SELF-ASSEMBLING GENES, VECTORS AND USES THEREOF

The invention relates to a method for directing the self-assembly of a gene or gene assembly having three and preferably six or more fragments in a directionally and spatially ordered fashion to produce a gene, gene vector or large nucleic acid molecule. The method can be used to create libraries, such as combinatorial libraries. In another embodiment of the invention a vector is described for the incorporation and screening of endogenous mouse promoter elements for the identification of cell-specific promoters.
<table>
<thead>
<tr>
<th>Code</th>
<th>Country</th>
<th>Code</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>Albania</td>
<td>ES</td>
<td>Spain</td>
</tr>
<tr>
<td>AM</td>
<td>Armenia</td>
<td>FI</td>
<td>Finland</td>
</tr>
<tr>
<td>AT</td>
<td>Austria</td>
<td>FR</td>
<td>France</td>
</tr>
<tr>
<td>AU</td>
<td>Australia</td>
<td>GA</td>
<td>Gabon</td>
</tr>
<tr>
<td>AZ</td>
<td>Azerbaijan</td>
<td>GB</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>BA</td>
<td>Bosnia and Herzegovina</td>
<td>GE</td>
<td>Georgia</td>
</tr>
<tr>
<td>BB</td>
<td>Barbados</td>
<td>GH</td>
<td>Ghana</td>
</tr>
<tr>
<td>BE</td>
<td>Belgium</td>
<td>GN</td>
<td>Guinea</td>
</tr>
<tr>
<td>BF</td>
<td>Burkina Faso</td>
<td>GR</td>
<td>Greece</td>
</tr>
<tr>
<td>BG</td>
<td>Bulgaria</td>
<td>HU</td>
<td>Hungary</td>
</tr>
<tr>
<td>BJ</td>
<td>Benin</td>
<td>IE</td>
<td>Ireland</td>
</tr>
<tr>
<td>BR</td>
<td>Brazil</td>
<td>IL</td>
<td>Israel</td>
</tr>
<tr>
<td>BY</td>
<td>Belarus</td>
<td>IS</td>
<td>Iceland</td>
</tr>
<tr>
<td>CA</td>
<td>Canada</td>
<td>IT</td>
<td>Italy</td>
</tr>
<tr>
<td>CF</td>
<td>Central African Republic</td>
<td>JP</td>
<td>Japan</td>
</tr>
<tr>
<td>CG</td>
<td>Congo</td>
<td>KE</td>
<td>Kenya</td>
</tr>
<tr>
<td>CH</td>
<td>Switzerland</td>
<td>KG</td>
<td>Kyrgyzstan</td>
</tr>
<tr>
<td>CI</td>
<td>Côte d’Ivoire</td>
<td>KP</td>
<td>Democratic People’s</td>
</tr>
<tr>
<td>CM</td>
<td>Cameroon</td>
<td>KR</td>
<td>Republic of Korea</td>
</tr>
<tr>
<td>CN</td>
<td>China</td>
<td>KZ</td>
<td>Kazakhstan</td>
</tr>
<tr>
<td>CU</td>
<td>Cuba</td>
<td>LC</td>
<td>Saint Lucia</td>
</tr>
<tr>
<td>CZ</td>
<td>Czech Republic</td>
<td>LI</td>
<td>Liechtenstein</td>
</tr>
<tr>
<td>DE</td>
<td>Germany</td>
<td>LK</td>
<td>Sri Lanka</td>
</tr>
<tr>
<td>DK</td>
<td>Denmark</td>
<td>LR</td>
<td>Liberia</td>
</tr>
<tr>
<td>EE</td>
<td>Estonia</td>
<td>LS</td>
<td>Lesotho</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LT</td>
<td>Lithuania</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LU</td>
<td>Luxembourg</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LV</td>
<td>Latvia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MC</td>
<td>Monaco</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MD</td>
<td>Republic of Moldova</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MG</td>
<td>Madagascar</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MK</td>
<td>The former Yugoslav</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ML</td>
<td>Mali</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MN</td>
<td>Mongolia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MR</td>
<td>Mauritania</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MW</td>
<td>Malawi</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MX</td>
<td>Mexico</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NE</td>
<td>Niger</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NL</td>
<td>Netherlands</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NO</td>
<td>Norway</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NZ</td>
<td>New Zealand</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PL</td>
<td>Poland</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PT</td>
<td>Portugal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RO</td>
<td>Romania</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RU</td>
<td>Russian Federation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD</td>
<td>Sudan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SE</td>
<td>Sweden</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SG</td>
<td>Singapore</td>
</tr>
<tr>
<td>SI</td>
<td>Slovenia</td>
<td>SK</td>
<td>Slovakia</td>
</tr>
<tr>
<td>SN</td>
<td>Senegal</td>
<td>SZ</td>
<td>Swaziland</td>
</tr>
<tr>
<td>TD</td>
<td>Chad</td>
<td>TG</td>
<td>Togo</td>
</tr>
<tr>
<td>TJ</td>
<td>Tajikistan</td>
<td>TM</td>
<td>Turkmenistan</td>
</tr>
<tr>
<td>TR</td>
<td>Turkey</td>
<td>TT</td>
<td>Trinidad and Tobago</td>
</tr>
<tr>
<td>UA</td>
<td>Ukraine</td>
<td>UG</td>
<td>Uganda</td>
</tr>
<tr>
<td>US</td>
<td>United States of America</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UZ</td>
<td>Uzbekistan</td>
<td>VN</td>
<td>Viet Nam</td>
</tr>
<tr>
<td>YU</td>
<td>Yugoslavia</td>
<td>ZW</td>
<td>Zimbabwe</td>
</tr>
</tbody>
</table>
SELF-ASSEMBLING GENES, VECTORS AND USES THEREOF

Field of the Invention

This invention relates to the construction and usage of synthetic genes for genetic engineering and gene therapy.

Background of the invention

This application claims the benefit of a provisional application U.S. Serial No. 60/070,910, filed on February 28, 1997, entitled “Self-Assembling Genes.”

Recombination at the genetic level is important for generating diversity and adaptive change within genomes of virtually all organisms. Recombinant DNA technology is based upon simple ‘cut-and-paste’ methods for manipulating nucleic acid molecules in vitro. The pieces of genetic material, or DNA are first digested with a restriction endonuclease enzyme which recognizes specific sequences within the DNA. After preparation of two or more pieces of DNA, the ends of the DNA are further manipulated, if necessary, to make them compatible for ligation or joining together. DNA ligase, together with adenosine triphosphate (ATP) is added to the genes, ligating them back together. The genetic assembly containing an origin of DNA replication and a selectable gene is then inserted into a living cell, is grown up, and is positively selected to yield a pure culture capable of providing high yields of individual recombinant DNA molecules, or their products such as RNA or protein.

Significant improvements have been made to this technology over the last two and a half decades. Numerous enzymes, end-linkers and adapter molecules have been made commercially available, which facilitate in the construction of recombinant DNA molecules. By using two restriction enzymes with different single-stranded termini or blunt ends, it is possible to directionally assemble genes (forced cloning). This reduces the amount of screening required to determine orientation. Procedures have been automated for synthesis of single-stranded gene fragments up to 200 or more nucleotides in length by means of phosphoramidite chemistry, and the instrumentation is readily available through Applied Biosystems, Inc., Foster City, CA. Such single-stranded fragments can be joined by annealing overlapping complimentary phosphorylated strands, and by enzymatically filling in the ends with DNA polymerase and DNA precursors. In this way, multiple, overlapping, single-stranded fragments can be assembled into a larger, double-stranded superstructure.
Whole genes have been synthesized by similar methods. However, it becomes increasingly difficult to use synthetic DNA strands when making genes larger than approximately one kilobase. Using gene amplification methods (e.g. polymerase chain reaction (PCR), Mullis et al., U.S. Patent 4,683,195), together with synthetic oligonucleotides, it is possible to make biologically active, synthetic retro-vectors that are capable of RNA transcription, reverse-transcription, viral packaging, and integration into genomic DNA (see for example, Hodgson, WO94/20608). Hodgson, supra, also disclosed methods for cloning of transcriptional promoters into such a vector using traditional recombinant DNA technology.

Modified restriction enzyme sites, linkers, and adapters can change the primary or secondary structure of complex nucleic acid sequences thereby altering or obliterating a desired biological activity. For example, small mutations can drastically modify transcriptional promoters or change the reading frame of coding DNA. A logical goal of vectorology is to make exact constructs, without need of fortuitous restriction sites, adapters, or linkers.

Restriction endonucleases can be grouped based on similar characteristics. In general there are three major types or classes: I, II (including IIS) and III. Class I enzymes cuts at a somewhat random site from the enzyme recognition sites (see Old and Primrose, 1994. *Principles of Gene Manipulation*. Blackwell Sciences, Inc., Cambridge, MA, p.24). Most enzymes used in molecular biology are type II enzymes. These enzymes recognize a particular target sequence (i.e., restriction endonuclease recognition site) and break the polynucleotide chains within or near to the recognition site. The type II recognition sequences are continuous or interrupted. Class IIS enzymes (i.e., type IIS enzymes) have asymmetric recognition sequences. Cleavage occurs at a distance from the recognition site.

These enzymes have been reviewed by Szybalski et al. *Gene* 100:13-26, 1991. Class III restriction enzymes are rare and are not commonly used in molecular biology.

oligonucleotides required synthesis, cloning excision and fragment purification. The oligonucleotides were used to create a complete plasmid.

Lebedenko et al. (Nucl. Acids Res. 19(24):6757-6771) illustrated the class IIS enzymes and PCR for precisely joining 3 nucleic acid molecules for convention sub-cloning using BamHI. Tomic et al. (Nucleic Acids Res., 18:1656, 1990), reported a method for site-directed mutagenesis using the polymerase chain reaction and class IIS enzymes to join two nucleic acid molecules. Two overlapping PCR primers were used where the primers included class IIS recognition sites. The primers included a region of complementarity to the template DNA and include one to a few site-directed mutations. Stemmer et al. (U.S. Patent No. 5,514,568) employed overlapping primers with class IIS enzymes to amplify a plasmid and to introduce specific mutations into DNA leaving all other positions unaltered.

There remains a need for the ordering and assembly of complex genes to overcome the problems associated with sequential sub-cloning such as multiple purification steps, the potential for sample loss, and the like. Moreover there is a need for eliminating the use of prokaryotic hosts and for minimizing or avoiding the risks associated with bacterial contamination resulting from the use of bacteria as intermediaries in the cloning process. Further, there remains a need for efficient methods to assemble large nucleic acid molecules or many-fragmented nucleic acid assemblies with precision.

**Brief Description of the Figures**

Fig. 1A. provides one schematic of six double stranded DNA fragments, each terminus comprising a unique overhanging two-nucleotide sequence complementary to only one other terminus.

Fig. 1B. illustrates a three-piece ligation where 100% of the clones tested contained the predicted fragment order and desired fragment orientation.

Fig. 2. illustrates the use of a class IIS restriction endonuclease (as one example, Bpm1), restriction endonuclease recognition site and the selection of cohesive overhanging ends.

Fig. 3A. illustrates an exemplary retrotransposon-derived vector including a murine VL30 LTR (NLV-3) and packaging signal, an internal ribosome entry site (IRES) from encephalomyocarditis virus (EMCV), a gene encoding a green fluorescent protein (GFP), additional internal VL30 sequences (solid bar), SV40 early region promoter and Tn5
aminoglycosidase phosphotransferase (neo) gene, PBR322 plasmid origin of replication and a plus-strand primer binding site (VL30). An exemplary vector sequence is provided as VLBPGN (SEQ ID NO:1). Fig 3B is an illustration of an LTR with the insertion of a U3 (transcriptional promoter) region rescued by reverse transcriptase-polymerase chain reaction (RT-PCR). The promoter is amplified from the RNA of a cell expressing the VL30 U3 region. Complementary overhanging ends are created using class IIS restriction endonuclease digestion sites within the LTR and within the promoter. Fig. 3C provides the linear structure of a VL30 RNA transcript from a mouse cell with a U3 region near the 3’-terminus of the RNA molecule. PCR primers include a class IIS enzyme recognition site to amplify the U3 region from the RNA resulting in a double stranded DNA molecule. Cleavage with a class IIS enzyme (here BpmI), results in a double-stranded DNA molecule with end complementary to a site in the vector of Fig. 3A.

Fig. 4A. is a schematic illustrating steps for assembling a combinatorial library of cis- or trans-acting nucleic acid sequences for assembly and screening, useful for the rescue of biologically active species. Fig. 4b is a diagram of a U3 (transcriptional enhancer and promoter region of an LTR illustrating several sub-divisions of the transcriptional control region, including a distal enhancer region, an enhancer repeat region, a medial promoter and a proximal promoter. These regions have been described for other vectors in Hodgson et al. (1996. “Construction, Transmission and Expression of Synthetic VL30 Vectors” in Hodgson ed. Retro-vectors for Human Gene Therapy. RG Landes Company, Austin TX). Segments of these regions are amplified using primers for highly conserved sequences. Highly conserved sequences are determine based on a comparison of known VL30 sequences such as provided in Fig. 4.2 of Hodgson, 1996, infra). The parts are joined by annealing and ligation to provide an ordered assembly. Each construct is an allele or a representative of allelic variation in the combinatorial library.

Fig. 5 discloses two transcriptional promoters that have been rescued from mouse VL30 RNA sequences isolated from a mouse T-helper cell library. These promoters were assembled into a vector and introduced into retroviral helper cells and packaged into recombinant retrovirus for introduction into human T-cells. After transduction to human T cells, a β-galactosidase reporter gene was expressed from the T cell-derived promoters.

Fig. 6 discloses 10 biologically active mouse VL30 promoters obtained from mouse liver RNA. These promoters were introduced into the vector of SEQ ID NO:1. The vectors
were introduced into retroviral helper cells and then packaged into retrovirus where they were introduced into human liver cells. The cells expressed the green fluorescent protein.

**Fig. 7** illustrates a similarity plot of nucleotide sequences found in VL30 U3 regions.

**Fig. 8** illustrates a retro-vector comprising six double-stranded DNA fragments that were self-assembled into a circular structure using unique overlapping termini created using class IIS restriction endonucleases. Three templates and twelve primers were used in conjunction with three class IIS enzymes to make the six fragments that were ligated in a single step. The vector was efficiently self-assembled and was effectively transmitted by both DNA transfection as well as by retroviral transduction of the self-assembled DNA, without molecular cloning through a prokaryotic host (see Example 2).

**BRIEF SUMMARY OF THE INVENTION**

The invention described herein provides seamless, directional, ordered construction of complex DNA molecules, vectors and libraries. More particularly, it enables gene constructs to be assembled with greater efficiency and precision, and it enables multiple gene fragments to be assembled in the correct order and orientation without disturbing the internal structure of the gene. The method utilizes *in vitro* assembly of nucleic acid fragments and relies upon the unusual ability of certain enzymes to digest nucleic acid molecules at pre-determined sites without disrupting the structure of the gene. It is especially useful for the construction of genetic vectors for gene therapy or genetic engineering of cells and organisms. A particular application of the invention is in combinatorial, or evolutionary genetics, where it enables a large number of non-random, self-assembled constructs to be screened simultaneously for function.

In a preferred embodiment of this invention, the invention relates to a method for assembling a gene or gene vector comprising the steps of: a) designing at least 6 primers to produce to amplify at least three fragments in at least three separate polymerase chain reactions wherein each primer comprises at least one predetermined restriction endonuclease recognition site that recognizes a restriction endonuclease that cleaves at a distance from the recognition site, a sequence complementary to a template nucleic acid for amplification, and bases positioned at the restriction endonuclease cleavage site that are selected to be complementary to only one other overhanging created from enzymatic cleavage of the fragments; b) combining the primers with template nucleic acid and performing the
polymerase chain reaction to produce multiple copies of an amplified template fragment incorporating the restriction endonuclease recognition site; c) digesting the amplified template fragments with one or more restriction endonucleases that recognize the restriction endonuclease recognition site of the primers to create overhanging termini wherein each overhanging termini is complementary to only one other overhanging termini on another fragment; and d) combining the amplified and digested template fragments in a ligation reaction to produce a directionally ordered gene, nucleic acid fragment or gene vector.

In a preferred aspect of this embodiment, the restriction endonuclease is at least one class IIS restriction endonuclease and preferably, the class IIS restriction endonuclease is selected from the group consisting of: AlwI, Alw26I, BbsI, BbvI, BbvII, BpmI, BsmAI, BsmI, BsmBI, BspMI, BsrI, BsrDI, Eco57I, EarI, FokI, GsuI, HgaI, HphI, MboII, MnlI, PseI, SapI, SfaNI, TaqII, Tth1111I. Still more preferably, class II restriction endonuclease recognition sites (to be distinguished from class IIS restriction endonuclease recognition sites), linkers, or adapters are not used to create the gene or gene vector. In one embodiment, the product of the ligation reaction is introduced into prokaryotic or eukaryotic cells. Preferably, at least one template nucleic acid sequence is chosen from the group consisting of: transcriptional regulatory sequences; genetic vectors; introns and/or exons; viral encapsidation sequences; integration signals intended for introducing nucleic acid molecules into other nucleic acid molecules; retrotransposon(s); VL30 elements; or multiple allelic forms of a sequence.

In another preferred aspect of this embodiment, the method is used to generate combinatorial libraries of a target sequence. Preferably, the target sequence is part or all of a gene. In one embodiment, the gene encodes a protein. In one embodiment, the primers amplify allelic variants of part or all of a gene.

In still another preferred aspect of this embodiment, the product of the ligation reaction is passed between eukaryotic cells using a virus particle, by cell fusion, or by transfection. Preferably the product of the ligation reaction is not introduced into prokaryotic cells. Moreover, the method further comprises combining at least one screening or selection step to select the products of the ligation reaction. In one embodiment, the product of the ligation reaction is mutated during passage in cells in order to generate genetic diversity and preferably the product of the ligation reaction is mutated by homologous recombination during passage in cells.
In another aspect of this embodiment, the method is used to isolate and identify regulatory sequences from a cell. In another aspect of this embodiment, cells containing the product of the ligation reaction are selected for enhanced biological activity. Preferably, the cells containing the product of the ligation reaction are selected for tissue-specific, hormone-specific or developmental-specific gene expression. Also preferably, the ligation reaction is a circularized gene vector.

In another embodiment of this invention, the invention relates to a nucleic acid primer having a 5’ and a 3’ end to amplify a nucleic acid fragment for the ligation of at least two fragments comprising: a restriction endonuclease recognition site that recognizes a restriction endonuclease, wherein the restriction endonuclease cleaves at a distance from the recognition site and creates overhanging termini; a sequence complementary to a template sequence to be amplified to produce the nucleic acid fragment; at least two nucleic acid bases positioned at the restriction endonuclease cleavage site and that form an overhanging terminus after cleavage by the restriction endonuclease, wherein the at least two nucleic acid bases are selected to be complementary to only one other overhanging terminus on another fragment of the ligation; and an affinity handle on the 5’ end of the primer. Preferably the primer further comprises an anchor to provide stability to the restriction enzyme at the restriction enzyme recognition site.

In yet another embodiment of this invention, the invention relates to a method for isolating and identifying promoters comprising the steps of: a) obtaining a vector comprising at least a portion of a promoter region from a retrovirus transposon LTR and having two non-complementary overhanging termini; b) designing at least two PCR primers to amplify at least one region of a retrovirus transposon LTR from template nucleic acid to produce at least one nucleic acid fragment wherein each primer comprises at least one predetermined restriction endonuclease recognition site that recognizes a restriction endonuclease that cleaves at a distance from the recognition site, a sequence complementary to a template sequence from a retrovirus transposon, and bases positioned at the restriction endonuclease cleavage site that are selected to be complementary to only one other overhanging terminus of the vector wherein the restriction endonuclease cleavage site is created from enzymatic cleavage of the fragments; b) combining the primers with template nucleic acid and performing a polymerase chain reaction to produce multiple copies of an amplified template fragment incorporating the restriction endonuclease recognition site; c)
digesting the amplified template fragments with one or more restriction endonuclease that recognize the restriction endonuclease recognition site of the primer to create overhanging termini; and combining the amplified and digested template fragment in a ligation reaction with the vector to produce a gene vector with an intact LTR sequence. In one embodiment of this aspect of the invention, the template nucleic acid is DNA or RNA. In another embodiment of this aspect of the invention, the method further comprises the step of sequencing the insert to identify the promoter sequence. In one embodiment promoter sequences of SEQ ID NOS:1-13 identified using the methods of claim.

Detailed Description of the Invention

In one embodiment of this invention, the invention relates to the seamless, oriented self-assembly of at least three DNA fragments having overlapping unique cohesive ends generated by the enzymatic cleavage of at least one restriction endonuclease that is capable of cleaving at a site distant to the restriction enzyme recognition site. Preferably the restriction endonucleases employed in this invention are class IIS restriction endonucleases. These enzymes recognize a predetermined group of nucleotides and cleave at a distance characteristic of the particular endonuclease from the recognition site. The term “unique cohesive ends” is used herein to refer to the notion that the cleavage site for the endonucleases of this invention can be manipulated to produce overhanging ends with unique termini selected by the investigator. The term “complementary” as used herein in reference to the overhanging ends of the fragments of this invention refers to standard complementarity recognized in the field of molecular biology. For example, the nucleotides sequence 5’-TAG-3’ is said to be complementary to the nucleotide sequence 5’-CTA-3’. The term “PCR” is used generally to refer to the polymerase chain reaction and its variations, including RT-PCR as well as other gene amplification techniques employing primers.

In a first step for practicing one embodiment of this invention, a series of at least three overlapping fragments are created through the selection and creation of primers incorporating at least one class IIS restriction enzyme recognition sequence. The oligonucleotide primers of this invention are designed to amplify one or more nucleic acid fragments and comprise a sequence complementary to a target sequence for gene amplification, a recognition sequence for a restriction endonuclease that cleaves DNA at a distance from the recognition sequence (such as a class IIS restriction enzyme) and bases
positioned at the restriction endonuclease cleavage site that are preferably unique and complementary to only one other overhanging termini in the annealing/ligation reaction that generates the complex nucleic acid molecules. Optionally, the primers of this invention can include an “affinity handle for cleanup” at the 5’end. These sequences can be of any length, preferably at least about 6 bp and the sequences extend the primer in the 5’ direction from the restriction enzyme recognition site. This extra length gives many enzymes greater stability and improved activity. In addition, the sequence can be used for recognition and removal of the ends of the primers (either undigested fragments or digested ends of primers) using complementary nucleotide sequences bound to a solid support (such as cellulose, nitrocellulose or silica). Incubation with, or passage over a column or support containing the complementary sequences can be used to remove the tags by allowing them to anneal or hybridize. The nucleic acid can then be eluted from the column. Adapters can also be used in this invention. For purposes of this invention, adapters refer to double stranded fragments containing an enzyme recognition site, according to this invention. The adapters are ligated to double stranded DNA molecules, creating a fragment analogous to a PCR fragment with similar sites derived from a primer. The primers or adapters can be prepared using a number of methods for synthesizing oligonucleotides known in the art. For example instruments for producing oligonucleotides are available from Applied Biosystems, Inc., Foster City, CA.

In one example, for the design of an oligonucleotide primer for use in this invention, the particular complementary bases that will form the site for hybridization of the primer to template (i.e., target DNA or RNA) are selected. A restriction endonuclease recognition site is selected followed by a number of nucleotides to be positioned between the recognition site and the cleavage site. The nucleotides of the cleavage site are selected to include overhanging regions formed from the restriction endonuclease cleavage that are complementary to the overhanging regions of an adjacent fragment in the annealing/ligation reaction.

The length of the primer used in this invention can vary, but preferably the primer length is up to about 80 bases and preferably up to about 50 bases. In addition the primers are preferably at least about 15 bases in length and preferably at least about 25 bases in length. The 5’ region of the primer contains preferably at least about 6, preferably at least about 10 and still more preferably at least about 16-18 bases that are not complementary to the template DNA or RNA. Further, the primer incorporates a restriction endonuclease
recognition site preferably 5' to the region of complementarity and a restriction endonuclease digestion site preferably 5' to the region of complementarity or within the region of complementarity. There are a variety of restriction endonucleases that cleave at a distance from the restriction endonuclease recognition site of a DNA strand and a variety of enzymes that are commercially available from New England Biolabs are provided in Table 1.

Table 1. Restriction endonucleases useful in the construction of self-assembling genes

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Site size (bp)</th>
<th>Distance to overlap</th>
<th>Size of overlap</th>
<th>Overlap type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alw26I</td>
<td>5</td>
<td>1-5bp</td>
<td>4bp</td>
<td>5'-Overhang</td>
</tr>
<tr>
<td>BbsI</td>
<td>6</td>
<td>2-6bp</td>
<td>4bp</td>
<td>5'-overhang</td>
</tr>
<tr>
<td>BpmI</td>
<td>6</td>
<td>16-14bp</td>
<td>2bp</td>
<td>3'-overhang</td>
</tr>
<tr>
<td>BsmBI</td>
<td>6</td>
<td>1-5bp</td>
<td>4bp</td>
<td>5'-overhang</td>
</tr>
<tr>
<td>BspMI</td>
<td>6</td>
<td>4-8bp</td>
<td>4bp</td>
<td>5'-overhang</td>
</tr>
<tr>
<td>BsrDI</td>
<td>6</td>
<td>0-2bp</td>
<td>2bp</td>
<td>3'-overhang</td>
</tr>
<tr>
<td>Eco57I</td>
<td>6</td>
<td>16-14bp</td>
<td>2bp</td>
<td>3'-overhang</td>
</tr>
<tr>
<td>FokI</td>
<td>5</td>
<td>9-13bp</td>
<td>4bp</td>
<td>5'-overhang</td>
</tr>
<tr>
<td>Hgal</td>
<td>5</td>
<td>5-10bp</td>
<td>5bp</td>
<td>5'-overhang</td>
</tr>
<tr>
<td>HphI</td>
<td>5</td>
<td>8-7bp</td>
<td>1bp</td>
<td>3'-overhang</td>
</tr>
<tr>
<td>MnlI</td>
<td>5</td>
<td>7-6bp</td>
<td>1bp</td>
<td>3'-overhang</td>
</tr>
<tr>
<td>Psel</td>
<td>5</td>
<td>4-5bp</td>
<td>1bp</td>
<td>5'-overhang</td>
</tr>
<tr>
<td>SapI</td>
<td>7</td>
<td>1-4bp</td>
<td>3bp</td>
<td>5'-overhang</td>
</tr>
<tr>
<td>SfaNI</td>
<td>5</td>
<td>5-9bp</td>
<td>4bp</td>
<td>5'-overhang</td>
</tr>
</tbody>
</table>

In addition to the enzymes provided in Table 1, other restriction endonucleases that cleave at a distance from their restriction endonuclease recognition site include, but are not limited to, AlwI, BbsI, BbvI, BbvII, BsmAI, BsmI, BsrI, Earl, GsuI, MboII, TaqII, Tth1111II and their respective isoschizomers. These and other enzymes are known in the art and many are available from other manufacturers. The primers can be prepared to produce either 5'-overlapping ends or 3'-overlapping ends, as long as they are both are either 5'-overlapping ends or 3'-overlapping ends and are complementary to one other set of overlapping ends.

In the case of BpmI (see Example 1), the enzyme digests asymmetrically, 14-16 bp from the 3'-nucleotide of the recognition site. The resulting cleavage has a 3'-overhanging end of 2 bp. A second primer is then designed with a complementary
overhanging end, and it is used to generate the adjoining fragment terminus. At the opposite ends of the two fragments that are to be joined, similar complementary, overhanging ends are designed.

The oligonucleotides are then combined with template nucleic acid (either DNA or RNA, e.g., such as for reverse transcriptase polymerase chain reaction (RT-PCR)) containing bases complementary to at least a 3' portion of the primers (also referred to herein as "templates"). In one embodiment, the fragments are gene-amplified by PCR, RT-PCR or another gene amplification process using established PCR protocols such as those provided with PCR amplification kits, including those available from Perkin-Elmer Corp. (Emeryville, California). Preferably, the PCR products are analyzed by electrophoresis on a gel, such as an agarose gel and still more preferably the fragments of the predicted size are purified free of excess primers and small byproducts (such as by purification through a small column, such as a Qiagen™ column (Qiagen, Valencia, CA)). Following amplification or purification, the fragments are digested with the restriction endonuclease recognizing the restriction endonuclease recognition site in the primers. The digested fragments are then purified from the digested ends of the primers, preferably by preparative agarose gel electrophoresis. The fragments are combined, annealed and are ligated using standard hybridization and ligation conditions known for cloning (see Ausubel et al., Current Protocols in Molecular Biology, John Wiley & Sons, 1994).

Fig. 1A illustrates an example of a self-assembling gene construct (SEQ ID NO:1) comprising six fragments, each having unique overhanging dinucleotide ends. In this example, the ends of the fragments prepared by the methods of this invention are constructed using primers that include BpmI restriction endonuclease recognition sites. It will be understood by those of ordinary skill in the art that one or more other restriction endonucleases (such as those of Table 1) could similarly be used for the self-assembling product of Fig. 1A. In a preferred embodiment, the primers were created as described above and preferably the 3' ends of the primers are non-palindromic (i.e., non self-complementary) to prevent self-annealing of such fragments. Each fragment in this example preferably joins to only one other dinucleotide overhang in the annealing/ligation mixture, assuring ligation only to the intended fragment partner. An advantage of this strategy is that the formation of concatamers or multimers is minimal. The restriction endonuclease site is removed by
digestion with the restriction endonuclease, leaving the junction free of the extra DNA sequences associated with the site.

Using a single restriction endonuclease with a dinucleotide overhang (for example, using the enzyme *BpmI*) up to six pieces of genetic material can be joined together in a linear or circular form (such as a vector) without the need to perform sub-cloning procedures or detailed analysis of individual products because six unique combinations of dinucleotide overhangs create a directional clone with extremely high fidelity. With enzymes digesting single-base overlaps, only two fragments can be joined with positional and directional precision. With enzymes digesting three-base overlaps, 4^3/2, or 32 fragments can be so joined in the correct order and orientation. Therefore, this invention also relates to the use of restriction endonuclease recognition sites that facilitate cleavage by restriction endonucleases with three-base overlaps and self-assembly gene constructs including 32 fragments. Alternatively, a combination of restriction endonuclease recognition sites for use with a combination of restriction enzymes that create two-base or three-base overlaps can be used. Each enzyme has its characteristic limits to self-assembly imposed by the size of the overlap. For example, there are sixteen dinucleotides, therefore *BpmI* fragments (which have two dinucleotide ends each) are limited to eight for the purpose of self-assembly; therefore in another embodiment of this invention an assembly comprising eight fragments is contemplated. However, four of the sixteen dinucleotides are palindromes. Use of these palindromic dinucleotides can create some infidelity in the annealing/ligation reaction. The enzyme *HgaI* has a five base overlap, and there are 1,024 pentanucleotide combinations, permitting 512 fragments to be ligated together directionally and in order (no palindromes). The fragments to be joined at a particular place are designed to have their cut sites aligned, so that the overlapping region fits together. In some cases, the target sequences will contain natural restriction endonuclease recognition sites for the enzyme that is being used, such as one or more internal *BpmI* sites. These sites have the potential to self-relicate during vector or gene construction or they can be by passed by using a substitute enzyme in the primers (for example, *Eco 571* can substitute for *BpmI*). Alternatively, these sites can be removed by site-directed mutagenesis after consideration to the consequences of the mutagenized sequence to the gene or vector.

In addition to class IIS enzymes, class II restriction endonucleases can be used. These enzymes have intrinsic methylation activity that affects the outcome in either a
negative or a positive way, depending on the purpose for which it is used. In a preferred embodiment, the methylation activity of class II enzymes is ablated by mutation or by genetic engineering to convert the enzyme to an effective class IIS enzyme to expand the repertoire of useful enzymes for this invention.

In another aspect of this invention, the primer design and target fragment sequence selection can be automated (see Example 5) using a computer to assist in the selection of unique overhanging ends that have complementarity only to the overhanging end of an adjacent fragment.

Therefore, this invention permits high-fidelity annealing and ligation of six or more fragments with unique overhanging termini complementary to a single other overhanging termini. Any multitude of combinations can be created by combining the type of overhanging termini that can be created. Moreover, if one is willing to sacrifice the fidelity of the reaction, a variety of combinations can be used to anneal a variety of fragment numbers. In these cases, some selection may be necessary, such as size selection of the resulting fragment based on electrophoretic migration or restriction endonuclease profiling, both methods well known to those of ordinary skill in the art.

It is also necessary to have a high per-step efficiency (e.g., each step in the process is performed with an efficiency of at least 80%) to effectively ligate large numbers of fragments without error. Where large numbers of fragments are used, the purity of the fragments becomes important. This means that for large numbers of fragments, the digested DNA fragments for annealing and ligation should be substantially pure. If undigested fragments, digested ends of primers, degraded or partially degraded molecules are present, they can decrease the purity and affect the fidelity of the product. Therefore, it is particularly desirable to ensure complete digestion of both ends of each fragment and to remove all of the digested ends from the fragments prior to including the fragments in an annealing and ligation reaction. The use of Qiagen columns for oligonucleotide removal prior to digestion is generally sufficient to permit efficient digestion of the fragments. Agarose gel isolation is desirable after digestion particularly where the product contains some fragments that do not appear to be full length. The use of an analytical gel before and after digestion helps in determining whether both oligonucleotide tags have been removed. The isolation of fragments from agarose gels preferably avoids the use of ultraviolet light and exposure of the
DNA to ethidium bromide is also preferably avoided. These methods can be avoided by running replicate lanes and staining only a portion of the gel.

The fragments and vector are then digested to yield fully complementary ends, and the fragments are preferably again purified, as described above (such as through a Qiagen column or by gel isolation). The purified fragments are ligated together in a test tube, under standard conditions, such as using bacteriophage T4 DNA ligase and ATP. Preferred ligations include at least 20μg/ml total DNA concentration in the ligation mix to favor intermolecular interactions, and an equimolar ratio of fragments to be ligated. Where a prokaryotic intermediary is used, the ligated assemblage is transformed into a bacterium, such as an E. coli host, and the colonies are: selected with a drug (such as an ampicillin, tetracycline, or kanamycin marker). The colonies can then be selected either by individually selecting colonies or growing a mass culture, such as where a vector library has been created. Restriction enzyme analysis can be used to determine the identity of individual constructs or to assess the validation of the combination of plasmids. The plasmids can then be grown up and used as needed.

In one embodiment of this invention, at least a portion of a vector is used as one of the fragments for the ligation of at least three fragments according to this invention. In one example, where a vector is used as one of the starting fragments, two restriction endonuclease recognition sites recognizing an enzyme that cleaves at a distance from the recognition site, such as at least one BpmI site, can also be introduced into the vector. This permits the vector to be digested with the restriction endonuclease to produce a product having ends complementary to two ends of the insert DNA fragments. The vector can be made by amplifying a plasmid or portion thereof using the primers of this invention. Thus, the vector can also be constructed to include a variety of restriction endonuclease recognition sites using a variety of restriction endonucleases, including a variety of class II restriction endonucleases. In some cases, the target fragments for amplification will contain natural restriction endonuclease recognition sites for the enzyme that is being used for the self-assembly, such as for example, a fragment that includes one or more internal BpmI sites. Care should be taken either to utilize the complementarity of the naturally occurring site to reform the fragment as it originally existed or to eliminate the restriction endonuclease recognition site using, for example, site-directed mutagenesis. Preferably, the restriction endonuclease recognition site is be substituted for a different enzyme (in the case of BpmI.
substituting *Eco57I* or *BsrDI*) that has an equivalent structure at its ends. Two or more fragments of insert or two or more fragments of vector with at least one insert are amplified using primers according to this invention.

The exemplary enzyme, *BpmI* digests DNA 14-16 base pairs (bp) from the 3’-nucleotide of the recognition sequence (RS). Thus, by placing the RS exactly 14-16 bp from the desired dinucleotide cut site, the practitioner tags the dinucleotide for ligation with another dinucleotide that is exactly complementary to it. Such a complementary dinucleotide can be inserted by using the same enzyme and RS to make another fragment which fits the first exactly, as illustrated in Fig. 1. Because there are sixteen possible dinucleotide combinations (including twelve combinations that do not have palindromic ends), it is possible to create up to six fragments with unique dinucleotides, and it is also possible to join them all together in a predetermined order and orientation (Fig 1A). In addition, the palindromic sequences (such as AT, CG, TA, and GC) could also be used, although inefficiency and incorrect ligation will result from the self-complementarity of these sequences. It is furthermore possible and desirable to have three or more fragments joined in this way, such that the construct is circular as in Fig. 1, comprising a vector that may be grown in a bacterial and/or eukaryotic host cell. If the genetic construct is to be used as a vector, the vector should be designed to include a proper origin of replication to enable it to replicate in a particular cell. For example, a prokaryotic origin of replication such as a coliform plasmid origin of replication enables circular DNAs to be propagated in *E. coli* host cells. It is desirable to have at least one selectable marker, such as a neomycin marker that enables recovery of the clone through a selection process. It is also desirable, but not essential, to have two or more selectable genetic elements, to permit dual selection. For example, if one of the fragments contains a prokaryotic plasmid origin of replication, and another fragment contains a selectable marker, then the two fragments are both selectable, since the construct will grow in prokaryotic cells in the presence of a selection drug (such as ampicillin) only when both fragments are present. Drug selection can be combined with the methods of directed self-assembly to assure a high percentage of correct products. Because of the unique complementarity of the fragments, each contributes a selectable element that leads to recovery of a high percentage of correct products.

For prokaryotic vector construction, at least one fragment should contain a prokaryotic origin of replication and one fragment should contain a drug resistance marker.
gene. However, an advantage of the methods of this invention is that the construct can be introduced directly into eukaryotic cells. Here no plasmid origin of replication is necessary and no prokaryotic selectable marker or other prokaryotic nucleic acid sequence is necessary. In cases where the vector is subject to regulatory approval or where optimal gene function is necessary, it may be undesirable to include prokaryotic sequences, such as extraneous plasmids or expressed prokaryotic fragments particularly if the sequences contain immunostimulatory sites that can lead to activation of the intracellular immune system and inactivation of a gene product (see Krieg et al., J. Lab. Clin. Med., 128:128-133, 1996) or to avoid risks of endotoxin contamination. Moreover, the use of self-assembled product, according to the methods of this invention saves labor and time involved in the screening process.

Thus, in a preferred embodiment of the invention, the nucleic acid fragments are self-assembled in vitro, and are transferred directly into eukaryotic cells, by transfection, injection, or other methods known in the art. In one embodiment the cells receiving the assembled product of this invention are helper cells for recombinant virus assembly (including, but not limited to retroviral helper cells for retroviral or retrotransposon vectors, adenovirus helper cells for adenovirus vectors or herpes simplex virus helper cells for herpes simplex vectors). Alternatively, the assembled product can be introduced into cells along with a helper virus or the assembled product can be introduced into target cells for direct expression. The assembled product can be a vector, a minichromosome vector, a portion of a chromosome, or the like. In the preferred case of a retroviral vector, the genes are first transfected into a first helper cell line (such as ecotropic helper cells, GP+E86 (Markowitz et al. J. Virol. 862:1120-1124, 1988). The retrovirus-containing supernatant from these cells is then filtered (0.45mm Nalgene filters) preferably 48-72 hours after transfection and the filtrate is transferred to a second complementation retroviral helper cell line (such as PA317 retroviral helper cells, Miller et al., Mol. Cell. Biol. 6:2895-2902, 1986). After an additional 48 h, the second helper cell line is selected with the marker drug (such as the drug G418 for the selectable neomycin (neo) marker gene), until only drug-resistant cells remain. These cells contain stably integrated vectors that can be used to repeatedly transduce human cells. Advantageously, in the case of adenovirus vectors or other large eukaryotic -derived vectors including eukaryotic virus-derived vectors, it may be impossible to propagate them in prokaryotic hosts. The gene self-assembly method of the instant invention provides an
alternative to in vitro recombination method of gene construction by permitting large constructs to be constructed.

One advantage of introducing the assembled product of this invention into a helper cell line to produce recombinant virus for the introduction of a gene or nucleic acid complex into a cell is that the assembled product will be auto-selected by the cells during the packaging process. Therefore, even where the overhanging termini have palindromic sequences, where there is more than one (but preferably less than four) unique complementary matches for a particular overhanging termini, or where concatamers have formed, only the correct or functional assembled products are expressed, transmitted, and assembled into virus. When the virus is then introduced into cells, the use of a reporter gene or another selectable marker provides yet a second layer of security for the selection of cells containing a properly assembled construct. For example, where a retrovirus helper cell line is used to produce a recombinant retrovirus containing the product of this invention (for retrovirus, RNA transcribed from the DNA product of the invention becomes packaged into the virus particle), a retrovirus-derived vector is transcribed as RNA and transmitted by packaging the RNA in a retrovirus particle. In order to be properly transmitted as a virus, the construct must be: 1) transcribed as RNA in a vector producer cell; 2) packaged into viral particles; 3) reverse transcribed into double-stranded DNA (in the recipient cell); and 4) integrated into the host chromosome. Each of these steps requires specific cis-acting sequences that must be correctly positioned within the vector. Thus, passage via retrovirus (or by other virus) is a means of auto-selection for the essential sequences.

In one application of the methods of this invention, the methods are used to rescue expressed sequences from RNA, or genomic sequences from cell DNA without disrupting the promoter sequences. Cellular transcriptional promoters are typically difficult to identify and isolate because they are generally not included in the RNA molecule and often extend over a considerable distance in a chromosome. One application of this invention relates to a promoter rescue technique that permits the entire promoter, or a fragment of a promoter to be isolated and cloned directly in to an expression vector without disruption of the flanking sequences. Promoter rescue techniques are known and include WO 94/20608 to Hodgson.

In a preferred embodiment of the invention, transcriptional promoters are cloned in a transcriptionally active manner for the selection and identification of new and/or
of tissue or cell-specific promoters enabling them to be used, selected, or screened for activity directly. For example, Fig. 3 illustrates one example of the formation of a vector for the incorporation of promoter sequences and the ultimate identification of those sequences using an exemplary plasmid VLBPGN (SEQ ID NO:1) as provided in Example 3, with Bpm1 sites located within the locus of a retrotransposon (VL30) long terminal repeat (LTR). These methods preserve the structure and functionality of transcription factor response elements. The characteristic secondary structure of the LTR RNA remains very similar to the original LTR from which the promoter was rescued, thus preserving the important features of the original RNA/DNA molecule. Those of ordinary skill in the art will recognize that any of a variety of primers can be used with a variety of vectors and that the constructs of Figs 2 and 3 are exemplary and not limiting.

Fig. 2 illustrates the primers used to amplify the promoter insert (identified at a and c in Fig.2), and the insert region of the LTR (boxed), both of which can be digested at the same nucleotide position with Bpm1, to ensure a proper and seamless fit. In this example, after digestion of the vector, the two Bpm1 sites leave non-complementary ends (a 3' CC overhang on one end, and a 3' GC overhang on the other). Thus, the ends will not efficiently anneal or ligate to one another. However, the complementary termini of the insert serves as linkage, enabling the plasmid to be completed by ligation.

In the example illustrated in Fig. 2, the terminus on the 3'-side (GC) is palindromic. Palindromic termini are self-complementary and can therefore ligate to themselves or to an identical terminus facing the opposite way (forming concatamers in the opposite direction). Despite the presence of palindromic termini and despite the potential for reduced fidelity in the self-assembling process, a large percentage of clones obtained by inserting promoter sequences into VLBPGN were assembled correctly (20/23). These levels are reduced somewhat when three or more fragments are combined for self-assembly, according to this invention and preferably, the use of palindromic termini are avoided when even numbers of nucleotides are exposed as overhanging termini because with even numbers of nucleotides there is an axis of symmetry. As noted above, where five base overhangs are used there are 1024 possible combinations of five nucleotides [(4)⁵], yet none of them is palindromic.

The vector of Fig. 3 is an example of a particular type of vector that is known as a retrotransposon vector. Retrotransposon vectors are described and reviewed in Hodgson
et al., 1996 *Retro-Vectors for Human Gene Therapy*. RG Landes Company, Austin TX, chapter 5 and see US Patent 5,354,674 to Hodgson. This type of vector is derived from a mouse cellular retro-transposon element that has no essential viral or cellular genes, and that has little sequence similarity to a retrovirus. However, this RNA (known as VL30 [virus-like, 30S]) has all the necessary cis-acting structural elements (such as LTRs and primer binding sites) required for efficient transmission by a type C murine or primate retrovirus. Thus, it is a parasite transmitted by retroviruses that is also expressed as a cellular RNA in most mouse cells and tissues. This RNA becomes packaged into retroviral particles when the mouse cells become infected by retrovirus. The retrovirus then transmits the VL30 (or a VL30 vector) to the next infected cell (which can be a human cell). The RNA is then reverse transcribed and integrated into the DNA of the host cell.

Some advantages of VL30 vectors (over retrovirus-derived vectors) are: 1) lack of viral genes and other sequence homology that could lead to replication competent retrovirus (RCR); 2) ability to be expressed long-term *in vivo*; 3) a variety of LTR transcriptional promoters that can be expressed in various tissues and under the influence of various hormones and other stimuli; and 4) the ability to express genes in a number of cell types that are targets of gene therapy. An additional advantage is that VL30 parts can be switched with those of classical retrovirus-derived vectors. For example, the LTR or packaging signal of VL30 can be used in place of the equivalent retroviral signal. The ability to make mixed, or chimeric retro-vectors is a special application of gene self assembly technology.

Using a specific primer set, such as that shown in Fig. 2, or others, as taught in this invention, it is possible to amplify the U3 sequences expressed in the RNA of many different types of mouse cells. This is done using standard RNA isolation methods (Ausubel *et al.*, *supra*), coupled with extensive digestion with ribonuclease-free dexoxygenribonuclease, to eliminate residual DNA. Thus, to obtain a promoter that is expressed in the liver, one isolates RNA from liver and uses an RT-PCR procedure, such as those known in the art, with the primers to amplify the desired promoters. Fig. 6 illustrates liver RNA-derived promoters obtained using the methods of this invention. However, the promoters can also be derived by conventional PCR from cDNA libraries (Fig. 5 illustrates T cell-derived promoters that were obtained in this manner). It is also possible to use the well-known hormonal and pharmacological inducibility of VL30 LTRs to find LTRs that are responsive to peptides.
hormones, and cytokines (for a table and description of VL30 pharmacologic responses (see Hodgson et al., 1996 *Retro-Vectors for Human Gene Therapy*. RG Landes Company, Austin TX, chapter 4, and Fig. 4.2). Examples of substances inducing various VL30 promoters to high levels include: epidermal growth factor, basic fibroblast growth factor, insulin, erythropoietin, glucocorticoid hormones, activators of cyclic 3'\'-5'\'AMP, and others. To rescue promoters with pharmacological responsiveness, cells or animals stimulated with the desired pharmacological agent are subjected to the RT-PCR procedure and the resulting U3 regions are cloned into a vector, (such as the exemplary VLBPGN) and are tested for inducibility. Standard RNA blotting procedures can be used before isolating VL30 promoters, to determine whether a particular drug or hormone causes induction of VL30 RNA expression in a particular mouse cell or tissue. After the promoter has been rescued, the vector is transmitted via retrovirus to the target cell (possibly a human equivalent of the mouse cell from which the promoter was rescued). After selection with the drug G418 (400-700 µg/ml, for 7-10 days) to select against cells not containing the vector, the target cell population is challenged with the pharmacological agent of choice. Reporter gene expression (in the example, GFP) or RNA expression, as determined by RNA blotting, can be used as an assay of gene inducibility by the agent (for exemplary gene expression methods, see Chakraborty et al., *Biochem. Biophys Res. Commun*. 209:677-683, 1995).

Using any specific primer set designed for use with VL30 retro-elements and using total cellular RNA from a particular mouse cell type as a template for RT-PCR, (using commercially available kits and methods therein) candidate promoter elements can be amplified. This method is useful for the identification of mouse-derived promoters and in particular the method is useful for the identification of cell-type specific or tissue-specific promoters from a mouse and for the selection of these promoters and the identification of tissue-specific or cell-specific promoters that function in human cells. Thus, these types of vectors and the methods for using these vectors permits the identification of promoters to permit controlled transcription of a foreign gene. The promoters, originally obtained from the mouse, can be used to effect tissue-specific or cell-specific expression in a human or animal liver cell such as a hepatocyte, or in a human blood cell such as a T-helper cell or in an erythrocyte (red blood cell). Methods are disclosed in Example 2 for the screening and selection of the promoters from a library of amplified promoter sequences. Other methods are well known to those of ordinary skill in the art. The specificity of the selected promoter
can be assessed, for example, by introducing a selectable marker under the control of the test promoter in question and introducing this construct into various cells to assess the ability of the promoter to selectively regulate expression.

The amplified fragments represent U3 promoter regions from any RNA species expressed in the originating cells and their abundance will be in approximate proportion to the number of expressed copies of RNA in the original mixture. Example 3 illustrates one example using a mouse T-helper cell cDNA library to produce amplified fragments representing U3 regions expressed in T cells. The vectors were efficiently expressed as RNA and protein in PA317 helper cells, and were transmitted by retrovirus into human T-helper cells, where they were integrated and expressed as protein in the form of a β-galactosidase reporter gene, as visualized by X-gal staining. The products of this experiment are provided in Fig. 5 and as SEQ ID NOS: 2 and 3 from T-helper RNA. The products of another experiment are shown in Fig. 6 as SEQ ID NOS: 4-13 from mouse liver RNA (by RT-PCR).

Examination of the different U3 sequences isolated from T cells and from liver revealed several things. First, the T cell U3 sequences were related to each other, as were the liver sequences. However, the two types of U3 sequences were quite different between the two sources (T-cell, Figure 5 and liver, Figure 6). Specifically, the liver sequences (Figure 6) appeared to be a closely related group, differing mostly by single point mutations, some of which may affect transcription factor binding sites. Some of the polymorphic sites included: a phorbol ester response element (VLTRE); a Rel/NFκb binding region, and a possible glucocorticoid response element (GRE). Some of these polymorphisms are illustrated in Fig. 6. The T cell-derived sequences (Fig. 5, SEQ ID NO:2 and 3), on the other hand, differed significantly in length, with SEQ ID NO:3 missing more than 120 bases (compared with SEQ ID NO:2) including putative binding sites for retinoids (RAR/RXR) and several elements contained within the enhancer repeat region (including a cAMP response element (VLCRE, or CREB/jun binding site), and putative serum response element (SRE, CARG, and NF1/IL6). SEQ ID NO:3 represented one out of five clones sequenced, while SEQ ID NO:2 represented four out of five. Possible sites of interactions between transcription factors and DNA can be observed by comparing the experimentally derived U3 sequences with those in Hodgson et al. (Retro-Vectors for Human Gene Therapy, 1996 Fig. 4.2 supra). In addition
to the deleted sequences of SEQ ID NO:2, there are a number of single base differences within the conserved regions of the two T cell-derived sequences.

Advantageously, a number of new VL30 promoter sequences (SEQ ID NOS: 2-13, supra) were identified using these methods despite the fact that VL30 RNA comprises only about 0.3% of cell mRNA represented in a cDNA library. Moreover, in each case, the cloned insert was isolated without the need to use linkers, adapters, or multiple cloning sequences such as those that are typically use for other library construction methods. The promoter sequences can be used in the vectors disclosed here to express inserted foreign genes or the promoter sequences can be substituted into other retroviral vectors, such as MoMLV-derived vectors or other VL30-derived vectors. Further, vectors containing the promoter sequences can be propagated in retroviral helper cells, such as PA317 (U.S. Patent 4,861,719 to Miller) or introduced into cells by chemical or physical transfection.

In another application of the methods of this invention, libraries of amplified sequences can be incorporated into vectors using two or more fragments and using the restriction endonucleases cleaving at a distance from their recognition sites. Preferably the vectors are created using six or more fragments and preferably greater than 10 or more fragments. For example, as applied to VL30 promoter sequences, because there are over a hundred VL30 retro-elements in the mouse genome, it is possible to amplify all of the promoter sequences en masse, and propagate them en masse, enabling screening by serial passage through helper cells (such as the PA317 helper cell line) or by means of a replication competent retrovirus, as illustrated in Examples 3 and 4. Conversely, the promoter region may be broken down into several sub-domains and permutations of each could be combined and screened to enhance the chances of generating a superior construct (Fig. 4B).

As an example of breaking a promoter region down into several sub-domains, Fig. 7 illustrates a similarity plot of nucleotide sequences found in VL30 U3 regions. Plot similarity was performed using the Plot Similarity program (Wisconsin Sequence Analysis Package, release 8.1, Genetics Computer Group, Madison, WI). This program plots the running average of the similarity among the sequences in a multiple sequence alignment. The sequences compared were those found in Fig. 4.2 of Hodgson, 1996, chapter 4 (infra). That is, the plot disclosures the degree of conservation of VL30 promoter sequences among known VL30 promoters. From the figure, it can be seen that conserved sequences (close to 100% conserved) can be used as primer binding sites to amplify the adjacent sequences by PCR.
An allelic mixture of three fragment sets is then created to make a combinatorial library of promoters that can be positively selected, such as by using retroviral amplification of the active sequences. This, used in combination with the Fig. 4.2 (Hodgson, 1996, chapter 4 supra) can be used to determine regions of high similarity. Regions of high similarity within the U3 region can be replaced with one another. Therefore, a library of permutations of these sections can be made by combining allelic pools obtained by amplifying the sequences from individual subsections, followed by ligating the subsections in the correct order using the methods of the instant invention for gene self-assembly. For example, sub-section 1 can include the distal enhancer (from the LTR 5’-end to the site of insert primer 2, see for example the region defined by the insert primers 1 and 2 (SEQ ID NOS 55 and 56 of Example 4). In this way, using a plot similarity (such as Fig. 7), within each sub-section, the primers position fragments within a region of nearly 100% identity. Degenerate primers can also be used in these experiments to account for multiple nucleic acid base combinations along a particular sequence. In each case, the primers preferably are designed to have a melting temperature that is compatible with the RT-PCR conditions being used, and the conditions should be those recommended by the manufacturer (preferably Perkin Elmer Corp., Emeryville, CA). In Example 4, a set of primers is given that can be used to amplify different U3 subsections, together with directions for assembling a combinatorial library.

It will be appreciated by persons of ordinary skill in the art that the methods of the instant invention can thus be used to make allelic libraries of a variety of genes. For example, different allelic portions of a gene can be combined in a predetermined order and orientation to produce combinatorial libraries, without the need for fortuitous restriction sites separating the parts in the original construct, and without perturbing the important sequences joining the parts using the methods of this invention.

In this invention primers are constructed as described above. However, for the generation of allelic libraries or more complex library constructs it may be helpful to include 5’tags into the 5’ end of the primer. The purposes of the tag sequence are: 1) to provide extra nucleotides on both sides of the restriction endonuclease recognition sites (for more efficient digestion); and 2) to enable recovery of sequence tags or undigested fragments by means of an affinity reagent (such as silica, magnetic beads, or nitro-cellulose containing the complementary sequences) for purification. The use of an affinity reagent permits the digested ends to be purified away from the digested fragments. Furthermore, if any
undigested ends remain after thorough digestion, the affinity reagent will remove them, further aiding in the purification. In one embodiment, affinity purification of the digested fragments is used in place of gel isolation, eliminating possible damage caused by ultraviolet light as well as possible damage caused by dye (e.g., ethidium bromide) binding to the DNA.

It will also be appreciated that a number of other variations to the primer sequences can be employed. For example, as discussed above, the enzyme recognition site for an enzyme that digests outside of its recognition sequence is included in the primer, so that the DNA digest creates an overlapping end that is complementary to one other terminus to which it will be joined. The enzyme recognition site can be moved to any location within the primer so as to digest the DNA at the exact location desired. The primer can also be programmed with a novel enzyme recognition sequence to add any desired sequences between the two sequences to be joined or to incorporate a linker or adapter if desired. If the sequences to be amplified contain the enzyme recognition site of the primers, it may be necessary to switch to a different enzyme usage. The use of several different enzymes is possible and has been discussed above. As with other PCR procedures, after the initial primer selections have been made the primers are assessed for their ability to fold back on themselves or to create internal secondary structure. The primers are preferably modified to avoid palindromic sequences or the potential for self folding within a primer. Nucleic acid analytical software (such as the Wisconsin GCG package, Oxford Biomolecular, Oxford, UK) is available to perform this analysis and aid in the selection of alternative primers.

In addition, as with all PCR processes, it is necessary to determine the melting temperatures ($T_m$), and to adjust the annealing temperature of the PCR reactions to compensate for such temperatures. Finally, it is important to perform a sequence redundancy search, to determine whether the target sequence (the sequence complementary to the primer) is found more than once in the region to be amplified. If the sequence is repeated, it will be necessary to use a different primer in order to establish the single, correct priming site. Preferably, no more than 6-8 bases of incorrect target complementarity at the 3'-end of the complementary region is used and to allow a difference of at least 10° C between the $T_m$'s of the correct and the incorrect target. The annealing temperature should always be at least 5° C lower than the $T_m$ of the correct target and 5° C above the $T_m$ of the incorrect target. Again, the necessary software and instructions are readily available from the cited sources (Wisconsin Gene Computer Group and Oxford Biomolecular, *supra*).
Next, a vector is constructed to include the appropriate elements for expression in the desired cell type. For example, the plasmid of Fig. 3A can be used for the creation of a promoter library or a vector can be created using a commercially available vector and primers to create a three or more fragment annealing and ligation reaction as provided above. Preferably, the inclusion of a dominant negative selectable marker on the vector (e.g., the neomycin phosphotransferase gene, conferring G418 drug resistance) can be used to reduce the likelihood that cells without the vector are being maintained in culture.

Multiple allelic copies of DNA (cell derived or cDNA) can be amplified in separate reactions as a set of potential inserts with each set having its own unique overlap sequence following digestion with a restriction endonuclease, according to this invention. The fragments can then be ligated into an existing vector or in a single reaction of three or more fragments to form a combinatorial collection of potential alleles. For example, if six adjacent regions are amplified from five separate alleles, the number of combinations would be $5^6$, or 15,625 potential combinations. The combinations can then be grown en masse, and selected in vitro or in vivo. A variety of screening strategies can be used in this invention and those of ordinary skill in the art will appreciate that the type of screen will match the type of library being generation. Therefore, for the promoter library, introducing members of the library into particular cell types to assess for expression in one or more cell types versus the absence of expression in another cell type is evidence of tissue-specific or cell-specific expression. For screening purposes, the libraries of this invention function like other libraries created through other methods. A variety of screening methods for a variety of libraries have been described in the art. For example, selective screens are reviewed by Hodgson et al. (1996, RG Landes Company, supra). Reporter protein production is well known in the art as is dominant selectable marker (e.g. drug) selection and selection by fluorescence activated cell sorting, antibody affinity selection, phage display selection (such as commercially available from Amersham, Milwaukee, WI), and the like can be used without detracting from this invention.

In this way, it is possible to isolate multiple forms of genes, gene fragments or regulatory regions such as transcriptional promoters or packaging signals (for example, in a retro-vector system). The individual constructs may then be tested in vitro or in vivo to further characterize a particular phenotype.
In one example the method is used to create a library of complementarity determining regions (e.g., allelic variations that give rise to antibody diversity) of antibodies or from receptors, including T-cell receptors, epitopes, antigens, ligands and the like. For example, where a library of T-cell receptors is created, the introduction of a vector designed to create a functioning T-cell receptor can be introduced into T cells or T-cell progenitors and the cells can be tested for their ability to bind to a particular test ligand. The ligand-recognizing cells can then be isolated from the ligand and grown in the presence of cytokines to produce specialized T cell clones. Where a library of antibodies or antibody fragments is created, the antigen reactive portions, for example, can be recombined in a vector containing the remaining portions of an antibody molecule to generate antibodies or antibody fragments in a cell. In other examples, the methods of this invention can be used to create allelic domains of receptor families (such as the steroid receptor super-family); libraries with related regions from peptide hormones; cytochromes P450; or other protein families that have shared domains or sub-sections with similar structures. The methods of the instant invention allow the joining of allelic sub-sections in an ordered fashion. In each case, it will be necessary to design primers, and to keep track of the uniqueness of joining overlaps and the presence of internal restriction sites as described above. While these will be different in each case, here are listed some general guidelines that are incorporated into the method of the instant invention.

As discussed above, although described as it relates to promoter libraries, libraries of other nucleic acid sequences can be created using the methods of this invention. These libraries include, introns and/or exons and/or functional domains libraries, libraries of potential alleles for a particular gene sequence, and the like. These sequences can be amplified from cell DNA or RNA using the primers of this invention and incorporated into a variety of vectors. For example, one vector of this invention, VLBPGN, has a portion of LTR removed and can be used to create a variety of libraries following digestion with Bpm1.

Selected or screened products of the combinatorial library can be used for gene expression, such as the promoters of Figs. 5 and 6. In addition, the exploitation of these sequences for the expression of a variety of genes, the LTR fragment containing the promoter can be joined to one or more functional retroviral packaging signals, internal ribosome entry sites, additional promoters, coding regions, processing sites, and the like.
Advantageously, there are almost no spatial constraints upon the joining of molecules by the method of the instant invention and other methods have not taken advantage of the combination of PCR to isolate genes or gene fragments; enzymes cleaving at a site distant from their restriction endonuclease recognition site to combine three or more fragments with precision; and, the use of unique overlapping non-palindromic termini to ensure fidelity of multi-fragment ligations. This combination permits the artisan to prepare complex gene constructions in one ligation step and does not require sequential sub-cloning into a vector or propagation in a prokaryotic host. Added to this the combination by these methods of fragment pools facilitates recombinatorial genetics.

The ability to recombine (in the correct order and direction) and screen a large number of allelic variants (whether as a simple library or as a combinatorial library), resulting in increased abundance (by amplification in the RNA, and subsequently in the DNA) is a special characteristic of this invention. Particular advantages of this system are obtained when the methods of this invention are combined with retrovirus vector technology or other virus vector technology. For example, the combination provides a form of *in vitro* evolution whereby the passage of the library through virus and through cells selects functioning sequences and increases the abundance of the surviving RNA and DNA molecules.

For example, consider the consequences of screening several different promoters expressing RNA in a donor cell (*i.e.*, a cell producing virus particles), but at differing levels of RNA abundance. In the following example, the least abundant RNA species is expressed at 0.1 copy of RNA per cell, while six others are expressed at 1 copy, 10 copies, 100 copies, 1,000 copies, or 10,000 copies, or 100,000 copies/cell, respectively. After a single passage, the DNA copy number in the recipient cells now reflects the approximate RNA copy number in the donor cells. These numbers are further amplified in the relative abundance of RNA species produced in the recipient cells. Disallowing for factors such as position effects, transcription factor depletion, etc., (which may be considerable), the same relative ratios of expression would be expected. Taking into consideration position effects, the disparity between abundance caused by changing insertion loci should average out. The most abundant RNA species after two passages is then many orders of magnitude more abundant than the least abundant.
<table>
<thead>
<tr>
<th>Species</th>
<th>RNA abundance</th>
<th>DNA copy RNA</th>
<th>DNA copy RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P=0</td>
<td>P=1</td>
<td>P=2</td>
</tr>
<tr>
<td>A</td>
<td>0.1 copy/cell</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>D</td>
<td>100</td>
<td>100</td>
<td>10,000</td>
</tr>
<tr>
<td>E</td>
<td>1,000</td>
<td>1,000</td>
<td>10^6</td>
</tr>
<tr>
<td>F</td>
<td>10,000</td>
<td>10,000</td>
<td>10^8</td>
</tr>
<tr>
<td>G</td>
<td>100,000</td>
<td>100,000</td>
<td>10^10</td>
</tr>
</tbody>
</table>

**Table 2.** Enhancement of DNA and RNA copy number as a result of different RNA expression levels, after retroviral passage. P= (no. of passages). Numbers are interpreted as relative ratios within a column.

The present invention is able to efficiently create a library of RNA or DNA sequences whether or not they are in low abundance. The kinetics of screening for RNA abundance of a promoter can be appreciated best in the following discussion. For the purposes of this discussion, position effects have been ignored. An equation describing the kinetics of screening for RNA abundance is:

\[
(1) \quad R_{rel\chi} = A\chi / \sum A_{n=0}
\]

The above equation (1) can be stated in plain English: The relative abundance of an RNA species \(\chi\) ([R\(_{rel\chi}\)] within a population of RNA molecules expressed in a single cell or within a population of cells) is equal to the RNA copy number of RNA species \(\chi\) (\(A\chi\)) divided by the sum of the RNA copies of all RNA species present, including \(\chi\).

The relative abundance number of any given species changes as the number of passages change, according to the following approximation:

\[
(2) \quad R_{xpy} = D_{xP0} R^{P+1}
\]

In the simplest of terms, equation two (2) can be expressed as: The abundance of RNA species \(\chi\) after \(Y\) passages (\(R_{xpy}\)) is equal to the initial abundance of the DNA for species \(\chi\) at passage=0 (\(D_{xP0}\)), multiplied by the RNA abundance/DNA copy, raised to the power of the number of passages plus one. Thus, a typical RNA species that starts out as a
single copy of DNA, after zero passages (i.e., in the donor cell) expresses 10 copies of RNA/cell. After one passage it is amplified at the DNA level to a relative ten copies (the same as the RNA abundance at P=0), and at the RNA level to 100 copies (10 copies per DNA copy). The reason for the amplification is that viral packaging and passage is based upon the number of RNA copies present in the donor cell. These calculations can be used to arrive at approximate abundance determinations for any given passage. The actual results of any given experiment, of course, will be biological rather than physical or mathematical. This means that other variables such as RNA efficiency of transmission and longevity, availability of transcription factors, experimental variation, etc. also come into play. The underlying purpose of the approximating equations, however, is to illustrate that RNA is amplified in DNA in proportion to the abundance of the template (RNA) within the cell.

The abundance of mRNA in cells can vary continuously from less than a copy per cell to nearly 100,000 copies/cell in actively transcribing, highly-specialized cells such as reticulocytes, the chicken oviduct, the silk moth silk gland, etc. Therefore, the spectrum of RNA abundance from 0-10^4/cell is within the biological window of interest. For most practical purposes, such as biotechnological expression of genes in specific cells, only the higher end of this abundance range is desired. Therefore, using a viral selection system, as disclosed in this invention, it may be possible to disregard those species with less than a threshold level, such as <0.1 copies per cell. The selection through virus will lead to the recovery of the more abundant species. Furthermore, because the vector is likely to be the only considered sequence, it may be considered as a proportion of the whole of RNAs expressed in the target cell. The situation is more complex when a large number of permutations and combinations is generated, for example by self-assembling thousands or millions of fragments in a predetermined order using the self-assembly technique of the instant invention. Consider the assembly of allelic variants of four promoter subregions: distal enhancer, proximal enhancer, distal promoter and proximal promoter. If 100 varieties of each of the four groups were amplified and combined using the instant process along with a single vector, 10^8 resultant combinations could occur. However, a sufficient number of molecules to start out a combinatorial screening program might be a million. The problem can be simplified by considering these in groups as follows:
Table 3. Grouped abundance of RNA molecules derived from combinations.

<table>
<thead>
<tr>
<th>No. of species in group:</th>
<th>RNA abundance: molec. at P=0:</th>
<th>Total No. RNA at P=0:</th>
<th>RNA at P=1</th>
<th>RNA at P=2</th>
<th>RNA at P=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 X 10^5</td>
<td>1</td>
<td>9 X 10^5</td>
<td>9 X 10^5</td>
<td>9 X 10^5</td>
<td>9 X 10^5</td>
</tr>
<tr>
<td>2 X 10^5</td>
<td>10</td>
<td>2 X 10^6</td>
<td>2 X 10^7</td>
<td>2 X 10^8</td>
<td>2 X 10^9</td>
</tr>
<tr>
<td>2 X 10^4</td>
<td>1,00</td>
<td>2 X 10^6</td>
<td>2 X 10^8</td>
<td>2 X 10^10</td>
<td>2 X 10^12</td>
</tr>
<tr>
<td>1 X 10^3</td>
<td>1000</td>
<td>1 X 10^6</td>
<td>1 X 10^9</td>
<td>2 X 10^12</td>
<td>2 X 10^15</td>
</tr>
<tr>
<td>1 X 10^1</td>
<td>10,000</td>
<td>1 X 10^6</td>
<td>1 X 10^8</td>
<td>1 X 10^13</td>
<td>1 X 10^17</td>
</tr>
<tr>
<td>1</td>
<td>100,000</td>
<td>1 X 10^6</td>
<td>1 X 10^10</td>
<td>1 X 10^15</td>
<td>1 X 10^20</td>
</tr>
<tr>
<td><strong>Sum Total:</strong></td>
<td><strong>6.6 X 10^6</strong></td>
<td><strong>1.11 X 10^10</strong></td>
<td><strong>1.01 X 10^15</strong></td>
<td><strong>1 X 10^20</strong></td>
<td></td>
</tr>
</tbody>
</table>

Thus, it follows that in the example population (Table 3) of over a million constructs (equally represented in the DNA), a single construct expressing 10^5 copies of RNA per DNA copy will increase to approximately 99% of the total expressed RNA sequences in two passages. Using similar procedures in combination with drug and/or hormonal stimulation, and after consideration of the possible transcription factor binding sites within the sequence family (Figs. 5 & 6), it is within the intended scope of the invention to select for hormonal or pharmacological controls of transcription such as have been described herein. The factors contributing to the outcome are not only the input constructs, but recombinants and mutants as well. These secondary contributors to molecular diversity will be enhanced if multiple rounds of infections are allowed to occur, as oftentimes the difference between a particular transcription factor being able to bind (or not) may depend upon a single base change. Because viral infection is progressive and competitive, molecular evolution can be used to generate gene constructs de novo in the tissue culture dish in short time periods.

Advantageously, the use primers to generate amplified fragments with uniquely complementary cohesive ends (i.e., that the ends will preferably only hybridize with the intended 5’ and 3’ fragments) to ligate three or more fragments as taught in this invention improves the potential for obtaining a diverse library.

Although the examples particularly point out a transcriptional promoter as the product of the process, the skilled artisan can appreciate that a particular selection technique can be applied to other cis- and trans-acting genetic sequences as well. Although a virus is used to propagate the selective advantage of a preferred embodiment, it can also be appreciated that any selective screen, such as drug selection, cell survival, phenotypic selection, cell sorting, antibody selection, and the like (see Ausuble et al., supra) could be
substituted without changing the intended scope of the invention. Likewise, transfection or cell fusion could be used in place of viral infection. Furthermore, substitution of different viruses, retrotransposons, or functional groups are likewise within the intended scope of the invention. The described embodiments are to be considered only as illustrative and not restrictive, and the scope of the invention is indicated by the claims rather than by the narrative description. All references and publications, cited herein, are incorporated by reference into this disclosure.

Like the embodiments detailed above, the method of library production is also conducive to assembly and transfer of genetic material directly into eukaryotic cells, saving the step of propagation in bacteria that is standard in bacteria. An advantage of direct transfer of the libraries of this invention to eukaryotic cells, including the exemplary retroviral vector producer cells, is that certain essential cis-acting structural features will be under positive selection (i.e., if they are not present, the molecule will be lost due to its non-functionality). As discussed above, it is often advantageous to eliminate bacterial and plasmid DNA sequences, endotoxin, and other bacterial contaminants by introducing the constructs directly into eukaryotic cells.

In addition to providing a method for constructing complex DNA molecules efficiently (as in the examples of three piece and six piece constructs), the methods of this invention permit the assembly of constructs that are larger than those conventionally propagated in E. coli. Examples of these types of vectors include adenovirus vectors, herpes simplex vectors and artificial minichromosomes. In order to insert genes into such vectors that are too large for conventional molecular cloning procedures, in the past it was often necessary to resort to in vivo recombination, wherein the genes of interest are cloned into a suitable vector and the flanking homologous regions are used to target the foreign genes to a homologous site within the larger viral or minichromosome vector. However, the methods of this invention permit PCR fragments of any size (up to the limits of PCR capability, 20-30 kb per fragment) to be joined together. Thus, it is feasible to precisely construct adenovirus vectors by amplifying larger sequences, and combining them by ligation. For example, several sections of adenovirus (5-10 kb each) can be ligated using the methods of this invention, up to for example, about 37 kb, and then transformed directly into human cells. Only the correctly recombined vectors are capable of replicating. Hence, the DNA is autoselecting. A similar procedure is used for generating herpes virus vectors, which are
approximately 150 kb. The precision of the methods of this invention permit non-essential viral genes to be more easily eliminated from the construct. After transfection into appropriate cells, the DNA replicates and virus particles are formed.

Some special considerations apply to larger vectors, however. First, it is desirable to use enzymes that do not cut within the large DNA fragments. To prevent excessive fragmentation of the DNA by internal sites, it is desirable to use enzymes that cut rarely or infrequently, such as CpG-containing enzymes recognizing six bases, or enzymes such as Sap1, recognizing seven bases and digesting a three bp overhang (thus permitting up to 32 fragments to be joined in order). It is also desirable to avoid shearing the DNA once large segments have been joined by ligation. One method of avoiding shear is to add the transfection agent, such as Superfect™ reagent (dendrimers, Qiagen) or Lipofectamine™ (liposomes, Life Technologies, Gaithersburg, MD) directly to the ligation reaction, and then add the cells to be transfected to the mixture. This, or a similar method avoids the need to physically move the ligated DNA, and thus prevents shearing. Another method is to add a DNA condensing reagent (dendrimers, polycations [such as polyethyleneamine] histones or liposomes) directly to the DNA ligation reaction, and then move the DNA by pipette after it has condensed (thus reducing shearing of the DNA). Once inside the cell, viral DNA can replicate (as in the examples of partially replication-competent adenovirus and herpes simplex virus vectors).

Artificial minichromosomes have been under development for years. True artificial chromosomes require a centromere, at least one origin of DNA replication, and in the case of linear molecules, telomeric repeats at the chromosomal termini. In addition, to be very effective it is desirable to have a selectable marker gene, one or more therapeutic genes, and/or reporter genes.

In reality, the use of minichromosomes has been delayed by the inability to effectively manipulate the larger DNA molecules in vitro. Yeast and bacterial artificial chromosomes have been used with little success in mammals, and the addition of telomeres to the ends of linear chromosomes is also a special problem, as there is no prokaryotic host that can tolerate large linear DNA. The methods of this invention offers the opportunity to assemble human or mammalian minichromosomes in vitro, by using large segments (10-30 kb) of synthetic, gene-amplified DNA as ligation starting materials. For example, up to 32 Sap1 fragments (up to 30 kb each, containing the essential cis- and trans-acting sequences),
or 512 shorter HgaI fragments can be combined using these methods. As with the other examples, several enzymes suitable for this invention (e.g., such as class IIIS enzymes) can be combined (possibly with different termini lengths) to simplify the task. The methods of this invention also facilitate construction of telomeric repeats, because the constructs of this invention do not need to be circular. Thus, the methods of this invention can be used to make telomeres of any length, by adding additional segments onto the ends of molecules. One way to do this is using self-assembling genes that employ a repeating overhang sequence (self-complementatory molecule, such as AG-3’ at one end, and CT-3’ at the other end), permitting the telomeres to be lengthened to the extent desired by adding the required molar excess of the telomeric repeat-containing fragment. This technique gives the investigator some control over the relative length of the telomeres, although the self-complementarity indicates that many repeats will be lost due to self-ligation. This can be alleviated by using higher starting concentrations of DNA to favor inter-molecular ligations over intra-molecular ligations (e.g., >20 μg/ml starting concentration of DNA).

A two fold molar excess of telomeric fragments gives approximately twice the average length of telomere as a strictly 1:1 molar ratio of all fragments. By using a higher molar ratio of shorter telomeric repeats it is possible to give greater uniformity to the overall length of the molecules, which will vary from one terminus to the other. Thus, in addition to providing a way to build large molecules with precision, the methods of this invention provides for a way to control the telomere length (or potential life-span) of the artificial chromosome. To prevent damage during handling, the minichromosome DNA can be condensed with polyacations, adenovirus particles, dendrimers, histones, or liposomes prior to transfection, similar to larger viral vectors.

The methods of this invention can be used to create recombinant virus. One example of this is an adenovirus vector self-assembling gene system. This system can include three parts: 1) vector; 2) helper virus; and 3) helper cells. The vector part is a self-assembling fragment set of at least three fragments comprising the essential cis-acting sequences (left and right inverted terminal repeats, which are the 103 bp at both ends of the genome that are required for replication [ITRs] and packaging sequences [Y, base pairs 194-358) and central 'baggage' area, comprising one or more self-assembling fragments including therapeutic genes, marker genes, and reporter genes. The baggage area is thus flanked by the cis-acting sequences in the vector. Because the synthetic oligonucleotide sequences
comprising the 5' and 3' termini of the helper virus are not phosphorylated, they will not ligate together creating multimers. Thus, the Ad5 vector region will assemble only into monomers. The helper virus part comprises all Ad5 trans-acting genes except for the E1A and E1B genes. The helper virus part has no cis-acting sequences, and it is amplified in several sections. In this preferred embodiment, the virus is amplified using primers that exclude the ITRs, packaging region and E1A&B genes. The helper virus is digested by Sap1 digestion, creating seven uniquely terminated fragments comprising the trans-acting viral genome, with dephosphorylated, blunt 5' and 3' ends on the terminating fragments. The primers are designed so as to amplify the internal virus sequences without changing them, except for the 5' and 3' ends of the virus. The PCR-amplified fragments are digested with Sap1 and are religated in their natural order after gel isolation and Qiagen column purification. The 5' end of the helper virus genome starts at 3.2 kb (in the E1A gene) so as not to overlap the vector sequences, which could otherwise cause replication competent adenovirus (RCA). Because the 5' and 3' ends of the helper virus do not contain Sap1 sites, they remain intact after digestion with Sap1. Because the synthetic oligonucleotide sequences comprising the 5' and 3' termini of the helper virus are not phosphorylated, they will not ligate. Thus, the Ad5 helper virus genome assembles only into preferred monomers during ligation.

In a preferred embodiment, non-essential genes are deleted from the Ad5 genome by means of the method of self-assembling genes. In another preferred embodiment, the helper virus genome is approximately 30 kb after deletion of E1A, E1B and E3 gene sequences from the helper virus, and it is amplified as a single long fragment using the eLONGase Amplification System (Life Technologies or a similar strategy for creating long PCR fragments with high fidelity). It is not of great importance that occasional PCR errors may occur, because multiple copies of the Ad5 helper virus are transfected into target cells, thus providing trans-complementation. The helper cells are preferably 293 cells, a human kidney cell line expressing E1A and E1B genes (ATCC). The vector part and the helper virus part are combined in equimolar ratios after ligation has been performed separately on each fragment set. The Superfect protocol (Qiagen) is used to transfec the vector part and the helper part into the helper cells. The helper cells lyse, releasing high-titer adenovirus particles that are capable of infecting a variety of human cells. The resulting defective virus is incapable of forming RCA, and it transmits up to 34 kb of foreign genes in the baggage area. Unlike conventional Ad5 vectors that require separate constructs for E. coli propagation of
insert genes, and recombination in vivo, the present vectors are relatively easy to make and provide a precise, safe alternative to first generation and second generation adenovirus vectors.

Exemplary methods for producing self-assembling vectors and genes are provided below. Further, the Examples provide methods for producing libraries of nucleic acid sequences using the methods of this invention. A number of nucleic acid sequences identified using the methods of this invention are described. The examples provided below are exemplary and not limiting. All references and publications provided herein are incorporated by reference into this disclosure.

Example 1
Three-Piece Gene Self-Assembly with 100% efficiency

Using 6 primers (SEQ ID NOS:24 and 63-67), three PCR fragments were amplified from templates VLMG (SEQ ID NO:22) and VLBPGN (SEQ ID NO: 1). PCR reactions were carried out using the hot start technique, according to the manufacturer’s instructions (Perkin Elmer) using Pfu DNA polymerase (Stratagene). To amplify specific portions of the above templates, each primer contained a class IIS enzyme site capable of digesting a unique overhanging end that was complementary to only one other terminus in the subsequent ligation. The class IIS enzymes used were BpmI and Eco 57I (the latter was used to copy a fragment that contained an internal BpmI site). The reactions were carried out as follows: 1) the lower reaction was assembled according to the protocol for PCR Gems (Perkin Elmer); 2) the lower reaction was heated to 80°C, 5 min, then cooled to 4°C for 5 min; 3) the upper reaction was prepared according to PCR Gems protocol and was added to the lower reaction (separated by cooled wax). The primer concentration was 0.3 µM (final). The dNTP concentration was 200µM (final). 5 Units of Pfu polymerase was used. All fragments were amplified using the following conditions: 96°C, 45 sec; (then followed by 30 cycles of the following) 96°C 45 sec, 52°C 45 sec, 72°C, 6 min; then followed by a single incubation at 72°C for 10 min; then hold at 4°C. All fragments were successfully amplified. The PCR fragments were purified using the Qiaquick PCR purification protocol (Qiagen). The fragments were digested with an excess of the appropriate restriction enzyme (BpmI or Eco57I). The digested fragments were run on a 1% agarose gel and were excised using minimal irradiation from a hand-held 365 nm ultraviolet light. The fragments were purified
using the Qiagen Qiaquick Gel Purification Protocol. The fragments were ligated at an equimolar ratio at a concentration of >20 μg/ml with T4 DNA ligase (Boehringer Mannheim) overnight at 4°C. Competent *E. coli* SCS110 cells (Stratagene) were transformed with the ligated DNA. Eight colonies were characterized by restriction enzyme analysis, and all eight contained the correct order and orientation of the three fragments. The experiment was repeated independently by another investigator, and the same result was obtained (8/8=100%). Thus, the procedure resulted in a high percentage of correctly assembled vectors.

This three-piece vector was VLΔBP. The deletion extended from the distal enhancer region to the TATA box near the start of transcription. The deletion region was a pair of *Bpm1* sites that permitted U3 sequences to be cloned into the insert.

One validated *E. coli* clone of VLΔBP was transfected into retroviral helper cells. After 48 h, the vector was transduced into amphotropic helper cells. After selection for two weeks with the drug G418, drug resistant colonies were grown up in a mass culture and the vector was transduced from the amphotropic helper cells into a human HT1080 cell line (ATCC, Rockville, MD). Surprisingly, even with a large deletion in the LTR promoter, the basal TATA box-containing VLΔBP was transmitted as a retrovector and was permanently inserted into the human cell line, thus establishing the validity of the self-assembly technique for the construction of functional eukaryotic vectors.

**Example 2**

Production of a Six Piece Self-Assembling Expression Vector

Due to the high efficiency of the gene self assembly process for the three piece assembly, a complex vector containing six fragments was constructed. The results here were extended to determine whether such a self-assembled vector would also have biological activity in human cells without being cloned and grown in a prokaryotic cell.

Six fragments were individually constructed by PCR using three different templates and twelve primers (as illustrated in Fig.8). The primers used three different class IIS enzymes. The enzymes were chosen so as to give 2 base pair, 3'-overhanging ends. Three enzymes were used in order to avoid the use of enzymes that had additional sites internal to the fragments being amplified. Thus, *Bpm1* was used unless there was an internal *Bpm1* site. If such a site existed, *Eco571* was used. If there was also an internal *Eco571* site, then *BsrD1*
was used. However, it is alternatively possible to use an enzyme such as Eam11041, where
the Eam11041 sites in the primers are unmethylated (therefore susceptible to digestion by the
enzyme), and wherein the \textsuperscript{3}dCTP analog of dCTP is used in the PCR reaction, methylating
all internal sites (and protecting them from digestion by Eam11041), as suggested by Padgett

Using 12 primers, 6 fragments were amplified from 3 templates: pBK-CMV
(SEQ ID NO:26) , pVLMB (SEQ ID NO:23) and pVLOVhGH-900 (SEQ ID NO:21).
Fragment 1 was amplified from pBK-CMV using primers 1 and 2 (SEQ ID NOS:31 and 32).
Fragment 2 was amplified from pVLMB using primers 3 and 4 (SEQ ID NOS:33 and 34).
Fragment 3 was amplified from pVLOVhGH-900 using primers 5 and 6 (SEQ ID NOS:35
and 36). Fragment 4 was amplified from pVLMB using primers 7 and 8 (SEQ ID NOS:37
and 38). Fragment 5 was amplified from pVLMB using primers 9 and 10 (SEQ ID NOS:39
and 40). Fragment 6 was amplified from pVLMB using primers 11 and 12 (SEQ ID NOS:41
and 42). PCR reactions were carried out using the hot start technique, according to the
manufacturer's instructions (Perkin Elmer AmpliWax PCR GEMS 100). The lower reaction
was heated to 80 °C for 5 min, then cooled to 20 °C for 5 min. The upper reaction was
prepared according to PCR gems protocol and was added to the lower reaction (separated by
cooled wax). The primer concentration was 0.3 micromolar (final). The dNTP concentration
was 200 μM (final). 5 U of \textit{Pfu} polymerase (Stratagene) was used per reaction. 100 ng of
template was used for each reaction 14 rounds of PCR amplification were used to reduce
mutagenesis of the templates. The PCR cycling protocol was 96 °C 45 sec; then two cycles
of (96°C 45 sec, 52°C 45 sec, 72°C 6 min); then 12 cycles of (96°C 45 sec, 58°C 45 sec,
72°C 6 min) followed by a 72°C soak for 10 min, then to 4°C hold.

The six PCR fragments were designed to self-assemble into a retro-vector after
digestion with the correct class IIS restriction enzyme (Fig. 8). After transfection into
retroviral helper cells, the vector DNA is transcribed as RNA by means of the
cytomegalovirus immediate early promoter (fragment 1). This promoter replaces the
retroviral or VL30 LTR in this vector. The RNA transcript region begins with the R and U5
regions of the Moloney murine leukemia virus (MoMLV) LTR, the viral packaging signals
(Ψ) region of MoMLV, the packaging enhancer (Ψ+) region of mouse VL30 and the IRES
region of EMCV fragment 2. Fragment 3 consists of the human growth hormone (hGH)
cDNA sequence. Fragment 4 consists of the SV40 virus early region promoter driving
expression of the neomycin phosphotransferase (neo) gene. Fragment five consists of the (+)-strand primer binding site of the MoMLV LTR, the U3 region of the MoMLV LTR, the repeat (or R) region, and a portion of the U5 region. Fragment 6 consists of the PBR322 plasmid origin of replication.

5
Fragment 1: CMV early region promoter
  Template: pBK-CMV plasmid DNA (Stratagene, LaJolla, CA) *Bpm* 1 (SEQ ID NO:26)

  PCR primer 1 (SEQ ID NO:31)

  GACTAACCCTGATTCCA**CTGGAG**CCGTATTACC GCCATAGTTATTAATAG

  PCR primer 2 (SEQ ID NO:32)

  GACTAACCCTGATTCCA**CTGGAG**TAATTCGCGGTAGCGGATCTGACG

10

Fragment 2: R-U5-Psi-Psi(+) IRES *Bpm* 1

  Template: pVLMB plasmid DNA (SEQ ID NO:23)

  PCR primer 3: SEQ ID NO:33

  GACTAACCCTGATTCCA**CTGGAG**ACACTTGACCTCTACCGCCAGTCCCTCCGAT

  TGACTGAGTCG

  PCR primer 4: SEQ ID NO:34

15

GACTAACCCTGATTCCA**CTGGAG**GATCCCGCGCCCCATGATTATTTACT

Fragment 3: human growth hormone (hGH) *Bsr* DI

  Template: pVLCNOVhGH plasmid DNA (SEQ ID NO:21)

  PCR primer 5: SEQ ID NO:35

  GACTAACCCTGATTCCA**GCAATG**TCGGTTAGTTTTTTTACTGGTTTGTC

20

  PCR primer 6: SEQ ID NO:36

GACTAACCCTGATTCCA**GCAATG**TTAGGACAAGGGCTGTTGGGCACCTGG

25

Fragment 4: SV40 early promoter-neomycin phosphotransferase

  Template: VLMB plasmid (SEQ ID NO:23)

30

  PCR primer 7: SEQ ID NO:37

GACTAACCCTGATTCCA**CTGGAG**GGTCGACCCGTTGGAAATGTGTTGTACAG
PCR primer 8: SEQ ID NO:38
GACTAACCTTGGATTCCACTGGAGAAATCTCGTGATGGCAGGGTTGGGCGT

Fragment 5: MLV(+)PBS-U3-R-U5
Template: VLMB plasmid (SEQ ID NO:23)
PCR primer 9: SEQ ID NO:39
GACTAACCTTGGATTCCACTGAAGAGATTATATTATGTCTCCAGAAAAAGGGGGG
PCR primer 10: SEQ ID NO:40
GACTAACCTTGGATTCCACTGAAGCCCCCAAATGAAAGACCACCCCGCTGACG

Fragment 6: PBR322 origin of replication
Template: VLMB plasmid (SEQ ID NO:23)
PCR primer 11: SEQ ID NO:41
GACTAACCTTGGATTCCACTGGAGCGGGGACGGAATTCTGATCTGCTG
PCR primer 12: SEQ ID NO:42
GACTAACCTTGGATTCCACTGGAGTTCTCGAGCAGCGGCATCTCGGCC

Procedure: The twelve primers were prepared by the following procedure: 1) oligonucleotides were synthesized with trityls off. After deprotection and lyophilization, the samples were resuspended in 5 microliters deionized formamide and loaded onto a polyacrylamide gel (12% polyacrylamide, 250V). The samples were excised under short wave UV irradiation and eluted overnight in 600 microliters of sample elution buffer (0.5 M ammonium acetate, 10 mM Mg acetate, 1 mM EDTA, 0.1% SDS). The contents were loaded onto a BioRad Chromatography column (Cat. # 732-6008) and centrifuged into an Eppendorf tube at low speed (2000 RPM, 5 min). After washing the column with 500 microliters TE buffer (10 mM Tris, 1 mM EDTA), pH 8.0 and recentrifugation (2000 RPM, 5 min), the pooled eluate was ethanol precipitated, washed with 100% ethanol, resuspended in TE buffer and quantitated by spectrophotometry of a small sample, which was then discarded.

Fragments were cleaned using the Qiaquick PCR cleanup procedure. The fragments were digested with their respective class IIS restriction enzyme. The digested fragments were run on 1% agarose gels, and the fragments were excised and cleaned using the Qiaquick gel cleanup procedure. Fragments were combined in an equimolar mixture and
ligated overnight at 4°C with T4 ligase and ATP. An analytical gel was run with the ligated DNA, as well as with controls including unligated fragments and ligated fragments with a single fragment missing. As opposed to the controls, the complete ligation included bands equivalent to the full-length supercoiled monomer (refer to as GENSA 981, SEQ ID NO:29), as well as bands possibly representing multimers (up to six bands were observed).

In order to assess the efficiency of the method, eleven nanograms of DNA were transfected into SCS1 supercompetent cells. Thirteen kanamycin resistant colonies were harvested, and plasmid DNA preps indicated 10 out of thirteen that appeared to be the correct length. All ten gave the expected bands when digested with PstI, SnaB1, and BamHI. 1.35 μg of the ligated DNA was purified by phenol-chloroform-isoamyl alcohol extraction, followed by two extractions with chloroform-isoamyl alcohol, and was precipitated in ethanol. The DNA was washed in 70% ethanol and re-suspended in 50 μl of sterile phosphate buffered saline (for transfection). The DNA was transfected (using the Qiagen Superfect protocol) into HTam1 (amphotropic human helper cells). 24 h after transfection, the target cells were washed and fresh culture media was added. 48 h after transfection, the supernatant from the vector producer cells was filtered (0.45 μm, Nalgene) and transferred to PG13 helper cells (ATCC) and HT1080 human fibrosarcoma cells. This procedure was repeated after 72 h. 48 h after transduction, recipient cells were started on G418 drug selection (500 μg/ml). The appearance of G418 drug-resistant colonies on transduced PG13 and HT 1080 cells after 6 days of selection indicated successful transmission via retrovirus particles. The transfect HTam cells were also selected with G418. After six days of drug treatment, 45 colonies of resistant cells were counted. Thus, the six fragment gene assembly was effectively transmitted and expressed as either a DNA (transfection) vector or a retro-vector.

Example 3
Design and Construction of Single LTR Vectors

Background: In order to manipulate the interior of the VL30 LTR sequences using a promoter rescue technique, single LTR vectors were constructed. The mouse VL30 element NVL-3 was used as the starting material as it is constitutively and abundantly expressed in most mouse tissues. Single LTR vectors are circular and behave as if they contained two LTRs. Thus, in these vectors RNA transcription begins at the start of the R region (see Fig.
3B), and continues through the polyadenylation site after completing the second round of transcription of the R sequences (Fig. 3A). In previous studies, these vectors were expressed transiently in vector producer cells and the DNA did not integrate into cell DNA as a standard two LTR vector. Therefore, the vectors were usually passed to a second complementation helper cell line via retroviral transduction of the vector RNA transcribed in the first helper cell. This process resulted in the vector regenerating a correct (two LTR) structure upon integration into the recipient cell DNA.

*Experimental method:* The plasmid pNVL-3 (SEQ ID NO:25, kindly provided by Dr. J. Norton, Manchester, UK), containing a complete copy of the NVL-3 (mouse VL30) genome (Adams et al., 1989), was digested with *Xho*I (which cuts in the LTRs), releasing the 4.27 kb VL30 genome with one copy of the LTR. This fragment was circularized using T4 DNA ligase and ATP. The circular DNA was linearized by digestion with SnaBlI, 187 bp from the 3'-LTR. A 2.3 kb fragment containing the SV40 virus early region promoter and the aminoglycoside phosphotransferase (*neo*) gene, together with the PBR322 plasmid origin of replication, was excised from the BAG retrovirus vector (Price et al., *Proc. Natl. Acad. Sci.* 84:156-160, 1987, kindly provided by C. Cepko, Cambridge, MA). BAG is also obtainable in a retrovirus helper cell line from American Type Culture Collection (ATCC), Rockville, MD by digestion with *Xho*I and *Bam*HI. This fragment was blunt ended with T4 DNA polymerase and dephosphorylated with calf intestinal alkaline phosphatase (CIP). The fragment was then ligated to the linearized SnaBlI fragment of NVL-3. The resulting plasmid (containing a circularly permuted NVL-3 genome with the SV-*neo-ori* region) was designated VLSNO2 (SEQ ID NO:30).

In order to facilitate the switching of LTR sequences by means of the class IIIS enzyme *Bpm*1, VLSNO2 was digested with *Bpm*1 (six sites). The region containing four *Bpm*1 sites was removed and replaced with a 19 bp linker (SEQ ID NOS: 1 and 52, see below), 921 bp beyond the LTR. The linker contained *Sna* BI, *Cla*I and *Bam* HI cloning sites.

- Linker (top strand): 5'-'TACGTATCGATGGATCCGA-3' (SEQ ID NO:51)
- Linker (bottom strand): 5'-'GGATCCATCGATACGTAAG-3' (SEQ ID NO:52)
The remaining two of the Bpm1 sites had complementary ends, which permitted their ligation and resulted in eradication of all Bpm1 sites within the resulting vector VLSNO3 (SEQ ID NO:20).

In order to facilitate reporter/therapeutic gene function, a 3.7 kb fragment containing the internal ribosome entry site (IRES) from encephalomyocarditis virus, together with the β-galactosidase reporter gene, was excised from the plasmid pVLSAIBAG (kindly provided by Mr. James Grunkemeyer, Omaha, NE) by means of a partial digestion of the plasmid with Bam HI. This region was inserted into the Bam HI site of VLSNO3, resulting in the vector VLSNOSIB (SEQ ID NO:14).

A second reporter construct, pVLSNOG (5774 bp, SEQ ID NO:19) contained the green fluorescent protein (GFP, Clontech, Palo Alto, CA) gene was constructed by inserting a Bgl2-Bcl1 fragment (800 bp) from plasmid pGFP-N1. This sequence, containing the GFP gene, was treated with mung bean exonuclease and inserted into the unique Sna B1 site of pVLSNO3.

In order to enhance GFP fluorescence from the reporter plasmid pVLSNOG, the serine-65 codon in the GFP gene was mutated into threonine by a site-directed mutagenesis procedure with the Transformer™ Site-Directed Mutagenesis kit from Clontech. A Bpm1 site in the GFP gene (threonine-9) was mutated at the same time without changing the amino acid (ACT to ACA). The resulting plasmid was pVLSNOGM (SEQ ID NO:18).

An Nco1-Xho1 fragment (585 bp) from plasmid pG1IL2EN (kindly provided by Dr. Steven Rosenberg, Bethesda, MD), containing the internal ribosome entry site (IRES) from encephalomyocarditis virus (EMCV) was inserted into the Apa1 site upstream of the GFP gene in pVLSNOGM, resulting in pVLSNOGMI (SEQ ID NO:17). Both insert and plasmid fragments were blunted with mung bean exonuclease. One variant version of pVLSNOGMI with an IRES tandem dimer was also constructed and designated pVLSNOGMI2 (SEQ ID NO:16).

Oligonucleotides (SEQ ID NO:53 and 54) containing a splice acceptor (SA) of AKV virus (in bold) was inserted into pVLSNOGMI at the unique Sac 2 site just before the IRES, resulting in pVLSNOGMI5 (SEQ ID NO:15).

Oligo: (SEQ ID NO:53)

5' -GGCCGCTAACTATAAGCCCATTCCTCCAAAGTGCTAGC-3'
3' - CGCGGCGATTGATTAACGGGTAAGAGTTCCATGCAT-5'  
(SEQ ID NO:54, bottom Oligo)

Recovery of LTR promoter sequences from mouse CD4+ T-helper cells

In order to facilitate the recovery of VL.30 promoter sequences expressed in mouse T-helper cells, a mouse CD4+ T-helper cell cDNA library (Stratagene, San Diego, CA, Catalog # 937311) was screened by plaque hybridization. Approximately 2 x 10^4 bacteriophage λ-ZAP clones were plated on a lawn of E. coli cells according to the manufacturer's instructions. Two nylon filters were sequentially layered onto the lawn of E. coli cells and bacteriophage. The filters were hybridized to a ^32P-labelled (Prime-It RmT Random Primer Labeling Kit, Stratagene), 4.2 kb internal XhoI fragment of NVL-3 (containing the NVL-3 genome). 55 plaques (or approximately 0.3% of the total phage) reacted positively on both filters. 18 VL.30 cDNA sequences were cloned from the plate, which was used to identify U3 promoters that are actively expressed in the RNA of mouse T-cells. Five of the 18 clones contained intact U3 sequences, representing four of one molecular species, named TH1 (SEQ ID NO: 2) and one of another species, named TH2 (SEQ ID NO: 3) also provided in Fig. 5. TH1 contained approximately 120 bp more DNA than did TH2. Because TH1 was more abundant (4 out of 5 clones), the additional sequences in the enhancer region were implicated to be a possible reason for the stronger expression in mouse T cells. Examination of the known and putative transcription factor binding sites in the VL.30 LTR (Hodgson, 1996, chapter 4, Fig. 4.2 supra) revealed several interesting features of TH1 and TH2. First, the extra sequences of TH1 that were missing in TH2 included an extra copy of the enhancer repeat region as well as a potential retinoid (RAR/RXR) binding site. Several transcription factor binding sites in the enhancer repeat region that differed between the two elements included: a cyclic 3'-5'AMP response element (VLCRE, a potential CREB/jun binding site), a serum response element (SRE), and a potential NF1/IL6 binding site (although there were additional sites for these factors in other enhancer repeats). These factors could possibly explain why VLTH1 appeared to be expressed at higher levels, both in the source cells and into transduced cells. Together, the VL30 sequences represented 0.3% of the mRNA expressed in the T cells, and TH1 appeared to be most abundant VL30.
Sequencing Primers:
(SK, SEQ ID NO:49) 5’-CGCTCTAGAACTAGTGGATC (20 mers, Tm 60°C).
(T7, SEQ ID NO:50) 5’-GTAATACGACTCAGCTATAGGG (21 mers, Tm 60°C).

Seamless Rescue of T cell promoters using class II S restriction enzymes

Two sets of primers containing offset Bpm1 restriction sites were designed and synthesized. One set was for amplification of the plasmid sequences, and another was for the amplification of the inserts.

Insert Primers: (Bpm1 site bold)

ITA (43 mer, Tm: 67.2 °C, SEQ ID NO:45)
CGATCCACTGGAGCTCGGAGCCACACCCCCCTCCCCATCTAGAGGT

ITB (43 mers, Tm: 66.3 °C, SEQ ID NO:46)
CGTCCTCCTGGAGACACAGGGTAGAGGAGTCTCGACGCTAG

Vector primers: (Bpm1 site bold)

VLA (43 mers, Tm: 68.2 °C, SEQ ID NO:47)
CGCAACCTCTGGAGACCTCTAGATGGAGGGGTTGGCTCCGAG

VLB (43 mers, Tm: 66.3 °C, SEQ ID NO:48)
GCAGGACCTGGAGCTGACCCTCGAGACTCTCTCTACCCCTGTG

To amplify vector sequences more efficiently, vector templates were shortened by deleting marker genes from vectors. pVLSNOSIB (SEQ ID NO:14) was cut with Kpn I and a 4201 bp fragment containing β-gal gene was removed. The remaining vector has 3923 bp.

The U3-promoter inserts (357 bp for TH1 and 240 bp for TH2) were PCR-amplified from TH1 and TH2 promoters with primers ITA and ITB. The vector cassettes (~4.2 kb for pVLSNOSIB and ~3.7 kb for pVLSNOGMIS) were PCR-amplified from the shortened vector templates using primers VLA and VLB, (supra). The PCR-amplification was done with high-fidelity *Pfu* DNA polymerase from Stratagene (La Jolla, CA). The amplified products were gel-purified (1% agarose gel). The inserts were then cut with Bpm1 to produce complementary ends. The vector cassette products were phosphorylated with
PNK, then circularized with T4 ligase, and transformed into SCS 110 cells. Recovered plasmids were then digested with Bpm 1 and treated with CIP to produce complementary ends. Bpm 1 treated inserts and vector cassettes were ligated, and T-cell tissue-specific VL 30 vectors VLTH1 and VLTH2 were produced. The marker β-gal gene and GFP gene were put back into those vectors at the original unique sites Kpn 1 and Sal 1 respectively.

**Transmission and expression of single LTR vectors and T cell U3 sequences**

Vector DNA constructs were transfected into GP+E86 retroviral helper cells (Markowitz et al., 1988, *supra*) using the Lipofectamine protocol (Life Technologies, Gaithersburg, MD). The culture media from these cells (supernatant), containing defective transducing particles (72 h post-transfection), was transmitted to PA317 (Miller, US Patent, cited supra) amphotropic helper cells, using Lipofectamine to enhance transduction efficiency (Hodgson et al., 1996. Synthetic Retrotransposon Vectors and Gene Targeting pp. 3-14, in : Felgner et al., eds. *Artificial Self-Assembling Systems for Gene Delivery*. American Chemical Soc. Books, Washington, D.C.). A similar procedure was used to transmit VLTH1 and VLTH2 to the PG13 helper cell line (Miller et al., 1991. *J. Virol.* 65:2220-2224). 24 h post-transfection, the recipient cells were selected with the drug G418 (500μg/ml, 2 weeks) to enrich for stably transduced cell populations.

All of the single LTR vectors, including VLTH1 and VLTH2 were transmitted by this method, indicating that single LTR vectors can be used for promoter switching and yet revert to dual LTR vectors after a single passage. Vectors VLSNO2, VLSNO3, and VLSNOSIB were then titered on NIH 3T3 cells (using the PA317 vector producer cell lines). VLTH1 and VLTH2 vectors were titered on human HT1080 cells (PG13 cell lines). Surprisingly, all of the single LTR vectors were transmitted effectively. However the titers of stably transduced TH1 and TH2 cell lines were $5.5 \times 10^2$-1.1 x $10^3$ TU/ml, compared to $0.4$-$3.0 \times 10^4$ TU/ml for the VLSNO2, VLSNO3 and VLSNOSIB cell lines. Thus, switching from the NVL-3 transcriptional promoter (originally isolated from NIH 3T3 fibroblast cells) to VL30 promoters derived from T helper cells, appeared to have a negative effect on RNA expression in fibroblast cells, as determined by the transmissibility of the RNA.

In order to study the usefulness of rescued promoters as DNA transfection vectors (as opposed to retro-vectors), VLSNOSIB, VLTH1 and VLTH2 were also transfected into a number of cell lines (using Lipofectamine), including NIH 3T3, PA317, GP+E86,
PG13, HT1080, SW480 and HeLa (available from ATCC). RNA expression in these cell lines is shown in Table 4, wherein gene expression from the LTR promoter (as determined by β-gal staining) is normalized to VLSNOSIB (100).

<table>
<thead>
<tr>
<th>Cell line:</th>
<th>NIH 3T3</th>
<th>PA317</th>
<th>GP+E86</th>
<th>PG13</th>
<th>HT1080</th>
<th>SW480</th>
<th>HeLa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VLSNOSIB</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>VLTH1</td>
<td>39.3</td>
<td>18.7</td>
<td>0.1</td>
<td>21</td>
<td>25.5</td>
<td>156</td>
<td>156</td>
</tr>
<tr>
<td>VLTH2</td>
<td>28.6</td>
<td>7.1</td>
<td>5.5</td>
<td>11.5</td>
<td>46.8</td>
<td>82</td>
<td>156</td>
</tr>
</tbody>
</table>

Table 4. Transient expression of a β-gal marker gene by three VL30 promoters: NVL3 (VLSNOSIB), VLTH1 and VLTH2. Cells were transfected using the Lipofectamine procedure. Total blue cells were counted from each well in 6-well plates, and the number of blue cells from VLSNOSIB was normalized to 100%.

The expression of both the VLTH1 and VLTH2 promoters was significantly reduced compared to VLSNOSIB in cell lines of fibroblastic origin, whereas in SW480 colorectal cancer cells and HeLa cells, it was comparable to or better than VLSNOSIB (the NVL-3 promoter). However, VLSNOSIB was expressed poorly in the non-fibroblastic cell lines, so a direct comparison was difficult to interpret. Unfortunately, the human T cell lines (Jurkat and MOLT4 [obtained from ATCC]) were not transfected by Lipofectamine, and they were poorly transduced by VLTH1 and VLTH2 retro-vectors. In the Jurkat and MOLT4 cells transduced with VLTH1 and VLTH2, only a small percentage (1-10%) of cells that were stably transduced by the vectors stained positively for β-gal expression. However, the marker gene (neo) continued to be expressed from an internal promoter, as evidenced by drug selection.

Taken together, the results demonstrated: 1) the ability of the promoter rescue technique to seamlessly capture functional transcriptional promoters from specialized cells; 2) the ability of single LTR vectors to introduce the rescued promoters into standard transducing vectors; 3) the ability of the rescued promoters to be expressed at differing levels in several different cell types, including T cells; and 4) screening and selection established the efficacy, or lack thereof, of individual promoter sequences.

Although the general method of promoter rescue was demonstrated by the foregoing experiments, the titers obtained from the sLTR VL30 vectors may not be useful where selection systems are not available.
Additional experimentation led to the development of a chimeric packaging signal, combining the essential packaging signal from Moloney murine leukemia virus (Ψ), and the enhanced packaging signal (Ψ⁺) from a mouse VL30 element. A vector embodiment of this packaging system is VLMB (SEQ ID NO:23). One advantage of the chimeric packaging system was the elimination of retroviral gag gene sequences that were present in previous high-titer MLV-based vectors (viral gag sequences contribute to the generation of replication competent retrovirus outbreaks). The titers of VLMB-based vectors ranged from approximately 1 x 10⁵ to 4 x 10⁶ TU/ml.

Construction of a cloning vector for promoter rescue

Using pVLSNOGMIS as a template, and primers (SEQ ID NOS:28 and 68), a 6.4 kb plasmid fragment was PCR amplified (Using Hot Start AmpliWax PCR Gems 100, Perkin Elmer). 30 cycles of PCR were performed by following the manufacturer’s instructions, with the following input conditions: lower reaction, 80°C, 5 min., then add upper reaction and template, 96°C, 1 min. Each reaction vial contained 50 ng template, 0.5 µM each primer, 200 µM dNTPs and 5U (2µl) Pfu polymerase (Stratagene, LaJolla, CA). 30 repeating cycles of: 96°C, 45 sec; 50°C, 45 sec; 75°C, 1 min. A final incubation of 75°C, 10 min, then hold at 4°C. After amplification, the reactions were purified using Qiaquick PCR Purification Kits (Qiagen). The PCR products were digested with Pac1, heat inactivated (65°C, 20 min) and ligated together using T4 DNA ligase (overnight at 4°C in a 5 µl vol). The ligated DNA was transfected into SCS110 E. coli cells (Stratagene) with kanamycin (50 µg/ml) antibiotic added to the agar plates. The cells were den't, dam't (to prevent methylation of Bpm1 sites). The resulting plasmid, pVLBPNGN (SEQ ID NO:1, Figs 2 &3) has a deletion in the U3 region of the LTR. A linker containing a central Pac1 site flanked by two outwardly-digesting Bpm1 sites occupies the site of the deleted U3 sequences. The Bpm1 sites enable the plasmid to be digested with Bpm1, resulting in two 2 bp 3'-overhanging ends that are complementary to the U3-derived RT-PCR inserts described below. The digested plasmid was purified free from the intervening linker sequences from an agarose gel after digestion with Bpm1, using the Qiaquick gel purification kit (Qiagen).
Procedure for amplification of liver U3 promoter region

Purified mouse liver total tissue RNA was purchased from Ambion, Inc., (Austin, TX). Total liver RNA was treated with RQ1 Rnase-free (Promega, Madison, WI). Using Perkin Elmer Gene Amp thermostable rTth reverse transcriptase RNA PCR kit (P/N N808-0069), the following conditions for RT-PCR were used: RT-PCR A 70° (hot start); RT-PCR B, 95°C, 60 sec, then 35 cycles (95°C 10 sec, 58°C, 15 sec) then a final 58°C incubation for 7 min, then 4°C and hold. Additional conditions were: primer concentration 0.15 micromolar, template 100 ng/reaction, dNTPs 200 micromolar (final) and MgCl₂ 3.5 mM(final). The primers for insert amplification were SEQ ID NOS:28 and 68)

The amplified U3 sequences were purified using Qiaquick. The pVLBPNG plasmid was digested with Bpm1, isolated from a 1% agarose gel and purified using the Qiaquick method. The purified U3 sequences were ligated at 1:2, 1:4 and 1:6 molar ratios of VLBPGN plasmid:insert using T4 DNA ligase and a 5 microliter reaction volume overnight at 4°C (100 ng plasmid: 16 ng insert = 1:1 molar ratio). 1 microliter of each ligation reaction was transformed into E. coli SCS 110 competent cells (Stratagene). 26 colonies were recovered in total. Out of 23 clones grown overnight in the presence of kanamycin, 20 had sequences that appeared to be mouse VL30 sequences, representing 10 different VL30 species (Fig. 6, SEQ ID NOS: 4-13). One of these (Hep 10, SEQ ID NO: 13) was transiently transfected into Hep G2 liver hepatocellular carcinoma cells. 48 h after transfection, intense GFP fluorescence was observed, indicating strong expression of the Hep 10 U3 promoter region.

Example 4

Creating a combinatorial library of mouse VL30 U3 sub-regions.

Using Fig. 7 and Hodgson, 1996, supra, Fig. 4.2 as a guide, the following three sub-regions of the VL30 U3 region were empirically established: Distal (1); medial (2); and proximal (3). Peaks of similarity were used to guide the following choice of primers: (+) primer binding site-5'-LTR boundary; ~80 bp (defines sub-region 1); ~80-210 bp (sub-region 2); ~210-430 (sub-region 3). The following primers were selected to amplify the vector VLBPGN or a similar VL30, NVL-3 LTR-containing vector:

P1 (going left from the 5'-end of the LTR to amplify the plasmid)
(SEQ ID NO:55)
GACTAACCTTGATTCACCTGAGTTTTC(CT)(CT)ATTCTTCAATCCCACTTC
TTC

P2 (going right from the 3'-end of the promoter region to amplify the plasmid)

(SEQ ID NO:56)
GACTAACCTTGATTCACCTGAGAAATCTGGACCAATTCTATTATAAGCCTG
TGAAAAATTT

The six primers selected to amplify the inserts are as follows:

Fragment 1, primer 1 (going right from the LTR terminus into U3) (SEQ ID NO:57)
GACTAACCTTGATTCACCTGAGAAGAAAGAAGTGGGGGAATGAAGAA

Fragment 1, primer 2 (going left from the end of fragment 1) (SEQ ID NO:58)
GACTAACCTTGATTCACCTGAGATCTCTAGATGGGAGGGGTGTG(CT)GGG

Fragment 2, primer 1 (going right from the left end of fragment 2) (SEQ ID NO:59)
GACTAACCTTGATTCACCTGAGCTCGGAGCACCACCCCTCCCATCT

Fragment 2, primer 2 (going left from the right end of fragment 2) (SEQ ID NO:60)
GACTAACCTTGATTCACCTGAGGGGAGCCCTATCTCAAAAAATGTT

Fragment 3, primer 1 (going right from the left end of fragment 3) (SEQ ID NO:61)
GACTAACCTTGATTCACCTGAGGTCTAAAGAACATTTTTGAGTAAGGGCC
T

Fragment 3, primer 2 (going left from the right end of fragment 3) (SEQ ID NO:62)
GACTAACCTTGATTCACCTGAGCTCAGGCTTATATAG(TG)AAA

100 ng of genomic DNA from *Mus musculus* is used as a template (the mouse genome
bears 100-200 copies of VL30 elements). Standard PCR procedures for *Pfu* polymerase are
used. Fragments are amplified 35 rounds of PCR to obtain single-copy genomic DNA
amplification. Samples of Qiagen column purified DNA are examined on analytical agarose
gels to determine the approximate size. The remainder of each reaction is digested with the
appropriate enzyme and run on an acrylamide or agarose gel. The digested fragments are
purified by standard gel purification procedures and are ligated to the plasmid fragment at an
equimolar ratio of the four PCR fragments (three inserts and one plasmid). The ligation mix
is transformed into *E. coli* SCS1 and is grown on kanamycin. The number of colonies is used to establish the size of the combinatorial library, and the pooled colonies are grown in *E. coli* and the DNA is harvested *en masse*. A dozen or more colonies are characterized by DNA sequencing to determine the approximate fidelity of the reaction. A library of 1,000 or more, but preferably 100,000 or more members is used for combinatorial screening procedures.

**Screening the combinatorial libraries for expression in specific cell types using a replication defective helper virus**

The U3 library DNA is transfected into the desired target cells in which expression is desired. Along with the library, approximately 25% of the total DNA should include retroviral helper sequences. The latter sequences can be a helper plasmid (such as pPAM3, Miller *et al.*, US Patent 4,861,719). The virus is amphotropic, permitting it to infect most human cells. The RNA from individual clones that are transcribed in the target cells will be packaged into retroviral virions made by the helper virus, and the virions can be harvested as the cell free filtrate (0.45 mm) from the vector producer cells. This virus (containing the expressed sequences) can be transmitted to fresh target cells that do not contain helper virus. 48 hours after passage, the DNA form of the transcriptionally active clones will be integrated in the recipient cells, and these transcriptionally active loci will produce more RNA, and protein. After G418 drug selection to increase the proportion of cells expressing the vector sequences, helper virus DNA is again transfected into the recipient cells, transforming them into vector producer cells. The virus from these cells should contain increased amounts of the RNA from clones that are transcriptionally active in those cells. Passage of the virus is continued for two or three rounds to permit recombination and mutation to take place, enhancing the effect of *in vitro* evolution of promoters. The actual degree of enhancement attainable at each step is illustrated in Table 2 (*supra*). After several passages, the actual level of RNA expressed by several clones is determined by RNA blotting, or by the amount of a reporter gene expressed as protein (determined visually or by the appropriate assay). Because human cells do not naturally contain VL30 DNA or RNA, the sequences that remain in the human cells are those with the most transcriptionally active promoters. These sequences can be amplified and re-cloned using the methods of the instant invention, or they can be rescued by virus packaging, reverse transcribed by the endogenous reverse
transcriptase reaction, and grown as plasmids (due to their plasmid origin of replication and the selectable kanamycin marker gene).

In addition to using a replication defective helper virus, such as the clone pPAM3, it is also possible to use a replication competent retrovirus, such as Moloney murine leukemia virus to passage the library. For use in human cells, however, the virus should have a tropism that is compatible with human cells (gibbon ape leukemia virus and amphotropic [4070A] murine retroviruses are acceptable).

In addition to being useful for generating active transcriptional promoters de novo, a small variation on the above procedures may enable the isolation of hormone responsive promoters. In it, the cells are treated with the hormone (which could be a steroid, a peptide hormone known to affect the cells, a drug, a drug agonist or antagonist, etc.) during passage. After isolation of surviving VL30 vector-containing cells, individual clones of drug resistant cells are tested for reporter gene expression with and without drug treatment to determine relative protein expression. Likewise, RNA expression can be determined by blot analysis or a similar method. A useful list of known VL30 responses to pharmacological agents is listed in Fig. 4.2 of Hodgson, 1996, supra, and can be used as a guide to help assess the potential agents known to have an effect on VL30 transcription.

Once the transcriptional promoters with the known specificity have been obtained, they can be used to obtain expression of genes from a variety of types of vectors. For example, in addition to retrovirus particles, the promoters can be incorporated into all other major groups of vectors: adenoviruses, herpes simplex virus vectors, DNA transfection vectors, etc. It will be apparent to persons of ordinary skill in the art that similar combinatorial libraries can also be used to screen for other characteristics than transcription activity in a particular cell. For example, combinatorial libraries of complementarity determining regions (CDRs) of antibodies or T cell receptors can be so screened using antibody screening methods, such as the phage display screening method (Pharmacia, Milwaukee, WI). Thus, the methods of this invention, particularly the combinatorial simplicity of this invention is a significant improvement over many in vivo recombination methods including those of (Stemmer, US Patent 5,605,793; 1997) that have described for the production of CDR combinatorial libraries.
Example 5
Gene Assembly Line

From the above examples of 3 and 6 fragment gene self-assemblies, it is evident that assembly of genes by means of gene amplification, the use of offset restriction enzymes and incorporating unique, non-palindromic ends is a highly efficient process compared to conventional cloning methods. However, in addition to the considerations already discussed, it will be apparent to a person of ordinary skill in the art that the various procedures, protocols, methods and material of the instant invention become more difficult to use as the number of fragments increases. For example, if the efficiency of combining each fragment in an assemblage is 99%, then the overall efficiency of combining ten fragments will be 90%, the efficiency of combining 100 fragments will be 37%, etc. Therefore, a small drop in efficiency of any step or fragment, or a large increase in the complexity of the project, will be sufficient to reduce the overall efficiency. Fastidious procedures permit one to achieve success with more complex projects.

Foremost in its potential for inducing failure is human error in primer design where large numbers of fragments are used. Fortunately, the instant invention is suited to automation of most of the steps. This allows human input to be focused on design, analysis, and quality control. For the purposes of generating large vectors or chromosomes, it is desirable to provide an automated environment. One method to achieve this goal is a gene assembly line.

In a gene assembly line, multiple tasks are controlled by a machine or machines working together to increase speed and efficiency and to reduce human error. For example, computer aided design (CAD) and computer aided manufacturing (CAM) are incorporated and combined with the methods of this invention. The computers accept inputs in the form of template and primer sequences, together with preferences of regions to be copied and joined. The preferences include at least the sequences of the primer regions and information about the known restriction sites and maps of the sequences to be assembled, but ideally include the entire sequence. The preferences also include the number of sequences to be joined, the desired Tm for the primers, the list of potential restriction enzymes capable of offset digestion that are potential candidates for use in the assembly process, the desired end structures for each fragment terminus, a tag sequence (if any), whether circular or linear ends
are desired, and additional design considerations. The computer algorithm then searches the sequences to determine the candidate enzymes and specific primers that match the criteria of the input. Candidates for selection of unique non-palindromic overlaps are selected. The computer then posts selections or preferences for the type and order of end structures, the primer binding sites, their potential for primer-dimer and intra-molecular interaction artifacts, and the potential conflicts with repeat sequences within the templates that could lead to incorrect polymerization. Based upon the selections made by the operator, the computer then determine the T<sub>m</sub> for each primer, and makes adjustments (with suitable inputs from the investigator) to achieve a suitable T<sub>m</sub> for the appropriate DNA synthesis or gene amplification reaction. Ideally, the primers should have similar T<sub>m</sub> so that all amplification reactions can be performed at once with one set of amplification instructions. In reality, it may be difficult to do this with complex projects. The output of this portion of the program, which can be in a generic format, such as a Microsoft Excel spreadsheet is then downloaded to a computerized oligonucleotide synthesizer, such as the Applied Biosystems 3928 nucleic acid synthesizer.

One advantage of using a computerized synthesizer is its robotic capability to de-protect and purify the oligonucleotides automatically. In addition this synthesizer can accept computerized input.

The quantity of individual oligos recovered is then determined spectrophotometrically. It is desirable to purify the oligonucleotides by high performance liquid chromatography or by polyacrylamide gel. In a preferred embodiment, the oligonucleotides and templates are then assembled robotically using an automated nucleic acid handling system such as the Qiagen BioRobot 9600. The BioRobot is capable of accepting input from a computer and can combine the gene amplification reactions based upon the assignments of templates, primer and reagents provided in the input. The assembled reactions are then amplified for example by PCR. In a preferred embodiment, the PCR heat block is incorporated into the robotic workspace and genes are assembled robotically but with minimal human intervention to change buffers, rearrange the platform, change programs, and the like. The resulting amplified products are also purified by the BioRobot or a similar robotic device. In a preferred embodiment, the robotic device uses Qiaquick cleanup procedures, or a similar method and then assembles restriction endonuclease reactions to digest the purified gene amplification products. The gene amplification products are loaded onto a gel and electrophoresed. Human intervention may be necessary to analyze the
products and excise the correct fragments from the gel. At this point, the results are assessed and missing or incorrect sized fragments are resynthesized. The robotic device is preferably used to purify the gel fragments using Quiagen or similar cleanup procedures. After spectrophotometric quantitation of the purified fragments, the robotic device is preferably used to assemble the ligation. Ideally the fragments are combined in an equimolar ratio of 1:1. However it is not necessary to use equimolar ratios in order to achieve gene self-assembly. For automated gene assembly, it may be desirable not to use equimolar ratios of input fragments, particularly if it simplified the task of quantitation. After ligation, the assemblies can be purified and ethanol precipitated or they can be added to the appropriate host cells. Automation aids in maintaining the sterility of the reaction.

Several additional considerations can assist in the construction of long genes using gene assembly. First the number of fragments and the length of constructs are limiting factors. In addition to maintaining high standards of purify of both the oligonucleotide primers and gene amplification products, it is important to keep the error rate low during copying. Thus, one can optimally start with 100 ng of template use only five rounds of gene amplification and finish with nearly 2 micrograms of product. This is more desirable for reducing errors than using a large number of amplification steps. It is also desirable to use a special copying enzyme such as Pfu DNA polymerase that has a low intrinsic error rate. Further it is desirable to use in vivo selection (in eukaryotic cells or tissues) rather than E. coli cloning to reduce the incorporation of errors into the vectors. For example, a viral vector such as an adenoviral vector or the retro-vectors of the preceding examples are auto-selecting. A single correctly-assembled adenovirus vector molecule, for example, leads to a lytic infection (the viral products of which are cloned by limiting dilution on the appropriate eukaryotic cells), even though it may be combined in a ligation mix with a large excess of incorrectly assembled molecules that are non-functional. Thus, it is not necessary to have a high efficiency, although high efficiency has been demonstrated in this system, in order to achieve success in making, for example gene therapy vectors.

For long fragments (3-30 kb), it is desirable to use enzymes and procedures that are designed or facilitate replication of long fragments, one such example is the eLONGase system (Life Technologies). This system can copy up to 30 kb on a fragment with proofreading. Considerations for long PCR are reviewed in Beck, 1998. (The Scientist 6 January, 1998, pp. 16-18).
Internal restriction sites are a potential problem, particularly with large constructs and can be overcome in a number of ways. Use of alternate enzymes, methylation of internal restrictions sites (such as by using methylated DNA precursors during synthesis to leave the sites in primers unaffected, incorporation of the internal sites into the construct (if they are non-palindromic), or mutagenesis of internal sites, are exemplary ways to deal with some of these issues.

For very large constructs, it is desirable to use enzymes such as SapI (recognizing 7 nucleotides and leaving a 3 bp overhang). This enzyme digests every 16,384 bp on average. There are 64 nucleotide triplet combinations, meaning that up to 32 fragments can be ligated in a circle using SapI. FokI and HgaI are other examples of class IIS enzymes that are useful for making large constructs. HgaI has 5 bp overhangs, permitting more than 500 HgaI fragments to be ligated. FokI includes a Kozak ATG start codon. In a preferred embodiment, a FokI site is inserted at the PuXXATG start site of a cDNA encoding region. The cDNA is inserted in frame, providing a site for inserting and switching coding sequences within a vector.

It will be readily understood by those skilled in the art that the foregoing description has been for purposes of illustration only and that a variety of embodiments can be envisioned without departing from the scope of the invention. Therefore, it is intended that the invention not be limited except by the claims.
(1) GENERAL INFORMATION:

(i) APPLICANT: NATURE TECHNOLOGY CORPORATION, ET AL.

(ii) TITLE OF INVENTION: SELF-ASSEMBLING GENES, VECTORS AND USES THEREOF

(iii) NUMBER OF SEQUENCES: 68

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: MÜETING, RAASCH & GEBHARDT, P.A.
(B) STREET: 119 NORTH FOURTH STREET, SUITE 203
(C) CITY: MINNEAPOLIS
(D) STATE: MINNESOTA
(E) COUNTRY: USA
(F) ZIP: 55401

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.30

(vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER: Not Assigned
(B) FILING DATE: 28-FEB-1998
(C) CLASSIFICATION:

(vii) PRIORITY APPLICATION DATA:
(A) APPLICATION NUMBER: 60/070,910
(B) FILING DATE: 28-FEB-1997
(C) CLASSIFICATION:

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: MCCORMACK, MYRA M.
(B) REGISTRATION NUMBER: 36,602
(C) REFERENCE/DOCKET NUMBER: 228.00010201

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: 612-305-1225
(B) TELEFAX: 612-305-1228

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6225 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:
TGAAGAATAA AAAATTACTG GCCCTCTTGAG AGAACATGAA CTTTCACCTC GGAGCCGACC  60

CCCTCCGATC TGGAAAATCT CAGTTATAAG TGGAGTTTTT CCTTTAAAAAG CTTGCGAAAA  120
ATTTCAGGCG TGCTGAGAGG TCCTCATTCC TGTGCAAAGG TGATGAGTT TGGACCCGAG  180
AAATTTATTT GCACTACTGG AAAACTACCT GTCCTCAGTG CCACACTTTG CACTACTTTG
ACTTATGGGT TTCAATGGCTT TTCAAGATAAC CGAGTACCTTT GAAAACGGCA TGACTTTTTCT
AAAGCTGCAA TGCGCGAGGG TTATGTGACAC GAAAGAACAC TATTTTCTCAA AGATGACGGG
AATCAGACAG CACGTCGCTGA AGTCAGTTTG GAAGTGTATA CCCCGTATAC TAAGATCACG
TTAAAAAGTA TTGATTTTTTAA AGAAGATGGGA AACCATTCTTG GACACAAATT GGAATAACAC
TATAACTCAG ACAATGCTATA CATCATGCGCA GACAAACAAA AGAAATGGAC CAAAGTATAC
TTCAAAATTA GACGAACATC TGGAGATGGA AGCCTTCAAC TACGACACCA TTATCAACAA
AATACCCTAA TTGGCGGTATGG CCTGTGTCTT TTACGACACA ACCATTATCT GTCCACACAA
TCTGCCCTTT CGAAGAGATCC CAAGAAAGAG AGAGAGCACA TGCGTCTTTCT TGAGTTGTGA
ACAGCTGCTGT GGATTTACACA TGCGTATGATG GAATTAATACA AGTCCGGATC TAGATAACTG
TATCGAGTGG TCAGGAAGCG CCGACAGCAG TGCACTGGTG GACAGAAAGC AAGTGATCTA
GGCCAGCCGC CTCCCCACAG GCAGGCTACCC AAACTTGTGG CTTTTAATAC
AGCTCTGTAA ATGGTTAAAA AAAAAGAATG TACGCAGCGAC GCAGATTGCG TCTTGCCACT
GTACAGACGA ATATACAGAC AGAAAGATCTG CACCGACATC GTGACATCT TGGAGAACAG
CTTGGCTAA AAAGAACTAG ATGGCAAAATC TAAAGCGCCCA GCCATCTTAA AGAGCAATGA
TCTTGACGTG CGTGAAGACTA TCAAGTTATA GACAAATTTA GACTGGTAAA AAAAACCTTG
TATAAAATAC TAAAAACGTA AAAAAGAATA CATGTCCTTT CATGGAAGA CAGACCTTAC
ATCTACTGAA AAATAGACTT TACTGGAAA AATATGTGTA TGAATACCTT TCAGTTTTTG
TGAACGGTTCT CAAGATGGAT AAAAGCTTTCT CCTGTGAAAA CAGACGATG CAGATATGCA
TCAGAGACAGAT TGTTAAAGGA AAATTITTCAAA GGTGTGAGGT GCCAAAASGC ATAGTGTCAG
ATAATGCTCC TGGCCTTTGT GGGAGGGTTA GTGAGGGGTG GGCCAAGATTT TTAGGATCT
AAGGAAATT TGAGGCTGGA TACGAGACTC AGAGCTCAGG AAAGATAAAA AAGAATAAAT
AAACCTCTAA ACAGACCTTG ACACAAATTA TCTGTAGACG TCAGCAGACG TTACTTGGTA
CTCCCCCTCC TGCGCCCTATT AGAATCAGTG ATACCTCTCT TTACGATCTG TCTTTTCCTC
TCTAAGATCC TTATGGGGGC TTCTATGGCCT TCACTGCTTT AAAGATGATGTT TTTAAACCTA
TGTTGTTATA AATATGACTCT ATATGTGTAAG TAAAAGGGGT TGGAGGTTGT GCAGAAGAAA
GTCGAGGCAC AGTGGACACA GTGGGTACCC CAAGGACATC TTACGAGTTC
CAGCCAGAGA TCTGATCTAC GATCCCCGGG TCGACCAGGG TGCAAACCTGT GGAATGTTGT
TCAGTGTAGG TGTTGAAGAT CCCAGGGCTC CCCAGCATGC AGAAGTATGC AAAAGCATCA
TCTCAATTAG TCAGCAACCA GGTGTGAGAA GTCGCCAGGC TCCCGCAGGC GCAGAAGATAT
GCAAAAAGATG CATCTCAATT AGTCAGACAC CATAGTGTCG GCCCTAACTC GCAGGCAATCC
GCCCTAATCT CCGCGAGATT CCGCCGACCT CCGGCCCATC TGGCTGACTAA TTTTTTTTAT
TTATGACGAG GCAGAGGCGG CCTGCGCTTC TGAGCTATTG CAGAAGTACG GAGGAGGTTT
TTTTGAGGCG CTAGGCCTTTT GCAAAAAGCT TACGCCTGCC GCAAGCACCT AGGCGCAAG

---

58

PCT/US98/003918

WO 98/38326
GGCTGCTAAA GGAACCAGAA CACGTAGAAA GCCAGTCGCC AGAAACGGTG CTGACCACGG 4260
ATGAAAAGTCA GCTACTGGGC TATCTGGGACA AGGGGAACGG CAAGGCACAAG GAGAAAGCAG 4320
GTAGCTTGCA GTGGGCTTAC ATGGGATAG CTTAGACTGG CGGTTTATAG GACAGCAGAAC 4380
GAACCGGAAAT TGGCAGCTGCG GGGCCCTCCT GGTAAAGTGG GGAAGCCCTG CAAAGTTAAC 4440
TGGATGGGCTT TCTTTCGCCGC AAGGATTCTG TGGCCAGGGG GATCAAGATC TGGATAGAG 4500
ACAGGATGAG CACGCTTTGC CATGATTGAA CAAGATGGAT TGGACGCAGG TTCTCCGCCC 4560
GCTTGGGTGG AGAGGCTATG CGGCTATGAC TGCCCACACAAG AGAACAATTCG CTGCCACCTGAT 4620
GCCGCCGCTGT TCCGGCTGTCC AGCCGACGAGG GCCGCCTTGC TTCACCTGCAA GACCAGACCTG 4680
TCCGGTGCCC TGAATGACAT GCAGACAGAG GCAGACGCAGG TACCAGTGGC GCCAACGAGG 4740
GGCGTTCTCT CGCGACAGTG TCTCGAGCTT GTCACTGAAG CGGGAAGGGA CTGGCGTCTA 4800
TGGGCGGAAG TGGCAGCGCA GGATCTCTTG TCACTCTACC TTGCTCCTGC CGGAAAGATA 4860
TCCATCATGG CGTGATGCAAT GCAGCCGCTG CTACAGCTTG ATCCCGCTAC CTGGCCATCTC 4920
GACCACAAG CGCAACATCG CATCGAAGCG CAGCAGTATCG GGTGAAGAGC CGGTCTTGTGC 4980
GACACAGG ATCTGGGAGA AGAGACATCG GGGCTCGCGC CAGCAGAATCT GTCGGACAGG 5040
CTCAAGGGGC GCAGTGCGGCA CGCGAGGATG CTCGCTGTGA CCAAGTGGGA TGCGCTGTTG 5100
CCGAAATATCA TGGTGGAAGA TGGCCGCTTT TCTGGATATCA TCGACTTGAG CGGCGCTGGT 5160
GTGGCGGGAC GCCATACGAGA CATACGCTTG GCTACCCGTG ATAPTGCTGA AGAGCCTTGGC 5220
GCCGAATGGG CTGACCCGTT CCTCGTCTCTT ACATGATGCG CCGTCTCCAG ATCCGAGGCC 5280
ATCCGGCTCT ATCGCCTTCT TGACAGTGTT TTCTGAGGCG GACTCTGGGC TTCCGAATGA 5340
CGAACCACGC GGGCCGCAAC CTGCGCTACAC GAGATTTGCG TFCACCGCCG GCCTCTTATG 5400
AAAAAGTTTGGG CTCTGGGAATC GTTTTTCGGG ACGGAATTCG TAATCTCGTG CTTGCAACA 5460
AAAAAAACCC CGTGACCCAGC GGTGTTTTTG TGGCCGGATC AAGAGTACC AACCTTTTTT 5520
CGGAAGGTAA CTGGCTCTCG CAGGGCGCGAG ATACCCAAATC CTGTCCCTCT ATGTAGAGGGC 5580
TAGTGGGCCC ACCATCTCAA GAACTCTGTA GCAGCCGCTA CATACCTGCG TCTGCTAAAC 5640
CTGTACCACG TGCCGCTCTGC CAGGCAGGAT AAGCTGCTGTC TTACCAGGGT GGACTCAAGA 5700
CGATAGCATTAC GGGATAAGGC GCAGGCGCGG GGGCGAAGCG GGGGTCTGCG CGACACCCGC 5760
AGGCTGGAGC GAAACGACCTA CACCGAACTG AGATACCATC AGCGTGAGCA TTTGAAGAAC 5820
GCCACCCCTTG CGAAGGGGAG AAGAGGCGGC AGGTTATCGG TAAGCGGCAG GGTGGAACAA 5880
GAGAGAGGAA CGAGGCAGCT TCCAGGGGA AAGCGCTCGT ATCTTTTATG TCCGCTGGGG 5940
TTTGGCCACC TCTGACTTGA GCAGCGGATT TGTGTAGTGC CTTGACGCGGG GGGGACCTGA 6000
TGGAAAAACGC CCAGACCGCG CCGAGTGCGG CGCCTCGAGT AAGCTCAGCTG CAGCGTAAAG 6060
CCTCAAGGCC TCACTAAAGG GCTCCTCGCC TAGTCTCTTT TACTAATCTG CTTTATCTCG 6120
TTTTGTTTCC CATGTTAGAA ATAGAGTAAA TGCAGATATC TCCACATAAG GATATAGACT 6180
TCTGAAATTC TAAGATTAGA ATTATTTACA AGAAGAAAGT GGGAA

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 487 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:
CCTCCCATCT AGAAAGTTGT TCTGGAACAC TCCTAAACTTT TTCAACCCCAA AACTCCCCAC 60
CCTAAAGGTC GAAAAAATCG TTCAAGAAGAC ATTTTTGAGA TAAGGCTCTC CTGAACAAC 120
CTCAAAAATGA CATTGGCAAAG TGAATAGCA CTGAACCTCTT TACGCTAGG TTCTCTGATA 180
GGACATGACT CCTTAGTTAC GTAGGTCTCT TGAAGGACAC TGACTCCCTA GTAACGTAGA 240
TTCTTTTGCT AGAAACTCCCT ACGATGTAAC ACTTGCTACTT TCCCTGCCCA GTTCTCCCA 300
TTGGAGTTTT ACTATATAAG CCTGAAAAAC ATTTTTGCTG ACCGTGAGA CTCCCTTACC 360
CTTGGTCTAGG GTGATGAGTA TTGACCCCA GAGCTGCTGT TGCTTCTCATG TTGCTGCTTT 420
ATTTCGACC CAGAGCTCTG GTCTGTTGTC TTCTATCTCG CTGCTTTATT AAATCTGACC 480
TTCTACA
487

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 366 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:
CCTCCCATCT AGAAAGAATT TTTTAGATAA AGGCTCCTCT GAACAACCCTC AAAATGAACC 60
AGTACTCCT TAGTTAGTGA GTTTCCCTGA TAGGACATGC CTCCTTACTG ATACAGATTC 120
CTTGGCCAGA ACTCCTAGT GTAGTTAAACT TGTTACTCTCC CTGCACCTTT CTCCCCCTTT 180
GATTATTACT ATATAAGCCT GTGAAAAATT TTGGCTGACC GTGAGACTC CTCTACCCCTG 240
TGCTAAGTGT TATGAGTTTC GAACCCAGAG CTCTGTGTCG TTCCATGTGG CTGCTTTATT 300
TGGACCACAG AGCTCTGGTC TGTGTTTTTT CATGTTGCTG CCTATTAAAA TCTTGCTCTC 360
TACATT
366

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 304 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

CCTCCCATCT AGAGATTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAAG AATGCATGCC

TGAACTCCTC ACCTTAGAGT TCGAACCCTC CCAACTAAAG ACTGTTCGAA GACCATTTTT

GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAAACGGG TACATGGCCA AATAATAGGA

CATGACCCTCT TAGTTACGTA AAATCCCTTG GCAGAAACCC TTCGTCCCTTG GCAGAACCCCC

TTAGTTATGT AAACCTTGTAC TTTCCCTACC CCGCTCTCCC CCTTGAGTT TTTCTATAT

AAGC

304

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 304 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

CCTCCCATCT AGAGACTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAAG AATGCATTCC

TGAACTCCTC ACCTTAGAGT TCGAACCCTC CCAACTAAAG ACTGTTCGAA GACCATTTTT

GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAAACGGG TACATGGCCA AATAATAGGA

CATGACCCTCT TAGTTACGTA AAATCCCTTG GCAGAAACCC TTCGTCCCTTG GCAGAACCCCC

TTAGTTATGT AAACCTTGTAC TTTCCCTACC CCGCTCTCCC CCTTGAGTT TTTCTATAT

AAGC

304

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 304 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

CCTCCCATCT AGAGAGTTT CCCAGAACAC TCCTGAACTC TTCACCCCAAG AATGCATTTCC

TGAACTCCTC ACCTTAGAGT TCGAACCCTC CCAACTAAAG ACTGTTCGAA GACCATTTTT

GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAAACGGG TACATGGCCA AATAATAGGA

CATGACCCTCT TAGTTACGTA AAATCCCTTG GCAGAAACCC TTCGTCCCTTG GCAGAACCCCC

TTAGTTATGT AAACCTTGTAC TTTCCCTACC CCGCTCTCCC CCTTGAGTT TTTCTATAT

AAGC

304
TGACTTCTC ATCTAGAGT TCGAACCCTC CCAACTAAAG ACTGGTCCAA GAACATTTTT
GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCTGG TACATTGCAA AATAAATAGGA
CATGACCCCTT TAGTTAGGTA GAATCCCCGG GCAGAACCCTC TTGTCCTCCTG GCAGAACCCTC
TTAGTTATGC AAACCTTGCAC TTCTCTGCC CCGCTCTCCC CCCTTGAGTT TTTCTATAT
AAGC

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 304 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

CCTCCCATCT AGAGATGTT CCCAGAACAC TTGCTGAACC TTTACCATCAG AATTGCAATCC
TGAAACTCCTC ACCCTAGAGT TCGAACCCTC CCAACTAAAG ACTGGTCCAA GAACATTTTT
GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCTGG TACATTGCAA AATAAATAGGA
CATGACCCCTT TAGTTAGGTA GAATCCCCGG GCAGAACCCTC TTGTCCTCCTG GCAGAACCCTC
TTAGTTATGC AAACCTTGCAC TTCTCTGCC CCGCTCTCCC CCCTTGAGTT TTTCTATAT
AAGC

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 305 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

CCTCCCATCT AGAGATGTT CCCAGAACAC TTGCTGAACC TTTACCATCAG AATTGCAATCC
TGAAACTCCTC ACCCTAGAGT TCGAACCCTC CCAACTAAAG ACTGGTCCAA GAACATTTTT
GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCTGG TACATTGCAA AATAAATAGGA
CATGACCCCTT TAGTTAGGTA GAATCCCCGG GCAGAACCCTC TTGTCCTCCTG GCAGAACCCTC
TTAGTTATGC AAACCTTGCAC TTCTCTGCC CCGCTCTCCC CCCTTGAGTT TTTCTATAT
TAAGC

(2) INFORMATION FOR SEQ ID NO:9:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 305 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

CCTCCCATCT AGAGACTGTT CCCAGAACA TTCCTGACTTC TCCACCCAG AGATGCATTCC 60
TGAACCCCTC ACCTAGAGTT TGAAACCTTC CCAACTAAAG ACTGTTCCAA GAACTTTTT 120
GAGATAAGGG CCTCCCTGAA CAACCTCAGA ATGAACCCAG TACATTGCCA AATAATAGGA 180
CATGACCCCT TAGTTAGCCTA GAATTTCCCTT GGCAGAACC CCCTGTCTCTT GGCAGAACC 240
CTTAGTTATG CGAACTCTGTA CCTTCCCTGC CCCGCTCTCC CCCCTTGAGT TTTTCTATA 300
TAAGC 305

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 306 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

CCCTCCCATC TAGAGAGGTT TCCCAGAACA TTCCTGAACTT TCCATCCCA GAATGCATTCC 60
CTGAACCTCT CACCCTATAG TTGGAACCCT CCAACTAAAG GACTGTTCGA AGRAGATTTTT 120
TGATAGATAG CTCCCTGGA ACAACCTCAG AATGAACCGG GTACATTGCC AAATAATAGGA 180
ACATGACCCCT TAGTTAGCCTA AGAAATTTCTT TGCCAGAACC CCTGCTGCTG TGGCAGAACC 240
CTTAGTTATG GTAAACTCTGTT ACTTTCCCTG CCCGCTCTCC CCCCTTGAGT TTTTCTAT 300
GATGC 306

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 305 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

CCTCCCATCT AGAGAGTTT CCACAACAC TCTTGAACCTC TTCACCCCAAG AATGCAATCC 60
5
TGAACCTCTC ACCCTAAAGT TGAACCCCTC CCAACTAAAG ACTGTTCAAA GAACATTTTT 120
GAGATAAGGG CCTCCTGAAA CAACCTCAGA ATGAACCGGG TACATGCGCCA AATAATAGGA 180
CATGACCCCT TAGTTACACA GAATTCCTTT GGCACAACCC CTTGTCCTTT GGCAGAACCC 240
CTTAGTTATG CAAACTTGTA CTTTCCCTGC CCAGCTCTCC CCCCTTGAGT TTTTCTTATA 300
TAAGC
305
(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 304 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

CCTCCCATCT AGAGAGTTT CCACAACAC TCTTGAACCTC TTCACCCCAAG AATGCAATCC 60
30
TGAACCTCTC ACCCTAAAGT TGAACCCCTC CCAACTAAAG ACTGTTCAAA GAACATTTTT 120
GAGATAAGGG CCTCCTGAAA CAACCTCAGA ATGAACCGGG TACATGCGCCA AATAATAGGA 180
CATGACCCCT TAGTTACACA GAATTCCTTT GGCACAACCC CTTGTCCTTT GGCAGAACCC 240
CTTAGTTATG CAAACTTGTA CTTTCCCTGC CCAGCTCTCC CCCCTTGAGT TTTTCTTATA 300
40
AAGC
304
(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 303 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

CCTCCCATCT AGAGAGTTT CCACAACAC TCTTGAACCTC TTCACCCCAAG AATGCAATCC 60
55
TGAACCTCTC ACCCTAAAGT TGAACCCCTC CCAACTAAAG ACTGTTCAAA GAACATTTTT 120
GAGATAAGGG CCTCCTGAAA CAACCTCAGA ATGAACCGGG TACATGCGCCA AATAATAGGA 180
CATGACCCCT TAGTTACACA GAATTCCTTT GGCACAACCC CTTGTCCTTT GGCAGAACCC 240
CTTAGTGATG AAACCTTGTA CTTTCCCTGC CCAGCTCTCC CCCCTTGAGT TTTTCTTATA 300
AGC
303
(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 8657 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

TGAAGAATAA AAAATTACTG GCCCTCTTGTG AGAACATGAA CCTTCACCTC GGAGCCCAAC 60
CCCTCCCATC TGGAAACAT ACTTGAGAAA AACATTTTCT GGAACAAACA CAGAATTGTTT 120
CAACAGGGCA GATGTATGCG CAAACACAGG ATATGACTCT TTGGTTTGGT AAATTTTGTG 180
TTTTAAACT TCCCCATTCC CCTCCCCAT CCCCCCTCCA GTTTGTGGTT TTTCTCTTTA 240
AAAGCTTGTG AAAAATTTGA TGGCGTGTCG AGACTTCCTT ACCCTGTGCA AAGGTTATG 300
AGTTTCGACC CCCAGGCTCT GGTGTGCTTTCC TGGTCTGCTT CTTATTTGCAG CCCAGGCTC 360
TTGTCTGTTGT GCTTCTCATG CGTGTCTCTTA TTAAAATCTTA CCCCCTCATC TTATTTATG 420
GTCTCAGTGT TCTTCTGGGT ACAGCGGTGT CCCCAGACTT GAGTGTCGTTA GTGAGCCTTCT 480
TCCCTCGAGG GTCTTTCCAT TGGTACATGG GCCGGGAATT CGAGAATCTT TCATTTGTG 540
CATTTGCGCG GAATTCGAAA ATCCTTCATC TTGGTGATTT GGCCGGAAAC AGCGCCAGCA 600
CCCAGAGGTC CTAGACCCAC TTAGAGTAA GATCCTTTGT CTTGTTTTGG TCTGATGTCT 660
GTGTTCTGAT GTCTGTGTTCC TGGTCTCAAG TCTGCTGCGA TCGCAGTTTC AGTTTTCGGC 720
ACCTCTAGTG AGACCCGGCT CCCAGAGGGA GTCCGGGTTT GTAAGGATA GACGTGTCCA 780
GGTGTTCCAC GTCCGTGCAC CCTGGGAGAC GTCCCGAGAG GAACGGGGA GGATCGGGGA 840
CGCTGTTGCG ACCCTTTGGA AGCGCAAGAG ACCATTTGGG GTGGCGAGAT CTGGGGTTGC 900
AGTCCACCTC CTGTCGGCCCG TCGGAGATCG TGGTTTCCAG TCCCCCTCCT GCCTTTTGTG 960
CGAGATCTTG GGGTGACGTC CCACCTCGCC TCTGCTGACG GGATGTGGGG TCGGCTGCC 1020
ACCTCTGTTT TTGTCTGCGG AGTGTTGGTT CCAGTCGCCAC CTGCGTGTCG GTCACTGGTT 1080
CTGGGGTTGC AGTCCACCTC GTGCGAGAG AGTCTAAGTT GCGGGCCTTA GAGAGGCGAT 1140
CGATTCTCTC TTGTTTCTTT TTTGTCCTTA GTCTCGTGTCG CGCTCTTTTG GTGACTACTG 1200
TTTTTCTTAA AATGCGCCAT CCTGCTGCTCA CCTCCCCTTC TCTGACTCTG GTCTGTGC 1260
TGAGTTTTTT TTGTTTGTATA CGTTTTTTTG TTGGAGTGAT CATAGTGTCG TTGACTATC 1320
TGCTCTTTTTT TGGTTGTGTT ACGTTTGTCTT GTGCTGGCTT TTGGTCTCTT TTGTTTCAGA 1380
CTGGAACGTA TGACTGACGA CTGTTTTTAA GTATGCCCTT CTAAAATAG CTCAAAAATC 1440
CTGTCGATCT CCTATGTGTA CCACCTCCTT TCAGATCAAC AGCTGCCCATT AGTATGTG 1500
GGATCCCTCG ACTRACCTAT AGCCCCATTCT CCAAGGTGCA GCAGGATCAA TTCCGCCCCC  
CCCTAAGCT TACTGGCGGA AGCCGCTTTGG AATAAGCAGG GTGTCGTTTT GTCTATATGT  
TATTTCACAC CATATGGCGG TCTTTTGGCA ATGTGAGGGC CCGGAAGAAG GCCCCCGTCT  
TCCTTGAGAG CATCCTAGAG GTGCTCTCCC CTCTGCCCAA AGGAATGCAA GTGCTGTGGA  
ATGTCGTAAG GGAAGCGATT CCTCTGGAAG CCTTCTTGAG ACAAACACGC TCTGTACGGA  
CCCTTTGCAG GCAGCGGAAC CCCCAACCCTG GGCACAGGTT CCTCTGCGGC CAAAGCGCAC  
GTGTATAGAG TACACCTGCA AACGGCGCAC AACCCCAATG CCACGTGTGTG AGTGTGATAG  
TGTTGGAAAG AGTCAAAAGT CGCTCTCCAA CGGTATTTCA AAAGGGGCTG AAGGATGCC  
AGAAGGTACC CCATTGGTACG GGAATCTGATG GGGGGCTTTG GTGGCACTAGC TTCAGATGG  
TTTATGCAG GTTTAAGAAA CGTCTAGGGCC CCCCGAAGCG CGGGGACAGT GTTTTCTTTT  
GAAAAACAGC ATAATAATCA TGGGCCGAGG TCCCGTCTTT TPACAAGCGTC GTGACTGCGA  
AAACCCCGG GTTACCCCAAC TTAATGCGCT TCGAAGCACAT CCCCTTTCG CCACGTGCGC  
TAATAGCAGA GACCCGCCCA GCATGCGTTC TCTCCTCAAC AGTCGCAGC TGAATGCGGC  
ATGGCGGCTTT GCCCTGCTTTT CGCGCACAGA AGCGGTGCCG GAAACGCTTG TGGAGTCGCA  
TCTTCCTGAG GCCGATCTCG TCGTCGCCCT CTCAAACTGG CAGATGCAAG GTTACGATGC  
GCCCATCTAC ACCAGCTGAAG CTTATCCCAAT TACGGTCAAT CGCGCGTTTG TCTCCACGGA  
GAATCCCGACG GTTGGTACTT GCCTCACATTT TAATGTGTGAT GAAAGCTGCC TACAGGAAGG  
CCAGACCGCA ATATTTTTTG ATGGCGGTTAA CTGGCGGTTTT CATCTGTTGGT GCAAGCGGCG  
CTGGGCTGCT TACCGCCAGG ACATGCTTTT GCCGTCTGAA TTTGACCTGA GCCCATTTTT  
ACGCACCGGA GAAACGGCGG TCCCGTGTAAT GTGGCTGCGT TGAGTGAGC GCAAGTTACT  
GGGAAGATCG GATATGTGGC GGATGACGCG CATTTTCCCG GACGTCTCGT TGCTGCATAAA  
ACCGACTACA CAAATCAGCG ATTTCGCTAT TGCCACTCGC TTTAATGATG ATTTTCAGCG  
CGCTGTACTG GAGGCGTGAAG TGCAAGATGG CGCCGAGTGT CGTGACTACC TACGGGTAGA  
AGTTTCTTTA TGGCCAGGTT AAACCGCGGT GCCCGCGCCTT GTCGGGCTGTA  
AATTATCGAT GAGGCTGTTTG GTTATGCGCA TGCCGTCACA TACGTCGTA ACGCTCGAAA  
CCCGAAACTG TGAGGCGCGC AAATCGCGAA TCTCTATCGT GCAGGTGTTTG AACTGCAAC  
CCCGACGCG ACACGCTGTTG GACGAGAAGC GTGGTCTGCC CGTTTCGCCG AGGATCAGAT  
TGAAAGTGGT CTGTGCCTCG TGGACGCGCA GCCGGTCGTTG ATCGACCGGG TTTAAGCGCT  
CGAGACTCAT CACTCTCAGT GCAGACGGCT TGACGAGGTC AGTGAATGCC AAGTAGATCT  
GCTGATGAGG CAGAACAAACT TTAAACGCGG GCCGCTTGTGC CATATCCCGA ACCATCGCGT  
GTTGGTACAG CTTGCGGACC GCTACGGCCT GAATGTGGTG GATAAGCGCA ATAATTGAGAC  
CCACGGCATG GGGGCGATCG ATCGCTGACG CGATGATCGC GCCTGGCTAC CGCGGATGAG  
CGAACCGGTA ACACCGAATGG TGCCAGCGCA TCCTAATACAC CGAGGTGTGA TCAGCTGCTC  
GCTGGGGGAAT GAAATCCGCG ACACCGCTAAA TCACGACCGC CTGTATCGGCT GGATCAAATC
TGTCGATCCT TCCGCGCGG TGCACTATGA AGGGCGCCGA GCCGACACCA CGGCCACCGA 3600
TATTATTTCG CCCATGTACG CGCGGCGGGA TGGAGACGAG CGCCGCGCTGA TCCCTTGGCA 3660
ATGCCATCTGAAAATAAGGAC TGGTCTACGGT GGGAGAGCTGGT AATATTGCCG ATGCCGTTCC 3720
ATAGCCGACC GCGATGGGTA AGACCTTCTGG CCGTTCGCTA AAAATCTGGG AGGGCTTTGG 3780
TCAGTATCCG GTTTACAGGGG CCGGCTCTGG CTGGAACCTG ATTGGACATG CGCTGATTAA 3840
GCTATGATGAA AGCGGCCAAC CGTGCTGCAC TTTACGCGGT GATTGTTGCG ATACGCCGAA 3900
CGATGCAGACG TTTGGTATGA ACGTCTGGCT CCTTGGCGCC ACCAGCCGGCC AGTCCACGCT 3960
GACGGAGGCC AACACCGGGC AGAACATTTC CAGTTCGCTA TTATCGGGCC AAACACACGTG 4020
AGTGACCCAGC GAATACCTGT TCCGTCAAGC GTAACAACGAG CTCTGCTACT GGATGTTGCG 4080
GCTGGATGCTT AAGCCGGCTTG CAGCGGGTGA AGTGCCTCTG GATGTCGCTC CCAAGGTAA 4140
ACAGTTGATT GCACTCCTTG AACCTACGCA GCCGGAGAGC GCCGGGCAAC TCGGGCCTAC 4200
AGTACGGCTTA GTGCAACCCGA AGCGCGCAGC ATGGTCAAGA GCCGGCAACA TCACGGCCTG 4260
GCACGAGTGG CGTCTCCGGGC AAAACCCACG TGTCAGCGCTC CCCGCCGGGC CCCACCGCAT 4320
CACCAGTCG ACCACCCGGC AAATGAAATT TCTGATCGAG CTGGGTATATA AGCGTGGACA 4380
ATTAGGGCAG CAGTCAGGCT TCTTTTACCA GATGTTGGATG GGGGATAAAA AACAACGTCT 4440
GACGCCGCTG CGCCATACGT TACCCCGCTG ACCGCTGGAT AAGCAGATAG GCGTAAGTGA 4500
AGCGGACTGC ATGGACCCCTA AGCGCTCAGG CCGAAAGGCG AAGCGCGCGG GCCCATACCA 4560
GGCCGAGGCA GCCGTTGTCAG ATGTCACCGC AGATACACTT GTGATGCAGG TGCTGATTAC 4620
GACCGCTCAC GCCTGGCGAG ATCAAAGGCA AACTTATTT ATCAACGGGGA AAACACTACG 4680
GATGATGGCTT AGTGGTCAAAG TGGCGATTAC CGTGATGTTT GAAGTGGCGA GCGGATACCC 4740
GCATCCGGCG CGGATTGGGC TGAACCTGCGA GCTCGCGCAC GTAGCGAGAG GGGTAACCTG 4800
GCTGGATATT CCGCGCGGGA AAAACTATCC CGAAGCCTTCT ACTGCGCGCT GTTTTGGACCG 4860
CTGGGAATCG CCATTGTCAG ACCAGTATGC ATCCGTCAGC TCCCGGCGCG AAAACGGTCT 4920
GCCGCTCGGG ACGCGGGAAT TGAATTATGG CCCACACAG TGCCGGCGCG ACTTCCAGTT 4980
CAACATCAAGC CGCTTACAGTC AACGACACTG GAGTTGAAAC CGACATCGCC ATCCTGCTCA 5040
GCGGGAAGAA GGCACATTGG CGAATTATCGA CGTTTCCCAT ATGGGGTTTG GTGGGGACGA 5100
CTTCTGGAGC CGGTACAGAT GCGGAAAAAT CGACGTGACG GCCGGTCTGG ACCATATCCCA 5160
GGCTGTTCTGG TGTCAAATTTA AATATTACCA GGGACAAGGG AGTCCGGAAG GCCGGGACAGC 5220
AGTGCAGTGG TGGAGAGAAA GCAAGTGAAC TAAAGGACGA GGCTCCTATAA AAGGACTTCA 5280
GCCGACAAAG CAAACCTTTG GGGTTAAATA CAAAGCTCTGT AAATGGTTA AAAAAAAAAG 5340
TCTACGAGGA CAGCGAGATA GCTCTGCGCA CTGTACAGAG CAAATTACAG ACAAGAGAAA 5400
CTGTTGACAT CTGCAGAGAA AGACCAAGAG TGCTGGGCT TAAAAGAATC AGATGGCAGA 5460
TCTACCGGCG CAGGATCCTT CAAGAGCAAT GATCCGTACA GTCTGAGAGC TATCGAGTTA 5520
TAGACAAATT AAGACTGCTA AAAAAAAAAA TGTATAAAAT AGTAAAAACT GAAAAAGAA
AACTAGTCT CTCATGAGAA GACGAGACTG AGACTACTGT AAAAAAGTAC TTTACTGGAA
5 AAAATATGTG TATGAAATACC TTCTTAGTAAAA TGGAGACGTT CTCAAGATGTT ATAAAGACCTT
TTCTTGTAAA AACGAGACTG ATCGAGATGT CATCAAAGAG ATTTGAAGA AAAATTTCCTC
5700
AAGGTTGCGA GTGCCCAAAAG CAATAGTGTG AGATAATGTT CTCGCTTCTT TTGCCCCAGTT
5820
AAGTCAGGCGT TGGCCGCAAGT ATTTAGGAGT CAATGAAAAA TTCAATTGTG TGTAACAGACC
5880
TCAGAGCTCA GAAAGATATA AAAAAATATATA ATAAAACCTTT AAACAGACCCT TGACAAAATTT
5940
AATCCTATAGAG ACTGGCCACAG ACTTACTCGG TACTCTCTCC CTCCTGCCCCTA TTAGAAGCTG
6000
AGAATACCTC CTTCTGGATTC GGTTTTAATCT TTTTTAAAGAT CTCTTTATGGC GCTCCCTAATG
6060
CATCAGCTGT TTTTTAATGTG TGTGTTAAAC TTGGTTAATA TAAATATGTG CTAATAGTTA
6120
AGTTAAAAGG GTTGCAGGTG GTGCAAGAAC AAGTCTGTGC ACAATCGGCT ACAGTGAACA
6180
AGCTGGGTAC CCACAGGACG TCTTACCCGT TCAGCCGAGA GATCTGATCT ACAGATCCCCCG
6240
GGTTCGACCC GGTCGACCCCT GTGGAAATCTG GTCTCGATGG GTGTTGAGAA GTCCCAGACCC
6300
TCCTCGAGCG CCAAGAGTAG CAAAGAGATCT ATCTATCAATT AGTCTAGAGCC CAGGTTGGAG
6360
AAGTCTCGCCAG GTCCGCTGGAAG AGCAGAAGACT AGGATACAGAA GTATCTCAAA TTAGTCGACA
6420
ACCATACTCC CGGCCAAAAT CGCCACCTAC TCCCAGCGAT CCGCCCTCAA CGCCCGCGATC TTGGCCCAT
6480
TCCGCGCCC ATGGGCGTACT ATATTTTTTTT ATTTATGAGC AGGCGCGAAC GCCTCGGCCC
6540
TCCTGACATGT CTCGGAAATA GTGAGGAGGC TTTTTTGAGG GCTTAGGCTT TPCCAAAGAG
6600
CTTCACGCTG CGCCGACGCGT CAGCGGCGCA AGGGCTCGTA AAGGAGCGG AACAGTGAGA
6660
AAGCCAGCTGG GCGCAAGAGCT CCCTGACCCAG GAGCAAGATGT CAGCTACTGG GCTATCTGGA
6720
CAAGGAAAAG GCGAAGCGCA AAGAGAAGAC AGGTAGCCTTG CAGGGGCTCT ATCATGGGCTAT
6780
AGCTAGACTG GCCGGGTATA TGGACAGCAA GCGAACCGGA ATTTGCGACGTGGGCGCCCT
6840
CTGGTAAGGTT TGGGAAGGCC TGCAGAACTAA ACTGGAATTCC TTTCTTCCGC CGCAAGATCT
6900
GATGCGCGGAC GGAGATGAAG TCGTATCAGG AGGACGATGG AAGATGTTGG CATGATGGT
6960
AACAGAGATTG ATTCAGAGCCA GTTCTCTCCG CGGCTTTGAGG GGAAGCTCTTA TCCGCTATGG
7020
ACTGGGCCACA ACAGACAATC GCCTGCTCTG ATGCGCGCCGT TGGCGGCGTAC TCGCCAGG
7080
GGCCGCCGGT TCTTTTTTGTG AAGACGACC GTTCGCGGTC CTCGTAATGCA CTGCAGAGGC
7140
AGCCAGCGCG GTCTATGCTGG TGCCGCAAGA GGCGCGGCTCC TGGCCGAGCT TGCTCGGACG
7200
TTGTCAGCAG GACGCGGAAGG GACTGGCTGC TATTGGGCGA AGTGGCGGAGG CAGGATCTCC
7260
TGTCACTCTA CTTGGCTCTC GCCGAGAAAAT TTCCCATTCA GCGTATGCGA ATGGGGGCGG
7320
TGCTACCGCT TGATCCGCGGT AACTTCTCCGT TGCACTCGAC AGCGAAGAC CAGCATCGAGC
7380
GAGCAGTACG TCAGATGGAA GCCGGTGTCTG TCGATCGAGA TATTGACGGA GAAGAGACTC
7440
AGGGGCCGTC GCCAGCGCGA GTTGGCGCCA GCTGCTCAGC GCCGCGGCAC CGCCGCGGAGG
7500
ATCTGGTGCAT GACCCATGCG AGTGGCGCTCT TGCCGAAATAT CATGGGCGGAA AATGGGCCGCT
7560
TTTCTGGATT CATCGACTGT GCCCGGGCTGG GTGTGGCGGA CGGCTATCCG GACATAGCCT 7620
TGCTACCCCG TGGATATTGCT GAAGAGCTTG GCCGCCGAAAT GGCTGACCCG TTCTCTGCTSC 7680
TTAGCAGTTAT CGCCGCTCCC GGATCCGACG GCAATGCCCT GATATCCGCTT CTGAGACGAT 7740
CTCTCGACG GGAGCTCTGG GGTCCGAATA AGCGAACCAC ACCGCCACCA AGCAGCACAC 7800
ACGAGATTC GATTCCACCG CGCGCCTTCTA TGGAAAGGTTG GGCTTCGGAAG TGTTTTTCGG 7860
GGACCGAATT CTAGATCTGC TGCTTGCAAA CAAAAAACC ACCGCTACCA GCGGTGGGTTT 7920
GGTGGCCGGA TACAGAGCTA CCAACCTCTT TTCCGAGAAT AACTGGCTTC AGCAGAGCCG 7980
AGATACCAAA TACTGATCCTT CTAGGTCAGC CGTAGTTGAG CAACCCTCCG AAGAACCTCTG 8040
TAGGACCCGC TACGATACCTC CTCGACGTTAA TCCTGTTTACG AGTGGCTGCT GCCAGTGCCG 8100
ATAAGTGCTG TCTTACCGGG TTGGACTCAA GACGATAGTT ACCGAGTAAAG GCAGCAGGCT 8160
CGGGCTGAAC GGGGGTTCG TGCAACACAG CCAGCTTGGG GCGACAGACC TACRACCAAC 8220
TGAGATACCT CACAGCTGAG CATTGGAATA GCCCGACCGT TCCCGAAGGG AGAAAGGCGG 8280
ACAGGTATCC GTGTAACGGGC AGGCGCGGA CAGGAGAGCG CAGCGAGGGG CTCAGCAGGG 8340
GAAAACGCTG GTATCTTATT AGTCTGTCTGC GGTTCGCCCA CCTCTGACTT GAGGCTGCGAT 8400
TTTCTGATCT CGCTCAGGGG GGCAGGACGC TATGGAAAAA GCACCAACAC GCCGAGATGC 8460
GGCGCCCTGA GTACACTGCG GTATGCTGA GACCCTCAAG CCTCACAATA AGGTCCCTGTG 8520
CCTAGTTCTG TTTAATATCC TGGTCCCATC TGTTTTGATT CCCATGTTAA AGATAGATGA 8580
AAATCGATAT TCTCCACATA GAGATATAGA CTCTTGGAAT TCTAAGATTA GAATTATTTA 8640
CAAGAGAAAG TGGGAA 8657

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6359 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xii) SEQUENCE DESCRIPTION: SEQ ID NO:15:

TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGGCCACC 60
CCCTCCCACT TGGAAAACAT ACTGGAAAAA AACATTTTCT GGAACAACCA CAGAATGTFTT 120
CAACAGGCCA GATGTATTGC CAAACACAGG ATAGACTCTT TTGGTGTAGT AAAATTGGG 180
TTGTTAAAAT TCCCTTATTCC CTCGCCCACT CCCCCCCTCC GTTTGTGATA TTTTCCCTTA 240
AAAGCTTCTT GAAAAATGGA GTGCTCGTGC AGACTCCCTC ACCCTGTGCA AAAGGTGTAG 300
AGTTTCGACC CCAGACTCTT GTGTGCTTTC TTTTCGGAAC CCCAGAGCTC 360
TGTCCTGCTGT GCTTTGATGTT CATGCGTCTA TCTATATCTT ATTCTACAT AATATGTA  420
GTTCTGATGCT CTCTCTTGAG ACCGGCGTGT CCCGCTGACTGT GAGTGTGGAT AGGAGGCTT  480
TCCCCCCAGT GTTCTCTCATG TGTTATCTGG CCCGAAATTT TGGAGATTCT CATTGTTGCT  540
CATTGGCCCT GAAACATGAA AATTTCTTTT TGTGCTGATT GGGCAGAAGA ACCGCGAGCA  600
CCGAGCTGTC CTAGAGCCAC TTAGAGGTTA GATTTTCTCT TCTGTTTTGG TCTGATGTCT  660
GTGTTCTCTGT GTCTGTTCTC TGTTCATAGG TCTGTCGTTG TCAGTGTTCC AGTTTTGCGG  720
AGCTGTCAGT AGGCCGCGC TCGAGGCGGG GTGCGGCGGG GATAGAATG ATCGTCTTCA  780
GGTGTCACCC GTCCGTTGCG CCTGGGAGAG GTCCCGGAGA GAACAGGGGA GAATCGAGGA  840
CGCTGTCCTGG ACGCCCTTCA AGGCCCAAGA AGCAATTTTG GGGGACGAGA CTGGCGTTTG  900
AGCTCCACCT CCTGCCCGAT CGGAGAGGCT TTGTTGTCAG TCCACCCCTC TGTTTTGTTG  960
CGAGATCTGG GTGTTGACGT CCACCTCGTG TCTGTCACGG GATCGGCTGG TCTGAGTCCC 1020
ACCTGTCGTT TTGTGCGAGA CGATGTTGCTT CTGCCCGGTCG TCGCGAGGRT 1080
CCTGGGCTCG AGTCCCACCT CGTGCAGAGG CTCTCAATTG GGGCAGCGTA GAGAGCCCAT 1140
CTGATCTTTC TGGTTTCTGAC TTTGTCCTTTA GTCTGCTGTC GCCTCTTCTT GTGACTCTG 1200
TTTCTTCTAAA AAGGGGCAA AATCTGTCGA CTCCTCTCTTC TCTGACTCTT GTCTGCTGC 1260
TTGGATATTG TGGTGTGTGA GAGGGTGCTT TGAGTCTGAT CTAGTTGTTG TGTTACTATC 1320
TTGTTTTTTG TTGTGGGTTA CGTTCTCTAG TGTGTCCTTG TGTGTTCTTT TGTGTCGACA 1380
TTGGACTGTA TGACTGAGCA CGTTTTTAAA GTTATGCGTT CTTAAAATAG GCCAAATATC 1440
CTGTCAGATC CCTATGCTGA CCACTTCTCT TCAGATCAGA AGCTGCTTTCT ACTGAGGCT 1500
AAAGCTCGAAA TTCTCGAGTC GACGGTACCG CCGGCGCTAA CTAATAGCCC ATTCTGCAAG 1560
GTACGGAGGG GGGATCAATG CCCGCCGCC TCTGACGATA CTGGCAGGAAG CCGCTTGAA 1620
TAAAGGCGGG GTGGGTTTGG CTAGATGTTA TTTTCCACCA TATGGCGGCT TTTTGGCAAT 1680
GTGAGGGGGC GGAACCGTGG CCGTCTCTTC TTGGACAGCA TCTCTAGGGG TCTTCTCCCC 1740
CTGCGCAAAG GAATGCAAGG TGTTGTGAAT GTGTCGAGAG AAGCAGGTCC TCTGAGGCT 1800
TCTTGAAGAC AAAACAGGCG TGATAGCGCC CTTTACGGCAG ACCGCGACAG CCCACCGGCC 1860
GACACGGTGCC TCTGGGCGCA AAAGCCACGT GATAAGATA CACCTGCAA GGGCGCAAGA 1920
CCCCCAGTGCC AGGTTGCTAG TGTTAGATGT GGGAGAGGAG CTAAAGGTCT CTCTCAGAGC 1980
GTATTCACCA AGGGGCTGAA GGATGCCCCG AAGGTATCGA ATCTGATCCTG 2040
GGGCGCTGGT GCACTGCTTT GACATTGTTA TAGTGGGTGG TAAAAACCG CTTAGGCCCC 2100
CGCCACACCG GGGGCTGTGG TTTTCTTCTG AAAACAGGAT ACGCGGACTCA CCGTGCACCA 2160
CCATGGGTAA AGGGAGAGAG CTATTACAGA GAGAGGGCTCC AATTACAGTT GAATTTAGATG 2220
GTGATTTTAA TGGGCGACA AATTTGCTGT GCGAGGGAG TGGAGGTGTG GCAAATACCG 2280
GAAACTTATA CCTAAATTTT ATTGTGCTCA CTGGAAACAT CTGGTGTCGA TGCCCAACAC 2340
TTGCTACCTAC TTTCTCTTTG GGTGTCAGAT GCTTTTCAGA ATACCCGAGA CATATGAAAC 2400
GGCATGACTT TTTCAAGAGT GCCATGCCCG AAGGTTATGT ACAGGAAAGA ACTATATATTT
TCAARAGATGA CGGGACTAC AAGACAGCTG CTGAAGTCAA GTTTGAAGGTT GATACCCCTTG
TAAATAGATG CGAGTTAAAA GGTATTTGAT TTAAGAAGAA TGGGAAACTT CTTGGAACACA
AAATTGAATA CAACATAAAC TCACAACATG TATAACATCAT GCCGACAAA CAAAAGAATTG
GAAACCAAGT TAATCTAAA AATTAGACAA ACACTTGAAG TGGGAGCCTT GAACTACGAG
ACCATTAC ACAAAATACT CCAATGCGG ATGGCCCCGT CCTTTTACCA GACCAAACATT
AATCTGACAC ACAATCTGCC CCTCCGAAAG ATCCCAACGA AAGAGGAAGG CACATGGTCCC
TTCTTGAGTT TGTACACGCT GCTGGGATTA CACATGGCAT GATGAATCA TACAAAGTCG
GATCTAGATA ACTGCTATCGA TGAGCTCGAA GCCGCGGACA GCAAGTGCA GGTGGAAGA
AAACGAGTGA TCTAGGGCAG CAGCCTCCTCT AAAAAAGCTT CAGCCCCACA AAGCAAACTT
GTGCGTTGCA TACAGATCTC GTAAAAATGTA AAAAAAAAAGTCTACCA GACAGACAGT
AGTCTCCTGC CACTGTACAG AGCAATATAC AGCAAAAGAG AACTCTTTCG ATCTGGCAAG
AAAGACCTAA GATGGCTGTG CTAaaaaGAA TCAGATGGCA AATCTAACC CCGAGGCATC
CTAAAGACCA ATGACTTCGA CAGTCTGAGA ACTCATCAAGT TATAGAACAA TTAAGACTGG
TAAATGAAAA TGCTATAAAA ATAGTaaaaT CTGAAAAAG AAAAAACTGTC CTCTCATGAG
AAGACAGACC TGACATCTAC TGAAAAATAG ACTTTACTGG AAAAAATATG TGATGAATA
CCTCTAGTT TTTTGTGACG TTCTCAAGAT GGATAAAAGC TTTTCTGTGG AAAAAAGGAC
TGATCAGATA GTCATCAAGA AGATGTGTAA AGAAATTTTT CAAGGGTTCG GAGTGCCAAA
AGCAATATGG TGACATTAATG GTCTGTCCTT TGTGTCGCCAG GAAAGTCAGG GTGGGCGCAA
GTAATTAGAG GTCAAAATGA AATTTCCTTT GTTTGACAGA CCTCAAGACT CAGAAAAAGAT
AAATAAAGAA TAAATAAAC CT TAAACAGAC CCTGCAAAAA TTAATCCTAG AGACTGCCAC
AGACTTACTT GGTACTCCCT CCCCTGCGCC TATTTGACG TAGAAATACT CCCCCTTATG
TCGTTTTTAC TCTTTTAAAG ATCTCTTTATG GGCTCTCTAT GCCCATCAGT TCTTTAATGA
TGTGTTTAAA CCTATGTGTT TATAATAATG ATCTATATGT TAAATATAGG GCCGTGCAAG
TGGTGCAAGAA AGAAGTCTGG TCAACAATGG TCAAGTGAAG ACACCCAGGA GAACTGGGCT
CATCTTACCA GTCCAGCCCA GAGATCCTGG ATCAATGATG CGAGGTCGCC CCGGTTGCGG
CTGTGGAAATG TGTTGTCAAT AGGGTGTGGA AAGTCCCCAG GCTCCCTCAG AGGCAAAGGT
ATGCAAGACA TGACATCTCA TTAGTCAAGA ACCRGGTGCT GGAATGACC AGGCCCTCCC
GGAGGCAGAA GTATGCACAG ATACATCTGC AATTAGTCAG CAAACCATACT CCCGCCCTTA
ACTCCGGCCA TCCGCGCCCT AACTCCGCCG AGTCCGCGCC ATTCTCCGCC CCATGGGCTGA
CTTATTTTTA TATATTGAGC AGAGGCGCCAG GCCCGTCTGG CTCCTGAGCT ATCCACAGAG
TAGTGAGGAG GTTTTTTTGG AAGCCTAGGC TTTGGCAAAA AGCTTCAGGC TGCCGCAAGC
ACTCAGGGCG CGAACGGCAG TAAAGGAAGC GAAACACGTA GAAAGGACGTC CGCGAGAAC
GGTGCTGACC CCGATGAAAT GTCAGCTACT GGCTATCTTG ACAAGGGGAA AAGCGAAGCG 4440
CAAAGAGAAA GAGCTGAGCT TGCAAGTGCC TTACATGGGC TTAGCTAGAC TGCGCGGTGTTT 4500
5 TATGGACAGC AAGCAAGCCG GAATTGCGAG CTGGCGCGCC CCTGTGTTAAG GGTGGGAAAGC 4560
CCTGCAAGTG AAACCTGGATG GCTTTCTTGC CGCCAAGGAT CTTGATGCGC AGGGGAATCAA 4620
GATCTGTACA AGAGACAGGA TTAGGATGGT TGGGACTGCA TGAACAGATG GGATTTGACG 4680
10 CAGGTCTCCT GCCGCCTTGG GTTTAGAGGC ATGTCCGCTA TGACTGGGCA CAACAGACAA 4740
TCGGCTGCTC TGATGCAGCC TGTTCCGGCC TGCTACGGCA GGGCGCGCCG GTTCTTTTGG 4800
15 TCAAGACCGA CCTGCTCCGG GCCCTGAAATG AACCTCAGAG CAGAGCAGCG GGGATACGT 4860
GGCTGCACAC GACGCGGTTT CCTTGGCGAG CTGTTGCTCA GGTGACTCAAC AAGAGGGGAA 4920
20 GGGACGTGCTG CTATTTGCGG GAAGTGCGGC GGCGAGATCT CCTGTGCTAC CACCTTGGCTC 4980
CTGCCGAGAA AGTATCCATC ATGGCTGATG CAATGCGGCCG GCTGCACACCG CTGTGATCGGG 5040
CTAACCCTCC ATTCGCACAC CAAGCGAACG ATCGGATCGA CGACGACATG ACTCAGGTGG 5100
25 AACGCGGCTT TGTCCTACGAG GATGATCTGG AGCAGAGACA TCAGGGCCCT GGGCCAGCGC 5160
AACCTGTCGC CAGGTCAAGC GCGACCGGCC GAGATCTCGTC GTCAGGGAATG 5220
GGAGATGGCT CGTGTCCGAA ATCAGGGGTT AAAATGGCCG CTTTTGCTGA TTATCAGGAC 5280
30 GTGGCGGGGGT GGGTTGGGGC GACCCCGACG AGGACATAGC GTTTGGCTAC CGTGATATTG 5340
CTGAAGAGCT TGGCGCGGAA TGGGCTGACG GCTTTCTGACG GCTTTACGGT ATCGCGCCCTC 5400
35 CCGATTCCGA GCCGATCCCG TTCTATCGCC TTCTTGACGA GTCTTTGACGA CGGGGACTCT 5460
GGGGTCGCAA ATGACGGAGA AAGCGACGCC CAACCTGGCA TCAGGAGATT TGCAATTCCAC 5520
40 CGCCGCCTTC TATGAAGAGT TGGGCTCTGG AATCGTTTTT CCGGAGGCAA TTCTGTAATCT 5580
GCTGTGTTCA AACARAOAOA CCACCCCTAC CAGCGGTTGGT TTGGTTGGCC GATCAAGAGC 5640
TACCAACTCT TTTTCCGAAG GTAACCTGGG TCAGCAGAGC GCAGATAACA AATACTGGCC 5700
45 TTCTAGTGAC GCCGATGTTA GCACACCATC TCAGAAACTC TGAGTACACC CGCATGACCAC 5760
TCGGCTGCTT AATCTGCTTG CAGTGGCTGG CGAGATCTCT GTCCTTACCGGC 5820
50 GGTGGGACTGC AAGCGATAGA TTACCAGGTA AGGGCGAAGG GTCGGCGCTG ACGGCGGGTT 5880
CGTCACACA GCCCGCCTTG GAGCGAAGCA CCTACACGGA ACTGAGATAC TCACGCCGGT 5940
AGCATGGA GAGCGCCACG CTCTCCGAAG GAGAAGAGGC GACAGCGTAT CGGTAAGCG 6000
55 GCAGGGTCGG AACAGGAGAC CGACAGAGCG AGCTTTGCGG GGAGGAGCGC TGTATCTTTT 6060
ATAGTCCTGT CGGTTTCCGC CACCTCGACG TGAGGTCTCG ATTTTTGTGA TGCTGTCGAC 6120
60 GGGGGCGGAG CCTATGGGAA AAGCGCCAGC AGCCGGCGATG GGCAGCGCCG CAGTACACCT 6180
GCGTCATTGC GAGACCTCGA AGCCTGACTA AAAGGGTCCC TGCGTGATTT AATAGATGAG 6240
TCTGGCTTTAT TCAGTTTTTT TCCCGATTGG AAAGATAGAG TAAATGACGT ATTTTCCAAA 6300
65 TAGAGATATA GACTTGCGA ATTCATAGAT TAGAATATT ACAAAGAAGAG ACTGCGGGAA 6359

(2) INFORMATION FOR SEQ ID NO:16:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6891 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xii) SEQUENCE DESCRIPTION: SEQ ID NO:16:

TGAAGAATRA AAAAAATCTG GCTCTTCTGT AGAACATGAA CTTTCACCTC GGAGCCACCC
CCCTCCCACCA TGGAACACAT ACTTGAGAAA AACATTTTCT GGAAACAACCA CAGAAATTTT
CAACAGGGCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTCTAGT AAATTTGTTG
TTGTAAACTC TCCCTCATTG CCTCCCCATTT CCCCTCCCAC ACCTTGGTTT TTTTCTTTTA
AAAGCTTGTG AAAAAATTTG GAACGTCTCG AGACTCTCTT ACCTCTGTCA AAGGTGTATG
AGTTTCGACC CCAGACGCTT GTGCTGGTTC TGGTTGCTGCT TTATTTCCAC CCCAGAGCTC
TGCTCGTGCT GTCTCTCGTG CTCCTTACGG TTTAACTTTA CTTTTTCTAT TTTTTATTTG
GTCTCAGTGT CTCTTTGCTG AGCCGGCTGT CCCCAGGCTT GATGTCTGTA GACGAGGCC
TCCCTCGGAG GTCTTTTATT GTGTAATGAG GGGGAGGATT CAGAAATATT TCATTTGGTG
CATTGGGGGG GAATTCGAAA ATCTTCCTCT TGCTGCTATT GGGGGGAAAC AGCGGACCA
CCAGACGGTC CTAGACCCAC TTAGAGCTAA GTCTTTTGT TCTGTTTTTG TCTGATGTCT
GTGTCTGTAT GTCTGTCTCC TGTCTTTCAAG TCTGTTGCGA TGCCAGTACC AGTTTTGCGG
AGCTCAGTGT AGACCGCGCTT CGAGAGGGGA GTGCGGGGTTT GTAAGAGATA GACGGTGCTCA
GGTGGCCACCC CTCGCGTCCGG CCTGGGAGGAT CTGCCAGGAG GAAACGGGGG GGATCAGGGA
GCCTCGGTGG ACCCGCGTGG ACCCGTGGGG ACCCTTTTGA AGGGCAAGG ACCATTTGGA GTCGGGAGAT CGTGTTGCTGC
AGTCCGCCAT CGTGGCCAGTT TGCCAGATCG TGGGTTGAG TCCACCTCC TGTCTTGTGG
CGAGATCGTG GTTGGTCGGA CCACTCGCCG TCGTTGCGAG GATGTTGCTGG TCGAGTCCCA
ACCTCGTGTT TGTTGGGATT ATCGTGGGTT CGAGTCCGAC CTCCGCGTCTG GTCACGGGAT
CGTGTTTCTCG AGCTCGACAG GTGGATCTAG TGCCGGGATTT GAGGGGCTTA GAGAGGCCAT
CTGATTTCTCC TGGTTTCCTC TTTTTGTTTA GTCTGCTGTG TGTTGCTTTA GTGAACACTG
TTTTTCTAAA AATGGGAAAC TCTGTTGCTCA TCCCTCTTTC TCTAGCTCTG TGGCTGTGCG
TTGGTAATTG TGGTTTTGTTA CGTGGGCTTG GTGCTGCTTG TGGCTGCTTT TGGTTGTCGA
CTGGGACTCA TGACTGACCA CTTGTTTATA GTTAGTCTTT CTAATATAG CGTAAATATCT
CTGGCTGATC CCTATGCGGA CCACCTCCCT TCACGATCAAC AGCTGGGCTT ACTGGAGCTC
AAGCTTGGAA TTCTCGATCG CAGCGTACCG GCGGGATGCAA TTCCGGCCCCC CCCCTAAACGT
TACCTGCCGA AGCCGCTTGG AATAGGGCCG GTGTGCCTTT GCTCATATGT TATTTTCCAC 1620
CATATTGCG TCTTTTGGCA ATGTGAGGGG CGGAAACCTT GGCCTGTGCT TCTTGCAGG 1680
5 CATTCCTAGG GTCTCTTCCC CTCTCGCCAA AGGAAAGCAAA GTGCTGTGTA AGTCTGTGAA 1740
GGAGACCTTT CTCCTGGAAG CTCTCTTGAAG ACAAACACG TCGTACGGA CCCCACGACG 1800
GCAGGCGAAC CCCCAACCTG CGGACAGTGT CCTCTGCGGG CAAAGGCAAC GTGTATAAGA 1860
TACACCTGCA AAGGGCGCAC AACCCCGCTT CCAGGTTGTTG AGTGGAGTAG TTGTGGGAA 1920
AGTCAAATGG CTCTCTCACA GCGTATTCAAA CAAGGGGCTGG AAGGATGACC AGAGGGATT 1980
5 CCATTGATAG GGGATCGATC TGGGGCCCTG GTGCAACATGC TTACATGTTG TTATTGCGA 2040
GTAAAAAAAC GTCTAGGGCC CCGGAGACCAG GGGACGCTGG TTTCCTGTTT AAAACACGGA 2100
20 GCGGGATCAA TTCCGCCCCC CCCCCAAGCT TACTGGCCGA AGCCTGTTTG AATAGGGCCG 2160
GTGTGCGTTT GTCTATAATG TATTTTCCAC CATATTGCGG TCTTTTGGCA ATGTGAGGGG 2220
CGGAAACCTT GCCTCCTGCT TCTGCTGAGA CATTCTTAGG GTTGCTTTCC CTCTGCGCAA 2280
AGGAATGCAA GTCTCTGTTG AGAAGCTGGA GGAACGATT GCTCTGGGAAG TTTCTGGAAG 2340
ACAACCAACGT TCGTACGGCA CCCCACCTGG CGACCAGAAC CCCCCACCTGG 2400
30 CCCCCTGCCC CAAAGGCCAC GTGTATATAAGA TACAATGCCA AAGGGCGCAC AACCCCGCTT 2460
CCAGGTGTGT AGTTGATAGT TTGTTGAGAA AGTCAAATGGG CTCTCCCTCA CGCCTATTCA 2520
CAAGGGGCTG AAGGATGACC AAGAAGTACC CCATTTGATG GGAATGGATC TGGGGCTTGG 2580
35 GTGCACATGC TTTATGATGT TTTAGTGGAG GTAAAAAAA CGTCTAGGGCC CCGGAGACCAG 2640
CGGGAGCGTG GTTTCTCTTT GAAAACACGG ATACCGGATC CACCGTGCCG CACCATGCGG 2700
40 AAAGGAGAAG AACCTTTTAC AAGAGTTGTC CCAATTCTTG TTGAATATAGA TGTTGATAGT 2760
AAAGGGCAACA AAATTCTCTGT CAGTGGAGAG GGTGAAGTGG AGCAGACATA CGGAAACCTT 2820
ACCCTTAAAT TTATGTCAGA TCTGGAAGAC CTCCTGGTGC CAGGCCCACCA ACTGGTCCA 2880
35 ACTTTACATT ATGGTGTTCA ATGCTTTTCA AGATACCACG ATCATAAGCA AGGCACATGAC 2940
45 TTTTTCAGGA GTGCCAGCCT CGGAGGTAT TACAGCCGGAA GAACATATTATT TTATGCAAGAT 3000
GACGGGAACG ACAAGACACC TGCTGCAAGTC AAGGTGGAAG GTGATACCTG TTATGAGATAG 3060
50 ATCGAGTTAA AAGGGTATTG TTATAAAGAA GATGGAAAAA TTCTGGGACA CAAATTTGGAA 3120
TACACTATA ACTCAGCAAA TGTATACATG AGGCAAGACA ACAAAGAGAA TGGAACCAGAA 3180
55 GTTAACTTCA AAATTTAGCA CAACATTGGG GATGGAAAGCG TTCAACTAGG AGACCATATT 3240
CAACAAAAATA CTCGATTTTG CGTTGACCCG CTGCTTCTGAC CAGACAAACA TTACCGTCC 3300
ACCAACATCTG CCCCCCTCTGA AGATCCCAAAC GAAAGAGAGG ACCACATGGT CTTCTGGAAG 3360
60 TTTTGACAG CGCTGCGGAT TACCATGGGC ATGGATGAACT TATACAGTCG GGGATTGAG 3420
TAACACGTATC GATGGAGGAG AAGCGGCGGG CAGCCAGTGCA GTGTGGGACA GAAAGCAAGT 3480
GATCCTAGGCG ACCAGCCTCC GTAAAAAGAC TTCAGCCCAAA AAGGCAAAC GTGCGGGCTTT 3540
65 ATACCAAGCT CTGTAAATTG TAAAAAAA AAGTCATACA CGGACAGCAG GTATGCTCTT 3600
GCCACTGTAC AGAGCAATAT ACAGACAAAG AGAATCTGGT AGACTCTGCAC AGAAGACACT 3660
AAAGATGCTGT GGCTAAAAGA AATCCAGAAGT GAAATCTAATAC GCCCCAGGCA TCCTAAAGAG 3720
CAATGACCTTAC GAGAGCTCTGA AGACTATAACA GTTATAGACA AATTAGAAGCT GTGAIAAAAAA 3780
ACCGCTTATAA AATTAGTAAA AACTGAAAAA AGAAAATCTAG TCCCTCTCATG AGAAGACAGA 3840
CTGAGACCATG ACTGAAAATAG AGACTTTACT GGAAAATATA GTGATGATAA TACCTTCTAG 3900
TTTTTGGAAG CTTTCTCAGA ATGGAATAAA GCTTCTTCCTT GTAAACGAGC ACTGATCAGA 3960
TATTCATCAGA GAGATTGCGT TTGGAATGATG TOKGTTGAGTAC CGAGATGGCA AAAGACAATA 4020
TGTCAGATGAA TGGTCTCGGC TTTGTTGCCC AGTGAAGTCA GGGTGTTGCA GTACATTTAG 4080
AGGTCAAGTG AAAATTCCTAT GTGTTGTCATA ACCTCTCCAGAG GCAGAAAGAA ATGAATAAAGA 4140
ATAAAATAAA CTCTAAACAG ACCTTGACAA GATTAATCATG AGAGACTGGG AGACAATTAC 4200
TTGGTACTCC TTTTCTTGGC CCATTTAGA AACTGAAAATAA AGCTCAGAAGA ATCTTTCTTG 4260
ACTCTTTTTA AGATCCTTTTA TGGGCTCTCT ATGCAATCAGA TGCTTTAATAA GTATGTTATA 4320
AAGCTTATGTT TTATATATAC TGAATCTATAT GTGAAATGTA AAGCTTGCAG ATGGGTGGCA 4380
AAAGAAGCTCT GTTGCACACT GCCTAGCTGT GACAAGCTGG GTACCCCAAGG GCATACCTTC 4440
CGATTCGCCGC CCGAGATCTGT ATCTACGATTC CCGGGTGCAG CCGGGTGCAG CCGGGTGGAA 4500
TGATGTGCTAG ATGATTGTTG GAAATGTCCTG ACCCTCTCTGC TTGATGCAAAG GTATGAAAGA 4560
CAGCAGTCGTC CAACCAGTGG TGGAAAGTCC CCAAGGCTTCC CAGCAGGCTG 4620
AGTATGCACC AGCATGACATC TCAATTAGTC AGCAACCATA GTCCCCGCCCT TAACCCCGGC 4680
CATCAGCGGCG CTAACCTCCGC CCAAGGTCCGC CCAATCTCCAG CCAATCTCCG CCAATCTCCAG 4740
TTTTATTTAT GCAGAGCCGC AGGCCGCCCTC GCGCTCTGAG CTTATTCAGA AGTATGAGG 4800
AGCTTTTTTT GGAGGGCTTAG GCTTTTGCAA AAGCTTCCAC GCTGCCGGAA GCACCTCAGG 4860
CGCAAGGGCT GCTAAAGGAA GGCAGACAGG TAGAAAGCACA GTGCCCAAGG GCAGGTGGCTA 4920
CCCGGATGCA ATCTCAGCTA CTTGAGCTATC TGGCAAAAGG AAAACGCAGG GCAGAAGACAG 4980
AAGCAGGCTGC ATCTGAGTGG CTTAAGTCATC CGTATGCTAG ACTGCGCTTGG GTTATTGCAC 5040
GCAAGCAAC CGGAATTCGA AGCTGGGGGC CCGCTCTGTA AGGTTGGGA GCCTGGCACA 5100
GAAAATGCTA TGCGTCTCTT GCCGCAAGGG ATCTGATAGC GCAGGGGTGC AAGATCTGAT 5160
CAAGAGACAG GATGAGGATG GTCGCCAGATG ATGCAACAGC ATGGATGTCAG CGCAGGCTTT 5220
CCGGCGGCTT GGGTGAGGAG GCTATTCGCG TATGACTGGG CACAACAGAC AATCGGCTGC 5280
TCTGATGCAG CGTGCTCCGC GCTGTCAAGCG CAGGGGCAGG CGGCTGTTTTT GTTCAAGACC 5340
GACCTGCGCG GTGCCTCGGA TGAACTGACG GACGAGGGAG CGGGCTATCG GTGCTGGTCCG 5400
ACGACGGCGG TCTGCTGGGG AGCTGTGGACG AGCTGTTGCA CGTGAAGGCGG AAGGACTGG 5460
CTGCTATTGG GCAGAGTGGG GGCGGGAGAT CTCTGCTGCA CTCTGCTTGC TCCCTAGGAG 5520
AAAGTATCCA TCATGGCGTA TGCAATGGCG CGGCTGCAATA GCCTGATGAC GCCTCATCCTG 5580
(2) INFORMATION FOR SEQ ID NO:17:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6321 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xii) SEQUENCE DESCRIPTION: SEQ ID NO:17:
TGAAGAATAAAAAATTACTG GCCCTCTCTG AGAACAAGAA CTTTCACCTC GGAGCCCACC
60
CCCTCCCCCC TGGAAAAACT AACTTGAGAAA AACATTTCCT GGAACACCAA CAAATGTTTT
120
CAACAGGCCA GATGTATTTGC AAACACAGGG AATATGACTCT TTTGTTGGAG AATTGTTGG
180
TTGGTTAATT CCCCCTATTC CCCCCCATT CCCCCCACC GTTGGTGTATT TTTCCTTTTA
240
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGTGAAGGTG ATGCAACATA CGGAACACTT ACCCTTAARAT TTATTTGCAC TACTGGAAAA</td>
<td>2280</td>
</tr>
<tr>
<td>CTACCTGTTT CATGCGCAAC ACTTGTGCACT ACTTTTCACT TTATGTTCAA ATGCTTTTCA</td>
<td>2340</td>
</tr>
<tr>
<td>AGATACCCCA CATATATGGA AGCCGCATGAC YYTTTTCAGA GTGCCATGCC CAGAGGTAT</td>
<td>2400</td>
</tr>
<tr>
<td>GTACAGAAA GAATATATT TTTTCAAGAT GACGGGAACT ACAAGACACG TGCTGAAATC</td>
<td>2460</td>
</tr>
<tr>
<td>AAGTTTGAAG GTGATACCCCT TTGATAGA AGCTGATTTA AAGTATTTGA TTATTTAAGAA</td>
<td>2520</td>
</tr>
<tr>
<td>GATGGGAACA TTCTTGGACA CAAATGGAA TACCACTATA ACTCACAAA CTTATACATC</td>
<td>2580</td>
</tr>
<tr>
<td>ATGGCCAGACA ACAAAGAAA TGGAACCAAA GTTAACTCTCA AAATAGACAA CAACATGGA</td>
<td>2640</td>
</tr>
<tr>
<td>GGATGGAGCG TTTAAGCTTC GCACATTTAT CAACAAATA CTCCAAATGA CGATGGCCCC</td>
<td>2700</td>
</tr>
<tr>
<td>GTCTTTTCTAC TCTGACAACTA TACCTGTCCC GACCAATCTCG CCCTTTGGAA AGATCCCAAC</td>
<td>2760</td>
</tr>
<tr>
<td>GAAAGAGAGA ACCATCTGGT CCTTCTTGAAG TTATGTAACAC CTGCTGGGAT TACACATGCC</td>
<td>2820</td>
</tr>
<tr>
<td>ATGGATGAAAT TATACAAGTC CGAGTCTAGA TAACTGATTC GATGGAATCG AGGCGCGGGA</td>
<td>2880</td>
</tr>
<tr>
<td>CAGCATGCGA GGTTTGTGACA GAAAGAAGT GCTATTAGGC AGCAGCATCCT CTAAAGGAGC</td>
<td>2940</td>
</tr>
<tr>
<td>TCCAGCCCAAG AAAGACAAAC TCTGGTACTT AATCTAACGCT CTGTATAGG TAAATAAAAAA</td>
<td>3000</td>
</tr>
<tr>
<td>AAAATCTTACA CAGGACGGGA GATGCTCCTTC GCCACTGTTAC AGGAACAATAT ACAGACAAAG</td>
<td>3060</td>
</tr>
<tr>
<td>AGAAACTTGTTG ACGTCTGCGA AGAAGACTCCT AGATGTCTGT GCTTTAAAGA AATCGAGATG</td>
<td>3120</td>
</tr>
<tr>
<td>CAAATCTAAC CGGCGCGCAGA TTCTAAAAAG CAGTATCTG CCCGACTGCTG AGACTATCAA</td>
<td>3180</td>
</tr>
<tr>
<td>GCTATAAGAC AATTATAAGACT GTGAAAAAAA ACCCTGTATA AAATAGTAAA AACCTGAAAA</td>
<td>3240</td>
</tr>
<tr>
<td>AGAAACACTAG TCTCTCTCATG AGAAGCAAGA CTCGACACTT ACTGAAAAAT AGACTTTTACT</td>
<td>3300</td>
</tr>
<tr>
<td>GAAAAAAATA TGATGTTAGAA TACCTTTCTAG TTTTTTGAGA CTTTCTCACAG ATGATTAAAAA</td>
<td>3360</td>
</tr>
<tr>
<td>GCTTTTTCTTT GTTAAAACAGA ACTGATCAGA TATGCTACA AAAGATGTGT AAAGAAAATATT</td>
<td>3420</td>
</tr>
<tr>
<td>TCCAGGAGTT CGAGGATGGCA AGAAGCTAAG TGTCAAGATAA TGGTCTGCCC TTTTTTGGCC</td>
<td>3480</td>
</tr>
<tr>
<td>AGGTAAGTCA GGGTGTTGCA AAGTTATTTAG AGTGCAATATG AAAATTTTCA AGGGTGTTACA</td>
<td>3540</td>
</tr>
<tr>
<td>GAGCCTGAGA CTGGAGGAAAG AAAAATAATAA TAAAATAATAA TTACAAAAGAC ACGCTGCCAA</td>
<td>3600</td>
</tr>
<tr>
<td>AATTAACCTC AGAACGTGGC ACAGACCTTC TGGTACTTCC TCCCTCTTGC CTTATTTTAGA</td>
<td>3660</td>
</tr>
<tr>
<td>ACTGAGAAATA CTACCCTTCTG ATTCCGGATT ACTCTTTTTA AGATCCCTTA TGGGCTCCT</td>
<td>3720</td>
</tr>
<tr>
<td>AGTGGCATCAG TGTCTTTAAAAT GATGTTTTTA AAACATATGT TATATAATAA TGAATCTATAT</td>
<td>3780</td>
</tr>
<tr>
<td>GTTAGGTTAA AAGGGTCTCA GGTGCACAGA AGAAGAATCT GGCACAACAT GCTACAGTAG</td>
<td>3840</td>
</tr>
<tr>
<td>ACAAGCTGG GTACCCCAAG AAGCTCTTTAC CAGTCCAGAC CAGAGATCTG ATCTACGATC</td>
<td>3900</td>
</tr>
<tr>
<td>CCCGGTCTGA CCGGGTGCGA CCTGAGGGAA TTGGTGTGAC TTAGGGTGGA AAGAGCTCCC</td>
<td>3960</td>
</tr>
<tr>
<td>AGGCTCCCC GAGCGCAGAA GTATGCAAGA CAGTGACTCTC AAATAGTCAAG CAACCCAGTT</td>
<td>4020</td>
</tr>
<tr>
<td>TGGGAAGTCC CGAGGCTTCC CGAGCGCAGA AAGATGCAAA AGCCATGACAT TCAATTGCTG</td>
<td>4080</td>
</tr>
<tr>
<td>AGCCAGCATACTTG TCTCCCGCCCA TACTCCCGCC CATC CCCGGCCCCT CAATCTGCCC CGAGTTCCGC</td>
<td>4140</td>
</tr>
<tr>
<td>CCATCTCCTG CCCATGCTGT CACTAATTTT TTTTATTATG TGCAAGGCGGC AGGGCGCTTC</td>
<td>4200</td>
</tr>
<tr>
<td>GCCCTCTGAG CTAATTCAGA AGTAGTGGAG AGGGTCTTT AAGGCCTAAG CTGGTGCAAA</td>
<td>4260</td>
</tr>
</tbody>
</table>
AAAGCTTCAC GCACCTCAGG CCAGACTGGT GCTAAAGGAA GCCGAACAGC
TAGAAAGCACA GCTCCGCAAA AACGTTCTGAA GCCCGGTAGA ATGTACGCTA CTGGGCTTAC
TGAGCAGGGA AAAACGCAAG GCCAAGAGGA AAGAGGTATC CGTGGACCTG CTTTACATGG
CGATGCTGCTT TTATGCGACA GCAACAGGAA CGGAAATTCC AGGCTGGGGCG
CCCTCTGGTA AGGTGCTGGGA GCCCTGCAAA GTAAACTGCA TGGCTTCTT GCCGCGCAAG
ATCTGATGGC CGAGGGGATC AGAGATCTGT CAAGATGACAG GATGAGGATC GTTTGCAGT
ATTGACCAAG ATGGATGGCA CGCTTGGCCT CGTGCGTCCG GCTGTCGACG
CAGGGCCGCC CGGTCTTTTG TGTCAAGACG GACGCTGTCG CGTGCGTCCG TGAACTTACG
GAGGAGGCAG CGCGGTATCTC GGTGCTTCGC ACGACGGGGG TTTCTTCTGC AGCTTTGCTC
GACGTAGTCA GTGGAAGGGG AGGAGGTGTG CGCTCATTTG GGCAAGTTGG GGGGCAGGT
CTCGCTGCTCCT CTCACCTGCT CGTGCGGAGG AAATATCCCC TCAATGGTCG TGCAATGGGC
CGCTGGCCATA CGCTTGATCC CGCTACCTGC CCGCGCCGAGG CGTGGGCTGA CAGGACGAGT
GAGGAGGCAC GTACTCGGAT GGAAGGGCGT CTTCTGGATC AGATGATGCT GGGAGGAGG
CATCAGGGGC TGCCGCCAGC CGAAGCTGTC GCCGAGCTCA AGGGCCGACT GCCGACGGGC
GAGGTATGC TGGCTGACCA TGCCGATGCC TGGTCCGCAA ATATCATGTG GGAAATGGCC
CGTTGCTCAG AATATGACGA CTGTTGGGCGG CTGGTGTGGC CGAGCGGCTT TCAGGACATA
GCCGTTGCTA CCGTGTATGA TGTATTGGAG CGTGGGCGGCG AATGCGCCTG CGTGGTGTC
CTCGGGCTAG GTATCGCGGC CCCGGATTGC CAGGGCATCG CTCTCTATCG CGTCTTTCG
GAGTTCTTCTC GAGGGGACTC CTGGGATTCG AAATGACCGA CCAAGGCCGGA CCAACCTTGC
CATCACGAGA TTTGATCCAC ACCCGGCCTC TCTATGGAAG GGTGGGCTGC GGAATCGTTT
TGGCAGGGGA ATTCGCTAAT CGTCTGCTTG CAAACCAAAA AACCACCGCT ACCGGCGGCT
GTTTGTTCGG CGGATCAGGA GCCAACCACT CTTTCTCCGA AGGTAACTGG CGTCAGCGA
GCCGAGATAA CAAATACGTG TAGCCGTCAG TTTCCGACCA CTAAGCCAGAC
TCTGTACGAG CGCCCTACATA CCTGCCCTCTG CTAATTTGTG TACCAGTGGA GGCGCCAGGT
GGGCGATAGT CTTGTGGCAG CGGCTGTCGA ATAGACGGAT TGGACTGGGA CAAAGCGGAC
CGGGTCGGCTT ACTGTTGACG CAGGAGGCTC GGAACAGGAG AGGGCGAGG GAGCTCTCCA
GGGGGAAAGG CCTGGATCTC TTTAGCTTCC GTGGGGGTTC GCCACCTCAG ACTTGAAGTC
CGATTCTGGT GAGCTGTCGT AGGGGCGGGG AGCTTATGGG AAAAGCGGAG CACGCGGAG
ATGCCGCGCC TGGAGTGACG CTGCGTGCTG CTGAGACCTG CAGGCTACAC TAAAAGGGTC
CCTGCGCTAG TCTGTTTACT AATCTGCTCT ATCTGTTTTT TGTTCGCGATG TTAAGATAG
AGTAAATGCA GTATTCTCCA CATAGAGATA TAGACTTCTG AAAATTCAAG ATTAGAATTA
TTTACAAGAA GAAGTGGGGA A

(2) INFORMATION FOR SEQ ID NO:18:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5754 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

TGRAAGATTA AAAATTAATG GCCCTCTTGTG AAGAACATGAA CTTCACACTC GGAAGCCAAAC 60
CCCTCCCATC TGGAAGCAAT ACTTGAGAAA AACAATTTCTT CCAAACCAACA CAGAAATTTT 120
CAACAGGCCA GATGTATGCT CAACACAGGG ATATACTTCT TTGATGTGAGT AAAATTTGTG 180
TTGTTAACC ACTCCCTATTC CTCCTCCATT CTTCCCTCCCA GCTTTGTGTTT TTTTCTTTTA 240
AAAGCTTGTG AAAAATTGGA GTGGTGTGTCG AGACTTCTCT ACCCTGTGCA AAGGTGATAG 300
AGTTTCGACC CCAGAGCTCT GTGCTGTCTTC CTTATGCTCT TATATCTGAC CCCAGACTC 360
TGGCTCTGTG GTCTCTATGT CCGTGTTTTA TAAATCTCTA CCTCTCATTT TTTATGTGAT 420
GTCTCAGTGT CTTCTTTGGT ACGCCGGCTG CCCCCAGCCTT GAGTGTGCTG GTGAGGCTC 480
TCCCTCGAGG GTCTTTCATT TTGCTATCGG CCGCGGAAAT CGAAGATCTT TCTATGGTGT 540
CATGTCGGCG GAATTCGAAA ATCTTCATTT TGCTGTCTTG GCCGGGAAAAC AGCGCCACCA 600
CCCAGAGGCCCTG CTAGACCCCA TTAGAGCCGA GATTCTTTGT TCTGTTTTGG TCTGATGCTC 660
GTGTTTCTGTG GTCTGTCTTC TTGTTCTTGG TCTGAGCTCAG TCGCAGTTCT AGTTTGCGG 720
ACGCCTCAGTG AGACCAGCCCT TGGAGGAGCA GTGGCGGCGTG GTAAGAGGATG GACGTGGTCCA 780
GGTGTCACCC GTCTCAGTCC CTCGGGACGC GTCAGCAGAG GAACAGGGGA GGTCAAGGGA 840
CGCTGTGGGG ACCCTTTCTA AGGCCGAGAG ACCATTTGGGT GGGCGAGGATC TCGGGTTGCG 900
AGTCCCGACT CGTGGCAGAGG ATGGGTCGAG TCGGCTACCG TCCCCCTTGTG TTGGCGGGTGT 960
CGATATCGTG GGTGGGAGCC CCCACCTCGTT TCGAGACTCA GATGGGCTCG CTGGGAACGTCC 1020
ACCTCTGCTT TTGTGGGCAG ATGCTGGGTTG CGAGTCCCAC CTCGCTCTTC GTGAGGGGAT 1080
CGTGGGCTGG AGTCCACCCCT CTGCGAGAGG GTCTCAATTTG CCGCCGTTTA GAGAGGCCCATT 1140
CTGATTTCCTC GTGGTCTCTT TTTTCTCTA GTCTGTCTGG GTGACTATGCTC 1200
TTTTTCCTAAA AATGGGACAA TCCGCTGCCTA CTCCTTTTTCC TTGACTCTGTG GTCTCTGGC 1260
TTGGATATTTT TGTGCTTGAT AATGGTTTTT TGTGACGCTT GAAATGTTGTC TGGTGATATC 1320
TTGTTTTTGT TGTTGCTTTA CGTTCTTCTGT GTGGTCTCCTG TGTGTTCTCGA 1380
CTGGACTGGA TGACTGAGCA CTGTTTTTA AAATGCGCCT CCTAAAATAAG CCTAAAATATC 1440
CTGTCAGATC CCTATGCTGA CCACCTCCCT TCAGATCAAC AGCTGCCCTT ACTGCGAGTC
1500
AAGCTTCGAA TTCTGCAGTC GACGGTACCG GGCGGCCGGG ATCCACAAGG CTCACAAGATG
1560
GGAATAAGAG AGAAGCTTTT CACAGGAGTT GTCCCAATTT TGGTGAATTT AGATGCTGAT
1620
GTTAATGGGC ACAAAATTTTC TGTCACTGGA GAGGGTGGA GTGATGCAAC ATACGGAAAA
1680
CTTACCCTTA AATTATTTG CACTCTGGA AACACTACTG TTCCATGGCC AACACTTGTC
1740
ACTACTTTCAT CTATGCTGTT TACAGCTCCT TCAAGATACG CAGATCATAT GAAACGGCAT
1800
GACTTTTCTA AGAGTGCCGA GCCGAGCTGA TATGGTACAG AAAGAACCAT ATTTTCTCAA
1860
GATGACGGGA ACTACAAGAC AGCTCTGGGA GTCAGTTGTT AAGGATGATC CTTGCTTAAT
1920
AGAATCTGAT TAAAGATTAT TGATATTTAAA GAAGATGGAA ACATTCTTGG ACACAAATTG
1980
GAATACAACCT ATAACCTACA CAATGTATAC ATCATGCGAG ACAAAAAAAG GATGGGACC
2040
AAAGTTAACCT TCAAAATTAG ACACAAACATT GAAGATGGAA GCATTCAACT AGCACACCAT
2100
TATCAACAAA ATACTCCAAAT TGCGGATGCG CCTGTCTTCT TACCAGCAGA CCATTCTCTG
2160
TCCACACAATT CTGCCTTTAC GAAGATCCCC AACGAAAGGA GAGACCACAT GGTCTCTCTT
2220
GAGTTTGAAA CAGCTCTCTTT GATTACATG GGCAATGATG AACTTACCAA GGCGGACTCT
2280
AGATACTGCT ATCGATGCTAT CCAGAGCGCC GCAGACAGCC GCAGTGTGTTG AACGAAGACA
2340
AGTATCTATCG GCCAGCGAGCC TCCCTAAAGG GACTTCAGCC CACAAAGCCA AACCTGTGCG
2400
TTTAATACAA GCTCTGTAAA TGTTAAAAAA AAAAAAGTCT ACACGGACAG CAGTATGCTT
2460
CTTGCCACTGT TACAGGCAAA TATACAGACA AAGAGAAGTG TGGACATCTG CAGAGAAAGA
2520
CCTAAGATGC TGTTGGGCTAA AGAAATCAGA TGCGCAATTCT AAACCGCCGAG GATCCCTAAA
2580
GAGCAATGAT CCTGACAGTC TGAAGACTAT CAAGTATAAG AGAAATTAAG ACTGGTAAAA
2640
AAAACCCCGT ATAAATAGTT AAAAACTGGAA AAAAAAAACG TAGTCCCTCTC ATGAAAGAGC
2700
AGACCTGACA TCTACTGAAAA AATAGACTTT ACTGGAAAAA ATATGTGATG GAATACCTTC
2760
TAGTGGGTTGT GAACCTGCTC TAGATGGATA AAGCTTTTCT TCTGTAAAAA GAGACTGAC
2820
AGATAGTCTAT CAGAAGATTG GTTAAGAAAA ATTTTCAAG GTTCGGAGTG CCAAAAGCAA
2880
TAGTGTGAGA TAAATGCTCCT GCCTTGGTTG CCCAGGTGAG TCGAGGGTGT GCACAGATT
2940
TAAAGTGTCGA ATGGAAAAATCAT GATGCTGTGT ACAGACCTCA GAGCTCGGA AAGATAAAAA
3000
AGAATAAATA AAACCTCATAA CAGACCCTGA CAAATTTAAT CTTAGAGACT GCACAGACT
3060
TACTTTGCTAC TCTCTCCCCT TGCCCTATTT AGAAGCTGAG ATACTCCCTC TTGGATCGGT
3120
TTTCATCCCTTT TTAAGATCCCT TTATGGGCGT CCTATGCCCAT CACTGTCTTA AATGATGTTT
3180
TTAAACCTTAT GTGTTTAAAA TAATGATCTA TGTGTAAAGT TAAAAAGGCT GCAGGGTGTTG
3240
CACAAAAAGG CTGGGTCTAA ACGGAGCTCA TGGAACAGCG TGGTGACCAC AAGGAGACCT
3300
TACCAGTTCC AGCCAGGAT CGATCTACGG ATCCCCGGGT CGACCCCGGGT CGACCCTGTT
3360
GAATGTGTGTT CAGTTAGGTT GTGGAAGACTC CCCAGGCTCC CCAGGGACGA GAATGATGCA
3420
CCTGTGGGTT TCGCCACCT CAGACTTGAG CTCGATTTT TGTTGATGCT GTCAGGGG 5520
CGAGGCTAT GGAAAAACCG CAGCAACGCG GAGATCGGCC GCCTCGAGTA CACCTGCTC 5580
ATGTGGAGAC CTCCTGGCT CACTAAAGGG GTCTCCGACT AGTTCTGGTT ACTAATCGC 5640
CTTTATCTGT TTTGGTTCCC ATGTAAAGA TAGGATTAAT GGAGTATTTT CCACATAG 5700
ATAAGACTCT TGGAAATTCT AGAGATGAA TATATTACAA GAAGAGTGG GGAA 5754

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5754 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:
TGAAGAATAA AAAATTACTG GCTCTTTGTT AGAACATGAA CTTTCACCTC GGAAGCCACC 60
CCCTCCCATC TGGAAAAACAT ACTGAGAAGA AAACATTCTT GGAACACCA CAGAAATTTT 120
CAACAGGGCA GATGATATGC CAAACACAGG ATATGACTCT TTGTTGAGTT AAATTGTTG 180
TGTTAAACTC TGGGCTATTC CCTCCTCCATT CCCCCCTCCA GTTTGGTTT TTTTCCCTTA 240
AAAGCTTGTG AAAAATTTGA AGCTGTGCTG AGACTCTCTT ACCCTCTGCA AGGGTTATG 300
AGTTTGACCC CCAAGGCTCT GTGTCGCTTC TGTTGCTGTG TTAATTCTGAC CCCAGAGCTC 360
TTGCTCTGTG GCTTTCTAGT GTGGCTCTTA TTAATCTTA CTTTCTCATT TTTAGTTATG 420
GTCTCAGGTG CTCTTCGGGT AGCGCGCTGT CCGGGACTT GAGTTCTGGA GTCAGGGCTC 480
TCCCTCGAGG GTCTTTCACT GTGTACATGG GCCGGGAAAT CGGAATATT CTATTGTGG 540
CATTCGCGGG GAATTCGAAA ATCTTTCAAT TTGTTGCTAT GCCGGGAAAC AGCCGGCAAC 600
CCAGAGGCCG CTAGACCCAC TTAGGTGTTA GTTTTTTCT CGTGTTGCTT 660
GGTGTTCGAT CTCTGTTTCTC TTTTCTAAG CTGTTGCTGA CCAGCAGTTTC ATGTTTGG 720
AGCTCTAGTG AGAGCGCGCT CCGAGAGGA GTGCGGGGTT GATAAGGATA GCAGGTCTCA 780
GGTGTCGACC GTCGGTTGCG CCTGGGAGAG GTCCAGAGAC GAACAGGGGA GATCAAGGA 840
CGCCCTGTTG ACCCCCTTGA AGGCCAAGAG ACCATTTGGG GTTGCGGAGT CGTGGCC TCTTGGACGCAG 900
AGTCCCACCT CGTGGCCAGT TGCGAGATCG TGTTGTCCAG TGCACCACTGG TGTGGGTTCG 960
CGAGATCGTG GTTGGCGATC CCACTCTGGG CTCTGCTACG GGATCTGGGT GGCGGTCCGC 1020
ACCTGGTTGT TTGGTCGAGG ATGCTGGGTT CAGCAGCCAC CTCGGTGCTG GTCACGGGAT 1080
CGTGGGTTCG AGTCCGACCT CGTCGAGAG GTCTCAGTTG GCCGGCCTTA GAGAGGCAAT 1140
CTGATTTCTC TGGTTTCTCT TTGTTGCTTA GTCTGCTGTC GCCTCTTGGT GTGACACTCTG 1200
TTTTTCTAAA AATGGGACAA TCTGTGTCCA CTCGCCCTTC TCTGACTCTG GTTCTGTGCG 1260
TTGGTAATTT TGTTTTTTTA CGTTPTTTTT TGTTGAGTCGT CTATTGTGTC TGTTACTATC 1320
TTGTTTTTTCT TGTTGTGTTA CCGGTTCTGT TGTTGTCTGT TGTTGTCTTT TGTTGTCGAGA 1380
CTTGGGACTGA TGACTGAGCA CTTTTTTTTA GATATGCTCTT CTAAAATAAG CCTAAAATCT 1440
CTGTCAGATC CCTATGCTGA CCACTCTCCT TCAGATCAAC AGCTGCCCTT ACTCGAGGTC 1500
AAGCTTCGAA TTCTGCAGTC GACGGTACCG CCGGCCGGGG ATCCACCGGT CGCCACCATG 1560
GGTAAAGGAG AAGAACCTTCA CACTGGAGTT GTCCAATTTTC TGTGTGAATT AGATGGTGAT 1620
GTTAATGGGC ACAAAATTTC GTGCTAGTGGA GAGGGTGAAG GTGATGCAAC ATACGGAATAA 1680
CTTACCTTCA AATTTATTGG CATCTAAGGA AAATCTCTGG TTCCATGGCC AAACACTCTG 1740
ACTACCTTCT CTATTGTGTTG TCAATGCTTT TCAAGATACC CAGATCATAT GAACACGACAT 1800
GACTTTTTTA AGAATGCGCA GCCGAGAAG GTATGTACAG AAAGAAGTAT ATTTTTCAAA 1860
GATGCAGCGGA ACTACAAGAC AGCTGCAGCA GCTAGATTTT AAGGTGATAK CTTCTGTTAT 1920
AGAATCGAGTT AACAGATGAA GAAGATGGAA ACATCTTGGG ACAACAAATTG 1980
GAATACAACAT ATAAACTCACA CAAATGTATT ATCCTGCGA CCACAACAAA GAAACTGACC 2040
AAAGATTTAATCT TAAAAATTAG ACACAACATT GAAGATGGAA GCGCTTCACT ACTGACAGCAT 2100
TATCAACAAA ATACTCCCAAT TGCCGATGTC CCTGTCTCCT TACAGACACAA CCAATTACTT 2160
TCCACACAAT CTGACCTTCT CAAAGATCCC AACAGAAGAG AGACACCAAT GTCTCTCCTT 2220
GAGTTTGTAA CAGCTGTCGG GATTACACAT GCAGATGAGT AACTATACAA GCACCGATCT 2280
AGATACGTGT ATCGATGGAT CGGAGAAGGG GAAGACAGCT GCATGTGGTG AGCAGAAAGCA 2340
AGTGATCTAG CCGCAGACGC TTCCTAAAAG GACTTCAGCC CACAAAGCGA AACGTTGAGC 2400
TTTAAATCAA GCTCTGTTAAG TGGTTAATAA AAAAAGCTCT ACACGGAAGAG CAGGTATGCT 2460
CTTGCCACTG TACAGACGCA TATACAGACA AAGAGAATCTG TGCACTCTG CAGAGAAGAGA 2520
CCTAAGATGC TGTGCGTAAAG AGAAATCAGAG TGGCAATCTC AACCGCCGAG GCATCCTAA 2580
GAGCAATGAT CCTGACAGTC TGAAGACTAT CAAGTTTATG ACABAATQAG ACTGTTGATF 2640
AAAACCCCTGT ATAAATATGT AAAAATCTGA AAAAAGAAAC GATCTCTCCCT ATGAGAGAAC 2700
AGACCTGGACA TCTACTGCAA AATAGACTTT ACTGGAAAAA ATATGTTGAT GAATACCTTC 2760
TAGTTTTTTG GAACGGTCTCA AAGATGGATA AAAACTTTCCT TGTGAAACA GAGACTGATC 2820
AGATAGTGTA CAAGAGAGTT GTTTACGAAAA ATTTTCCACAG GTCGGGTGTC CCAAAGGACA 2880
TAGTGTCAAC TAATGTTCTC GCCTTGGGTG CCGAAGTGAAG TCAGGGTGTTG GCGAAGATT 2940
TAGAGGGCTAA ATGAAATTTG CATTGTTGCT AGACCCCTCA GAGCTCAGGA AACATATAAA 3000
AGAATAAAAATA AAACCTTCAA CAGACCTTCA AAAATTTAT GCATAGGACT GCCACAGACT 3060
TACCTGTGAC TCTCTCCCTC TGCCTATTTT AAGAACTGAGA ATACTCCCCTTG TAGATGCTGA 3120
TTTACTCTTT TAAGATGCTT TTTAGGGGCT CCTAATGCGA CACTGTCTTA AATGATCGCT 3180
TTAACCTTATA GTTGTATATA TAATGATCTA TATGTTAAGT TAAAAGGCTG GCAGGGTGTTG 3240
CAGAAAGAAG TCTGGTCACA ACTGGCTACA GTGAAACAGC TGGGTACCC C AAGGACATCT 3300
TACCCAGTCC AGGCCAGGAT CGAGATCTACG ATCCCCGGGT CGACCCCGGGT CGACCCTTGTG 3360
GAAATGTGTG CAGTTAGGGT GTGGAAGGTC CCCAGGCTCC CCAGCAGCAGA GAAGATGAGC 3420
AAGCACATG CTCAATTAGT CAGCACCAGA GTGTTGGAAGA CCCCGAGGCT CCCGCCAGGG 3480
CAGAAGTATG CAAAGCAGTC ATCTCAATTA GTCAGCAACC ATAGTCCGCC CCTAATCACC 3540
GCCCATCCCG CCCCTAATCTC GCAGCCATTC GCAGGCCATTG CGCTGACTAA T 3600
TTTTTTATTAT TATGCGAGGG CCTGGCCGCGG CTCGCGCTCTT GAGCTATCCG AGAAGTAG 3660
AGAGGGCTTTT TTTGAGGCGG TAGGCTTTTTA CAAAAAGCTT CACGGTCGCGG CAGGACACTCA 3720
GGCGGCAAGG GCTGCTTAAAG GAAGGGGAAAC AGTGAAGAAG CCCGGCTGCC GTAAGGGTTG 3780
TGCCCCCGAG TGATATCGAC CTATCGGGCT ATCTGGCAAA GGGAAACGCG AAGGCACAGG 3840
AAGAACAGGG TAGTCTGCAG TGGGCTTACA TGGGCCATAG TACACTGGCC GTTTTTTATGG 3900
ACAGCAAGGG AACGGGAAATT CGGCAGCTGGG GCGCCTCCTGT GAAAGTTGGA GAGAGGCTGC 3960
AAGATGAACT GGATGGCTTT CTTGCGGCCAA AGGATCTGAG GGGGAGGAGG ATCAGATCCT 4020
GATCAAAGAGA CAGGATGAGG ATCGTTTGGC ATGATTGCAAC AAGATGGATT GCCAGCAGGT 4080
TCCTGGGCGC CTTGGGTGGA GAGGCTATCC GTCTATGACT GGGCAGACAAC ACAATCGGCCC 4140
GCTCTCTGATG CCAGGGCTTGG CCAGCCGTCA GGGGCGGCCC GCCGCCGTTC TTGTGTCAAG 4200
ACGCAGGTGTG CGGTGTGGCCT GAATAGAATGT CAGGACGGAG GAGGCAGGGT ATCGTGGCTG 4260
GCCACGACGG GCCTTGCTTTG CGAGACGTGGG CTGAGACGTTG TCACTGAGAC GGGAGAGGGC 4320
TGCTGCTGATG TGGGCGGAAGT GCCGGGGCAG GATCTCCTGT CATCTCAACTT GCTGCTGCTG 4380
GAGAAAGTATG CCATCAGTGC TGATGCAATT CGGCGCGGCTG ATACGCTTTGA TCCCCCTGAC 4440
TGCCCCATCGG ACCAACAAGC GAAACAAGGCA ATGAGACAGA CAGCTACTGC GATGGGAACC 4500
GCTTTGTGCC ATCAGGATGA TCTGGACGAA GAGCATCAGG GCTGCGGCGGC AGCGCAACTG 4560
TCCGCGAGGC TCAAGGCGCC CATGGCACGC GGGAGGAGTC TGGCTGTCAG CCATGGCGAT 4620
GCCGCTTGGG GCAAATATCATGG GCCGGGCTTTT CTGGATATGC CATCTGTTGGG 4680
GGGCTGCGGTG TGCCGCAGCCG CTATCAGGAC ATAGCGTTGG CTACCCGCTG ATATGGTGAA 4740
GGACGTTTGCG CGCAATGGGC TGGACGCTCTT CTGGTGCCTGT ACAGCATGCG GCCTGGGCGAT 4800
TGCGACGGCA TCAGCTTCAA TCGCCCTTCTG GACGAGTCTG TCTCGAGCGG ACTCTGGGGT 4860
TGGAAATGAC CGACCAACGC AGGCCAAAACT CGGCCATACACG AGATTCTGAG TCCACGGCGC 4920
CCTCTATAGA AAGGTTGGGC TCGGNNATCAG TTTTGGGGGA CGGAATAGCT AATCTGCTGC 4980
TTGCAAACAA AAAAACACC CCTAGACGGC GTGTTTGTGTC TCCGGGATCA AGAACGATCC 5040
ACTCTTTTTCC CGAAGGTAAC TGGCTCTAGG AGAGGCGAGA TACCAAAATAC TGGCTCTCTA 5100
GTGTAAGCGGT AGTTAGGCCG CCAACTTCAAAG AACTGGTAGG CACGGCTTACC ATACCTGCTG 5160
CTGCTAATCC TGTTACAGGT GGGCTGTGCC AGTGGCGATA AGTCTGCTGT TACCGGGTGG 5220
GACTCAAGAC GATAGTTACC GGATAAGGGC CAGCGGTCGG GCTGAACGGG GGGTTGGTG 5280
ACACAGCCCA GCTGGAGCG AACGACCTAC ACCGAACTGA GATACTACAACGGTGAAGCAT 5340
5TGAGAAAGGC CCAAGCTTCC CGAAGGGAGA AAGCGCAGACA GGTATCGGTA AAGCGGCAAGG 5400
GTCGGGACAG GAGAGCCGCA GGGAGCTTCCAGGGGGAAGGCGCTTGTA TCTTTTATG 5460
CTGTTCCGGGT TCGCCCACTCT GTGACCTGAG CGTGTGGATTT TGTGATGCTC GTCAGGGGGG 5520
10CGAGGCTAT GGAAAAACGCC CAGCAACGCC GAGATGCGCC GCCTCAGGTA CACCTGCGTC 5580
ATGCTGAGAC CCTCAAGCCT CACTAAAGGG GTCCCTGCGCT AGTCTGTTTT ACTAATCTGC 5640
15CTTTAATGT TTTAGTTCCC ATGTATAGGA TAGATGAAAT GCAGATATCCT CCACATAGG 5700
ATATAGACTT CTGAAATTTCT CAAAGTGGAA TTATTTTACAA GAAGAGGTTGG GGAAG 5754

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 4958 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

AGGGGGGAGC AGCCAGTCGAG TGCTGGACAG AGAACCAAGTG ATCTAGGCCA GCAGCCTCCC 60
TAAAGGGACT TCAGCCCAACA AAGCCAAACT TGTTGGCTTTTA ATACAAAGCTC TGGATAATGGT 120
AAAAAAAAAA AAGTCTACAC GGACAGCAGA TAGGCTCTGG CACTGTACA GAGCAATATA 180
40CAGACAAAGA GAACCTGGGA CATCTCGCAA GAAAGACTTA AGATGCTTGCT GTAAAAGAAA 240
ATCGAGATGC AACATCAACC GCCAGGGCAT CTAAGAGAGC AATGATCTTG AGTACCTGAA 300
GACTATCAAG TTATAGCAAA ATTAAGACGTG CTAAAAAAATCCGTATTTA AATAGAATAA 360
45ACTGAAAAAA GAAAAGATGT CCTCTCATGA GAGAGACAGC CTGCACATCTA CTGAAAAATA 420
GACCTTACTG GAAAATATAT GTGATGAATG ACCCTCTAGTT TTTTGTAAGAC GCTTCACAAGA 480
50TGGATAAAAG CTTTTCTTCT TAAACGAGA CTGTACGATG ATGTCATCAAG AAGATGTTTA 540
AAGAAAAATTT TCAAGGTTGC GAGTGCCAAG AAGCAATAGT GTCGATAATTG GGTCTGCT 600
55TTGTTGCCCA GGGTAAGTCAG GGTGTTGCGCA AGTATTTTAGA GTGCAATATG AAAATCCATT 660
GTGTGTACAG ACCTCAGACC TCAGGAAAGA TAAACACAAAATAAAGA CTCGACAGA 720
CCTTGACAAA ATTAATCCTT CAGACTGCTCA CAGACTCTCT TTGGATTTCC CCAACC 780
60CTATTAGAA CTGAGAATAC TCCCTCTTTGA TCCGTTTTTA CTCTTTTTAA GTACCTTTAT 840
GGGCTCTCTA TGCCATCCTG TCTTTAATG ATGTGTTTAA ACCTATGGTG CTGATAATAAT 900
GATCTATAGTT AAGGCTTGCG AGTGTTGCGAA AAAGATGCTG AGCTAACCTG 960
65GCTACGTGAA AGAAGCTGGG TACCCCAAGG ACATTTTACC AGTTCCAGCC AGAGATCTGA 1020
ACCTACACCG AACTGAGATA CCTACACGGT GAGCATTAGG AAAGCGCCAC GCTTCCCGAA
GGGAGAAAGG CGCAGACGGTA TCGCTRAACCC GGCAGGTCG GAACAGGAGA GCGCAGAGG
5
CTGAGCGGTC GATTTTTTGT AGTCTCTGTA GGGGGCCGGA GCTTATGGA AAACGCCAGC
AAGCCGCAGA TGCCGCCGCT CGAGTACACC TGCGTCAATGC TGAGACCCTC AAGCCTCAC
10
AAAAAGGTGCT CTGCTTAAGT TTGTCTTAATA ATCTGCTCTTA TTCTGTCTTTTA GTTCCCATGT
TAAAGATAGA GATAATGAGC TATTTCTCAC ATAGAGATAT AGACTCTCTGA AATCTTAA
15
TTAGAAATTAT TTACAAAGAG AAAGGGAAGA TGAAGAATAA AAATATTACT GCCTCTTTGG
AGAAGAGTAA CTTTCACCTC GGACGCCACC CCGTCCACAC TGAGAAACAT ACTTGLGAAA
20
AACATTTTCT GGAACACCCA CAACAGGTTC CACACGGCA GATGTATGGC CAAACAGCAG
ATATGACTCT TTGGTGGAGT AAAATTTTGG TGTTAATACCT TCCCCATTTCC CTCCTCTATT
25
CCCTCCTCCA GTGTTGTGTT TTTTCTTTTA AAAGGCTTGGTG GAAATTTGGA TGCTGCTCTG
AGACTCTCTCT ACCCTGGGCA AAGGTGTAAG AGTTTCGACC CGCAGACTCT GTGTGCTTTC
30
TGTTGCTCTGTT TATATTGGAG CCGAGAGGTC TGGCTCTGTT GCTTTACGTT CCGCTCTTTA
TTAAATCTTTA CTTCTCTACTT TTTATGTTAG GTCTCAGTGT TTTCTTGAGT AGCCGGCGTG
35
CCCCGGAGCTT GAGTGTCGTA GTCAGGGGCT TCCCTCGGAG GCTTCTATTG TGTTAGATCG
GCCGGRATT CGGAAATCTT TCGATTGGTTG CTTGSGCCGG GAATTTGCAAA ATCTTTTATT
40
TGCTGCAATGG GCAGCGCAGACC AGCGGCAAC CGCGGCAGGGT CAAGAGAGTTCAA TTGCTCCCGG
AGCTCTCTCGA TTCTGCTTTGG TGCTAGTGTCT TGCTGTTGAT AGTTGTTCTAG
45
TCTGAGGGGTA CGTAAGGGATT GACCGTGCTCA GGTGTCACCG GTCCGGTCGC CCGGGAGAG
GTGCGGCTTGTA CACGAGGAT CGCGTGTCCG CTTGTCGTCG CCGGGACGGC CGGGGAGGAG
50
GTCCGGGGCTA GACGCCTGAG GCAGCTGCTAG CGGGGTCGCTG TCCTGTCGCTA GACAGGCTCT
ACCTACACCG CTCGCGCCTCT GTCAGCCGAT CGTGGGTCG CGTCCACCTCG GTCGAGGAGG
55
GTCTCAATTTA GGCGGCCCTA GAGAGGCATT CTGATGCTGCT TGTTTCTCTT TTTGCTTTTA
GTCTCAGAGTC GCCTCTTTGT GTGACTCTCG TTTTTCTTAA GATGGGCGAA TGTGTGCCTCA
CTCCCTCTCT TCTGACTCTTG CTGTCACGCTG TGTTGAATTG TTTGTTTTTGA CGTTGGTTTT
60
TGTTAGTGGT CTAAGGCTTG GTTACAGTCT TCTGGTTTGT TTTGTTCTTG GGGTTTCTGT
GTAGTTTCTTG TGGTGTCGCTG GTCAGTACCGA ATGGAACTGA TGACTGAGCA CTTGTTTTAA
65
GTATGCTTTC AAAAAATAG CCTAAAAATGC CTGTCAGATC CCTATGCTGA CCACTTCCCTT
TCAGATCAAC AGCTGCCCCG AGCTATCCGAG CCGGAG
(2) INFORMATION FOR SEQ ID NO:21:
(i) SEQUENCE CHARACTERISTICS:
   (A) LENGTH: 7080 base pairs
   (B) TYPE: nucleic acid
   (C) STRANDEDNESS: single
   (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xii) SEQUENCE DESCRIPTION: SEQ ID NO:21:

```
GAATACAAGC TTGCTATGCC TCGAGTGACG TCTAGAGGAT CTTGAGAAAT AAAAAATTAC
60
TGGCCTCTTG TGGAGAATGT AGCTTCACAC TCGAGGCCCA CCCCTCTCCC TCTGGAAAC
120
ATACTTGAGA AAAAAATTTT CTCGAGAAC AAGAAGATGT TTAAACAGAC CAGATGATT
180
GCCAATCAC AAGATGACT CTTTGCGTTGA TCTAAATTTGT GGGTTGTTAA CATCCCTAT
240
TCCCTCCCCA TTCCCCCTCC CAGTTTGTGG TTTTTTCTTT TAAAAAGCTT TGAAAAATT
300
GAGTCGTGCT CGAGACTCCT CTACCCGCTG TCGATTTTGA TGAAGCTGCA CCCAGAGCT
360
CTGTGCTCTT TCGTGTCTGT CTATTTCCTG ACGACAGACG TCTGAGCCTGT GTGCTTTCAT
420
GTCGCTATCT TATATAACTC TACCTTCTAC ATTTATGTAC TGGTTCTCAGT GTCTTCTTGG
480
GTCGACGGCT GTCGGGCGAC TCTAGTGTGT CAGTGGGAGG CCTCCCTCAGA GGGCTTTCA
540
TTGCTGACAT GGGCCCGGAA TCTACGAGATC TATCCTTTGG TGATTTGCGC GGAATCTCG
600
AAATCTTTCA TTTGCGTCAT TGGCGCGGAA ACAGGCGGAC CACCCAGAGC TCCTAGACCC
660
ACTTACAGGT AAGATTCTTT GTCTGTGTGT GTGTCTAGGT GTGCTGTCTG AGGTCTGCTG
720
TCGGTTCTTA AGTGCTGTTG CAGCCGACTT TCACGTTTGC GGACGCTCGA TAGGACCGCG
780
CTCAGAGGAG GAGTCCGGGG TGGATAAGGA TAGACGTGCT CAGGTTCTGA CCGTCGGGCT
840
GCCGCGGGAG AAGACCGGAG GAGAGCGGAG GAGCTCGTGG AGACCCCGTT GTGAGGCTG
900
GAGGGCGAAG AGACCTTTGC GGGTTGCGCAT ATGGTTGCTT CTAGCTTGAC ATCAGATGTT
960
GCAGAGGGTC TCAAATGGC GCCCGTATAG GCTACGATCT AGCAGATGG TAAAGAGTCG
1020
ATTGCCAGCG GCCGCGTGGT GCAGGGGGTA TCCGTCATG ACTGCGGACA
1080
ACAGACAACT CAGGTGCTCT ATGCCCGCGG GTTCGGGTGT TCAGCGCGGC GGGCCGGGT
1140
TCTGTTTGTC AAGACCGACC TGGCGCGGTC CCTGAATTGA CTGACGAGAC AGGACCGCG
1200
GCTACGCTGG CTGGCGACA GCAGCGTCCC TTCCGCAGCT GTGCTGACG TTGCATCTGA
1260
AGCGGGAAAG GACTGCGCGG TATCTGCGGA ATGCGCGGCG TGAGATCTCG TCTGATCTCA
1320
CTGATATCGGT GGGAGGAAAG TATCAGATCT GGCTGATGCA ATGCCGCGCG GCGCCTACG
1380
TGTACGCGGCT ATCGCGCCAT CTGACGACCA AGGAAACAT CGCATGAGC GACGACGTC
1440
TGAGATGCGA GCCGCGGTCG TCGACGAGC TGATCTGGAG GAGAGATCAT CGGACCGGTC
1500
GCCAGCGGAA CTGGCTGCCA GCAGCATGGC GCCCGCGGAC ATCGCTGGTC
1560
```
GACCCATGGC GATGCGCTGCT GCAGGAATAT CATGCTGGAA AAGTCCGCCGCT TTCTCTGGATT 1620
CATCGACTCTTG GCCCGGGCTGG GTGTTGGCGGA CCGCTATCAG GACATAGCGT TGCGTACCCG 1680
5
TGATATGCGT GAAGAAGCTTG GCGGCGAAGT GGCTGACCCGC TTCCCTCGTC ATTTACGGAT 1740
CGCCGCCTGGC GATTGCAGGC GCATGCGCCTT CATAAGCCTT GCTTACGGATC 1800
10
GGGACTCGTG GGTTGGTAAT GACGACAAAA CGGCGCCTCC AAGTCCGATC ACGGATTTAC 1860
GATTTCCACCG CGCGCTCTCAG TGAAGGGTTG GGCTCTGGAG TTAGCTGTTT TCTTATCAGT 1920
TTGTCAGATTTC TATATTTCGA ATACAGAAACA ATAGCTTCTA TAATCAAAAT ATATTGAATCT 1980
15
TTGATATATTA TGATTTCTCCCG TCGAACCAGT AAACACTTCTC CAGCTGGAAT TCAACATTCC 2040
TCTGTCATCT GCCAGGCCCCA TATGTTATTAC ATGGAGATCT TTTGGAGAAC ACTGCAAGTT 2100
20
CATATCATATAG CAGAGATTTGA AATAGGTATG TGTGGGCGAT TGATGAGAGCT ATGTTTTGC 2160
TGATATCCATC GAAAAAGGAT TTGTATAAAA GCATTTAACA CCATATATAG ATAGATTAAA 2220
ATATTCGGAC CATTAAAAGA AAAGTGAGTC TGCTCATCAG TCTGCTCTCC ATGGCTCCCT 2280
25
TCACATGCGAT CCTCTCTTATT TTCTCCATTG TTGTCAAGAA AAATATTAGCT CAGCTTCTGT 2340
TCTCAGTTAT TCCTGTCCCTA GCATGCGCTCA ATGGGACGGT GTGATAAGAA AGGGGATCAA 2400
AGAAAGCGAG CTTCCTGCTGT GCTACTACAA CCATAAGTAT AAGCAAGCCTA ACTAATTATT 2460
30
GCTAATTAGT TTTTTCACTCT CTAAGGGTCC CACATTATTTC TGGTTTCCTTA AAGATCCCCT 2520
TATCTTGTTT TCACTGAGAT CAAATCCCAG AAGGGACGTC TCTCTCCCTA 2580
35
TCAACAGGTG CTGGATTAGT AGGCGAACAG CGGAAAAGCA CATTGTTACAG CAGGATAAAA 2640
AACAAATATT TGCTCTCCCAG TCAATGCCAA ATGGCGCTAT TGAAGACTAAA ATCTAACCAA 2700
ATCCCATATAG ATGATTTCTA TGCCGTCAAAG GCACAAATCT CTGAAGGGGA CCCTGGGTTG 2760
40
GGTCAAGATT CAGGCTTATA ATCCCCCAAG GCTACAGCCAG TGCTCTGATA TACACACCGG 2820
ATCTGTGGAA CAGTCTCACT AGCTGGCAATG GTACAGGCCT CCACGCCGGTC CCTGCTCTTG 2880
45
GCTTTGCGCC TGCTCTGCTC GCCTGTTGCTT CAAGAGGGCA GTGCTCTCCCC AACCATTTCCG 2940
TTATCCGAGCC TTTTTCGCAG CCGTATGCTC CGGCCCGCTAC GTCTCGCAAC GCTGGCCCTTT 3000
GACACCTACC AGGAGTTTGA AGAAGGCTAT ATCCCAAAAG AAGCAGAGAGT TCACTTCCTG 3060
50
CGAAGACCCCG AGACCTCCTCT CGTTCTTCGTA GAGCTTATTG CAGACACCTC CAAAGGGGAG 3120
GAAACACAAAC AGAAATCCCAA CCTAGAGCTG CTCCGCACTCT CCGTTCGCTAT CATCCAGTGG 3180
55
TGCTGGGAGC CGGTGCAGTT CCTAGAGGTG GTCTTCGCGCA AAGACGCTCT GTACGGCGCC 3240
TCTGACAGCA AGCTCCTATGA CCTCTAAAGG GACCTAGAG AAGGACATCC AAACGCTGATG 3300
GGGAGGCTGG AAGATGGGAC CCCCGGACAG GGCAGATATC TCAAGCGAGC CTACAGCAAG 3360
60
TACCAACAACT ACTACACAGA CGATGACGCA CTACTCAAGA ACTACGGGCTG CTCTTACTCG 3420
TTACAGGAAGG AGATGAGCAGA GTGAGCGGCAA TCTCTGGGCA TGTCCTGATG CGGCTCTTG 3480
65
GAGGGCGAGT GTGCTCTTCTA CCTGCCCGGG TGGCATCTCG TGACCCCTCC CGAGTCCCTC TCCTGGCCCT GGAAGTTGCC ACTCCAGTGCT CCACCAGCTT TGCTCTAATG AAATAAGTT 3540
3600
CATCTAAAGT CAGGCGGCGC CTAGACTTCTT GAATTCTAA GATTGAGATT
ATTACAAGA AGAAGTGGGG AATGAGAAT AAAAAATTAC TGGCCCTTTG TGAAGACATG
AACCTTCACC TGGAGCCCCA CCCCTCCCCA TGTGGAAAAC ATACCTGAGA AAAAACATTTT
CTGGAAACAC CAGACAGTGT TCCAACAGCC CAGATGTATT GCCAAGACACA GGAATGACT
CTTTGGTTGA GTAAATTTGT GGGTGTAAA AATCCCCCTA TCCCTCCCA TCCCCCCTCC
CAGTTTGTTG TTTAATCTTT TAAAGAAGTT TAGAAAGATTT GAGTCTCGTG CGAGACTCTC
CTACCCCTGTG CAAGAGTTGA TGAGTTCTGA CCCGACGGCT CTGATGCTTCT TCTGGTGCTG
CTTAAATTGC ACCCCAGAGC TCGTGGCTGT GTGGTTCTAG GTCTGGCCTT TATTTAATCT
TACCTTCCAC ATTTATGTA TGGTCTCGAT GTCTTCTTGG GTACCGCGGT GTCGGCCGAC
TTGAGTGTCT CGAGTGAGGT CTTCTCTGGA GGCTTTCTCA TTTGGTACAT GGGCCGGAA
TTCCGGAATTC TTTCAATTTGG TGCAATTGGG GGGAAACGTA AAATCTTTCA GATCCCCGGG
TACCAGCTCC GAATCTCCGT CTCTCTATAG TGAGTCTCGAT TAAATTCGAT AGGCAAGCTG
CATTAAAGGA TGCGCCAAAG CCGCGGGGGA GGGCGTTTGC GTAATTGGGC CTCTCTCCGT
TCCTCGCTCA CTGACTCGCT GGGCTCTGGTC GTTGGGCTGAC GGGCCGCGAT ATCACGTCAC
TCAAGGGCGG TAATACGGTT ATCCACAGAA TCAGGGGATA ACGCAAGAAA GAACATGTTA
GCAAAAGCC AGCAAAAGGC CAGGAACCGC AAAAAAGGCC CTGTTCTCGGC GTCTTCCAT
AGGCTCCGCC CCCCTGACGA GCATCACAAA AATCGAGGCT CAGATCGAGG GTGCCGCAAAC
CGACAGGAC TATTAAGATA CAGGGCGGTTT CCCCTGGGAA GCTCTCTCGT GGCGCTCTCT
GTTCCGACCC TGCCGGCTAC CGGATACCTG TCCGCTCTTC TCCCTCTGGG AAGGCGTGCGG
CTTTCTCATA GTCAGCGCTG TAGGATATTC AGGCTGCTTG AGTTGTGCGG CTCACAGCTG
GGCTGTGGGC ACAGAACCCC GGTGACACC ACGCCGTGCG CCTCTATCCG TAACCTATCGT
CTTGAATCCA ACCCGGTAAG ACAGCACTTA TGGCCACTGG CAGCGCCAC TGTAACAGG
ATTAGCAGAG CGAGGTATGT AGGCGGTGCT ACAGACCTTG TGAAGTGGTG GCCTAACTAC
GGCTACACTA GAAGGAGCT ATTTGTGATG TCCGCTCGTC TGAAGGCAGG TACCTCCGGGA
AAAAGAGTGG TGATGCTCTTG ATCCGGCACA CAACACCGCG TGGGTGTGGT 5100
GGTTGCAGAC AGCAATACAC GCCAGAAGAA AAAAACTC GAGAGATCCC TTTATCTTTT 5160
TCTACGGGGT CTGAGCGCTCA GTGGAAGCGA AACTCAGTTT AAGGGATTTT GTGTACCCAG 5220
TTATCAAAAA GGAATCTCAG CTAGATCCTT TTAATATTTA AATGAGTTTT TAAATCAATC 5280
TAAGGTATAT ATGAGTAAC GGTTGCTGAC AGTTAACAACT GCTAAATCCG TGAGGCACCT 5340
ATCTACGGGA TGGCTCTATT TGTGTGACAC ATAGTTGCGT GACCTCCCGGT GTGGTAGATA 5400
ACAGTGATAC GGGAGGGGCT ACCATCGCTG CCCAGTGCTG CAATGATACC GCGAGACCA 5460
CGCTCAGCGG CTCCAGATTG ATCAGCAAATA AACCAGCGCC CGGGAAGGGC CGAGGCGAGA 5520
AGTGGGTCTG CAACTTTATC CGCCTCCATC CTGCTTATTTA ATTTGGTCGG GGAAGCTAGA 5580
GTAAGTAGTT CGCCAGTTAA TAGTTTGCAG CACGGTGTTG CCAATGCTAC AGGCATCGTG
5    GTTCAGCGCT CTCGTTTGG TATGGCTTCA TCTGAATCCC GTCGCCCAAC ATCAAGGGCA
10   GTTACCATGAT CCCCCATGGTT GTGCAAAAATA GGCGGTTAGCT CTTCCGTTTG TCCGATCTTG
15   GTGACAGAGTA AGTGGGCGCC AGTGTATACCA CTCATGTTGCA TGCCGCCACT GCCAATATTCT
20   CTTAAGTCTCA TGCCCATCCTG AAGATGCTTAT TCTAGTGCAGT GTGATGCTACG AACACGAGTCA
25   TCTCGAGAAT AGTGTATCCG GGCGCCGAGTG CTGCCTTGCC CGCCGCTCAAAG GGGGAATAAG
30   ACCGCGCCAC ATACGCGAAC GTTAAAGTG CTGACGGATGG AAAAGCTTGGG TCCGGGAGCA
35   AACTCTTCAAG GTGATTACC GCTGTTGAGA TCCGGTTGCA TGTAACCCAC TCGTGCAACC
40   AACATGATT CAGCACATTT TACCTTCACC AGCGTTTCTC GGTGAGCAGAA AACAGGAAGG
45   CAAAAATCCG CAAAAAGGG TAAATGACGCA CACCGGAAAT GTGGAATACGC CATACCTCTC
50   CTTTTTTCAAT ATTATGAAG CATTTACAG GGTATTGCTC TCTGAGCGCC ATACATATTT
55   GAAATTTATG AGAAGAAATA CAAGGTAGGG GTAACCAGCCA CATTCCGCCAA AAAAAAGCAG
60   CCTGAGCTCT AAGAAACCAG TTTATCATG ACATTAACCT ATAAAAATAG GCGGTATGAC
65   AGGCCCTTTC GTCTCGAGCC TTGGGGGATG GACGCTGAAA ACCTCTGAGA CATGCGGACTC
70   CCGGAGAGCG TCACAGCTTG TCTGTAAGCG GATGGCCGGA GCAGACACAGG CGGTCCAGGG
75   GGTGCGCGGG GTGGTGCGGGG TGCGTTAATCT ATGGGCGATCT AGAGCAGATT
80   GTACATGAGA TGGCACCATT GACAGCTTTC CCTTATGCAG TCTCGACATT AGGAACGACG
85   CCAGTAGTAG GTTGGAGCCG TTTAGCACCC CCGGGCCGAA GAATGGGTCCA AGGAGATGGC
90   GCGCAACAGT CCCCCGGCCG CCCGGCTCGC CACCATACCC ACGCCGAAAC AAGCGCTCAT
95   GAGCCGCAAG TGCGGAGCCC GATCTTCCCC ATCGGTGATG TCGCGCATAT AGGCCGGCAGC
100  AACGCGACCT GTGGCGCGCG GTATGCGGCG CACGATGCGT CCGGCGTAGA GGAAGTCGGC
105  AGGAGATGCC CTGCTGTATG GTTCCGTGAC CATTCCGGGG GTGCGGAACG GTGTTACCAG
110  AAACATGAGA GTTGGCGCTCA ACCAAGACGA CTCTGACCGC AGTTAACGAG AGAGATGATA
115  GIIGTCTGCCT CAGTAGGCCA GATGCTACAC AATTAGGCTT TCTGACATTG TCTGGTACGAC
120  GCGGCTACAA TTAATACATA ACCTATGTTA TCATACACAT ACGATTAGG TGACATTATA
125  (2) INFORMATION FOR SEQ ID NO:22:
130   (i) SEQUENCE CHARACTERISTICS:
135   (A) LENGTH: 6795 base pairs
140   (B) TYPE: nucleic acid
145   (C) STRANDEDNESS: single
150   (D) TOPOLOGY: linear
155   (ii) MOLECULE TYPE: DNA (genomic)
160   (xii) SEQUENCE DESCRIPTION: SEQ ID NO:22:
165   AATGAAAGAC CCCACCTGTA GGTGGCGCAA GCTAGCTTTA GTAACGCCT ATTTGGAAGGC
<table>
<thead>
<tr>
<th>Position</th>
<th>Sequence</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ATGGAAAAT ACATAACTGA GAATAGGAA GTTCAGATCA AGGTCAAGGA CAGATGGAAC</td>
<td>120</td>
</tr>
<tr>
<td>5</td>
<td>AGCTGAAATAT GGGCCAACA GGTATCTGT GTTAAGCAGT TCCTGCCCGC GCTCAGGGCC</td>
<td>180</td>
</tr>
<tr>
<td>10</td>
<td>AGAGACAGAT GGAACAGCTG AATATGGGCC AAACAGGATA TCTGTGTTAA GCAGTTCTTG</td>
<td>240</td>
</tr>
<tr>
<td>15</td>
<td>CCCCCGCTCA GGGCAAGAAG CAGATGGTCG CCGAGTGGCG GCCAGGCCCTC AGCCAGTTCT</td>
<td>300</td>
</tr>
<tr>
<td>20</td>
<td>AGAGAACCAT CAGATGTTTC CAGGGTGCCC CAAGGACCTG AAATGACCCT GTGCTCTATT</td>
<td>360</td>
</tr>
<tr>
<td>25</td>
<td>TGAACCTAACC AATCAGTGCC CTCTCTGCTT CGTCTGCGGC GCTCTGTCCT CCCAGCCTCA</td>
<td>420</td>
</tr>
<tr>
<td>30</td>
<td>ATAAAGAGAC CCACAAACCC TCACCTGCGG CGCGACTTCT CGGATTGACT GAGTCGGCGG</td>
<td>480</td>
</tr>
<tr>
<td>35</td>
<td>GGTTACCTTG TATTTAATCAA ACTTTTGGTA CAGTTGATGG CAGCCGACTT CGTGGTTCGAG</td>
<td>540</td>
</tr>
<tr>
<td>40</td>
<td>GCCACCTCG GTTTGCTTGG CAGATCTGCT GGTCTGTCGTC CGACCCGTTCC CTGTCGACG</td>
<td>600</td>
</tr>
<tr>
<td>45</td>
<td>GGATCGTGAA TCCAGACGGCA AGAGATGTCA GAGGATGGCC AGAGACTTCTT GTAAGAGGAG</td>
<td>660</td>
</tr>
<tr>
<td>50</td>
<td>GAGTTCCGAA CACCCGGCG CGACCCGTTCC GAGGCAGTGG TGGGATCGAC</td>
<td>720</td>
</tr>
<tr>
<td>55</td>
<td>GGGCCACGGG TCTCTGGTCC CACCTGTGCG AGGATGGTCG TGGGATCGAC</td>
<td>780</td>
</tr>
<tr>
<td>60</td>
<td>GTTCCGAGAT CGTGGTCTGC AGTCCCACCT CGGCTGCTGG CAGGGATCG TGGGATCGAC</td>
<td>840</td>
</tr>
<tr>
<td>65</td>
<td>TCCACCTCG GTTTGCGTTC CAGATCTGCT GGTCTGTCGTC CGACCCGTTCC CTGTCGACG</td>
<td>900</td>
</tr>
<tr>
<td></td>
<td>GGTCAGCTCC ACCCTGCTGC ATCTGCTGCT TGGCTGGTTTC CAGGGATCG TGGGATCGAC</td>
<td>960</td>
</tr>
</tbody>
</table>

**Total:** 93
TCTGGGCCT CGGTGCACAT GCTTTACATG TGTTTATGCG AGGTTAAAAA AAGCTCTAGG 2100
CCCCCGAACACRGGGGGAC TGGTTTCTCTT TGGAAAAACA CGATACGGGA TCCACCGTGTC 2160
GCCACCATGG GATAAAGGAGA AGAACCTTCC TCAGGAGTGTG TCACCAATCTC TGTGAATTA 2220
GATGGTGATG TTAATGGGCC CAAATTCTCCT GCTAGTGGAG AGGGTGAAGG TGATGCAACA 2280
TACGGGAAAC TTACCTCTTAA ATTTATTGCG ACTAAGCGAA AACTACCTGT TCCATGGGCA 2340
ACACTTTTCA CTACTTTTAC TATGTTGGTT CAATGCTTTTT CAAGATACCC AGATCTATA 2400
AAACGCGACTG ACTTTTCTAA GAGTCCTGATG CCCGAAGGTT ATGAGCAAGA AAGAACTATA 2460
TTTTCAAGG ATGACGGGAA CTACAAGACA CTGTCGTGAA TCAAGTTTGA AGGTGATACCC 2520
CTTGTTATAA GAATCGAGTT AAAAAATTGTT GATTAAAAAG AAGATGGGAA CATTCTTGGG 2580
CACAAGATGG AATACAGACTA TAATCAGACAC AATGTATACA TCACTGGCGA CAACAACAAG 2640
AATGGAAACA AGTTTACTT CAAAATAGAA CACAACATTG AAGATGGGAA CGTTCACATA 2700
GCAGACCATG ATACAAACAAA ACTTCAACTT GGCGATGCGG CTGTCCTTTTT ACCAGACAC 2760
CATACCTGG CCCACACATC TGCCCTCTCTG AAAGATCCCCA ACAGGAAAGG AGACCACTTG 2820
GTCCCTCTTG AGTTTGGTAC AGCTGCGTGG ATTCACATAG GCATGGAGTA ACTATACAAG 2880
TCCGGATCTA GATACTGTA TCGATTGATC CGAAGGCGGG GACAGCGTGC CAGTGCGTGCA 2940
CAGAAAGCAAG TGATCTAGGG CACAGCGCCT CCCCTAAGGG ACTTACGGCC ACAAAGGCAA 3000
ACTTGGTGGCT TTAATACAGA CTCGTTAAAAT GGTAAAAAAA AAAAAGCTCA CACGGACAGC 3060
AGGTATGTCG TTGCACTGATG AGCAGACCAAT ATACAGACAA AGAGAACGTG TGACATCTGC 3120
AGAGAAAGAC CTAAAGATCT GTGCGTAAAAA GAAATCAGAT GCGAAATCTA ACCGCCCGAG 3180
CATCCTAAAG AGCAATGACAT CTGACAGCTT GAAACTATGCA AAGTATAAGAA CAAATTAAA 3240
CTGGTAAAGAA AAACCTCGTGA TAAATAAGTA AAAAATGGAA AAAGAAAATCT AGTCTCTCCA 3300
TGGAGAAGACA GACCTGACAT CTACTGAAAA ATAGACTTTA CTGGAATTTT TATGTTGATG 3360
AATACCTTCT AGTTTTTGGT AACGTTTCTCA AGATGGAATA AAGCTTTTTT TTTGAAAAAG 3420
AGACTGATCA GATAGTACATG AAGAAAGATG TTAAGAAGAA TTTTCCAAAGG TTTGGAGTGC 3480
CAGAAAGCAAT AGTGTCCGAT AATGGTTCTGG CCTTTGTGAC CACGGTAAGT CAGGGTGGTG 3540
CCAAGTATTT AGAGTGACAA TGAAAAATCC ATTTGTTGTA CAGACCTCAG AGCTCAGGAA 3600
AGATAAAAAA GAATAAATAA AACCTCTAAC AGACCTTGAC AAAATTAATC CTAGAGACTG 3660
GCAACAGACTT ACTGTTGACT CCTCTCGCCT GCCCTATTAA GAACAGGAAA TACTCCCTCT 3720
TAGATTGGTT TTAACCTCTT TAAAGATCCT ATATGCGATC ATGTGCTTAA 3780
ATAGTGGTT TAAACCTATG TTTGTATAAT ATGGATCTATG AATAAGCTGT 3840
CAGGTTGTGC AGAAAGAAAG CTGAGTCACA CTGAGTACAG TGAAACAGCT GGTACCCCA 3900
AGGCATCTCT ACCAGTTCCA GCCAGAGATC TGATCTACAG TCCCCGGGGT GACCCGCGGC 3960
GACCCCTGAG AGTTAGGTGG ATGGAGGGTT TGGAAAAGCT CAGAGCTCCC CAGCAGGCGA 4020
AAGTATGCAA AGCAGTCCATC TCAATTAGTC AGCAACCAGG TGTTGAAAGT CCCAGGCTCC 4080
CTTTATAGTC CTGTCGGGTT TCGCCACCTC TGACTTGAGC GTGATTTTTT GTGATGCTCG 6120
TCAGGGGGGC GAGACTATTG GAAAACGCC AGCAACGCGG AGATGGGCGC CCTCGAGAAC 6180
CCTGCCCTTA TATTGGGGT GACTAAACAT GGGGGGAAAT GGGGTGGGAA TAAAGGACAGG 6240
GACTACGCT CTAATGGGCA CTCAGCAGTTT CAGCCACGC CAAAGCCAGC TACAGGATGA 6300
TCTCAGGAG GTTGAAAAAT CAATCTCTAA CCTAGAAGAAG CTCTCCTCCTT CCCCTGCTGA 6360
AGTGTGCTCTA CAGAATGCAA GGGCCCTAGA CTTGTTATTTT CTAACAAGAG GAGGCTGCTG 6420
TGCTGCTCTA AAAGAGAARAT GGTGCTTCTA TGCGGACCAC ACAGGACTAG TGAGAGACAG 6480
CATGCGCAAA TTGAGAGAGA GGCTTAATCA GAGACAGAAA CTGTTTAGAT CAACTCAAGG 6540
ATGGTTTGAG GAGACTTGTAAA ACAGATCCCC TTGGTTTACC ACCTTGAATAT CTACCATTAT 6600
GGGACCCATTCT ATTGTATTCT TAATAGGTTT GTCTCTCGGA CCCCTGACTCC TTAATCGATT 6660
AGTCCAATTCT GTTAAGAAGCA GGAATATCAGT GTGCAGGCGG CTTAGTTTGA CTCAACACAATA 6720
TCACCAGCTG AGGCTCATTAG AGTACGAGGC ATAGATAAAA TAAAAGATT TATTTAGCTC 6780
CCAGAAAAAG GGGG 6795

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 9093 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

AATGAAAGAAC CCCAAGCTTA GGTGGCGCAA GCTAGCTTAA GTAACCCCAT TTTGCAAGGC 60
ATGAAAAAT ACATAACTGA GAATAAGAGA GTCAGATCA AGGTCCAGGA CAGATGGAAC 120
AGCCTGAAATG GGGCCAAACA GGAATACCTT GTGTAAGCGT GTCTGGCCCG GCTCGGGGCC 180
AAGAACAGAT GGAACAGCTG AATAGGGGCA AAACAGAGATA TCTGGTTGAA CGACTTCTCG 240
CCCGGCTCTCA GGGCGCAAGAA CAGATGGTCAC CAGAGATGCGG TCCAGGCTCTC AGCGATTCT 300
AGAGAACCAAT CAGATTTTTC CAGGGGCGGA CAAGACCTGG AAAATGACCT GGTGCCTATT 360
TGAACTAACC AATCAGTTGG CTTCTGGCTT CTGTTGGCGC GCCTTCTGCT CCACCTGCTCA 420
ATAAAAAGAGC CCAACACCCCC TACTCGGGG CGGCGATCCT CGGATGACT GAGTGCGGCG 480
GGTACCCGGTT TATCCAATAT ACCCTCTTCG AGTACCATCC GACTTGTGAT TCTGGCTGGCT 540
CTTGAGGGGG TCTCTTGCTA GTGATTGACT ACCCGTACGC GGGGCTTCTTT CATTGGGCGG 600
CTCGTCGGGG ATGGGAGGC CCCGTCCGCA GGAACCCAGA CCCACACGCG GAGGTGAAGC 660
TGCCCGCAAA CTTATCGTTG TCTGTCGAT TGTATGTGT CTATGACGTG TTTTATGGCGC 720
CTCGGTGGCT ACTAGTGGGC TAATCTAGCT CTTATCGGTC GGACCCTGGG TGGAACTGAC 780
<table>
<thead>
<tr>
<th>Position</th>
<th>Sequence</th>
<th>Length</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GAGTTTGGGA CACCCGCGG CAACCCCTGGG AGACGTCACCA GGAGGACAG GGAGGATCA</td>
<td>840</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>GGAGCAGCTG CTGAGACCCCT TTGGAAGGCA AGAGACCATTT TGGGGTGGG AGATCGTGGG</td>
<td>900</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>TCCAGTCCCC ACTCTGTCGCG CAGTGTGGAG ATCTGTCGTT CGAGTCCCAC CTCGTCCTTT</td>
<td>960</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>GTGGCAGGAT GTCGTTGCTGG AGTCCCCACT CGGGCTCTCT GAGGGGCTAG TGGGTTCGAG</td>
<td>1020</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>TCCACCTCG TGTTTTGTGT CGAGATCTGT GGTTGAGTGC CACCTCGCGG TCTGTCAGCAG</td>
<td>1140</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>GGATCGTG GG TTGAGTCGCC ACCTGTCGCA GAGGGTCTCA ATGCGCCGGC CTAGAGAGG</td>
<td>1200</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>CATCCTGTAT CTCTCTGTT CTCTTTTTGT CTTAGTGCTG TGTCGGCTTCTCTGTCGACT</td>
<td>1260</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>TATCTTTGTT TGTTTTGCTT TTTACGCTTG TTTGCTCTGT TCTTTGTGTCT</td>
<td>1280</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>CAGACTGCGA CTGTAGACTG ACGACTTTTT TTAAGTTATG CCTCTCTAAA TAAAGCTAAA</td>
<td>1320</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>CACGCTGTCA GATGCCCTATG CTGAGCAGCTT CTTCTGCAGAT CAACAGCTGC CTTACGCTAT</td>
<td>1380</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>CAGATGACCC CGAGACTAAC TAATAGCCCA TCTCTCAGGG TCAGACGGAG TCAATTCGCCGC</td>
<td>1380</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>CCCCCCTCTA AGTTACTGCG CCGAGGCGGC TGAGTTATAG GGCGGTCGTC GTTTCTCTAT</td>
<td>1440</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>ATGTATTTTT CACCATATT GCCGCTCTTTT GCCAAGCTGAA GGCGCGCGGAA ACTCTGCCCCT</td>
<td>1450</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>GTCTCTCTGA CGGACATCCC TATGGGGCTTT TCCCCCTCTC GCAAGAAGAT GCAAGGCTCG</td>
<td>1500</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>TTGAATGTGC TGAAGGGAGC AGTCTCCCTCT GAGCTCTCTG GAAGACAAAC AAGTCTCCTGTA</td>
<td>1560</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>GCCACCTTTT GCACGCGGCG GAACCCCCCA CTCGGGCGAA GTGCCCTCTG CGGCAAAAG</td>
<td>1560</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>CCACGCTGAT AAGATACACC TGCAAGGCGC GCACRACCCCC ATGCGCCACTG TGGAATTTGC</td>
<td>1620</td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>ATAGTTGCGG AAAGACTCAA ATGGCTCTCC TCAAGGCTAT TCAACAAGGG GCTGAAGGAT</td>
<td>1740</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>GCCAGAAAG TACCCCATTT TATGGGATCT GAATCGGGGC CTCGGTGCA ACGCTTTACA</td>
<td>1740</td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>TGTTTTATA G CTGGATTAAA AAAACGCTTA GGGCCCGCGA ACCACGGGGA CGTGGTTTCT</td>
<td>1920</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>CTTGGAATAA ACAGATAATA ATCATGGGGCG CGGATTCCCCG CTGTTTACA A C GTGCGTGA CT</td>
<td>2100</td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>GGGAAACCC TGGCCGTTACC CAACCTAAT C GCTTTGCGAC ACATCCCCCT TGCAGGCAGCT</td>
<td>2100</td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>GCCCTAATAG CGAAGGCTAC CGCGAGCTAC GCCTTTCCCA ACAGTTGGGC AGGCTGAATG</td>
<td>2100</td>
<td></td>
</tr>
<tr>
<td>115</td>
<td>GCCAATTGGCG CTGTCGCCCTG TTCCGGCGAC CAGAAAGGCGT GCCGGAAAGC TGCTTGAGG</td>
<td>2340</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>GCCATCTTCC TGAGCAGGGT ACGTCGTCGCG TCCCTCCTAAA CTGGGCAATG CACGGCTTAGC</td>
<td>2400</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>ATGCGCCCAT CTACCAACAT GTAACTATC CCATTACGGT CAATCCGGCG TTTGGTCACC</td>
<td>2460</td>
<td></td>
</tr>
<tr>
<td>130</td>
<td>CGGAGAATCC GACGCTGGTG TACTCGCTCA CATTATAATGG TGATGAAGGC TGCTACAGG</td>
<td>2520</td>
<td></td>
</tr>
<tr>
<td>135</td>
<td>AAGGCCAGAC GGAATATTAT TTTGATGGCGC TTAACGCGGC GTCTACGTGG TGCTGCCAGC</td>
<td>2580</td>
<td></td>
</tr>
<tr>
<td>140</td>
<td>GGGCGTTGGGT CGGTTACCGC CAGGACGGCT CTTGGCGCTGC TGAATTTGAC CTGAGGCGCAT</td>
<td>2640</td>
<td></td>
</tr>
<tr>
<td>145</td>
<td>TTTTACCCGC CGGAGAAAAAC GCCTCGCGG TGATGTTGGCT CCGTTTGAAGT GACGGCAGTT</td>
<td>2700</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>ATCTGGAAGA TCAGGATATG TGCCGGATAGA GGCGCATTTT CCGTACGC TC</td>
<td>2760</td>
<td></td>
</tr>
</tbody>
</table>
ATAAACCAGAC TACACAAATC AGCGATTTCG ATGGTGACAC TCGCTTATAA GATGATTCCA
GCGCCGGCTGT ACTGGAGGCT GAAGTTCCA GAATGCGGCA GTTGCGGTCAG GATCTACGG
5   GCTAAGTTAT CGATGAGGCT GTGTTTAGT CCGATGCGGAT CACACTAGCT CGAAGACGTG
AAAACCGGA ACTGTTGAGGC GCGAAATCTC CGAATCTCCTA TCAGGCGGGC GTTGGAACCTGC
10  ACACCCGGCA CGCCAGCTGCT ATGGAGCAATG AAGCTTACGA TGGTGCTTTTC CCGGAGGGGC
GGATGAAAAT TGGCTGCTGT CTGGTGAACGC GCAAGCGGTT GCTGATTGGA GCCCTTAACC
15  GCACCGAGCA TCATCCTCTG CATGTCCAGG TCATGATTGA GCAGACAGAG GTGAGGAATA
TCCTGACGTG GAAGCAAGAAC AACTTTAACG CGGTGCGCTT GTGGCATATA CGGACCATCG
20  CGCTGTTGTA CAGCCTGGGC GAGCCCTAGC GCTCTGTATGT GGGTTAGATG GCCAATATTT
AAAACCCAGG CATTGTGCGA ATGAACTGGC TGGCCGATGG CTACCGGGGA
25  TGACGACAGCG CTAGGTGCAGC ATGGGCGCTC GCAGGCAGTG GCTAGCATCT GTGATCCTCG
GGTGCAGTGG GAAAGTATCT GCCACCTGGC CTATACGAGC GGCCTGGTATG TGCTGATGCA
AATCTGCTGA TCTCTCCCGC CCAGTGCAGT AGTAAAGCGG CGGAGGCGCC ACCACCGCC
30  CGAGATTATAT TGGCCCGATG TACGCGCCCG TGGATGAAGA CCAGGCTTTCC CGGGCTGTG
CGAAATGGTC CATCAAAAAA TGGCCTTGCC TACCCTGGAA GAGCCCGCGC CGTAGCTTTT
35  CGGATATGCC CCACCGAGTG GTAAACAGTC TGGCCTGTTT GCCTAAATAC TGCCAGGGCT
TTGCTCAGTA TCCCGCTTTA CAGGGCGCCT TGCTCTGGGA CTGGTTGGAT CAGTGCCTGA
TTAAAATAGTA TGAAACGCGC AACCCTGTGG CGGTCTACGG GGGTGATTGT GCCGATACGC
30  CGAACGATGC CCAAGTTGTG ATGAACTGGC TGGCCGATGG CTACCGGGGA
CGCTGAGCGA AGGAACAGAC CAGCAGCACTT TTTTCACTGT CCGTATTCGC GGCGACACCA
40  TGGGAAAGTC CACTCAGGCTG TGGCAGATT AATTTTGCG ATAGCTAGCT CAGCTAGCGT
TGCGCGTGGA TGGTAAAGCG CTCGCGCCGC GTGGAAGTCC TGGTATTGCT GCCCAACAG
45  TAAACAGTTG ATGTGCACTG CCTGGAATAC GCAGGCGCGA GAGCGCGGGG CAACATCGGC
TCAGAGTACG CGTAGTCCGA CGGACGCCGA CCCATTGGTC AGAAGCGCGGG CACATCGCC
CCGCCGAGCA GTGGCCGCTG GCCGGAAAAAC TGACTGTCAG GCCTCCCGCG GCCTCCACG
50  CACACCGCGA TGCCGCTCTG GCCGGAAAAAC TGACTGTCAG GCCTCCCGCG GCCTCCACG
CCATCCCGCA TCTGACACCG AGCCGATATGG ATTTTTGGCA CGAGCTGGAT AATAGCGTGGT
55  GCGAAATTAA CGCGAGTCAA GCTTTTTTCT CACAGATGGT GATTGGCGAT AAAAAAACAC
TGCTGAGGCC GCTGGCGGAT CAGTTCACTC GTGACCGCCTG GTAACAAGCG ATTTGGCTTA
GTGAGACGC CGCGATTTGC CAACTTCGCTT GGTCACGCGA CTGGAGACGC GCCGGCGATT
60  ACCAGGCGA AGCACCGTTG TGGGAGTGTCA GCGAGATAC ACTGTTGGTAT GCGGGTGCTGA
TTACGACGGC TACCGCGTTG CAGCACAATG AAAAAACTT ATTATCAGC GCAGCAACCGC
65  ACCGGATTGA TGTTAGTGTT CAAATGCGCA TTACCGTTGA TGGTGGATAC GCGGAGCGAT
CACCCGACAG GGGCGCGATT GCCCTGAATC GCCAGCTTGC GCAGTTAGCA GAGCGGGCTA
98
PCT/US98/03918
CGATAGCTAG ACTGGCGGT TTTATGGACA GCAAGCGAAC CGAATTGCC AGCTGGGGCG 6840
CCCTCTGTTA AGGTGGGAAA GGCCTCGAAA GTAAACTGGA TGCTTCTCTT GCCGCGAAAG 6900
5
ATCTGATGAC GAGGGGATGC AAGATCTGAT CAAGAGACAG GATGAGGATC GTTTCGCAATG 6960
ATTGAAACAG ATGAGATTGC ACAGGGTTCT CGCGCGCGCTT GGGTGGAGAG GCTATTTCGC 7020
TATGACTGGA CACACGACAG AATCGCTGCG TCGTATGGCG CGGTGTTCCG GCTGTCAGCG 7080
CAGGGGGCGC CGGGCTTCTT TGATCAAGACC GACCTGTGCC GTGCCCTGAA TGAACCTGCA 7140
GAGGAGGCGG CGGGGCTATC GTGCGTGCCC ACGACGCGGC TTCTCTCGCC AGCTGTTGTC 7200
15
GAGCTTGTCG CTGAAAGGGG AAGGGACTGG TGCTTATGGG CCAGATGCGG GGGCCAGATT 7260
CTCTCTGTCAT CTCACTTTCG TTCTCGCGAG AAAGTATACC TACATGGCTGA TGCAATCGGG 7320
20
CGGCGCGATA CGGTCTGATCC GGCATCCTGC ACCATCGACC ACCAAGCGAA ACAATCCGAT 7380
GAGCGAGGAC GTACTCGGAT GGAAGCGCGT CTGTGTGATCG AGATATGACT GGACACAGAG 7440
CATCAAGGGC TCGCGCGCGC AGACCTGCCT GCAGCGCTCA AGCGCGCGCC GCCGCGCGGC 7500
GAGGATCTCG TCGTGACCAC TTGCGATGCT GTCTTCCGAA ATATCATGGT GGAAAGGCTC 7560
25
CGCTCTCTCT GATTATCGCA CGTGTGCGGCG CTGATGCTGG CGGACCGCTA TCGGACATA 7620
GGCTGGGGCTCA CCGGTGATAT TGCTGAGAAG CTGGGCACGG AATTGGCTGA CCGCTTCTCTC 7680
30
GTGCTTTTACG GTATCGCGGC TCGCGATTCC CAGCGCATCG CCTCTCTATCG CCTCTCTTGC 7740
GAGCTTCTCT GACGCGGACT CTGGGGTTCG AAATGACCGA CCAAGCCGAG CCCACCTGGC 7800
CATCAAGAGA GATGGATATCC GTCTGCGATC CAAACAAGAAA AACACGGGCT ACCAGCGGTTG 7860
35
TCGGGGCGGG AATCTGGAAT CTCGCTGCTT CCCCAAAAA AACACGGGCT ACCAGCGGTTG 7920
GTTGTTTGGC CGGATCAAGA GCTACCAACT CTCTCTCCGA AGTAAAACTGG GTTCAAGAGA 7980
40
CCCGAGATAC CAAAATCTGT CTTCTCTAGT TACCGGATGT TAGGGCCACA CTTCAAAGAAC 8040
TCTTTGACAC CCGCTTACAT CTTGCGCTCT TAAATATCGT TACCAGTGGC TGCTGCCGAT 8100
GCGCATAAGT CCTGTGCTTC CGTTGGTGAC TCAAGACGAT ATTTACCGGA TAAGGGCGAG 8160
45
CGGTGGGCTT GAGCGGGGAG TTCTTGCACA CAGCCGGACT TGAGGCGGAC GACCTACACC 8220
GAACCTGAGAT ACCCTACGCG TGAGCTTCTGA GAAAGCCGCA CCGTCTCGCGA AGGGAAGAAG 8280
50
GGGGACAGGT ATCCGCTTAAG CGCGAGGGTC GGAACAGGAG AGGCGCAGAC AGAGCTTTCCA 8340
GGGGGAAAGG CCTGGTATCT TTTAATCTCT GTCTGGGTTC GCCACCCCTG ACTTGGAGGT 8400
55
CGATTGTTGT GATGCTGCTG AGGGGGGCAG AGCCCAAGGA AAAAAACGCA AGACGCAGAG 8460
ATGCGGCCGC TCAGAAACC TGGCCCATTT ATTTGTTGGA CTAACCATGG GGGAATTGCC 8520
CGCTGGAATA GAAAGAAGGA CTCATTGCTT AACGCCCACT CAGCAATTCC AGCAGCTCCA 8580
60
AGCCGCAGTA CGAGAGATGC TCAGGAGGAT TGGAAAAATCA ATCTCTTACC TAGAAAAAGTC 8640
TCTCACTTCC CTGTCTGGAAG TTGCTTTACA GAATCGAAGG GCCCTGAGCT TGTTATTCTTC 8700
65
AAAGAAGAAG GGGCTGTTGT CTGGTCTAAA AGAAAGATGT TGCTTTATGC CGGACCACAC 8760
AGGACTAGTG AGAGACAGCA TGGCCAAATT GAGAGAGGCG TTTATCCAGA GACAGAACT 8820
GTTGAGTCA ACTCAAGGAT GGTGAGGG ACTGTTTAAC AGATCCCTT GGTTCACCAC 8880
CTTGATATCT ACCATTATGG CACCCCTCAT TGATATCCTA ATGATTTTG CTTTCGGACC 8940
CTGCCATTCT AATGCATTAG TCCAAATTTTGA TAAAGACAGG ATATCAGTGG TCCAGGCTCT 9000
AGTTTTGAGT CAACAATATC ACCAGCCTGA GCCTATAGG TACGAGCCAT AGATAAATAA 9060
AAAGATTTTT TTAGCTTTCG AGAAAAAGGG GGG 9093

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 46 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:
GACTAACCTT GATTCCTGG AGGGGGGGTT CTTTCATTG GGGGT 46

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 4834 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:
TGAAGAATAA AAAATTACTG GCCCTTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC 60
CCCTCCCATC TGAAAACAT ACTTGAGAAA AACATTTTCT GGAACAACCCA CAGAATGTCT 120
CAACAGGCCA GATGTATTCG CAAACACAGAT ATATGAACCTT TTGGTTGAGT AAATTTGGG 180
TGTTAACATT TTCCCCATTC CTTCCCCCATT CCCCCTCCCA GTTTGTTGTT TTTCCTTTA 240
AAAGGCTGTTG AAAAAATTGA GTGTGCTTCG AGACCTCCTC ACCTGTCAGA AAGGTTATAG 300
AGTTTCGACC CCAGAGCTCT GTGGGCTTTC TGTTGCTGCT TTTATTGCCA CCAAGAGTC 360
TGTTCCTGTG GCTTTTCACTG CGCTGCTTGA TAAATCTTA CTTCTACAT TTTATGTATG 420
GTCTCAATGT CTTCCTGCTG AGCGGCGCTG CCCGGACCTT GAGTGTCTGA GTGAGGCTCT 480
TCCCCGAGG GTCTTTCATT TGTTGACATGG GCCGGGAATT CGGAAATTCT TCATTGGTGT 540
CATTGGCCGG GAAATCCAAA ATTCCTACTG TTGGTGCATTG GCCGGAAAAAC AGCGCGACCA 600
CCCAAGGGTC CTGAGCCCCA TTAGGCTTAA GATTCTTGTG TCTGTTTTTG TCTGATTGCT 660
GTGTTCCTGAT GTCTGCTTGC TGTTTCTAAG TCTGGTGGCGA TCGCAGTTTC AGTTTTGCGG 720
ACGCTCAGTG AGACCACGCT CCGAGAGGGA GTGCGGGGTA GATAAGGATA GACGTTTCCA 780
GGGTCCACC CTCGTTTGCC CCTGGAGAAC GTCCGACGGG AAGCAGCGGA GGATCGAGGA 840
5
CGCTGGTTGG AGGCCTCTGG AGGAGGAAGG ACCATTGGGT GTGGGAGAGT CAGGTGTTCCG 900
AGTTCCACCT CTGGCCAGGT TGGCAGATGC TGGTTTCGAC TCCACACTCG TGTTTGTGTC 960
10
CAGGAATCGTG GGGTGGAGGC CCACTCCTCAG CTGTGTCGAC GTATGGGGTG TCTGAGTCCC 1020
ACCTCGTTGT TTGTGGGGAG ATCGTTGGTT CGATGCTGAG TCTCGCTGTC GTACGGGAT 1080
CGTGCGTTGC AGTCCACCT CTGTCGAGAG GTCTCAATTG GCCGGCCTTA GACAGGCGAT 1140
15
CTGATTTCTTC TGTTTTCTCT TTCGTCTTTA CTGCTCTGTCG CTGCTTTGTC GTGACTACTG 1200
TTTTCTTTAAAA AATGGGACAA CAGTGTCTCCA CCTCCCCCTTCT CTTGACCTCTG TTGGTCTGCCG 1260
20
TTGGAAATTTG TTGGTGGTTGA CTGGTTGTTTA GTGAGGTGCT TCTAGTTGTCG TGTTACTATC 1320
TTGTTTTTGT TTGTTGTTGA CGGTCTTTTTTG TGGTCTGCTG TCTTGTTTGC TGGTTTCAAGA 1380
CTTGGACTGA TGACTCACGA CTGTTTTTTA GATATCGGTT CTAATAAAGT CCTAAAATACT 1440
CTGTCAGATC CCTATGCTGA CCACCTCTCTC TCAAGATCAAC AGCTGAGCTG CTCGCCACCT 1500
CAACCTCAGA GACGGCGACG CGGTCGAGAC GTGGCTGGGAC CTGAGCCACC CATGAGTGGG 1560
AAAATGAGC TCGGAAATTC GAGGCAAGAT AGGAGTTGCT TCTGAGAAGT CAGTGCGCTTA 1620
30
AATGTTGTTG CGTCTGAGAC AAAAGAGAAG GAGGCTTTTTG GAGGCGCTTG GCCAGAGGCGC 1680
CGGGGCTGCC CAGCGAGCCT CTATTCCCC TTGCCCTCAGT AATCGGCTTCC TGTATACGAG 1740
35
AAAACCTCCT TTTTCTTTTG GATATGAGTG GCCGACTACA GCCGCCTCTG AAGAGCTGTAC 1800
TCCCTCTCCT GCCCGACCTGG TTTTCTCTTC TCAAGCGGCC ACCCCCGCCAG CTGGCTGCCAG 1860
TGAGTCTGTTT CTAAAGCTCC CGAGGAGGAC GTATCGCAGCA CGTGGGCTTCT CTGGTACAG 1920
40
TAAAGGACAC CTGTTGATCT GACTGCCAGG CTGCTGAGG AATTTGGTCT CTTGAGGCTG CTGGAGCT 1980
AGGAAGTGTC CCAACAATCT GACCAAGAGA ACAGGAGATG TAAGAACAAA CAGAGACAA 2040
45
AGTCAGAATC AGAGCTGTGC TGTGACACAA AAGATAAAA AAAAAAAATT GAGGGCCACA 2100
AAAGTCAGGA AAACCAACAA ACCTAGATAG TACCTGGCAA CAAAGAAGAG CTTTTGCTCA 2160
50
AGATCACAGG TGTAAGCAGT AAGAAGAATT AGCAGAGGGT GAGGAGACTGC CCGCAAAAAC 2220
TAAACAAAGG GGGCTTACA CAGAAGCTGA GTCCTAAAAA GAGTGGGCTG GCTTACCTGT 2280
55
TGAAGCAGCT AAAAAGAGGA CTGTTGTTCA TACTCTTCCA GTGCCAGTGA AAAAAACAGC 2340
TAAAAGTTGC CTGGCCACTG CCGGCTTTTG CAGATGTTGG ATTCGAGGTT TTGCTGAGTT 2400
AAAGAGAATA ACAAGCCTTC GTATAGAAAT AAAAAAAAAC ACCTGGATGT TCTTTGATG 2460
50
CTATTCAGAC TGCCCTGAATG TTGTCCCCAT CTATGGGACT CTTAGATGTG GACTGAGACA 2520
AAAGTATACGG CAAAGAAGTT CTATACAGA GATTTGGCAC CTGAAAAAAGA CCGTGAGCAT 2580
ACTTGTAAGA AATTGACCT CGTGCGCTGA AGATGGCCTG TTGGTCTGCA CATAGCTGGCT 2700
60
TCTGTTCAAG GAGCAGAGAT AATTGACTCT GAGACAAAAA TTGGCACATG TCCCATGAAAG 2760
TGTGGTTACG CCCCCATGAC CGATGGCTGA CTAACGCTCT TGAAACATT ATCCAACTGT

TCCCCCTACC GTAGGAGACA TTTTCGACACG TTTTTTCGAC TGAAGCGAGF ACCTTCGCTC

CCCTGCTAT CACGTGCTCT AACTCTGCTT GAGACTTCCAG CTACTCATCA TTGGGCTGAC

ATTGGGGAG AAGAACAACCA TACTGGACAT GATCTGGAAG ATCAGATCAG CTTGGGCTGT

AGAGTTTGAG CTGGTACAGG GATGGCAGTA GCTGGGAGGT TAAAGGTAAG CGGAAGCCGG

GGACAGCAGT CGAGTGGTGG AGCAAGAACG AGTATGCCTG GCCAGACGCC TCCTCTAAAG

GACTTCAAGC CACAAAGCCA AACTTGTGGC TTTRATACAA GCTCTGTAAG TGGTAAAAAA

AAAAAAGCTT ACACGGCAGC CAGGGTATGCT CTGGCAACTG TACAGAGCAA TATACAGACA

AAGAGAACCT GCTGACTCTG CAGAGAAAGA CTTAAGATCG TGTGGCTAAA AGAAATACGAA

TGGCAAAAAT CACGGCCGAG GATCTCTAAA GAGCAGATGT CCTGACAGTC TGAAGACTAT

CAAGTATTAG ACAAAATTAAG ACTGGTAAAA AAAACCCCTGTT ATAAAAATGTT AAAAATCGAA

AAAAGAACAA TAGTCTCTCTC ATGGAAGAGC AGACCTGAGA TCTACTGAAA AATAGAAGCTT

ACTGGAAAAA ATATGTGTAT GAATACCTTC TACTTGGTTG GAACGTTCTC AAGATGGATA

AAAGCTTTTC CTTGTTAATC ATCAGTCTGAT ATGAGAAGTT GATAAAAGAAA

APTTTCGAGG GTTCCGAGTG CCAAAAGCAA TAGTTGCAA GATAGTTGCT GCTTTTTGTG

CCAGGTTAG TCAGGGGGGT GCAAGATATT TAGGAAGCTCA ATGAAATTTT CATTTGTGT

ACAGACCTCA GAGTCTAGAA AGATAAAAA AAGATTAATA AAACCTAAA CAGACCTTG

AAAAATTTAT CCTAGAGACT GCCAGACACT TACTGGTAC TCCTCCCTCT TGCCCTATTT

AGAACTGAGA ATACTCCCCC TTGTATTGCT TTTAATCAC TTTGAGATCTT TATGAGGACT

CCTATCTCAA CACTGTCTTA AATGATGTTT TTAACCTTAT GGGTTTATAA TAATGATCCTA

TATGTTAAGTT TAAAAAGCTG GGGATGCTGG CAGAAAGCGA TCTTGCTACA ACTTGCTACA

GGGAACAGAC CTTGGTACCC AAGGACATCT TACCGTTCTC AGCCAGAGAT CGTGAATGCG

TACACCTTGG CCAAAGGCAGG ACTCCTCAACG CTCACCTAAA GGGCTCCCTTC CTAGTTGCTT

TTACTAATCT GCCATATTTA GTTTTTTTGC CCATTTAAA GATAGATCAA ATGCAAGATT

CTCCACATA GAGATATAGC TCTGAAATTT CTAAGATGTT AATTATTTAC AAGAAGAGCT

GGGGAATGAA GAATAAAAAA TATTCTGCCC CTTGGAAAAA CATGAACCTT CACCTCGGAG

CCACCCCCCT CCAACTGGAA AAAACATACTT GAGAAAACCA TTTTCTGGAA CAACCCACAG

ATGTGTTCAAC AGGCCGATATG TATGGACAAC CACAGGATA GACTCTTTTG TTAGTTAAT

TTGTTGTTGT TAAACTCCCA CTAATCCCTC CCCATCCCCC CTCCCAGTTT GTGTTTTTTT

CCTTTAAAAG CTGGGGAAAA ATTTGAGTCG TCAGTCGAGAC TTCTCTACC TGGCAAGG

TGTATGACCT TGAGCCCAAG AGCTCTGTTT GCTTTCTTTT GCTGCTTTAT TGGACCCCA

GAGCTCTGTTT TCATGCTGCC GCCTTTAAA ATCTTACCTT CTACACCTTA

TGTATGCTCT CAGTTGCTTC TGGGTGACGC GCAGCTGCGGC GAACCTGATG GTCTAGTGA

2820
2880
2940
3000
3060
3120
3180
3240
3300
3360
3420
3480
3540
3600
3660
3720
3780
3840
3900
3960
4020
4080
4140
4200
4260
4320
4380
4440
4500
4560
4620
4680
4740
GGTCTTCCC TCGAGGTCT TTCAATTGGT ACATGGGCGG GGRATTGCAG AATCTTTCCAT
TTGGTGCACT GGGCAGGAAAT TGCAAAATCT TTCA

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 4518 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

CACCTGACGC GCCCTGTAGC GCACGTAAAC ACCGCCGACG TGTGGTGTTT AGCCGCAGGC
TGACCCGCTAC AGCTAGCCGC CGCCCTGCCCC CTCTCTCCTC CTTCTCTCTC
TCGCCACGTT CGGCGGCTTT CCCGGTCAAG CCTTAAATAGC GGGGCTCCCT TTAGGTTGTT
GATTTATGCG TTCAGCCGAC CTGCACTCCCA AAAACTTGGTA TTAGGTTGAT GGTGGCAGTA
GTTGGCCCATC GCCTGTAGT AAGCCTTTTCC CCCCTTGGAC GGGGAGTGCAGC TTTTTCCCTA
ATATGCTGCT ACTTTCTCAA ACTGAACCAAA CAATCCACCT TACCTCCGTC TATTTTTTTG
ATTAAATGAG CATTGCGGCC ATTTGCTTAGA ATGGGTTTAA AATAGAGCTG ATTTAACAAG
AATTTAAGGC GAATTATAC AAAATATTA CGGTCCAAT TTACGGCTTA AGTATACATTG
ATGAGGGTTG GAAACCACCA ATCTAGAAGCC AGTGAAAAAT ATGCTTTTATG TGTGAAATTT
GTGATGCTAT GTCTTTATTT GTAACCATTA TAAGCGGAAA TAAAACATAAT ACAAACAGCA
ATTGCCATCA TTTATGTTT CAGGTTCAGA GGGAGGTTGT GGAGTTTTTT TAAGACAGAT
AAAAACCTCTA CAAATGTTGT ATGGGTCAAT ATGATCAAGA AGCACACTTG AGGACTGAG
GCCGTAAAAT GAGCCCTGGG ACTGTGAAAC TAAAATACAC AAACAAATAG AATCACTGAT
TTTACACATT ATACACTTAA AATTTGGACT TCATGGCCGC ATTCAGGCTG CGCAACTTGT
GGGAAGGGCG ATCGGGGCTG CCTCTCGTC ATTACGCGA CCGGCGAAA GGGGAGTG
CTGCAAGGC ATTAAGTGGG GTAAGCGGC AGGTTTTCGAAA GTCAGACGTT TGTAAAACG
CGCCAGTGGT ATGGTAGATAC GACTCAGTAT AGGCGGAATT GGGTACACTT ACCTAGTACC
CCACCCGGGT GGAATCTGA TGGGCGCGCG GGGCTCTTAG AAGTACTCCTC GAGAAGCTTT
TTGAAATTCTT TGATCTCACT AGTGTCGACC TCGAGGCGCG CGAAGCTTAC CTTTTGGTCC
CTTTAGTGAG GTTATTTTATG GAGGTGTTGG TAATCAAAGT CATAGCCTTT TCTGTGTTGA
AATTGTATTG CGCTCACAAT TCCACAAATA ATACAGGGCC GAAGTTAATAA GTGAATAGCC
TGAGGGCGCTT ATTAGCTGAG CTAATACAGA GTTAAGTGGG ATGCGGAACT TGACGGTTCA
CTAAAGGCGG TCGGCTTTAA TAGAACTCCC AACGTAACCG CTTAAGCTGCA ATTTCGCTCA
ATGGGGCGGA GTTTATTAGA TTTTTGGAA AGTCCCGGTTG ATTTTGTTGC CAAAACAAAC
<table>
<thead>
<tr>
<th>105</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCCCAT TGAC GTCAATGGGG TGGAGACTTG GAAATCCCG CGTACCAAAC 1500</td>
</tr>
<tr>
<td>GCCCAT TTGAT TAATGGC GACG CAGGACCTACA CCATGGGTAAT AGCGATGACT AATACGTAGA 1560</td>
</tr>
<tr>
<td>TGACTGGCAA AGTAGGGGAAG TCCACTAGG TACATGACTG GCGATAAT GCAGGCGGCCC 1620</td>
</tr>
<tr>
<td>ATTTACCGTC ATTAACGGCTA ATAGGCCGGC TACTTGGCAT ATGATACACT GATGACTCTG 1680</td>
</tr>
<tr>
<td>CCAA GTTTAGG CTAGTTACCT AAATACCTCA CCCATTGACG TGAATGGAAA GTCCCCATTG 1740</td>
</tr>
<tr>
<td>GCGTTACTAT GAGGAACTAC GTCAATTATG AGCTCAATGG GCGGGGGCGT GGGCGCGGTC 1800</td>
</tr>
<tr>
<td>AGCAGGCGGG GCCATTTACCG GAATAGTATG TAACCGGCGA CTCCATTATG GGGCTATGAA 1860</td>
</tr>
<tr>
<td>CTAATGACCC GTGAAATGAT TACTATTAAT AACTATGCA TGGCGGTAAT AGCGGTATCC 1920</td>
</tr>
<tr>
<td>ACAGAATCAC GGGTAACGGC AGGAAGAAC ATGTGACGAA AAGGCCGACG AAGGCGGACG 1980</td>
</tr>
<tr>
<td>AGCCGTAAAA AGGCGGCCGT GTCGGCGTTTT TCCCATAGGC TGCGCCCCCC TGAGAGCAT 2040</td>
</tr>
<tr>
<td>CA CAAAATTC GACGCTCAAG TCAGAGTTGG CGAACCACCGA CAGGACTATA AAGATAACAG 2100</td>
</tr>
<tr>
<td>GCGTTTCCCC CTGGAAGCTC CCTGTGCGGC TCTCTCTGTC CGACCTGGCC GCTTACGCGA 2160</td>
</tr>
<tr>
<td>TACCGGTCCG CTTTCCTCGG TTTGCGAAAGC CTCAATGCTC AGCGTGTAAGG 2220</td>
</tr>
<tr>
<td>TATCTCAGTT GCTGCGTCCG GTCGAGGGCA AAGCTGCGGT GTGTGCACGA ACCCCCAGTT 2280</td>
</tr>
<tr>
<td>CAGCCCCGACC GCTGCGCCCT ATCCGCTAAC TATCTGCTTG AGTTCAACCC GGTAGACAC 2340</td>
</tr>
<tr>
<td>GACTTATGCC CACCTGGGACG AGCCACTGTT AACAGGGATA CGAGAGCGGAG GATTGTAAGG 2400</td>
</tr>
<tr>
<td>GG TGCCTACAG AGTTCTGGAA GTTTGCGGT CTACAGGGCT AACTACGGCT AACTAGAAG CAGAGTAAT 2460</td>
</tr>
<tr>
<td>GTATCTGCGC TCTGCGTCGA GCCAGTTACC TTCGAAAAGG AGTTCTGTAG CTTCTGATCC 2520</td>
</tr>
<tr>
<td>GGCAAAACAA CCACCCTGGG TAGGCGGGTG TTTTTGGGT GCAAACGCAGA GATTACGGCC 2580</td>
</tr>
<tr>
<td>AGAAAAAAAG GATTCGCAAG AGATCTCCTTG ATCTTTCTTA CGGGTGCTGA CGCTCAGTGG 2640</td>
</tr>
<tr>
<td>AACGAAAACT CAGGTTAAGG GATTTTGTGC ATGAGATATT CA AAAAGGAT CTTCCACCTG 2700</td>
</tr>
<tr>
<td>ATCTTTTTAA ATTTAAATG AGTTTTAAA TCAATCTAAA GTATATATGA GTAACCTGAG 2760</td>
</tr>
<tr>
<td>GCTATGCGAG CCCTGCGCGC CCGACGTTGG GTGCGCGGCC TACGGCGTCC ACCCGAATT 2820</td>
</tr>
<tr>
<td>GGGGGTGAGG TGTTGCGGGG CAGAAGCGAT TTGGCGATTG GGGTTCTGGT 2880</td>
</tr>
<tr>
<td>GGGGTATCGA CAGAGTGCAC GCCCTGCGGAC CAGAACCGGC GTTTATGAAC AACAGCACCA 2940</td>
</tr>
<tr>
<td>ACACCCTGGG TTTTATTCCT TCTTTTTATT GGGGTTCTAG CCGGGGTTCG TTCCGGTATT 3000</td>
</tr>
<tr>
<td>GTCTCTCTCC GTGTTTCTGT TAGGCTCCCC CTAGGGGTGG CAGAAGACTC CAGCATAGA 3060</td>
</tr>
<tr>
<td>TCCCCGCCGGCT GGAGGATCAT CCACGCGGCG TCCCGGAAA CAAGACCGAA GCCCAACCTT 3120</td>
</tr>
<tr>
<td>TCATGAGGG CGCGGCTGGGA ATCGGAAATCT CTGTGATGGCA GGTGGGCGGT CGTGGTGGCG 3180</td>
</tr>
<tr>
<td>GTCAATTGCA ACCCCGAGTT CCCGTCAGA AGAAGCTGTC AAAGAAGGGCA TAGAAGGCGA 3240</td>
</tr>
<tr>
<td>TGCGTCCGCA ATCAGGAGCC CGGATACCGT AAGACCGGAG GAAGCGGCTCA ACCATCCGCG 3300</td>
</tr>
<tr>
<td>CGCCAAGCTCA TTCAAGCAAT TAACGCTGTTA CCAACGGCAT GTCTCTGTAAG CGGTCCCCCA 3360</td>
</tr>
<tr>
<td>CACCCAGCCG GCCACAGTCG ATGAACTCAG AAAAGCGGCC ATTTTCCACC ATGATATCG 3420</td>
</tr>
</tbody>
</table>
GCAAGCAGGC ATGCCATGG GTACGCAGCA GATCCCTCGCC GTCGGGCATG CTCGCTTTGA 3480
GCCTGGGCAA CAGTTCCGCT GGCAGCAGCC CCTGATGTCT TTAGCTCCAGA TCATCTGTGAT 3540
5
CGACAGAGCC GCTTCCATCT CGAGTACCTG GTCGCTCGAT GCGATGTTTC TCTTGGTGTTG 3600
CGAATGGGCA GGTAGGCCGA TCAAGCTGTAT GCAGCCGGCC CATTGATCATG GCCATGATCG 3660
10
ATACCTTTCTC GCCAGGAGCA AGGTGAGATG ACAGGAGATC CTGCCCCGGCC ACTTCGCCCAC 3720
ATAGCAAGCA GTCTCCCTCC GCTTCAGTGA CAACGTCCGAG CACAGCTGGC CAAGGAACGC 3780
CCGTCGTGCG CAGCCAGATG AGCCGGCCTG CCTGCTCTTG CAGTTGATTT AGGGACCCGG 3840
15
ACAGGTGCGTT CTTGACAAAAG AGAACCAGGGGC GGCTCTGGCG TGACAGCGGCG AAGCCGGCGG 3900
CATCAGAGCA GCCGATTTGCT TGTTTGCGCC AGTCATAGCC GAATAGCCATC TCCACCAGGAAG 3960
CGGGGAGAGGAA CTGCGGTGTC AATTTATCTT GTTCAATCAT GCAGAAAGAT CCTCATACTTG 4020
20
TCCTCTGATAC GATCTTCTGA AAAGCCTTAGG CTTCCAAAAG AGGCCCCTCTCA CTACTTCTGG 4080
AAATAGCTCAG AGGCGCCAGCC GGCTCCGGCC TCTGCTATAA TAAAAAAT TACGTCAGCCA 4140
25
TGGCGGCGAG AATGGGGCCGA ACTGGGCGGA GTTAGGGCCG GCGAAGCGCCG AGTTAGGGGC 4200
GGACTATGG TTGCTGACATG ATGGAGATGC ATGCTTTTGCA TACTTCTGCG TGCTGGGAGG 4260
CCTCCGCACT TCTCCACACT GGTGCTGATT CAATTTTGTAG GCACTGTTTGC TACACTCTCG 4320
30
CCGTGCGGCG GAGCCTGGGGGA CTTTCCAGAG CCTAAGTGAG ACACAGAGTCC AAGCTGGTTC 4380
TTTCCGCGTTC AGGACTCTCTC CTTTTTGATG ATTAATGAGA CATTATAGAG GGTATATTGTC 4440
35
TCATGACGGTT ATACATATTT GAAATGTATTA AGAATAAATAA ACAAAATAGGG GTTCCGGCAG 4500
CATCTTCTCC AAAAGTGC 4518

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 25 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:
CTCCACATAG AGATATAGAC TTTCTG 25

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 45 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

CGATCTTAAT ATTAACTGG AGTTTTGAGC CCRMCCCCCTC CCATC

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 5594 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xii) SEQUENCE DESCRIPTION: SEQ ID NO:29:

TGCATTAGTT ATTAATAGTA ATCAATTACG GGGTCATTAG TTCA TACGCCC ATATAGGAG

TTCGCGGTTA CATACCTAC GTAATAGGCC CGGCCCTGCT GACCGCCCCA CCACCGCGCG

CCATTGAGCT CAAATAGGAC TGTAGCTTCG ATAGTAAACG CAATAGGAC TTTCCATTGA

CGTCAATGGG TGGAGTATTT AGCGTAAGACT GCCCACTGG CGATACATCA AGTGATATCAT

ATGCCAACTGA CCGGGCCCATG GACCGTAAAT GCGCCGCCTG GCATATTGCC

CAGTACATGA CCTTATGGGA CTTTCCACTC TGCGAATACA TCTACGTATTT AGTCACTGGCT

ATTACATTGG TGAATGCGTT TGGCGATATCA ATCAATGGGC GTGGATAGGG GTTGGATCTCA

CGGGGATTTCC AAGATCTCCCA CCCATGGGAC GTCAATGGGA GGGTTTGGGG GCACCAAAAT

CAACGGGACT TTCCAAAATG TCGTAAACAC TCCGGCCCAT TGACGCAAT ATGCGGTAGG

CGTGTACGGT GGAGGGTCTA TATAAGCAAG GCTGGTTTATG TGAACCGCAT GCACCGGCC

AGTCTCTCGGA TGACTGAGT CCGGCGGTTA CCGGTCGATA TAATAAACAC TCTTGCAGSTT

GCATCGGACT TGTGCGTCTCG CTGTTCTCTG GGAGGGTCTC CTGTGAATTG TGGACTACCC

GTCAGCGGGG GTCTTTCTATT TGCGGCGTCG TCCGGATCGT GGAACGCCCT GCCAGGGGAC

CACCGACCCA CCCAGGGGAG GTAAGCTGCC CAGCAAATCA TCTTGCTGTC TCGGATTCGT

TAGTGCTCAT GACGTGATTGG AGCGGTGCTG GTCGGTCTCA GTGACGCTAC TAGCTCTGTA

TCTGGCGGAC CGGCGGGTGA ACTGACGAGT TCGGAACACC CGGCGCGCAAC CCTGGGAGAC

GTCCCGAGAG GACAGGGGGA GAATCGGGGA CGCGCTGGTG AGCCCTTTTG AGGCCAACAG

ACCATTGGG GTGGCGGATG CGGCTGGCTG ATGCCCACTG GTGCACGCTG

TGGGTGCAG TCCGACCTCG TGGTTTGTGG GAGATGCTGC GTTCGACGT CCACCTCGC

TCTTGTCAGC GCATGCGTTTG TCAGGTGCC ATCCTGTTG TGGTGCGGAG ATCGGGGTTT

CGAGTCCCAAC TCGGCCGTCG GTGACCGGAT CGTGGTTTGC CGTCCGACTCC GTGCAAGGG

GTCTCAATTG GCCGCGCTTA GAGAGGCCCT CTGATTCTTG TGGTTTCTCT TTTTGTCTTTA

GTCGCCTGTGC CGCTCTGTGT GTGACTACGT TTTTTCTAAA AATGGGACAA TCTGCTGCCA

CTCCCTTTCCT TCTGACCTCTG GGTGGGCGTGG GTGGTAAATT TGGTTGTTTA CGTTTGGTTT
ACGGTGCTGA CCCCAGATGA ATGTCAGCTA CTGGGCTATGC TGGACAAGGG AAAACGCCAG
3480
CGCAAAAGAG AAGGAGCTTG CTTGACAGTG GCCCTACTAGG CGATAGCTAG ACTGGGCCGT
3540
TTATGGGACA CCAGGAGAAC CGGAATTGGC AGCTGGGCCC CCCTCTGTGA AGGTGAGGAA
3600
GCCCTGCATA GTAAACTGTG CTGGCTTTCTCT GGGCCCAAGG ATCTGATAGGC GCAGGGGATC
3660
10 AAGATCTGTA CAAGGAACAG GTAGAGGACTT GTTTGCGATGG ATGGAAACAG ATGGATTGCA
3720
CGCAGGTTCTT CGGCGGCTTTT GGGTGAGAGACTTATCTGGA TATGACTGGC CACAACAGAC
3780
AACCGGTGCTC TTGATGGCAGCG CACTGCTGCCG CTGGTGACTGTG CAGGGGCGCC CGTTCTTTTT
3840
15 TGTCAAGACC GACCTGTCGG TGGGCTGAA TGAACGTGAC GACGGAGCCG CGCGGTCTATC
3900
GTGGCTGCGC ACCGACGGGC TTCTCTGCGC AGCTGTGTCC GAGCTGGTA CAGGAACGCGG
3960
20 AAGGGCATGG TTCTGTATTG CCGGAATGCC GGGCCAGGAT CTCTGTCTATC TCGACCTGGC
4020
TCCTGCAGAC AAAGATCCCA TCTACGTGCTA TGCAATTGGC CCGTCATAGA CCGTTAGATCC
4080
GGCTACCTGC CCATTGACCC ACCAGCCGAA ACATCGCATC GACGGGACCG TACTCGATAT
4140
25 GGAACGGGTT CTGTTGCTATG ATGAGATCTC GGACGAGAGT CATCAGGCCC TCGCGGCAAGC
4200
CGAACTGTTC CGAGAGGTCTC AGGGCGCTGCA CGGGAGGATCC GTAGTGACCGA
4260
TGGCGATGCC TGGCTGCGGAG ATATCATGGA GGAAATGGC CGTGGTTCGTG ATTCACTCGA
4320
CTTGGCCGAG CTGGTGTGGC CGGACCGCTA TCAGGACATC GCGTGGACCA CCCGTGATAT
4380
TGGCTGAAAGG CTGGCCGCGC AATGCGTGG CGGCTTCTTG GTGCTTTAGG GATACGCCGC
4440
35 TCCGGATTGG CAGGCGCATCG CTTCTCATCG CTTCTGTGAC GAGTTCTTCT GACGGGAGCT
4500
CTGGGGTTCG AAATGACGCA CCGACCGAGC CCCACCCCTCC AGAAAAAGGG GGAATGAAAA
4560
40 GACCCCAACCT GTAGGTTGGA CAAGCTAGCT TAAGTAAAGC CATTTTACAA GCCATGGAAA
4620
AAATACATAAC TGAGATAGA GAAAGTTCGA TCAGGCTAG GAAAGATAGG AACAGCTGAA
4680
TATGGCCCAA ACCAGATATCC TGTGGTAAGC AGTCTCGGCC CGGGTCAGCC CCAAAGAACG
4740
GATGGGAACAG CTGAAATTTG GCCAAACAGG ATATCTGTGG TACGCCTTGC CTGCCCAGGG
4800
TCAAGGCCCA AAGCAAGATCG CGTTCAGCCG CTAGCAGATT TCTAGAAGAC
4860
CAPCAGATGT TTCCAGGTTG CCCCCAGGAC CTGAAATGAC CTGGCTGCTT ATTGGAACTA
4920
50 ACCATAGGCT TCCGTCCTTC CTCTCTGTGG CGGGCTTTCG CTCTGGCAAG TCAGAATTAAG
4980
AGCCCACAAC CCTCAGTCC CGGCGCCAGT AATCTCGGTG TTGCAAAAGA AAAAACACC
5040
55 GCTACGAGGC GTGTTGGTTT TGCCGGGATCA AGAGCTACCA ACTCTTTTTT CGAAGGTAC
5100
TGGCTCTGCG AGAGCGCAGA TACCACAATAC TGCTCTCTTA GTGTAGCGGT AGTGAAGCCA
5160
60 CCACTTCAGG AACCTCAGGT CACCCGCCTAC ATACCCTCGCT CTGCTTAATCC TGTTACCAGT
5220
GGCTGGCTGC AGTGGCGATA AGTGCTGGCT TACCGGGGTTG GACTCAAGAC GATAGTTACC
5280
GGATAAGGCG CAGCGGTGGG GCTGAACGGG GGGTTGGTCG ACACAGCCA GCTTGAGGCG
5340
65 AAGGACCTAG ACCGAAGCTGA GATACCTACA CGTGAGGACT TGAGAAGACG CCACGCTTCC
5400
(2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6561 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

GATCCCCGGG TCGACCCGGG TCGACCCGTT GAAGTGTGAG TGAGTTAAGG TGTTGAAAGT 60
CCCCAGGCCT CCCCCAGGC AGAAGTATGC AAGCCATGCA TCTCAATTAG TGAGCAACCA 120
GCTGTGGAAC GTCCCCAGGC TCCCCAGGC AGCAAGTATG GCAAAGCATG CATCTCAATT 180
AGTCGCAAC CATACTCCCG CCCCCCTAATCC CCGGGACATCC GCGCTGTATT 240
CGGGCCATTG TCGCCCCCAT GGCAGATCAG TTTTTTATT TTATGCAGAG GCGGGCGCG 300
CCTCGGCCCTC TGAGCTATCC CAGAAGTAGT GAGGAGCTAT TTTGGGAGG CTAGGCTTTT 360
GCAAAAAAGCT TCACGCTGCG GCAAGCACTG AGGGCAGGAG GCTGCTAAA GGAAGCGAAA 420
CAGTAGAAGGCCAGTCGCGG CAGAACGAGT CGAGCCCGGG ATGAAGTATCA GCTACTGCGG 480
TATCTGGACA AGGGAAAGCG CAGGGCAGAA GAGAAGACAG ATGAGCTTGG ATGGACCGTAC 540
ATGGCAGTAG ATGACTGGG CGTAAAAATG GACAGCAAGC GAACCCGAAT TGCCAGCTGG 600
GGGCCCTCTG GGAAGGTGGA GAAGCGGGCG CAAAGTAAAC TGGATGCGTT TCTGCGGCGC 660
AAGGATCTGA TGCCGCGGAT GATCAAGACT TGATCAGAGG AACAGTACAGG ACAGTCTTCG 720
CATGATGAA CAAGATGGAT TGCCGCGGG TTCTCGGCG GCCGTGGTGG AGAGCTATT 780
CGGCTATGAC TGCCAGGCAA AGCAATCTCG CAGCTCTCTG GCCGCCGCTG TCCGGCTGTC 840
AGCCAGGGG CGCCCGGTTT TTTTTGTCAA GACCGAGCCTG TCCGGGCTGC TGGAAATGAC 900
GCCAGAGGAG GACGGCGGCC TATGCTGGCT GCGCCAGCGG GGGTCCTCCC GGGCGCTGGT 960
GCTCGACGGT GTCACTGGAAG CGGAAAGGGA CTGCGTCTAT TGGGGCAAGA TGGCGGCGGCA 1020
GGATCTCCGG TATCTCCAGC TGGCTCGTGC CGGAAAGATA TCAGACTGAG CTGGCATGAT 1080
GCGGGCGCTGT CATTCCGCTT ATCCGCGTAC CTCGCCCATT GACCACCAA GGAACACTCG 1140
CATCGCAAG ACCCGTACTG GAGTGAAGGC CGTCTTGTGC GATCAGGATG ATGCGGTGCA 1200
AGAGCACAG GGCGCGCCG CAGCGCCACT GTCCGCGAG CTGCGGCGC CGATCGGCGA 1260
CGCGAGGAT CTGCTCGGAG CCGATGCGGA TGCGCTGTTG CGGAAATGCA TGGCGGCAAAA 1320
TGCCGCTTT TCTGGATTCA TCGACTGGG CCGGCTGGGT GTGGCGGACC GCTATCAGGA 1380
CATAGCGTTG GCTACCCCGT ATATGCTGTA AGAGCTTGGG GGGCAGATTGG GTGACGGCTT 1440
CTCGTGCTTT TACGGTACGG CCGCTCCCGA TTGGCAGCGG ATCAGCTTCT ATCAGCTTCT 1500
TGACGAGTTC TCTGGAAGGG GACTCTGGGT TTTGAATAAG CCGACCAAGC GACGCCCAAC 1560
CTGCCCATCAC GAAGATCGGG TACCCCGGGC GCCTCTATAG AAGGTTGGG CTTGGAATCT 1620
GGTTTCCGGG AGCCGGCTGT GATGATCCTC CATCAGCTGCT GAGGTTTCTC 1680
GCCACCCCCA GAATTTGCAA TCTGCTGCTT GCAAAACAACA AAACCACGGC TACACCGGT 1740
GGTTTGGTTT GCGGATCAAG AGCTACCAAC TCTTTTCCGG AAGGAAACTG GTCTCAGCAG 1800
AGGCCGAATA CCAATAACTG TCCTTCTAGT GTAGGCGGAT TTAGGCCACC ACTCTCAAGAA 1860
CTCCTGTAAG CCGCCCATAC ACTCCGCCTCT GCTAAACCTG TTACCAGTGG CTGCTGCGCA 1920
TGCCGATAAG TCGTGCTTCA CCGGGTGGGA CTCAGAAGGA TGGTTACCGG ATAAAGCAGCA 1980
GGGTCGCGCC TGAACGGGCTT TTGGTCGACAC ATACCGGCGG TTGGAGGAGA CGACTACAC 2040
CGAACCTGAG AACTCTACCG GTGACGATTTG AGAAGACGCG ACAGCTCCCG AAGGAGAAA 2100
GGGGGACAGC TAGCCGCTGA CCGGACGGGT CCGAAGAGGA GAGGCAGCAG GGGCAGTCCC 2160
AGGGGGAAAC GCTGCTGATAC TTTATAGCTC TGTCGGGTTT CGCACCCTCT AACTTGAGGG 2220
TGATTTTTTTT TCACTTGGGT CTGGCGTGTTT CAGCCACTTT GATTGAGGG 2280
GATGCCGGCG CTCGAGTACA CTTGCGCTAT GCTAGAGCCG TTAAACCTGCT CTTAGGGG 2340
CCCTGCTTAG TTCTGTTTTAC TAAATCAGCT TATTCGTTTT TTGTCTCCGT GTPAAAGATA 2400
GAGTAAATAG AGTATTCTCC ACAATAAGAT ATAGACTTCTT GAAATTTCAA GATTAGAATT 2460
ATTGACAGG AGAAGGGGTTT AATGAGAATA AAAAAATTAC TGGCCTATTG TGAGAACAGT 2520
AACCTTACCG TCGGACCGCA CCCCCTCCCA TCGTGAACAT ATACCTTGGAG AAAAACTTTT 2580
CTGGAAACAC CACAGATGTT TCAACAGGCG CAGATGTATT GCAAAACACAA GGATATGACT 2640
CTTTGTGTGA GTAAATTTGT GGTGGTTAAA CTTCCCTCAT TCTCCCTCCA TTCCCCCTCC 2700
CAGTTTTGGG TTTTCTCTTT TAAAGCTTGG GCAAGATTTG TGGTCTGCAG CGAGACTCCT 2760
CTACCCCTGTTT CAAAGGTGTGA TGATTTTCGA CCCCAGACGT CTTGTGCTTT TCTGTGCTTG 2820
CTTTATTTTG ACCCCAGAGC TCTGGTCGTT GTGCTTCTCAT GTCAGCTGCTT TATTTAATCTT 2880
TACCTCTTAC ATTTATGTTA TGGTCTCAGT GTCTTCTTGG GTACGCGGCT GTCCCGGGAC 2940
TTGAGTG CCTC TCTCCTCGAG GGGTCTTTCA TTTGCTACAT GGGCGCGGAA 3000
TTGGAGAACCT TTTCTTCGGC TGGATTGCCG AAATCTTTCA TTTGTTGCTA 3060
TGGCCGGGAA ACAGCGCCAC CAGCCAGAGG TTCTAGACCC ACTTGAAGGG AAGATCTTCTT 3120
GTTGCTTTTTT GTGCTGATGT GTGCTGCTTGT ATGGTCTGGTT TCTGTTTCTCA AGTCTGTGTT 3180
GATCGACAGTT TCACTTCTTG GAGCAGCTCA TGAGCGCGG CTCCGAGAGG GAGTGCGGAA 3240
TGGTAAAGGA TAGAGGGTGTC CGGCTGCTCA CGGCTGGGCT GCCCTGGGAG ACGTCCCGAG 3300
AGGAACAGGG GAGGATCAGG GACGCCCTGGT GGACCCCTTTT GAAGGCCAAG AGACCATTTG 3360
GGGTTCGGAG ATCGTGGTGT CGAGTCCCAAC CTGCTGGCCAAG TTGGACAGATG CTGGAATCTG 3420
AGTCCCAACCT CGTTTGTCTGT GCCGAGATCG TGGGTTCCAG TCCCCACTGCT CGCTCGTCTCA 3480
CGGGATCGTG GTTTCAGAGTC CAACTCGTGT TTTTGTGGCG AGATCTGTTGG TTCCAGTCCC 3540
ACCTCGCGTC TGCTGACGGG ATGTTGGGTTT CGAGTCCCAAC CTGCTGCAAGA GGTTCCTAAT 3600
TGCCGGCCCT TAGAAGGAGCC ATCTGATTCTG CTGGTTTCTT CTGTTTCTCT TAGTCTGCTG 3660
TCGCTCTCTT TTGCTGACTAC TGGTTTTTCTAA AAAATGGGAC AAATCTGTGTC CACTCCCTT 3720
TCCTCTGACTC TGTTCTGCTGT GCTGTTGAAT TTGGTTGGT TTAGTTTTGT TACGTTTCTTG 3780
TGCTATGTTG TGCTTACTATA TCTGTTTTTT TTGTTGTGTTG TACGTTTCTTG TGCTGGTCTT 3840
TGTTGTGTCCT TTGGTGTTCTG ACGTGGGACT GAGGAGTGGC GACGTGTGGGGT AAAGTTTGGC 3900
TTCTAAATA AGCCCTAATAA TCTCTGCAAGA TCCCATATCTG ACCACCTTCC AGGACATCCTA 3960
ACAGCTGCCGC TGCTCCTCCAC TCCAACTCCAA GAGAGCAAGC AGCGGCTCAC AGTGTTCCCG 4020
CCCATGAACC TGGAGGCTAG GGGAAAAATGA GCTGCGGAAAT AGCGGACAAA TGAGGATTGG 4080
TCCTGAGAAA GCACTGGGCA TAAATGGTGT GCCGGTTGCA GCAAAAAGAG AGAGGCGCTG 4140
TGGAGATGCC GGGCAAGAGGC GGGGGGCTGT CCCAGGCGAC TTCTCATTTCC CTGGCCTCCT 4200
CCATCCCGTC TCTGTGTAAC AGAAAAACTG CTTTCACTTTT GAGATATGAG TGGCCGCGATA 4260
CAGCGAGCTG TGAGAGCCTT ACTCCCTTTCC CTGCCCCACAG TGTTTCTCTCT TCTGACCGCA 4320
CCCTCGCTTG AGCTGCTGAGC AGTGGATCTGT TCTCTAGACT CGAGTGAGGG AACATCGGCC 4380
CAGTTGGGCG TCTCTGCAAGA GCTAAGGAGC AACCTGAGAT GTAGATCCGAG 4440
GGGGTCGGAG TCTCTGTTGGA CTGAGAAGTGT TCCAAACACAT CTGACCAAGG TACACAGAG 4500
TCAAGACAAA GACAGAGACC AAAGTGCAAGA TCAGAGCTGT GCTGTTGAGAC AAAAAGATAA 4560
AAAAATAAA AGTCTGGGCA CAAAGATCTG CAAAGACTGAA AAAAACTGGA AGTACCGGAG 4620
AACAAAAGAA AGCCTTTTGGC TAAAGATCAA CGTGTATACT GTAAGAGAAA TGAGCAGCTT 4680
GTGAGAGACT GCCCCAACAAG AAAAGAGGAG AGCCCCCCTC ATGCAACAAAC CTTGCACTGG 4740
TTCTGGGCTA AAAGTAAAGGA GATAACAAAG GGGTGCTCAA CACAGAGCGT GAGTCTCTCA 4800
AAGAGTCCGG TGCCCTACCT GTGGAACAGA CTAAAAAAAGA GACTGTTTCTT CATACCTCTC 4860
CAGCAGACAG TGGAAAAACAA GCTAAAAAGT TCTGGGGCAC TGCGGGCTTTT TGGAGATTGT 4920
GGATCCAGG TTTTGCTGAG TTAAGAGATA AAACAGCGCT TGATGATAGAA AAATAAAAAA 4980
CAACCTGTTG GTCCTTTGGA TGCTATTGAG ACTGGCCTTA TGGTTGCTCCGC AGCTATGGGA 5040
CTCCTAGAGC TGACTGGAAG CAAAGTATT GCAAAAGAGA TCTTTACTCA GAGATGGCA 5100
CCCTGAAAAA GACCTGCTGGG ATACCTGTTA AAAAAAGAC TGGTGCTGGAG TAGAGAGGCC 5160
TGCTTGCTCTG CACATAGTGG CTCTGCTGCA AGAGCGCAGA TAAATGACTG CTGAGACAAA 5220
ACTTGGCACA GTGCTGCTAGA AGTTGCTGTC AGCCCCCAGT GCAGATGGCT GACTAAGCT 5280
CTTGAARACA TTATCCCAAC TCTCCCTGCA CGAGTACGACA CATTGTCAGA GGTCTTTTTG 5340
ACTGAACGAG TGACCTTCGC TCCCCCTGCT ATCCTCGATC TCACCTACTGC CTGAGACTTC
ACCTACTCAT CATTGTGCTG ACAATCTGCG AGAAGAAACT CATACTGCGA AATGACTGAA
GGATCAGATC AGCCCTGCCC TGAGAGTTTG AGCTGGTACA CGGATGCGAC TAGGCTGGAG
GTTAAGGGTA AGCGGAGGCG GGGGACAGCA GTGCACTGGT GGACAGAAGA CAAGTGATCT
AGGCCAGCAG CCTCCCCAAA GGGACCCTCC CCCCCAAAGC CAAACTGTGT CACCTTAATAC
AAGCTCTGTA AATGGTAAA AAAAAAAGT CTACACGGAG AGCAGGTATG CTCCGCCCAC
TGTTACAGAGC AATATACAGA CAAAGAGAAC TGTTGACATC TCCAGAGAAA GACCTAGAT
GCTGTTGCTA AAAGAATACTA GATGGCAAAT CTAACCGCCC AGGCACTCCTA AAGAGCAAAT
ATCCCTAGAC TCGGAAGACT ATCAAGTTAT AGCAAATATAG AGACTCGTAA AAAAACCCT
GTATAAAATA GTAAAACCTG AAAAAGAAAA ACTAGTCCTC TCAGGAGAAG ACAGACCTGA
CATCTACTGA AAAATAGACT TTACTGGAAA AAATATGTGT ATGAATACCT TCTAGAAAA
GTGAACGGTC TCAAGATGGA AAAAGCTTT TCTCTGTAAA ACAGAGCTGA TCGAGATGTC
ATCAAGAAAG TTGTTAAAGA AAAATTTCGAG AAGTTCCGAG TGCCAAAAGC AATAGTGCTA
GATAATGGTC CTGCTTTGTG TGGCCAGATT GAAAGGTGCT TTAGAGGTC
AAATGAAAAT TCCATTGCTG TTACAGACCT CAGAGCTCAG GAAAGATAAA AAGAAATRA
TAAAACKCTA AACAGACCTT GACAAAAATA ATCCCTAGAGA CTGCACAGAC CTACTTGGG
ACTCCTCCC CTTGCCCTAT TTGAAGACTG GAATACTGCC TCTTGATCTG GTTTTACCTT
TTTAAAGATC CTGTATTGGG CTCCTATGCC ATCAGTGCTT AATATGATGT GTTTAAACCT
ATGTTGTTAT AAATAAGATC TATATGTTTA GTAAAAGGCT TTGCAGGTTG TGCAAGAAGA
AGTCTGCTCA CAACTGGCTA CAGTGAACAA GCTGGGTACC CCAAGGACAT CTATTACGTT
CCAGCCAGAG ATCCTGACCTA C

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:
   (A) LENGTH: 55 base pairs
   (B) TYPE: nucleic acid
   (C) STRANDEDNESS: single
   (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

GACTAACCTT GATTCCACTG GAGCCGTATT ACCGCCATGC ATTAGTTATT AATAG

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:
   (A) LENGTH: 47 base pairs
   (B) TYPE: nucleic acid
   (C) STRANDEDNESS: single
   (D) TOPOLOGY: linear

(55)
(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:
GAC'TAACC'CTT GAT'TCCACTG GAG'TAATT'G GGC'TAGC'GGA TCTGACG

(2) INFORMATION FOR SEQ ID NO:33:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 66 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:
GAC'TAACC'CTT GAT'TCCACTG GAG'ACACT'G GAC'TCTACCG CAC'CCAGTCCT CGA'TTGAC'T

GAG'TCG

(2) INFORMATION FOR SEQ ID NO:34:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 48 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:
GAC'TAACC'CTT GAT'TCCACTG GAGG'GATCCG CGG'CAGTATG TATTATCG

(2) INFORMATION FOR SEQ ID NO:35:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 55 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:
GAC'TAACC'CTT GAT'TCCAGCA ATG'CATGGC TAC'AGGCTCC CGGACGTC'CC TGC'TC

(2) INFORMATION FOR SEQ ID NO:36:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 48 base pairs
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear  

(ii) MOLECULE TYPE: DNA (genomic)  

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:  

GACTAACCTT GATTCCAGCA ATGTTAGGAC AAGGCTGGTG GCCACTGG  

(2) INFORMATION FOR SEQ ID NO:37:  

(i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 49 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear  

(ii) MOLECULE TYPE: DNA (genomic)  

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:  

GACTAACCTT GATTCCACTG GAGGTCGAC CCTGTGGAAT GTGTGCAG  

(2) INFORMATION FOR SEQ ID NO:38:  

(i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 48 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear  

(ii) MOLECULE TYPE: DNA (genomic)  

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:  

GACTAACCTT GATTCCACTG GAGAATCTCG TGATGGCAGG TTGGGCGT  

(2) INFORMATION FOR SEQ ID NO:39:  

(i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 54 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear  

(ii) MOLECULE TYPE: DNA (genomic)  

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:  

GACTAACCTT GATTCCACTG AAGAGATTTT ATTTATCTTC CAGAAAAAG GGGG  

(2) INFORMATION FOR SEQ ID NO:40:  

(i) SEQUENCE CHARACTERISTICS:  


5 (ii) MOLECULE TYPE: DNA (genomic)

10 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:
GACT AACCTT GATTC CACTG AAGCCCCCAA ATGAAAGACC CCGCTGACG

15 (2) INFORMATION FOR SEQ ID NO:41:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 49 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

30 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:
GACTAACCTT GATTC CACTG GAGCCGGGAC GGAATTGC TA ATCTGCTGC

35 (2) INFORMATION FOR SEQ ID NO:42:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 47 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

40 (ii) MOLECULE TYPE: DNA (genomic)

45 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:
GACT AACCTT GATTC CACTG GAGTTCTCGA GCCGGC GCAT CTCG GCG

50 (2) INFORMATION FOR SEQ ID NO:43:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

60 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:
CGCTCTAGAA CTA GTGGATC

65 (2) INFORMATION FOR SEQ ID NO:44:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:
GTAATACGAC TCACTATAGG G

(2) INFORMATION FOR SEQ ID NO:45:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 43 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:
CGATCCACTG GAGCTCGGAG CCCACCCCCCT CCAATCTAGA GGT

(2) INFORMATION FOR SEQ ID NO:46:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 43 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:46:
CGTCCTCCTG GAGAGCACAG GGTAGAGGAG TCTGAGCGGT CAG

(2) INFORMATION FOR SEQ ID NO:47:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 43 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:47:
CGCAACCCCTG GAGACCTCTA GATGGGAGGG GGTGGGCTCC GAG
(2) INFORMATION FOR SEQ ID NO:48:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 43 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:48:
GCAGGACCTG GACGCTACCC TCTACCTGT GCT

(2) INFORMATION FOR SEQ ID NO:49:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:49:
CGCTCTAGAA CTAGTGATC

(2) INFORMATION FOR SEQ ID NO:50:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:50:
GTAATACGAC TCACTATAGG G

(2) INFORMATION FOR SEQ ID NO:51:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:51:
(2) INFORMATION FOR SEQ ID NO:52:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(x) SEQUENCE DESCRIPTION: SEQ ID NO:52:
GGATCCATCG ATACGTAAG

(2) INFORMATION FOR SEQ ID NO:53:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 38 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(x) SEQUENCE DESCRIPTION: SEQ ID NO:53:
GGCCGCTAAC TAAATTGCCA TTCTCCAGG TACGTAGC

(2) INFORMATION FOR SEQ ID NO:54:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 38 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(x) SEQUENCE DESCRIPTION: SEQ ID NO:54:
TAGTACCTT GGAGAATGGG CTATATTTA GCGGCGGC

(2) INFORMATION FOR SEQ ID NO:55:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 55 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:55:
GACTAACCTT GATTCACCTG GAGTTTTTCTC TATTCCTCAT TCCCCACTCTC TTCTT 55

(2) INFORMATION FOR SEQ ID NO:56:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 60 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:56:
GACTAACCTT GATTCACCTG GAGAATCTGG ACCAATTCTA TATAGCCTG TGAATAATT 60

(2) INFORMATION FOR SEQ ID NO:57:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 46 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:57:
GACTAACCTT GATTCACCTG GAGAAGAAGA AGTGGGAAT GAAGA 46

(2) INFORMATION FOR SEQ ID NO:58:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 51 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:58:
GACTAACCTT GATTCACCTG GAGATCTCTA GAMGGAGGG GGTCTGGGCT C 51

(2) INFORMATION FOR SEQ ID NO:59:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 47 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:59:

GACTAACCTT GATTCCACTG GAGCTCGGAG CCCACCCCT CCCATCT

47

(2) INFORMATION FOR SEQ ID NO:60:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 47 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

20

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:60:

GACTAACCTT GATTCCACTG GAGGGAGGCC CTTATCTCAA AAATGTT

47

(2) INFORMATION FOR SEQ ID NO:61:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 51 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

35

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:61:

GACTAACCTT GATTCCACTG GAGTCTAAGA ACATTTTGA GATAAGGGCC T

51

(2) INFORMATION FOR SEQ ID NO:62:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 44 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

40

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:62:

GACTAACCTT GATTCCACTG GAGTCAGGG CTTATATACT GAAA

44

(2) INFORMATION FOR SEQ ID NO:63:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 46 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

55
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:63:
GACTAACCTT GATTCCCTGG AGACTGCACT GCTGTCCCCG CCTCG
46
(2) INFORMATION FOR SEQ ID NO:64:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 53 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xii) SEQUENCE DESCRIPTION: SEQ ID NO:64:
GAGTAACCTT GATTCCCTGG AGATTTCTCA GACCCGGGTC GACCCCTGTTG AAT
53
(2) INFORMATION FOR SEQ ID NO:65:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 44 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xii) SEQUENCE DESCRIPTION: SEQ ID NO:65:
GACTAACCTT GATTCCCTGG AGCTCGAGGC GGCACATCTC GGGG
44
(2) INFORMATION FOR SEQ ID NO:66:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 47 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xii) SEQUENCE DESCRIPTION: SEQ ID NO:66:
GACTAACCTT GATTCCTGGA AGACTGCGT CATGCTGAGA CCTCAA
47
(2) INFORMATION FOR SEQ ID NO:67:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 50 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:67:

GACTAACCTT GATTCCTGAA GCGGCAAT GCACCCAATG AAAGATTTTC

(2) INFORMATION FOR SEQ ID NO:68:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 50 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:68:

CGCATCTTTT AATTAACCTGG AGARAAATTT TYACAGGCTT ATATAGKAAA
We claim:

1. A method for assembling a gene or gene vector comprising the steps of:
   a) designing at least 6 primers to produce at least three fragments in at least
      three separate polymerase chain reactions wherein each primer comprises at least one
      predetermined restriction endonuclease recognition site that recognizes a restriction
      endonuclease that cleaves at a distance from the recognition site, a sequence complementary
      to a template sequence for amplification, and bases positioned at the restriction endonuclease
      cleavage site that are selected to be complementary to only one other overhanging created
      from enzymatic cleavage of the fragments;
   b) combining the primers with template nucleic acid and performing a gene
      amplification reaction to produce multiple copies of an amplified template fragment
      incorporating the restriction endonuclease recognition site;
   c) digesting the amplified template fragments with one or more restriction
      endonucleases that recognize the restriction endonuclease recognition site of the
      primers to create overhanging termini wherein each overhanging termini is
      complementary to only one other overhanging termini on another fragment; and
   d) combining the amplified and digested template fragments in a ligation
      reaction to produce a directionally ordered gene, nucleic acid fragment or gene vector.

2. The method of claim 1 wherein the restriction endonuclease is at least one class IIS
   restriction endonuclease.

3. The method of claim 2 wherein the class IIS restriction endonuclease is selected from the
   group consisting of: AlwI, Alw26I, BbsI, BbvI, BbvII, BpmI, BsmAI, BsmI, BsmBI, BspMI,
   BsrI, BsrDI, Eco57I, EarI, FokI, GsuI, Hgal, HphI, MboII, MniI, PleI, SapI, SfaNI,
   TaqII, Tth111II.

4. The method of claim 1 wherein class II restriction endonuclease recognition sites,
   linkers, or adapters are not used to create the gene or gene vector.
5. The method of claim 1 wherein the product of the ligation reaction is introduced into prokaryotic or eukaryotic cells.

6. The method of claim 1 wherein at least one target nucleic acid sequence is chosen from the group consisting of: transcriptional regulatory sequences; genetic vectors; introns and/or exons; viral encapsidation sequences; integration signals intended for introducing nucleic acid molecules into other nucleic acid molecules; retrotransposon(s); VL30 elements; or multiple allelic forms of a sequence.

7. The method of claim 1 wherein the method is used to generate combinatorial libraries of a target sequence.

8. The method of claim 7 wherein the target sequence is part or all of a gene.

9. The method of claim 8 wherein the gene encodes a protein.

10. The method of claim 8 wherein the primers amplify allelic variants of part or all of a gene.

11. The method of claim 1 wherein the product of the ligation reaction is passed between eukaryotic cells using a virus particle, by cell fusion, or by transfection.

12. The method of claim 1 wherein the product of the ligation reaction is not introduced into prokaryotic cells.

13. The method of claim 1 further combining at least one screening or selection step to select the products of the ligation reaction.

14. The method of claim 1 wherein the product of the ligation reaction is mutated during passage in cells in order to generate genetic diversity.
15. The method of claim 14 wherein the product of the ligation reaction is mutated by homologous recombination during passage in cells.

16. The method of claim 1, wherein the method is used to isolate and identify regulatory sequences from a cell.

17. The method of claim 11, wherein cells containing the product of the ligation reaction are selected for enhanced biological activity.

18. The method of claim 17, wherein the cells containing the product of the ligation reaction are selected for tissue-specific, hormone-specific or developmental-specific gene expression.

19. The method of claim 1 wherein the product of the ligation reaction is a circularized gene vector.

20. A nucleic acid primer having a 5' and a 3' end to amplify a nucleic acid fragment for the ligation of at least two fragments comprising:
   a restriction endonuclease recognition site that recognizes a restriction endonuclease,
   wherein the restriction endonuclease cleaves at a distance from the recognition site and creates overhanging termini;
   a sequence complementary to a template sequence to be amplified to produce the nucleic acid fragment;
   at least two nucleic acid bases positioned at the restriction endonuclease cleavage site
   and that form an overhanging terminus after cleavage by the restriction endonuclease,
   wherein the at least two nucleic acid bases are selected to be complementary to only one other overhanging terminus on another fragment of the ligation; and
   an affinity handle on the 5' end of the primer.

21. The primer of claim 20 further comprising an anchor to provide stability to the restriction enzyme at the restriction enzyme recognition site.
22. A method for isolating and identifying promoters comprising the steps of:
   a) obtaining a vector comprising at least a portion of a promoter region from a retrovirus transposon LTR and having two non-complementary overhanging termini;
   b) designing at least two PCR primers to amplify at least one region of a retro-transposon LTR from template nucleic acid to produce at least one nucleic acid fragment wherein each primer comprises at least one predetermined restriction endonuclease recognition site that recognizes a restriction endonuclease that cleaves at a distance from the recognition site, a sequence complementary to a template sequence from a retrovirus transposon, and bases positioned at the restriction endonuclease cleavage site that are selected to be complementary to only one other overhanging terminus of the vector wherein the restriction endonuclease cleavage site is created from enzymatic cleavage of the fragments;
   c) combining the primers with template nucleic acid and performing a gene amplification reaction to produce multiple copies of an amplified template fragment incorporating the restriction endonuclease recognition site;
   d) digesting the amplified template fragments with one or more restriction endonuclease that recognize the restriction endonuclease recognition site of the primer to create overhanging termini; and
   e) combining the amplified and digested template fragment in a ligation reaction with the vector to produce a gene vector with an intact LTR sequence.

23. The method of claim 22 wherein the template nucleic acid is DNA or RNA.

24. The method of claim 22 further comprising the step of sequencing the insert to identify the promoter sequence.

25. Promoter sequences of SEQ ID NOS:2-13 identified using the methods of claim 22.

Fig 1A
**Fig. 4 A.**

Genomic DNA or cellular RNA

Amplification of allelic parts via PCR or RT-PCR

Combine the parts in defined order using self-assembling genes

Grow constructs *en masse*

Transfect cells with constructs + replication competent retrovirus

Passage vectors that are expressed in mass cultures

Reisolate vectors after several passages

---

**Fig. 4 B.**

<table>
<thead>
<tr>
<th>LTR</th>
<th>U3</th>
<th>R</th>
<th>U5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Variable region:
- distal enh/pre
- ER1
- ER2
- ER3
- ER4
- medial promoter
- proximal promoter

PCR products:
1 2 3 4 5 6 7

Primers: - - - - - - -
<table>
<thead>
<tr>
<th>ID</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CCTCCACACT AGAGTTGTT CTGGAAACAC TTGTAACCTTT TCCACCCCCTAA AACTCTCTAC CCTTTTGGTC CATAAAACCTTG TTGTAAGAAGC</td>
</tr>
<tr>
<td>2</td>
<td>ATTTTAGA TAAGGCTAC CTAGGCCCTC CTGGAAACAC CTGTACAACAG CATGGGCTAAG TGATAGACAA TGACGTCTTA GTGACGTAGG TTGCTCTTGA</td>
</tr>
<tr>
<td>3</td>
<td>ACCCATGACT CTTCTAACCCT AGAGTTGTT CTGGTAACCTTT TCCACCCCCTAA AACTCTCTAC CCTTTTGGTC CATAAAACCTTG TTGTAAGAAGC</td>
</tr>
</tbody>
</table>

Fig. 5
INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12N15/86 C12N15/10

According to International Patent Classification (IPC) or to both national classification and IPC.

B. FIELDS SEARCHED
Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched.

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>PADGETT K A ET AL: &quot;Creating seamless junctions independent of restriction sites in PCR cloning&quot; GENE, vol. 168, no. 1, 2 February 1996, page 31-35 X0004042930 see the whole document</td>
<td>1, 2, 4-14, 19-21</td>
</tr>
<tr>
<td>Y</td>
<td>TOMIC, M. ET AL.: &quot;A rapid and simple method for introducing specific mutations into any position of DNA leaving all other positions unaltered&quot; NUCLEIC ACIDS RESEARCH, vol. 18, no. 6, 1990, OXFORD GB, page 1656 X000209445 cited in the application see the whole document</td>
<td>3</td>
</tr>
</tbody>
</table>

Further documents are listed in the continuation of box C.

Patient family members are listed in annex.

"A" document defining the general state of the art which is not considered to be of particular relevance.

"E" earlier document but published on or after the international filing date.

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified).

"O" document referring to an oral disclosure, use, exhibition or other means.

"P" document published prior to the international filing date but later than the priority date claimed.

Year of the actual completion of the international search

26 June 1998

Date of mailing of the international search report

09/07/1998

Name and mailing address of the ISA
European Patent Office, P. B. 5818 Patentbaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31651 epo nl.
Fax (+31-70) 340-3015

Authorized officer
Chambonnet, F.
<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication where appropriate of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>LEBEDEKO, E.N. ET AL.: &quot;Method of artificial DNA splicing by directed ligation&quot; NUCLEIC ACIDS RESEARCH, vol. 19, no. 24, 1991, OXFORD GB, pages 6757-6761, XP002069446 cited in the application see the whole document</td>
<td>1</td>
</tr>
<tr>
<td>P,X</td>
<td>WO 97 28282 A (STRATAGENE INC) 7 August 1997</td>
<td>1,2, 4-14, 19-21</td>
</tr>
<tr>
<td>P,Y</td>
<td>see the whole document</td>
<td>3</td>
</tr>
<tr>
<td>Patent document cited in search report</td>
<td>Publication date</td>
<td>Patent family member(s)</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>-----------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>WO 9728282 A</td>
<td>07-08-1997</td>
<td>NONE</td>
</tr>
</tbody>
</table>