including different intron/exon boundaries and locations of start codons, a variety of activation constructs is provided to activate the maximum number of different genes within a population of cells.

[0181] These constructs can be transfected separately into cells to produce libraries. Each library contains cells with a unique set of activated genes. Some genes will be activated by several different activation constructs. In addition, portions of a gene can be activated to produce truncated, biologically active proteins. Truncated proteins can be produced, for example, by integration of an activation construct into introns or exons in the middle of an endogenous gene rather than upstream of the second exon.

[0182] Use of different constructs also allows the activated gene to be modified to contain new sequences. For example, a secretion signal sequence can be included on the activation construct to facilitate the secretion of the activated gene. In some cases, depending on the intron/exon structure or the gene of interest, the secretion signal sequence can replace all or part of the signal sequence of the endogenous gene. In other cases, the signal sequence will allow a protein which is normally located intracellularly to be secreted.

[0183] The regulatory sequence on the vector can be a constitutive promoter. Alternatively, the promoter may be inducible. Use of inducible promoters will allow low basal levels of activated protein to be produced by the cell during routine culturing and expansion. The cells may then be induced to produce large amounts of the desired proteins, for example, during manufacturing or screening. Examples of inducible promoters include, but are not limited to, the tetracycline inducible promoter and the metallothionein promoter.

[0184] In preferred embodiments of the invention, the regulatory sequence on the vectors of the invention may be a promoter, an enhancer, or a repressor, any of which may be tissue specific.

[0185] The regulatory sequence on the vector can be isolated from cellular or viral genomes. Examples of cellular regulatory sequences include, but are not limited to, regulatory elements from the actin gene, metallothionein I gene, immunoglobulin genes, casein I gene, serum albumin gene, collagen gene, globin genes, laminin gene, spectrin gene, ankyrin gene, sodium/potassium ATPase gene, and tubulin gene. Examples of viral regulatory sequences include, but are not limited to, regulatory elements from Cytomegalovirus (CMV) immediate early gene, adenovirus late genes, SV-40 genes, retroviral LTRs, and Herpesvirus genes. Typically, regulatory sequences contain binding sites for transcription factors such as NF-kB, SP-1, TATA binding protein, AP-1, and CAAT binding protein. Functionally, the regulatory sequence is defined by its ability to promote, enhance, or otherwise alter transcription of an endogenous gene.

[0186] In certain preferred embodiments, the regulatory sequence is a viral promoter. In particularly preferred embodiments, the promoter is the CMV immediate early gene promoter. In alternative embodiments, the regulatory element is a cellular, non-viral promoter.

[0187] In alternative preferred embodiments, the regulatory element may be or may contain an enhancer. In particularly preferred such embodiments, the enhancer is the cytomegalovirus immediate early gene enhancer. In alternative embodiments, the enhancer is a cellular, non-viral enhancer.

[0188] In alternative preferred embodiments, the regulatory element may be or may contain a repressor. In particularly preferred such embodiments, the repressor may be a viral repressor or a cellular, non-viral repressor.

[0189] The transcriptional regulatory sequence can also comprise one or more scaffold-attachment regions or matrix attachment sites, negative regulatory elements, and transcription factor binding sites. Regulatory sequences can also include locus control regions.

[0190] The invention also encompasses the use of retrovirus transcriptional regulatory sequences, e.g., long terminal repeats. Where these are used, however, they are not necessarily linked to any retrovirus sequence that materially affects the function of the transcriptional regulatory sequence as a promoter or enhancer of transcription of the endogenous gene to be activated (i.e., the cellular gene with which the transcriptional regulatory sequence recombines to activate).

[0191] The vector constructs of the invention may also comprise a regulatory sequence which is not operably linked to exonic sequences on the vector. For example, when the regulatory element is an enhancer, it can integrate near an endogenous gene (e.g., upstream, downstream, or in an intron) and stimulate expression of the gene from its endogenous promoter. By this mechanism of activation, exonic sequences from the vector are absent in the transcript of the activated gene.

[0192] Alternatively, the regulatory element may be operably linked to an exon. The exon may be a naturally
occurring sequence or may be non-naturally occurring (e.g., produced synthetically). To activate endogenous genes lacking a start codon in their first exon (e.g., follicle stimulating hormone-β), a start codon is preferably omitted from the exon on the vector. To activate endogenous genes containing a start codon in the first exon (e.g., erythropoietin and growth hormone), the exon on the vector preferably contains a start codon, usually ATG and preferably an efficient translation initiation site (Kozak, J. Mol. Biol. 196: 947 (1987)). The exon may contain additional codons following the start codon. These codons may be derived from a naturally occurring gene or may be non-naturally occurring (e.g., synthetic). The codons may be the same as the codons present in the first exon of the endogenous gene to be activated. Alternatively, the codons may be different than the codons present in the first exon of the endogenous gene. For example, the codons may encode an epitope tag, signal secretion sequence, transmembrane domain, selectable marker, or screenable marker. Optionally, an unpaired splice donor site may be present immediately 3' of the exonic sequence. When the structure of the gene to be activated is known, the splice donor site should be placed adjacent to the vector exon in a location such that the codons in the vector will be in frame with the codons of the second exon of the endogenous gene following splicing. When the structure of the endogenous gene to be activated is not known, separate constructs, each containing a different reading frame, are used.

Splicing of primary transcripts, the process by which introns are removed, is directed by a splice donor site and a splice acceptor site, located at the 5' and 3' ends of introns, respectively. The consensus sequence for splice donor sites is (A/C)AG UAGGU (where R represents a purine nucleotide) with nucleotides in positions 1-3 located in the exon and nucleotides UAGGU located in the intron.

An unpaired splice donor site is defined herein as a splice donor site present on the activation construct without a downstream splice acceptor site. When the vector is integrated by nonhomologous recombination into a host cell's genome, the unpaired splice donor site becomes paired with a splice acceptor site from an endogenous gene. The splice donor site from the vector, in conjunction with the splice acceptor site from the endogenous gene, will then direct the excision of all of the sequences between the vector splice donor site and the endogenous splice acceptor site. Excision of these intervening sequences removes sequences that interfere with translation of the endogenous protein.

The terms upstream and downstream, as used herein, are intended to mean in the 5' or in the 3' direction, respectively, relative to the coding strand. The term "upstream region" of a gene is defined as the nucleic acid sequence 5' of its second exon (relative to the coding strand) up to and including the last exon of the first adjacent gene having the same coding strand. Functionally, the upstream region is any site 5' of the second exon of an endogenous gene capable of allowing a nonhomologously integrated vector to become operably linked to the endogenous gene.

The vector construct can contain a selectable marker to facilitate the identification and isolation of cells containing a nonhomologously integrated activation construct. Examples of selectable markers include genes encoding neomycin resistance (neo), hygromycin phosphoribosyl transferase (HPRT), puromycin (puc), dihydro-orotase glutamine synthetase (GS), histidine D (his D), carbamyl phosphate synthase (CAD), dihydrofolate reductase (DHFR), multidrug resistance 1 (mdr1), aspartate transcarbamylase, xanthine-guanine phosphoribosyl transferase (gpt), and adenosine deaminase (ada).

Alternatively, the vector can contain a screenable marker, in place of or in addition to, the selectable marker. A screenable marker allows the cells containing the vector to be isolated without placing them under drug or other selective pressures. Examples of screenable markers include genes encoding cell surface proteins, fluorescent proteins, and enzymes. The vector containing cells may be isolated, for example, by FACS using fluorescently-tagged antibodies to the cell surface protein or substrates that can be converted to fluorescent products by a vector encoded enzyme.

Alternatively, selection can be effected by phenotypic selection for a trait provided by the endogenous gene product. The activation construct, therefore, can lack a selectable marker other than the "marker" provided by the endogenous gene itself. In this embodiment, activated cells can be selected based on a phenotype conferred by the activated gene. Examples of selectable phenotypes include cellular proliferation, growth factor independent growth, colony formation, cellular differentiation (e.g., differentiation into a neuronal cell, muscle cell, epithelial cell, etc.), anchorage independent growth, activation of cellular factors (e.g., kinases, transcription factors, nucleases, etc.), expres-
sion of cell surface receptors/proteins, gain or loss of cell-cell adhesion, migration, and cellular activation (e.g., resting versus activated T cells).

0201 A selectable marker may also be omitted from the construct when transfected cells are screened for gene activation products without selecting for the stable integrants. This is particularly useful when the efficiency of stable integration is high.

0202 The vector may contain one or more (i.e., one, two, three, four, five, or more, and most preferably one or two) amplifiable markers to allow for selection of cells containing increased copies of the integrated vector and the adjacent activated endogenous gene. Examples of amplifiable markers include but are not limited to dihydrofolate reductase (DHFR), adenosine deaminase (ada), dihydro-orotate glutamine synthetase (GS), and carbamyl phosphate synthase (CAD).

0203 The vector may contain eukaryotic viral origins of replication useful for gene amplification. These origins may be present in place of, or in conjunction with, an amplifiable marker.

0204 The vector may also contain genetic elements useful for the propagation of the construct in micro-organisms. Examples of useful genetic elements include microbial origins of replication and antibiotic resistance markers.

0205 These vectors, and any of the vectors disclosed herein, and obvious variants recognized by one of ordinary skill in the art, can be used in any of the methods described herein to form any of the compositions producible by those methods.

0206 Nonhomologous integration of the construct into the genome of a cell results in the operable linkage between the regulatory elements from the vector and the exons from an endogenous gene. In preferred embodiments, the insertion of the vector regulatory sequences is used to upregulate expression of the endogenous gene. Upregulation of gene expression includes converting a transcriptionally silent gene to a transcriptionally active gene. It also includes enhancement of gene expression for genes that are already transcriptionally active, but produce protein at levels lower than desired. In other embodiments, expression of the endogenous gene may be affected in other ways such as downregulation of expression, creation of an inducible phenotype, or changing the tissue specificity of expression.

0207 According to the invention, in vitro methods of production of a gene expression product may comprise, for example, (a) introducing a vector of the invention into a cell; (b) allowing the vector to integrate into the genome of the cell by non-homologous recombination; (c) allowing overexpression of an endogenous gene in the cell by upregulation of the gene by the transcriptional regulatory sequence contained on the vector; (d) screening the cell for overexpression of the endogenous gene; and (e) culturing the cell under conditions favoring the production of the expression product of the endogenous gene by the cell. Such in vitro methods of the invention may further comprise isolating the expression product to produce an isolated gene expression product. In such methods, any art-known method of protein isolation may be advantageously used, including but not limited to chromatography (e.g., HPLC, FPLC, IEC, ion exchange, affinity, size exclusion, and the like), precipitation (e.g., ammonium sulfate precipitation, immunoprecipitation, and the like), electrophoresis, and other methods of protein isolation and purification that will be familiar to one of ordinary skill in the art.

0208 Analogously, in vivo methods of production of a gene expression product may comprise, for example, (a) introducing a vector of the invention into a cell; (b) allowing the vector to integrate into the genome of the cell by non-homologous recombination; (c) allowing overexpression of an endogenous gene in the cell by upregulation of the gene by the transcriptional regulatory sequence contained on the vector; (d) screening the cell for overexpression of the endogenous gene; and (e) introducing the isolated and cloned cell into a eukaryote under conditions favoring the overexpression of the endogenous gene by the cell in vivo in the eukaryote. According to this aspect of the invention, any eukaryote may be advantageously used, including fungi (particularly yeasts), plants, and animals, more preferably animals, still more preferably vertebrates, and most preferably mammals, particularly humans. In certain related embodiments, the invention provides such methods which further comprise isolating and cloning the cell prior to introducing it into the eukaryote.

0209 As used herein the phrases “conditions favoring the production” of an expression product, “conditions favoring the overexpression” of a gene, and “conditions favoring the activation” of a gene, in a cell or by a cell in vitro refer to any and all suitable environmental, physical, nutritional or biochemical parameters that allow, facilitate, or promote production of an expression product, or overexpression or activation of a gene, by a cell in vitro. Such conditions may, of course, include the use of culture media, incubation, lighting, humidity, etc., that are optimal or that allow, facilitate, or promote production of an expression product, or overexpression or activation of a gene, by a cell in vitro. Analogously, as used herein the phrases “conditions favoring the production” of an expression product, “conditions favoring the overexpression” of a gene, and “conditions favoring the activation” of a gene, in a cell or by a cell in vivo refer to any and all suitable environmental, physical, nutritional, biochemical, behavioral, genetic, and emotional parameters under which an animal containing a cell is maintained, that allow, facilitate, or promote production of an expression product, or overexpression or activation of a gene, by a cell in a eukaryote in vivo. Whether a given set of conditions are favorable for gene expression, activation, or overexpression, in vitro or in vivo, may be determined by one of ordinary skill using the screening methods described and exemplified below, or other methods for measuring gene expression, activation, or overexpression that are routine in the art.

0210 As used herein, the phrase “activating an endogenous gene” means inducing the production of a transcript encoding the endogenous gene at levels higher than those normally found in the cell containing the endogenous gene. In some applications, “activating an endogenous gene” may also mean producing the protein, or a portion of the protein, encoded by the endogenous gene at levels higher than those normally found in the cell containing the endogenous gene.

0211 The invention also encompasses cells made by any of the above methods. The invention encompasses cells containing the vector constructs, cells in which the vector
constructs have integrated, and cells which are over-expressing desired gene products from an endogenous gene, over-expression being driven by the introduced transcriptional regulatory sequence.

[0212] Cells used in this invention can be derived from any eukaryotic species and can be primary, secondary, or immortalized. Furthermore, the cells can be derived from any tissue in the organism. Examples of useful tissues from which cells can be isolated and activated include, but are not limited to, liver, kidney, spleen, bone marrow, thymus, heart, muscle, lung, brain, testes, ovary, islet, intestinal, bone marrow, skin, bone, gall bladder, prostate, bladder, embryos, and the immune and hematopoietic systems. Cell types include fibroblast, epithelial, neuronal, stem, and follicular. However, any cell or cell type can be used to activate gene expression using this invention.

[0213] The methods can be carried out in any cell of eukaryotic origin, such as fungal, plant or animal. Preferred embodiments include vertebtrates and particularly mammals, and more particularly, humans.

[0214] The construct can be integrated into primary, secondary, or immortalized cells. Primary cells are cells that have been isolated from a vertebrate and have not been passaged. Secondary cells are primary cells that have been passaged, but are not immortalized. Immortalized cells are cell lines that can be passaged, apparently indefinitely.

[0215] In preferred embodiments, the cells are immortalized cell lines. Examples of immortalized cell lines include, but are not limited to, HT1080, HeLa, Jurkat, 293 cells, KB carcinoma, T84 colonic epithelial cell line, Raji, HEP G2 or Hep 3B hepatoma cell lines, A2058 melanoma, U937 lymphoma, and WI38 fibroblast cell line, somatic cell hybrids, and hybridomas.

[0216] Cells used in this invention can be derived from any eukaryotic species, including but not limited to mammalian cells (such as rat, mouse, bovine, porcine, sheep, goat, and human), avian cells, fish cells, amphibian cells, reptilian cells, plant cells, and yeast cells. Preferably, over-expression of an endogenous gene or gene product from a particular species is accomplished by activating gene expression in a cell from that species. For example, to overexpress endogenous human proteins, human cells are used. Similarly, to overexpress endogenous bovine proteins, for example bovine growth hormone, bovine cells are used.

[0217] The cells can be derived from any tissue in the eukaryotic organism. Examples of useful vertebrate tissues from which cells can be isolated and activated include, but are not limited to, liver, kidney, spleen, bone marrow, thymus, heart, muscle, lung, brain, immune system (including lymphatic), testes, ovary, islet, intestinal, stomach, bone marrow, skin, bone, gall bladder, prostate, bladder, zygotes, embryos, and hematopoietic tissue. Useful vertebrate cell types include, but are not limited to, fibroblasts, epithelial cells, neuronal cells, germ cells (i.e., spermatocytes/spermatid, spermatozoa and oocytes), stem cells, and follicular cells. Examples of plant tissues from which cells can be isolated and activated include, but are not limited to, leaf tissue, ovary tissue, stamen tissue, pistil tissue, root tissue, tubers, gametes, seeds, embryos, and the like. One of ordinary skill will appreciate, however, that any eukaryotic cell or cell type can be used to activate gene expression using the present invention.

[0218] Any of the cells produced by any of the methods described are useful for screening for expression of a desired gene product and for providing desired amounts of a gene product that is over-expressed in the cell. The cells can be isolated and cloned.

[0219] Cells produced by this method can be used to produce protein in vitro (e.g., for use as a protein therapeutic) or in vivo (e.g., for use in cell therapy).

[0220] Commercial growth and production conditions often vary from the conditions used to grow and prepare cells for analytical use (e.g., cloning, protein or nucleotide acid sequencing, raising antibodies, X-ray crystallography analysis, enzymatic analysis, and the like). Scale up of cells for growth in roller bottles involves increase in the surface area on which cells can attach. Microcarrier beads are, therefore, often added to increase the surface area for commercial growth. Scale up of cells in spinner culture may involve large increases in volume. Five liters or greater can be required for both microcarrier and spinner growth. Depending on the inherent potency (specific activity) of the protein of interest, the volume can be as low as 1-10 liters. 10-15 liters is more common. However, up to 50-100 liters may be necessary and volume can be as high as 10,000-15,000 liters. In some cases, higher volumes may be required. Cells can also be grown in large numbers of T flasks, for example 50-100.

[0221] Despite growth conditions, protein purification on a commercial scale can also vary considerably from purification for analytic purposes. Protein purification in a commercial practical context can be initially the mass equivalent of 10 liters of cells at approximately 10⁶ cells/ml. Cell mass equivalent to begin protein purification can also be as high as 10 liters of cells at up to 10⁸ or 10⁹ cells/ml. As one of ordinary skill will appreciate, however, a higher or lower initial cell mass equivalent may also be advantageously used in the present methods.

[0222] Another commercial growth condition, especially when the ultimate product is used clinically, is cell growth in serum-free medium, by which is intended medium containing no serum or not in amounts that are required for cell growth. This obviously avoids the undesired co-purification of toxic contaminants (e.g., viruses) or other types of contaminants, for example, proteins that would complicate purification. Serum-free media for growth of cells, commercial sources for such media, and methods for cultivation of cells in serum-free media, are well-known to those of ordinary skill in the art.

[0223] A single cell made by the methods described above can over-express a single gene or more than one gene. More than one gene can be activated by the integration of a single construct or by the integration of multiple constructs in the same cell (i.e., more than one type of construct). Therefore, a cell can contain only one type of vector construct or different types of constructs, each capable of activating an endogenous gene.

[0224] The invention is also directed to methods for making the cells described above by one or more of the following: introducing one or more of the vector constructs; allowing the introduced construct(s) to integrate into the genome of the cell by non-homologous recombination; allowing over-expression of one or more endogenous genes in the cell; and isolating and cloning the cell.
The term “transfection” has been used herein for convenience when discussing introducing a polynucleotide into a cell. However, it is to be understood that the specific use of this term has been applied to generally refer to the introduction of the polynucleotide into a cell and is also intended to refer to the introduction by other methods described herein such as electroporation, liposome-mediated introduction, retrovirus-mediated introduction, and the like (as well as according to its own specific meaning).

The vector can be introduced into the cell by a number of methods known in the art. These include, but are not limited to, electroporation, calcium phosphate precipitation, DEAE dextran, lipofection, and receptor mediated endocytosis, polybrene, particle bombardment, and microinjection. Alternatively, the vector can be delivered to the cell as a viral particle (either replication competent or deficient). Examples of viruses useful for the delivery of nucleic acid include, but are not limited to, adenoviruses, adeno-associated viruses, retroviruses, Herpesviruses, and vaccinia viruses. Other viruses suitable for delivery of nucleic acid molecules into cells that are known to one of ordinary skill may be equivalently used in the present methods.

Following transfection, the cells are cultured under conditions, as known in the art, suitable for nonhomologous integration between the vector and the host cell’s genome. Cells containing the nonhomologously integrated vector can be further cultured under conditions, as known in the art, allowing expression of activated endogenous genes.

The vector construct can be introduced into cells on a single DNA construct or on separate constructs and allowed to concatemerize.

Whereas in preferred embodiments, the vector construct is a double-stranded DNA vector construct, vector constructs also include single-stranded DNA, combinations of single- and double-stranded DNA, single-stranded RNA, double-stranded RNA, and combinations of single- and double-stranded RNA. Thus, for example, the vector construct could be single-stranded RNA which is converted to cDNA by reverse transcriptase, the cDNA converted to double-stranded DNA, and the double-stranded DNA ultimately recombining with the host cell genome.

In preferred embodiments, the constructs are linearized prior to introduction into the cell. Linearization of the activation construct creates free DNA ends capable of reacting with chromosomal ends during the integration process. In general, the construct is linearized downstream of the regulatory element (and exonic and splice donor sequences, if present). Linearization can be facilitated by, for example, placing a unique restriction site downstream of the regulatory sequences and treating the construct with the corresponding restriction enzyme prior to transfection. While not required, it is advantageous to place a “spacer” sequence between the linearization site and the proximal most functional element (e.g., the unpaired splice donor site) on the construct. When present, the spacer sequence protects the important functional elements on the vector from exonucleolytic degradation during the transfection process. The spacer can be composed of any nucleotide sequence that does not change the essential functions of the vector as described herein.

Circular constructs can also be used to activate endogenous gene expression. It is known in the art that circular plasmids, upon transfection into cells, can integrate into the host cell genome. Presumably, DNA breaks occur in the circular plasmid during the transfection process, thereby generating free DNA ends capable of joining to chromosome ends. Some of these breaks in the construct will occur in a location that does not destroy essential vector functions (e.g., the break will occur downstream of the regulatory sequence), and therefore, will allow the construct to be integrated into a chromosome in a configuration capable of activating an endogenous gene. As described above, spacer sequences may be placed on the construct (e.g., downstream of the regulatory sequences). During transfection, breaks that occur in the spacer region will create free ends at a site in the construct suitable for activation of an endogenous gene following integration into the host cell genome.

The invention also encompasses libraries of cells made by the above described methods. A library can encompass all of the clones from a single transfection experiment or a subset of clones from a single transfection experiment. The subset can over-express the same gene or more than one gene, for example, a class of genes. The transfection can have been done with a single type of construct or with more than one type of construct.

A library can also be formed by combining all of the recombinant cells from two or more transfection experiments, by combining one or more subsets of cells from a single transfection experiment or by combining subsets of cells from separate transfection experiments. The resulting library can express the same gene, or more than one gene, for example, a class of genes. Again, in each of these individual transfections, a unique construct or more than one construct can be used.

Libraries can be formed from the same cell type or different cell types.

The library can be composed of a single type of cell containing a single type of activation construct which has been integrated into chromosomes at spontaneous DNA breaks or at breaks generated by radiation, restriction enzymes, and/or DNA breaking agents, applied either together (to the same cells) or separately (applied to individual groups of cells and then combining the cells together to produce the library). The library can be composed of multiple types of cells containing a single or multiple constructs which were integrated into the genome of a cell treated with radiation, restriction enzymes, and/or DNA breaking agents, applied either together (to the same cells) or separately (applied to individual groups of cells and then combining the cells together to produce the library).

The invention is also directed to methods for making libraries by selecting various subsets of cells from the same or different transfection experiments. For example, all of the cells expressing nuclear factors (as determined by the presence of nuclear green fluorescent protein in cells transfected with construct 20) can be pooled to create a library of cells with activated nuclear factors. Similarly, cells expressing membrane or secreted proteins can be pooled. Cells can also be grouped by phenotype, for example, growth factor independent growth, growth factor independent proliferation, colony formation, cellular differentiation (e.g., differentiation into a neuronal cell, muscle cell, epithelial cell, etc.), anchorage independent growth, activation of cellular factors (e.g., kinases, transcription factors,
nucleases, etc.), gain or loss of cell-cell adhesion, migration, or cellular activation (e.g., resting versus activated T cells).

[0237] The invention is also directed to methods of using libraries of cells to over-express an endogenous gene. The library is screened for the expression of the gene and cells are selected that express the desired gene product. The cell can then be used to purify the gene product for subsequent use. Expression of the cell can occur by culturing the cell in vitro or by allowing the cell to express the gene in vivo.

[0238] The invention is also directed to methods of using libraries to identify novel gene and gene products.

[0239] The invention is also directed to methods for increasing the efficiency of gene activation by treating the cells with agents that stimulate or effect the patterns of non-homologous integration. It has been demonstrated that gene expression patterns, chromatin structure, and methylation patterns can differ dramatically from cell type to cell type. Even different cell lines from the same cell type can have significant differences. These differences can impact the patterns of non-homologous integration by affecting both the DNA breakage pattern and the repair process. For example, chromatinized stretches of DNA (characteristics likely associated with inactive genes) may be more resistant to breakage by restriction enzymes and chemical agents, whereas they may be susceptible to breakage by radiation.

[0240] Furthermore, inactive genes can be methylated. In this case, restriction enzymes that are blocked by CpG methylation will be unable to cleave methylated sites near the inactive gene, making it more difficult to activate that gene using methylation-sensitive enzymes. These problems can be circumvented by creating activation libraries in cell lines using a variety of DNA breakage agents. By doing this, a more complete integration pattern can be created and the probability of activating a given gene maximized.

[0241] The methods of the invention can include introducing double strand breaks into the DNA of the cell containing the endogenous gene to be over-expressed. These methods introduce double-strand breaks into the genomic DNA in the cell prior to or simultaneously with vector integration. The mechanism of DNA breakage can have a significant effect on the pattern of DNA breaks in the genome. As a result, DNA breaks produced spontaneously or artificially with radiation, restriction enzymes, bleomycin, or other breaking agents, can occur in different locations.

[0242] In order to increase integration efficiency and to improve the random distribution of integration sites, cells can be treated with low, intermediate, or high doses of radiation prior to or following transfection. By artificially inducing double strand breaks, the transplanted DNA can now integrate into the host cell chromosome as part of the DNA repair process. Normally, creation of double strand breaks to serve as the site of integration is the rate limiting step. Thus, by increasing chromosome breaks using radiation (or other DNA damaging agents), a larger number of integrants can be obtained in a given transfection. Furthermore, the mechanism of DNA breakage by radiation is different than by spontaneous breakage.

[0243] Radiation can induce DNA breaks directly when a high energy photon hits the DNA molecule. Alternatively, radiation can activate compounds in the cell which in turn, react with and break the DNA strand. Spontaneous breaks, on the other hand, are thought to occur by the interaction between reactive compounds produced in the cell (such as superoxides and peroxides) and the DNA molecule. However, DNA in the cell is not present as a naked, deproteinized polymer, but instead is bound to chromatin and present in a condensed state. As a result, some regions are not accessible to agents in the cell that cause double strand breaks. The photons produced by radiation have wave lengths short enough to hit highly condensed regions of DNA, thereby inducing breaks in DNA regions that are under represented in spontaneous breaks. Thus, radiation is capable of creating different DNA breakage patterns, which in turn, should lead to different integration patterns.

[0244] As a result, libraries produced using the same activation construct in cells with and without radiation treatment will potentially contain different sets of activated genes. Finally, radiation treatment increases efficiency of nonhomologous integration by up to 5-10 fold, allowing complete libraries to be created using fewer cells. Thus, radiation treatment increases the efficiency of gene activation and generates new integration and activation patterns in transfected cells. Useful types of radiation include α, β, γ, x-ray, and ultraviolet radiation. Useful doses of radiation vary for different cell types, but in general, dose ranges resulting in cell viabilities of 0.1% to >99% are useful. For HT1080 cells, this corresponds to radiation doses from a 137 CS source of approximately 0.1 rads to 1000 rads. Other doses may also be useful as long as the dose either increases the integration frequency or changes the pattern of integration sites.

[0245] In addition to radiation, restriction enzymes can be used to artificially induce chromosome breaks in transfected cells. As with radiation, DNA restriction enzymes can create chromosome breaks which, in turn, serve as integration sites for the transfected DNA. This larger number of DNA breaks increases the overall efficiency of integration of the activation construct. Furthermore, the mechanism of breakage by restriction enzymes differs from that by radiation, the pattern of chromosome breaks is also likely to be different.

[0246] Restriction enzymes are relatively large molecules compared to photons and small metabolites capable of damaging DNA. As a result, restriction enzymes will tend to break regions that are less condensed then the genome as a whole. If the gene of interest lies within an accessible region of the genome, then treatment of the cells with a restriction enzyme can increase the probability of integrating the activation construct upstream of the gene of interest. Since restriction enzymes recognize specific sequences, and since a given restriction site may not lie upstream of the gene of interest, a variety of restriction enzymes can be used. It may also be important to use a variety of restriction enzymes since each enzyme has different properties (e.g., size, stability, ability to cleave methylated sites, and optimal reaction conditions) that affect which sites in the host chromosome will be cleaved. Each enzyme, due to the different distribution of cleavable restriction sites, will create a different integration pattern.

[0247] Therefore, introduction of restriction enzymes (or plasmids capable of expressing restriction enzymes) before, during, or after introduction of the activation construct will result in the activation of different sets of genes. Finally,
Restriction enzyme-induced breaks increase the integration efficiency by up to 5-10 fold (Yorifuji et al., Mut. Res. 243:121 (1990)), allowing fewer cells to be transfected to produce a complete library. Thus, restriction enzymes can be used to create new integration patterns, allowing activation of genes which failed to be activated in libraries produced by non-homologous recombination at spontaneous breaks or at other artificially induced breaks.

Restriction enzymes can also be used to bias integration of the activation construct to a desired site in the genome. For example, several rare restriction enzymes have been described which cleave eukaryotic DNA every 50-10000 kilobases, on average. If a rare restriction recognition sequence happens to be located upstream of a gene of interest, by introducing the restriction enzyme at the time of transfection along with the activation construct, DNA breaks can be preferentially upstream of the gene of interest. These breaks can then serve as sites for integration of the activation construct. Any enzyme can be that cleaves in an appropriate location in or near the gene of interest and its site is under-represented in the rest of the genome or its site is over-represented near genes (e.g., restriction sites containing CpG). For genes that have not been previously identified, restriction enzymes with 8 bp recognition sites (e.g., NolI, SfiI, Pmel, Swal, Ssel, Srl, SgrAI, Paci, Ascl, Sgfl, and Sse837I), enzymes recognizing CpG containing sites (e.g., Eagl, Bsi-WI, MluI, and BssHII) and other rare cutting enzymes can be used.

In this way, “biased” libraries can be created which are enriched for certain types of activated genes. In this respect, restriction enzyme sites containing CpG dinucleotides are particularly useful since these sites are under-represented in the genome at large, but over-represented in the form of CpG islands at the 5' end of many genes, the very location that is useful for gene activation. Enzymes recognizing these sites, therefore, will preferentially cleave at the 5' end of genomic sequences.

Restriction enzymes can be introduced into the host cell by several methods. First, restriction enzymes can be introduced into the cell by electroporation (Yorifuji et al., Mut. Res. 243:121 (1990); Winager et al., Mut. Res. 225:49 (1989)). In general, the amount of restriction enzyme introduced into the cell is proportional to its concentration in the electroporation media. The pulse conditions must be optimized for each cell line by adjusting the voltage, capacitance, and resistance. Second, the restriction enzyme can be expressed transiently from a plasmid encoding the enzyme under the control of eukaryotic regulatory elements. The level of enzyme produced can be controlled by using inducible promoters, and varying the strength of induction. In some cases, it may be desirable to limit the amount of restriction enzyme produced (due to its toxicity). In these cases, weak or mutant promoters, splice sites, translation start codons, and poly(A) tails can be utilized to lower the amount of restriction enzyme produced. Third, restriction enzymes can be introduced by agents that fuse with or permeabilize the cell membrane. Liposomes and streptolysin O (Pimplikar et al., J. Cell Biol. 125:1025 (1994)) are examples of this type of agent. Finally, mechanical perforation (Beckers et al., Cell 50:523-534 (1987)) and microinjection can also be used to introduce nucleases and other proteins into cells. However, any method capable of delivering active enzymes to a living cell is suitable.

DNA breaks induced by bleomycin and other DNA damaging agents can also produce DNA breakage patterns that are different. Thus, any agent or incubation condition capable of generating double strand breaks in cells is useful for increasing the efficiency and/or altering the sites of non-homologous recombination. Examples of classes of chemical DNA breaking agents include, but are not limited to, peroxides and other free radical generating compounds, alkylating agents, topoisomerase inhibitors, anti-neoplastic drugs, acids, substituted nucleotides, and enediyne antibiotics.

Specific chemical DNA breaking agents include, but are not limited to, bleomycin, hydrogen peroxide, cumene hydroperoxide, tert-butyl hydroperoxide, hypochlorous acid (reacted with amine, 1-naphthylamine or 1-naphthol), nitric acid, phosphoric acid, dioxidobacitin, 9-deoxydoxorubicin, demethyl-6-deoxyxurubcin, 5-iminodaunorubicin, adriamycin, 4 °-(9-acridinylamino)ethanesulfon-m-anisidide, neocarzinostatin, 8-methoxycaline, etoposide, ellipticine, iododeoxyuridine, and bromodeoxyuridine.

It has been shown that DNA repair machinery in the cell can be induced by pre-exposing the cell to low doses of a DNA breaking agent such as radiation or bleomycin. By pretreating cells with these agents approximately 24 hours prior to transfection, the cell will be more efficient at repairing DNA breaks and integrating DNA following transfection. In addition, higher doses of radiation or other DNA breaking agents can be used since the L50 (the dose that results in lethality in 50% of the exposed cells) is higher following pretreatment. This allows random activation libraries to be created at multiple doses and results in a different distribution of integration sites within the host cell's chromosomes.

Once an activation library (or libraries) is created, it can be screened using a number of assays. Depending on the characteristics of the protein(s) of interest (e.g., secreted versus intracellular proteins) and the nature of the activation construct used to create the library, any or all of the assays described below can be utilized. Other assay formats can also be used.

ELISA. Activated proteins can be detected using the enzyme-linked immunosorbent assay (ELISA). If the activated gene product is secreted, culture supernatants from pools of activation library cells are incubated in wells containing bound antibody specific for the protein of interest. If a cell or group of cells has activated the gene of interest, then the protein will be secreted into the culture media. By screening pools of library clones (the pools can be from 1 to greater than 100,000 library members), pools containing a cell(s) that has activated the gene of interest can be identified. The cell of interest can then be purified away from the other library members by screening, limiting dilution, or other techniques known in the art. In addition to secreted proteins, ELISA can be used to screen for cells expressing intracellular and membrane-bound proteins. In these cases, instead of screening culture supernatants, a small number of cells is removed from the library pool (each cell is represented at least 100-1000 times in each pool), lysed, clarified, and added to the antibody-coated wells.

ELISA SpotAssay. ELISA spot are coated with antibodies specific for the protein of interest. Following
coating, the wells are blocked with 1% BSA/PBS for 1 hour at 37°C. Following blocking, 100,000 to 500,000 cells from the random activation library are applied to each well (representing ~10% of the total pool). In general, one pool is applied to each well. If the frequency of a cell expressing the protein of interest is 1 in 10,000 (i.e., the pool consists of 10,000 individual clones, one of which expresses the protein of interest), then plating 500,000 cells per well will yield 50 specific cells. Cells are incubated in the wells at 37°C for 24 to 48 hours without being moved or disturbed. At the end of the incubation, the cells are removed and the plate is washed 3 times with PBS/0.05% Tween 20 and 3 times with PBS/1% BSA. Secondary antibodies are applied to the wells at the appropriate concentration and incubated for 2 hours at room temperature or 16 hours at 4°C. These antibodies can be biotinylated or labeled directly with horse-radish peroxidase (HRP). The secondary antibodies are removed and the plate is washed with PBS/1% BSA. The tertiary antibody or streptavidin labeled with HRP is added and incubated for 1 hour at room temperature.

[0257] FACS assay. The fluorescence-activated cell sorter (FACS) can be used to screen the random activation library in a number of ways. If the gene of interest encodes a cell surface protein, then fluorescently-labeled antibodies are incubated with cells from the activation library. If the gene of interest encodes a secreted protein, then cells can be biotinylated and incubated with streptavidin conjugated to an antibody specific to the protein of interest (Manz et al., Proc. Natl. Acad. Sci. (USA) 92:1921 (1995)). Following incubation, the cells are placed in a high concentration of gelatin (or other polymer such as agarose or methycellulose) to limit diffusion of the secreted protein. As protein is secreted by the cell, it is captured by the antibody bound to the cell surface. The presence of the protein of interest is then detected by a second antibody which is fluorescently labeled. For both secreted and membrane bound proteins, the cells can then be sorted according to their fluorescence signal. Fluorescent cells can then be isolated, expanded, and further enriched by FACS, limiting dilution, or other cell purification techniques known in the art.

[0258] Magnetic Bead Separation. The principle of this technique is similar to FACS. Membrane bound proteins and captured secreted proteins (as described above) are detected by incubating the activation library with an antibody-conjugated magnetic beads that are specific for the protein of interest. If the protein is present on the surface of a cell, the magnetic beads will bind to that cell. Using a magnet, the cells expressing the protein of interest can be purified away from the other cells in the library. The cells are then released from the beads, expanded, analyzed, and further purified if necessary.

[0259] RT-PCR. A small number of cells (equivalent to at least the number of individual clones in the pool) is harvested and lysed to allow purification of the RNA. Following isolation, the RNA is reversed-transcribed using reverse transcriptase. PCR is then carried out using primers specific for the cDNA of the gene of interest.

[0260] Alternatively, primers can be used that span the synthetic exon in the activation construct and the exon of the endogenous gene. This primer will not hybridize to and amplify the endogenously expressed gene of interest. Conversely, if the activation construct has integrated upstream of the gene of interest and activated gene expression, then this primer, in conjunction with a second primer specific for the gene will amplify the activated gene by virtue of the presence of the synthetic exon spliced onto the exon from the endogenous gene. Thus, this method can be used to detect activated genes in cells that normally express the gene of interest at lower than desired levels.

[0261] Phenotypic Section. In this embodiment, cells can be selected based on a phenotype conferred by the activated gene. Examples of phenotypes that can be selected for include proliferation, growth factor independent growth, colony formation, cellular differentiation (e.g., differentiation into a neuronal cell, muscle cell, epithelial cell, etc.), anchorage independent growth, activation of cellular factors (e.g., kinases, transcription factors, nucleases, etc.), gain or loss of cell-cell adhesion, migration, and cellular activation (e.g., resting versus activated T cells). Isolation of activated cells demonstrating a phenotype, such as those described above, is important because the activation of an endogenous gene by the integrated construct is presumably responsible for the observed cellular phenotype. Thus, the activated gene may be an important therapeutic drug or drug target for treating or inducing the observed phenotype.

[0262] The sensitivity of each of the above assays can be effectively increased by transiently upregulating gene expression in the library cells. This can be accomplished for NF-κB site-containing promoters (on the activation construct) by adding PMA and tumor necrosis factor-α to the library. Separately, or in conjunction with PMA and TNF-α, sodium butyrate can be added to further enhance gene expression. Addition of these reagents can increase expression of the protein of interest, thereby allowing a lower sensitivity assay to be used to identify the gene activated cell of interest.

[0263] Since large activation libraries are created to maximize activation of many genes, it is advantageous to organize the library clones in pools. Each pool can consist of 1 to greater than 100,000 individual clones. Thus, in a given pool, many activated proteins are produced, often in dilute concentrations (due to the overall size of the pool and the limited number of cells within the pool that produce a given activated protein). Thus, concentration of the proteins prior to screening effectively increases the ability to detect the activated proteins in the screening assay. One particularly useful method of concentration is ultrafiltration; however, other methods can also be used. For example, proteins can be concentrated non-specifically, or semi-specifically by adsorption onto ion exchange, hydrophobic, dye, hydroxyapatite, lectin, and other suitable resins under conditions that bind most or all proteins present. The bound proteins can then be removed in a small volume prior to screening. It is advantageous to grow the cells in serum free media to facilitate the concentration of proteins.

[0264] In another embodiment, a useful sequence that can be included on the activation construct is an epitope tag. The epitope tag can consist of an amino acid sequence that allows affinity purification of the activated protein (e.g., on immunoaffinity or chelating matrices). Thus, by including an epitope tag on the activation construct, all of the activated proteins from an activation library can be purified. By purifying the activated proteins away from other cellular and media proteins, screening for novel proteins and enzyme
activities can be facilitated. In some instances, it may be desirable to remove the epitope tag following purification of the activated protein. This can be accomplished by including a protease recognition sequence (e.g., Factor IIa or enterokinase cleavage site) downstream from the epitope tag on the activation construct. Incubation of the purified, activated protein(s) with the appropriate protease will release the epitope tag from the protein(s).

[0265] In libraries in which an epitope tag sequence is located on the activation construct, all of the activated proteins can be purified away from all other cellular and media proteins using affinity purification. This not only concentrates the activated proteins, but also purifies them away from other activities that can interfere with the assay used to screen the library.

[0266] Once a pool of clones containing cells over-expressing the gene of interest is identified, steps can be taken to isolate the activated cell. Isolation of the activated cell can be accomplished by a variety of methods known in the art. Examples of cell purification methods include limiting dilution, fluorescence activated cell sorting, magnetic bead separation, sib selection, and single colony purification using cloning rings.

[0267] In preferred embodiments of the invention, the methods include a process wherein the expression product is purified. In highly preferred embodiments, the cells expressing the endogenous gene product are cultured so as to produce amounts of gene product feasible for commercial application, and especially diagnostic and therapeutic and drug discovery uses.

[0268] Any vector used in the methods described herein can include an amplifiable marker. Thereby, amplification of both the vector and the DNA of interest (i.e., containing the over-expressed gene) occurs in the cell, and further enhanced expression of the endogenous gene is obtained. Accordingly, methods can include a step in which the endogenous gene is amplified.

[0269] Once the activated cell has been isolated, expression can be further increased by amplifying the locus containing both the gene of interest and the activation construct. This can be accomplished by each of the methods described below, either separately or in combination.

[0270] Amplifiable markers are genes that can be selected for higher copy number. Examples of amplifiable markers include dihydrofolate reductase, adenosine deaminase, aspartate transcarbamylase, dihydro- orotase, and carbamyl phosphate synthase. For these examples, the elevated copy number of the amplifiable marker and flanking sequences (including the gene of interest) can be selected for using a drug or toxic metabolite which is acted upon by the amplifiable marker. In general, as the drug or toxic metabolite concentration increases, cells containing fewer copies of the amplifiable marker die, whereas cells containing increased copies of the marker survive and form colonies. These colonies can be isolated, expanded, and analyzed for increased levels of production of the gene of interest.

[0271] Placement of an amplifiable marker on the activation construct results in the juxtaposition of the gene of interest and the amplifiable marker in the activated cell. Selection for activated cells containing increased copy number of the amplifiable marker and gene of interest can be achieved by growing the cells in the presence of increasing amounts of selective agent (usually a drug or metabolite). For example, amplification of dihydrofolate reductase (DHFR) can be selected using methotrexate.

[0272] As drug-resistant colonies arise at each increasing drug concentration, individual colonies can be selected and characterized for copy number of the amplifiable marker and gene of interest, and analyzed for expression of the gene of interest. Individual colonies with the highest levels of activated gene expression can be selected for further amplification in higher drug concentrations. At the highest drug concentrations, the clones will express greatly increased amounts of the protein of interest.

[0273] When amplifying DHFR, it is convenient to plate approximately 1×10⁷ cells at several different concentrations of methotrexate. Useful initial concentrations of methotrexate range from approximately 5 nM to 100 nM. However, the optimal concentration of methotrexate must be determined empirically for each cell line and integration site. Following growth in methotrexate containing media, colonies from the highest concentration of methotrexate are picked and analyzed for increased expression of the gene of interest. The clone(s) with the highest concentration of methotrexate are then grown in higher concentrations of methotrexate to select for further amplification of DHFR and the gene of interest. Methotrexate concentrations in the micromolar and millimolar range can be used for clones containing the highest degree of gene amplification.

[0274] Placement of a viral origin of replication(s) (e.g., ori P or SV40 in human cells, and polyoma ori in mouse cells) on the activation construct will result in the juxtaposition of the gene of interest and the viral origin of replication in the activated cell. The origin and flanking sequences can then be amplified by introducing the viral replication protein(s) in trans. For example, when ori P (the origin of replication on Epstein-Barr virus) is utilized, EBNA-1 can be expressed transiently or stably. EBNA-1 will initiate replication from the integrated ori P locus. The replication will extend from the origin bi-directionally. As each replication product is created, it too can initiate replication. As a result, many copies of the viral origin and flanking genomic sequences including the gene of interest are created. This higher copy number allows the cells to produce larger amounts of the gene of interest.

[0275] At some frequency, the replication product will recombine to form a circular molecule containing flanking genomic sequences, including the gene of interest. Cells that contain circular molecules with the gene of interest can be isolated by single cell cloning and analysis of Hirt extraction and Southern blotting. Once purified, the cell containing the episomal genomic locus at elevated copy number (typically 10-50 copies) can be propagated in culture. To achieve higher amplification, the episome can be further boosted by including a second origin adjacent to the first in the original construct. For example, T antigen can be used to boost the copy number of ori P/SV40 episomes to a copy number of ~1000 (Heinzell et al., J. Virol. 62:3738 (1988)). This substantial increase in copy number can dramatically increase protein expression.

[0276] The invention encompasses over-expression of endogenous genes both in vivo and in vitro. Therefore, the
cells could be used in vitro to produce desired amounts of a gene product or could be used in vivo to provide that gene product in the intact animal.

[0277] The invention also encompasses the proteins produced by the methods described herein. The proteins can be produced from either known, or previously unknown genes. Examples of known proteins that can be produced by this method include, but are not limited to, erythropoietin, insulin, growth hormone, glucocerebrosidase, tissue plasminogen activator, granulocyte-macrophage colony stimulating factor, granulocyte/macrophage colony stimulating factor, interferon α, interferon β, interferon γ, interleukin-2, interleukin-6, interleukin-11, interleukin-12, TGF β, blood clotting factor V, blood clotting factor VII, blood clotting factor VII, blood clotting factor IX, blood clotting factor X, TSH-β, bone growth factor 2, bone growth factor 7, tumor necrosis factor, alpha-1 antitrypsin, anti-thrombin III, leukemia inhibitory factor, glucagon, Protein C, protein kinase C, macrophage colony stimulating factor, stem cell factor, follicle stimulating hormone β, uricase, nerve growth factors, insulin-like growth factors, insulin receptors, parathyroid hormone, lactoferrin, complement inhibitors, platelet derived growth factor, keratinocyte growth factor, neutrophin-3, thrombopoietin, choric gonadotropin, thrombomodulin, alpha glucosidase, epidermal growth factor, FGF, macrophage colony stimulating factor, and cell surface receptors for each of the above-described proteins.

[0278] Where the protein product from the activated cell is purified, any method of protein purification known in the art may be employed.

Isolation of Cells Containing Activated Membrane Protein-Encoding Genes

[0279] Genes that encode membrane associated proteins are particularly interesting from a drug development standpoint. These genes and the proteins they encode can be used, for example, to develop small molecule drugs using combinatorial chemistry libraries and high-throughput screening assays. Alternatively, the proteins or soluble forms of the proteins (e.g., truncated proteins lacking the transmembrane region) can be used as therapeutically active agents in humans or animals. Identification of membrane proteins can also be used to identify new ligands (e.g., cytokines, growth factors, and other effector molecules) using two hybrid approaches or affinity capture techniques. Many other uses of membrane proteins are also possible.

[0280] Current approaches to identifying genes that encode integral membrane proteins involve isolation and sequencing of genes from cDNA libraries. Integral membrane proteins are then identified by ORF analysis using hydrophobicity plots capable of identifying the transmembrane region of the protein. Unfortunately, using this approach a gene encoding an integral membrane protein cannot be identified unless the gene is expressed in the cells used to produce the cDNA library. Furthermore, many genes are only expressed in very rare cells, during short developmental windows, and/or at very low levels. As a result, these genes can not be efficiently identified using the currently available approaches.

[0281] The present invention allows endogenous genes to be activated without any knowledge of the sequence, structure, function, or expression profile of the genes. Using the disclosed methods, genes may be activated at the transcription level only, or at both the transcription and translation levels. As a result, proteins encoded by the activated endogenous gene can be produced in cells containing the integrated vector. Furthermore, using specific vectors disclosed herein, the protein produced from the activated endogenous gene can be modified, for example, to include an epitope tag. Other vectors (e.g., vectors 12-17 described above) may encode a signal peptide followed by an epitope tag. This vector can be used to isolate cells that have activated expression of an integral membrane protein (see Example 5 below). This vector can also be used to direct secretion of proteins that are not normally secreted.

[0282] Thus, the invention also is directed to methods for identifying an endogenous gene encoding a cellular integral membrane protein or a transmembrane protein. Such methods of the invention may comprise one or more steps. For example, one such method of the invention may comprise (a) introducing one or more vectors of the invention into a cell; (b) allowing the vector to integrate into the genome of the cell by non-homologous recombination; (c) allowing over-expression of an endogenous gene in the cell by upregulation of the gene by the transcriptional regulatory sequence contained on the integrated vector construct; (d) screening the cell for over-expression of the endogenous gene; and (e) characterizing the activated gene to determine its identity as a gene encoding a cellular integral membrane protein. In related embodiments, the invention provides such methods further comprising isolating the activated gene from the cell prior to characterizing the activated gene.

[0283] To identify genes that encode integral membrane proteins, vectors integrated into the genome of cells will comprise a regulatory sequence linked to an exonic sequence containing a start codon, a signal sequence, and an epitope tag, followed by an unpaired splice donor site. Upon integration and activation of an endogenous gene, a chimeric protein is produced containing the signal peptide and epitope tag from the vector fused to the protein encoded by the downstream exons of the endogenous gene. This chimeric protein, by virtue of the presence of the vector encoded signal peptide, is directed to the secretory pathway where translation of the protein is completed and the protein is secreted. If, however, the activated endogenous gene encodes an integral membrane protein, and the transmembrane region of that gene is encoded by exons located 3' of the vector integration site, then the chimeric protein will go to the cell surface, and the epitope tag will be displayed on the cell surface. Using known methods of cell isolation (for example flow cytometric sorting, magnetic bead cell sorting, immunoadsorption, or other methods that will be familiar to one of ordinary skill in the art), antibodies to the epitope tag can then be used to isolate the cells from the population that display the epitope tag and have activated an integral membrane encoding gene. These cells can then be used to study the function of the membrane protein. Alternatively, the activated gene may then be isolated from these cells using any art-known method, e.g., through hybridization with a DNA probe specific to the vector-encoded exon to screen a cDNA library produced from these cells, or using the genetic constructs described herein.

[0284] The epitope tag encoded by the vector exon may be a short peptide capable of binding to an antibody, a short peptide capable of binding to a substance (e.g., poly histi-
dine/divalent metal ion supports, maltose binding protein/maltose supports, glutathione S-transferase/glutathione support), or an extracellular domain (lacking a transmembrane domain) from an integral membrane protein for which an antibody or ligand exists. It will be understood, however, that other types of epitope tags that are familiar to one of ordinary skill in the art may be used equivalently in accordance with the invention.

Vectors for Non-targeted Activation of Endogenous Genes

[0285] As noted above, non-targeted gene activation has a number of important applications, including activating endogenous genes in host cells which provides a powerful approach to discovering and isolating new genes and proteins, and to producing large amounts of specific proteins for commercialization. For some applications of non-targeted gene activation, it is desirable to create libraries of cells in which each member of the library contains an activation vector integrated into a unique location in the host cell genome, and in which each member of the library has activated a different endogenous gene. Furthermore, it would be desirable to remove cells from the library that contain an integrated vector, but fail to activate an endogenous gene. Since eukaryotic genomes often contain large regions that lack genes, integration of an activation vector into a region devoid of genes can occur frequently. These integrated vectors, however, fail to activate an endogenous gene, and yet are capable of conferring drug resistance on the host cells when a selectable marker (driven by a suitable promoter and followed by a polyadenylation signal) is included on the activation vector. Even more problematic for gene discovery applications, a transcript containing vector sequences is produced in these cells regardless of whether or not a gene has been activated. In cases where a gene has not been activated, these vector sequence-containing transcripts contain non-genic genomic DNA sequences. As a result, when isolating activated genes, one cannot isolate all RNA (or cDNA) molecules that are derived from the integrated vector (i.e. transcripts containing vector sequences), since many of these transcripts do not encode an endogenous gene. To overcome these difficulties, the present invention provides highly specific vectors and methods that facilitate isolation of vector-activated genes.

[0286] These vectors of the invention are useful for activating expression of endogenous genes and for isolating the mRNA and cDNA corresponding to the activated genes. One such vector reduces the number of cells in which the vector integrated into the genome but failed to activate expression from (or transcription through) an endogenous gene. By removing these cells, fewer library members can be created and screened to isolate a given number of activated genes. Furthermore, vector-containing cells that fail to activate gene expression produce an RNA molecule that can interfere with isolation of bona fide activated genes. Thus, the vectors disclosed herein are particularly useful for producing cells suitable for protein over-expression and/or for isolating cDNA molecules corresponding to activated genes. The second type of vector of the invention is useful for isolating exon I from activated endogenous genes. As a result, these vectors can be used to obtain full-length genes from activated RNA transcripts. Each of the functional vector components described herein may be used separately, or in combination with each other.

Poly(A) Trap Activation Vectors

[0287] To facilitate isolation of activated genes, the present invention provides novel gene activation vectors that are capable of producing a drug resistant colony, preferentially upon activation of an endogenous gene. Such vectors are referred to herein as “poly(A) trap vectors.” Examples of poly(A) trap vectors are shown in FIG. 8A-8E. The nucleotide sequence of one such dual poly(A) trap vector, designated pRIG21b, is shown in FIG. 15A-15B (SEQ ID NO:19). These vectors contain a transcriptional regulatory sequence (which may be any transcriptional regulatory sequence, including but not limited to the promoters, enhancers, and repressors described herein, and which preferably is a promoter or an enhancer, and most preferably a promoter such as a CMV immediate early gene promoter, an SV40 T antigen promoter, a tetracycline-inducible promoter, or a β-actin promoter) operably linked to a selectable marker gene lacking a poly(A) signal. Since the selectable marker gene lacks a polyadenylation signal, its message will not be stable, and the marker gene product will not be efficiently produced. However, if the activation vector integrates upstream of an endogenous gene, the selectable marker can utilize the polyadenylation signal of the endogenous gene, thereby allowing production of the selectable marker protein in sufficient amounts to confer drug resistance. Thus, cells that integrate this activation vector generally form a drug resistant colony only if an endogenous gene has been activated.

[0288] The poly(A) trap activation vectors can include any selectable or screenable marker. Furthermore, the selectable marker can be expressed from any promoter that is functional in the cells used to create the integration library. Thus, the selectable marker can be expressed by viral or non-viral promoters. Optionally, an unpaired splice donor site may be included in the construct, preferably 3' of the selectable marker to allow the exon encoding the selectable marker to be spliced directly to the exons of the endogenous gene. When a downstream transcriptional regulatory sequence and a splice donor site is included on the vector, the inclusion of a splice donor site adjacent to the selectable marker results in the removal of these downstream elements from the messenger RNA.

[0289] In a related embodiment, a second transcriptional regulatory sequence (which may be any transcriptional regulatory sequence, including but not limited to the promoters, enhancers, and repressors described herein, and which preferably is a promoter or an enhancer, and most preferably a promoter) may be located downstream of, and in the same orientation as, the selectable marker. Optionally, an unpaired splice donor site may be linked to the downstream transcriptional regulatory sequence. In this configuration, the poly(A) trap vector is capable of producing a message containing the downstream vector-encoded exon spliced to endogenous exons. As described below, these chimeric transcripts can be translated into native or modified protein, depending on the nature of the vector-encoded exon.

[0290] As used herein, a “vector-encoded exon” means a region of a vector downstream of the transcriptional regulatory sequence and between the transcription start site and the unpaired splice donor site found on the vector. The vector-encoded exon is present at the 5' end of the transcript containing the endogenous gene in the fully processed
message. Analogously, as used herein, a “vector-encoded intron” is the region of the vector located downstream of the unpaired splice donor site. When a linearization site is present on the vector, the vector-encoded intron is the region of the vector that is downstream of the vector-encoded exon between the unpaired splice donor site and the linearization site. The vector-encoded intron is removed from the activated gene transcript during RNA processing.

Splice Acceptor Trap (SAT) Vectors

[0291] As an alternative approach for removing cells that fail to activate an endogenous gene, the invention provides additional vectors designated herein as “Splice Acceptor Trap” (SAT) vectors. These vectors are designed to splice from a vector encoded splice donor site to an endogenous splice acceptor. Furthermore, the vectors are designed to produce a product that is toxic to the host cells (or a product that can be selected against) if splicing does not occur. Thus, these vectors facilitate elimination of cells in which the vector-encoded exon failed to splice to an endogenous exon.

[0292] The splice acceptor trap vectors can contain both a positive selectable marker and a negative selectable marker gene oriented in the same direction on the vector. As used herein, a positive selectable marker is a gene that, upon expression, produces a protein capable of facilitating the isolation of cells expressing the marker. Analogously, as used herein, a negative selectable marker is a gene that, upon expression, produces a protein capable of facilitating removal of cells expressing the marker.

[0293] The positive selectable marker and the negative selectable marker are preferably separated in the vector construct by an unpaired splice donor site. In other embodiments, however, the positive selectable marker may be fused to the negative selectable marker gene. In this configuration, an unpaired splice donor site is located between the positive and negative selectable marker, such that the reading frame of the negative selectable marker is preserved. The unpaired splice donor site is preferably located at the junction of the positive and negative selectable markers. However, the unpaired splice donor site may be located anywhere in the fusion gene such that upon splicing to an endogenous splice acceptor site, the positive selectable marker will be expressed in an active form and the negative selectable marker will be expressed in an inactive form, or not at all. In this configuration, the positive selectable marker is located upstream of the negative selectable marker.

[0294] It will also be apparent to one of ordinary skill in view of the description contained herein that the positive and negative selectable markers on the SAT vector need not be expressed as a fusion protein. In one embodiment, an internal ribosomal entry site (IRES) is inserted between the positive selectable marker and the negative selectable marker. In this configuration, the unpaired splice donor site can be positioned between the two markers, or in the open reading frame of either marker gene such that, upon splicing, the positive selectable marker will be expressed in an active form and the negative selectable marker will be expressed in an inactive form, or not at all. In another embodiment, the positive selectable marker may be driven from a different transcriptional regulatory sequence than the negative selectable marker. In this configuration, the unpaired splice donor site is located in the 5' untranslated region of the negative selectable marker or anywhere in the open reading frame of the negative selectable marker such that, upon splicing, the negative selectable marker will be produced in an inactive form or not at all. Furthermore, when the positive and negative markers are driven from different transcriptional regulatory sequences, the positive selectable marker may be located upstream or downstream of the negative selectable marker, and the positive selectable marker may contain or lack a splice donor site at its 3' end.

[0295] The vectors described herein may contain any positive selectable marker. Examples of positive selectable markers useful in this invention include genes encoding neomycin (neo), hypoxanthine phosphoribosyl transferase (HPRT), puromycin (pac), dihydro-orotase, glutamine synthetase (GS), histidine D (his D), carbamyl phosphate synthase (CAD), dihydrofolate reductase (DHFR), multidrug resistance 1 (mdr1), aspartate transcarbamylase, xanthine-guanine phosphoribosyl transferase (gpt), and adenozine deaminase (ada). Alternatively, the vectors may contain a screenable marker in place of the positive selectable marker. Screenable markers include any protein capable of producing a recognizable phenotype in the host cell. Examples of screenable markers included cell surface epitopes (such as CD2) and enzymes (such as β-galactosidase).

[0296] The vectors described herein may also, or alternatively, contain any negative selectable marker that can be selected against. Examples of negative selectable markers include hypoxanthine phosphoribosyl transferase (HPRT), thymidine kinase (TK), and diphtheria toxin. The negative selectable marker can also be a screenable marker, such as a cell surface protein or an enzyme. Cells expressing the negative screenable marker may be removed by, for example, Fluorescence Activated Cell Sorting (FACS) or magnetic bead cell sorting.

[0297] To isolate cells that have activated expression of an endogenous gene, the cells containing the integrated vector can be placed under the appropriate drug selection. Selection for the positive selectable marker and against the negative selectable marker can occur simultaneously. In another embodiment, selection can occur sequentially. When selection occurs sequentially, selection for the positive selectable marker can occur first, followed by selection against the negative selectable marker. Alternatively, selection against the negative selectable marker can occur first, followed by selection for the positive selectable marker.

[0298] The positive and negative markers are expressed by a transcriptional regulatory element located upstream of the translation start site of each gene. When a positive/negative marker fusion gene or an ires sequence is used, a single transcriptional regulatory element drives expression of both markers. A poly(A) signal may be placed 3' of each selectable marker. If a positive/negative fusion gene is used a single poly(A) signal is positioned 3' of the markers. Alternatively, a poly(A) signal may be excluded from the vector to provide additional specificity for a gene activation event (see dual poly(A)/splice acceptor trap below).

Dual Poly(A)/Splice Acceptor Trap Vectors

[0299] To further reduce the number of cells that lack a gene activation event, the invention also provides vectors that confers host cell survival only if the vector-encoded exon has spliced to an exon from an endogenous gene and has acquired a poly(A) signal. These vectors are designated
herein as “dual poly(A)/splice acceptor trap vectors” or as “dual poly(A)/SAT vectors.” By requiring both splicing and polyadenylation to occur for cell survival, cells that fail to activate an endogenous gene are more efficiently eliminated from the activation library.

[0300] The dual poly(A)/splice acceptor trap vectors contain a positive selectable marker and a negative selectable marker configured as described for the SAT vectors; however, neither gene contains a functional poly(A) signal. Thus, the positive selectable marker will not be expressed at high levels unless splicing occurs to capture an endogenous poly(A) signal. Aside from the lack of a poly(A) signal, all other features and embodiments of this type of vector are the same as those of the SAT vectors as described herein. Examples of dual poly(A)/SAT vectors are shown in FIGS. 9A-9F and 10A-10F. The nucleotide sequence of one such dual poly(A)/SAT vector, designated pRIG22b, is shown in FIG. 16A-16D (SEQ ID NO:20).

Vectors for Activating Protein Expression from Endogenous Genes

[0301] In many applications of non-targeted gene activation, it is desirable to produce protein from the activated endogenous gene. To accomplish this, a second transcriptional regulatory sequence (which may be any transcriptional regulatory sequence, including but not limited to the promoters, enhancers, and repressors described herein, and which is preferably a promoter or an enhancer, and most preferably a promoter) can be placed downstream of the selectable marker(s) on any of the vectors described herein. When poly(A) trap vectors, SAT vectors, or dual poly(A) trap/SAT vectors are used, the downstream transcriptional regulatory sequence is positioned to drive expression in the same direction as the upstream selectable marker(s). To activate expression of full-length protein with this type of vector, however, the vector must integrate into the 5' UTR of the endogenous gene to avoid cryptic start ATG codons upstream of exon I.

[0302] Alternatively, to increase the frequency of protein expression using non-targeted gene activation, the downstream transcriptional regulatory sequence on the vector may be operably linked to an exonic sequence followed by a splice donor site. In a preferred embodiment, the vector exon lacks a start codon. This vector is particularly useful for activating protein expression from genes that do not encode the translation start codon in exon I. In an alternative preferred embodiment, the vector exon contains a start codon. Additional codons can be located between the translational start codon and the splice donor site. For example, a partial signal sequence can be encoded on the vector exon. The partial signal sequence can be any amino acid sequence capable of complementing a partial signal sequence from an endogenous gene to produce a functional signal sequence. The partial sequence may encode between one and one hundred amino acids, and may be derived from existing genes, or may consist of novel sequences. Thus, this vector is useful for producing and secreting protein from genes that encode part of the endogenous signal sequence in exon I, and the remainder in subsequent exons. In another example of a vector useful for activating a particular type of endogenous gene, a functional signal sequence can be encoded on the vector exon. This vector allows protein to be produced and secreted from genes that encode a signal sequence in exon I. It can also be used to produce secreted forms of proteins that are not normally secreted.

[0303] In cases where a start codon is included on the vector exon, it can be advantageous to produce a vector in each reading frame. This is achieved by varying the number of nucleotides between the start codon and the splice donor junction site. Together, the preferred vector configurations are capable of producing protein from endogenous genes, regardless of the exon/intron structure, location of the translation start codon, or reading frame.

Vectors for Isolating Exon I from Activated Endogenous Genes

[0304] The non-targeted gene activation vectors described above are useful for activating and isolating endogenous genes and for producing protein from endogenous genes. Upon integration upstream of an endogenous gene, however, each of these vectors produces a transcript that lacks exon I from the endogenous gene. Since the vectors are designed to produce a transcript containing the vector encoded exon spliced to the first splice acceptor site downstream of the vector integration site, and since the first exon of eukaryotic genes does not contain a splice acceptor site, normally, the first exon of endogenous genes will not be recovered on mRNA molecules derived from non-targeted gene activation. For some genes, such as genes that contain coding information in the first exon, there is a need to efficiently recover the first exon of the activated endogenous gene.

[0305] To recover the first exon of activated endogenous genes, a transcriptional regulatory sequence (which may be any transcriptional regulatory sequence, including but not limited to the promoters, enhancers, and repressors described herein, and which is preferably a promoter or an enhancer, and most preferably a promoter) is included on the activation vector downstream of a second transcriptional regulatory sequence (which may also be any transcriptional regulatory sequence, including but not limited to the promoters, enhancers, and repressors described herein, and which is preferably a promoter or an enhancer, and most preferably a promoter) which drives expression of a vector encoded exon. Thus, the upstream transcriptional regulatory sequence is linked to an unpaired splice donor site and the downstream transcriptional regulatory sequence is not linked to a splice donor site. Both transcriptional regulatory sequences are oriented to drive expression in the same direction. Examples of such exon I recovery vectors are shown in FIG. 12A-12G. The integration of this type of vector will create at least two different types of RNA transcripts (FIG. 13). The first transcript is derived from the upstream transcriptional regulatory sequence and contains the vector exon spliced to exon II of an endogenous gene. The second transcript is derived from the downstream transcriptional regulatory sequence and contains, from 5' to 3', the region between the vector and the transcription start site of the gene, exon I, exon II, and all downstream exons. Using methods described herein, both transcripts can be recovered and analyzed, allowing the characterization of exon I from genes isolated by non-targeted gene activation.

[0306] The exon located on the activation vector can encode a selectable marker, a protein, a portion of a protein, secretion signal sequences, a portion of a signal sequence, an epitope, or nothing. When a protein is encoded by the exon, a poly(A) signal may be included downstream of the vector
encoded gene. Alternatively, a poly(A) signal may be omitted. In another embodiment, a positive and negative selectable marker may be operably linked to the upstream transcriptional regulatory sequence(s). In this embodiment, the position of the unpaired splice donor site relative to the selectable markers is described above for the SAI vectors and the dual poly(A)SAI vectors.

Gene Activation Vectors for Single-Exon and Multi-Exon Gene Trapping

[0307] As noted above, in one embodiment the poly(A) trap vectors of the invention may contain a promoter operably linked to a selectable marker followed by an unpaired splice donor site. Such vectors, when integrated into or near a gene, produce transcripts containing the selectable marker spliced onto an endogenous gene. Since the endogenous gene encodes a poly(A) signal, the resulting mRNA is polyadenylated, thereby allowing the transcript to be translated at levels sufficient to confer drug resistance on the cell containing the integrated vector.

[0308] While the vectors described above are capable of "trapping" endogenous genes, the splice donor site downstream of a selectable marker cannot be used in, and in some cases can interfere with, several potential applications for such vectors. First, these vectors cannot be used to selectively trap single exon genes, since these genes do not contain a splice acceptor site. Second, these vectors often "trap" cryptic genes, since drug resistance relies solely on vector integration upstream of a poly(A) signal. Unfortunately, cryptic poly(A) signals exist in the genome, leading to formation of drug resistant cells and creation of non-genic transcripts containing the selectable marker. These cells and transcripts can interfere with gene discovery applications using these vectors. Third, without novel modifications such as those described herein (see above), these vectors are not capable of efficiently producing protein from the activated endogenous gene. Furthermore, protein expression from an endogenous gene can be poor even when an internal ribosome entry site (IRES) is included between the selectable marker and the splice donor site, since translation from an IRES is generally less efficient than translation from the first start codon at the 5' end of a transcript. Thus, there is a need for vectors that are capable of more specifically trapping endogenous genes, including single exon genes, and that are capable of efficiently expressing protein from the activated endogenous genes.

[0309] Thus, in additional embodiments, the present invention provides such vectors. In one such embodiment, the vector may contain a promoter operably linked to one or more (i.e., one, two, three, four, five, or more) selectable markers, wherein the selectable marker is not followed by a splice donor site or a poly(A) signal (see FIGS. 17A-17G). In general, upon integration into a host cell genome, this vector will fail to produce sufficient quantities of selectable marker since the marker transcript will not be polyadenylated. However, if the vector integrates in close proximity to, or into, a gene, including a single exon gene, the selectable marker will acquire a poly(A) signal from the endogenous gene, thereby stabilizing the marker transcript and conferring a drug resistant phenotype on the cell. In addition to selecting for vector integration into or near genes, vectors according to this aspect of the invention can also be used to recover exon 1 from the activated gene, as described in the section of this application entitled “Vectors for Isolating Exon 1 from Activated Endogenous Genes.”

[0310] In a preferred embodiment, the vector can contain a second selectable marker upstream of the first selectable marker (see FIG. 18). The upstream selectable marker is preferably operably linked to a transcriptional regulatory sequence, most preferably a promoter. Optionally, an unpaired splice donor site can be positioned between the transcription start site and the translation start site of the upstream selectable marker. Alternatively, the splice donor site may be located anywhere in the open reading frame of the upstream selectable marker, such that, following vector integration into a host cell genome, and upon splicing from the vector encoded splice donor site to an endogenous exon, the upstream selectable marker will be produced in an inactive form, or not at all. By selecting for cells that produce the downstream positive selectable marker in an active form, cells containing the vector integrated into or near a gene can be isolated. Furthermore, by selecting against cells producing the upstream selectable marker in the active form, cells in which the vector transcript has spliced to an exon from a multi-exon endogenous gene can be removed. In other words, these vectors can be used to isolate cells that contain a vector integrated into a single exon gene or into the 3' most exon of a multi-exon gene since, in these instances, a splice acceptor site is absent between the vector encoded splice donor site and the endogenous poly(A) signal. Thus, the majority of cells containing activated multi-exon genes will not survive selection, and as a result, cells containing activated single exon genes will be greatly enriched in the library.

[0311] In another preferred embodiment, vectors according to this aspect of the invention may contain one or more (i.e., one, two, three, four, five, or more, and preferably one) negative selectable marker(s) upstream of the first selectable marker (see FIGS. 19A and 19B). The negative selectable marker preferably is operably linked to a promoter. Optionally, an unpaired splice donor site may be positioned between the transcription start site and the translation start site of the negative selectable marker. Alternatively, the splice donor site may be located anywhere in the open reading frame of the negative selectable marker, such that, following vector integration into a host cell genome, and upon splicing from the vector encoded splice donor site to an endogenous exon, the negative selectable marker will be produced in an inactive form, or not at all. By selecting for cells that produce the positive selectable marker in an active form and selecting against cells producing the negative selectable marker in the active form, these vectors can be used to identify cells containing the vector integrated into or upstream of an endogenous gene. Since (1) splicing to an endogenous exon and (2) acquisition of a poly(A) signal are both required for cell survival, cells containing cryptic gene trap events are reduced within the library. The reason for this is that the probability of a vector integrating next to both a cryptic splice acceptor site and a cryptic poly(A) signal is substantially less than the probability of the vector integrating next to a single cryptic site. Thus, these vectors provide a higher degree of specificity for trapping genes than previous vectors.

[0312] It will also be recognized by one of ordinary skill in view of the teachings contained herein that vectors containing positive and negative selectable markers can be
used to produce protein from the activated endogenous gene. One vector configuration capable of directing protein production consists of the splice donor site positioned in the 5′ UTR of the negative selectable marker. Upon splicing, a chimeric transcript containing the 5′ UTR from the negative selectable marker linked to the second exon of an endogenous gene is produced. This vector is capable of activating protein production from genes that encode a translation start codon in the second or subsequent exon. Likewise, the splice donor site can be placed in the open reading frame of the negative selectable marker, in a position that does not interfere with the function of the marker unless splicing has occurred. Similar vectors containing the splice donor site positioned in different reading frames relative to the translation start codon can also be used. Upon splicing to an endogenous gene, these vectors will produce a chimeric transcript containing a start codon from the negative selectable marker fused to exon II of the activated endogenous gene. Thus, these vectors will be capable of activating protein expression from genes that encode a translation start codon in exon 1. Additional positive/negative selection vector designs capable of efficiently producing protein from activated endogenous genes are described below.

[0313] Any of the vectors of the invention can contain an internal ribosome entry site (ires) 3′ of the downstream selectable marker. The irs allows translation of the endogenous gene upon vector integration into an endogenous gene. Optionally, a translation start codon may be included between the selectable marker and the irs sequence. When a start codon is present, additional codons may be present on the exon. The start codon, and if present additional codons, may be present in any, and collectively all, reading frames relative to the splice donor site. Furthermore, the codons downstream of the translation start codon, if present, may encode, for example, a signal secretion signal, a partial signal sequence, a protein (including a full-length protein, a portion of a protein, a protein motif, an epitope tag, etc.), or a spacer region.

[0314] In additional preferred embodiments, any of the vectors described herein may contain, upstream of the selectable marker(s), a second transcriptional regulatory sequence (most preferably a promoter) operably linked to a eukaryotic region, followed by an unpaired splice donor site. This upstream exon is particularly useful for expressing protein from activated endogenous genes. The exon may lack a translation start codon. Alternatively, the exon may contain a translation start codon. When a start codon is present, additional codons may be present on the exon. The start codon, and if present additional codons, may be present in any, and collectively all, reading frames relative to the splice donor site. Furthermore, the codons downstream of the translation start codon, if present, may encode, for example, a signal secretion sequence, a partial signal sequence, a protein (including a full-length protein, a portion of a protein, a protein motif, an epitope tag, etc.), or a spacer region.

Activation Vectors Useful for Detecting Protein-Protein Interactions

[0315] Genetic approaches for detecting protein-protein interactions have previously been described (see, e.g., U.S. Pat. Nos. 5,283,173; 5,468,614; and 5,667,973, the disclosures of which are fully incorporated herein by reference). This approach relies on cloning a first cDNA molecule next to, and in frame with, a gene fragment encoding a DNA binding domain; and cloning a second cDNA molecule next to, and in frame with, a gene fragment encoding a transcription transactivation domain. Each chimeric gene is expressed from a promoter region located upstream of the chimeric gene. To detect expression, both chimeric genes are transfected into a reporter cell. If the first chimeric protein interacts with the second chimeric protein (via the proteins encoded by the cloned cDNA’s fused to the DNA binding and transcription activation domains), then the DNA binding domain and the transcription activation domain will be joined within a single protein complex. As a result, the protein-protein interaction complex can bind to the regulatory region of the reporter gene and activate its expression.

[0316] A limitation of this previous approach is that it is only capable of detecting protein-protein interactions between genes that have been cloned as cDNA. As described herein, many genes are expressed at very low levels, in rare cell types, or during short developmental windows; and therefore, these genes are typically absent from cDNA libraries. Furthermore, many genes are too large to be isolated efficiently as full-length clones, thereby making it difficult to use those previous approaches.

[0317] The present invention is capable of activating protein expression from endogenous genes or from transfected genomic DNA. Unlike previous approaches, virtually any gene can be efficiently expressed, regardless of its normal expression pattern. Furthermore, since the present invention is also capable of modifying the protein expressed from the endogenous gene (or from the transfected genomic DNA), it is also possible to produce chimeric proteins for use in protein-protein interaction assays.

[0318] To detect protein-protein interactions by the present invention, two vectors are used. The first vector, generally referred to as BD/SD (binding domain/splice donor), contains a promoter operably linked to a polynucleotide encoding a DNA binding domain and an unpaired splice donor site. The second vector, generally referred to as AD/SD (activation domain/splice donor), contains a promoter operably linked to a polynucleotide encoding a transcription activation domain and an unpaired splice donor site. To accommodate genes that have different reading frames, the binding domain and activation domain can be encoded in each of the three possible reading frames relative to the unpaired splice donor site. In addition, BD/SD and AD/SD vectors can have other functional elements, as described herein for other vectors, including selectable markers and amplifiable markers. The vectors may also contain selectable markers oriented in a configuration that permits selection for cells in which the vector has activated a gene. Multi-promoter/activation exon vectors are also useful. Several examples of BD/SD and AD/SD vectors are illustrated in FIG. 25. An example illustrating detection of a protein-protein interaction using these vectors is depicted in FIG. 26.

[0319] The DNA binding domain of the BD/SD vector may encode any protein domain capable of binding to a specific nucleotide sequence. When a transcription activation protein is used to supply the DNA binding domain, the transcription activation domain is omitted from the BD/SD vector. Examples of genes encoding proteins with DNA
binding domains include, but are not limited to, the yeast GAL4 gene, the yeast GCN4 gene, and the yeast ADR1 gene. Other genes from prokaryotic and eukaryotic sources may also be used to supply DNA binding domains.

[0320] The transcription activation domain of the AD/SD vector encodes a protein domain capable of enhancing transcription of a reporter gene when positioned near the promoter region of the reporter gene. When a transcription activation protein is used to supply the transcription activation domain, the DNA binding domain is omitted from the AD/SD vector. Examples of genes encoding proteins with transcription activation domains include, but are not limited to, the yeast GAL4 gene, the yeast GCN4 gene, and the yeast ADR1 gene. Other genes from prokaryotic and eukaryotic sources may also be used to supply transcription activation domains.

[0321] In the present invention, protein-protein interactions are detected using the BD/SD and AD/SD vectors described above, to activate expression of genes located in stretches of genomic DNA.

[0322] In one embodiment, the BD/SD vector is integrated randomly into the genome of a reporter cell line. As with other vectors described herein, the BD/SD vectors are capable of activating protein expression from genes located downstream of the vector integration site. Since the activation exon on the BD/SD vector encodes a DNA binding domain, the activated endogenous protein will be produced as a fusion protein containing the DNA binding domain at its N-terminus. Thus, by integrating the BD/SD vector into the genome of a host cell, a library of fusion proteins can be created, wherein each protein will contain a DNA binding domain at its N-terminus.

[0323] It is also recognized that the AD/SD vector can be integrated into the genome of a reporter cell line to produce a library of cells, wherein each member of the library is expressed as a different endogenous gene fused to a transcription activation domain.

[0324] Once created, the BD/SD library may be transfected with a vector expressing a specific gene (referred to below as gene X) fused to a transcription activation domain. This allows virtually any gene encoded in the genome to be tested for an interaction to gene X. Likewise, the AD/SD library may be transfected with a vector expressing a specific gene (e.g., gene X) fused to a DNA binding domain. This allows virtually any gene encoded in the genome to be tested for an interaction to gene X. It is also recognized that the specific gene may be stably expressed in the host cell prior to construction of the BD/SD or AD/SD libraries.

[0325] In an alternative embodiment, genomic DNA is cloned into the BD/SD and/or AD/SD vector(s) downstream of the DNA binding domain and activation domain, respectively. If a gene is present and correctly oriented in the genomic DNA, then the BD/SD vector (or the AD/SD vector) will be capable of expressing the gene as a fusion protein useful for detecting protein-protein interactions. Likewise integration of BD/SD (or AD/SD) vectors in situ, any gene can be tested regardless of whether it has been previously isolated as a cDNA molecule.

[0326] In another embodiment, a second library is created in the cells of the first library. For example, the AD/SD vector can be integrated into cells comprising the BD/SD library. Conversely, the BD/SD vector can be integrated into cells comprising the AD/SD library. This allows all proteins expressed as binding domain fusion proteins to be tested against all activation domain fusion protein. Since the present invention is capable of expressing substantially all of the proteins (as fusions with the binding and activation domains) in a eukaryotic organism, this approach, for the first time, allows all combinations of protein-protein interactions to be tested in a single library.

[0327] It is also recognized that library vs. library screens can be created in several ways. First, both libraries are produced, simultaneously or sequentially, by integrating BD/SD and AD/SD vectors into the genome of the same reporter cells. Second, a first library is created by integrating a BD/SD vector into the genome of a reporter cell, and a second library is produced by transfecting the AD/SD vector containing cloned genomic DNA. It is recognized that in this approach, the AD/SD library may be created first, followed by introduction of a BD/SD vector containing cloned genomic DNA. It is also recognized that the first library can be created by transfecting the BD/SD vector (or AD/SD vector) containing cloned genomic DNA, followed by integrating the second vector into the reporter cell genome. Third, both libraries are created, simultaneously or sequentially, by transfecting cells with a BD/SD and AD/SD vectors, wherein each vector contains a cloned fragment of genomic DNA. Fourth, it is recognized that when cloned genomic fragments are used in either the BD/SD vector or the AD/SD vector, a cDNA library may be created in the other vector and introduced into cells. This allows all of the genes present in the cDNA library to be tested for interaction with all other genes in the genome.

[0328] Since library/library screens involve the creation of large libraries of cells, it is important to maximize the frequency of gene activation and in frame fusion protein production among the members of the library. This can be accomplished in at least two ways. First, the BD/SD and AD/SD vectors can contain selectable markers in a configuration that “traps” genes. Examples of selection trap vectors are shown in FIGS. 8, 9, 10, 17, 19, 21, and 25. These vectors select for cells in which the activation vector has transcriptionally activated a gene. Second, multiple promoter/activation exon units can be included on the BD/SD and AD/SD vectors. Each promoter/activation exon unit encodes the binding domain (or activation domain) in a different reading frame relative to the unpaired splice donor site. An example of a multi-promoter/exon vector is illustrated in FIG. 25. This type of vector ensures that any gene
activated at the transcription level will be produced as an in frame fusion protein from on of the promoter/activation exon units on the vector. Third, the vectors can be introduced into the reporter cells using efficient transfection procedures. In this respect, insertion of BD/SD and AD/SD vectors by retroviral integration is advantageous.

[0329] Reporter cells useful in the present invention include any cell that is capable of properly splicing the transcripts produced by the BD/SD and AD/SD vectors. The reporter cells contain a reporter gene that is expressed at higher levels in the presence of a protein-protein interaction between proteins expressed from BD/SD and AD/SD vectors. The reporter gene may be a selectable marker, such as any of the markers described herein. Alternatively, the reporter gene may be a screenable marker. Examples of useful selectable markers and screenable markers are described herein.

[0330] In the reporter cells, a minimal promoter is operably linked to the reporter gene. To allow increased expression of the reporter gene in the presence of a protein-protein interaction, a DNA binding site is positioned in or near the minimal promoter, such that the DNA binding site is recognized by the protein encoded by the DNA binding domain region of the BD/SD vector. In the absence of a protein-protein interaction, the DNA binding domain fusion protein produced from BD/SD lacks a transcription activation domain, and therefore, cannot activate transcription from the minimal promoter of the reporter gene. If, however, the DNA binding domain fusion protein produced from BD/SD interacts with the activation domain fusion protein produced from the AD/SD vector, then the protein complex can activate expression of the reporter gene. Increased reporter gene expression can be detected using an assay for the screenable marker, or using drug selection for a selectable marker.

[0331] It is also recognized that other reporter systems can be used in conjunction with the present invention to detect protein-protein interactions. Specifically, any protein that contains two separable domains, each required to be in close proximity with the other to produce a biochemical or structural activity, can be used in conjunction with the present invention.

Multi-Promoter/Activation Exon Vectors

[0332] In applications of nontargeted gene activation in which the goal is to activate protein expression from an unknown gene, a collection of vectors typically must be used. Thus, in an additional embodiment, the invention provides vectors containing one or more promoter/activation exon units (see FIGS. 20A-20E).

[0333] To accommodate the variety of gene structures that exist in the genomes of eukaryotic cells, vectors according to this aspect of the invention preferably contain a transcriptional regulatory sequence (e.g., a promoter) operably linked to an activation exon with a different structure. Collectively, these activation exons are capable of activating protein expression from substantially all endogenous genes. For example, to activate protein expression from genes that encode a translation start codon in exon I (or exons downstream of exon II), one vector can contain a transcriptional regulatory sequence (e.g., a promoter) operably linked to an activation exon lacking a translation start codon. To activate protein expression from all types of genes that encode a translation start codon in exon I, three separate vectors must be used, each containing a transcriptional regulatory sequence (e.g., a promoter) operably linked to a different activation exon. Each activation exon encodes a start codon in a different reading frame. Additional activation exon configurations are also useful. For example, to activate protein expression and secretion from genes that encode a portion of their signal secretion sequence in exon I, three separate vectors must be used, each containing a transcriptional regulatory sequence (e.g., a promoter) operably linked to a different activation exon. Each activation exon encodes a partial signal sequence in a different reading frame. To activate protein expression and secretion from genes that encode their entire signal sequence in exon I, three vectors must be used, each containing a transcriptional regulatory sequence (e.g., a promoter) operably linked to a different activation exon. Each activation exon contains an entire signal secretion sequence in a different reading frame. In addition to activating expression of genes that encode secreted proteins, promoter/activation exons encoding entire signal sequences will also activate expression and secretion of proteins that are not normally secreted. This, for example, can facilitate protein purification of proteins that are normally intracellularly localized.

[0334] Other useful coding sequences can be included on the activation exon of vectors according to this aspect of the invention, including but not limited to sequences encoding proteins (including full length proteins, portions of proteins, protein motifs, and/or epitope tags). As described herein, vectors according to this aspect of the invention can be integrated, individually or collectively, into the genome of a host cell to produce a library of cells. Each member of the library will potentially overexpress a different endogenous protein. Thus, these collections of vectors make it possible to activate all or substantially all of the endogenous genes in a eukaryotic host cell.

[0335] When integrating a collection of vectors into host cells, as described above, activation of protein expression can be achieved from substantially any gene. Unfortunately, to produce protein from all endogenous genes, a large number of library members must be generated. In part, this is due to the large number of genes encoded by the host cell. In addition, using this approach, many cells will contain a vector integrated into or near an endogenous gene; however, the integrated vector will contain an activation exon with a structure that is incompatible with activating protein expression from the endogenous gene. For example, the vector exon may encode a start codon in reading frame 1 (relative to the splice junction), whereas the protein encoded by the first exon downstream of the integrated vector may be in reading frame 2 (relative to the splice junction). Thus, many library members will contain an integrated vector that has activated transcription of an endogenous gene, but that failed to produce the protein encoded by the endogenous gene.

[0336] To decrease the number of cells that fail to activate protein expression following vector integration into or near an endogenous gene, a vector containing multiple promoter/activation exon units can be used. On this vector, each promoter/activation exon unit can be capable of activating protein expression from an endogenous gene with a different structure. Since a single vector comprising multiple activation
exons is capable of producing multiple transcripts, each containing a different activation exon, a single vector integrated into or near a gene can be capable of activating protein expression, regardless of the structure of the endogenous gene (see FIG. 21).

[0337] Multi-promoter/activation exon vectors can contain two or more promoter/activation exons. Each promoter/activation exon unit may be followed by an unpaired splice donor site. In one such embodiment, two promoter/activation exons are included on the vector, wherein each promoter/activation exon is capable of activating protein expression from a different type of endogenous gene. In a preferred embodiment, the vector may contain three promoter/activation exons, wherein each exon encodes a translation start codon in a different reading frame. In another preferred embodiment, the vector may contain three promoter/activation exons, wherein each exon encodes a partial signal sequence in a different reading frame. In yet another preferred embodiment, the vector may contain three promoter/activation exons, wherein each exon encodes an entire signal sequence in a different reading frame. Additional embodiments include each of the vectors above containing a fourth promoter/activation exon, wherein the fourth activation exon does not encode a translation start codon.

[0338] Any number (e.g., one or more, two or more, three or more, four or more, five or more, etc.) of promoter/activation exon units may be included on the vector. When multiple promoter/activation exons are present on a single vector, they are preferably oriented in the same direction relative to one another (i.e., the promoters drive expression in the same direction).

[0339] The promoters that drive transcription of different activation exons may be the same as one another or one or more promoters may be different. The promoters may be viral, cellular, or synthetic. The promoters may be constitutive or inducible. Other types of promoters and regulatory sequences, recognizable to one skilled in the art or as described herein, may also be used in preparing the vectors according to this aspect of the invention.

[0340] Any of the vectors containing multiple promoter/activation exon units may optionally include one or more selectable marker(s) and/or amplifiable marker(s). The selectable and/or amplifiable markers may contain a poly(A) signal. Alternatively, the markers may lack a poly(A) signal. The selectable marker may be a positive or negative selectable marker. The selectable marker may contain an unpaired splice donor site upstream, within, or downstream of the marker. Alternatively, the selectable marker may lack an unpaired splice donor site. The selectable marker(s) and/or amplifiable marker(s), when present, may be located upstream, among, or downstream of the promoter/activation exon units. The selectable and/or amplifiable marker(s) may be located on the vector in any orientation relative to the promoter/activation exon units. When the purpose of the selectable marker is to trap endogenous genes, the selectable marker is preferably oriented in the same direction as the promoter/activation exons.

Amplifiable Markers

[0341] Any of the vectors described herein may also optionally comprise one or more (e.g., two, three, four, five, or more) amplifiable markers. Examples of amplifiable markers include those described in detail hereinabove. Preferably, the amplifiable marker(s) are located upstream of the positive/negative selectable marker(s). When using polyadenylation trap vectors, it may be advantageous to omit a polyadenylation signal from the amplifiable marker(s) to eliminate the possibility of capturing a vector-encoded poly(A) signal derived from vector concatemerization prior to integration.

[0342] When present, the amplifiable marker(s) may be located upstream of the activation transcriptional regulatory sequence (i.e. the promoter responsible for directing transcription from the vector through the endogenous gene). The amplifiable marker(s) may be present on the vector in any orientation (i.e. the open reading frame may be present on either DNA strand).

[0343] It is also understood that the amplifiable marker(s) can also be the same gene as the positive selectable marker. Examples of genes that can be used both as positive selectable markers and amplifiable markers include dihydrofolate reductase, adenosine deaminase (ada), dihydro- orotate, glutamine synthase (GS), and carbamyl phosphate synthase (CAD).

[0344] In some embodiments and for certain applications, it may be desirable to place multiple amplifiable markers on the vector. Use of more than one amplifiable marker allows dual selection, or alternatively sequential selection, for each amplifiable marker. This facilitates the isolation of cells that have amplified the vector and flanking genomic locus, including the gene of interest.

Promoters

[0345] It is understood that any promoter and regulatory element may be used on these activation vectors to drive expression of the selectable marker, amplifiable marker (if present), and/or the endogenous gene. In additional embodiments, the promoter driving expression of the endogenous gene is a strong promoter. The CMV immediate early gene promoter, SV40 T antigen promoter, and β-actin promoter are examples of this type of promoter. In another preferred embodiment, an inducible promoter is used to drive expression of the endogenous genes. This allows endogenous proteins to be expressed in a more controlled fashion. The Tetacycline inducible promoter, heat shock promoter, ecdysone promoter, and metallothionein promoter are examples of this type of promoter. In yet another embodiment, a tissue specific promoter is used to drive expression of endogenous genes. Examples of tissue specific promoters include, but are not limited to, immunoglobulin promoters, casein promoter, and growth hormone promoter.

Restriction Sites

[0346] The vectors of the invention can contain one or more restriction sites located downstream of the unpaired splice donor site in the vector. These restriction sites can be used to linearize plasmid vectors prior to transfection. In the linear configuration, the activation vector contains, from 5' to 3' relative to the transcribed strand, a promoter, a splice donor site, and a linearization site.

[0347] A restriction site(s) may also be included in the vector intron to facilitate removal of vector intron-containing cDNA molecules. In this embodiment, the vector con-
tains, from 5' to 3' relative to the transcribed strand, a promoter, a splice donor site, a restriction site, and a linearization site. By including a restriction site between the unpaired splice donor site and the linearization site, unspliced transcripts can be removed by digestion of cDNA with the appropriate restriction enzyme. cDNA molecules derived from gene activation have removed the vector intron containing the restriction site, and therefore, will not be digested. This allows gene activated transcripts to be preferentially enriched during amplification/Cloning, and greatly facilitates identification and analysis of endogenous genes.

0348] A restriction site(s) may also be included in the vector exon to facilitate cloning of activated genes. Following gene activation, mRNA is recovered from cells and synthesized into cDNA. By digesting the cDNA with a restriction enzyme that cuts in the vector exon, gene activated cDNA molecules will contain an appropriate overhang at the 5' end for subsequent cloning into a suitable vector. This facilitates isolation of gene activated cDNA molecules.

0349] In one embodiment, the restriction site located in the vector exon is different than the restriction site(s) located in the vector intron. This facilitates removal of cDNA molecules that contain a vector intron since the digested cDNA fragments from vector intron containing transcripts can be designed to have an overhang that is incompatible with the cloning vector (see below). Alternatively, degenerate restriction sites recognized by the same enzyme may be located in the vector exon and intron. Enzymes that cleave these sites are capable of cleaving multiple sites, sites with an odd number of bases in the recognition sequence, sites with interrupted palindromes, nonpalindromic sequences, or sites containing one or more degenerate bases. In other words, restriction sites recognized by the same restriction endonuclease may be used if the enzyme produces an overhang in the vector exon that is different from the overhang produced in the vector intron. Since different overhangs are produced, a cloning vector containing a site that is compatible with the vector exon overhang, and incompatible with the vector intron overhang may be used to preferentially clone vector exon containing and vector intron lacking cDNA molecules. Examples of useful degenerate restriction sites include DNA sequences recognized by Sfi I, Accl, Afl III, Spl II, Pse I, Tsp45 I, SstII, Tse I, PpuM I, Rsr II, and Sgr A I.

0350] The restriction site(s) located in the vector intron and/or exon can be a rare restriction site (e.g., an 8 bp restriction site) or an ultra-rare site (e.g., a site recognized by intron encoded nucleases). Examples of restriction enzymes with 8 bp recognitions sites include NotI, Sfi I, PacI, Ascl I, Fse I, Pmel, SspI, Srl II, Sph I, Sse 8387 I, and Swal. Examples of intron encoded restriction enzymes include I-Ppo I, Xcm I, CeuI, PI-Psp I, and PI-Thl. Alternatively, restriction sites smaller than 8 bp can be placed on the vector. For example, restriction sites composed of 7 bp, 6 bp, 5 bp, or 4 bp can be used. In general, the use of smaller restriction recognition sites will lead to the cloning of less than full-length genes. In some cases, such as creation of hybridization probes, isolation of smaller cDNA clones may be advantageous.

Bidirectional Activation Vectors

0351] The activation vectors described herein can also be bidirectional. When a single activation transcriptional regulatory sequence is present on the vector, gene activation occurs only when the vector integrates into an appropriate location (e.g., upstream of the gene) and in the correct orientation. That is, in order to activate an endogenous gene, the promoter on the activation construct must face the endogenous gene allowing transcription of the coding strand. As a result of this directionality requirement, only half of the integration events into a locus may result in the transcriptional activation of an endogenous gene. The other half of integration events result in the vector transcribing away from a gene of interest. Therefore, to increase the gene activation frequency by a factor of two, the present invention provides bidirectional vectors that may be used to activate an endogenous gene regardless of the orientation in which the vector integrates into the host cell genome.

0352] A bidirectional vector according to this aspect of the invention preferably comprises two transcriptional regulatory sequences (which may be any transcriptional regulatory sequences, including but not limited to the promoters, enhancers, and repressors described herein, and which preferably are promoters or enhancers, and most preferably promoters), two splice donor sites, and a linearization site. When a splice donor site is useful, each transcriptional regulatory sequence is operably linked to a separate splice donor site, and the transcriptional regulatory sequence/splice donor pairs may be in inverse orientation relative to each other (i.e., the first transcriptional regulatory sequence may be integrated into the host cell genome in an orientation that is inverse relative to the orientation in which the second transcriptional regulatory sequence has integrated into the host cell genome). The two opposing transcriptional regulatory sequence/splice donor sites can be separated by the linearization site. The function of the linearization site is to produce free DNA ends between the transcriptional regulatory sequence/splice donor sites (i.e., in a location suitable for activation of endogenous genes). Examples of bidirectional vectors of the invention are shown in FIG. 11A-11C.

0353] The two opposing transcriptional regulatory sequences may be the same transcriptional regulatory sequences or different transcriptional regulatory sequences. Optionally, a translational start codon (e.g., ATG) and one or more additional codons may be included on either or both vector encoded exons. When a translational start codon is present, either or both vector exons may encode a protein, a portion of a protein, a signal secretion sequence, a portion of a signal secretion sequence, a protein motif, or an epitope tag. Alternatively, either or both vector exons may lack a translational start codon.

0354] The bidirectional vectors according to this aspect of the invention may optionally include one or more selectable markers and one or more amplifiable markers, including those selectable markers and amplifiable markers described in detail herein. The bidirectional vectors may also be configured as poly(A) trap, splice acceptor trap, or dual poly(A)/splice acceptor trap vectors, as described above. Other vector configurations described for unidirectional vectors may also be incorporated into bidirectional vectors.

Co-Transfection of Genomic DNA with Non-targeted Activation Vectors

0355] It is recognized that any of the vectors described herein can be integrated into, or otherwise combined with, genomic DNA prior to transfection into a eukaryotic host.
cell. This permits high level expression from virtually any gene in the genome, regardless of the normal expression characteristics of the gene. Thus, the vectors of the invention can be used to activate expression from genes encoded by isolated genomic DNA fragments. To accomplish this, the vector is integrated into, or otherwise combined with, genomic DNA containing at least one gene, or portion of a gene. Typically, the activation vector must be positioned within or upstream of a gene in order to activate gene expression. Once inserted (or joined), the downstream gene may be expressed (as a transcript or a protein) by introducing the vector/genomic DNA into an appropriate eukaryotic host cell. Following introduction into the host cell, the vector encoded promoter drives expression through the gene encoded in the isolated DNA, and following splicing, produces a mature mRNA molecule. Using appropriate activation vectors, this process allows protein to be expressed from any gene encoded by the transected genomic DNA. In addition, using the methods described herein, cDNA molecules, corresponding to genes encoded by the transected genomic DNA, can be generated and isolated.

To achieve stable expression of the activated gene, the transected activation vector/genomic DNA can be integrated into the host cell genome. Alternatively, the transected activation vector/genomic DNA can be maintained as a stable episome (e.g., using a viral origin of replication and/or nuclear retention function—see below). In yet another embodiment, the activated gene may be expressed transiently, for example, from a plasmid.

As used herein, the term “genomic DNA” refers to the unspliced genetic material from a cell. Splicing refers to the process of removing introns from genes following transcription. Thus, genomic DNA, in contrast to mRNA and cDNA, contains exons and introns in an unspliced form. In the present invention, genomic DNA derived from eukaryotic cells is particularly useful since most eukaryotic genes contain exons and introns, and since many of the vectors of the present invention are designed to activate genes encoded in the genomic DNA by splicing to the first downstream exon, and removing intervening introns.

Genomic DNA useful in the present invention may be isolated using any method known in the art. A number of methods for isolating high molecular weight genomic DNA and ultra-high molecular weight genomic DNA (intact and encased in agarose plugs) have been described (Sambrook et al., Molecular Cloning, Cold Spring Harbor Laboratory Press, 1989). In addition, commercial kits for isolating genomic DNA of various sizes are also available (Gibco/ BRL, Stratagene, Clontech, etc.).

The genomic DNA used in the invention may encompass the entire genome of an organism. Alternatively, the genomic DNA may include only a portion of the entire genome from an organism. For example, the genomic DNA may contain multiple chromosomes, a single chromosome, a portion of a chromosome, a genetic locus, a single gene, or a portion of a gene.

Genomic DNA useful in the invention may be substantially intact (i.e. unfragmented) prior to introduction into a host cell. Alternatively, the genomic DNA may be fragmented prior to introduction into a host cell. This can be accomplished by, for example, mechanical shearing, nuclease treatment, chemical treatment, irradiation, or other methods known in the art. When the genomic DNA is fragmented, the fragmentation conditions may be adjusted to produce DNA fragments of any desirable size. Typically, DNA fragments should be large enough to contain at least one gene, or a portion of a gene (e.g., at least one exon). The genomic DNA may be introduced directly into an appropriate eukaryotic host cell without prior cloning. Alternatively, the genomic DNA (or genomic DNA fragments) may be cloned into a vector prior to transfection. Useful vectors include, but are not limited to, high and intermediate copy number plasmids (e.g., pUC, pBluescript, pACYC 184, pBR322, etc.), cosmids, bacterial artificial chromosomes (BAC’s), yeast artificial chromosomes (YAC’s), P1 artificial chromosomes (PAC’s), and phage (e.g., lambda, M13, etc.). Other cloning vectors known in the art may also be used. When genomic DNA has been cloned into a cloning vector, specific cloned DNA fragments may be isolated and used in the present invention. For example, YAC, BAC, PAC, or cosmid libraries can be screened by hybridization to identify clones that map to specific chromosomal regions. Optionally, once isolated, these clones can be ordered to produce a contig through the chromosomal region of interest. To rapidly isolate cDNA copies of the genes present in this contig, these genomic clones may be transected, separately or en masse, with the activation vector into a host cell. cDNA containing a vector encoded exon, and lacking a vector encoded intron, can then be isolated and analyzed. Thus, since all genes present in a contig can be rapidly isolated as cDNA clones, this approach greatly enhances the speed of positional cloning approaches.

Any activation vector described herein, including derivatives recognized by those skilled in the art, may be co-transfected with genomic DNA, and therefore, are useful in the present invention. In its simplest form, the vector can contain a promoter operably linked to an exon followed by an unpaired splice donor site. Examples of other useful vectors include, but are not limited to, poly A trap vectors (e.g. vectors illustrated in FIGS. 8, 9, 11C, 12F, and 17), dual poly (A)/Splice acceptor trap vectors (e.g. vectors illustrated in FIGS. 9, 10, 12G, 19, and 21), bi-directional vectors (e.g. vectors illustrated in FIG. 11), single exon trap vectors (e.g. the vector illustrated in FIG. 19), multi-promoter/activation exon vectors (e.g. the vector illustrated in FIG. 23), vectors for isolating cDNA’s corresponding to activated genes, and vectors for activating protein expression from activated genes (e.g. vectors illustrated in FIGS. 2, 3, 4, 8b-F, 9b-C, 9E-F, 10b-C, 10E-F, 11, 12, 17b-G, and 23).

The activation vector may also contain a viral origin of replication. The presence of a viral origin of replication allows vectors containing genomic fragments to be propagated as an episome in the host cell. Examples of useful viral origins of replication include ori P (Epstein Barr Virus), SV40 ori, BPV ori, and vaccinia ori. To facilitate replication from these origins, the appropriate viral replication proteins may be expressed from the vector. For example, EBV ori P and SV40 ori containing vectors may also encode and express EBNA-1 or T antigen, respectively. Alternatively, the vectors may be introduced into cells that are already expressing the viral replication protein (e.g. EBNA-1 or T antigen). Examples of cells expressing EBNA-1 and T antigen include human 293 cells transfected with an EBNA-1 expression unit (Clontech) and COS-7 cells (American Type Culture Collection; ATCC No. CRL-1651), respectively.
The activation vector may also contain an amplifiable marker. This enables cells containing increased copies of the vector and flanking genomic DNA, either episomal or integrated in the host cell genome, to be isolated. Cells containing increased copies of the vector and flanking genomic DNA express the activated gene at higher levels, facilitating gene isolation and protein production.

The activation vector and genomic DNA may be introduced into any host cell capable of splicing from the vector-encoded splice donor site to a splice acceptor site encoded by the genomic DNA. In a preferred embodiment, the genomic DNA/activation vector are transfected into a host cell from the same species as the cell from which the genomic DNA was isolated. In some instances, however, it is advantageous to transflect the genomic DNA into a host cell from a species that is different from the cell from which the genomic DNA was isolated. For example, transfection of genomic DNA from one species into a host cell of a second species can facilitate analysis of the genes activated in the transfected genomic DNA using hybridization techniques. Under high stringency hybridization, activated genes that were encoded by the transfected DNA can be distinguished from genes derived from the host cell. Transfection of genomic DNA from one species into a host cell from another species can also be used to produce protein in a heterologous cell. This may allow protein to be produced in heterologous cells that provide growth, protein modification, or manufacturing advantages.

The activation vector may be co-transfected into a host cell along with genomic DNA, wherein the vector is not attached to the genomic DNA prior to introduction into the cell. In this embodiment, the genomic DNA will become fragmented during the transfection process, thereby creating free DNA ends. These DNA ends can become joined to the co-transfected activation vector by the cell’s DNA repair machinery. Following joining to the activation vector, the genomic DNA and activation vector can be integrated into the host cell genome by the process of non-homologous recombination. If, during this process, a vector becomes joined to a gene encoded by the transfected genomic DNA, the vector will activate its expression.

Alternatively, the non-targeted activation vector may be physically linked to the genomic DNA prior to transfection. In a preferred embodiment, genomic DNA fragments are ligated to the vector prior to transfection. This is advantageous because it maximizes the probability of the vector becoming operably linked to a gene encoded by the genomic DNA, and minimizes the probability of the vector integrating into the host cell genome without the heterologous genomic DNA.

In a related embodiment, the genomic DNA may be cloned into the activation vector, downstream of the activation exon. In this embodiment, cloning of large genomic fragments can be facilitated in vectors capable of accommodating large genomic fragments. Thus, the activation vector may be constructed in BAC’s, YAC’s, PAC’s, cosmids, or similar vectors capable of propagating large fragments of genomic DNA.

Another method for joining the activation vector to genomic DNA involves transposition. In this embodiment, the activation vector is integrated into the genomic DNA by transposition or retroviral integration reactions prior to transfection into a cell. Accordingly, activation vectors can contain cis sequences necessary for facilitating transposition and/or retroviral integration. Examples of vectors containing transposon signals are illustrated in FIG. 27; however, it is recognized that any vector described herein may contain transposon signals.

Any transposition system capable of inserting foreign sequences into genomic DNA can be used in the present invention. In addition, transposons capable of facilitating inversions and deletions can also be used to practice the invention. While deletion and inversion systems do not integrate the activation vector into genomic DNA, they do allow the activation vector to change positions relative to cloned genomic DNA when the genomic DNA has been cloned into the activation vector. Thus, multiple genes within a given genomic fragment can be activated by shuffling the activation vector (by integration, inversion, or deletion) into multiple positions within, or outside of, the genomic fragment. Examples of transposition systems useful for the present invention include, but are not limited to ϕy, Tn3, Tn5, Tn7, Tn9, Tn10, Ty, retroviral integration and retro-transposons (Berg et al., Mobile DNA, ASM Press, Washington DC, pp. 879-925 (1989); Strathman et al., Proc. Natl. Acad. Sci. USA 88:1247 (1991); Berg et al., Gene 113:9 (1992); Liu et al., Nucl. Acids Res. 15:9461 (1987); Martin et al., Proc. Natl. Acad. Sci. USA 92:8398 (1995); Phadnis et al., Proc. Natl. Acad. Sci. USA 86:5908 (1989); Tomcsanyi et al., J. Bacteriol. 172:6348 (1990); Way et al., Gene 52:369 (1984); Bainton et al., Cell 62:805 (1991); Ahmed et al., J. Mol. Biol. 178:941 (1984); Benjamin et al., Cell 59:375 (1989); Brown et al., Cell 49:347 (1987); Eichinger et al., Cell 54:955 (1989); Eichinger et al., Genes Dev. 4:324 (1990); Braiterman et al., Mol. Cell. Biol. 14:5719 (1994); Braiterman et al., Mol. Cell. Biol. 14:5731 (1994); York et al., Nucl. Acids Res. 26:1927 (1998); Devine et al., Nucl. Acids Res. 18:3765 (1994); Goryshin et al., J. Biol. Chem. 273:7367 (1998).

Using transposition, an activation vector may be integrated into any form of genomic DNA. For example, the activation vector may be integrated into either intact or fragmented genomic DNA. Alternatively, the activation vector may be integrated into a cloned fragment of genomic DNA (FIG. 28). In this embodiment, the genomic DNA may reside in any cloning vector, including high and intermediate copy number plasmids (e.g. pUC, pBlueScript, pACYC184, pBR322, etc.), cosmids, bacterial artificial chromosomes (BAC’s), yeast artificial chromosomes (YAC’s), P1 artificial chromosomes (PAC’s), and phage (e.g. lambda, M13, etc.). Other cloning vectors known in the art may also be used. As described above, genomic fragments from specific genetic loci may be isolated and used as a substrate for activation vector integration.

Following integration of the activation vector, the genomic DNA may be introduced directly into a suitable host cell for expression of the activated gene. Alternatively, the genomic DNA may be introduced into and propagated in an intermediate host cell. For example, following integration of an activation vector into a BAC genomic library, the BAC library can be transformed into E. coli. This allows plasmids containing the transposon to be enriched by selecting for an antibiotic resistance marker residing on the activation vector. As a result, BAC plasmids lacking an integrated activation vector will be removed by antibiotic selection.
The transposition mediated activation vector integration may occur in vitro using purified enzymes. Alternatively, the transposition reaction may occur in vivo. For example, transposition may be carried out in bacteria, using a donor strain carrying the transposon either on a vector or as integrated copies in the genome. A target of interest is introduced into the transposon host where it receives integrations. Targets bearing insertions are then recovered from the host by genetic selection. Similarly, eukaryotic host cells, such as yeast, plant, insect, or mammalian cells, can be used to carry out the transposon mediated integration of an activation vector into a fragment of genomic DNA.

Isolation of mRNA and cDNA Produced From Activated Endogenous Genes

In addition to methods for isolating genes, particularly genes contained within the genome of a eukaryotic cell, that are activated using the vectors of the invention. These methods exploit the structure of the mRNA molecules produced using the non-targeted gene activation vectors of the invention. The methods of the invention described herein allow virtually any activated gene to be isolated, regardless of whether it has been previously isolated and characterized, and regardless of whether it has a known biological activity. This is made possible by the nature of the chimeric transcripts produced from the integrated vectors of the present invention. Using methods described herein, activation vectors can be integrated into the genome of a cell. Typically, the activation vectors, however, are integrated into the genome of many cells to produce a library of unique integration events. Each member of the library contains the vector located at a unique integration site(s), and potentially contains an activated endogenous gene. Gene activation occurs when the activation vector integrates upstream of the 3'-most exon of an endogenous gene and in an orientation capable of allowing transcription from the vector to proceed through the endogenous gene. The integration site may be in an intron or exon of the endogenous gene, or may be upstream of the transcription start site of the gene. Following integration, the activation constructs are designed to produce a transcript capable of blocking from an exon encoded by the activation vector to an exon encoded by the endogenous gene. As a result, a chimeric message is produced that contains the vector exon linked to the exons from an endogenous gene, wherein the endogenous exons are derived from the region located downstream of the vector integration site. The structure of this chimeric transcript can be exploited for gene discovery purposes. For example, the chimeric transcripts can be rapidly isolated to use as probes (to isolate the full length cDNA or genomic copy of the gene or to characterize the gene) or for direct sequencing and/or characterization.

To isolate the chimeric transcripts activated by vector insertion, cDNA is produced from a library member containing the activation event. It is also possible to isolate chimeric transcripts from pools of library members in order to decrease the through-put of the procedure. cDNA can then be produced from the mRNA harvested from the activated cells. Alternatively, total RNA may be used to produce cDNA. In either case, first strand synthesis can be carried out using an oligo dT primer, an oligo dT/poly(A) signal primer, or a random primer. To facilitate cloning of the cDNA product, a poly dT based primer can be used with the structure: 5'-Primer X(dT)10-33-3'. The oligo dT/poly(A) signal primer can have the structure 5'-(dT)30-Primer X-N30-3'. The random primer can have the structure: 5'-(Primer X)NNNNN-3'. In each primer, Primer X is any sequence that can be used to subsequently PCR amplify target nucleic acid molecules. Where the activated gene amplification product is to be cloned, it is useful to include one or more restriction sites within the primer X sequence to facilitate subsequent cloning. Other primers recognized by those skilled in the art can be used to create first strand cDNA products, including primers that lack a Primer X region.

In accordance with the invention, the primers may be conjugated with one or more hapten molecules to facilitate subsequent isolation of nucleic acid molecules (e.g., first and/or second strand cDNA products) comprising such primers. After the primer becomes associated with the nucleic acid molecule (via incorporation during cDNA synthesis), selective isolation of the molecule containing the haptenylated primer may be accomplished using a corresponding ligand which specifically interacts with and binds to the hapten via ligand-hapten interactions. In preferred such aspects, the ligand may be bound to, for example, a solid support. Once bound to the solid support, the molecules of interest (haptenylated primer-containing nucleic acid molecules) can be separated from contaminating nucleic acids and other materials by washing the support matrix with a solution, preferably a buffer or water. Cleavage of one or more of the cleavage sites within the primer, or by treatment of the solid support containing the nucleic acid molecule with a high ionic strength elution buffer, then allows for removal of the nucleic acid molecule of interest from the solid support.

Preferred solid supports for use in this aspect of the invention include, but are not limited to, nitrocellulose, diazocellulose, glass, polystyrene, polivinylchloride, polypropylene, polyethylene, dextran, Sepharose, agar, starch, nylon, latex beads, magnetic beads, paramagnetic beads, superparamagnetic beads or microtibre plates and most preferably a magnetic bead, a paramagnetic bead or a superparamagnetic bead, that comprises one or more ligand molecules specifically recognizing and binding to the hapten molecule on the primer.

Particularly preferred hapten molecules for use on the primer molecules of the invention include without limitation: (i) biotin; (ii) an antibody; (iii) an enzyme; (iv) lipopolysaccharide; (v) apotransferrin; (vi) ferotransferrin; (vii) insulin; (viii) cytokines (growth factors, interleukins or colony-stimulating factors); (ix) gp120; (x) β-actin; (xi) LFA-1; (xii) Mac-1; (xiii) glycoporin; (xiv) laminin; (xv) collagen; (xvi) fibronectin; (xvii) vitronectin; (xviii) integrins αβ1 and αβ3; (xix) integrins αβ1, αβ2, αβ4, β1β2, αβ5, β1β3, β1β5, and β2β3; (xx) integrins αβ1, αβ2, αβ3, β1β2, β1β3, and β2β3; (xxi) integrins αβ1, αβ2, αβ3, αβ7, β1β2, β1β3, and β2β3; (xxii) ankyrin; (xxiii) C3b, fibrinogen or Factor X; (xxiv) ICAM-1 or ICAM-2; (xxv) spectrin or fodrin; (xxvi) CD4; (xxvii) a cytokine (e.g., growth factor, interleukin or colony-stimulating factor) receptor; (xxviii) an insulin receptor; (xxix) a transferrin receptor; (xxx) Fe3+; (xxxi) polymyxin B or endotoxin-neutralizing protein (ENP); (xxxii) an enzyme-specific substrate; (xxxiii) protein A,
protein G, a cell-surface Fc receptor or an antibody-specific antigen; and (xxxiv) avidin and streptavidin. Particularly preferred is biotin.

[0378] Particularly preferred ligand molecules according to this aspect of the invention, which correspond in order to the above-described hapten molecules, include without limitation: (i) avidin and streptavidin; (ii) protein A, protein G, a cell-surface Fc receptor or an antibody-specific antigen; (iii) an enzyme-specific substrate; (iv) polymyxin B or endotoxin-neutralizing protein (ENP); (v) Fe***; (vi) a transferrin receptor; (vii) an insulin receptor; (viii) a cytokine (e.g., growth factor, interleukin or colony-stimulating factor) receptor; (ix) CD4; (x) spectrin or fodrin; (xi) ICAM-1 or ICAM-2; (xii) C3bi, fibrinogen or Factor X; (xiii) ankyrin; (xiv) integrins αβ1, αβ2, αβ3, αβ4, αβ5, and αβ6; (xv) integrins and αβ1, αβ2, αβ3, and αβ5; (xvi) integrins αβ1, αβ2, αβ3, αβ4, αβ4, αβ6, αβ6, and αβ6; (xvii) integrins αβ1, αβ2, αβ3, αβ4, αβ4, αβ6, αβ6, and αβ6; (xviii) vitronectin; (xix) fibroectin; (xx) collagen; (xxi) laminin; (xxii) glycophorin; (xxiii) Mac-1; (xxiv) LFA-1; (xxv) β-actin; (xxvi) gp 120; (xxvii) cytokines (growth factors, interleukins or colony-stimulating factors); (xxviii) insulin; (xxix) ferritin; (xxx) apotransferrin; (xxxi) lipopolysaccharide; (xxxii) an enzyme; (xxxiii) an antibody; and (xxxiv) biotin. Particularly preferred, for use with biotinylated primers of the invention, are avidin and streptavidin.

[0379] Following first strand synthesis, second strand cDNA synthesis may be carried out using a primer specific for the vector encoded exon. This creates double stranded cDNA from all transcripts that were derived from the vector encoded promoter. All cellular mRNA (and cDNA) produced from endogenous promoters remains single stranded since the transcript lacks a vector exon at it 3’ end. Once second strand synthesis is carried out, the cDNA may be digested with a restriction enzyme, cloned into a vector, and propagated.

[0380] To facilitate cloning, cDNA molecules containing the vector exon are amplified by PCR using a primer specific for the vector exon and a primer specific for the first strand cDNA primer (e.g., Primer X). PCR amplification results in the production of variable length DNA fragments representing different locations of priming during first strand synthesis and/or amplification of multiple chimeric transcripts from different genes. These amplification products can be cloned into plasmids for characterization, or can be labeled and used as a probe.

[0381] Other amplification techniques, such as linear amplification using RNA polymerase (Van Gelder, Proc. Natl. Acad. Sci. USA 87:1663-1667 (1990); Eberwine, Methods 10:283-288 (1996)), can be used. For example, when linear amplification by RNA polymerase is used, a promoter (e.g., T7 promoter) can be placed on the vector exon. As a result, gene activated transcripts will contain the promoter sequence at the 5’ end of the transcript. Alternatively, a promoter can be ligated onto the cDNA following first strand and second strand synthesis. Using either strategy, RNA polymerase is then incubated with cDNA in the presence of ribonucleotide triphosphates to create RNA transcripts from the cDNA. These transcripts are then reverse transcribed to produce cDNA. Since RNA polymerase can create several thousand transcripts from a single cDNA molecule, and since each of these transcripts can be reverse transcribed into cDNA, a large amplification can be achieved. As with PCR, amplification with RNA polymerase can facilitate cloning of activated genes. Other types of amplification strategies are also possible.

[0382] In another embodiment, the vector exon containing cDNA molecules are isolated without amplification. This may be useful in instances where biases occur during amplification (for example, when one DNA fragment amplifies more efficiently than another). To produce cDNA enriched for tagged messages, RNA is isolated from the activation library. A primer (e.g., a random hexamer, oligo(dT) or hybrid primers containing a primer linked to poly(dT) or a random nucleotides) is annealed to the RNA and used to direct first strand synthesis. The first strand cDNA molecules are then hybridized to a primer specific for the vector encoded exon. This primer directs second strand synthesis. Following second strand synthesis, the cDNA may be digested with restriction enzymes that cut in the vector exon and in the first strand primer (e.g. in Primer X—see above). The second strand products may then be cloned into a useful vector to allow them to be propagated.

[0383] It will be apparent to one of ordinary skill in view of the description contained herein that the cDNA products made according to the methods of the invention may also be cloned into a cloning vector suitable for transfection or transformation of a variety of prokaryotic (bacterial) or eukaryotic (yeast, plant or animal including human and other mammalian) cells. Such cloning vectors, which may be expression vectors, include but are not limited to chromosomal-, episomal- and virus-derived vectors, e.g., vectors derived from bacterial plasmids or bacteriophages, and vectors derived from combinations thereof, such as cosmids and phagemids, BACs, MACs, YACs, and the like. Other vectors suitable for use in accordance with this aspect of the invention, and methods for insertion of DNA fragments therein and transformation of host cells with such cloning vectors, will be familiar to those of ordinary skill in the art.

Removal of Unspliced Transcription Products

[0384] In some instances, the activation vector will integrate into the genome in a region lacking genes. Alternatively, it may integrate into a region containing a gene(s), but be oriented in a manner that results in the transcription of the non-coding strand. In each of these instances, gene activated transcripts are produced that contain normally untranscribed DNA sequences next to the vector encoded exon. These sequences would complicate identification and analysis of novel genes. Therefore, it would be advantageous to selectively remove these genomic molecules.

[0385] To remove cDNA molecules that contain a vector encoded intron, the double strand cDNA is treated with a restriction enzyme that recognizes a sequence located in the vector encoded intron. Preferably, the restriction enzyme creates an overhang that is different from the overhang produced by cleavage of the vector exon. This ensures the cloning of only activated genes by preventing the cleavage products from ligating into the cloning vector.

Recovery of Exon I from Activated Endogenous Genes

[0386] To recover exon I from activated genes, specialized vectors can be used to create non-targeted gene activation libraries. In its simplest form, this vector contains, from 5’ to 3’, a promoter, an unpaired splice donor site, and a second
promoter. The downstream promoter is oriented in the same direction as the upstream promoter. Upon integration upstream of an endogenous gene, this type of vector produces two types of transcripts. The first transcript contains the vector exon joined to exon II of the endogenous gene. Methods for isolating this transcript are described above. The second transcript contains the upstream region of the endogenous gene followed by exon I joined to exon II and other downstream exons from the endogenous gene (FIG. 6).

[0387] Using a two step process, exon I can be recovered from cells containing the integrated vector. First, vector exon containing transcripts (i.e. Transcript type #1, FIG. 13) are isolated using the methods described above. Once isolated, the 5' end of the transcript including exon II can be sequenced to determine the sequence of the flanking endogenous exons. Second, once the sequence of the flanking endogenous exons is known, PCR primers capable of annealing to exon II (or a downstream exon) of the activated gene can be developed. These primers can be used to amplify exon I from Transcript #2 (FIG. 13) using a modified form of inverse PCR (Zeiner, M., Biotechniques 17(6):1051-1053 (1994)). Briefly, amplification of exon I from the endogenous gene is achieved by carrying out first strand cDNA synthesis with a gene specific primer, based on the sequence information determined above. Second strand synthesis can be carried out using E. coli DNA polymerase I under conditions well known to those skilled in the art. The double strand cDNA is then digested with a restriction enzyme that cleaves at least once in the endogenous gene upstream of the first strand cDNA primer, and that does not cleave in the vector exon. Following digestion, the cDNA is self ligated to produce circular molecules. Using inverted PCR primers that anneal in the endogenous gene upstream of the restriction/circularization site, amplification by PCR produces a DNA product containing exon I sequences from the endogenous gene.

Method for Selecting Cells Containing Higher Levels of Gene Activated Transcripts/Protein

[0388] In several embodiments of the disclosed invention, the activation vector contains an amplifiable marker (e.g. DHFR) and a viral origin of replication (e.g. EBV ori P). In other embodiments, an amplifiable marker and viral origin of replication are present on a cloning vector containing a cloned fragment of genomic DNA. In yet another embodiment, the activation vector contains one element (e.g. DHFR) and a cloning vector carrying a genomic insert contains the other element (e.g. Ori P). Regardless of the initial location of the amplifiable marker and viral origin, the elements are combined on the same DNA molecule prior to or during introduction into a host cell.

[0389] In addition to the cis-acting elements, a trans-acting viral protein is generally required for efficient replication of the episomes. Examples of trans-acting viral proteins include EBNA-1 and SV40 T antigen. To promote efficient replication of episomes, the trans-acting viral protein can be expressed from the episome. Thus, the viral trans-acting protein may be expressed from the transposing activation vector, or may be positioned on the backbone of the cloning vector. Alternatively, the trans-acting viral protein may be expressed by the eukaryotic host cells into which the episome is introduced.

[0390] Once the amplifiable marker and viral origin of replication are on the same molecule and present in a host cell expressing the appropriate viral replication protein(s), the copy number of the episome can be increased. To increase the copy number of the episome, the cells can be placed under the appropriate selection. For example, if DHFR is present on the episome, methotrexate may be added to the culture. The selective agent may be applied at relatively high concentrations to isolate cells in the population that already have a high episome copy number. Alternatively, the selective agent may be applied at lower concentrations, and periodically increased in concentration. Two-fold increases in drug concentration will result in step-wise increases in copy number.

[0391] To reduce the frequency of non-specific drug resistance (i.e. drug resistance that is not associated with increased copy number of the episome), more than one amplifiable marker can be placed on the vector. Inclusion of multiple amplifiable markers on the episome allows cells to be selected with multiple drugs (either simultaneously or sequentially). Since non-specific drug resistance is a relatively rare event, the probability of a cell developing non-specific drug resistance to multiple drugs is exceedingly rare. Thus, the presence of multiple amplifiable markers on the episome facilitates isolation of cells that have a high episome copy number.

[0392] Amplification of episome copy number increases the number of transcripts derived from the vector activated gene. This, in turn, facilitates isolation of cDNA molecules derived from the activated gene. Furthermore, amplification of episome copy number can dramatically increase protein expression from the activated gene. Higher levels of protein production facilitate generation of proteins for bioassay screening, cell assay screening, and manufacturing purposes.

[0393] As a result of the highly desirable characteristics described above, vectors containing a viral origin of replication and an amplifiable marker, and the use of these vectors to rapidly amplify the copy number of episomal vectors, represent a break through that extends beyond the scope of activating expression of genes present in genomic DNA. For example, these vectors can be used to over-express cDNA encoded genes to produce high levels of protein expression without the need to integrate the gene into a host cell genome with an amplifiable marker. Furthermore, like amplification of chromosomal sequences, cell possessing several hundred to several thousand episomal copies of the vector can be isolated and maintained in culture. Thus, the vectors described herein, and their uses, allow high levels of cloned genomic DNA to be propagated in mammalian cells, facilitate isolation of cDNA copies of genes present on the vector as genomic inserts, and maximize protein production from cloned cDNA and genomic copies of eukaryotic genes.

[0394] Other suitable modifications and adaptations to the methods and applications described herein will be readily apparent to one of ordinary skill in the relevant arts and may be made without departing from the scope of the invention or any embodiment thereof. Having now described the present invention in detail, the same will be more clearly understood by reference to the following examples, which are included herewith for purposes of illustration only and are not intended to be limiting of the invention.
EXAMPLES

Example 1

Transfection of Cells for Activation of Endogenous Gene Expression

Method: Construction of pRIG-1

Human DHFR was amplified by PCR from cDNA produced from HT1080 cells by PCR using the primers DHFR-F1 (5' TCTCTCGAAGCTTGTCAATGGGTTCGCTAAACTGCA 3') (SEQ ID NO:1) and DHFR-R1 (5' AACCTTTAAGATCTGATATCTTCTCATATCTCAA 3') (SEQ ID NO:2), and cloned into the T site in pTARGET™ (Promega) to create pTARGET-DHFR. The RSV promoter was isolated from pREP9 by digestion with NheI and XbaI and inserted into the NheI sites of pTARGET-DHFR to create pTGt:RSV+DHFR. Oligonucleotides J1H169 (5' ATCCACCATGGCTACAAGGAGTAGCTCG 3') (SEQ ID NO:3) and J1H170 (5' GATCCGAGTACCTGATGAACTGTTAGGATTTA 3') (SEQ ID NO:4) were annealed and inserted into the lPpo-I and NheI sites of pTGt:RSV+DHFR to create pTGt:RSV+DHSV+Ex. A 279 bp region corresponding to nucleotides 230-508 of pBRR32 was PCR amplified using primers Tet F1 (5' GGCAGAATCTAGCCTATATGCTTTGGATCA 3') (SEQ ID NO:5) and Tet F2 (5' GCCGAGATCTGCTACTCTAAGACACCGGAAAACAGGGCTATGAGCCGAA 3') (SEQ ID NO:6). Amplification products were digested with BglII and cloned into the BamH1 site of pTGt:RSV+RSV+DHFR+Ex to create pRIG-1.

Transfection—Creation of pRIG-1 Gene Activation Library in HT1080 Cells

To activate gene expression, a suitable activation construct is selected from the group of constructs described above. The selected activation construct is then introduced into cells by any transfection method known in the art. Examples of transfection methods include electroporation, lipofection, calcium phosphate precipitation, DEAE dextran, and receptor mediated endocytosis. Following introduction into the cells, the DNA is allowed to integrate into the host cell’s genome via non-homologous recombination. Integration can occur at spontaneous chromosome breaks or at artificially induced chromosomal breaks.

Method: Transfection of human cells with pRIG1. 2×10^7 HH1 cells, an HPR1 subclone of HT1080 cells, was grown in 150 mm tissue culture plates to 90% confluence. Media was removed from the cells and saved as conditioned media (see below). Cells were removed from the plate by brief incubation with trypsin, added to media/10% fetal bovine serum to neutralize the trypsin, and pelleted at 1000 rpm in a Jouan centrifuge for 5 minutes. Cells were washed in 1×PBS, counted, and resuspended as above. The plate pellet was resuspended at 2.5×10^7 cells/ml in 1×PBS (Gibco BRL Cat #14200-075). Cells were then exposed to 50 rads of gamma irradiation from a 137Cs source. pRIG1 (FIG. 1A-14B; SEQ ID NO:18) was linearized with BamH1, purified with phenol/chloroform, precipitated with ethanol, and resuspended in PBS. Purified and linearized activation construct was added to the cell suspension to produce a final concentration of 40 µg/ml. The DNA:irradiated cell mixture was then mixed and 400 µl was placed into each 0.4 cm electroporation cuvettes (Biorad). The cuvettes were pulsed at 250 Volts, 600 µFarads, 50 Ohms using an electroporation apparatus (Biorad). Following the electric pulse, the cells were incubated at room temperature for 10 minutes, and then placed into αMEM/10% FBS containing penicillin/streptomycin (Gibco/BRL). The cells were then plated at approximately 7×10^5 cells/150 mm plate containing 35 ml αMEM/10% FBS/penstrep (33% conditioned media/67% fresh media). Following a 24 hour incubation at 37° C, G418 (Gibco/BRL) was added to each plate to a final concentration of 500 µg/ml from a 60 mg/ml stock. After 4 days of selection, the media was replaced with fresh αMEM/10% FBS/penstrep/500 µg/ml G418. The cells were then incubated for another 7-10 days and the culture supernatant assayed for the presence of new protein factors or stored at -80° C for later analysis. The drug resistant clones can be stored in liquid nitrogen for later analysis.

Example 2

Use of Ionizing Irradiation to Increase the Frequency and Randomness of DNA Integration

To induce integrated DNA constructs, HT1 cells were harvested at 90% confluence, washed in 1×PBS, and resuspended at a cell concentration of 7.5×10^6 cells/ml in 1×PBS. 15 µg linearized DNA (pRIG-1) was added to the cells and mixed. 400 µl was added to each electroporation cuvette and pulsed at 250 Volts, 600 µFarads, 50 Ohms using an electroporation apparatus (Biorad). Following the electric pulse, the cells were incubated at room temperature for 10 minutes, and then placed into 2.5 ml αMEM/10% FBS/1x penstrep. 300 µl of cells from each shock were then incubated at 0, 50, 500, and 5000 rads immediately prior to or at either 1 hour or 4 hours post transfection. Immediately following irradiation, the cells were plated onto tissue culture plates in complete medium. At 24 hours post plating, G418 was added to the culture to a final concentration of 500 µg/ml. At 7 days post-selection, the culture medium was replaced with fresh complete medium containing 500 µg/ml G418. At 10 days post selection, medium was removed from the plate, the colonies were stained with Coomassie Blue/90% methanol/10% acetic acid and colonies with greater than 50 cells were counted.

Example 3

Use of Restriction Enzymes to Generate Random, Semi-Random, or Targeted Breaks in the Genome

Method: HH1 cells were harvested at 90% confluence, washed in 1×PBS, and resuspended at a cell concentration of 7.5×10^6 cells/ml in 1×PBS. To test the efficiency of integration, 15 µg linearized DNA (PGK-fgeo) was added to each 400 µl aliquot of cells and mixed. To several aliquots of cells, restriction enzymes XbaI, NotI, HindIII, Ippol (10-500 units) were then added to separate cell/DNA mixture. 400 µl was added to each electroporation cuvette and pulsed at 250 Volts, 600 µFarads, 50 Ohms using an electroporation apparatus (Biorad). Following the electric pulse, the cells were incubated at room temperature for 10 minutes, and then placed into 2.5 ml αMEM/10% FBS/1x penstrep. 300 µl of 2.5 ml total cells from each shock were plated onto tissue culture plates in complete media. At 24 hours post plating, G418 was added to the culture to a final concentration of 500 µg/ml. At 7 days post selection, the media was replaced with fresh complete media containing 600 µg/ml G418. At 10 days post selection, media was
removed from the plate, the colonies were stained with Coomassie Blue/90% methanol/10% acetic acid and colonies with greater than 50 cells were counted.

Example 4
Amplification by Selecting for Two Amplifiable Markers Located on the Integrated Vector

Following integration of the vector into the genome of a host cell, the genetic locus may be amplified in copy number by simultaneous or sequential selection for one or more amplifiable markers located on the integrated vector. For example, a vector comprising two amplifiable markers may be integrated into the genome, and expression of a given gene (i.e., a gene located at the site of vector integration) can be increased by selecting for both amplifiable markers located on the vector. This approach generally facilitates the isolation of clones of cells that have amplified the correct locus (i.e., the locus containing the integrated vector).

Once the vector has been integrated into the genome by nonhomologous recombination, individual clones of cells containing the vector integrated in a unique location may be isolated from other cells containing the vector integrated at other locations in the genome. Alternatively, mixed populations of cells may be selected for amplification.

Cells containing the integrated vector are then cultured in the presence of a first selective agent that is specific for the first amplifiable marker. This agent selects for cells that have amplified the amplifiable marker either on the vector or on the endogenous chromosome. These cells are then selected for amplification of the second selectable marker by culturing the cells in the presence of a second selective agent that is specific for the second amplifiable marker. Cells that amplified the vector and flanking genomic DNA will survive this second selective step, whereas cells that amplified the endogenous first amplifiable marker or that developed non-specific resistance will not survive. Additional selections may be performed in similar fashion when vectors containing more than two (e.g., three, four, five, or more) amplifiable markers are integrated into the cell genome, by sequential culturing of the cells in the presence of selective agents that are specific for the additional amplifiable markers contained on the integrated vector. Following selection, surviving cells are assayed for level of expression of a desired gene, and the cells expressing the highest levels are chosen for further amplification. Alternatively, pools of cells resistant to both (if two amplifiable markers are used) or all (if more than two amplifiable markers are used) of the selective agents may be further cultured without isolation of individual clones. These cells are then expanded and cultured in the presence of higher concentrations of the first selective agent (usually twofold higher). The process is repeated until the desired expression level is obtained.

Alternatively, cells containing the integrated vector may be selected simultaneously for both (if two are used) or all (if more than two are used) of the amplifiable markers. Simultaneous selection is accomplished by incorporating both selection agents (if two markers are used) or all of the selection agents (if more than two markers are used) into the selection medium in which the transfected cells are cultured. The majority of surviving cells will have amplified the integrated vector. These clones can then be screened individually to identify the cells with the highest expression level, or they can be carried as a pool. A higher concentration of each selective agent (usually twofold higher) is then applied to the cells. Surviving cells are then assayed for expression levels. This process is repeated until the desired expression levels are obtained.

By either selection strategy (i.e., simultaneous or sequential selection), the initial concentration of selective agent is determined independently by titrating the agent from low concentrations with no cytototoxicity to high concentrations that result in cell death in the majority of cells. In general, a concentration that gives rise to discrete colonies (e.g., several hundred colonies per 100,000 cells plated) is chosen as the initial concentration.

Example 5
Isolation of cDNAs Encoding Transmembrane Proteins

pRIG8R1-CD2 (FIG. 5A-5D; SEQ ID NO:7), pRIG8R2-CD2 (FIG. 6A-6C; SEQ ID NO:8), and pRIG8R3-CD2 (FIG. 7A-7C; SEQ ID NO:9) vectors contain the CMV immediate early gene promoter operably linked to an exon followed by an unspliced splice donor site. The exon on the vector encodes a signal peptide linked to the extra-cellular domain of CD2 (lacking an in frame stop codon). Each vector encodes CD2 in a different reading frame relative to the splice donor site.

To create a library of activated genes, 2×10^7 cells were irradiated with 50 rads from a 57CS source and electroporated with 15 μg of linearized pRIG8R1-CD2 (SEQ ID NO:7). Separately, this was repeated with pRIG8R2-CD2 (SEQ ID NO:8), and again with pRIG8R3-CD2 (SEQ ID NO:9). Following transfection, the three groups of cells were combined and plated into 150 mm dishes at 5×10^6 transfected cells per dish to create library #1. At 24 hours post-transfection, library #1 was placed under 500 μg/ml G418 selection for 14 days. Drug resistant clones containing the vector integrated into the host cell genome were combined, aliquoted, and frozen for analysis. Library #2 was created as described above, except that 3×10^7 cells, 3×10^7 cells and 1×10^7 cells were transfected with pRIG8R1-CD2, pRIG8R2-CD2, and pRIG8R3-CD2, respectively.

To isolate cells containing activated genes encoding integral membrane proteins, 3×10^6 cells from each library were cultured and treated as follows:

Cells were trypsinized using 4 ml of Trypsin-EDTA.

After the cells had released, the trypsin was neutralized by addition of 8 ml of alpha MEM/10% FBS.

The cells were washed once with sterile PBS and collected by centrifugation at 800×g for 7 minutes.

The cell pellet was resuspended in 2ml of alpha MEM/10% FBS, 1 ml was used for sorting while the other 1 ml was replated in alpha MEM/10% FBS containing 500 μg/ml G-418, expanded and saved.

The cells used for sorting were washed once with sterile alpha MEM/10% FBS and collected by centrifugation at 800×g for 7 minutes.
The supernatant was removed and the pellet resuspended in 1 ml of alpha MEM/10% FBS. 100 μl of these cells was removed for staining with the isotype control.

200 μl of Anti-CD2 FITC (Pharminogen catalog #30054X) was added to the 900 μl of cells while 20 μl of the Mouse IgG1, isotype control (Pharminogen catalog #33814X) was added to the 100 μl of cells. The cells were incubated, on ice, for 20 minutes.

To the tube that contained the cells stained with the Anti-Human CD2 FITC, 5 ml of PBS/1% FBS were added. To the isotope control, 900 μl of PBS/1% FBS were added. The cells were collected by centrifugation at 6000g for 6 minutes.

The supernatant from the tubes was removed. The cells that had been stained with the isotype control were resuspended in 500 μl of alpha MEM/10% FBS, and the cells that had been stained with anti-CD2-FITC were resuspended in 1.5 ml alpha MEM/10% FBS.

Cells were sorted through five sequential sorts on a FACS Vantage Flow Cytometer (Becton Dickinson Immunocytochemistry Systems; Mountain View, Calif.). In each sort, the indicated percentage of total cells, representing the most strongly fluorescing cells (see below) were collected, expanded, and resorted. HT1080 cells were sorted as a negative control. The following populations were sorted and collected in each sort:

<table>
<thead>
<tr>
<th>Library #1</th>
<th>Library #2</th>
<th>Library #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sort #1 500,000 cells collected (top 10%)</td>
<td>100,000 cells collected (top 10%)</td>
<td>40,000 cells collected (top 10%)</td>
</tr>
<tr>
<td>Sort #2 300,000 cells collected (top 5%)</td>
<td>220,000 cells collected (top 11%)</td>
<td>14,000 cells collected (top 5%)</td>
</tr>
<tr>
<td>Sort #3 90,000 cells collected (top 5%)</td>
<td>40,000 cells collected (top 10%)</td>
<td>120,000 cells collected (top 10%)</td>
</tr>
<tr>
<td>Sort #4 600,000 cells collected (top 5%)</td>
<td>(a) 60,000 cells collected (top 5%); (b) 160,000 cells collected (next 5%); (c) from group (c) of sort #4, 100,000 cells collected (top 10%), and 350,000 cells collected (next 5%); (d) from group (b) of sort #4, 120,000 cells collected (top 10%)</td>
<td></td>
</tr>
<tr>
<td>Sort #5 (a) 260,000 cells collected (top 10%); (b) 530,000 cells collected (next 25%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cells from each of the final sorts for each library were expanded and stored in liquid nitrogen.

Isolation of Activated Genes from FACS-Sorted Cells

Once cells had been sorted as described above, activated endogenous genes from the sorted cells were isolated by PCR-based cloning. One of ordinary skill will appreciate, however, that any art-known method of cloning of genes may be equivalently used to isolate activated genes from FACS-sorted cells.

Genes were isolated by the following protocol:

1. Using PolyAtract System 1000 mRNA isolation kit (Promega), mRNA was isolated from 3×10⁷ CD2-cells (sorted 5 rounds by FACS, as described above) from libraries #1 and #2.

2. After mRNA isolation, the concentration of mRNA was determined by diluting 0.5 μl of isolated mRNA into 99.5 μl water and measuring OD₂₆₀. 21 μg of mRNA were recovered from the CD2-cells.

3. First strand cDNA synthesis was then carried out as follows:

(a) While the PCR machine was holding at 4ºC, first strand reaction mixtures were set up by sequential addition of the following components:

(b) 41 μl DEPC-treated ddH₂O

(c) 4 μl 10 mM each dNTP

(d) 8 μl 0.1 MDTT

(e) 16 μl 5× MMLV first strand buffer (Gibco-BRL)

(f) 5 μg (10 pmol/μl) of the consensus polyadenylation site primer GD.R1 (SEQ ID NO:10)*

(g) 1 μl RNAsin (Promega)

(h) 3 μl (1.25 μg/μl) mRNA.

*Note: GD.R1: 5’TTTTTTTTTTTCGTACGCCGCGC- CATCNAATTATT (SEQ ID NO:10), is a "Gene Discovery" primer for first strand cDNA synthesis of mRNA; this primer is designed to anneal to the poly-adenylation signal AATAAAA and downstream poly-A region. This primer will introduce a NotI site into the first strand.

Once samples had been made up, they were incubated as follows:

(b) 70º for 1 min.

(c) 42º hold.

2 μl of 400 U/μl SuperScript II (Gibco-BRL; Rockville, Md.) was then added to each sample, to give a final total volume of 82 μl. After approximately three minutes, samples were incubated as follows:

(d) 37º for 30 min.

(e) 94º for 2 min.

(f) 4º for 5 min.

2 μl of 20 U/μl RNase-IT (Stratagene) was then added to each sample, and samples were incubated at 37º for 10 min.

(4) Following first strand synthesis, cDNA was purified using a PCR cleanup kit (Qiagen) as follows:

(a) 80 μl of the first strand reaction were transferred to a 1.7 ml siliconized eppendorf tube and adding 400 μl of PB.

(b) Samples were then transferred to a PCR clean-up column and centrifuged for two minutes at 14,000 RPM.
(c) Columns were then disassembled, flowthrough decanted, 750 of µl PE were added to pellets, and tubes were centrifuged for two minutes at 14,000 RPM.

(d) Columns were disassembled and flowthrough decanted, and tubes then centrifuged for two minutes at 14,000 RPM to dry resin.

(e) cDNA was then eluted using 50 µl of EB through transferring column to a new siliconized eppendorf tube which was then centrifuged for two minutes at 14,000 RPM.

(5) Second strand cDNA synthesis was then carried out as follows:

(a) Second strand reaction mixtures were set up at RT, through the sequential addition of the following components:

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>dH₂O</td>
<td>55 µl</td>
</tr>
<tr>
<td>10x PCR buffer</td>
<td>10 µl</td>
</tr>
<tr>
<td>50 mM MgCl₂</td>
<td>5 µl</td>
</tr>
<tr>
<td>10 mM dNTPs</td>
<td>2 µl</td>
</tr>
<tr>
<td>25 pmol/µl R.I.G.F781</td>
<td>4 µl</td>
</tr>
<tr>
<td>25 pmol/µl GD.R2**</td>
<td>4 µl</td>
</tr>
<tr>
<td>First strand product</td>
<td>20 µl</td>
</tr>
</tbody>
</table>

(b) Start second strand synthesis:

94° C. for 1 min;

add 1 µl Taq (5U/µl, Gibco-BRL);

(c) Incubate at 63° C. for 2 min.

(d) Incubate at 72° C. for 3 min.

(e) Repeat step (b) four times.

(f) Incubate at 72° C. for 6 min.

(g) Incubate at 4° C. (hold)

(h) END

(6) 200 µl of 1 mg/ml Streptavidin-Paramagnetic Particles (SA-PMP) were then prepared by washing three times with STE.

(7) The products of the second strand reaction were added directly to the SA-PMPs and incubated at RT for 30 minutes.

(8) After binding, SA-PMPs were collected through use of the magnet, and flowthrough material recovered.

(9) Beads were washed three times with 500 µl STE.

(10) Beads were resuspended in 50 µl of STE and collected at the bottom of the tube using the magnet. STE supernatant was then carefully pipetted off.

(11) Beads were resuspended in 50 µl of dH₂O and placed into a 100°C water bath for two minutes, to release purified cDNA from PMPs.

(12) Purified cDNA was recovered by collecting PMPs on the magnet and carefully removing the supernatant containing the cDNA.

(13) Purified products were transferred to a clean tube and centrifuged at 14,000 RPM for two minutes to remove all of the residual PMPs.

(14) A PCR reaction was then carried out to specifically amplify RIG activated cDNAs, as follows:

(a) PCR reaction mixtures were set up at RT, through the sequential addition of the following components:

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₂O</td>
<td>59 µl</td>
</tr>
<tr>
<td>10x PCR buffer</td>
<td>10 µl</td>
</tr>
<tr>
<td>50 mM MgCl₂</td>
<td>5 µl</td>
</tr>
<tr>
<td>10 mM dNTPs</td>
<td>2 µl</td>
</tr>
<tr>
<td>25 pmol/µl R.I.G.F781</td>
<td>4 µl</td>
</tr>
<tr>
<td>25 pmol/µl GD.R2**</td>
<td>4 µl</td>
</tr>
<tr>
<td>Second strand product</td>
<td>20 µl</td>
</tr>
</tbody>
</table>

(b) Start thermal cycler:

94° C. for 5 min;

add 1 µl of Taq (SU/µl; Gibco-BRL);

add 1 µl of 0.1 U/µl Vent DNA polymerase (New England Biolabs)

(15) PCR was then carried out by 10 cycles of steps (c) to (e):

94° C. for 30 sec.

(16) 60° C. for 40 sec.

72° C. for 3 min.

PCR was then completed by carrying out the following steps:

72° C. for 30 sec.

60° C. for 40 sec.

72° C. for 3 min.

i) 72° C. +20 sec each cycle for 10 cycles

ii) 72° C. for 5 min

iii) 4° C. hold.

(17) After elution of library material with 50 µl EB, samples were digested by adding 10 µl of NEB Buffer 2, 40 µl of dH₂O and 2 µl of SfiI and digesting for 1 hour at 50° C., to cut the 5’ end of the cDNA at the SfiI site encoded by the forward primer (R.I.G.F781; SEQ ID NO:13).

(18) Following SfiI digestion, 5 µl of 1M NaCl and 2 µl of NotI were added to each sample, and samples digested for one hour at 37° C., to cut the 3’ end of the cDNA at the NotI site encoded by the first strand primer (GD.R1; SEQ ID NO:10).

(19) The digested cDNA was then separated on a 1% low melt agarose gel. cDNAs ranging in size from 1.2Kb to 8Kb were excised from the gel.
[0485] (18) cDNA was recovered from the excised agarose gel using Qiaex II Gel Extraction (Qiagen). 2 µl of cDNA (approximately 30ng) was ligated to 7 µl (35ng) of pBS-HS1 (linearized with SfiI/NotI) in a total volume of 10 µl of 1x T4 ligase buffer (NEB), using 400 units of T4 DNA ligase (NEB).

[0486] (19) 0.5 µl of the ligation reaction mixture from step (18) was transformed into E. coli DH10B.

[0487] (20) 103 colonies/0.5 µl ligated DNA were recovered.

[0488] (21) These colonies were screened for exons using the primers M13F20 and JH182 (RIG Exon1 specific) through PCR in 12.5 µl volumes as follows:

<table>
<thead>
<tr>
<th># of 96-Well Plates:</th>
<th>1 Plate</th>
<th>2 Plates</th>
<th>3 Plates</th>
<th>4 Plates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of 12.5 µl PCR mix:</td>
<td>96</td>
<td>192</td>
<td>288</td>
<td>384</td>
</tr>
<tr>
<td>dH2O</td>
<td>755 µl</td>
<td>1.47 ml</td>
<td>2.20 ml</td>
<td>2.94 ml</td>
</tr>
<tr>
<td>5x PCR Premix-4</td>
<td>250 µl</td>
<td>500 µl</td>
<td>750 µl</td>
<td>1.0 ml</td>
</tr>
<tr>
<td>F Primers premix (25 pmol/µl)</td>
<td>10 µl</td>
<td>20 µl</td>
<td>30 µl</td>
<td>40 µl</td>
</tr>
<tr>
<td>R Primers premix (25 pmol/µl)</td>
<td>10 µl</td>
<td>20 µl</td>
<td>30 µl</td>
<td>40 µl</td>
</tr>
<tr>
<td>RNase-H/Cocktail</td>
<td>3.2 µl</td>
<td>6.3 µl</td>
<td>9.6 µl</td>
<td>12.8 µl</td>
</tr>
<tr>
<td>Taq Polymerase (5 U/µl)</td>
<td>3.2 µl</td>
<td>6.3 µl</td>
<td>9.6 µl</td>
<td>12.8 µl</td>
</tr>
<tr>
<td>Total Volume (µl)</td>
<td>1.01</td>
<td>2.02</td>
<td>3.03</td>
<td>4.04</td>
</tr>
</tbody>
</table>

[0489] (a) 100 µl of LB (with selective antibiotic) were dispensed into the appropriate number of 96-well plates.

[0490] (b) Single colonies were picked and inoculated into individual wells of the 96-well plate, and the plate placed into a 37° C. incubator for 2-3 hours without shaking.

[0491] (c) A PCR reaction “master mix” was prepared on ice, as follows:

<table>
<thead>
<tr>
<th># of 96-Well Plates:</th>
<th>1 Plate</th>
<th>2 Plates</th>
<th>3 Plates</th>
<th>4 Plates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of 12.5 µl PCR mix:</td>
<td>96</td>
<td>192</td>
<td>288</td>
<td>384</td>
</tr>
<tr>
<td>dH2O</td>
<td>755 µl</td>
<td>1.47 ml</td>
<td>2.20 ml</td>
<td>2.94 ml</td>
</tr>
<tr>
<td>5x PCR Premix-4</td>
<td>250 µl</td>
<td>500 µl</td>
<td>750 µl</td>
<td>1.0 ml</td>
</tr>
<tr>
<td>F Primers premix (25 pmol/µl)</td>
<td>10 µl</td>
<td>20 µl</td>
<td>30 µl</td>
<td>40 µl</td>
</tr>
<tr>
<td>R Primers premix (25 pmol/µl)</td>
<td>10 µl</td>
<td>20 µl</td>
<td>30 µl</td>
<td>40 µl</td>
</tr>
<tr>
<td>RNase-H/Cocktail</td>
<td>3.2 µl</td>
<td>6.3 µl</td>
<td>9.6 µl</td>
<td>12.8 µl</td>
</tr>
<tr>
<td>Taq Polymerase (5 U/µl)</td>
<td>3.2 µl</td>
<td>6.3 µl</td>
<td>9.6 µl</td>
<td>12.8 µl</td>
</tr>
<tr>
<td>Total Volume (µl)</td>
<td>1.01</td>
<td>2.02</td>
<td>3.03</td>
<td>4.04</td>
</tr>
</tbody>
</table>

[0492] (d) 10 µl of the master mix were dispensed into each well of the PCR reaction plate.

[0493] (e) 2.5 µl from each 100 µl E. coli culture were transferred into the corresponding wells of the PCR reaction plate.

[0494] (f) PCR was performed, using typical PCR cycle conditions of:

[0495] (i) 94° C./2 min. (Bacterial lysis and plasmid denaturation)

[0496] (ii) 30 cycles of 92° C. denaturation for 15 sec; 60° C. primer annealing for 20 sec; and 72° C. primer extension for 40 sec

[0497] (iii) 72° C. final extension for 5 min.

[0498] (iv) 4° C. hold.

[0499] (g) Bromophenol blue was then added to the PCR reaction;

[0500] samples were mixed, centrifuged, and then the entire reaction mix was loaded onto an agarose gel.

[0501] 23) Of 200 clones screened, 78% were positive for the vector exon. 96 of these clones were grown as mini-preps and purified using Qiagen 96-well turbo-prep following the Qiagen Miniprep Handbook (April 1997). 24) Many duplicate clones were eliminated through simultaneous digestion of 2 µl of DNA with NotI, Bam HI, XhoI, XbaI, HindIII, EcoRI in NEB Buffer 3, in a total volume of 22 µl, followed by electrophoresis on a 1% agarose gel.

Results:

[0502] Two different cDNA libraries were screened using this protocol. In the first library (TMT #1), eight of the isolated activated genes were sequenced. Of these eight genes, four genes encoded known integral membrane proteins and six were novel genes. In the second library (TMT #2), 11 isolated activated genes were sequenced. Of these 11 genes, one gene encoded a known integral membrane protein, one gene encoded a partially sequenced gene homologous to an integral membrane protein, and nine were novel genes. In all cases where the isolated gene correspond to a characterized known gene, that gene was an integral membrane protein.

[0503] Exemplary significant alignments (obtained from GenBank) for genes isolated from each library are shown below:

TMT #1 Significant Alignments:

[0504] 179761/gb/M76559/HUMCANC1L Human neuronal DHP-sensitive voltage-dependent, calcium channel alpha-2 subunit mRNA complete CDs.

[0505] Length=3600

[0506] 663189374/emb/Y10183/HSMEMD H. sapiens mRNA for MEMD protein

[0507] Length=4235

TMT #2 Significant Alignments:

[0508] 476590/gb/U06715/HU06715 Human cytochrome B561, HCYTO B561, mRNA,

[0509] partial CDs.

[0510] Length=2463

[0511] 2184843/gb/AA459959/AA459959 zn66e01.s1 Soares total fetus Nb2H8 Fw Homo sapiens cDNA clone 796414 3' similar to gb:J03171 INTERFERON-ALPHA RECEPTOR PRECURSOR (HUMAN);

[0512] Length=431

Example 6

Activation of Endogenous Genes Using a Poly(A) Trap Vector

[0513] HT1080 cells (1x10^7 cells) were irradiated with 50 rads using a 137C Cs source and electroporated with 15 µl linearized pRIG14 (FIG. 29A-29B). Following transfection, the cells were plated into a 150 mm dish at 5x10^6 cells/dish.
At 24 hours, puromycin was added to 3 μg/ml. The cells were incubated at 37° C. for 12 days in the presence of 3 μg/ml puromycin. The media was replaced every 5 days. At 12 days, the number of colonies was counted, and the cells were trypsinized and replated onto a new dish. The cells were grown to 90% confluency and harvested for frozen storage and gene isolation. Typically, 1000-3000 colonies were produced per 1x10^7 cells transfected.

Example 7

Activation of Endogenous Genes Using a Dual Poly(A) Trap/SAI Vector

1x10^7 HH1 cells (HPRT-minus HT1080 cells) were irradiated with 50 rads using a ^137Cs source and electroporated with 15 μg linearized pRIG-22. Following transfection, the cells were plated into a 150 mm dish at 5x10^5 cells/dish. At 24 hours, neomycin was added to 500 μg/ml G481. The cells were incubated at 37° C. for 4 days in the presence of 500 μg/ml G481. The media was replaced with fresh media containing 500 μg/ml G481 and AgThg and grown in the presence of both drugs for an additional 7 days. Alternatively, as a control for HPRT activity, the media was replaced with fresh media containing 500 μg/ml G481 and HAT (available from Life Technologies, Inc., Rockville, Md., and used at manufacturer's recommended concentration) and grown in the presence of both drugs for an additional 7 days. At 12 days post transfection, the number of colonies was counted, and the cells were trypsinized and replated onto a new dish. The cells were grown to 90% confluency and harvested for frozen storage and gene isolation. Typically, cells subjected to G481/AgThg selection produced 1000-3000 colonies per 1x10^7 cells transfected. In contrast, cells subjected to G481/HAT selection produced approximated 100 colonies per 1x10^7 cells transfected.

Example 8

Isolation of Activated Genes

Non-targeted gene activation vectors are integrated into the genome of a eukaryotic cells using the methods of the invention. By integrating the vector into multiple cells, a library is created in which cells are expressing different vector activated genes. RNA is isolated from these cells using a commercial RNA isolation kit. In this example, RNA is isolated from cells using Poly(A) Trap 1000 (Promega). The RNA is converted into cDNA, amplified, size fractionated, and cloned into a plasmid for analysis and sequencing. A brief description of this process is presented.

1) Place 4 ml GTC Extraction buffer (Poly(A) tract 1000 Kit-Promega) in a 15 ml polycarbonate screw cap tube and add 168 μl 2-mercaptoethanol and place in a 70° C. water bath.

2) Place 8 ml dilution buffer in a 15 ml polycarbonate screw cap tube for every pellet processed and add 168 μl 2-mercaptoethanol and place in a 70° C. water bath.

3) Remove from ~80° C. storage cell pellets (1x10^7-1x10^8 cells) containing non-targeted gene activation vector integrated into their genome. Pipette 4ml GTC Extraction buffer immediately onto cell pellet. Pipette up-and-down several times until the pellet is resuspended and transfer into a 15 ml snap cap polypropylene tube.

4) Add the 8 ml dilution buffer and mix by inversion.

5) Add 10 μl (500 pmol) of the biotinylated oligo dT primer and mix.

6) Let sit at 70° C. for 5 minutes inverting every couple of minutes to ensure even heating.

7) Centrifugate in a Sorvall BB-6 rotor at 7800 rpm (10kxg) at 25° C. for 10 minutes. During this period of time wash 6 ml Streptavidin-Paramagnetic particles (SA-PMPs) 3x with 6 ml 0.5xSSC through use of the Poly(A) Tract system 1000 magnet.

8) After 3 washes resuspend the SA-PMPs in 6 ml 0.5xSSC.

9) Pipette to remove the supernatant from the RNA prep and add to the resuspended SA-PMPs (Be careful when removing supernatant so that you do not disrupt the pellet).

10) Let the SA-PMP/RNA mix and incubate for 2 minutes at room temperature.

11) Capture the magnetic beads through use of the Poly(A) Tract system 1000 magnet. Note that it takes some time for all of the beads to pellet due to the high viscosity of the liquid.

12) Pour off the supernatant and resuspend the beads in 1.7 ml of 0.5xSSC using a 2 ml pipette and transfer to a 2 ml screw cap tube.

13) Capture the SA-PMPs using the magnet and remove the supernatant by pipetting with a P1000.

14) Add 1.7 ml 0.5xSSC and invert the tube several times to mix.

15) Repeat steps 14 and 15 two more times.

16) Resuspend the SA-PMPs in 1 ml of nuclease free water and invert several times to mix.

17) Capture the SA-PMPs and pipette off the mRNA.

18) Place 0.5 ml of the mRNA into each of two siliconized eppendorf tubes and add 50 μl of DEPC-treated 3M NaOAc solution and 0.55 ml of isopropanol. Invert several times to mix and place at ~20° C. for at least 4 hours.

19) Centrifugate the mRNA for 10 minutes at max RPM (14 k).

20) Carefully pipette off the supernatants and wash pellets with 200 μl 80% ethanol through re-centrifugation for 2 minutes at 14K RPM. Note that the pellets are often brown or tan in color. This color results from residual SA-PMPs.

21) Remove wash and let pellets air dry for not more than 10 minutes at room temperature.

22) Resuspend pellets in 5 μl each and combine into a single tube.

23) Centrifugate at 14K RPM for 2 minutes to remove the residual SA-PMPs and carefully remove the mRNA.
[0539] 24) Determine the concentration of mRNA by diluting 0.5 µl into 99.5 µl water and measuring OD 260. Note that 1 OD 260 = 40 µg RNA.

[0540] 25) Set up first strand reaction for both the test sample and the negative control (HT1080) through the sequential addition of the following components while the PCR machine is holding at 4°C:

[0541] Step 1:
- 42 µl DEPC-treated ddH₂O
- 4 µl 10 mM each dNTP
- 8 µl 0.1 M DTT
- 16 µl 5xMMLV 1st strand buffer
- 5 µl (10 pmol/µl) GDR1
- 1 µl RNAsin (Promega)
- 4 µl (1.25 µg/µl) mRNA.

[0542] Step 2: 70°C/1 min

[0543] Step 3: 42°C/hold

[0544] Step 4: After 1 minute add 2 µl SUPERSCRIPT II® (Life Technologies, Inc.; Rockville, Md.) and incubate at 37°C for 30 min

[0545] Step 5: 94°C/2 min

[0546] Step 6: 45°C

[0547] Step 7: Add 2 µl RNase and incubate at 37°C for 10 min

[0548] Step 8: 40°C

[0549] 26) Analyze 8 µl of cDNA on a 1% agarose gel to check for cDNA synthesis and purify remaining cDNA using the PCR cleanup kit from Qiagen by transferring the 70 µl first strand reaction to a 1.5 ml siliconized eppendorf tube and adding 400 µl PB.

[0550] 27) Transfer to a PCR clean-up column and centrifuge 2 minutes at max RPM.

[0551] 28) Disassemble column and pour out Flow through. Add 750 µl PE and centrifuge 2 minutes at max RPM.

[0552] 29) Disassemble column and pour out Flow through then centrifuge 2 minutes at max RPM to dry resin.

[0553] 30) Elute using 50 µl of EB through transferring column to a new siliconized eppendorf tube and centrifuging for 2 minutes at max RPM.

[0554] 31) Second Strand cDNA synthesis set up at RT:

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₂O</td>
<td>8.5 µl</td>
</tr>
<tr>
<td>10x PCR buffer</td>
<td>5 µl</td>
</tr>
<tr>
<td>50 mM MgCl₂</td>
<td>2.5 µl</td>
</tr>
<tr>
<td>10 mM dNTPs</td>
<td>1 µl</td>
</tr>
<tr>
<td>25 pmol/µl GDE5Bio</td>
<td>10 µl</td>
</tr>
<tr>
<td>25 pmol/µl GDR2</td>
<td>10 µl</td>
</tr>
<tr>
<td>First strand product</td>
<td>15 µl</td>
</tr>
</tbody>
</table>

[0555] 32) Prepare 100 µl of SA-PMPs by washing 3x with STE and collection using a magnet. After the final wash, resuspend the beads in 150 µl STE.

[0556] 33) Purify the products of the second strand reaction using the PCR cleanup kit from Qiagen. Elute in 50 µl EB and add he products of the second strand reaction to 150 µl of the PMPs.

[0557] 34) Mix gently at RT for 30 minutes.

[0558] 35) After binding collect SA-PMPs through use of a magnet and recover flow through material (SAVE THIS MATERIAL!)

[0559] 36) Wash the beads 3x with 500 µl STE and 1x with NEB 2 (1x).

[0560] 37) Resuspend the beads in 100 µl NEB 2 (1x).

[0561] 38) Add 2 µl SfiI and digest at 50°C for 30 minutes with gentle mixing every 10 minutes.

[0562] 39) Recover purified cDNA through use of a magnet and carefully removing the supernatant.

[0563] 40) Transfer the products to a new tube and centrifuge at maximum RPM for 2 minutes to remove all of the beads.

[0564] 41) Set up a PCR reaction to specifically amplify RAGE activated cDNAs:

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₂O</td>
<td>37 µl</td>
</tr>
<tr>
<td>10x PCR buffer</td>
<td>10 µl</td>
</tr>
<tr>
<td>10 mM dNTPs</td>
<td>2 µl</td>
</tr>
<tr>
<td>25 pmol/µl GDE</td>
<td>10 µl</td>
</tr>
<tr>
<td>25 pmol/µl GDR2</td>
<td>10 µl</td>
</tr>
<tr>
<td>Second strand product</td>
<td>25 µl</td>
</tr>
</tbody>
</table>

[0565] Step 1: 94°C/2 min


[0567] Step 3: 60°C/10 min.

[0568] Add 0.5 µl Taq Polymerase

[0569] Step 4: 72°C/10 min.

[0570] Step 5: 60°C/2 min.

[0571] Step 6: 72°C/10 min.

[0572] Step 8: Cycle to step 5, 8 more times
[0591] Step 10: 60°C/2 min.
[0592] Step 11: 72°C/10 min.+20 sec each cycle
[0593] Step 12: Cycle to step 9, 14 more times
[0594] Step 13: 72°C/5 min.
[0595] Step 14: 4°C hold.

[0596] 42) Check specificity of PCR amplification of HT1080 versus library material through analysis on 1% agarose gel. If there is a high specificity of cDNA amplification, then use Qiagen PCR clean up kit to purify PCR products.

[0597] 43) After elution of library material with 50 µl EB add 10 µl NEB2, 40 µl dH2O and 2 µl SII and digest for 1 hour at 50°C.

[0598] 44) Add 5 µl of 1 M NaCl and 2 µl of NotI and digest for 1 hour at 37°C.

[0599] 45) Prepare and run a 1% L.M. agarose gel and run library material on gel. After visualization of material, cut out fragments ranging in size from 500 bp to 10 Kb.

[0600] 46) Recover the library DNA from agarose using Qiagen II Gel Extraction Protocol (Qiagen) and elute DNA in 10 µl EB. Digest 5 µl of this material to 4 µl PBS-HSBS (SII/NotI) or PBS-SNS in a total volume of 10 µl.

[0601] 47) Transform E. coli with 0.5 µl ligated DNA per 40 µl cells.


[0603] 49) Analyze gene activated cDNA inserts by restriction digest and DNA sequencing.

Example 9
Isolation of Activated Genes from Subtracted cDNA Pools

[0604] Purified mRNAs from non-transfected HT1080 cells was prepared using the Poly-A Tract 1000 system (Promega), as described in Example 8 steps 1-24, and were biotinylated using EZ-LinkTM Biotin LC-ASA reagent (Pierce), as follows:

[0605] 1.) 25 µl DEPC-treated dH2O and 15 µl containing 10 µg of HT1080 mRNA was added into a siliconized microfuge tube and held on ice.

[0606] 2.) Working under subdued light, 40 µl of prepared LC-ASA stock reagent (1 mg/ml in 100% ethanol) was added into the reaction tube.

[0607] 3.) A UV light (365 nm wavelength) was positioned 5 cm above the microfuge tube and used to irradiate the reaction mix for 15 minutes.

[0608] 4.) Unlinked biotin reagent was removed from the labeled HT1080 mRNA by passing the reaction mix through an RNase-free MicroSpin P-30 column (BioRad), as prescribed by the manufacturer.

[0609] HT1080 cells were transfected with a poly(A) trap pRKG activation vector and grown under selective media to produce a population of drug resistant colonies, as described in Example 1. Purified mRNAs were prepared from the pooled colonies using the Promega Poly-A Tract 1000 system, as described in Example 8. First strand cDNA was prepared from 5 µg of this mRNA using oligo GD.R1 (TTTTTTTTTTTGTCAGCCGCGG- CATCENNNNTTATT) (SEQ ID NO:10), as described in Example 8, Step 25. The reaction mix was passed through a Qiagen PCR Quick Clean-up column and the purified 1st strand cDNA was recovered in 100 µl EB.

[0610] The subtractive hybridization of biotinylated HT 1080 mRNAs (subtractor population) and 1 st strand cDNAs prepared from the superpool of pRKG-transfected colonies (target population) was performed as follows:

[0611] 1.) 9 µg of biotinylated mRNA was added into a 0.5 ml microfuge tube containing 0.5 µg 1st strand cDNA.

[0612] 2.) 1/100x volume of 10 mg/ml glycogen, 1/10x volume of 3 M sodium acetate, pH 5.5, and 2.6x volume of 100% ethanol were added into the tube and mixed.

[0613] 3.) The tube was placed at -80°C for 1 hr, then spun in a refrigerated microfuge for 20 minutes.

[0614] 4.) The pellet of precipitated nucleic acids was drained, washed once with 70% ethanol, then air-dried.

[0615] 5.) The pellet was solvated in 5 µl HBS (50 mM HEPES, pH 7.6; 2 mM EDTA; 0.2% SDS; 500 mM NaCl) and overlaid with 5 µl light mineral oil, then heated to 95°C for 2 minutes followed by 68°C for 24 hours.

[0616] 6.) Reaction mix was diluted with 100 µl HB (HBS without SDS) and extracted once with 100 µl chloroform to remove the oil.

[0617] 7.) The diluted hybridization mix was added to 300 µl streptavidin-coated paramagnetic particles (Promega) which had been pre-washed 3x in 300 µl HB.

[0618] 8.) The mix was incubated 10 minutes at room temperature and the SA-PMP’s and bound Biotin-mRNA/DNA hybrids were removed from solution by magnetic capture.

[0619] 9.) Steps 7 and 8 were repeated once.

[0620] 10.) The cleared solution was subjected to one additional round of subtractive hybridization and magnetic removal of captured hybrids (Steps 1-9), with the following exceptions:

[0621] 11.) Step 6: the hybridization reaction was diluted with 2xPCR Buffer (40 mM Tris-HCl, pH 8.4; 100 mM KCl).

[0622] 12.) Step 7: PMP’s were pre-washed in 1xPCR Buffer

[0623] The twice-subtracted 1st strand cDNA was used to generate 2nd strand cDNA by combining 45 µl of 1 st strand cDNA with 7 µl dH2O, 5 µl 50 mM MgCl2, 2 µl premix of 10 mM each dNTP, 1 µl 10xPCR Buffer, 20 µl of 12.5 pmol/µl GD 19F1 -Bio (5’-Biotin-CTCGGTTATGTGGCCGGC- GCTGAG-ATCAGTAACTTGACAGCCT) (SEQ ID NO:14), 20 µl of 12.5 pmol/µl GD.R2 (TTTTCTGCAGCCGCGGCCCATC) (SEQ ID NO:12), and 0.5 µl Taq Polymerase, with thermocycling as described in Example 8, Step 31. The second strand cDNA product was amplified and further processed for the production of an E. coli-based cDNA library, as described in Example 8, steps 32-49.
Selective Capture of RIG-Activated Transcripts

**Example 10**

[0624] HT1080 cells were transfected with pRIG19 activation vector (FIG. 30A-30C) and cultured for 2 weeks in selective media, as described in Example 6. Total RNA was prepared from a pellet comprised of 10⁶ cells using TRIzol® Reagent (Life Technologies, Inc.: Rockville, Md.) following the manufacturer's protocol, and was dissolved in 720 µl of DEPC-treated dH₂O (dH₂O DEPC). Contaminating genomic DNA was eliminated from the RNA preparation by mixing 80 µl NEB 10× Buffer 2, 8 µl Promega RNasin, and 20 µl RQ1 Promega RNase-free DNase, incubating at 37°C for 30 minutes, extracting sequentially with equal volumes of phenol/chloroform (1:1) and chloroform, mixing with 1/10x volume sodium acetate (pH 5.5), precipitating the RNA with 2x volume of 100% ethanol, and solvating the dried RNA pellet in dH₂O DEPC to a final concentration of 4.8 µg/ml.

[0625] mRNA transcripts derived from pRIG19-activated genes were selectively captured from a pool of total cellular RNAs by mixing in a 2 ml RNase-free microfuge tube 150 µl total RNA, 150 µl HBDEPC (50 mM HEPES, pH 7.6; 2 mM EDTA; 500 mM NaCl), 3 µl Promega RNasin, and 2.5 µl (25 pmol/µl) oligo GD19.R1-Bio (see Table 1), then incubating at 70°C for 5 minutes followed by 50°C for 15 minutes. One ml of Promega streptavidin coated paramagnetic particles (SA-PMPs) was magnetically captured and washed 3x each with 1.5 ml of 0.5xSSC and the SA-PMPs were left without being resuspended. The warm oligo: RNA hybridization reaction was added directly into the tube containing the semi-dry SA-PMPs. After incubating for 10 minutes at room temperature the SA-PMPs were washed 3x with 1 ml 0.5xSSC.

| Primer and Oligonucleotide Sequences | SEQ ID | Not
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer/Oligo Name</td>
<td>Sequence</td>
<td></td>
</tr>
<tr>
<td>Forward GD19.F1-Bio</td>
<td>5' Biotin-CTGGTTATAGGCCGCAGGCGAGGGCACTGAATTC</td>
<td>14</td>
</tr>
<tr>
<td>GD19.F2-Bio</td>
<td>5' Biotin-CTGGTTATAGGCCGCAGGCGAGGGCACTGAATTC</td>
<td>15</td>
</tr>
<tr>
<td>GD19.F2</td>
<td>GACCTACTGATTAAACGGCGCCAT</td>
<td>16</td>
</tr>
<tr>
<td>Reverse GD.R1</td>
<td>TTGGTTTGGTCGTTATGCCGCAGGCGAGGGCACTGAATTC</td>
<td>10</td>
</tr>
<tr>
<td>GD.R2</td>
<td>TTTGGTTTGGTCGTTATGCCGCAGGCGAGGGCACTGAATTC</td>
<td>12</td>
</tr>
<tr>
<td>mRNA Capture Oligo GD19.R1-Bio</td>
<td>TGGTCAGAAATGGTACAGAT-CT-3' Biotin</td>
<td>17</td>
</tr>
</tbody>
</table>

**Example 11**

Selective Capture of RIG-Activated Transcripts

[0634] HT 1080 cells were transfected with pRIG19 activation vector and cultured for 2 weeks in selective media, as described in Example 6. Total RNA was prepared from a pellet comprised of 10⁶ cells using TRIzol® Reagent (Life Technologies, Inc.) following the manufacturer's protocol, and was dissolved in 720 µl of DEPC treated dH₂O (dH₂O DEPC). Contaminating genomic DNA was eliminated from the RNA preparation by mixing 80 µl NEB 10× Buffer 2, 8 µl Promega RNasin, and 20 µl Promega RNase-free DNase, incubating at 37°C, for 30 minutes, extracting sequentially with equal volumes of phenol/chloroform (1:1) and chloroform, mixing with 1/10x volume sodium acetate (pH 5.5), precipitating the RNA with 2x volume of 100% ethanol, and solvating the dried RNA pellet in dH₂O DEPC to a final concentration of 4.8 µg/ml.

[0635] mRNA transcripts derived from pRIG 19-activated genes were selectively captured from the pool of total
cellular RNAs by mixing in a 2 ml RNase-free microfuge tube 150 μl total RNA, 150 μl HBDEPC (50 mM HEPES, pH 7.6; 2 mM EDTA; 500 mM NaCl), 3 μl Promega RNasin, and 2.5 μl (25 pmol/μl) oligo GD19.R1-Bio (see Table 1), then incubating at 70°C for 5 minutes followed by 50°C for 15 minutes. One ml of Promega streptavidin coated paramagnetic particles (SA-PMPs) was magnetically captured and washed 3x each with 1.5 ml of 0.5xSSC, and the SA-PMPs were left without being resuspended. The warm oligoRNA hybridization reaction was added directly into the tube containing the semi-dry SA-PMPs. After incubating for 10 minutes at room temperature the SA-PMPs were washed 3x with 1 ml 0.5xSSC. After the final magnetic capture the SA-PMP's were suspended in 190 μl dH2O2DPC and incubated at 68°C for 15 minutes. PMPs were immobilized by exposure to a magnetic and the cleared solution containing RIG-activated transcripts were transferred to a microfuge tube. 63 82 I of captured RIG-activated transcript were transferred to a PCR tube where first and second strand cDNA synthesis was performed using PCR program “4+2CDNA”, as follows:

[0636] Step1: 4°C ∞x; Add into the PCR tube containing the RIG-activated transcripts 20 μl 5x GirbeBRL RT Buffer, 1 μl Promega RNasin, 10 μl 100 mM DTT, 5 μl dNTP premix at 10 mM each, 1 μl oligo GD.R1 (see Table 1) at 25pmol/μl.

[0637] Step 2: 70°C C/3 minutes

[0638] Step 3: 42°C C/10 minutes

[0639] Step 4: Add 2.5 μl SUPERSCRIPT II® (Life Technologies, Inc.), then incubate at 37°C C/1 hour

[0640] Step 5: 94°C C/2 minutes

[0641] Step 6: 60°C C/∞; while holding temperature, the following were added: 2 μl 50 mM MgCl2, 1 μl oligo GD19.F1-Bio (Table 1) at 25 pmol/μl, and 2 μl Stratagene RNase-It. After 10 minutes, 0.5 μl Taq DNA Polymerase (Life Technologies, Inc.) was added and the cycling was continued:

[0642] Step 7: 72°C C/10 minutes

[0643] Step 8: 4°C C/∞.

[0644] The 100 μl volume cDNA reaction mix was transferred to a 1.5 ml siliconized microfuge tube and extracted sequentially with equal volumes of phenol:chloroform (1:1) and chloroform, and the aqueous phase was transferred to a new tube and placed in speed-vec for 5 minutes at 37°C. Restriction digestion of the cDNA was performed by adding 74 μl dH2O2, 20 μl NEB 10x Buffer 2, 2 μl 1 mg/ml BSA, 4 μl StII and incubating at 50°C for 1 hour, then adding 10 μl 1 M NaCl, 4 μl NolI and incubating an additional 37°C for 1 hour. The reaction mix was extracted sequentially with equal volumes of phenol:chloroform (1:1) and chloroform, and then cDNAs were precipitated by adding 1/100x volume 10 mg/ml glycogen, 1/30x volume 3 M sodium acetate (pH 7.5), 2x volume 100% absolute ethanol, and freezing at -80°C for 1 hour. The cDNA pellet was washed once with 70% ethanol and air dried for 15 minutes, then solvated in 5 μl dH2O, 1 μl 10xNEB Ligase Buffer, 4 μl pGBl5 vector DNA previously prepared by digestion with StII, NolI, and CIP, 0.5 μl T4 DNA Ligase was added, and the reaction mix was incubated at 16°C overnight. 10 μl dH2O was added to the ligation reaction and 0.5 μl was used to transform competent E. coli DH10B cells. Typically, 6 to 10 colonies per μl of transformed ligation mix were observed.

Example 12

Ligation of Activation Vectors to Genomic DNA and Transfection into Human Cells

[0645] Genomic DNA was harvested from a human cell line, HT1080 (106 cells), according to published procedures (Sambrook et al., Molecular Cloning, Cold Spring Harbor Laboratory Press, (1989)). The isolated genomic DNA was digested with BamHI under conditions that resulted in incomplete digestion. This was accomplished by titrating the amount of BamHI in the reaction. Each reaction contained 10 μg genomic DNA and BamHI at a concentration of either 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64, 1.28, 2.56, 5.62, or 11.24 units. After a one hour incubation at 37°C, the reactions were stopped by phenol extraction, followed by ethanol precipitation. The digested DNA from each reaction was separated by agarose gel electrophoresis. Reactions containing DNA predominantly in the range of 10 kb to 400 kb were combined for ligation to the activation vector. The pooled, digested genomic DNA was then added to BamHI linearized activation vector in 1x ligation buffer. Ligase (Life Technologies, Inc., 40 units) was added and the ligation reaction was incubated at 16°C for 24 hours. Following ligation, the genomic DNA/activation vector was transfected into HT1080 cells using LIPOFECTIN® (Life Technologies, Inc.) according to the manufacturer's procedures. Optionally, the HT1080 cells were irradiated prior to or after transfection. When cells were irradiated, doses in the range of 0.1 rads to 200 rads were found to be particularly useful. Following transfection, cells were grown in complete media. At 36 hours post-transfection, G418 (300 μg/ml) were added to the media. At 14 days post selection, the drug resistant clones were pooled, expanded, and harvested. Total RNA or mRNA was collected from the harvested cells. cDNA derived from vector activated genes was then synthesized and isolated using the methods described herein (see, e.g., Example 8 supra).

Example 1

Co-Transfections of BAC Contig Clones with the Activation Vector

[0646] Genomic libraries were created in pUniBAC (FIG. 34A-34B) according to published procedures (Shizuya et al., Proc. Natl. Acad. Sci. USA 89:8794 (1992)). Typically, the size of genomic fragments can be between 1 kb and 500 kb, and preferably between 50 kb and 500 kb. The BAC library was propagated in E. coli. To prepare plasmids for transfection, the library was plated onto LB agar plates containing 12.5 μg/ml chloramphenicol. Approximately 1000 clones were present on each 150 mm plate. Following growth and selection, the colonies from each plate were eluted from the agar plate through the addition of LB and pooled. Each pool (~10,000 clones) was grown in 1 liter LB/12.5 μg/ml chloramphenicol overnight. BAC plasmids were then isolated from each pool using a commercial kit (QIAGEN).

[0647] Purified BAC clones were digested with I-Ppo-I which cleaves a unique site in the BAC vector flanking the cloning site. Since I-Ppo-I is an ultra-cuts cutter, it will not digest the vast majority of genomic DNA inserts. Following
digestion, the linearized genomic library clones were cotransfected into HT1080 cells using LIPOFECTIN® (Life Technologies, Inc.) according to the manufacturer’s directions. Briefly, 10 µg of BAC genomic DNA was combined with 1 µg of linearized pRIG20 (Fig. 31A-31C) in α-MEM (no serum). 5 µg of LIPOFECTIN® was added to the DNA and the mixture was incubated at room temperature for 15 minutes. The DNA/LIPOFECTIN® mixture was then added to 10⁶ HT1080 cells in a 6 well dish. The cells were incubated with the DNA/LIPOFECTIN® in serum free α-MEM for 12 hours, washed, and placed in α-MEM/10% FBS for 36 hours. To select for cells that had integrated the vector and genomic DNA, the transfected cells were replated into a 10 cm dish and incubated in the presence of 300 µg/ml G418 for 10 days. Drug resistant clones were expanded and harvested to allow isolation of the activated cDNA molecules as described herein in Example 8.

Example 14

In vitro Integration of Activation Vector into Purified Genomic DNA and Transfection of the Integration Products into Host Cells

[0648] Genomic DNA was isolated and cloned into the Bacterial Artificial Chromosome, pUniBAC (Fig. 34A-34B), using published procedures (Sambrook et al., Molecular Cloning, Cold Spring Harbor Laboratory Press, (1989); Shizuya et al., Proc. Natl. Acad. Sci. USA 89:8794 (1992)). Following ligation of the genomic inserts into pUniBAC, the plasmids were transformed into the E. coli strain DH10B (Life Technologies, Inc.) and selected on tetracycline. Individual bacterial clones were combined into pools containing approximately 1000 members. Each pool was grown to saturation in 1 liter LB/tetracycline. pUniBAC plasmids containing genomic DNA inserts were isolated from the bacteria using a commercial kit (Qiagen).

[0649] For each pool of UniBAC clones, 2 µg of the library were incubated with 50 ng of the activation vector pRIG-T and 1 unit of mutant Tn5 transposase for 2 hours at 37°C. (transposase available from Epicentre Technologies). Following incubation, the pUniBAC clones were transformed into DH10B cells and selected on chloramphenicol. All colonies from each pool were combined and grown in 1 liter LB/chloramphenicol. Plasmids were harvested using Qiagen Tip-500 columns according to the manufacturer’s instructions.

[0650] For each pool, 20 µg of the library was transfected into 2x10⁶ HT1080 cells with 30 µg Ex-gen 500 (MBI Fermentas) according to the manufacturer’s instructions. At 48 hours post-transfection, the cells were placed into media containing 3 µg/ml puromycin. After 10 days of growth in the presence of puromycin, drug resistant clones were pooled, expanded and harvested for gene discovery. To isolate vector activated genes, mRNA from each pool of cells was isolated, converted to cDNA, and cloned into plasmids as described in Example 8. Individual cDNA clones were analyzed by restriction digestion and sequencing.

Example 15

Creation of Protein Expression Libraries from Cloned Genomic DNA

[0651] A genomic library containing genomic DNA inserts (100 kb avg. size) was created in pUniBAC as described in Examples 13 and 14. (Note: In some embodiments of the invention, the genomic fragments are cloned into the linearization site of an activation vector, wherein the activation vector is preferably a YAC, BAC, PAC, or Cosmid based vector.) In this example, the activation vector, pRIG-Tp, was integrated into the BAC genomic library using in vitro transposition as described in Example 14. pRIG-Tp is shown in FIG. 36. Following integration, the library plasmids were transformed into E. coli and BAC vectors containing an integrated pRIG-Tp vector were selected for on chloramphenicol plates. Colonies were pooled and grown to saturation in LB/Tetracycline. BAC plasmids were harvested using a commercial kit (Qiagen).

[0652] For each transfection, 20 µg of the BAC library was transfected into 2x10⁶ HT1080 cells using 30 µg Ex-gen 500 (MBI Fermentas) according to the manufacturer’s instructions. At 48 hours post transfection, the cells were placed into media containing 3 µg/ml puromycin. After 10 days of selection, drug resistant clones were pooled and expanded. The expanded pools of drug resistant clones were divided into separate groups for freezing, protein production, and episome amplification.

[0653] To isolate and test activated secreted proteins, culture supernatants were harvested and saved at −80°C until used in specific assays. Activated intracellular proteins were harvested from cell lysates (prepared by any method known in the art) and used in vitro assays.

[0654] To amplify the copy number of the BAC episomes, the cells were selected with increasing concentrations of methotrexate. In these experiments, the initial methotrexate concentration was 20 nM. Methotrexate concentrations were doubled every 7 days until cells resistant to 5 µM were obtained. At each methotrexate concentration, a portion of cells were removed for storage and protein production. Activated secreted and intracellular proteins were harvested from these cells as described for the non-methotrexate selected cells.

[0655] Having now fully described the present invention in some detail by way of illustration and example for purposes of clarity of understanding, it will be obvious to one of ordinary skill in the art that the same can be performed by modifying or changing the invention within a wide and equivalent range of conditions, formulations and other parameters without affecting the scope of the invention or any specific embodiment thereof, and that such modifications or changes are intended to be encompassed within the scope of the appended claims.

[0656] All publications, patents and patent applications mentioned in this specification are indicative of the level of skill of those skilled in the art to which this invention pertains, and are herein incorporated by reference to the same extent as if each individual publication, patent or patent application was specifically and individually indicated to be incorporated by reference.
SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOs: 33

<210> SEQ ID NO 1
<211> LENGTH: 39
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 1

ttccttcaag cttgtcatgg tgtgtgctc aacctgcct 39

<210> SEQ ID NO 2
<211> LENGTH: 40
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 2

aasctt exactgattatc atctttcata tsacttcas 40

<210> SEQ ID NO 3
<211> LENGTH: 28
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 3

atcactacag gctacaggtg agtactcg 28

<210> SEQ ID NO 4
<211> LENGTH: 36
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 4

gatcogaat gatcactgtga gocatgtgg attaa 36

<210> SEQ ID NO 5
<211> LENGTH: 33
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 5

ggcogatct agcogctatat cgogttgatgc ant 33

<210> SEQ ID NO 6
<211> LENGTH: 51
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 6

ggcogatct gtcactttaa gagagcogaa acaagcogctc atgagcogca a 51

<210> SEQ ID NO 7
<211> LENGTH: 6084
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 7

agatcttcaag ttgggcat tagcactatt atctattgtg tatataagat aatcataat 60
tgatgtatgg ccattgca aatggtatct atataaatg atgtcatattt attggctc 120
-continued

tgggcccct ggtgggcatg attatgtgacg aggttatctt catagtaac 180

tcggaattc tcaacacca cctggttctg ggttcagtg cgggtttttc 240

tggcgcctt acgtgtatga acgcgtcgct ttcacatgc cccttcgccg 300
tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 360

tgggcatc tttgggttcgc ctggtttcgc gcgtgttttc gcgtgttttc 420
tggtgtgagct acgttgtgagct acgttgtgagct acgttgtgagct 480

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 540

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 600

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 660

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 720

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 780

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 840

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 900

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 960

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1020

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1080

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1140

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1200

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1260

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1320

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1380

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1440

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1500

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1560

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1620

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1680

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1740

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1800

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1860

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1920

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 1980

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 2040

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 2100

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 2160

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 2220

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 2280

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 2340

tccatcagc ggtgtatcctg ttcgttcatg ggtgtatcctg 2400
-continued

tgagcctaat gggagtttgg ttgagcacc aatctaeccg gaccttccaa aatgtctgtaa 660
caacctggat gccccggcag gttggaagaa atgctgaagga ggtggtgccag tggggaggtc
720
tatataagca gacgctgtttg agttgacgct cagaccaacta gacgctctttg tcgagtgtt
780
tatataagct aatattgcgg ccgagtcgat gttctgacatt cagctcgtc gactatctaag 840
tgagagctc ctgctaaa taactcagcag tctcaactcct gttgctttcg ctgctacaag 900
tctcaactcctgctccttttg cagacgagc tacgacacaa agacgccaa agacgcctaa
960
gttgagcttt ctcgtaaat ttgtagccag ctctcctctg atcttctatg ttctcttccaa
1020
aggtggaccgc tctcagaaag ctcgacattc tspcagaaaaacc tcgagctctc ggtgggagga
1080
catcagctgc gacatcctca gtttcttcct cagctgagtt attgcagctata ttaaagggaa
1140
aaaactttgc gacagaaaat agatcggcc aagactaagc amagagagag gcacacagggaa
1200
aagagcatca taataaaatgt ttaaatgtgaa ctacttgcac attagcatac tgaagagcag
1260
tgctgactgt atctataagc tctcaactatg tgcattcagta gggacaaagag tgcaggaaa
1320
aattttagt gttgagattc aagagaggttc ctcagacaa aagacagcct gcagttctgt
1380
cacacacacg ctgctggcctg agtagtatgga tttatgactg caaacgcataa acagcataa
1440
agatgagaaaa cttcactaac gttcactaata ccaacgagcc caaacggtgaa ccaacagctc
1500
ggtggcacaat ttgcatgca ccagagggac cacaggtcag aagagatcaca gttgtagagc
1560
ctgcttgctg cccattgcac ggtatcgctg tttgctgggctg tctgatcgtccg cggctgggtc
1620
cattcatcagct gttgcttcgct gttcgttcatt aacatcagcc atctacgtgctt gcagttctgag
1680
tcggagagaga aagctttgctg acagttctgtg cttctctcctt tgcgctccttc tggagcagc
1740
attcacaagt tctgtctctaat ctaaataaatg tttgtcttcat tggccataaca atgggtcatat
1800
cattctcctg tggattgcttc ctatttcgctg caatctcctac ccacacatccag cccagagccg
1860
ataagctgta aagcctgggtg tctcaatagt ggagcatacg tctaatatatg tttcggtcgc
1920
agatgcttcc cttctgtatt gcgaacgc tggtaacatctg tggcaggtgagg 1980
ttttgacaca cccacacacg agttcagtgta aaaaaactgt ttaatttgtta aattttgtgt
2040
gtcatgctg tcttactacg aattactatacg gccttaatata acgctgataa caacacctggc
2100
attcatattg tgtctttctgt tttcttctag aatgtggggg ttttttaaatg cacgtaaaca
2160
tctactacgt gttgtaataat cccagatcag cccgactcctg agccgtgatag gcacatgagg
2220
ccaggtgagct gcagcttgaga aagagctggcg gttgtgttgg ttgagcagcag tcgaggcctgca
2280
cacattcctcg ccagcctctct ccttcttcttct cttccctccagc 2340
tgctgcgtt cctccgctaa gttcataact gggctctcttc tttggtttctt gcattactgtg
2400
ctttacggtacc cccagagacc aaaaaaccttg attaggtggt ggtagtcggt gattggtgcgt
2460
cgctggtaga cggctgtttgc ccgcttggtac caagcctcttt ctaatggtgac 2520	
tttgctgtcc aacgagccaa atctcactcct cctacctcttt cttctcctttt gatttttgaag 2580
ggatttgctg ctttgctgggctg tatgcctaa aaataaagact gttcttacat ccaaaatcagc
2640
cgatatttc gaaatatatg cccgatcctc ttcctccggcg gccagctgcc 2700
aaccgctgtgt gcagttgtcag tctagagcgc cccccatcagg cccagcggc 2760
agagatgctg aaactaatc gagattacgt attgcatacg gcagttgggaa gctcctcaggc
2820
tccctcacg gcagagatag ccaacatctat aacattccttgtcgtcagc cctagtcccg 2880
-continued

cvecctact cgcccatcct ggcctact cggcccatcct gcggccactt cggcccatcct gcggccacttt 2940
ggtcatacaa tttttttttt ttttcatgag gcgggcgcgc ctgctgcto ctgcatacac 3000
cagaagact ggggaggttt ttttggaggg ctggctttttt gggagaagc gtggcttttt 3060
ggcacacagct cttgcatcct aagggtagac ccccatcctg cccactgcag sgctgtgacgc 3120
caggtctccg gcggcccttc ggctgctgctt acctgtgggc caagacagcaca 3180
tggctgctcc ggctgctgctc gcggctgcgc ctgggccctt gcggccccct gccccttttgc 3240
tcagacacgct ctcctctggt gcgcctgtat ggcctgagag gcggccactc gcggctgctc 3300
ggcgccccgc gcggcctggt cttgctgctc gcggccactc gcgcctgctc gcggccactt 3360
ggcgcggcgc gtttttgggcc gcccggcgtt cttgctgctc gcgcctgctc gcgcctgctc 3420
cgtgcgccagc agatcatact atggctgctc ggcctgctgc gcggcgccgc gcgcctgctc 3480
cattccgtgcc atctgcacac ccggacagcgc gcccggcgc gcgcctgctc gcgcctgctc 3540
cgcgctgtct gtcgcctgtgc gcgcctgctc gcgcctgctc gcgcctgctc gcgcctgctc 3600
eaccgcttcgg gcgcctgctc gcgcctgctc gcgcctgctc gcgcctgctc gcgcctgctc 3660
gggctgacg gtggctgctc gcgcctgctc gcgcctgctc gcgcctgctc gcgcctgctc 3720
ggcgcggcgc gtttttgggcc gccccgctgc gcgcctgctc gcgcctgctc gcgcctgctc 3780
cgtggctgctc gtttttgggcc gccccgctgc gcgcctgctc gcgcctgctc gcgcctgctc 3840
cgcgctgcgc gcgcctgctc gcgcctgctc gcgcctgctc gcgcctgctc gcgcctgctc 3900
ggggcgccgc gcgcctgctc gcgcctgctc gcgcctgctc gcgcctgctc gcgcctgctc 3960
tactttattt ttttatcttttct ctcctttttt ctctttttttttttt tactttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
cyttctcgtag tttctctacac gacggggagt ceggcaacta tcggtggaacq aatagacag 5220
tagctgaga taggtgccgc aotgattaag catggyaacc tggctcaagca aagtttca 5280
tatatacttt aaggtagttt aaaaacctctt ttttaatatta anaggtatca ggtgtgactc 5340
tttttgtgata ctatctagtc gaaatctcctc ttactatggtg tttgttccaa ctagtctcag 5400
gacccgctag aagggtaaact gagaatctct tggagatctct ttttcttctg ctgaactctc 5460
ttgttgaana caaaaaacaac acgcctacca gcggcgcttt gttgctcgga tcaagagctta 5520
ccaaccttcttt ttccgagctt aactgtcttc agcagacgcc agatataacaa tctgctcctt 5580
cattgtgacg ctagcgagct cccacacttc aagactctcty tagcagcgcg cctataacttc 5640
gctctgttact cttcgctacag atggttgctgc gctcgttgcg ataatctgctct tcttacgagg 5700
ttggaactoa gacgtagcag atcgatattaag gcggagccag cgggcttgac gggggtctcg 5760
tggccagaac ccagcttgga gggaaacacac tcaaaactacgag tggatatcct acagctcgag 5820
cattgagaac ggcgcactagt ttcggagacgg aagagctccg acaggtatcct gcgaagcgcg 5880
eggctgagga caaggagccg caaggagggcattcagggg gaaarccgtct tctgatcttat 5940
tagctctagct ggttttgcacacctgcacttt ggggctgtat ctcggctcggg 6000
gggcggaggc tgtgctagaag cgcagaccaag gcggcgcttt ctggctcttc gcggcttttcg 6060
tggctcttgc ctcacagctgc ctagc 6085

<210> SEQ ID NO 9
<211> LENGTH: 6086
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<400> SEQUENCE: 9

agatotccaa tattggccag tagcctcttt attttatgtg ttataagcag ataatcactat 60

ttgctcagttag ccattcgtca gttttgatct atctocatatt atgtcatttt attttgtgctc 120

tagcctaatg ccggccacgg ctggctcaattg atttgaggct attatagctct atagtaacat 180

tcaggggtcattg tattccccag gcaccatatt gggctcctcc gcactactac ttcgctataa 240

tggccgctcg ggttcgcgcag ccaaggcacc ggggctcttg acctcataa tgcgctagtt 300

tccacatagc aaccacatag gtccatctga ttggctgatct ctctaggtctt cttactgctat 360

dagggccgac gttgggttag agtcgctcag ctgggcgcgg cttccccagtc ttccagggagt 420

daggtatcctg aatagccggc ccggccacgg cggctcctatt ggttctccag atggctgtcg 480

tacttgccag taattacactg tttggctgtag cctattacag gtttgctggtg gttttggcct 540

gttccacactg gggctgttag atggccgttag tctccaggtgtt cttccagctg ttcaccacacctt 600

tggctgactg ggtgggttgc tttggccaca aatgcaacag gacctgtccag aatgctgtaa 660

dacagccgct gcggccacgg ctggctgatctg aatggcgctg ggcgctgatc tgggggcgtc 720

tatatatagc gaggcttaggt gttggctgtag gcacacagat gaaagttttc tgcgctgagt 780

ttccacagctg aagggcagtt ctcggctcag cttccagctg ctggctgtaa 840

tgagatcctga ctctgctggtt tttgcctgcag cctccagcatc gatctcaagccttcagccctg 900

ttcaccactg gttttttgct ctgctgtcag atggggggcag gcaccacctcga 960

gttgggttgc tttgcctgtag cttccagctg ctggctgtaa aatgctgtaa 1020

dagttgggttgc tttgcctgtag cttccagctg ctggctgtaa 1080
cattcacttg gacattctta gttttcasaat ggtgtagtac attgacgata taasatggga 1140
aaaaacctca gcaacaaaaga aagttgcaga atacgaaaaa gagaaagaga ctctoaaaga 1200
aaagataca ttaaaaacct ttaaaaaattt gacctctgaa attaaccaca tgaagccgga 1260
tgtactgat atctcaaaagt taataataata taataacaaga ggaaagaaaaat gttgyggaaga 1320
aatattttga ttaggacatt aagagagagtt tcaaaacccca agatctcctt ggaacctgat 1380
cacacacacc ctgactctgtg aggtaaaggag tgaactgac ccgcgatttaa aactgdgtaca 1440
agatggggaa cactctaaaa ctctctcagag ggtctcatca acaagtgtag ccacccagct 1500
gagtgaacaa ttcacaggca ccagagggaa caagagcaga aagagatcga gttcctagcc 1560
tgctcgctgt ccaagagaaag ggtctcaacag gtgagtggg cccagtctct ctctagctcga 1620
gccctcttaa ggtgcaacag ttcaagaca gtttaaaggga gcaacataa aacctagcctt 1680
gctgagacag agaagactttc tgcgttctct gatggcccat attggtctta cgggagccgcg 1740
aatcacaacc tgcgacatt tattgycctt cccaaataac ccggcgttaa ctctgtcata 1800
tctgtctcct gttggaatatt gtgcttccct ccaacttaa ccaacacta cgcggcagga 1860
cataaaatgt aaacttcggg tgtcataatg aqtagagcata ccaactattaa tttggtctgcg 1920
cagggctcct attttcaggag tgtaagwtatgt tcggagagac gatcagagag actatttga 1980
gtggcagaac aacccacaaa aagttgagtg aaaaaatgc ttttatcgtt ataatgttga 2040
tgctatgtc tttacttggaa ccaattttaa ctcgcaaaaa caagatcaca cacaccaatttg 2100
cattttattt attggcaagg ccgcttaaaag gttgycggagc gtttttaaat ccgaagtttaas 2160
cctctctaaa atctgtttaaa tccgataaag atgacagtcg gacgtggaatt ggcctaatgga 2220
cgggcactgt agggcggcat taagagcggc ggggttggtg gttacgacga cgtgacgctg 2280
actggccgtag gggcgtctct ccggactcct ttcggctctct tctgttttcct cttgtcgaac 2340
tcgccggtct tttcctctcgag acgttttaaa ggggcgtcct cttcgggttt cgtgatgtat 2400
gtttaacgq accttgcaac caaaaactt gattaggggg atgttctcag ttagtgccca 2460
ttgctgctgt agaagtttcctt tggcctttctt aggctgagat ccagcttttt taaaagtttga 2520
cctgttgcctcc aactctggcc aaccacaaa ctcctctctgcttctcctc tttgattaa 2580
ggggttttcg gattcggcc ctattggtta aaaaatgcg tatttttaaa aacaacttaac 2640
ggaatattta acacaaaatg aaggtcctaa attgacgctg ttaactttct cctggyggaaaa 2700
gacaagcgtg ttggaagtgt ggtaagttgg gttgggaag cccagagctt cccacgagggg 2760
cacgattatc cccagctatg atctcsgccttt caacagctcg gcagttggcga atcctccgaaag 2820
cctctctgact ggctaacgta tgaacagcat gctatctcatt gttcctagcag ccaactttcc 2880
gcgcccttact ccgcccctcct ccgtcctctcctt ccggtgctt ccgcttcccaat 2940
tgtgcctgac aaaaaaaaaaaa tttaacctca ggccagagcct cccctggtct ttcagctatt 3000
ccaactattg tggagggcgt cttttgccttct tcaactattt cagtaaggra ttgatttccct 3060
tgaaacacca cgcttctcag ttacctgtag aacgacctga ccaaaagaag tyagggtcagc 3120
gcagctttctg ccggggttct gttggagggag tttaccggtgt aacatgggcc aacccaacaca 3180
atcagctctg tctgtccgct cgtgtccctgg ctctccggcgcc caggcgggccc gttcttttctt 3240
gctagaaccg actgttcggc tcggctcctt gaaagagcgg cagagcagc ggggtcctag 3300
tcgctgtggcc ccggcgggtt cttcctggtct ctgtgcttcct ccaaggggga 3360
<table>
<thead>
<tr>
<th></th>
<th>Sequence</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>3420</td>
<td>agggactggc tgcatttggg cgaagtgccg gggcaggctc ttcgctctc tcaaccttggct</td>
<td></td>
</tr>
<tr>
<td>3480</td>
<td>cctgcgcaga aagatcatct caggtgtat ggcaagtgccg gggcactgcag gcttgatgcg</td>
<td></td>
</tr>
<tr>
<td>3540</td>
<td>gctaacctgc accagcaca ccatgccaaa acatgcagct agcgaccacg tacctgctgag</td>
<td></td>
</tr>
<tr>
<td>3600</td>
<td>gaaagctgttc atgctgcata caggtgtacg gaggaagagc atacggtgtc cggccgacgc</td>
<td></td>
</tr>
<tr>
<td>3660</td>
<td>gaacgcctcg cccgctcata ggccggtgct cccgacccgc ggtgtgctgc acatgcagct</td>
<td></td>
</tr>
<tr>
<td>3720</td>
<td>cccggtgacgt gcacgtgcac atacggtacg gaaatgccg cgttctcttg attcacatgc</td>
<td></td>
</tr>
<tr>
<td>3780</td>
<td>tgctggtgcgc tcgtggtgcgc ggaagctctg cccgacccgc ggtgtgctgc acatgcagct</td>
<td></td>
</tr>
<tr>
<td>3840</td>
<td>cggagacgc ttcggtgcgc atgggtgcac gcgtctctct gcgttctctct gcgtgtgctgc</td>
<td></td>
</tr>
<tr>
<td>3900</td>
<td>ccagctgcag gcgcagctgc accagcaca ccatgccaaa acatgcagct agcgaccacg</td>
<td></td>
</tr>
<tr>
<td>3960</td>
<td>tgggttcgca aatgagccac ccatgcagct gcacgtgcac atacggtacg gaaatgccg</td>
<td></td>
</tr>
<tr>
<td>4020</td>
<td>atacgttat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
<tr>
<td>4080</td>
<td>ccaatctttct cctgtggtgcgc ctgttttgtct aatggttatatt aatggttatatt</td>
<td></td>
</tr>
<tr>
<td>4140</td>
<td>ttcggacgttc cggcagctgc accagcaca ccatgccaaa acatgcagct agcgaccacg</td>
<td></td>
</tr>
<tr>
<td>4200</td>
<td>cggcagctgc accagcaca ccatgccaaa acatgcagct agcgaccacg ctcggtgcgc</td>
<td></td>
</tr>
<tr>
<td>4260</td>
<td>accagcaca ccatgccaaa acatgcagct agcgaccacg ctcggtgcgc accagcaca</td>
<td></td>
</tr>
<tr>
<td>4320</td>
<td>ttcggacgttc cggcagctgc accagcaca ccatgccaaa acatgcagct agcgaccacg</td>
<td></td>
</tr>
<tr>
<td>4380</td>
<td>ccaatctttct cctgtggtgcgc ctgttttgtct aatggttatatt aatggttatatt</td>
<td></td>
</tr>
<tr>
<td>4440</td>
<td>atacgtatat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
<tr>
<td>4500</td>
<td>ttcggacgttc cggcagctgc accagcaca ccatgccaaa acatgcagct agcgaccacg</td>
<td></td>
</tr>
<tr>
<td>4560</td>
<td>ccaatctttct cctgtggtgcgc ctgttttgtct aatggttatatt aatggttatatt</td>
<td></td>
</tr>
<tr>
<td>4620</td>
<td>atacgtatat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
<tr>
<td>4680</td>
<td>ccaatctttct cctgtggtgcgc ctgttttgtct aatggttatatt aatggttatatt</td>
<td></td>
</tr>
<tr>
<td>4740</td>
<td>atacgtatat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
<tr>
<td>4800</td>
<td>ccaatctttct cctgtggtgcgc ctgttttgtct aatggttatatt aatggttatatt</td>
<td></td>
</tr>
<tr>
<td>4860</td>
<td>atacgtatat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
<tr>
<td>4920</td>
<td>ccaatctttct cctgtggtgcgc ctgttttgtct aatggttatatt aatggttatatt</td>
<td></td>
</tr>
<tr>
<td>4980</td>
<td>atacgtatat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
<tr>
<td>5040</td>
<td>ccaatctttct cctgtggtgcgc ctgttttgtct aatggttatatt aatggttatatt</td>
<td></td>
</tr>
<tr>
<td>5100</td>
<td>atacgtatat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
<tr>
<td>5160</td>
<td>ccaatctttct cctgtggtgcgc ctgttttgtct aatggttatatt aatggttatatt</td>
<td></td>
</tr>
<tr>
<td>5220</td>
<td>atacgtatat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
<tr>
<td>5280</td>
<td>ccaatctttct cctgtggtgcgc ctgttttgtct aatggttatatt aatggttatatt</td>
<td></td>
</tr>
<tr>
<td>5340</td>
<td>atacgtatat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
<tr>
<td>5400</td>
<td>ccaatctttct cctgtggtgcgc ctgttttgtct aatggttatatt aatggttatatt</td>
<td></td>
</tr>
<tr>
<td>5460</td>
<td>atacgtatat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
<tr>
<td>5520</td>
<td>ccaatctttct cctgtggtgcgc ctgttttgtct aatggttatatt aatggttatatt</td>
<td></td>
</tr>
<tr>
<td>5580</td>
<td>atacgtatat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
<tr>
<td>5640</td>
<td>atacgtatat ttgtattctg cgttctctct ggtttttttgt ggtaagaccc cggctatgctg</td>
<td></td>
</tr>
</tbody>
</table>
-continued

cgcttgctc atctggttaa cagtgagctgc tgcacaggtgc gtaaggtcgg gtctttactcg 5700
gtgcaactca aacgaaataa taccagtaaa ggcggacgag tgcggctgaa cggggggttc 5760
gtgccacag cccagcctgg agcccaacgc ttcacccgaa ctgagatacc tatacgtgaa 5820
gctatggaag gcgcacagcg ttcgctgaggg gaaagggcag gaaagttctcg cgttaaacgg 5880
caggggtgga acagggagac gcacggagga gttccaggg ggaacgcctt ggtatctttta 5940
tagctctgc gccttggcgacctgctgact tggcgcctgaa tttttgtgtat gctctgcaggg 6000
gggccgagc cctggaaaag aagccgcaaa cgccgctttt ttacggttcc tggccttttg 6060
tgctgcttttt gttcactcgg ctctgg 6086

VENTUS: 10

<210> SEQ ID NO 10
<211> LENGTH: 38
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (29..(32)
<223> OTHER INFORMATION: a, c, t, g, other or unknown
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic oligonucleotide

<400> SEQUENCE: 10

TTTTTTTTTT TTTGTACGCG GCACGATNMM NTTTTATT 38

VENTUS: 11

<210> SEQ ID NO 11
<211> LENGTH: 25
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic oligonucleotide

<400> SEQUENCE: 11

cagctcaca gaaacgttatt tgcgg 25

VENTUS: 12

<210> SEQ ID NO 12
<211> LENGTH: 20
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic oligonucleotide

<400> SEQUENCE: 12

TTTTGTACGCG GCACGATNMM 20

VENTUS: 13

<210> SEQ ID NO 13
<211> LENGTH: 45
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic oligonucleotide

<400> SEQUENCE: 13

ACTCTAGGCG TATCAGGTCT TACTACTTT AAATGGCTAA CGCCG 45

VENTUS: 14

<210> SEQ ID NO 14
<211> LENGTH: 43
<212> TYPE: DNA

<400> SEQUENCE: 14

Some DNA sequences are shown, with annotations indicating the type of sequence and the location of the features within the sequences. The sequences are described as synthetic oligonucleotides.
-continued

<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic oligonucleotide

<400> SEQUENCE: 14
ctcgtttagt ggcggcgtct agaacctga atctgacagc tgt  41

<210> SEQ ID NO 15
<211> LENGTH: 41
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic oligonucleotide

<400> SEQUENCE: 15
cctgtttagt ggccggcag atcaagtgaat tctcaggacc t  41

<210> SEQ ID NO 16
<211> LENGTH: 22
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic oligonucleotide

<400> SEQUENCE: 16
gacccactga ttaacgca ta  22

<210> SEQ ID NO 17
<211> LENGTH: 20
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic oligonucleotide

<400> SEQUENCE: 17
tcgccgagat tcagtgactc  20

<210> SEQ ID NO 18
<211> LENGTH: 6836
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 18
agatcttctag tattgacct tatacaatttt atcaatgac tataagat attac  60
ttgttagtt gcacctatat cgttataatt atatataactt attaggtggtc  120
atgccccct gcccaggt acttggtact tattgactt attagattta gtttagagt  180
tagcccagat ctgacctatat cgttataatt atatataactt attaggtggtc  240
tgcccaggt agatccgact ttaaatgatt cgttataatt atatataactt attaggtggtc  300
tctgagctgt cgaattccttat tggccgtctt tggcataagt gcttggtgtc agtac  360
aatgccctga tggccagat ctgacctatat cgttataatt atatataactt attaggtggtc  420
cagtgagct tgtggcagct gagctgcttt tggccgtctt tggcataagt gcttggtgtc agtac  480
tactgagct gatgctctgct cccttttttt cgcgggttct cggctttttc  540
gatgctagct cgtgctctgct cccttttttt cgcgggttct cggctttttc  600
-continued

tgacgctaat gggagtttgg tgtgtgccaa aacctaaacgg gactttccaa astgctcttaa 660
cacactgat cggccgccca gttggcagaa atggcgggta gggtgtcgcg tgtggaggctc 720
tatatagca ggtgctgctt aagtaacgct cacatcacta gaaactttatt tccggtgtatt 780
tatatggtt aatattctaa cggagcgcgatt gttgtgccaa caacgaattg caaacatctt 840
tgacgtgctt ctcttaaacca cacccattgtct caggtgtcgaat actggtgctct acggctatat 900
gctgtatgct aatccttatt cggcctgccct cggtggaccc cttgacgcc 960
cggcagacggc tgacgtgcttc gctgatggga gccaatcagc aactaacgat caggyccgcc 1020
acaacccgac tggtgatcct caacoaaaga ccagctggtt cggccatacc cggcgcgcac 1080
ggtggggttg ccggtgccca tatacgcgac aatcagcgggt gggagacgac gggtgacgcc 1140
tccagggtc cggcgtcgtg ttctggcctc cttgtaaggtc cacatcctgg ctggactcgcg 1200
cacatcatttata cttggtacag cttcactcct cagatcctgga gttggtattg tggacocctaa 1260
gtacactctc ctctgctgcc gcatactgtg gcaagcttct cttcctgcgt cttggttgtagg 1320
agagctgctga tgctgctcgg agccaaaaattt acgctcaaac acggcaccgg tggacocca 1380
atggactgtgaa gaggttagcc gttgggtgcct gttgagcttag gtaaggcccc 1440

gtataacgct tataagggtgc gatgggttag tctggtggagg ccaagggggtg cttgggtgtg 1500
acacggtctag gctgacccct ctaatacgatt acggtctgct cggcctcag tgggacaatt 1560

tgactttctcat ccaactaacgc cttgtaacgc gtagaagcag acaacacgc 1620
cctacacagcg gagaaaaacc acggctggac cggacggtgg gaagtaacgt ggtacggtcgc 1680
tgcttacttc gaagggcacc acaagtgttc gacatagttc gggacgcacc ctaattcccc 1740
cattgcccaag ataattgtct ctaataagct actagctata ctaaaacggc attggacactt 1800
tccaccaaat tgggtgctgcct tccaaagctgg acctaccgct cctagcgtgat 1860
ttttttgcga gttgggctag cggctgctacgtgactgttag cttccacaa 1920
tggccacgg ccagaaaaagc gatcggccct ggcacagcct caggtagcga ttcacgattt 1980
tttccagct gcacccacac cttcctcgata acggcactgt ctgcttagtct ctaatgtagta 2040
agaaagcgct gttccctctcg atggagcaacatg gatccggactat gccagtttcttggc 2100

ttcacacac ggcaacagac cggcactcgc aacggtggtgc tttcttctcc acgagtagct 2160
agtgtgtctt tgcaatctcg gcacactccag aattagcccc taaatgagacttggcgctgctg 2220

tgtgtggag cttgctcttg cctataggaac cccatgcaag ccaaggggct ccattacaatt 2280
ttgtaacggt gcacgctcgcc gacttccaga gttcagctggct cttcccggtt atggcttggtt 2340
agagatcata gcttcctactg cggcctgctcct cttggttgctc tgtgttcctgt gggagagggc 2400
ggtttgtaagc cccatgcat gcggcagtcg tgtgtgctcct cttggttgacg cggagttggc 2460

tttgacgag tgcggctac gcgtggctcgc gcggagacgct cttggttgacg cttggttgacg 2520

cctaacacgc tccctctctgt tgggtcgata ccccccttat tttttttttt tttttttttt 2580

caggaacagc cgggacgccgc ggttgcctgct cttggttgacg cttggttgacg 2640
agagatcata gcttcctactg cggcctgctcct cttggttgctc tgtgttcctgt gggagagggc 2700

ttggtggag cttgctcttg cctataggaac cccatgcaag ccaaggggct ccattacaatt 2760
aatgtagctgctcct cttggttgacg cttggttgacg 2820

cctaacacgc tccctctctgt tgggtcgata ccccccttat tttttttttt tttttttttt 2880
5220
5280
5340
5400
5460
5520
5580
5640
5700
5760
5820
5880
5940
6000
6060
6120
6180
6240
6300
6360
6420
6480
6540
6600
6660
6720
6780
6836
actgoccact tggcaatca acaatggatat ccataatgcac gctcgcoccc tatctgactc 420
aatacgccta aatagcgccg cctgcatatt tgcocatca tacacctag gcgttccttc 480
acttggcagt acctatcagt attttgcatc gtatttacca ttgggatgct gtttttgacg 540
taccaaacg ggcgctagta gcggctggtc tccagggatg tccccagctt ccaccaccat 600
gacgtctaat ggaggattgt tttgcaacaa aatacagcgg actttcacaat atggcctaac 660
aatgcctgac gcgggccccc tttgacccaa tgccggttag gccgctagcc tcggaggttc 720
atatacagc agatctgattg ttgacacgct gatatactgta aattctgca gctatatgttt 780
aaccggccta gacgcctctc tcagcatctg gaaatcctta ttgacgtag tttatcaagt 840
taaatggcctg aacgctgctc acctgtgctg aacactttga ctgtcgagtc gtttctcgtg 900
tctctttaac ccctcctggc taccagtggt tctctgctac ttaaccagag gcctatcttg 960
ccagtgctac gtcgaaaag gaagttttaag agagcgcgaa caagcgttca tgacccgga 1020
gtcgcgcgc cgccttcctc catgcgtttg gctgctgcta gtcgggcccc caacgcgcac 1080
tgctggcgcg tggatggcgg caagctgctg tcggcctgtag aggctcaca gaaaggggtg 1140
ggtccgctag atcgctagtc gctagggcag tccagtcagc gaaacgcgaag gacgtggcgcg 1200
gccggccag ggggctggca gtcgagcag aacgggtgag cttgtggat gataggggacg 1260
atatacgcg atacttcgct gattgctgag atctgtgact aacatgctag aacagttgac 1320
caaaagaa caaacgctt aacagggccc agttttgcgc cttttcgctg tttcccattg gcctatcttg 1380
cctgagctac atcaccacaa tccagcctca agttcggagt ggccgaacc cagacgttga 1440
taaagattc acggcttttc ccctgggaac tccctgctgg tgcgaactgg 1500
ccgctcttcg gatatactgc cggcttcctc cttcgggaaa gctggtgcct ttctctatog 1560
tcagcctgct gccttccttc ttgctgtag tgcgctgctc ccagctgagg cttgtggtcg 1620
gaaacggccg ttcaagcctg gcttggcgct tttccgctct aacatgtctg tggattcgca 1680
cggataagg aagactttac gccttcgctc gacacctgca gactggagct ttcacagctg 1740
agagttgctgc ggcacgcttc aagagctctg aacagttggc cttactacgg cttcctctga 1800
aggccccagtt tgtatctgct gcggctgcct aagccctgta cttccggaga aagaggtgtg 1860
agctttgtag cggagccaca aacccaccc ggttggcttg gttttttgtg tggacgctag 1920
cagatatgcg gcggcggcgc aggatcctca gagatccttt tgcctctttc tccgggtttg 1980
gagtcccgt ggagcggaa perggccgtacctg aagttcaggatt aacaaaaag 2040
agcttctcgg ggttgggag gggtgctctc gtcgctggct tttctctctc 2100
cggccttgg cggacctagt cggcggagtt tcggcatctc gttctcagaa cggctttaga 2160
gggctggcgc tggatacgct gcgcagcttc ggtgctggag cctgggaaac gttccttaga 2220
tccggccag gcgttccatt gcacatcgag ggttccgctc gcctgctgctg gttcctgctc 2280
cggcccttc gcggcttcgc agtctgtaga tccccagggc ctgcgctttg ccaccgctctg 2340
atccggcctcg cggctctgct gcctgtctcg ggtgctgctg gctgctgctg 2400
cgtggcctgc cggatcgcgg ggtgctgctg tgcctgctg tgcctgctg 2460
cgctctgctgc cggcttctcg cgccttctcg ggtgctgctg gtttctctctg 2520
gtctgtgacttc ggctgtcagc ggtgttttctc gctgtttttc tggacggctc 2580
gatgataact ttctggcctg gcggacaggt agatgacgtc gagctcctgcc ggctgtcttc 2640
-continued

gatcttcctt atggcccatt agccattatta ttcattggttt atatacataa atccaaatatt 60
ggatatggc catccagcact gttgtatctta taataataa tgaatcatat tataacctta 120
tgctcaatat gacccgacatg ttggccatgga ttatgacctga gttaatatata gtaaatcatt 180
agccgcttac gatttcagcct cacataatcct gcacccgctc tttaaactct actcagaa 240
ggcgggctcg gcgtgacgcc caagacccgg cgccggattga ggttaaatggt gatgtatgtt 300
ccagtaaga cgccactaga gatctttcct tgcgtatcct ggtggatcaggttattcatt 360
actgcoacact tgcgaatcata tcaaggtgtat cctatagccaa gtcgcccccc tattgacgc 420
aataaactca taatgcoacgct cgactttcatat ggcgactgca taacacttacag ggccatttct 480
actggtcagt acactcataact attagccatg gctataacca gttgtagtggc gtatttgccag 540
tccacacatg ggcgtctgtga gctgttggatg cacgaaggaat cttaaaagtct ccacccatt 600
gactcactag gcggttttgtt tggccaoaas acatcaaggg acttctcaaa atagttgaca 660
aactgcgtac ggccggccgg cgccgactgaa ccggccgatcc gcgtgtgagcc tgggaggtttt 720
atataagcag agctcttgaattgctaatcag agagtttctat ggcggttttt ctgggaggtttt 780
atcattgcatt aataactcaac gcagctcagtt cttccgacac aacagtctccg aacattaagc 840
gccgacact tcttttaacctt ccacccgctaga cgggtagtaa cttcagatcct taagagggc 900
catatgacg cagttcagct gagaagataa agtttaaagag acggcgacac agccgtagatc 960
agccggtatt gcggagggcc acttttccaa cttggtataggt acggatattc gccggccaga 1020
agccggattg tgcagccgctg gctgcctagtc caagcttgggt ccggcgagcc gatcagccagc 1080
agggccggta cggcgttactgc atagttggtgct taaagccacga aagagyagcag 1140
actgctgacgc gcgaaagacag gctcggcagc gcgtgacacag cgggctgcaac taaaacttgc 1200
aatacgcctat atcctctgcaat gttctgctagcc atgtgcgcgcat tttgtggcgta 1260
aacgggcaaa acacagccag aacagctttgg acggcagcttt ttcgattaacacacac 1320
tccggcccccc ctcggagccag ccccaactt ccaggcttatg cgggctttgg cggggggc 1380
cagctgctata aagcttacgc gcggcctcct ccctgacggtt cttctgctgctg 1440
cggccggatt cttgcgggag caactcctcct tcggggtatt gcggccggctt 1500
cctttgcaggatatagc tgggtggttt ccctggtgatt tcggttggctt 1560
gttgctgtag caacccagtt ccagcttgacgc gcgtgcggctt atccgtaacg tatcgtctct 1620
aggccgatc gcgtctagc tggagtcttct tcgatgactg atcagatctg gtagctggctt 1680
gccggggtt cccttctctt ccactcctgtg ccgtgagtaag cgggtttggc gctttcttcct 1740
acacttgaag gccaattttt gcgtgactgt gctgctgatatt gcagcttggag ctcggggggg 1800
aggctgctgt gcggatccct gccgagccttt ccagcgtcgg tagctttggct ttttcttggg 1860
gcgtccccag gcgtgtggag agagagagag ggtgcttcgag agctcttttg tagctttctca 1920
cggttccttg cccttcgcttg gcggcgctttt ggtcttgggtt gctgggtgtt atctgagttgct 1980
cgcagaccaag gcggcttgcg atcttttttc gcggcgccct gggcggcttg ggtctttctc 2040
agcgccactgc gtcggcctcg caatgggctc cgggcggctt ccggggaagc ctcggggggc 2100
cgaagctggcc gcgtttgctt ccagcggcct ctgcttggctt ccggggagcgg 2160
tccggcgctt ccagcgccag gcgggctttg tgggtttcct gcggctgcttt ccgggagcggc 2220
tcgggtgtc ccgagcctgc gcggggtttg tggtggggttt ccgggagcggc 2280
-continued

cacctgtatt tggcacaaga ggcacagcga tgggtccecga cggagcttctg gcgcgtcggcc 2340
atgctgcct tgaacctggc gaaagtttgc gctggcgcga ggcacgtgag ctctctccgc 2400
agatcactct gctacacaag aacggtctcc atccgagtaa gttgctgctc gatgcgcggc 2460
ttcgtcttggt gcacaacagc gacgatagcc gtcacagaa caaatgcagc gcaatgcgcac 2520
tccgacatga ttgatctctt ttcggaaga gcaacgtgag atgacagagg atccagcccc 2580
gggcttccgc ccaatagccg cccgctccgt tccgcttcag tgcacacgca gacagcaagc 2640
ggcacaggaac gcgcctctctgt ggcacagcac aactgaggcc ctcgctctgc ttcgagtcga 2700
tcagccccag gcgcagcgct gcctctccgc aaaaagaccc gcggccctctg gcgtgacagg 2760
cggaacaaag gcgcagcagc gacgcttctgc gcctgtcttg gcgtgagctct cgcgtatgca 2820
cctgccaccc aacgcgcagc agaaccgttg gcgtgatcct cttgcagagc ctcgaaagac 2880
gcatctcact gcgcgctggggcc atacagacaa gtcgcocctg ccgcagcttg aaccgcttgt 2940
ggacacgcgc cccacccctt tttgcaggggc tttgcaaccc tacacgctaa aaccgcttgt 3000
catgggaaac cgttcacact cgacggcagaa gcaccacgct gttgacagtg gtcgagtgtag 3060
ggtgatgaaaa gtcgccacggcc cccacgcaag cgcagatggc gcacacgctt catcatcatac 3120
agtcacagcac gcaggtgtggc aagtcctcccag gtcgccacggcc aggcaccagtt atgcagacaa 3180
tgcacatcct ttaagcagca acacatgctgc gcgcaccaac ctiocgcagct cgcgccttac 3240
cggccgctcc ttcggctccct cttgctgcct cttgctgatc atctgagccc aaccgctgatc 3300
aggcgcagcg cgccgcttgcct tctgagctat cttcgagcta gttgacccgg ctttttcttgag 3360
gctctggttc ttcgacaaaaa ctcttccttct ctaacacacac agtcacagata ttcgaaccag 3420
agtcacagcac atagagctgc ctgacggcttc cccgaagccc gggccggttc ggggctggag 3480
ctcttcgcatt ctgtgctctgc cctgacacaa cggccgccag gcgcctttgc gcgcctttgcg 3540
gctgtccgct cagggccggt gcgcctttgc gtcgcctttgc gcgcctgttg gtcgcctttgaa 3600
tgaactgcag gcggccgctct gcgcctcatt gtcgcctttgc gcgcctttgc gtcgcctttgaa 3660
agtctgtgcc gcgtcctgtca ctaagggccgc aagggcatcg gttcagttgg gcgcctgtgc 3720
ggggcagagt gcgcctctgc cttgctcctgct cttgacccgg ctcgacttaca cttgacctgtca 3780
tggagctgg gcgcctcctct gcgcctgtgac gcgcctcttc cccgctgaca acacaagccg 3840
acatcgaatc ggcgcacgcc cggctgtagct gcgccttacc gcgcctttgc gcgccttacc 3900
ggcacagag cttgacggtgc gcgcctttgc gcgcctttgc gcgcctttgc gcgcctttgc 3960
gccgcgctcc gcgcctttgc gcgcctttgc gcgcctttgc gcgcctttgc gcgcctttgc 4020
ggggcaagcc gcgtttctggc gattctatgc cttgacccgg cgccgctctgc gcgcctttgc 4080
tgcagcagct gcgtgcttctc cnaagccgcc gtcgcctttgc gcgcctttgc gcgcctttgc 4140
cggccgcatgc gcgctttttgc gattctttgc taccagttcc atctgagctgc gcgcctttgc 4200
cctctcctgc gcgctttttgc gattctttgc taccagttcc atctgagctgc gcgcctttgc 4260
tgctgtgcag gcgctttttgc gattctttgc taccagttcc atctgagctgc gcgcctttgc 4320
ctctctcctgc gcgctttttgc gattctttgc taccagttcc atctgagctgc gcgcctttgc 4380
agatgtgctgc agtgagccgc ggcgcctttgc gcgctttttgc gattctttgc taccagttcc 4440
cgcctttttgc gcgctttttgc gattctttgc taccagttcc atctgagctgc gcgcctttgc 4500
atccagaatc atacagttcc atctgagctgc gcgcctttttgc gcgcctttttgc gcgcctttttgc 4560
-continued

egggggacta aagtaatttg gtggeaatgta tcctctgccct ttaactggga aegatgtcttt
4620
gattgagaga atatgaatttg aacaatcggaa acacatcag aacatctgtt cctctgtcag
4680
gcaaatgatt ttaaaggtgg tcaagaattcc aagtttgtct gtttaaaaag acaacagcaag
4740
tatggtatct aacgccagact tttggtgtt ttagaatccta gaaacagttg ttttgaatag
4800
tgcctgttaa accaataagac acccagaggg acttaacggtgt tttggtctgt aacagatgag
4860
tgcgttaaaa taatcacaag acctacggaa ctttccacacc caaactagca
4920
catggttcttc atatctctgc tgtgtctggga gctctggyggac ttttccacacc caaactagca
4980
caggccacgc attgttctcc tacgcttcac gtaacgattgcc acacagatgag ctttccacacc
5040
actatggttgtt taaaatccttt gtagaatctct gtagaatctct taaaatccttt gtagaatctct
5100
cagcgcagc caacacccct taaaatccttt gtagaatctct gtagaatctct taaaatccttt gtagaatctct
5160
ttcagagttt gaaacagttg ccaactattaa aagaactgtgg ctttaacgctt aacccgac
5220
acacccctta tccggtccagt ggccacac
5287

<210> SEQ ID NO 21
<211> LENGTH: 5382
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (890)
<225> OTHER INFORMATION: a, c, t, g, other or unknown
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (1042)
<225> OTHER INFORMATION: a, c, t, g, other or unknown
<400> SEQUENCE: 21

caacataaa tctggtgcta atatctttttt aaaaattcgcg ttaaattttt gtaaatcctg
60
cctatatttt taaaccattgg cggaaatacg ccaatatcct taaaattcgg aagaaatcag
120
cagacagggg aggagtgggttt tttcaggtttg gaaacaggtt ccacatttaaa aagaactggg
180
ttcacacgct aacccgacgg aacccgctca tccggtccagt ggccacac
240
cacataaatca actttttttgg tggctaggttg cgtaaagcgc acatacttgga aacaagatgg
300
gagccacgct tgtgaggttg gacccggtta attcgcggta acagggcgcgc gttccctgctt gtcgggtcagc
360
gacgccaa caaaccacccc caaaccacccc ccacgcgccg cccggtctct tcccacgctt ccccacgcgc
420
cacccacccc cacccacccc cccggtctct tcccacgctt ccccacgcgc
480
ggcacaggtgt cgcggtggct gttcagaggtt gttccctgctt ccccacgcgc
540
aggggaggtgt gttttagggt gttttagggt gttttagggt gttttagggt gttttagggt gttttagggt
600
tggttaacag acagctaggt tatttgttaa cgagccgacta tagggcacta tgggtcacta
660
tcatcttgcc ctaaccgccg cttccggtggc cttcctccag cttcctggcg cttcctccag
720
ggcacaggtgt cgcggtggct gttcagaggtt gttccctgctt ccccacgcgc
780
cgacccacccc ccacccacccc ccacccacccc ccacccacccc ccacccacccc ccacccacccc ccacccacccc
840
ggcacaggtgt cgcggtggct gttcagaggtt gttccctgctt ccccacgcgc
900
ggcacaggtgt cgcggtggct gttcagaggtt gttccctgctt ccccacgcgc
960
ggcacaggtgt cgcggtggct gttcagaggtt gttccctgctt ccccacgcgc
1020
ccttcctggcc tccagaggtt gnnaggtgtgt cggccgctag gggccgctag
1080
---continued---

```
gggcggggcg ggccggccgca cgggtcttcgc ggcggccggc ctctgcccag ctccgggggg 1140

gcaccctgtg ccggctgcgcccc tccctctcct cgggctgccc cttgggtgcc gctctccagc 1200
tagctcttg cgcagtgcagc ctctaaactg cgggctgtac ggcggacgct gcctctggccc 1260
cgcggcc ggacgctgggc ggcgctgcgc ccggctgtac gcggccgggt gcctctgccc 1320
ccggcggcc caaagggtaag cgggctgccc cgggctgtac gcggccgggt gcctctgccc 1380
ttcctcttcg cgcgggccc gctggcttcgc gcagaaggtg gcggctgccc gcggccgggt 1440
cgctgcgtg cgctggccag ccgggctgccc cgggctgtac gcggccgggt gcctctgccc 1500
gcggcggcc gcggggggtgc ccgggctgccc cgggctgtac gcggccgggt gcctctgccc 1560
ttcctcttcg ccgggctgccc cgggctgtac tcggccccgc ccgggctgccc cgggctgtac 1620
ttcctcttcg ccgggctgccc cgggctgtac tcggccccgc ccgggctgccc cgggctgtac 1680
gggtggcccg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 1740
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 1800
gccggggcc cgctggccag ccgggctgccc cgggctgtac gcggccgggt gcctctgccc 1860
cggcggcc gcggggggtgc ccgggctgccc cgggctgtac gcggccgggt gcctctgccc 1920
gggtggcccg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 1980
gggtggcccg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2040
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2100
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2160
gggtggcccg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2220
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2280
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2340
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2400
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2460
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2520
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2580
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2640
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2700
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2760
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2820
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2880
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 2940
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 3000
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 3060
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 3120
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 3180
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 3240
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 3300
ttcctcttcg ccgggctgccc cgggctgtac gcggccgggt gcctctgccc gcggccgggt 3360
```
| gcaacctgoc accgccgcca tctgctgcc cgccagtttt cggctcgcgg  |
| aaagagcggc ggacggtggt tgcgagttgg cgctgctgtc ctcgagcagt |
| ccgctgccg agtctgctc cggctgtgct ctcgagcagt ctcgagcagt |
| ggcctggtg gcgcagctg cggccagcgg ggtgcgtgct ctcgagcagt |
| cgcagctg ccagctgagc cgcagctgagc cgcagctgagc cgcagctgagc |
| ggcctggtg gcgcagctg cggccagcgg ggtgcgtgct ctcgagcagt |
| cgcagctg ccagctgagc cgcagctgagc cgcagctgagc cgcagctgagc |
| ggcctggtg gcgcagctg cggccagcgg ggtgcgtgct ctcgagcagt |
| cgcagctg ccagctgagc cgcagctgagc cgcagctgagc cgcagctgagc |
| ggcctggtg gcgcagctg cggccagcgg ggtgcgtgct ctcgagcagt |
| cgcagctg ccagctgagc cgcagctgagc cgcagctgagc cgcagctgagc |

<210> SEQ_ID: 22
<211> LENGTH: 9737
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (8947)
-continued

<223> OTHER INFORMATION: a, c, t, g, other or unknown
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (8499)
<223> OTHER INFORMATION: a, c, t, g, other or unknown

<400> SEQUENCE: 22

gatcttcaat atggcacaatt agccatatta ttaattgtgtt atataccata aatacaatttt 60
ggcattggc cattgctataa gttgtgtctta tcatacataa tactatactata tattgtcacca 120
ttgcaaatg gagcgcgttg ttgggatatga ttatgacta tgtattaata gtaatcatatt 180
caggggtctact agattcataaga ccatcttgaag tgttcctccg ttacataact taccgtaaat 240
ggccccgccgcc gtcgacccgcc caagacaccc ggcgcgttgga cgtcaattga gacctatgtgtt 300
cocatagtaa cgcctaatg gccttctcatt gacaatgccacc cgtagcgtgagt 360
actgccacaat tggcaattcaca tcgaaggttat cttataagcaac gtcgacccgcc tattgacgct 420
aattggctgta atggccccgcc gccgcttacac gcgccttacag gcgtctttcctt 480
actggacytct acgtctatgggttat attgctcctatt gattaccaccttaccttggtggggtg 540
tacccacctag gcggctggata gggggggtctttcct ctctgggctattctt ccacccctaat 600
gactgtcaagt ggagtgttggtt ttggaacaac ccattacaggg catttccaaat cttgctaatc 660
acagggccag acggccccgccct tggcaacaaaa tgggaagtgtt cgggtgctggct 720
atattaacgcc cagttctttttta tgaacacgtcg aagagctgca acgtcagtttc 780
aaggccgca ggcgccttaccc gcgaacctttg ttcgctttggc tccctcag aactattcctt 840
atacagcgg cgtttaactc tcatcgcctcttt gccggtgcagcag tccgctggacg agcagccctc 900
gaccagggcct cggacacacac agcctgctctagt ggcgccagca ctcttattttttttttata 960
ggggtctcttg gcggcgctttacc ctgagttcttac ctcgcttttgg gtcgctttcct 1020
tttataacgtc atctagatccct tgtctgagct gcaccaattcctt tcatatgtgctc atctgattca 1080
tgcagcactc gaggagttgtt cttggtacactt cgcatacacttttc ggggcttgact 1140
taggagctag aggcttgctcatt ggtgctggtt tgggtgaccttg cgggctcctaat 1200
tttattagc cacagccacac caagccctggac gtcgccttttct gttttgttccgt 1260
ggggtttggc gggtcttttctt ggtctccttg cctgctggcttgtt gggctgctccttt 1320
gtggttgagc ctgctcgccgc ggtgcttgggg tgggtgggtgct gggtcttttttg 1380
agtggtttagc tcgggttcaggt cggagggctg aagtgacggct tattttctttttttttata 1440
tggcaaggcaggt cacgcttttcctt cggacaccctt atagttcttata ggtgggttcttctt 1500
attgccaggttg tggagaggtctc gattaggtgct cttgctttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
cagaattgct gaaatagcgtc gagctggttgg tggcaggtct tggattaaag
2100
aagccatgaa tccoccgagct catcttaaacc tatttgqac aaggacattg caagaccttg
2160
aaaagtacag ctttttttccaa gaaatttgtt tgagagaaaaa taaaaccttc gcagaattc
2220
cagtgctct cctcgtagtgc cagagggaga aaggctatcga gtaaaacaaat tgaagatagt
2280
agaagaattc taattaaggc ccccaataac tggcttaaaga aataatcgcc cggcccctgcc
2340
actcatacga gtaaatctgtt aactcattaa gacatctgccc gacatggaaag ccctcccaga
2400
cgtcagatgc aacactgtaac gcacaagggcag tggcgctcttg gtcagctgct gtaaatatta
2460
tgcagatgt taaaacaggg gcagaaagag tgcagcttcat ggcacgctttt aataccccca
2520
tgtagagatat ccccccagggg cttgctgctg aaaaacacaa attttcatca aacccccttg
2580
ggaatagcg ccgatttgc gcctaaacg acacatctgg cygatatata tggagaaact
2640
gcggagaaag cggcgtagtt tcacccagca gacagcagaa cgttccagtta tgctcaggtga
2700
aacgctgtga acgggtggttag acacatcccc ataatcaccag ctcacgctct ttcctggcct
2760
tacgagatgt ccgtagcagc ttcacccagc gcggcagata gtagaaaaa ctaaccctta cgtgcagcttt
2820
actctgtgct aatcccccaata cggcggttaa aaagacggttt ataatcaccag tgaaccggct
2880
ggtctatgtt aacactgtgaa actgacagca actgctccaa aatgctttta cagtgcatgt
2940
gggtaatata acgggtggttag acacatcccc ataatcaccag ctcacgctct ttcctggcct
3000
cggaaatct cggaaaaaa aaaaaacaag acaagcctcttg cttaggtgtat gcgatcagtag
3060
agttggagcc tctcagctgct gctaaacgacctcttgcc ccaaacaaat taaagcggcgcc
3120
cgctctctc gcctgtgcctg cggcagtcagc aacagtctcttc ggtctttcg ggtctgtggtag
3180
tgtggtctga taccagggct gtaagcctc ctctttaacc ctaagcggag gaaaccgct
3240
tgctatatgt tccacagct gtccttcgc ctccttggtt ggcggcctggc
3300
acggcgcca caaaccggtc gctggacagc gccatatcgat cggcttttgct tggcttttttt
3360
acaaaaacat atatatgctgct aagggggtga gcggagctt cggataacaa ttaagctcggc
3420
gcggctgcct acacggtata tgcaggggtgac tggctgaagc tggcgctgtg cggctgtggtat
3480
gctcctctg tttctctgctg ctggtggtgag cgggtgtgtag aggagagata
3540
cggagatcag gcggagagag gtagagaaaatcgggcagagagatgtg catgactgtt
3600
acggcagcgc ctcctctgct tttacgaaggcg tgcaccaccc cccctacatag aataatccct
3660
cagtycagat ctcagcggtct ggctggtgtat tttctcgctgc tagctgttgc acctatttcc
3720
tccctctac ctggtggtat cggccctacc cgtttttgctg ccctccctgctc ctgctgctc
3780
acagccctct cgtggtggag acataacagtac agctttttccct ggtctggacag cagccctgcg
3840
acaggtttag ccttggtgcc ttaaatcaca ctgaagtgtg cgaconaagc catcgagggg
3900
aaggggccag gcagaaactt cccgcccccct tggggagtgg cgccagcatgc aaagggagcc
3960
actccacat tactcggtgctg ttagtactgctgc tgtcgagata tagatgcgctg
4020
tatcctgata tagaggttag tagataatcc tccctgata tagaggttag tagataatcc
4080	tatcctgata tagaggttag tagataatcc tccctgata tagaggttag tagataatcc
4140	tatcctgata tagaggttag tagataatcc tccctgata tagaggttag tagataatcc
4200	tatcctgata tagaggttag tagataatcc tccctgata tagaggttag tagataatcc
4260	tatcctgata tagaggttag tagataatcc tccctgata tagaggttag tagataatcc
4320
ctastctcta ttaggtagc aatactgcc ccgctcagc ttaggtagc aatactgcc 4380
cagatatacg ctatgctaac cagatatacg ctatgctaac cagatatacg ctatgctaac 4440
cagatatacg ctatgctaac cagatatacg ctatgctaac cagatatacg ctatgctaac 4500
cagatatacg ctatgctaac cagatatacg ctatgctaac cagatatacg ctatgctaac 4560
cagatatttc ggtattatat gcatacctag gcgaacattg ccacccctgtcc tctctctccc 4620
cgtcctgact taagacaac cactccctct gttgccctgc aaggcccaagtg gttgctaat 4680
tgctcctcag atcgacagaa tctggccccc atctgggccg gccacactac ttagcaggt 4740
attcctccgg ggtcctcatag ttttttttgt ggaatggtgt tggacccacag tgggtctgag 4800
ggctcaacct ggcactgtag cttttgtatat cttttgtatat cttttgtatat cttttgtatat 4860
gtggggtcag acgcctgcct ccacccctct atttccaaaa aagagttgcttg cactctgtctt 4920
tggttcttggg cccactttggc gttggagccc ctttaatttt cccccctgctt agacacacag 4980
agctagcctag accgctctcg gctgcctctgcc tttttttttt gtttccata caggtgtcag 5040
aacgtgtgct acctggtctag tggcctgct tggggccaca tttttttttt cttttttttt 5100
tgctcctcag atcttttgtat gcacccctct atcttttgtat gcacccctct atcttttgtat 5160
ctacatagct attattttgga cactgtagt cttttttttt cttttttttt cttttttttt 5220
ctacatagct attattttgga cactgtagt cttttttttt cttttttttt cttttttttt 5280
gcgcctcact ccactcttctt aaccccttct tctggggtcag gcgcctcact ccactcttctt 5340
cgcgcctcact ccactcttctt aaccccttct tctggggtcag gcgcctcact ccactcttctt 5400
ggcgctacat gcacccctctt cttttttttt cttttttttt cttttttttt cttttttttt 5460
tggcacagca tgggcctag gcgggtctgcag gcgcctcact ccactcttctt aaccccttct 5520
ccactcttctt aaccccttct tctggggtcag gcgcctcact ccactcttctt aaccccttct 5580
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 5640
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 5700
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 5760
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 5820
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 5880
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 5940
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 6000
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 6060
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 6120
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 6180
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 6240
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 6300
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 6360
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 6420
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 6480
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 6540
tgggtctgcact cgtgggagct gcgggtgctgcact cgtgggagct gcgggtgctgcact cgtgggagct 6600
ctttcccata gctcagcctg taggtatctc agtctgggtg agtctgttgc ctctcaagctg 6660
ggtgtgctg acagacccc acgctccgacgcgcgcgcgccctggtaacattcgtct 6720
cctgcagttc aaccgcgtac acacgtacctgctctcgtcgtacccg ctcgaattcgg 6780
attagcagc aggatgtgtg aggattggtc acagaccttc tgcaggtgta gcctacactac 6840
ggcacactca ggaaggagct atttgtgtct tggtctctgc tggagccagt tacctctcga 6900
aaacagttg tgaagctttg atcgggaaa ccaacccccc cttggagctgc tggggggtttct 6960
gttggcaagc agcagttactac ggccagaaaa aaggaattctc aagaagactc tttgtatcttt 7020
ttcaacgggt ctgcagctca gctgggaacca aacctcaaggtt tgggagcg aagctgtct 7080
cttgttccaa ggtctctttc acatgtcctct tctcggttgt gaaaaaccgcc acaagacgcgt 7140
aaggagaaaa tgaacacttca gaaacttgtg agctttatgata atctgagcaag acctttcag 7200
aacggctagc aacaggatgc gtctgcaactc ggagccgtagc atacgtaaag gctagaggaag 7260
gcggctgacc cctcggccc aacgctcttc agcagactca cgggtacgca aacgcttgctgc 7320
tggagggcct cgcgacccac ccacgagcgccc acacgtagct aaaagttacagc aagggcattt 7380
tttgctgcc aatctccgctg acagctggtc gctcgtggct acagacagac ctgcttcgctct 7440
gggcagtctg gctggagcgc tgggaacagct ctcgcttgag cggacgccact ctggctctctc 7500
gtgcagtcctc tttgacggtc ccaacgggtc cttccacccc gtcgggctcg cttgatgctgcg 7560
agtctttgtg gttctggcga atcgccggtg acgagccgagc atgagctgca gcccggcgtct 7620
tgcagctcgac acagttggtctg cttcctctgc agggagccag tggagctgca gctccttgag 7680
cggccagctgc aacatcagct ccacgacgca gctcgtttct ccgtgaaaag cggggtacgc 7740
agctgcggca ggcgagcgcg cttggtgctcg ccacgagctgc ctgtcagctgc ggtctgctctg 7800
tttcttgctgc gacgccggtg cttctctttc gcaacggcgac cggggtggtg cttgacgctgg 7860
cagccggcct aacggcgtt cagacggcct gattggtcgtgt tggcgcact agtcgcaata 7920
tagactttc caccggctaagc cggcagctg tggcgcact gcaacctggct cctgtcctgct 7980
aaccgatctc cacgtctctct cttgatcagct gttctcccag cttcgccact aacgctctgg 8040
cggacagac gacacacggt ttcgctttcg ggcgcgcgct acctttccag ccataagctg 8100
tctcttggtt aaacaactctt acctggtcac ctcaggtctgg tcccagcggat 8160
ccgccaggc cagtctggggt catgttgctct gcacgcgccg cttggacgctt ggcggctacac 8220
aaggctggtc tggccgcact ccaacaccac acacccaggt aacggcggac cggcgtctctg 8280
cctcggtggt cccctcttcact ccttcctcacttc gcggcagact tcccctctgcc 8340
ccggctccatc gtcgctgtgc ggaagctgtc aatgggaatc gcagcgtctt cttcgctctg 8400
goatctggg agacagacgcct gacoagcggtg cggggtggtg cttcgggcc gggccaggtct 8460
ggcggagtg cttcttctct cttgggctag cggctgtgag aacccggcctg cttcgggcc 8520
ctgctggccc cccgagccca cgctgcgtgg cgtccttcctt cccatctctact cccgctggc 8580
tccagcttgc acatcaggct atcgtcagcag ctgcgtgctgg tgggagcagct cttcgggcc 8640
ttcgagagct aatccagctg cttcgggccgc ggtgcggcctt ccctcgtggtg cgggtccgct 8700
ccggccctac cccacgcagc gcgctgtgtg cccggtccggt gttccgtgcct tggcgcgcgc 8760
gcggcgac ccaccacgcc cggcgcggct gccggggtcgc gggccagagc gggccagagt 8820
cacggtgct cagcacttct ctcgggccgc cttgctgcgc gacagctggcagct ggtctgctggt 8880
---continued---
cgggaacgcc gcgcgcggcgg tgcgggctctg gcgcacaagcc gcagcgcctgc aagcgggggc
9440
ggtctgctag gctggctgct gcgcggccccc gcgtctgagc gcgtgtcggt ctggcgcaga
9000
gcacaagcgc gcagcgcgctgg cgcgcggcggc gcagcgcctgc aagcgggggc
9060
gcagaagcgc gcggcggcggc gcgcggcggc gcagcgcctgc aagcgggggc
9120
cggcgcgctgg gcgcgcggcggc gcgcggcggc gcagcgcctgc aagcgggggc
9180
cgcgcgcgc gcgcgcggcggc gcgcggcggc gcagcgcctgc aagcgggggc
9240
cggcgcggcgc gcgcgcggcggc gcgcggcggc gcagcgcctgc aagcgggggc
9300
gcgcgcggcgc gcgcgcggcggc gcgcggcggc gcagcgcctgc aagcgggggc
9360
taatgtgca cggcggtgcgc gcgggaggt cggaagaagt cgggggaatg
9420
attcttggctg gtgcgggac ggcggaggt ccagcgcggtc agcgggctg
9480
aggggttc ctgctggggcac ggcgggtagt ccagcgcggtc agcgggctg
9540
aattgctcgc ttggcgggac ggcggggaggt ccagcgcggtc agcgggctg
9600
cagggcagc ttcagcagc gcagcgcggtc agcgggctg ccagcgcggtc
9660
gcggggaggt ccagcgcggtc agcgggctg ccagcgcggtc
9720
gcgcgggagt gcgcgcggcgc gcgcgcggcggc gcagcgcctgc aagcgggggc
9737

<210> SEQ ID NO 23
<211> LENGTH: 9737
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (8347)
<223> OTHER INFORMATION: a, c, t, g, other or unknown
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (8499)
<223> OTHER INFORMATION: a, c, t, g, other or unknown

<400> SEQUENCE: 23

gatcttcaat attgcctcat agcgcattat ttcagttggt attagcgatt aatccattat
60
ggtatatgct atgcattgct atgtaaatgct attagctgat tattgtgcta
120
tgtcctaatt gcgcctgat tggctagct cctgatattc attaattgct
180
cagggctgct cggctagct cccagctagct ggcgcctgct cggcgctgct
240
ggcgcctgct cggctagct cccagctagct ggcgcctgct cggcgctgct
300
cccagctgct cggctagct ggcgcctgct cggcgctgct cggcgctgct
360
ctgctgctgct cggctagct cccagctagct ggcgcctgct cggcgctgct
420
ctgtgctgct cggctagct cccagctagct ggcgcctgct cggcgctgct
480
tctcggctgct cggctagct cccagctagct ggcgcctgct cggcgctgct
540
tctcggctgct cggctagct cccagctagct ggcgcctgct cggcgctgct
600
tctcggctgct cggctagct cccagctagct ggcgcctgct cggcgctgct
660
tctcggctgct cggctagct cccagctagct ggcgcctgct cggcgctgct
720
tctcggctgct cggctagct cccagctagct ggcgcctgct cggcgctgct
780
tctcggctgct cggctagct cccagctagct ggcgcctgct cggcgctgct
840
tctcggctgct cggctagct cccagctagct ggcgcctgct cggcgctgct
900

---end---
gaacatggct gcaagtaaac gactctacta agggcagcac gtgagtaacct gcatcatttaa
960
ggctatactg ggcgttctaa cagatgtgta taagagacag ctccttttaa gtagcctgtc
ttttcatac aactaatggc cttgtgaggt cgcacatcct cgtctttttga caggttattc
1020
ttcgacatc ggacagctga tgggtcaact gcgtctacct gcgtctctgc gcgtacctctg
taagcagatc ttgctgcttct ggtgttgcttg tggaggtgctgc gaggattgaaat
tgtttatatc aacaagcgcag aagactgacgc aacactgaca gaaacattct
1200
tttgctgttgc gctgctgctg gatgtctgctg ccagcatctg acgtatcttg aagggctag
1260
gttgttctag ggcctagcgg gcgtttcttg tggctgtttc ttaggtcccc tctcagatatt
1320
tagtttaggt cttttttgcatt ggggagggga acgtataagct ctttttttctctgtgac
1380
tggcagatcg gttgagttgc gttgagataag ttcacaactca ggcagagaca aagacattg
1440
tttcagtagc atagctagcgt cctcatctga ctctctttaat gcgattatgc ttcctggtg
1500
tccgatcgtg ggtgacagatc ttcgctttctt tttataagttt ttcctggtggtg
1560
tctagctgca tgcagttttagc gacttgccct aatcttgctctt cctcatcttt ttttattg
1620
ttttgactactg atggcgacag tggctggctgt ctcctctgtt ttcccctgtt
1680
tttgcgggtgct cgtgatgctt ggcagcgcct gctttttttt gaggattgat
1740
tgtttaatcg acgtataatg gctggtcctgc aatctcagcct ggggctagtcg
1800
ttctgcagcgc gctggtcctgc aatctcagcct ggggctagtcg
1860
tgcagataag ttcgctttctt ttttataagttt ttcctggtggtg
1920
tttgctgcttct ggtgagttgc gttgagcttgc ccatccttatt gtttttaagttt
1980
tgtttaatcg acgtataatg gctggtcctgc aatctcagcct ggggctagtcg
2040
tttgctgcttct ggtgagttgc gttgagcttgc ccatccttatt gtttttaagttt
2100
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2160
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2220
tgtgctcttg cgtggtcctgc aatctcagcct ggggctagtcg
2280
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2340
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2400
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2460
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2520
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2580
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2640
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2700
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2760
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2820
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2880
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
2940
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
3000
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
3060
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
3120
tttttaactcg gacatcgtgg ggaagttgcttg gttcattctt tttttttaagttt
3180
-continued

tgatctggta tttgggagta gtaagcctct ctctttattc aagcctgag gaggtgctc 3240
tgatctgcca cagccagct tggagctct cagtttgctt ggtgctgctg 3300
aacccggccca cagccagctt gctactggt caggttctt tggaggttc 3360
acaactcct atatccggtgc agatggtgct ggtgactgct tatactctat 3420
gcagcagcct aacattggtgt cttgccatac ctgcagatg cgcctagg 3480
gctttctggt cttcctaaa tctttgagag ggcagcctgct ggttgagttt 3540
cagcagcct gggagaggtg caggtggtgag ggcttgagtg aacagttgtg 3600
acggagcct gctaccattg tttgctactc ggttgctgctt ggtgagctct 3660
cagcagcttgctg gtttgctgtt tttgctactc ggttgctgctt ggtgagctct 3720
tctccaccct gcttggtcgt cttgacattc ggtgagctct gttttccct 3780
aacaagtttt ccagcctggag aaacctgatc agctttcatt ggttgagctct 3840
acggagcct gttttccctg ctgctgctt ggtgagctct gttttccct 3900
acggagcct gttttccctg ctgctgctt ggtgagctct gttttccct 3960
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 4020
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4080
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4140
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4200
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4260
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4320
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4380
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4440
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4500
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4560
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4620
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4680
tacaggata gctgctgctt tttgtctgtg gtttgctgtt tttgtctgtg 4740
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 4800
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 4860
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 4920
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 4980
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 5040
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 5100
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 5160
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 5220
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 5280
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 5340
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 5400
acccacactct tacatcggct tctgtctct tctgtctct tctgtctct 5460
-continued

gccaagtgccc ttgtgactaa aaaggtttcaat tcctctctgt gactctcgaac ccctcgtaaa 5520
taaacagctg aagcgcaagc gggtgaggaga ttagcgtggg cttcggaccat cttttbactaa 5580
ccctaatctg atagcatactg ctctccggttg ggttaacatt tctattgaat tagggtaagt 5640
tcctgataq aatacataa cccgggaaga actggtcacc ctgattgagtt taaaacggg 5700
gccttataa caactattgtc aatgcosctct tggaggtcccottagctgat gtacacacagg 5760
cccctctgt cgccggtggt gtagcctccct gtagctctcc ctggcccctcct ggaggtcact 5820
gtcocccacao attgtgtcgtt gcgctcagc caagagttac acataaaggg cagttttgtt 5880
tcctgcaacc aagcgcaacat gctttgctcga ggtgacagcc caactgctgtg tggcnaatgt 5940
gccactctct ttcataagggt gccuactaca atcagacacagc ccoctctctct tggctcctcc 6000
ccctgctgac atgcttggaac ggcagcctgt ggcaagttgt atcacaacac tgaagggatt 6060
aagactgctt cccggaataac aaaaacaagc gcgtttctctt accagcgaga aagggcgaag 6120
gattgcagtag ttgaggcttgg tccgtcgcgc ggcgggcccgc gcaagcgccc gcggctcaca 6180
cagcagcgggt ttgctttgcag ttggggtgat gtagatctgtg agcctaggtg 6240
caggcgtggg cgcgggcccgc ggcgggcccgc gcaagcgccc gcggctcaca 6300
tcggcataac gccatatagct ctgatatcttt gctagatctgg agatctgtcg agcctaggtg 6360
goaaacggcg acgaacagcct cggacgagcct aaaaccgcgc ggttggcccg ctttccccat 6420
aagctgcccgt cccctctctca acatcagctg caagctcgcag gtcgctgggct 6480
ccgccgctcat tacatttgta cccctctctca acatcagctg caagctcgcag gtcgctgggct 6540
gtcgccaagc tggcctttac cggatacctg ttggcctttac cccctctctca aagcgttggcg 6600
tctttcata gtcgctttac tggcctttac cccctctctca aagcgttggcg 6660
gggtggtggc acgaacagcct cccctctctca acatcagctg caagctcgcag gtcgctgggct 6720
cggcgcgctcat tacatttgta cccctctctca acatcagctg caagctcgcag gtcgctgggct 6780
attgacacag cagcggctgtga agcggctctt gcagtggtgtg gcctatctac 6840
ggctacataa gaagcggact atttggatgc tccgtctcgc tgaagcctg atccttggga 6900
aasagaggt gtcgctttag ctccggccaa ccaaccctcag ctgattgagtc gggttcctttt 6960
gtggcagcggc agcgaattaa aagttgcctc aagcgttcctg cttgtttttt 7020
ttcctggggt tctgctgctg gcggagcaga aaacagcggctt aagggattttt gtctgatgaga 7080
.ttcatataaa ggtcgctttt tggatgttgt tggcagttttt gagatacctg aagatgctgg 7140
aaggggacaa taccgccgat gcaacaggtta gcgtgatttttt ttcgctgttg aacgattgag 7200
aaggggatag aaggggacaa gtcggcagct ggagggcggcg atcctgttgaa gcgcgaggac 7260
gggtgctgtc catccgctgacc aaccgctgcggg cggagagtagc aaccgctgcggg cggagagtagc 7320
cggcgcgctcat tacatttgta cccctctctca acatcagctg caagctcgcag gtcgctgggct 7380
tttcgctctt ataccgccgag cagcggctgtga agcggctctt gcagtggtgtg gcctatctac 7440
gggctgtagcg gtcggagcctg ggcgttttttt ttcgctgttg aacgattgag 7500
gtgcgatcagt ttcggttcgt gtcggatggtag ccgtgccgtggt cttggctggtg gcctatctac 7560
atggttctct gcggccggctttt cgggctggcg ccgtgccgtggt cttggctggtg gcctatctac 7620
ttcggcggcttgcgctgctg ccggccggctttt cgggctggcg ccgtgccgtggt cttggctggtg gcctatctac 7680
cccggcgcgctcat tacatttgta cccctctctca acatcagctg caagctcgcag gtcgctgggct 7740
actgagccaa ggacgagcgc tctgggcca aagagcctag ccagctgacg ccctgatgagt gctctgtgcag
7800
tccatcagcc gacccgacca  gtgtgctgct gataaaagaa acgagcgccg cctggccggtg
7860
cagccgacac aagcggcgct cagagcgaccc gttgtcctgt gttgcccaggt cattacggaa
7920
tagttctcct acaccaagcg cggagacacc cgtgtgcaaat ccatcttctt caaatcttgcc
7980
aacagatctt catccctgctt tttgatgaga gttggatccac ctggcgccat agatccttttg
8040
cgagcagaaa gcccctcagct ttaaccttgcga ggcggtgctca acctttacac ctatgagttc
8100
tctatcattgg aaaaaatctca aatccgtgac ccctgaattt tacaggttgc aggaggtgcgt
8160

<210> SEQ ID NO: 24
<211> LENGTH: 9671
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (8481)
gatcttcact attgcccatt agcataatttt tcattggatt atatacattt aaactaatatt 60
ggtatttgc catcttctac gtgtgttctct ttcataaatcat ccgctatattt tattgctca 120
tgctcaat gacccgccag ttgctatgtg ttattgacta gttattaata gtaataaat 180
acggggttca tagttcatag ccctatattt ggttcccgct ttacataact taccgtaaat 240
ggcgcggcct cgtcagccgc cacaacccac ccggcccttg ccgctaatga gacctgttga 300
cocatactg ccgataatgg cactttcact tgcgttcaac ggtggtgatg tttacgtaaa 360
actgcccacactg tgtcacttttgtt acatgtccaa gtgaagccag gcctgggcgct 420
aattgactga actggcctcgc atgcctactct ggcacactaa gccacttcag ggtctttctt 480
actgctggct agctcttctct attatctcag gtttccacta tgggtgtgccg gtttggcgac 540
taacacaaag gcgggttctt gggtttgagtc ttcggggtgt ccacaaccaatt ggccttaga 600
gagtccagc gaggttttctt tgtccaccaaa attaacaggg aatcttttcaaa atgtacgtc 660
aatcggcagac gcccgcgccc tgcgaacaa gccgggttag ccggtgtaacg ccgggtgtct 720
ataacagc acgcttggat gtacacacgg atgatcaatcg attggtggttt ggtggtgtct 780
aagacatcaag atgtggcgac gcctacgctg cccttaaacc aagcgttaaag ccgcttcggc 840
gaacagttao acgtggggag gcgagctgcc cttgacggcg gcggccggcc gcggccggcc 900
tgcggacggct ctagccagct ccagctggcc ggaggtggtg gcggccggcc gcggccggcc 960
gggatcttacc cccacccactgg cggccgcttg ggtggtggtg gcgcttctct gcggtcga 1020
tgcgacagtag ttctatgaag cccttggaag ccggtgggtg acggtggtatt gacggtgtatg 1080
actgctgtaac ttcacacacg cttgcgagct ttcggtgtgt acggtgcctt gttgcgttgctg 1140
ataagctgccc ggtcctgtctt acatccttgt atcctgtggtg gctggcgaact aacggctctg 1200
agaagctgccc ctgacgctgct atctggtggc gcggggcatt gccggttctt gcggctgcctg 1260
tgcggtcagcttg cagctgtgtg ctgtgtcttg gcgggtggtg ctgtgtcttg gcggggctgtg 1320
gtgcggtcagcttg cagctgtgtg ctgtgtcttg gcgggtggtg ctgtgtcttg gcggggctgtg 1380
tgtgcttttgc gcgggttggtct gcggggtggtgc gggtggtgtgc gcggggtggtgc gcggggtggtgc 1440
cggggtggtgc gcgggttggtct gcggggtggtgc gcggggtggtgc gcggggtggtgc gcggggtggtgc 1500
tgcggggtggtgc gcgggttggtct gcggggtggtgc gcggggtggtgc gcggggtggtgc gcggggtggtgc 1560
ataacagcctgg gtttgccgccg ccctgggccg ccctgggccg ccctgggccg ccctgggccg 1620
aaaagctgtgc ctgggtgtgc ccctgggccg ccctgggccg ccctgggccg ccctgggccg 1680
agcggccagct ttggtgggctctg ggtgggttctt ccctgggccg ccctgggccg ccctgggccg 1740
attgatgcag gcggtgtggtg gcggggcttgt gcggggcttgt gcggggcttgt gcggggcttgt 1800
gtgcgcggcc aggctgggtg ccagtggtgtgc gcggggcttgt gcggggcttgt gcggggcttgt 1860
tgcggtgtgc gcggggcttgt gcggggcttgt gcggggcttgt gcggggcttgt gcggggcttgt 1920
gagagctggtgc ggcgggtggtgc ccctgggccg ccctgggccg ccctgggccg ccctgggccg 1980
ccacagttct gtcgctgtgc gcggggcttgt gcggggcttgt gcggggcttgt gcggggcttgt 2040
-continued

tcctattctt gagaagaagc gacottttaaa ggtagacttt accattgta cctgcagaga 2100
actcaagga caccccaagc gagctcattt tcttttccag aqtaaqagtt gctcctaaaa 2160
actactgca caacccagatt tagcaaataa agtagacagt gctttggtag tgggttgcag 2220
ttcgattat aagagaagc tgaactcaco acgcaccttt aaactatttg tgcaagacat 2280
catgcaagtc tttgaaagtgc acaagtttttc ttcagaagatt gattgtaga aatataac 2340
tctgagcagaa taccaggtgt ctctctctga tgcacaaggag gagaagagaa tttaagttc 2400
attggaatgt ttcagaagaa atcttaacta agggcgccca taactgctcat aaatatattt 2460
cggcgcgcgc tcgcatctct cgcaggtcttg ttcataactt ttaagctacttc tcggaactag 2520
gagcggcacta cagagggcgt gatgaggact agctgcgcag gggccgcgcct ctcgagcag 2580
ttcgtataaa tattttggcttg cggtaaagac gggagccgag aagttgtaac atctggccac 2640
gtttaaacta aacctgcgga aactcaacca ggggtggtct ggcagcagaa acaatattc 2700
ataaaacctt ttgaggactat aggcgggcttt tcctagcgtta cgagccacca cttgcaact 2760
tatggtgaga aacctgcgctg gttgttccct agacttcct cggacgcagt acaagttcttc 2820
agtttctgcttg tggaggaaccg tgcacacgcc ctgcagctact tcccatactt cgcagcag 2880
gttctctatt gcataccgta attcggagag actccctcct aagggggccaa gattgtgaa 2940
aaagggcgac taaactttgtt gatttttttt ctaaagctgc ccattaagtac aatttaactac 3000
caaggttcag gttcgggttat agtctaacag ggaactacgc taaaatggct taaatggctt 3060
ctttgcgag ttcgagcgcag ctactacgt ccataagttc agctggctcag tgcctttttt 3120
agtttctta acctgctgaa aatcgtgataa cttaaaaaa aagccocgtgc gttatttttt 3180
ttcattatttg tggaaagttg aacccctttc tgcccagcagc agctcctatt ttggcaaat 3240
ttaaatgac ggcgcgctct ccgggtggag atcaggtactt gagsagctac gcggcaccag 3300
actgggtgct cggttgtgtg gttctgtgagg tgaagagagga ctcccccactt cactctctg 3360
gaggaacgct cctctgtttat caaccaaatgt ctgctgcctc ctcggaagtc ttcacccctt 3420
ggagtggagcc ctggaccgag cccaccaacct ggcggtctaa ggggagcctat gcctctggt 3480
ttcgaggtct ttctacactat cttattttttt cggaggttgtt tggaggatgt gcattggacg 3540
cctttcttga caagccgcgcct cttaatcgtg aatcaagggg ttggctgtgtc gattttgtcc 3600
gtaaaggtac atttcttctg ccyygtttttc ccttattgttg aaggggttgg cgccggaggt 3660
gagcggagac agacggagag cggtagagag cggagagaga aagggagagg 3720
ctattgtactc tggtaggggtgc cgccttccat gcattttttc gctattttttt ccccccactt 3780
aaagataata cacccagtag atacctcagct gttcgtgttg gattttcttt gcaatctttt 3840
ttcgacactt ttcggggacc ctaaatgggg ctaattttttc tgcctttttt gacagtctac 3900
tcaatgccct ctctgctggtt cggagtaggtt cggagcttaa cctcttgtgtg 3960
caatcagca tgcagcggcct ttagcgtgccc ctcctctaatc caacctagttt gtcagctcag 4020
ccagagtcga gcgaagagac aagcgacact ttcctggagc ccctcgggctg gtggccggct 4080
atggcacaagc tagcataacc acctataacct tgggtctacttt atctgtagct gatatactg 4140
aggttagctg gctttccacc gatagagatt atcggctcagt ttgagatggt ctgataggitg 4200
aggttagctg gctttccacc gatagagatt atcggctcagt ttgagatggt ctgataggitg 4260
aggttagctg gctttccacc gatagagatt atcggctcagt ttgagatggt ctgataggitg 4320
aggtagacagct atgctaccgca gatagagtt aggtagacat atgctaccgca gatagagtt 4380
aggtagacat atgctatcag atatatttgg tagtatatgac taccagagata aatattgga 4440
tagctataac taccctaaac tcattattgga tagctataac taccctaaac agattgaga 4500
tagctataac taccctaaac tcattattgga tagctataac taccctaaac tagattgaga 4560
tagctatac taccctaaac tcattattgga tagctatac taccctaaac tagattgaga 4620
tagctatac taccctaaac tcattattgga tagctatac taccctaaac tagattgaga 4680
tagctatac taccctaaac tcattattgga tagctatac taccctaaac tagattgaga 4740
gtgctctac gcagcctcggt aataaaggca ccagacacccc tctgtcggtc gcctgaagccg 4800
aagtgttggt aatgtgtggt ccagatcctac gcagctgccg ccctatcttg gcgcctccc 4860
cctagttgctc agtatctgcc cgggtgctcc ttagtgtttt gcctgctgaa aggctttgctg 4920
gcgactgtta agggctgtac aatgaagccg gttattacac ctttttctta cagctacaaa 4980
cgcagcggcc ggcgtggtgg gcggagtggc gcggctgcctc cggagcttac cacagtttac 5040
tggcactgg tttgtttttag tggcccgccg tggggtgtaag cccggttttta tttttcttg 5100
ttgctggacg aagccgctgcagt cgtccggctg gttgccctcc actctttttg cccttttgg 5160
aataagagtc gcagatctgtcg cccctttcct cctttttcct gcgcctgcgc cccttttttt 5220
aaccacagac catttcttgc ttagattagg ttaaccttctgc aatcgtttta ccaggttt 5280
tgcagcattg gtttttacca gcccagcaggt gcagtttttg ctttctttcg aagctgtcctgc 5340
agagaggttc aacgagcggct cggagcattgt gcggctgatg cggctgcggc ctcgagtttt 5400
tgcagcattg gcagatccggt cctttttttt ccttttttt ttcttttgg gcggctgactgc 5460
tgaagccgctt tttttgtcag gtttaactac aactactacac tcaactata cagatcctgc 5520
tatatatgga aagggagcagc aggacagatc cggagttgggt taccttggctg 5580
cctttttag gcagctcctcgc taaattgtg cggagttgt cgtgagaggct 5640
tgcagcattg gcagatctgtcg cctcccttcgg ccacgtctctg cgggctggttc gcggctgctc 5700
aaccacagac catttcttgc ttagattagg ttaaccttctgc aatcgtttta ccaggttt 5760
gggagtttttc gagaggggtc aagttttttct gccatttttt gcgggtttttg 5820
agggggtcag cggggtgcct cagttttttt cctttttttt cctttttttt cctttttttt 5880
agggatatctgc ggcgcgggt cttccggttc ctgcggggtgg gcggctgctc 5940
cctggtggtgt gcagatctgtcg cctcccttcgg ccacgtctctg cgggctggttc gcggctgctc 6000
agggatatctgc ggcgcgggt cttccggttc ctgcggggtgg gcggctgctc 6060
gtgctgcgcg gcagatctgtcg cctcccttcgg ccacgtctctg cgggctggttc gcggctgctc 6120
gggggttcccg cgggggtggt ctggaggttg gcagatctgtcg cctcccttcgg ccacgtctctg 6180
gaagggggtt cgggggtggt ctggaggttg gcagatctgtcg cctcccttcgg ccacgtctctg 6240
agaagggggtt cgggggtggt ctggaggttg gcagatctgtcg cctcccttcgg ccacgtctctg 6300
gaagggggtt cgggggtggt ctggaggttg gcagatctgtcg cctcccttcgg ccacgtctctg 6360
agaagggggtt cgggggtggt ctggaggttg gcagatctgtcg cctcccttcgg ccacgtctctg 6420
agaagggggtt cgggggtggt ctggaggttg gcagatctgtcg cctcccttcgg ccacgtctctg 6480
agaagggggtt cgggggtggt ctggaggttg gcagatctgtcg cctcccttcgg ccacgtctctg 6540
agaagggggtt cgggggtggt ctggaggttg gcagatctgtcg cctcccttcgg ccacgtctctg 6600
-continued

tgctggtgc gaaacccgca gacactaaaa gcaccccgcc gttctccct ggagacctcc 6660
tgctggtgc gaaacccgca gacactaaaa gcaccccgcc gttctccct ggagacctcc 6720
cgggaagct gacggaggtct cctcactaac gctgtgagaa ttttcagttcg gtttgagttg 6780
tcgcctgcaac ggcacgctgct ggacgacgaa cccgctggcct aacgagcgtc tggccgatct 6840
cgcggcgtca cgcctggtag tcccaacgct gaccaacgctga ctgcaagcag 6900
cgcggcgtca cgcctggtag tcccaacgct gaccaacgctga ctgcaagcag 6960
ggcgcttacg ctcggtctac aacggaagct ctcggtctac aacggaagct 7020
cgcctgtctgc ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7080
ggcgcttacg ctcggtctac aacggaagct ctcggtctac aacggaagct 7140
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7200
tttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7260
tttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7320
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7380
tttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7440
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7500
tttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7560
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7620
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7680
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7740
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7800
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7860
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7920
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 7980
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8040
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8100
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8160
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8220
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8280
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8340
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8400
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8460
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8520
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8580
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8640
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8700
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8760
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8820
ctttttttttt ctgggccatac ctgggccatac ctgggccatac ctgggccatac 8880
-continued

acactgccg cgccgtcggc gcctcaoooc gccacgcgcc acacgcctcg caacgccg 0940
cacgacgg gaagcacaac gccgaaagaa ctctctccaa gcggcgttcg gcatgagcctc 9000
ggcaaggtgt gggcgtcggg ccaacggccc gcgggagcgg gctggggttc gggcaggac 9060
gtggacggg gcgggaggtt cgcggctgct gccgggcaag tggcggggtt gcaggccttc 9120
cggctgcccc gcagcccaac gatggaaggc ctctcggtgcc cggcagggcc caaggagc 9180
cggcgtgttc cttggcccac cgtggcggct ctggccggca ccacagggcc aaggtgcctg 9240
cagagcgcgt cggtgctccc gcggggagcg cgccccggggt gcggccctcc tggagacctc 9300
tggagacctg cgccggcccg aaccccccct tctacgcggc gctggcgtttc aacgcctccg 9360
cgcgggctg gggcgcggaa gcggggcgcag cctgggtgat gcggagcagc cggcgtgct 9420
ggagggggc cgccggagcg cgccggagcg cggagcgacgc cggagcgacgc cggagcgacgc cggagcgacgc cggagcgacgc 9480
tgggacaagg caggtgagg a ctggtgctaa cttggtgctg gactaattga 9540
gttgccatt cttggtctgt cggagtcgct ggggctccgt gcagaccattcc accctctact 9600
gcagacat cccagccctgt tctctctccg cttagaggt gccagggcga aatgtaagc 9660
gttaaatt ttttagttc cttgctcatt ttttaacc ccataccata 9720
tagccaaa taggcaaaa tccctataaa tcaaaaaat agagacagat aggttgtgat 9780
gttgtcagc tttggaacac gatgctcact ttaaagagcc tgagctccaa cggtagggg 9840
cgcacaccc tccataggg gcagtgccaa c 9871

<210> SEQ ID NO 25
<211> LENGTH: 10940
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (8670)
<223> OTHER INFORMATION: s, c, t, g, other or unknown
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (8822)
<223> OTHER INFORMATION: s, c, t, g, other or unknown

<400> SEQUENCE: 25

gatcttataat atggccactt agccatatatt tttcaatttt ttaatagcata ataatattt 60
ggtatatttc cattggtacg gttgtactca ttcataaat tctatatttt tgggtctca 120
tgctaaattt gacggtcaag tgggtatcgt tttggtacta gttatgattat gataataatt 180
caggggtgctg cctgggtgct gggctctggt ttcaattctt tgggctcat 240
ggaggggcgt cttgagctgcc ccacagccccc ccaccttataa ctgtctaatg gcgttctg 300
ccoatggtaa gcgggtgttatt ggttggtttt cttggttttattttctgctgctgctgctgctg 360
actggccat tggctctcagc tcaatgctat ctaatgctaa gctgctgtgct cttggtttagt 420
aatgcgctc tggctccttt cttggttttat cttggttttatt ccacagccccc ccaccttataa 480
actgggctc atgggactttag atatgctctt gttggtgctg cttggtttagt 540
tacgcctcgt ggggttgctg cgtggtgttat cttggttttatt ccacagccccc ccaccttataa 600
gcgttctgct gtgtggtgctg cttggttttatt ccacagccccc ccaccttataa 660
aatgcgctc tggctccttt cttggttttatt ccacagccccc ccaccttataa 720
atataagcag atgcgtttta gttgggctcct ctagctaatg atctgtgttttatt 780
aacygcaga tacactgtag cttctctgaa gctgagctct gctgttttcta tgaacaga 840
gtcgatatc tgtcggccta aacacactgaa gctgctatgg gaaagcc gactggg 900
gttctgagc ccaaaagccg aagcagccta cttctccccc accAACAGT tctcgcagc 960
ttgccgaaat cttcgaagaa tcgataaag cttgaaacag tattctcctt 1020
gatttccttc gagaagacac tgtgactgat tttgaaatgt gattttctac aggataaa 1080
agcattgtta aacaggttat tttctacaag taacttgaaat aagagtcgcg tcaacagt 1140
acattgctta cttgagacat aacacttctca accagcaagct gagcatagaa taagtggcag 1200
atcatctctc gagtcagaga ttaaaaagc tcaacagacag tggactttat cggcgagat 1260
caggtggtcct ctcgucacc taaagcgggt ctcgcccctgc aagagcctgt gtaataga 1320
cagctctctc aacagtagcct gctcttttct tcttcctagca cccctctgaa tccttccgat 1380
ttcgatagtt gtagcgtgta tctctcgaag ttcgtagcct gtaggttgaa ctcgagct 1440
aactctgtcct ttctctggcc acattgcag cctgctgcct tccttgctctgt gttgaggt 1500
cgagctagag gctgccagct aatattaagc tccacacgga cagggttagtc cagcacttg 1560
cagtcggctct cttctctacc gctgcttattg gctgctttctg ggggtgag 1620
atctggatat cttgagggct aaggttctgttt gttgccctca ggcggcggccttt gttgtagtc 1680
sggtgtggtc cccctcagag ttcggatttg taagggagag ggaatgttag 1740
ttcctctgct cagccctgaa atggaactg gtaagttgctg aagctctttgca atacgctttg 1800
caagccaga aaaacactcgc tggacccgct tggccggaga taaggtgtgta cagctgtgc 1860
ttttacgcag cggcagccag aaggtgaagc cgggttggag gaaacactga attocagt 1920
cagacgataaa ttgatttaac gtagctagct cgataacataa aagctctttc gacactcag 1980
cacacttggtg tggccctgacg ccggtttgctg ctcgacacgcc cttgagat 2040
cttcggtgtgt gctctgattct gttactcctt tctgcttgg ctctggttcc aagacattgg 2100
catcgggagcg aagggccccc cggccggtc cgcgtttggcc cggcactgag aatagatttctt catattccca 2160
gagacttgctt caataccttt cagtagaatg taaaagcatg ctcgagttta tggaggtgaa 2220
gagagttc ttcctccctct ctgagtatcc aacacttacg accttaaagc gtaagattta atttcccttt 2280
cagccagctaa cttctcagac gctcccaaatg agcactttct cttcagaga gactgatg 2340
tgcttcgtaa cttctcagac ccaccaggtt acgaaataaa cgggattctg tctagatgt 2400
tgcggttcaggt tccctttata aggcagcat gttacaccca gggcatcctaa aacctttttg 2460
gacagactgt cggacactct tttgaaactga cgggtttttttt ccgagattttgtt gtttcggag 2520
ataatacct gttcgaagat ttcccagttct gctctgtagc agaagagtcg 2580
taacagctaa atggagatgtc aataggtttct taaatattc aaagcgaat gaagattttta 2640
aacaataacc cggcagctcc gcactatcgc gcaagctgct gtaatctact ttaagaaattg 2700
gcagcagctt aagccacact acggcctactt aagcactctgct atgcagcacgg gcacactgctc 2760
ccttgcaggt ctgtgatatc attgacccag cttgaacag ggctggcagag cttggcag 2820
attggccacg cttacaatc aacggttcag gatggtcttt gcctagcaca gacggaaa 2880
catacatca cttacaatc tgagctacct ggccttgggct ctaacagcttc aacgacacag 2940
ttgaggtat atgtgtgagc aaggtgagctt gccttgctgg tagttctcct caagagatgta 3000
aacagcttca gttgctgctc gggaaagctt gtaacacagtt cccacattcc 3060
cagctcacg  tctttctttg  ccataagga  tttcgytgta  gctttcctca  ggccgygcaag  3120
aattgtgtaa  aagcccgat  aaactttgtg  cttttttttc  ttatggctct  ttaaaaaagc  3180
cgttaatccc  acggacggc  tctgggtata  ggtacattga  gcaactgacct  gaaatgctcc  3240
aataattgtc  tttcagctgcc  attgggatat  atcaacagttg  gttatatccg  ttgatgatta  3300
cxctcattttt  gctttcotta  ctctctgaaa  tttcgtataac  tcaaaaaata  cgcocggtag  3360
tgcttctta  tctattatgtt  gaaggttgga  acctttcagg  tgccacatca  ctgctcattt  3420
tgacaacctt  aaattaaggc  cgccgcttctc  ctggctagg  gtcacgtaga  aaggactacc  3480
gcgggacgga  ctcttggtgc  cggtgttgc  gtctactgag  gttgtaaacg  ctctccttcac  3540
aacaatagcg  ggagacgtgcc  ccctgtgatt  ccacaaatgc  tcctacattc  attgagctgt  3600
cctccctctg  gaccgccc  tgcaccgccc  ccacaaacctg  ggcocggcaag  gggtctcatt  3660
gctgtttatat  tctcctgttt  tttcagactc  catatatttc  tctgggtttt  gaaggttgcc  3720
attagcca  ctgctgatag  aaacccgctg  aacaattctca  gttcaggtgt  gcctgatggc  3780
eaagttgacg  atggagtaga  tttgct cocc  ttcgggtgga  aagggcttgc  3840
gcctggaggct  atgcggcga  tgcagcgagat  gaggagcttg  aacgggtcag  gcgggggaga  3900
gcgccggatg  gctggttcat  gttaggagac  gcccctcactc  gttattaagc  ccgctgatcc  3960
coccaacact  agaattact  ccccaagtga  catatagctc  gctgggttgttt  cttttctggc  4020
catccgttcg  tctcctgttt  ttcgctccac  aaccctggggc  aatgggcaatct  aacatgatg  4080
tgccgctc  cacccgctc  ctatagcact  cttcgggttg  gacacattact  ggacattta  4140
cctggggtagc  aatacgacat  gcgagcggct  taggttgtgc  tcttttaatt  cacaattaaat  4200
tggcgacacc  cacagcgcag  gaaaagccaa  aagcgacagaa  attcgctcgc  ccctgggagg  4260
tgacgcacag  tgcgctcag  gcacgagctc  cctgtctcgtg  ggttttaatt  gctgctgctg  4320
atatcgcag  ggagatacag  tcgctccagcg  atacaacag  gttacagcata  tactacccac  4380
atatagatc  ggatagcata  tcgctccagag  atatacagttg  ggtatagcata  tcgctccagag  4440
atatcattga  ggtatagcata  tcgctccagg  atatacgtgat  ggtatagcata  tcgctccagag  4500
atatagatc  ggatagcata  tcgctccagag  atatacgtgat  ggtatagcata  tcgctccagag  4560
atatagattga  ggtatagcata  tcgctccagg  atatacgtgatt  ggtatagcata  tcgctccagag  4620
aatagtaggtg  agcatataact  aaccoaatct  ctattagagct  agcatatactg  aaccoaatctg  4680
agagaggtc  agcatatact  aaccoaatct  atagagaggtc  agcatatactg  aaccoaatctg  4740
agagaggtctg  agcatatact  aaccoaatctg  atagagaggtctg  agcatatactg  aaccoaatctg  4800
agagaggtctcag  agcatatactc  agtagagaggtc  agcatatactc  aaccoaatctc  4860
agagaggtcatcg  agcatatactc  aaccoaatctc  atagagaggtcatcg  agcatatactc  4920
tggcctcaag  ttgcctccag  gacocggcaga  atagagaggtc  agcatatactc  aaccoaatctc  4980
ttttccgagca  ctgctgtgta  attttcttcctc  cagatccgcag  caagtcgcgc  ccctatttcct  5040
ccccgggcca  tactattgca  gttataggccc  ggggggcaatc  gatggtgtttt  gttgggcaagt  5100
ggggggccag  caagtcggtgta  gggggggcag  cttatattgca  atccggccag  ttatttacccg  5160
agctccacacct cgccggggggg  cgygttggggg  ctgctgcctgc  ccccccccctcc  ccaacattc  5220
aaaaagagct  ggcacaggct  atttgtgttat  gggcccccag  gggggtcagc  ccgggtttaat  5280
cttctggggct  ttggagacttc  gtcgggctgct  gtcggtcctg  tcgctttccc  ctctttcc  5340
-continued

ccttgtagc augaagagt tacaactctgg ttcacagcctg ttggcctctg ccttgagac 5400
aatctaaac ccctcagcct aatgtcagat aggaatttg tggctgctct gcataaaatt 5460
cgttgtgaag ggtcctccag tctcctccgg ttgcttcccc ccctctcgttt ctctctgtta 5520
agactatccaa gactttcctt ccgcctcttc agaataggct tttttcggac 5580
ttatatagtt gcatcaagat aatctcaca ctaaccaccc ctaaccacat ctttctctgg 5640
ggtcttcaac ggaactgggt tagggagcct ctcacataca ccttacctgtt cacaacctag 5700
cagttatcat attagctcag cgaaggagaa tgaagaagca ggcaagagtt cagagagttt 5760
ccaggcagct tcctgtagct tcagctcagct cctctgtagc taataagtctt cactccttc 5820
gttgacactc gcocggcagt aaatgccccc gttgacagct atggggtgag agatatagct 5880
gttccttacac accctttact caacccctat tcttagacat gttgcctcct cttgcctcctc 5940
tatgtagatt aattgagatt aagttggaata gtatatctta ctctcctgaca agctatagct 6000
accctcttgg ggtacacag gggcgcttac caccacatat gcctagctcc ctcttcgggtt 6060
cgctttctag ctgccttcac aagccctctgt gtttcagcct ggttcacctc ccctcttgttt 6120
tctggagcct cttgagagct cagtcccccc cccctggttat gttcaagtttccc gttcctgcgc 6180
tacactaaaa gccatcggct tgtgtagctc cacacagctc aaagttgctct ccaggtgtaa 6240
agccctgctcg tgtgcagcaca tggcagcaca cttcttataag gatgcctcact acagctcagq 6300
aacccctcg tgttctgctc ccoccggtgt cccagtgtgaca aacggccccc gttggaacgt 6360
tgtacaaacc aacagcaggg aacatcagca ctcggcccaac tacaacaca aagctgctct 6420
cgcagcagcg agacagagcc agagctgcagcc ttgctggacag cgtctgctcag gcgggcgggc 6480
gggccagacg gggccgctct gcacgacgct ggttggtctg cctaggctgct gtagctgata 6540
ggtcctacac gtagcctcag ggcagccagcc ggccgcggtg caacggcttg aacggctgtg 6600
cgcaagacg tagctgcagtt aatgccatca aacgcatata gctgtatcct cttgctcagq 6660
tgacatacg tagctgcagtt tgggacaaag gcagacaaaa gcagagagaa cgtgaaagac 6720
cgcgctttgt gttgccttcct cctacagcct gcggccctga cgccacactcc aaaaatgcac 6780
gctcagtccc agggttgcca aacggccagc gcgtacatag atccagcggc cttcctccct 6840
gaagacatac ccggctgtct cctgtctgca cccgtgctct taccgcatac gcttgctgctt 6900
ccctccctg ggaccagyttt gctctctctc atagctcagct ctgctagctat cttcgcttcg 6960
ttgaggtgtg cgtgctcagct cgcggtcttg tggacagaca cccgctctag ccggacagct 7020
gccgctatt cgtgtgctgt cgtctctgct ccaaccgcgt aacagacacg tttttcccac 7080
tgctgagaga cctctctgtaa gaggctgctgc gtaggtgcag tguatcaggt 7140
tttttaggt gttgcctcagct ctagagggct agttcttctg atctgcgtctc 7200
tggctgacg ccaggctgctg cggcagaggag cggctgtctc tgggctgctg cggcccttcg 7260
ccgcgtctgc cgcggtcttt gttgtagagc acacgctagc tgcggcagc aaaaatgcgcg 7320
cctacagcag ctttctgttg cttttctgac ggctctcagc cggccacagc aacaggttcg 7380
gttctctctg ttgctcagat agattcaca aagattctct ccactgtacct cttttctctg 7440
tggtgctcag ccgtgcttacgcc aacaggaagga cgccttctggat agttcttctg atctgcgtctc 7500
atattctaga aacacggtgc aacagagcgc tagaagccgaga tggctgctgc aacagggcgc 7560
ggtacacagt aaacgcagag gcagctctct gcaccctcag ctcagctcta 7620
-continued

tcaacggtac ccacacgctat gtctcgttag cgctgccgca ccaccagcgg gcaccagctg 7680
ataaatagc aaagagcggc aatttcgacc atagatagct gcaccagcgg aatgcagctg 7740
gtacagcg gatcctcgcc gcctgctagt cttgtccttg gaactggcag cctccggcct 7800
ggctggaca cctgctagc tgctcgctca cttcctgcat gcacacggtc ggtgccacct 7860
cagatcagtg ctctgctgtat gcgtatgttc gctgtggtgt ggaatgggga ggtagagcga 7920
tcaagctgtat gcaccagcgg cattgctaca gcagctagct aatcctttctc gcaccagcgg 7980
aggtgaagat acagagatcg ctgccccggc accagcggca atacgacgca gtcccttccc 8040
gtcaactagga caacgtgcaag caacgtgcgtg caaaggagcc caagctcggc cagacacagat 8100
agccggcggt ccctgctctg caagctgtgct agagcggcag acctggctgt ctgtccaca 8160
agcacgccc gcctcgtcgc tgcacagcgg aacaagcggc cattacgacgc ggcagattggc 8220
tgtggtgccc agtcataagc gaataccttc tctcacaagc aggccgaga acccctgctg 8280
aatccaccttt gttcatcagat gccaacagcgt cctcatctctg tctctgtgat caagcttggat 8340
cctgtgacac atcagatctct tggcgccag aaaaaccact atttacctctt gacgcgcttg 8400
tcaactctac aatgcttcag tcgctctctac ttcaaaaccct tcaatctctgc gcacccgaa 8460
তত্ত্বাদিত্ব তাত্ত্বিকতা তহদ্ধ তহদ্ধ তহদ্ধ তহদ্ধ তহদ্ধ তহদ্ধ তহদ্ধ তহদ্ধ 8520
cgccgctgca ctggctgcatt caagctgccg cctggctctg gcacacatct ccacactac 8580
gttgagggcc aacccgctgct gtttctctgg ggcctctcag gcgcacctcc tacatccttc 8640
ctagttgag acgtcctccgc ggcctccgac ctggcctgct gcggcagctg acaaaatggas 8700
atgcgggtct ctgcactctg atgtgcagat tgcacagcgg aatgctgttt ggcagctttg 8760
ggcctttgag ggtcgccagat atagcagcttt tggcctctgg ctgcctgctgc caaagacgctg 8820
গন্ধর্ভ গন্ধার্থ গন্ধার্থ গন্ধার্থ গন্ধার্থ গন্ধার্থ গন্ধার্থ গন্ধার্থ 8880
agtcctccgc gagggcctccg atcttctgac cttaaaaacc gcagctttgg cgcgcttttc 8940
tctcttctct cattcctccgc catttccgac tggacctctc tagatctctg caagcttgag 9000
ccttcacatga cgagtttgcga gcacaagcgt cgcgtgcgtag cggccaga cggccagggg 9060
ggcgactcgc cccagctggt gcctgctgga ccacacccgca ccacagccagcc caacgctgag 9120
cgggagcagc atacaagcgg gttcagccag atcagaaacc tttctctcct cggcgcgggg 9180
cgccacagc gcaagtcttg gcgtgagcag gcgccgctctg gggtgggtct ctggctccag 9240
cggagagcgc tcggcagctt gcgcagactg gcgcgctcaag gcgcgtgctg 9300
agcggtcgc gcgcgcgcgg cggccagcag atggagaccgc tccgtggccc gaccagcggc 9360
ccgagacgc gcgggtcttc tgcgcgcgac cgggcgcttc tggccgcagac cggccggcagc 9420
agggctcgcg aggagggctt accgctctct cttctgagcc ggccctgctc gccggagggc 9480
agccctttc cggccgagccc gcggagccc acccctccc tccagagcgg ccctgctcgcg 9540
cgccactg gcgcggcgc gcgccgcggc gccgcgcgc gcgcgcgcgc gcgcgcgcgcgc 9600
ccgggtgcag aggccgcgac cgccgcagccc gcgggcgcgg cgggcgcggc gcgggcgcgc 9660
atctctggag acggggcttc tggctctgct gggttgctggc gcctctcctc gcctcctcct 9720
actatctctg atctctctc tggctctgct gggttgctggc gcctctcctc gcctcctcct 9780
acccactctc acacactctc cagctctgtg ttttccggcc tcaaggttgc acacggtgac 9840
atcgttagct tcatacttct gccctgcttc gggttgctggc gcctctcctc gcctcctcct 9900
tttatcact aagcagcaga cggcagacat ctttacetaa cagaagagata gacgacgata
9960

ggtgaggta cttggtacaat tgggaacacg agtcagcatat taaagaaagtt ggagtcacacc
10020
gatgaaagtc gcaaacatcgt ctatacggggc gatggtccac
10060

<210> SEQ ID NO: 26
<211> LENGTH: 7714
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 26

tcaagcaacg gacgacgagc atgacacccg gttgacqgga cccacagctg cccgagatgc
60
gccgcgctcg gcgcgtggag atggcgaccg cgatggtatg tggctgccaa ggtgtgtgttt
120
ggcattcagc agtttccggca aagaattgat ttggcctcct ctggtgaggt tgaatcagctt
180
tagcggagtt cggcgggtt ctatccggt cgaagttgccc cggcctgatg cagcgcagac
240
cacccggggg aagggcagsca gttatgagggc ggcgcctaca atcctaggcaca acgcgtcaca
300
tgctgctgac gggcgtgacct aatcgcagat ggtgcagtcga tgccagtttag
360
gctgtgaaga gaggcggcag atccttgaag cttccctgta tggctgtcat ctacccgtct
420
ggacgacagt gggcagcacc cggcgtccccc gttgcgctcg gggcgcagag gagaatcataat
480
ggggacaggg atcgcgtcct ggcggtggaa cggacagcat ccgagctgca ggcgtgaccc
540
cggagtcgga ggcgtagat cgtgcttccgc cggcagaaat ttggcgggct ggcgctgac
600
ggcgggtgga ggcgggtgga cgggccaggg gacagcagctt ggtgccgcc ggcgcgcgcgc
660
cggcggtggag atgcagagcg aacagcggccct ggcgcgtggag atgcagagcg
720
tggagcgggtga ggtggaggtg cgggggtggag gttgccggtg cgggggtggag
780
cacccgggac gggcgggttg ggggtggag cctttcctgct gtcggtctgc
840
atgcgggctc tgcggagctc agcagcggcagt cattaggttg gaggctgtgca ggcgctgac
900
gcaacaggg atgtgtggtg cgtgcttccgc cggcagaaat ttggcgggct ggcgctgac
960
caacaccccc cggcgtgagct gggccaggg gccgcttccgc cggcagaaat ttggcgggct ggcgctgac
1020
atgcgggctc tgcggagctc agcagcggcagt cattaggttg gaggctgtgca ggcgctgac
1080
caacaccccc cggcgtgagct gggccaggg gccgcttccgc cggcagaaat ttggcgggct ggcgctgac
1140

ggcgggtgga ggcgggtgga cgggccaggg gacagcagctt ggtgccgcc ggcgcgcgcgc
1200
tggagcgggtga ggtggaggtg cgggggtggag gttgccggtg cgggggtggag
1260
attgctagc ttcacactgc agcagcggcagt cattaggttg gaggctgtgca ggcgctgac
1320
catacactgg tgtcgtggag acagcagtat cattaggttg gaggctgtgca ggcgctgac
1380
atgtgagat cttccacactgc gggccagagc cgggggtggag gttgccggtg cgggggtggag
1440

gcctgagctt gcgcgacgag cagcaggtcct aactgctggt cttttatgct gcctgagctt
1500
tagcagactc cgcggcggcgc gcgcgctgcct cggggtcgct tgcggtgggt tgcggtgggt
1560
tggggcggtc cttcactgag cgcgcgctgcct cggggtcgct tgcggtgggt tgcggtgggt
1620
gttgataa cgcgcgctgcct cggggtcgct tgcggtgggt tgcggtgggt
1680
cggggtggag atgcagagcg aacagcggccct ggcgcgtggag atgcagagcg
1740
ggcgggtgga ggcgggtgga cgggccaggg gacagcagctt ggtgccgcc ggcgcgcgcgc
1800
acggtggag atgcagagcg aacagcggccct ggcgcgtggag atgcagagcg
1860
gtactcgtct ctgctcct ctgctctact gctgtctgtc aacagtatttt cgtgtt cc aacagtcttt ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt 1920
atggtaattta tgggtggcct ctgcttctct tgggtg gtatt cgtgct tctt tacgtct ctgcttctct 1980
tgtaagtctt ctgcttctct taaagtggcc aacagtatttt cgtgtt cc aacagtcttt ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt 2040
aagcgtgct tctgctgctt cgtgcttct ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2100
cctcagctt cggagagcttc tctgctgctt cgtgcttct ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2160
gtacccctt aacgcttgaatt tctgtgtgca gctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2220
cggagagcttc tctgctgctt cgtgcttt ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2280
cctcagctt cggagagcttc tctgctgctt cgtgcttt ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2340
gggttattactt tgggtgggctt cgtgcttct ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2400
gggtttattacactt tgggtgggctt cgtgcttct ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2460
gagctctcttt cagagcttt cctgcttct ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2520
gtacccctt aacgcttgaatt tctgtgtgca gctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2580
ttcgctctct cgtgcttt ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2640
taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2700
gggtttattactt tgggtgggctt cgtgcttct ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2760
tttttatttttt cctttttcctttt cgtgcttct ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2820
cttattttatttttt cctttttcctttt cgtgcttct ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2880
cctttttcctttt cgtgcttct ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 2940
cctttttcctttt cgtgcttct ctgcttctct taaagtggcc cgtgtggtgca aacagtcttt ctgcttctct 3000
tttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
ctctacctgg agtcttttccc acgtggtctaa tttctctcttg cgctgagcgt eaaagccttc 4200
tgacaaagaac gtctctcttt gttcttcccc aactgctgo cgctatctggctg gttacagcgc 4260
tggaggagac gcagtagata aataagttact gagttatgct cctctctcttgct ctctcttttg 4320
gaggtctct tattttttttttt gttttttttt gtttttttttgaa ctttttttttctg 4380
gttttgcagt aaaaacctaa atttaatattt aaaaaaacaag gaaattgataa aagagagtcg 4440
agagaataaac tocgttgaac acctaacccg tcgtacacatta caagaccgaag 4500
gctatccgca ttgcaagcttt tcgtgatatg agacccgggaag cggaggggaa ataccggggcgc 4560
tgagaataag cggaagggg caccttctttg gttttttttt ctctctctact cgagagtgcc 4620
gagaacgag ggagccactcc gcaaccccgaa aaggtatttgc ggacacgggct tagcaaacgt 4680
gttggttata caatttgaataaatctatgtag ttggttattgcc gctttttttgca 4740
cgggtgtgag aaggtattttca aacagtgttac ggggttctctg ccattaaaggctgtccctct 4800
aagacctctcg tttctctctgct ccctgcttgct 4860
tcgtatggaat gtacacgacc ccacgggaaac cgttactactct aacgcggctg 4920
cctcatctttc acctctctttctcttcttttc tttgggaaaaa ggcagctgtgc 4980
acttagacat caagacocacat ttgctggcag ggtagttagca ttattttctt cttttctgtt 5040
cctccagcata aacggtacgg acattttggtt aagtttaattt ggcaccggcgct 5100
caacccctga gctgctgctg ccaccggggaa accgttttttc accacgggag cgttctctctc 5160
attgcaacgc cggactcactt ggtttctcgtg aaggtattttc cttctctcttgct 5220
cgtatcttc accacggcct ttgattttttt gacttaaccct cggactcactt gttttttttttt 5280
gcgtatact acgtttctttt cttttttttttt ccattaaagttt cggaggggaa 4920
tgctgttttc aacgtgctca acgttttctttt ctctctctttt ccattaaacggtg 5340
tgctacacca attgattttttt ccagcctttgcttg ccaggtggtaa 4980
cggatttccgt ggtctttttgt gcccttttctttt cggaggggaa 5400
aapagttsga aacgtttttttt cggaggggaa 5460
aaggttgaac ccgtttttttttt ccaggtggtaa 5520
ccgtttttttttt ccaggtggtaa 5580
ccggtcttggga aatattttttt cttttttttttt ccaggtggtaa 5640
ccgggtttttttt ccaggtggtaa 5700
attctctcttg cggaggttcttg ccgtttttttttt ccaggtggtaa 5760
ctctctctctg ccaggtggtaa 5820
ctctctctctg ccaggtggtaa 5880
atccacgattttttttt ccaggtggtaa 5940
ctctctctctg ccaggtggtaa 6000
ctctctctctg ccaggtggtaa 6060
tgctttttttttt ccaggtggtaa 6120
tgctttttttttt ccaggtggtaa 6180
agatttcatg cggaggttcttg ccaggtggtaa 6240
ctctctctctg ccaggtggtaa 6300
ctctctctctg ccaggtggtaa 6360
ctctctctctg ccaggtggtaa 6420
caagaactcg tttaaagcca cgacactcgt ttgctcttgg agcgacgta ttgtaaag 6480
ggtataaat ggtgcttaca atggcagagc tctggtccac aacgtgttgt atagagaaa 6540
ttsggacct ttttaagga ctttgaaagc cagacccctg atgcgaccaac gttttgctctt 6600
gctttcctcc gtttaaacc ttgaacgccg aagaaagctta actggtcctga 6660
ataattcct ccgctcgagc ggttggccca ctgctccact ttttaagcttg gttgtgataat 6720
cgaaactgga ccggctgtcc actggtacg cttctctgtat tattagctggt ggaacccggt 6780
cccaactgta tctggtgctct gattatagt ctgggaccac gttcacaectc gttagcctgg 6840
tcggtgataa agacggtgac ccggctgctca ctttgacgctt ccgggtcagg aatagctgct 6900
gacaagttgc ccactggtat ccggttctgt atattatgcg tgggaccacg ccggctacgct 6960
tctggtgcgt ctgggtgattt tttgaagacc acgtgctccac ctttgatctc ggtttgatt 7020
cttagctggg aaccagctcc cactgtatgt gcggctcttgctttctat gggacccaga 7080
tccacgctgt ttttaggtgtgc tttgcggcggg ccggtccactc tttaggtccttg 7140
atacagccct cagcgtgacg atcgtgacct atcctagtcc tctaaggggca agttattgcc 7200
ttgcgtaagt aaccagttgc cggtgatata cccggtgctc tttgtgtcct ggggtgtgat 7260
tggcgcctgt ccgtgatttt tccaaaccatc tgggcgctgg ttatggtgagc aaataaccctg 7320
cggtagagac aagagttggt agaagccgca atcggccaga tgtgtggtgc gtttttctgt 7380
taaattgatg ctgggaggtt tttgggaggc gttcgtggcc cccacccctc cggcgtggtt 7440
tgcgattcgt ctggctgccg ttttatccac cgggaggtgc ctcgggaca 7500
aacacaagat aacagaaagc ggcggcttctt cctaatagcct ctttggtttt atctgtgoc 7560
tgcgggtct ccactgtgctt atgggagac cccacatac cttcgccgcc acgggttggc 7620
actctgagtt tgctgatgggg gtcgattggg accacggcgc ttcggtcggc aggcaagctt 7680
tggttttcat aacggcctgct gcgggtggg cgcggc 7714

<210> SEQ ID NO 27
<211> LENGTH: 5314
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<400> SEQUENCE: 27

gatcctcaat cccgccccatgccagccttt ttccttatgct atatagctt actgctattttt 60

ggtctgctgtt ctgtgatata ctactata aatccttattt tagctgtgt 120
tgcctactg ggtgctctcc cctttgctt gtttttctctt atgtctg 180

gcgggtctgt actgttgtac ggcgacttgatt ctgggctttg atgttgataa 240
gcgcggtgtgctgacttgctc gggacccatc gttctcttctt gtagctgt 300

cctagtgata gacggtctcc gccacgtctgttg ctccacgat cttcagctagtt 360

gctttgctt cggctgtctt cctgcatctct gacggtttggc gacggttcctg 420

tgctgggttgt cttctcttctt atgtgctctg gggaggtgctc gtttttgct 480

tgctgggtgtc cttcagctgttt ggttctgttg cctggggtat ctcctgctct ccaggtctttt 540

gacggtctct gacggtctttc cctggtgttcttt tgggctttgctt 600

gccgctgctt ccaggtctttc tgggctttgctt ccaggtctttt 660

<210> SEQ ID NO 27
<211> LENGTH: 5314
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<400> SEQUENCE: 27

gatcctcaat cccgccccatgccagccttt ttccttatgct atatagctt actgctattttt 60

ggtctgctgtt ctgtgatata ctactata aatccttattt tagctgtgt 120
tgcctactg ggtgctctcc cctttgctt gtttttctctt atgtctg 180

gcgggtctgt actgttgtac ggcgacttgatt ctgggctttg atgttgataa 240
gcgcggtgtgctgacttgctc gggacccatc gttctcttctt gtagctgt 300

cctagtgata gacggtctcc gccacgtctgttg ctccacgat cttcagctagtt 360

gctttgctt cggctgtctt cctgcatctct gacggtttggc gacggttcctg 420

tgctgggttgt cttcagctgttt ggttctgttg cctggggtat ctcctgctct ccaggtctttt 480

tgctgggtgtc cttcagctgttt ggttctgttg cctggggtat ctcctgctct ccaggtctttt 540

gacggtctct gacggtctttc cctggtgttcttt tgggctttgctt 600

gccgctgctt ccaggtctttc tgggctttgctt ccaggtctttt 660

<210> SEQ ID NO 27
<211> LENGTH: 5314
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<400> SEQUENCE: 27

gatcctcaat cccgccccatgccagccttt ttccttatgct atatagctt actgctattttt 60

ggtctgctgtt ctgtgatata ctactata aatccttattt tagctgtgt 120
tgcctactg ggtgctctcc cctttgctt gtttttctctt atgtctg 180

gcgggtctgt actgttgtac ggcgacttgatt ctgggctttg atgttgataa 240
gcgcggtgtgctgacttgctc gggacccatc gttctcttctt gtagctgt 300

cctagtgata gacggtctcc gccacgtctgttg ctccacgat cttcagctagtt 360

gctttgctt cggctgtctt cctgcatctct gacggtttggc gacggttcctg 420

tgctgggttgt cttcagctgttt ggttctgttg cctggggtat ctcctgctct ccaggtctttt 480

tgctgggtgtc cttcagctgttt ggttctgttg cctggggtat ctcctgctct ccaggtctttt 540

gacggtctct gacggtctttc cctggtgttcttt tgggctttgctt 600

gccgctgctt ccaggtctttc tgggctttgctt ccaggtctttt 660
atataaacg agctgcgatt gtagaacgcg ctagcactga aattcgtagc ccatacgtatt 780
aatgcgtat ggcggcccgct ggctgata cggatccttg gctgatccttg gctgatccttg 840
atagcagctg atgtaatttg tcaacagctg tctgatgagc agcagaacgcg 900
gcaactggtg caccaatacc gacatcatac agagagccag ctgagactgcgttct 960
agcaagctgct ctcggaactgc agaacaagtc ctaagacagcg ctaagacagcg 1020
gtctcagcgg cggcaggtgcc ggcccctgag cttggacgtcc cggatttcagc 1080
ggcaagcag cggcaggtgcc ccagctgtcc ccagctgtcc ccagctgtcc 1140
cagcagcag cggcaggtgcc gtgcagctgc gcgcctgag gcgcctgag gcgcctgag 1200
gatggctgca gcgcctgag gcgcctgag gcgcctgag gcgcctgag gcgcctgag 1260
aatcagata aaccatacag gcgcctgag gcgcctgag gcgcctgag gcgcctgag 1320
tgcaaaaaa cggcgaagaat ccacattgcct ccacattgcct ccacattgcct 1380
tcagagctgc gcgacctctgc ccagcagcag ccagctgtcc ccagctgtcc ccagctgtcc 1440
aatgcagctg attaaccagcg ttaacccctg gaaagcctgct gatggctgct 1500
cgtgacgta atacgctgca cttggcctgat cttggcctgat cttggcctgat 1560
gcgcctgag atagcagctg atagcagctg atagcagctg atagcagctg atagcagctg 1620
cgtgcagctg atagcagctg atagcagctg atagcagctg atagcagctg atagcagctg 1680
cgctgctgca gcgcctgag gcgcctgag gcgcctgag gcgcctgag gcgcctgag 1740
agctgctgca gcgcctgag gcgcctgag gcgcctgag gcgcctgag gcgcctgag 1800
tagctgctgca gcgcctgag gcgcctgag gcgcctgag gcgcctgag gcgcctgag 1860
ggcaaaaaa cggcgaagaat ccacattgcct ccacattgcct ccacattgcct 1920
ttttttttttt gcttttttttt gcttttttttt gcttttttttt gcttttttttt gcttttttttt 1980
ttttttttttt gcttttttttt gcttttttttt gcttttttttt gcttttttttt gcttttttttt 2040
agacttccg aagacagcgc cccacttgtcg ggcgcgcttcc gcggctgtcct 2100
cgcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2160
cgcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2220
gcgcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2280
gcgcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2340
ggcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2400
ggcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2460
ggcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2520
ggcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2580
ggcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2640
ggcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2700
ggcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2760
ggcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2820
ggcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2880
ggcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 2940
ggcgctgtcct ggcgcgcttcc gcggctgtcct ggcgcgcttcc gcggctgtcct 3000
---continued

tggcggcgag acctcacaac agtttaacctt gcaacggcttg ctaaaccttac cagtaaaaag 3060
tgtagattgtag ttagggatgg gcaagcctag cagggagaac gagaagagtgt ctgggttt 3120
caggttaggt gtaggtggtt ctcaagctag agctgtgtga cagcttggtg 3180
tctatattctc atgcacagcg atgggtgatgc gctggcaaaa tggctaaccc 3240
cagaaagatgc ctaacattgg tctgctagcc tctgctagcc ctaacattgg 3300
cactcatcagcc ctaaacatcc atggggagcc aacagctggc cggagccagc 3360
tattgatgttc caacagtggt gcaagcctag cagggagaac gagaagagtgt ctgggttt 3420
ttagggagcc tagcttcttt ccaagaccc gatcccctct agaacacagc tcgggtgtca 3480
eagctagac gcctcaggtt caacagctgg gctggcaaaa tggctaaccc 3540
tcgaggtgtg atcggttatc gactggttga cacgacatcc gatgggttgt cgtggctggc 3600
tgcgctagct gtagctcagc tctgctatcc cctctggtga ctgtggctggc 3660
tcgggtgagc cggcgctagc tctgctatcc cctctggtga ctgtggctggc 3720
tcgggtgagc tctgctatcc cctctggtga ctgtggctggc 3780
aagctgctgagc cggcgctagc tctgctatcc cctctggtga ctgtggctggc 3840
tgcggtgaga cagatccagc tctgctatcc cctctggtga ctgtggctggc 3900
eagctgctgagc cagatccagc tctgctatcc cctctggtga ctgtggctggc 3960
agctagac gcctcaggtt caacagctgg gctggcaaaa tggctaaccc 4020
cgtggtgagc cggcgctagc tctgctatcc cctctggtga ctgtggctggc 4080
taagcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 4140
agcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 4200
ggtggtgatg cttccctctc gttcatgta ctaacagttt cttccctctc gttcatgta 4260
tattgatgttc caacagtggt gcaagcctag cagggagaac gagaagagtgt ctgggttt 4320
tgtagattgtag ttagggatgg gcaagcctag cagggagaac gagaagagtgt ctgggttt 4380
tgtagattgtag ttagggatgg gcaagcctag cagggagaac gagaagagtgt ctgggttt 4440
ttagggatg ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 4500
agcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 4560
ggtggtgatg cttccctctc gttcatgta ctaacagttt cttccctctc gttcatgta 4620
agcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 4680
agcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 4740
ttagggatg ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 4800
agcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 4860
ttagggatg ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 4920
agcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 4980
agcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 5040
agcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 5100
agcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 5160
agcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 5220
agcttga aagatcggcc ttttcgtgtg atcgagctcg cggcggctgg cggcggctgg 5280
<210> SEQ ID NO: 28
<211> LENGTH: 8737
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (8347)
<223> OTHER INFORMATION: a, c, t, g, other or unknown
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (8499)
<223> OTHER INFORMATION: a, c, t, g, other or unknown

<400> SEQUENCE: 28

```
gatacttcat attgccccatt agccatatta ttcaatgggt atataagctca aataaatatt 60
ggccatgccc catgctatca tttataataa ttcataattta attggctaa 120
gcctcaatatt ccgtgcctga ttttaggtt aagcttagct gatagaaatt 180
gctgttttag ctctcttagg ccatatggcg gctgtggaggac tagtgcttaa 240
ggtcctggt acctgtgtg gcctataat attagatgtag attgttgtctg 300
tccagatg ctccttattt gactgtaag gcctggtctgc tcttggtta tcttggtta 360
```

<400> SEQUENCE: 26

```
gatacttcat attgccccatt agccatatta ttcaatgggt atataagctca aataaatatt 60
ggccatgccc catgctatca tttataataa ttcataattta attggctaa 120
gcctcaatatt ccgtgcctga ttttaggtt aagcttagct gatagaaatt 180
gctgttttag ctctcttagg ccatatggcg gctgtggaggac tagtgcttaa 240
ggtcctggt acctgtgtg gcctataat attagatgtag attgttgtctg 300
tccagatg ctccttattt gactgtaag gcctggtctgc tcttggtta tcttggtta 360
```

<400> SEQUENCE: 26
ctgaactctg ctcgtctgct ctgtcctcag aactgggcat ccggcaagas gggyaacctgc 1800
cctcgcaaco ctgcaagaat gaattcaag attccacag aatgcccaac acctcctcaag 1860
tagaaagtta acagaaatct gtggattatt ggtaaaagag cttggctctc atttcctaga 1920
agatgcaaco ttcaggagtta aagasttatt tgattctacc cagagaacot aagaaacctc 1980
cacaagggcct tcaatattctt tccagaagtcc ttagatagttc cttaaaactct actgaacaaaco 2040
cgaatatagcc aataaagatgt gcatagttgg tggccgctct ttcttaaacgg 2100
aagccagcagc ctcgccagcc cacttttaac tatttgtgcc aaagatctctg cagagctttg 2160
aagtacgactgtggctt ctaaagattc acacacgttt ggtgctttct cagaaacgacc 2220
caggttctct ctctgtctgc cagggagca aagctgttaag gtaactatttt gtagactttgc 2280
agaagatgta aataaggg cccacaataac tgccttaaaa aattacgccc cggccctggc 2340
acctcgccgtaaatggttct aatatttgg gcctgtgcca gacagttgg acctacaga 2400
cggccatgtgc aacctgcgacc gcgggctcag cgcgctcct gttgctgttc gtaaatattt 2460
tgccaggtg gaaaagggcc gccaggaaggt tgtgttatatt gccacaggttta acaaataaaa 2520
tgggcaacgc cccaggggcc tgggcttgtga cgaanacact attoctcaaat aaccttattc 2580
ggaaatggc caggttttcct gctacaccc cccctctcttg cgaatatagtt tggagaactt 2640
gccggaaactctcgctcttgcta gccgatagaaa gtttctttgt ctggctgattg 2700
aaacggygta aacagggtca aacatctccc atatoaccag acctacgctct ttaattgcca 2760	ccagaaatgtc cagatgagca ttcacagcc gggcagaaaa tggataaacc ggctgtaaa 2820
actagttcttt atttatttta aaggtttttta aaaaagccgt aataatagcc tgaagctgtct 2880
gcttattaggt aactattggcc aactgaataa atgctctaaag atgtctattta ctatgcgaatt 2940
gggtatatagc aagctgtggtgtat ctacgtatag ctttattttcc cattttatat cctttatttgc 3000
cctgaaactc cttactacaag aaaatactcag cgcggtagta cttattttcct tataagtgata 3060
agtgggaacctc tccagctgct acaaacaatgg tccacaaattg gcaagggcct 3120
cgcgtcttgct gtctaggggtct acgaagaaag gcagctccag gccgaacctc cgggctccgg 3180
tgctgtgtta tattggaggt gtaaaacttc cttctacacc gtaaggggag gaactgcccct 3240
tettctaatc caagtctgcg ttaaaactca gattgcctga ccctttgagga ggcggcttcc 3300
aggccgcctgg acatccgctgggc gctagttagc ctgtttatatt tggttttttt 3360
acaaatctgtct tgttttggaa ggtatcctgt cattgatctgta 3420
gcgcgctgtct ccgattgtgc tggctgctgt cattgtggat gagttcattt 3480
gctcttcct tttcctcccc ctgtgggggg ggtgcgctcg ggggggtatc aacagatgtga 3540
ccggagatgaa gggaggagtgg gagatggagag agggaaggg aagagttgat gaatacttttg 3600
agggacacccacatcgtgaattc ttagatagttc cttaaaactttt gcaatctgatgctaatctctctctcc 3660
cagtacattc cgcgtggctgt ctttggccttat cgtctgtttgc acccttattc 3720
tctctccact tgggacactgc tgggagcgca cgggtgtaaag ggcggagag gcattcccc 3780
aactatccttt ctcgggtgcag aacttacttcctgc ggtgctaatg ccagaatctgctt 3840
aczggcttgac ccctctggatt ctataactcc cctactatgg gaaggtgagcg tgggctgctctcctc 3900
agggacaagc cggacaatctt ctcggcctt gggaggtgctt gggagttgctt gggagttgcttttttt 3960
acccagcaactc taactctggg tattcacta sgtagtata tgtcatgaggg tgtcatgatgtgc 4020
-continued

tacocggata cagattaggaa tagcataaac tacccagata tagattaggga tagcataagc 4080
tacocagata tagattaggga tagcataagc tacccagata tagattaggga tagcataagc 4140
tacocagata tagattaggga tagcataagc tacccagata tagattaggga tagcataagc 4200
tacocagata tagattaggga tagcataagc tacccagata tagattaggga tagcataagc 4260
tacocagata ttggggtaggt atatgtcacc cccataaaa ttggtagatgc atatactacc 4320
tacocagata ttggtagatgc atatactacc cccataaaa ttggtagatgc atatactacc 4380
tacocagata ttggtagatgc atatactacc cccataaaa ttggtagatgc atatactacc 4440
tacocagata ttggtagatgc atatactacc cccataaaa ttggtagatgc atatactacc 4500
tacocagata ttggtagatgc atatactacc cccataaaa ttggtagatgc atatactacc 4560
tacocagata ttggtagatgc atatactacc cccataaaa ttggtagatgc atatactacc 4620
tacocagata ttggtagatgc atatactacc cccataaaa ttggtagatgc atatactacc 4680
tacocagata ttggtagatgc atatactacc cccataaaa ttggtagatgc atatactacc 4740
tacocagata ttggtagatgc atatactacc cccataaaa ttggtagatgc atatactacc 4800
tacocagata ttggtagatgc atatactacc cccataaaa ttggtagatgc atatactacc 4860
tacocagata ttggtagatgc atatactacc cccataaaa ttggtagatgc atatactacc 4920
tacocagata ttggtagatgc atatactacc cccataaaa ttggtagatgc atatactacc 4980

tactctcta ctgttagatgc atatactacc cccataaaa ttggtagatgc atatactacc 5040
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5100
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5160
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5220
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5280
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5340
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5400
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5460
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5520
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5580
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5640
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5700
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5760
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5820
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5880
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 5940
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 6000
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 6060
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 6120
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 6180
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 6240
ntgtagaaa attgtagaa caaaggtggct aattaggtgc tttggtcagag tttttgattt 6300
ttgcatcac gctatagcg ctagttcctt gctagagcg agatctgctg agccatgta 6360
gcasaagc gcaaaaaac ggcaaacact aaaaggaagg gctggttggc gttttctcat 6420
agcctcgc cccctgagaa gctctacaaa aatgacagct caagtcagag gttgcaaac 6480
cgcacagca tatataagata cagggctttt cccctggaac gctcctctgc ggtggctctt 6540
gttgcagcct gcggctttac gcgatacttg tctgactttt cttcctgagag agggcttgcg 6600
ccttccata gcgtacggat taggtatctc aggtcgggtc aggctcggtg cttcaagctg 6660
ggcgtgtgct acgcaccgcc cgggtcgcgc gcggctgctc ctttatccgg taacttcctg 6720
cctgagtcct accccagtaag acaaacatta ttctacagctg caacgaccc cagtttacag 6780
attacacgag cgggggtgct aggggtctact acagtttgc tggatgctgg gcgtacacag 6840
ggtacataa gaggcagctt attggtatct tcttgctctgc tggagcagct tacctttcgg 6900
aagaaggtg gtagtttgtg atcggcacc aaaaacccag cttggtacgg tgggcttttt 6960
gttccagac aggacgattc gcgcagaaaa aacctctctc agagacgtc ctttgctctt 7020
tacgctggct gcgtcctgca gttgcagcct ACTGAAAGA 7080
tgttcataaa ggtctccatt ctgatctctt ttatggtgta gataaacgct acagatctcg 7140
aagggacca taccgcattc ggaatatctg acggttataa attcagaaag actgctcaag 7200
aagggcattg aagcgcctgc gttgcgctact gggagcccgg atacgctaaaa gcaagggagaa 7260
ggcgcagcct cattgcgcat cgggcttttc aaccaattctc cgggctaccg aagctgatgc 7320
cygtatccgg cttccgaccc caacgctgctg acgcttagt gattccagaaaa aggcggact 7380
tttcaccatt atctggggtg gcaagcgggtg cggctggtct gcttgcagttc 7440
ggcgctgtc gctgttcgaa ttggcgcctg tggggcttc catggcttgg gcagcgocct cattgttctt 7500
gttcgctata tctgtgctgc gcggccgctg gcgatgctgc gcgaagcggct gcgggcagct 7560
agttttcgct tggggtgctg atgggagct acggtagactg aagctgatgc ggcgcgccat 7620
tgctacagc ctagtggata cttctccggtt gcgaagcggct gcgaagcggct gcggagctgtgg 7680
cccgcgcact gcgcaccagc cttccgctcc tctgactaac cgcgcgact acgctgcgct 7740
agctggcag gcggcccccgc tcggctgctg cagcgcagct gcggtgtgtc cgtgggtcgcgc 7800
tctatgttc gcggcgcagc tgggcttgct ggccaaaaa aagcgcggccc cttctggtgta 7860
cggcggcagc acgctgctgc cagcgcagct gcggcagct gcgggtcgtgt gcctggctggct 7920
tgctgttctgg caccgacgg ctcgagacgg ctcgagacgg ctcgagacgg ctcgagacgg 7980
aacagctct ccctcggctcttc gttctcctcag cttcctcctcag aataaagctg 8040
cgcgcgtcc agcactccct gcgggtgtctt gcgcctgccgc gcgcctgccgc gcgcctgccgc 8100
tctgacatt gcggggatac ttgctgctgt gcggggttgg gcggggttgg gcggggttgg 8160
ctggcagct gcgcggtcctg ctcgagacgg ctcgagacgg ctcgagacgg ctcgagacgg 8220
agggctggtc cggctggtcctg ctcgagacgg ctcgagacgg ctcgagacgg ctcgagacgg 8280
ctcgcggtgc ctcgagacgg ctcgagacgg ctcgagacgg ctcgagacgg ctcgagacgg 8340
cgggctgcg ctcgagacgg ctcgagacgg ctcgagacgg ctcgagacgg ctcgagacgg 8400
gcgtgtgct gcggtgtgct gcggggttgg gcggggttgg gcggggttgg gcggggttgg 8460
gcggtgcag ccggggaccc cggggggtg cttggggggc gcggggtttgc gcggggtttgc 8520
gcgctgggct gcggggaggg gcggggggg ccggggggg cgggggggg cgggggggg 8580
ctgacagctt caaaaagcga cgtcgtgcgc gctgyttctcc tccctoctct ctcogggcct 9640
ttcgaotctg atcatactag atctcagcga gctgaagcct acacagtccg agtacaacgc 8700
cacaggtgcg ctcrcocaccct gcgagcgact cccaggggcgc tgaacgaccc tgcgacgcccg 8760
gtctgcagcc taaccgccgca cgcgtacgcc gcaggtcagc tgcgcgcgtcagctgggt 8820
caccgcagg cacagctctt tcccgacgcg ctgcgcggtc gacagctgcc aagtgtaggt 8880
cgcggagca gggcgcccg gctgggtctct gcacgcgcgc cccgggcccc盅 gacagccgtc aacgcggcccg 9040
gtggtctgcc gcgcgtgccc cggctagggc gcgttgaggg ccggtcgcgg ctgcgcgccc 9000
gcagtcagct gcagctctctt gcgctgcgca cccggcccgg ggcgggcccg tgtgcctcttg 9060
gcgcgcgcgc gcgctgcgga cccggcggag gttgtggcag gcgtgcgcgcgc gtcggctgtg 9120
catgtaggag gcggtgccc ggcgtgccc gcggtgccc gcggtgcggt cgcgtgcgcg 9180
cccggcccgc cacagctttta cggcggctc gcgtgcgcgc gcgtgcgcgc gcgtgcgcgc 9240
ctcgcagccc gcgagcgagc gctgcgcgcgc gcggtgccc gcggtgccc gcggtgcgcg 9300
gacagccgc gcgcgcggga aaggggcga cggcgcggga cgggcgcagc ggagggggcg 9360
tggtaaaaag cggaaaaagc gctcgcggtc gttgtacagt atagtagtag aagcttgggtc 9420
atgttgcag ctgcgcggggt gcgtggggac cttccagcct ctaactgaca cacattgca 9480
aggtgtgtct ttcggcgcgt cggagtcata cggcgagact ctatgtgagt atcattggct 9540
aagtttcgc ttaatttatt gttgtacag tctatttatt aacatgagc cgaatcctcg 9600
cgagacggt cattacagct cgcagatggg tggatgttgg ttcagatgg 9660
gacaagcgt acactattta gaaagctgga ctgcgcgctc aaggggcgaa aagcttgctg 9720
tcgcagttg gcgccac 9737

<210> SEQ ID NO 29
<211> LENGTH: 12
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Vector Promoter

<400> SEQUENCE: 29

aacccaggtga 12

<210> SEQ ID NO 30
<211> LENGTH: 15
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Vector Promoter

<400> SEQUENCE: 30

accagccag ttgag 15

<210> SEQ ID NO 31
<211> LENGTH: 16
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Vector Promoter

<400> SEQUENCE: 31
1-57. (canceled)
58. A method for over-expressing a desired protein encoded by a desired endogenous gene or portion thereof in a cell introduced into a non-human animal, said method comprising:

(a) introducing a vector comprising a transcriptional regulatory sequence into one or more cells in vitro;

(b) maintaining the one or more cells containing the introduced vector from step (a) under conditions appropriate for non-homologous recombination of the vector with the genome of the one or more cells thereby producing one or more non-homologously recombinant cells, wherein the transcriptional regulatory sequence on the vector is operably linked to an endogenous gene thereby over-expressing the endogenous gene and protein encoded by said endogenous gene;

(c) screening said one or more cells from step (b) for expression of a desired endogenous gene;

(d) isolating and cloning said cell from step (c) expressing said desired endogenous gene; and

(e) introducing said non-homologously recombinant cell of step (d) into a non-human animal and maintaining the non-homologously recombinant cell in said animal under conditions appropriate for over-expression of said desired protein encoded by said desired endogenous cellular gene or portion thereof.

59. The method of claim 58, wherein said transcriptional regulatory sequence is a promoter.
60. The method of claim 59, wherein said promoter is a viral promoter.
61. The method of claim 60, wherein said viral promoter is the cytomegalovirus immediate early promoter.
62. The method of claim 59, wherein said promoter is a non-viral promoter.
63. The method of claim 59, wherein said promoter is inducible.
64. The method of claim 58, further comprising introducing double strand breaks into the genomic DNA of said cell prior to or simultaneously with integration of said vector.
65. The method of claim 58, wherein said vector is linear.
66. The method of claim 58, wherein said endogenous cellular gene encodes a transmembrane protein.
67. The method of claim 58, wherein said promoter is operably linked to a splice donor sequence on the vector and wherein the splice donor is spliced to a splice acceptor in the endogenous gene that encodes the desired protein.
68. The method of claim 58, wherein said animal is a mammal.
69. The method of claim 58, wherein said mammal is a human.