Title: METHODS OF IDENTIFYING ACTIVE GENES

Abstract: This invention relates to a method of enhancer trapping in an organism, comprising introducing a retroviral vector into said organism, wherein said vector comprises a basal promoter and a detectably expressible gene. Retroviral vectors, transgenic organisms obtainable by introducing a retroviral vector into said organism and kits comprising the retroviral vectors or transgenic organisms are also included in the invention.
Methods of identifying active genes

This invention relates to methods of identifying genes that are active in a particular cell or tissue at a defined time in the lifespan of an organism. The invention also relates to new vectors for use in these methods.

Large scale sequencing projects, such as the sequencing of the human genome, the mouse genome, the Drosophila genome and the zebrafish genome amongst others, are producing a vast amount of sequence data. Whilst this sequence data is very useful, the data is currently being produced at a rate which far exceeds the rate at which the function of the new genes that are identified within these sequences can be determined.

Traditionally, information about the function of genes has been derived from the analysis of genetic mutations in whole organisms, and the effects of the mutations on the phenotypes of the animals. For example, genes affecting various diseases have been identified in humans through studying family histories of diseases, and the collection and annotation of spontaneous mutations in mice has been ongoing for many years. These mutations are however identified by virtue of their disease association or by the association of the mutations with a striking visual phenotype, and the vast majority of genes and their functions cannot be identified in this manner. Furthermore, spontaneous mutations occur with relatively low frequency and thus it would be an immense task to elucidate even a single genetic pathway by relying on the appearance of spontaneous mutations. The frequency of spontaneous mutations is around 5x10^-6 per locus.

These spontaneous mutations may arise in the coding sequence of the gene or alternatively may affect regulatory sequences, and may range from single base changes to the insertion of large retroviral sequences.

In order to overcome these difficulties, a variety of strategies have been adopted. These strategies generally involve the artificial generation of mutations, either randomly or in a targeted manner, and the subsequent study of the resulting mutant animals or embryos.

The application of mutagens, such as X-rays or chemical mutagens (e.g. chlorambucil and ethyl nitrosourea) has been widely used to artificially generate genetic mutations at a higher frequency than is possible by spontaneous mutagenesis. X-ray mutagenesis causes mutations at 20-100 times higher
frequency than spontaneous mutagenesis, and as the mechanism of mutation that is induced by X-rays is chromosomal rearrangement, it is relatively easy to identify the affected region of genome. The disadvantage of such techniques lies in the fact that these chromosomal rearrangements may affect multiple genes, resulting in difficulties in identifying the gene of interest.

Mutations resulting from the application of the chemical chlorambucil similarly result in chromosomal disruption and such techniques are thus not suitable for rapid single gene identification.

EthylNitrosourea is a chemical mutagen which is easy to apply and which generates point mutations i.e. single base or nucleotide changes in the genetic sequence of animals. This chemical has therefore been widely used to generate mutations and hence mutant animals *in vivo*. Once mutant animals with a visual phenotype have been identified it is however difficult to identify the mutated genes that are associated with the visual phenotype as the single nucleotide change or mutation does not provide any molecular landmarks from which the gene may be readily identified. The process of cloning the newly identified gene therefore remains laborious.

Transgenic or retroviral insertional mutagenesis, whereby an exogenous DNA sequence is inserted into the germline nucleic acid of the organism being studied has also been widely used. The insertion and integration of exogenous DNA into the genome of the animal which takes place in the cell of these animals may occur in a location that disrupts normal endogenous gene function and may in turn lead to a visually identifiable phenotype. The locus of insertion may be identified and cloned using the exogenous sequence as a tag, and the gene identified in this way may be correlated with and related to the phenotype. Transgenic insertional mutagenesis is however labour intensive, particularly in mice, requiring many rounds of screening and breeding.

A further technique that has been used to generate mutants is that of homologous recombination in ES cells. In this technique, regions of homology between an artificial construct and a known endogenous nucleic acid sequence allow recombination between these two nucleic acid molecules to occur. The ability to introduce mutations or deletions into the sequence of the artificial construct means that, should homologous recombination occur, the mutation or deletion appears in the gene of interest. A selectable marker is often also included in this region, to aid the identification of cells in which homologous recombination has occurred. The ES cells are then implanted into the uterus of surrogate females and the individual animals that are derived from this are
identified, bred and characterised. This is a highly targeted approach that is particularly labour intensive as the frequency of homologous recombination is low and many ES cells have to be screened. Then, mice have to be generated from the ES cells and these have to be bred and screened. Furthermore, the amount of useful information that may be derived from these studies is difficult to predict in view of genetic redundancy (i.e. the absence of a functional copy of one gene may be compensated for by the presence of another different gene) and the potential occurrence of embryonic lethal phenotypes. Furthermore, the gene into which the mutation is introduced is already known and this technique is therefore not suitable to identify or characterise new genes.

Trapping is a further mutagenesis and screening system that has been used in mice and Drosophila and which is considered to be more efficient than the above methods with respect to the ability to characterise gene function and expression. Trapping methods are useful with respect to both the identification of active genes and in the mutagenesis of these genes. These systems depend on the use of genetic vectors that contain reporter genes (whose presence may be readily detected) and, following insertion of vector DNA into the host genome, utilise the host's transcription system to direct expression of the reporter gene.

There are three types of strategy that have been developed for this kind of mutagenesis and screening. The first, the enhancer-trap, utilises a vector that contains a minimal (or basal) promoter sequence and a reporter gene such as \textit{lacZ}. Insertion of vector nucleotide sequences near to an active \textit{cis}-acting enhancer element will cause expression of the reporter gene. This expression does not occur if the vector nucleotide sequences insert in other locations. This expression of the reporter gene is then detected and the gene associated with this expression can be cloned based on the presence of the nucleotide sequence of the trap vector. If the insertion has occurred in a gene, any phenotype (or change in phenotype) of the organism can be studied in relation to the newly identified gene.

Promoter trap vectors lack a promoter and contain a reporter gene (which may also act as a selectable marker). If the reporter gene does not act as a selectable marker then a further selectable marker is also required. If the vector inserts into a transcribed region of a gene downstream of the first exon, then a fusion is generated which is made up of the upstream endogenous transcribed sequence fused to the reporter (and marker) gene. Transcribed genes are identified by detection of the reporter gene, however the frequency of insertion
is low. The selectable marker gene is therefore used to identify cells that contain vector insertions.

The third type of vector is the gene trap vector. This contains a splice acceptor site upstream of a promoterless reporter gene. If the vector inserts near to a cis-acting promoter and enhancer and downstream of the first exon, a fusion transcript results from the upstream coding sequence and the reporter gene. The trapped gene is thus mutated and its gene expression pattern can be mapped.

The trap vectors used to date have been either injected into the pronucleus of fertilised mouse oocytes or transfected into ES cells of various organisms. The disadvantage of this technique is the need for preselection in ES cells. For a gene to be identified, it must be active both in ES cells and in the cells of interest within the organism, and the fusion must occur so that the correct reading frame is maintained. Alternatively, gene trap vectors have been introduced using retroviral infection methods.

Enhancer trapping in Drosophila is generally carried out using transposon-mediated enhancer traps i.e. a transposable element carrying a reporter gene under the control of an internal basal promoter (O’Kane and Gehring (1987) P.N.A.S. USA 84(24): 245-253, Bier et al. (1989) Genes Dev 3: 1273-1287, Bellen et al. (1989) Genes Dev 3(9): 1301-1313). These allow flanking genomic sequences to be identified rapidly, when the element inserts near to a gene enhancer. In addition, 10 to 15% of the insertions are mutagenic (Bellen et al., supra).

Transposon based systems have also been used in zebrafish in order to generate transgenic animals. Transposon systems derived from Tc3 of C.elegans (Raz et al. (1997) Current Biology 8: 82-88), and mariner of Drosophila (Fadool et al. (1998), P.N.A.S. USA, 95: 5182-5186) have been used and a Tc12 transposon system has been identified in the medaka fish which has been considered to be useful to develop transgenesis and mutagenesis methods (Kawakami et al. (2000) P.N.A.S. USA, 97(21): 11403-11408). The significant drawback of the use of a transposon based technique is the rate of insertion. Only a single insertion is achieved per fish using transposon based techniques. Higher levels of insertion events are required to generate fish that are useful in screening systems.

Zebrafish have been subjected to gene-trapping experiments and retroviral mediated insertional mutagenesis. A retrovirus based gene trap technique using a splice-acceptor-based gene trap vector has been developed by Chen et al. ((2002) J. Virol. 76: 2192-8). None of the observed gene trap events
however led to a mutation, as it seems that the transcriptional machinery spliced over the insert at a certain rate, precluding complete loss of function of the relevant gene. The mutagenicity of these insertions was not based on gene trapping, nor did gene trapping aid detection of any mutagenic insertion inside genes.

A recent large scale retroviral insertional screen generated around 500 embryonic lethal recessive mutations, and cloning of these is progressing at a rate of around 3 per week (Amsterdam et al. (1999) Genes Dev 13: 2713-2724).

Zebrafish are a very good model system on which to conduct functional genome analyses, particularly given the high levels of homology of zebrafish genes to mammals and the short generation time of the fish.

When the insertional mutagenesis technique of Gaiano et al. ((1996) Nature 383: 829-32) is compared to a previous study using chemical mutagenesis it can be seen that although the chemical mutagenesis approach generated 1800 mutations, the fact that only about 100 of these have been cloned to date demonstrates that insertional mutagenesis confers a massive advantage with respect to the ability to clone the mutated genes.

However, in the insertional mutagenesis technique of Gaiano et al., one in ten insertions were into genes, and one in eighty insertions resulted in a visible phenotype. The mutagenic insertions were mainly in the 5' regions of the affected genes (Amsterdam (2003) Developmental Dynamics 228: 523-534).

The prior art techniques thus all have disadvantages, which are overcome by the method of the invention. The enhancer trap system using transposons is very inefficient with respect to the number of insertions. The gene trap technique of Chen et al. ((2002) supra) has not itself identified any mutations and the standard retroviral based insertional mutagenesis technique of Gaiano et al. is very labour intensive, and the used vector is non-expressing and therefore does not identify active genes. Thus what is required is an efficient method of identifying active genes that overcomes these disadvantages.

The Gaiano et al. technique relies on the generation of an F₀ generation, containing insertions. These are pair-mated to grow F₁ families. This F₁ generation is tail clipped to isolate DNA, on which quantitative PCR is performed to identify fish with multiple insertions, and those with 8 or more insertions are pair-mated. Around 50 fish per pair-mating are raised and 6 matings per family are then performed. The resulting embryos are then examined. The techniques of identifying fish that contain the mutation are thus time consuming. It can be seen from this that if it were possible to preselect
insertions that have occurred within genes before the zebrafish undergo breeding to homozygosity, it would be possible to identify ten times as many unique mutations. Similarly, if it were possible to preselect insertions that have occurred near or in genes it would be possible to identify a larger number of active genes and thus a wider range of genes can be identified and studied by virtue of their expression patterns, without necessarily being mutated or resulting in a visible phenotype.

Thus, it can be seen that another feature that would be advantageous is a more efficient trap technique that allows the preselection of insertions near or within genes. This obviates the need to perform time-consuming crosses and the generation of homozygous fish to analyse mutations.

If expression of an inserted reporter gene can be detected in the F1 generation by microscopic inspection then crosses to generate individuals homozygous for the insertion are not necessary for the detection of active genes. The detected active genes can be cloned from these heterozygous F1 animals and the nature of the insertion inspected for mutagenicity before inbreeding is attempted.

The inventors have therefore designed an improved method of enhancer trapping, which preferably generates insertional mutants. This new method, using retroviral vectors, increases the efficiency and sensitivity of such screens in identifying active genes (and also mutagenesis of said genes) in a wide variety of developmental processes and hence improves the ability to clone these genes.

The method has the further advantage in that it is possible to affect the types of genes which are detected in such a screen (with respect to the normal temporal or spatial expression pattern of the gene) by selecting various components of the retroviral vector.

Whilst initially developed with respect to zebrafish, the new method of the invention may be applied to other organisms, particularly other animals and especially vertebrate animals, particularly those useful as model animals or in genomic studies, e.g. rodents, fish, reptiles, frogs (e.g. Xenopus or other amphibians). The new method may also be applied to invertebrate organisms, e.g. Drosophila and C. elegans, chordates, crustaceans and mollusks.

The new technique has numerous advantages including the following; it can allow the selection of live embryos of interest (i.e. containing the insertion at a site which is being transcribed) at an early stage of the screen, thus reducing the number of animals required. In addition, the technique identifies cells in which the identified gene would be active in a wild type organism. The
embryos that are selected may then be bred, and only those that are expressing proviral insertions are used to identify and isolate the genes of interest. Furthermore the sensitivity of the screen can be enhanced significantly using new indirect methods of detecting the expression of a reporter gene.

The enhancer trap technique designed by the inventors has the further advantage that insertions tend to occur towards the 5' end of genes, in view of the fact that the vectors are based on retroviral sequences. This further increases the efficiency of the screen as the chances of the basal promoter being activated are increased (Wu et al. (2003) Science 300 (5626): 1749-51). A very low percentage of insertions are also seen in housekeeping genes (i.e. genes which are constitutively expressed and encode proteins which are required for the normal function of the cell) in contrast to standard insertional mutagenesis, in which the majority of insertions detected are insertions into housekeeping genes. This means that the new technique is significantly more efficient in identifying developmentally regulated genes, or genes which are important or expressed in a particular cell, organ or tissue type.

Furthermore, it is possible to increase the chance of identifying a gene which is usually expressed at a particular developmental stage of an organism by carefully selecting the basal promoter which is used in the retroviral vector. This means that the particular screen or mutagenesis experiment can be skewed towards identifying genes expressed at a defined developmental stage (e.g. early in development) and as such the efficiency of the screen is increased.

This novel method may also take advantage of the properties of fluorescent proteins such as GFP or other reporters which can be detected in a live animal or cell, in combination with the use of a basal promoter in the retroviral vector.

Thus in a first aspect, the invention provides a method of enhancer trapping in an organism, comprising introducing a retroviral vector into said organism (e.g. into a cell of said organism), wherein said vector comprises a basal promoter and a detectably expressible gene.

A retroviral system for performing enhancer trapping has not been used before. The reason for this is that retroviral LTRs contain strong promoters. The presence of such a promoter in the construct would be expected to override the presence of the basal promoter in the enhancer trap vector and cause expression of the detectably expressible gene in any cells in which the retroviral sequences have integrated. The method would thus lack the crucial feature of enhancer trapping techniques, i.e. being able to distinguish between active and
inactive genes by virtue of the fact that expression of the exogenous gene only occurs when insertion near an active cis-acting host enhancer occurs. It should be noted that the problem with the LTR of a retroviral vector overriding a basal promoter would not be problematic in other trap systems, e.g. gene trap systems, in which the vectors used do not contain or require any promoter sequences but instead use the promoter sequences from the host organism as the reporter gene is spliced into the transcript generated from the gene that the vector has integrated into.

Surprisingly however, the inventors have demonstrated that the promoters found in retroviral LTRs are not generally active. In particular, they have shown that an LTR from a retrovirus that infects one species is surprisingly inactive as a promoter when integrated into the genome of a different species, and thus does not cause the exogenous expression of the detectably expressible gene irrespective of the locus of integration, as would have been expected. Specifically they have shown that the LTR from a murine retrovirus, e.g. the mouse retrovirus MLV, is inactive as a promoter in lower vertebrates such as fish, and in particular zebrafish and medaka fish. It is thus possible to use retroviral vectors in methods to detect expression of active genes and thereby study expression patterns of these genes.

Thus, in a preferred embodiment, the LTR of the retroviral vector is substantially or essentially inactive or non-functional or not recognised or not utilised as a promoter when integrated into the genome of the organism being studied.

The LTR of the retroviral vector should however be active as a promoter in the packaging cells that are used to produce vector for infection. The generation of retroviral particles for infection depends on this activity. Thus in a preferred embodiment the retroviral LTR functions as a promoter i.e. is an active promoter, preferably a strong promoter, in the packaging cell line (which is generally a mammalian cell line), and is inactive or non-functional or not recognised or not utilised as a promoter when integrated into the genome of the host organism being studied. Alternatively stated, the LTR is capable of driving transcription, or contains sequences capable of driving transcription in the packaging cell line, but does not drive transcription in the host organism.

In a preferred embodiment, transcription of the detectably expressible gene directly or indirectly causes expression of a marker protein that can be detected in a live said organism.
In a further aspect, the present invention provides a retroviral vector for use in enhancer trapping in an organism, said vector comprising a basal promoter and a detectably expressible gene and preferably wherein transcription of said detectably expressible gene directly or indirectly causes expression of a marker protein or polypeptide that can be detected in a live said organism. Said retroviral vector is preferably for use in a method of enhancer trapping of the invention, as described herein. Kits comprising said retroviral vectors as described herein form yet further aspects of the invention. Kits according to the invention may comprise one or more retroviral vectors as described herein and other components as described elsewhere herein.

The kits according to the invention may be used for methods of enhancer trapping and/or methods of identifying an active gene, and in particular can be used for carrying out the methods of the invention as described herein.

The detectably expressible gene may thus cause expression of (e.g. it may encode or it may control the expression of) a marker protein, which may be directly or indirectly detected.

The marker protein may be a polypeptide e.g. may be any polypeptide, including also shorter peptide sequences, proteins and RNA.

A polypeptide may be detected by the effect it has on the host or the host cell in which it is expressed. The detectably expressible gene may thus be viewed as a marker gene, which may include not only conventional marker genes such as reporter genes but also a therapeutic gene. As described in more detail below, the detectably expressible gene may be a reporter gene which may be directly detectable, i.e. it may itself be directly signal-giving or it may itself directly be detected by a signal-giving mechanism or process, or it may be indirectly detected by taking part in a further, or secondary process or mechanism, which may then itself be directly or indirectly signal-giving, for example by activating, inactivating or causing the expression of a further gene.

In a particular embodiment, the invention provides a method of enhancer trapping in an organism, comprising introducing a retroviral vector into said organism (e.g. into a cell of said organism), wherein said vector comprises a basal promoter and a reporter gene. The LTR of said retroviral vector preferably has the properties discussed above. Thus, for example the LTR is preferably a strong promoter or contains a strong promoter which acts in the packaging cell line, and is inactive as a promoter when integrated into the genome of the organism being studied.
Further preferably, transcription of said reporter gene directly or indirectly causes expression of a marker protein that can be detected in a live said organism.

Preferably the organism is assayed to determine whether expression of the detectably expressible gene, e.g. expression of the marker protein, has occurred. Preferably the cells in which the detectably expressible gene/reporter protein is expressed are identified.

Also provided is a retroviral vector for use in enhancer trapping in a vertebrate or invertebrate, said vector comprising a basal promoter and a reporter gene. Preferably transcription of said reporter gene directly or indirectly causes expression of a marker protein that can be detected in a live said vertebrate or invertebrate.

Preferably the retroviral vectors contain an LTR sequence which has the properties discussed above. Thus, for example the LTR sequence is a strong promoter in the packaging cell line, and is inactive as a promoter when integrated into the genome of the organism being studied.

Alternatively viewed, the invention can be seen to provide a method of identifying an active gene in an organism, said method comprising (a) introducing a retroviral vector into said organism, wherein said vector is an enhancer trap vector comprising a basal promoter and a detectably expressible gene, and (b) detecting the expression of said detectably expressible gene whereby said expression is indicative of said active gene.

Also provided is a method of identifying an active gene in an organism, said method comprising (a) introducing a retroviral vector into said organism, wherein said vector is an enhancer trap vector which comprises (i) a basal promoter which is inactive in said organism in the absence of an enhancer, but which may become active when inserted into the genome of said organism in a position which may trap (e.g. near to) an enhancer of said active gene; and (ii) a detectably expressible gene, e.g. a reporter gene, wherein transcription of said detectably expressible gene/reporter gene directly or indirectly causes expression of a marker protein; and (b) detecting the expression of said marker protein, whereby said expression is indicative of said active gene.

Also provided is an enhancer trap retroviral vector as defined in step (a) above, for use in a method of identifying an active gene in an organism.

In these aspects of the invention, preferably the retroviral vector contains an LTR sequence having the properties discussed elsewhere herein.
Thus, in the aspects of the invention discussed herein, the basal promoter of the enhancer trap vector is insufficient in itself to direct or drive the detectable expression of the detectably expressible gene/reporter gene in the organism. However, if the vector inserts into the genome of the organism in a position where the promoter may be "activated" by an enhancer of an active gene (i.e. a gene which is actively being expressed or transcribed) (e.g. near to an enhancer of an active gene), then the basal promoter may be activated. To be able to "use" an enhancer of an endogenous gene of said host organism in this manner, the enhancer must be near to a gene which is active. Thus, by "trapping" the enhancer of an active gene, the detectably expressible gene/reporter gene may be expressed, leading to expression of the marker protein. Detection of the marker protein may thus lead to, or achieve, identification of the active enhancer which drives expression of a particular active gene. For example, the locus or position of the active gene and enhancer may be detected, and/or they may be identified by other means, for example by sequencing or other genetic analysis techniques. The marker protein thus acts as a "tag" or marker for the active gene or enhancer, through which it may be identified. It will be appreciated that such a method has many uses in genetic studies, e.g. in genomic studies, functional analysis of gene expression, in diagnostics or prognostics, drug discovery, genome annotation etc. It may be used to determine which genes are active at particular times (e.g. the life stages or cell cycle stages), in particular conditions, e.g. environmental conditions, disease states etc, or spatially within a given organism, for example in different cells or tissues. The expression of the marker protein may thus also be used to identify, isolate or observe these different cells or tissues for the purpose of cell culture, RNA expression profiling or developmental studies or mapping of these cells. Comparative studies may thus be carried out. The methods may also be used as a very efficient means for quickly generating large numbers of transgenic lines that express marker genes in specific subsets of cells. These in turn are important to isolate expressing cells. The isolated expressing cells may be used to make tissue specific expression libraries for microchip expression profiling. They may also be used for growth in culture, or may be observed in vivo, to observe their fate in the organism as in organogenesis or circuitry of the central nervous system for the purpose of neuron subtype identification or mapping different neuronal populations.

By "inactive" in the context above and as used elsewhere herein, it is meant that the basal promoter on its own is unable to direct the expression of the
detectably expressible gene/reporter gene (i.e. it has no activity in the host organism) or has low activity, i.e. is active at low levels only, for example such as to cause negligible, insignificant or undetectable expression of the detectably expressible gene/reporter gene, e.g. the promoter is substantially or essentially inactive. When the basal promoter is "activated" by trapping an enhancer in the host organism, it becomes "active" in the sense of being able to direct detectable or significant expression of the detectably expressible gene/reporter gene. The terms "active" and "inactive" (together with the alternative terms "non-functional", "not recognised" or "not utilised") are also used in this way herein to refer to the differential promoter activity of the LTR sequence in different organisms.

As described in more detail below, when a reporter gene is used, in a preferred embodiment the transcription of said reporter gene causes expression of the marker protein indirectly. Such an indirect mechanism allows the signal ultimately generated from the reporter gene expression to be amplified.

Thus, in an even more particular embodiment, the present invention provides a method of enhancer trapping in an organism, comprising introducing a retroviral vector into said organism (e.g. into a cell of said organism) wherein said vector comprises a basal promoter and a reporter gene and wherein transcription of said reporter gene indirectly causes expression of a marker protein that can be detected in a live said organism.

Preferably the organism is assayed to determine whether expression of the marker protein has occurred. Preferably the cells in which the marker protein is expressed are identified.

This, more particular, method of enhancer trapping may also be alternatively viewed as a method of identifying an active gene in an organism, said method comprising (a) introducing a retroviral vector into said organism, wherein said vector is an enhancer trap vector which comprises (i) a basal promoter which is inactive in said organism in the absence of an enhancer, but which may become active when inserted into the genome of said organism in a position which may trap (e.g. near to) an enhancer of said active gene; and (ii) a reporter gene, wherein transcription of said reporter gene indirectly causes expression of a marker protein; and (b) detecting the expression of said marker protein, whereby said expression is indicative of said active gene.

Also provided is an enhancer trap vector as defined in step (a) above, for use in identifying an active gene in an organism.
In these aspects of the invention, preferably the retroviral vector contains an LTR sequence having the properties discussed elsewhere herein.

By "enhancer trapping" is meant the process by which an exogenous gene, such as a reporter gene or other detectably expressible gene, under the control of a basal promoter is inserted into the genome of a cell and expressed by virtue of an endogenous enhancer present in said host cell, which allows the basal promoter to direct the expression of the exogenous gene. The exogenous gene (e.g. reporter gene) and the basal promoter are functionally linked, i.e. present on the same molecule and the basal promoter controls transcription of the gene. As discussed above, the basal promoter alone is however not sufficient to drive transcription of the exogenous gene and therefore in the absence of insertion into a location that is close to an enhancer sequence no or negligible transcription occurs. If the insertion occurs sufficiently close to an enhancer sequence, then transcription and translation of the exogenous gene can take place, leading to the production of the product of the exogenous gene, e.g. an exogenous protein. The detection of the exogenous peptide or protein encoded by the exogenous gene (e.g. the reporter gene), or its transcript, by any means, indicates that the insertion of the exogenous DNA has occurred in or near an active gene, i.e. a gene which is being actively transcribed. As the inserted exogenous sequence is known, it is possible to use standard cloning e.g. PCR techniques to clone and sequence the genomic DNA sequences at the site of insertion in order to identify or obtain the sequence of the genetic locus at which the insertion occurred.

(As discussed above, enhancer trapping techniques and the vectors used therein are distinct from other trapping techniques such as gene trapping techniques (which use promoterless vectors containing splice acceptor sites) and promoter trapping techniques (which also use promoterless vectors.).)

The retroviral vector is introduced into said organism in any convenient or desired way, usually by infection. Thus, as mentioned above it may be introduced into a cell of said organism. The cell may be any cell of said organism, for example a mature, developed or differentiated cell, a stem cell or progenitor or precursor cell or a germ cell e.g. an oocyte or egg cell or sperm cell. The organism may be at any life stage. For example, where the organism is a vertebrate, appropriate life stages include ante-, post- and neonatal, infant, adult, juvenile etc. Particularly included are vertebrate embryos, including also embryonic stem cells (although preferably human embryos or cells derived therefrom, e.g. human embryonic stem cells are not used unless this is necessary.
for the particular application), particularly where the reporter gene/gene identification aspects of the invention are concerned. In such a situation, introduction into a vertebrate germ cell, embryonic cell or embryo is preferred. Similar life stages are also appropriate if an invertebrate organism is used.

Preferably, where the reporter gene/gene identification aspects are concerned, the insertion of the exogenous nucleic acid sequence occurs within a gene i.e. within the nucleic acid sequences that make up the introns, exons and regulatory sequences of the gene and therefore preferably interrupts the normal or wild type function or activity of said gene and/or the normal or wild type function or activity of the protein or peptide or RNA encoded by said gene. The insertion thus causes a mutation or change in the sequence of the endogenous nucleic acid, e.g. DNA, by the addition of a nucleic acid sequence that is not found in the normal or wild-type nucleic acid sequence. In such aspects, the protein encoded by said exogenous nucleic acid therefore does not contain the same sequence as the normal or wild-type sequence.

A particular advantage of this method is that insertions are detected in non-housekeeping genes at much higher frequencies than the previously used methods of insertional mutagenesis. In fact, insertions into housekeeping genes are rarely detected when this method is used. This means that each time the method is performed, proportionally more developmentally regulated genes are likely to be identified, and the method is therefore more efficient than previously used methods of insertional mutagenesis at identifying developmentally regulated (i.e. non-housekeeping) genes. This is surprising as there will be many housekeeping or ubiquitously expressed genes that are expressed at any time.

Various different mutations may occur; nonsense mutations i.e. the generation of a stop codon in the reading frame will cause truncation of the protein encoded by the sequence into which the exogenous nucleic acid has inserted. Truncated proteins may not function correctly, depending on the location of the truncation. The insertion of exogenous DNA may also lead to a protein with additional or incorrect sequences. A protein that does not have the correct normal or wild-type sequence may also not function correctly.

Insertion into non-coding parts of the gene may also occur and may affect the normal or wild type function or activity of the protein or peptide. The insertion generated by this method may thus also disrupt non-coding portions of the gene, such as intronic sequences, promoter sequences or the sequences encoding the polyA tract of mRNA. If this occurs, then despite the absence of
an insertion into a part of the gene that affects the primary sequence of the protein and hence affects its primary structure, the insertion may disrupt other essential processes of protein synthesis such as splicing events or mRNA stability.

This type of mutagenesis is referred to as insertional mutagenesis as it is caused by the insertion of exogenous nucleic acid. By mutation it is meant any change in nucleotide sequence of the genetic material i.e. DNA or RNA of the cell. It is thus preferred, particularly in the context of the use of a reporter gene, (e.g. in a method of identifying an active gene) for the introduction of said retroviral vector to result in insertional mutagenesis of an endogenous gene in said organism (for example of an active gene).

The term "insertional mutation" or "insertional mutagenesis" therefore means that a nucleotide sequence is present in a gene, preferably an exon, which is not found in the normal, or wild-type nucleotide sequence, i.e. it is present in addition to the wild-type or endogenous sequence and does not replace an endogenous nucleic acid sequence. The extra or additional nucleotide sequence is derived from the exogenous vector, and results from the integration of the vector nucleotide sequence into the gene in question. When the exogenous nucleotide sequence is inserted in an exon of a gene, the presence of this new, or exogenous sequence disrupts the normal sequence of the expressed gene and the resulting protein produced by transcription and translation of the gene, and a truncated or non-functional product or protein may be generated, as described above. In addition, the additional sequence may be transcribed and translated and a protein that is not expressed in the non-mutated cell may be expressed, which may interfere with normal gene or protein function.

As used herein, "endogenous" refers to something that is naturally occurring in a wild-type cell, tissue organ or animal, and in enhancer traps is not part of the inserted vector itself.

As used herein, "exogenous" refers to something which is introduced by the technique. For example, any retroviral vector nucleic acid sequence or product thereof would be exogenous. In general, exogenous nucleic acid sequences or proteins are not usually found in the wild-type cell, tissue, organ or animal which has not been subjected to enhancer trapping and they are thus foreign or alien to the wild-type cell, tissue, organ or animal.

In a retroviral vector, the nucleotide sequence that is inserted is that which is flanked, i.e. surrounded by the long terminal repeat (LTR) sequences. The LTR sequences are also inserted into the organism's genomic sequence.
Thus, if an insertional mutant is generated, and the gene into which the insertion was made is active, then the additional sequence, which serves to disrupt the normal expression of the gene, should also be expressed. If the additional sequence is selected such that it encodes a peptide or protein or other product that can be detected, directly or indirectly, without interfering with the normal function of the cell, then it will be possible to select and/or identify the cell or cells into which the vector has inserted (i.e. the "positive" cells) on the basis of the expression of the exogenous gene (i.e. the detectably expressible gene) in the "positive cells".

If the method is being used in a screen to generate mutant organisms, only those organisms (e.g. embryos) containing cells that express the exogenous gene (e.g. reporter) gene can then be selected to be subjected to breeding and further analysis.

This allows the identification of genes that would not be detected by conventional phenotype-driven screens. Such genes include those that function at more than one point during development and that would, in recessive screens where the generation of homozygous embryos is required, lead to early lethal defects before later phenotypes could manifest. Genes that are only active in a small subset of cells may also not be detected in a phenotype based screen as the phenotype may be too subtle to detect. Phenotype driven screens suffer from the drawback that if a phenotype is not visible when the organism e.g. the embryo or whole animal is examined under the microscope, then it may not be selected for further examination.

As mentioned above, the vector may be introduced into an embryo, or a cell of an embryo, an ES cell, or indeed any suitable cell of an organism. Preferably the cell is a cell of an embryo. Infection of cells is a routine undertaking in the art, and generation of retroviral particles pseudotyped with the VSVG glycoprotein as well as the infection are described in Burns, J.C. et al. (1993), Proc. Natl. Acad. Sci. US A 90, 8033-8037.

The organism for use in the methods of the invention described herein may be any organism for example any vertebrate or non-vertebrate organism. For use in methods involving reporter genes, or methods of gene identification etc, the organism is preferably a lower vertebrate, for example a fish, reptile or amphibian, more preferably a fish or amphibian. Alternatively the organism is an avian vertebrate, for example a chick or chicken, or a rodent (e.g. mouse, rat, guinea pig, hamster etc) or other laboratory animal. The organism may thus be
mammalian or non-mammalian. Preferred non-vertebrate organisms for use in the invention include insects, molluscs, chordates and crustaceans.

Preferably, in such contexts, the vertebrate is an organism or species that has been genetically well-studied and/or documented, e.g. an animal commonly used in genomic studies, or for which animal models are desired. Accordingly, the organism is preferably a zebrafish (Danio rerio), Japanese Medaka fish, Xenopus or a mouse. The zebrafish and Medaka fish are particularly preferred as the sequencing of their genome is almost complete, and a large number of phenotypic changes induced by mutation have already been identified. The embryos are transparent and it is thus easy to detect phenotypic changes and to observe reporter gene expression. The similarity of developmental programs in all vertebrates means that the study of zebrafish also provides a good model for other vertebrates, such as mammals, including humans.

The retroviral vector may be any retroviral vector known in the art, or may be based on any retrovirus according to choice, and the precise system involved (e.g. the particular vertebrate or non-vertebrate host, basal promoter and/or exogenous gene selected etc). Thus, any desired or appropriate retroviral vector may be used, and may differ according to the host species or organism.

The various components of the retroviral vectors to be used will thus generally be selected using functional criteria as described elsewhere herein. For example, the basal promoter of the retroviral vector used in the enhancer trapping methods of the invention described herein will generally be selected so that it functions as a basal promoter in the host organism concerned, i.e. a promoter will be selected which alone is unable to direct expression of an exogenous gene in the host organism but can direct expression of an exogenous gene when it becomes integrated into the host organism in a position such that it can be acted on by a host enhancer sequence. The choice of basal promoter also has an influence on the types of gene in which insertion is detected, and hence on the genes which are identified by the method. For example, when a Gata2 basal promoter was used to perform the method on zebrafish embryos almost all of the genes identified were early genes, i.e. genes expressed on or before day one (e.g. up to 24 hours post fertilisation) of zebrafish embryonic development. Furthermore, all of the insertions that were detected were into genes which are developmentally regulated (see Example 1), and no fish were identified in which an insertion was made into or near a gene which is expressed only in the adult fish. In contrast, when a keratin 8 basal promoter is used, insertions were only detected when they were in genes which are expressed after day 2 of
embryonic development. Whilst not wishing to be bound by theory, it appears that different basal promoters will function with different efficiency with different enhancers. Thus, changing the basal promoter would not necessarily change the location or frequency of insertion but might influence the nature of genes which are identified in the method. For example, the Gata2 basal promoter functions with (i.e. is driven by) enhancers that are active at one developmental stage, and the keratin 8 basal promoter functions with (i.e. is driven by) enhancers that are active at a different developmental stage, and a non-overlapping spectrum of genes was identified in screens performed with these two different promoters. This probably reflects the fact that promoters function by virtue of binding of proteins to sequences within the promoter. The nucleotide sequence of a promoter thus determines whether it will be activated by a particular protein, and whether it will function with a certain enhancer. The skilled person can therefore select a basal promoter to use in the retroviral construct that is likely to allow the detection of the genes which are expressed at a particular developmental stage, or tissue or cell type. This choice can be made based on the usual expression pattern of genes under the control of a particular basal promoter.

Thus, the basal promoter is preferably selected such that it is capable of functioning i.e. directs expression in a particular developmental pattern, either spatially or temporally.

The methods of the invention are therefore particularly suited to identify developmentally expressed genes e.g. early genes or late genes. Promoters from structural proteins or other proteins expressed in differentiated cells are particularly preferred to identify late genes (e.g. genes expressed after day 2 of zebrafish embryogenesis) and promoters from early developmentally regulated genes e.g. early transcription factors are particularly preferred to identify early genes (e.g. genes expressed in the first day e.g. in the first 24 hours of zebrafish embryogenesis.

Similarly, the LTR of the retroviral vector used is generally selected such that it is inactive or non-functional or not recognised or not utilised as a promoter when integrated into the genome of the host organism. Preferably however, the LTR will be active as a promoter in the packaging cells, e.g. a mammalian cell line, used to produce the retroviral vector for infection.

Methods of testing whether a particular LTR is inactive or active as a promoter in various hosts and thereby identifying an appropriate LTR for use is well within the skill of a person skilled in the art.
Representative retrovirus groups upon which retroviral vectors may be based include lentiviruses and oncoretroviruses. Preferred retroviruses are those which infect mammalian cells, i.e. are mammalian retroviruses, preferably human or murine retroviruses.

Exemplary vectors include those based on or derived from murine leukaemia virus (MLV), particularly murine Moloney leukaemia virus (M-MLV) for example the vector pCL (Naviaux et al, 1996, J. Virol., 70(8), 5701-5705), those based on or derived from HIV (Miyoshi et al, 1998, J. Virol., 72(10), 8150-8157), those based on or derived from Rous Sarcoma Virus (RSV) (Chen et al. (1999), Dev. Biol. 214, 370-384) and derivatives of Avian Leukosis Virus (ALV), such as RCAS (Pao et al PNAS USA (2003) 15: 8764-8760). Preferred retroviral vectors for use in the present invention are those based on or derived from MLV and in particular M-MLV (e.g. those based on the pCL vector described by Naviaux, supra).

The retroviral vector may be introduced into the host organism in any convenient or desired way. For example, it may be injected into the organism, e.g. into an embryo or cell of said organism.

Preferably the vector is introduced in the form of a retroviral particle which contains the vector nucleic acid. The use of retroviruses as a means to insert exogenous genes or nucleic acid into cells is well known in the art and takes advantage of the ability of viruses to pass natural cellular barriers to exogenous nucleic acid uptake and integration into the genome, thus providing long term gene delivery (Coffin, J. (1992) Curr. Top. Microbiol. Immunol. 176: 143-64). Briefly, retroviral vectors are generally used which are replication incompetent, i.e. they do not possess all of the viral genes that are required to form a fully replicating vector. These vectors are transfected into "packaging cell lines". Packaging cell lines are capable of producing all of the necessary packaging trans proteins e.g. gag, pol and env, that are needed for the virus to be packaged, processed, reverse transcribed and integrated.

The presence of these trans proteins in the packaging cell line, in combination with the presence of the nucleic acid sequences of the vector, leads to the generation of infective viruses. These viruses are collected from the supernatant of the packaging cells and assessed for their infectivity by performing standard titre experiments.

Various means are known in the art to modulate the infectivity of the produced viruses or the number of colony forming units, and to manipulate the ability of the produced virus to infect certain cell types and host organisms. For
example it has been shown by Burns et al. ((1993) P.N.A.S. USA 90(17): 8033-7)) that it is possible to "pseudotype" retroviruses. This involves the substitution of the normal coat protein of one virus, with that of another, for example the normal coat protein of the MLV genome may be substituted with the glycoprotein of the vesicular stomatitis virus (VSVG). The presence of the VSVG glycoprotein allows zebrafish cells to be infected as they generate an active MLV core with the host specificity of VSV. Such retroviral modification is also appropriate to enable the infection of other non-mammalian cells, e.g. other fish cells such as medaka fish.

Thus it is possible to modify the properties of the virus by manipulation of the packaging cell line. Any packaging cell line may be used to generate the virus with which the cells are infected. Preferred packaging cell lines are those in which the LTR of the retroviral vector is an active promoter. Generally such packaging cell lines are mammalian cell lines, preferably human cell lines e.g. N1H3T3 (Markowitz et al. (1988) Virology 167(2): 400-6) or PA317 (Miller et al. (1988) Mol. Cell. Biol. 5: 431-7). Preferred examples include the human cell line 293, more preferably 293gp/bsr (Miyoshi et al 1998, supra). The packaging cell lines are preferably easy to transfec using standard techniques.

The titre of the virus may be detected by standard techniques, such as those described in Burgess & Hopkins ((2000) Methods Enzymol; 327: 145-61). Any means of determining the titre may be used.

The retroviral vector contains sequences from the long terminal repeats (LTRs) of the virus. These LTRs flank the nucleic acid sequence which is to be integrated into the genome of the cell. The LTR sequences are essential for reverse transcription and insertion of the sequences into the host genome. LTRs are found at the ends of the viral genome and can be divided into three elements. U3 is derived from the sequence unique to the 3' end of the RNA. R is derived from the sequence repeated at both ends of the RNA and U5 is derived from the sequence unique to the 5' end of the RNA.

The LTR sequences used in the vector may be derived from any retrovirus or may be synthetic, or may be modified versions of LTR sequences from any retrovirus, so long as they act to cause the reverse transcription of the retroviral RNA and integration of the nucleotide sequences that they flank. In addition they should preferably be active as promoters in the packaging cell line and inactive as promoters in the organism into which they are introduced. LTRs also act to control gene expression in integrated proviruses and therefore LTR sequences may also contain regulatory elements that act to control transcription,
such as promoters and enhancers. These regulatory elements may be responsive to cellular transactivating factors. Preferred LTRs are the MLV LTR or the HIV LTR.

It should be noted that although the two LTR sequences may have identical sequences, this is not necessary. The LTR sequences may also be generated by filling in one or more LTR sequence or fragment thereof. In a preferred embodiment the 5’ LTR is from pCL (Naviaux et al, supra) and contains a CMV promoter. The CMV promoter promotes strong transcription of the viral genome in the packaging cell line after transient transfection. As the 5’LTR is replaced with a copy of the 3’LTR during viral replication, the CMV promoter is not included in the viral RNA.

In addition to the LTR sequences required for reverse transcription of the viral RNA and integration, the retroviral vector preferably comprises an optimised packaging signal. The packaging signal functions to aid the assembly of the viral particles in the packaging cell line. In a preferred embodiment the packaging signal is from pCL (Naviaux et al., supra).

As mentioned above, the vector may be a lentiviral vector. Lentiviruses are a subgroup of retroviruses. The use of lentiviral vectors presents an advantage in that, unlike retroviral vectors in general, they are capable of infecting non-proliferating cells, making them useful in situations where other retroviral vectors are not. For injection into embryos, both systems are of use as the majority of cells are dividing in the embryo. It has been shown that lentiviral vectors are suitable for germline transmission in mice and rats (Lois et al (2002) Science 295 (5556): 868-72).

As mentioned above, it is preferred that the LTR acts as a promoter in the packaging cell line and is inactive as a promoter in the organism into which the vector is being introduced.

A preferred combination is a mammalian retroviral LTR sequence i.e. an LTR from a retrovirus which infects mammals and a non-mammalian test organism. Even more preferably, a rodent retroviral LTR sequence i.e. an LTR from a retrovirus which infects rodents is used with a non-mammalian test organism. The non-mammalian test organism is preferably a lower invertebrate e.g. a frog or other amphibian, or reptile, or fish or invertebrate. Any combination of retroviral LTR may be made with the test organism so long as the retroviral LTR sequence does not function or is inactive as a promoter in the organism, and does function as a promoter in the packaging cells. The LTR may also be synthetic. Preferred mammalian LTRs for use in the invention are
murine LTRs, in particular LTRs from MLV, particularly M-MLV, for example the LTR as used in the vector pCL. Preferred non-mammalian test organisms are lower invertebrates such as fish, and in particular zebrafish or medaka fish, or amphibians.

Thus, in a preferred embodiment of the invention the LTR of the retroviral vector is active in mammalian cells (such as those described above) and is inactive in lower vertebrates or invertebrates (such as those described above).

The term "promoter" as used herein refers to a nucleic acid sequence that acts as a recognition signal for one or more transcription factors. Once the necessary transcription factors have bound to the nucleic acid, RNA polymerase binds to the nucleic acid-transcription factor complex to initiate RNA synthesis and transcription. The promoter sequence is usually found adjacent to, next to, or functionally linked to the gene under its control and whose transcription it drives. Promoters may comprise one or more transcription factor binding elements and the particular combination of these elements determines the spatial and temporal transcription of the gene under its control. Thus, it can be seen that promoters may be of different strengths under different conditions. A strong promoter is one that activates transcription at high levels, whereas a weak one will activate transcription at low levels or not at all.

As mentioned above, promoters may also be dependent on "enhancer sequences". These are nucleotide sequences that are cis-acting (i.e. they are found on the same nucleic acid molecule as the promoter with which they interact) that interact with promoter sequences to enhance or increase the transcriptional activity of genes. Unlike promoter elements, which are found in close proximity to the gene, the position of enhancers is variable and they may be found some distance from the gene on which they act. Enhancers are also activated by binding of regulatory proteins. The binding of the relevant gene regulatory protein is thought to interact with the transcription factors bound to the promoter sequences, by looping out the intervening nucleotide sequences. Thus, the strength of a promoter may also be affected by its dependency or proximity to enhancer sequences.

As used herein, the term "basal promoter" refers to a promoter that has no associated enhancer sequences and which exhibits no promoter activity or only low activity. Particularly, the promoter exhibits no or low promoter activity in the organism (e.g. cell) into which the vector is introduced. By low
activity is meant that no RNA is transcribed, or only low levels or negligible or insignificant levels of RNA are transcribed.

The basal promoter may thus be viewed as "inactive" in the absence of an enhancer. As mentioned above, different basal promoters are believed to function differentially with different enhancers and therefore it is possible to influence the nature of the genes in which insertion is detected by selecting the basal promoter according to the type of genes (e.g. with a particular spatial or temporal expression pattern) which are under investigation. In other words a basal promoter can be selected which functions with enhancers which are present in particular cell types or particular developmental stages which are under investigation.

Examples of such promoters will also depend on the system into which the vector is to be introduced. As mentioned above, the promoter sequence is made up of a number of different elements. The presence or absence of the particular combination of transcriptions factors required to drive transcription in a particular cell, organ, tissue or animal will determine whether transcription occurs. Thus a basal promoter for zebrafish may be different with respect to its nucleotide sequence to a basal promoter for mice and the basal promoter should be chosen in line with the host or target system in which the method of the invention is to be performed. However to be useful in the methods described herein, the promoter must be capable of driving transcription when sufficiently close to an active enhancer in the cell, organ, tissue or animal of interest, yet is inactive (i.e. does not cause transcription, or only causes low, negligible or insignificant levels of transcription) when inserted into other locations.

Examples of basal promoters suitable for use in a zebrafish system include the Gata2 core promoter (Meng et al., 1997, P.N.A.S. USA 94(12): 6267-72) and the heatshock hsp 70 core promoter (Halloran et al., 2000, Development, 127, 1953-1960). These are two well characterised examples, however it would be straightforward to identify other such promoters and any basal promoter may be used. For example the zebrafish Gata 2 core promoter can also act as a basal promoter in medaka fish. A further example is the epithelial promoter from the keratin 8 gene (Gong et al 2002, Dev. Dyn 223: 204-2115).

Promoters of use in the invention include those which have been designed or artificially created, or promoters that are naturally occurring. The promoters may also have been additionally modified e.g. nucleotide sequence
changes have been introduced according to standard techniques known in the art, in order to manipulate the activity or function of the promoter.

The at least partially random integration of the vector nucleic acid, including the two LTR sequences, into the genome of the cell, organ, tissue or animal in question may lead to transcription of the detectably expressible exogenous (e.g. reporter gene). As mentioned above, this only occurs if the promoter sequence present on the vector nucleic acid is brought sufficiently close to an endogenous enhancer that is active in that cell, organ, tissue or animal at that time. Thus, the detection of the expression of the detectably expressible exogenous gene in a cell, or organ, or tissue or even whole animal, at different time points indicates the normal expression pattern of a gene that is, under normal or wild type conditions, influenced by the enhancer. For this reason, it is necessary that the basal promoter is inactive or active at low or undetectable levels when the insertion occurs in or near an inactive gene, however the promoter must become active when induced to by the enhancer.

Preferably the promoter can be active in a wide range of tissues in conjunction with the required enhancers. An example of a preferred promoter is the Gata2 core promoter (Meng et al 1997 supra), the hsp70 core promoter (Halloran et al 2000 supra) of zebrafish or the epithelial keratin 8 promoter (Gong et al., 2002 supra). Alternatively, if the method is to be used to identify genes that are expressed in a particular tissue or cell type, or at a particular developmental stage, then the basal promoter can be chosen to increase the chance of detecting a gene with that particular expression pattern.

As used herein, the term "reporter gene" refers to any gene or nucleic acid sequence (such as a cDNA sequence) which is capable of being transcribed and translated into a protein or a peptide whose presence may be directly or indirectly detected (e.g. measured). As mentioned above, the transcription of the reporter gene, i.e. into a protein or peptide product of the reporter gene, occurs only when the insertion is made in or near an endogenous gene which is being actively transcribed under the control of an endogenous promoter and/or enhancer in that cell at that time. Thus, the transcription of said reporter gene in a cell, tissue organ or animal indicates that there is an active gene in this location at this time.

Transcription of the reporter gene directly or indirectly leads to the presence of the marker protein, in the cell. The marker protein is a protein that is detected by the investigator.
The marker protein may be the protein or peptide product of the reporter gene. If this is the case, this represents the case whereby the transcription of the reporter gene directly results in the production of marker proteins.

Alternatively, the reporter gene may encode a "transactivating protein". In other words, transcription and translation of the reporter gene produces a transactivating protein. The transactivating protein is e.g. a transcription factor or another protein that is involved in controlling transcription. This transactivating protein can bind to an upstream regulating site which controls the expression of a further reporter gene or gene encoding a marker protein i.e. the upstream regulating site may be a promoter for a further reporter gene or gene encoding a marker protein. If this is the case, transcription of the reporter gene indirectly results in the production of the marker protein. The upstream regulatory site is preferably an exogenous nucleotide sequence i.e. is not normally found in the cell, tissue, organ or organism. The upstream regulatory sequence may be a synthetic upstream regulatory sequence e.g. one which has been designed or artificially created, or it may be an upstream regulatory sequence that is naturally occurring, but which is not found in the normal or wild-type cell, tissue, organ or organism. A further possibility is that it is an exogenous upstream regulatory sequence, but which has been additionally modified e.g. sequence changes have been introduced according to standard techniques known in the art, in order to manipulate the activity of function of the upstream regulatory sequence.

The upstream regulatory sequence may control the transcription or translation of a further transactivator protein or may control the production of a marker protein.

In this way, the "signal" or the output of the assay may be amplified. This is therefore of particular use if the expression of the endogenous gene, whose expression is being detected by the enhancer trap method, is low. The generation of only a small amount of transactivator protein may lead to higher levels of the marker protein or further transactivator protein. Thus detection of the marker protein may be possible at lower levels of endogenous gene expression.

The degree of amplification will depend on the properties of the transactivation protein and the regulatory sequence to which it binds. The stability of the transactivation protein will also influence the degree of amplification; the more stable and longer the half life of the transactivation
protein, the longer it will be able to bind the regulatory sequence, and the more marker protein will be produced.

If the vector of use in the invention contains a sequence encoding a transactivation protein as the reporter gene then the nucleic acid encoding the marker protein which is expressed under the control of the transactivation protein upstream regulatory sequence must also be present in the cell, tissue, organ or animal of interest. This may be achieved in a number of ways e.g. using an effector line, which contains the transactivation protein upstream regulatory sequence and the gene encoding the marker protein. The effector line may therefore provide the animal, cell or embryo into which the vector of use in the invention is introduced.

Alternatively, animals of the effector line may be crossed or bred with animals into which the vector of use in the invention has been introduced.

A further alternative is that the transactivation protein upstream regulatory sequence together with the further reporter gene (e.g. gene encoding a marker protein) are introduced (e.g. administered) to the cell, tissue, organ or animal in parallel with the first (i.e. retroviral) vector i.e. the vectors or constructs are administered separately, simultaneously or sequentially. For example, the regulatory sequence and further reporter gene may be introduced separately or in the form of a genetic construct containing the transactivation protein upstream regulatory sequence and the reporter gene (e.g. encoding the marker protein).

Several transactivating systems are known in the art. One example is the GAL4 system, which is based on the yeast system. The GAL4 method has been adapted for use in Drosophila (Brand and Perrimon, 1993, Development, 118, 401-415) and zebrafish (Scheer et al, 2001, Development 128, 1099-1107).

Thus, in a preferred embodiment of the invention the reporter gene encodes a transactivation protein, preferably the yeast transactivator GAL4, and the organism (e.g. cell or embryo) into which the vector is introduced contains a GAL4-responsive promoter termed Upstream activating sequence (UAS) functionally linked to a further reporter gene which may encode a marker protein.

Furthermore, as the sequence of the reporter gene is known, it may be used to clone the genomic sequences around the insertion. The known sequence of the reporter gene thus acts as a molecular marker to enable genetic cloning of the sequence surrounding the reporter gene in addition to encoding a reporter gene that can be detected directly or indirectly. In this way, i.e. by using
standard cloning and sequencing techniques that are well known in the art e.g. PCR, the area surrounding the reporter gene, the active gene i.e. the gene which is transcribed and translated under normal conditions in the cell at that time may be identified. Thus in a preferred embodiment the method further comprises the steps of cloning and sequencing the region of the genome into which the insertion of exogenous nucleic acid (i.e. the retroviral vector) has occurred. The method may further comprise the identification, isolation, cloning and/or production of the active gene into or near which the retroviral vector has been inserted. Such active genes may be known genes or may be novel, in which case sequencing or functional analysis to try and establish the function of these novel genes may be appropriate.

The marker protein as described herein may be any gene product, including also shorter peptides as well as longer polypeptides proteins and RNA molecules, which may be detected, i.e. which gives, or may give rise to, a signal. Thus the marker protein may itself by directly signal-giving and hence directly detectable, for example by giving a visually or spectrophotometrically detectable signal (e.g. colour, luminescence, fluorescence, absorbance etc. or any detectable enzymatic reaction). Alternatively, the marker protein may be indirectly signal-giving, for example it may give rise to a signal by taking part in a further signal-giving reaction, mechanism or process (e.g. by binding to a labelled or otherwise signal-giving antibody, or by taking part in an enzymatic reaction, etc.). Thus, for example, an indirectly signal-giving marker protein may be -galactosidase, as encoded for example by lacZ. Examples of other proteins useful in these cases are alkaline phosphatase and luciferase. RNA molecules may also be detected.

The marker protein is however, preferably a protein, polypeptide or peptide or other gene product that can be detected in a live vertebrate, for example live cells, or embryos i.e. it is possible to determine whether it is present without fixing, disrupting or killing the vertebrate or cell. This also includes isolating protein or nucleic acid from the cells. Preferably the marker protein is fluorescent and thus its presence can be detected by subjecting the appropriate cell to light of the relevant wavelength and inspecting said cell under a fluorescent microscope. More preferably the marker protein is green fluorescent protein (GFP), yellow fluorescent protein (YFP), red fluorescent protein (RFP), enhanced green fluorescent protein (eGFP) (Chalfie et al. (1994), Science 11 (263): 802-5) or any other fluorescent protein. It is well known in
the art that GFP and other fluorescent proteins can be expressed in zebrafish, Drosophila and mice for example and retain their fluorescent properties.

It is important that the marker protein is not normally expressed in the cell, tissue, organ or animal subjected to the method of the invention. This means that it is not required to distinguish between "normal" and "exogenous" expression of the marker protein, and false positives are not achieved.

Thus the detection step i.e. the step of ascertaining whether any marker protein is expressed in any cell, tissue, organ or organism, may be carried out by standard means known in the art, such as exposing the cell, tissue, organ or animal to light of the appropriate wavelength and observing whether fluorescence (or any other spectrophotometric signal) is seen. The presence of said marker protein is thus indicative of expression of an active gene in that cell, tissue, organ or organism, at that time during development.

It will be seen that the methods described herein result in the production of a transgenic organism, i.e. an organism containing an insertion of foreign or exogenous nucleic acid in its genome. The foreign or exogenous nucleic acid which is inserted is the part of the retroviral vector which is flanked, i.e. surrounded, by the LTR sequences. Thus, the present invention provides methods of producing transgenic organisms wherein said methods are the methods of the invention as described herein.

Accordingly, a transgenic organism obtained or obtainable by the methods as described herein is a further embodiment of the invention. For example, a transgenic organism obtained or obtainable by introducing a retroviral vector into said organism or a cell of said organism wherein said vector comprises a basal promoter and a detectably expressible gene is an embodiment of the invention. The progeny of said organism which contains said insertion of foreign or exogenous nucleic acid in its genome also forms part of the invention.

Kits comprising a transgenic organism as described herein are also contemplated and form further aspects of the invention. Kits according to the invention may comprise one or more transgenic organisms as described herein and other components as described elsewhere herein.

Transgenic organisms as described above or kits comprising one or more transgenic organisms may be used in assays, for example assays to identify compounds or conditions which affect (e.g. increase or decrease) gene expression or in assays to identify genes which are potential drug targets. As it is possible to detect an active gene in these transgenic organisms it will also be
possible to detect changes in the level of gene expression of the active gene, for example by comparing the expression of the detectably expressible gene in the presence or the absence of such a compound or condition. A detectable increase in the expression of the detectably expressible gene indicates that the compound or condition being tested increases the expression of the active gene and a detectable decrease in the expression of the detectably expressible gene indicates that the compound or condition decreases the expression of the active gene. The detection steps will generally be performed as described elsewhere herein. It is then simply necessary to compare expression levels in the presence or absence of the compound or condition being tested. Such assays can thus be used to screen and identify candidate compounds, e.g. candidate drugs, for use in therapy, e.g. to identify candidate compounds which can affect the expression of a particular gene in a desirable manner.

The present invention thus provides the use of the methods of the invention or the transgenic organisms produced by the methods of the invention in methods of screening for compounds or conditions which affect gene expression. The present invention can in particular be used to identify compounds or conditions which affect early gene expression. Such compounds may for example have use in therapy, in particular in cancer therapy.

A sample containing the compound being tested may be administered to the transgenic organism directly, e.g. by injection into the organism, preferably into the site of normal gene expression, or by administering the sample to the organism's food, or by exposing the organism to the sample, e.g. by adding the compound to the medium in which the organisms is submerged or swimming, in the case of fish or amphibians. When different conditions are being tested, e.g. different environmental conditions, then similar methods of exposure could be used.

In addition, the present invention provides the use of the methods of the invention or the transgenic organisms produced by the methods of the invention in the identification of genes which can be targeted in therapy, e.g. genes which are potential drug targets.

In an alternative application of this technology, it is also possible that the exogenous or detectably expressible gene may encode a therapeutic polypeptide. By "therapeutic" it is meant that it is of use in treating or preventing a disease or syndrome, or clinical condition.

Gene therapy using retroviral vectors has been shown recently to be linked to the development of certain cancers such as leukaemia (Kaiser 2003,
Science, 299: 495). Gene therapy vectors utilise strong, tissue specific or ubiquitous promoters. The tendency of retroviral vectors to insert into the enhancer/promoter region of genes, combined with the risk of this occurring in a growth controlling gene, thus activating such a gene, is one explanation for the development of leukaemia that has been seen in a number of gene therapy patients.

By using vectors as described herein, i.e. containing a basal promoter controlling an exogenous or detectably expressible gene, wherein the said gene encodes a therapeutic polypeptide one would reduce or even eliminate the risk of activating an endogenous growth control gene.

In this regard, by using a basal promoter, expression of the therapeutic gene will depend upon its site of insertion - the therapeutic gene will only be expressed if the vector inserts or integrates at a site where it may "trap" or utilise an enhancer of an active gene, thus reducing or eliminating the likelihood that an endogenous oncogene, or growth-controlling gene or other gene likely to give rise to cancer is expressed due to insertion of the vector.

Thus, a further aspect of the invention relates to method of gene therapy comprising introducing (e.g. administrating) to a subject in need of said therapy a retroviral vector comprising a basal promoter and a therapeutic gene.

Also provided according to the present invention is a retroviral vector comprising a basal promoter and a therapeutic gene.

Alternatively stated, the invention further relates to the use of a retroviral vector comprising a basal promoter and a therapeutic gene in the manufacture of a medicament for gene therapy.

Alternatively stated, the invention relates to retroviral vector comprising a basal promoter and a therapeutic gene for use in gene therapy.

Alternatively stated, the invention further includes a pharmaceutical composition comprising a retroviral vector, wherein said vector comprises a basal promoter and a therapeutic gene, together with a pharmaceutical acceptable carrier or diluent.

The therapeutic gene may encode a therapeutic polypeptide or gene product as defined above.

As explained and discussed above, such a retroviral vector as used in these therapeutic aspects of the invention is an enhancer trap vector as described elsewhere herein.
The term "pharmaceutically acceptable" means that the composition ingredient or component is compatible with other ingredients or components of the composition as well as physiologically acceptable to the recipient.

The invention will now be described in more detail by virtue of the following non-limiting Examples, with reference to the drawings in which:

Figure 1 shows the sequence of the construct pCL-GATA2-YFP. This is the pCL-retrovirus plasmid backbone with a GATA2 promoter + YFP insert. First, a YFP fragment (BamHI/BglIII) was cloned into the BamHI digested pCL vector. Second, a GATA2 fragment (BamHI/BamHI) was cloned into the pCL-YFP (BamHI/BamHI) to obtain pCL-GATA2-YFP. Sequence is 2753 bases long, sequenced using LE-1, LE-2, YFP rev and pMV3 reverse primers. Figure 1 also shows the sequence of the pCL vector construct without the GATA-YFP insert.

Figure 2 shows selection of stable transgenic lines showing spatially and temporally restricted expression of YFP (Yellow Fluorescent Protein) in various tissues or organs. In all cases, anterior is to the left and dorsal is at the top. Lines were generated by infection of a pseudotyped, viral enhancer-trap vector (CLGY) into zebrafish embryos at the 500-2000 cell-stage. Infected embryos were raised (F0) and crossed against wild-type fish. Progeny from such crosses (F1) were screened using a TE2000-S inverted microscope (Nikon), with a 500/20 nm excitation filter and a 515 nm BP emission filter. Main domains of expression are: 5, Retina and spinal chord. 7, Ubiquitous. 11, Subset of neurons in brain. 32, Notochord. 22, Retina and hindbrain. 129, Hatching glands. 151, Optic tectum. 183, Telencephalon. 186, Central nervous system. 298, Sensory neurons. 374, Dorsal epidermis along anterior/posterior axis. 375, Ventral part of forebrain/hindbrain/spinal chord.

EXAMPLE 1

VECTOR CONSTRUCTION
The enhancer trap vector is based on a Murine Moloney Leukemia Virus (M-MLV) type retroviral vector (pCL; Naviaux RK et al., supra). The reporter element of the vector (YFP) was ligated into BamHI site of the empty pCL vector. Secondly the minimal promoter (Gata2 fragment) was ligated into the vector immediately upstream of the reporter element. The inserted promoter-reporter cassette is oriented in a 5'-3' orientation in respect to the viral vector (Fig. 1).
TRANSFECTION FOR RETROVIRUS PRODUCTION
The day before transfection, one 100mm plate of 80% confluent 293gp/bsr cells
(Myoshi et al., supra) was split into two 100 mm poly-L-lysine coated plates (5x10^6 cells/plate).

On the day of transfection the following two mixes were prepared:

A: 10 μg viral vector
    4 μg VSV-G
    1.5ml OptiMem (Gibco-BRL)

B: 40ul LF2000 (Invitrogen, CA)
    1.5ml OptiMem

- Solution "B" was incubated for 5 min at RT, then mixed with solution "A".
- The combined cocktail was incubated for 20 minutes at RT.
- During this 20 minute incubation, the medium was changed in the plates to a
  fresh 8ml of growth medium (DMEM, 10% FCS, 2 mM L-Glu) without
  antibiotics.
- At 20-25 minutes, 3ml of transfection-cocktail was added to each 10cm plate.
- 20 to 24 hours later, the medium again was changed, this time to 8ml fresh
  growth medium (DMEM, 10% FCS, gentamicin, 2 mM L-Glu) with antibiotics.

Harvest & concentration of virus 48/72 hours post transfection:
- Ultracentrifugation tubes were sterilised in UV-light.
- Supernatant was harvested at appropriate time, and filtered through a 0,2 μm
  filter to remove cell debris.
- Ultracentrifugation: SW41Ti rotor
  Ultra-clear centrifuge tubes (14x89 mm)
  25 000 rpm (50 000 x g)
  4°C
  90 minutes

The supernatant was removed and the tube inverted over paper towels for 5 min.
Excess liquid was aspirated off. Virus was resuspended overnight at 4 C in
appropriate volume of 1XPBS.
- This virus may be used immediately, or stored at 4°C for up to 3-4 days, or frozen at -80°C for later use (although this will reduce the titer by 50%)

**INJECTION OF VIRUS INTO ZEBRAFISH EMBRYOS**

Zebrafish embryos are dechorinated before the blastula stage in 1X Holtfreter's solution (1L dH₂O: 0.1g CaCl₂; 3.5g NaCl; 0.05g KCl; 1 ml 1M HEPES pH 7.5) containing pronaseE (Sigma). 10 to 20 nl of the concentrated virus, containing 4 µg of polybrene per ml, is injected at 3-4 locations among the blastomeres of blastula-stage zebrafish embryos (500-2000 cell stage). After injection embryos were incubated at 37°C for 2-4 hours before being transferred to 28.5°C for further raising.

**MONITORING OF TRANSGENESIS**

Embryo assay was performed, using real-time PCR, on a fraction of embryos for each new batch of virus, to evaluate the infection efficiency for zebrafish embryos. Furthermore, F1 progeny from crosses of founders are also analysed using standard PCR, to calculate the frequency of germ-line transmission of the integrated provirus (Table 1).

Table 1 - selected frequencies of F1 transgenic progeny from founders crossed against wild type.

<table>
<thead>
<tr>
<th>Founder ID</th>
<th>Freq. of germline transmission from Fo to F1</th>
<th>Enhancer trap patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLGY080802#1</td>
<td>4/40 (10%)</td>
<td>0</td>
</tr>
<tr>
<td>CLGY080802#9 male</td>
<td>33/40 (82.5%)</td>
<td>1</td>
</tr>
<tr>
<td>CLGY080802#9 female</td>
<td>5/40 (12.5%)</td>
<td>0</td>
</tr>
<tr>
<td>CLGY080802#12</td>
<td>42/80 (52.5%)</td>
<td>1</td>
</tr>
<tr>
<td>CLGY080802#13</td>
<td>35/40 (87.5%)</td>
<td>2</td>
</tr>
<tr>
<td>CLGY080802#15</td>
<td>17/40 (42.5%)</td>
<td>1</td>
</tr>
<tr>
<td>CLGY080802#18</td>
<td>28/40 (70%)</td>
<td>0</td>
</tr>
<tr>
<td>CLGY080802#19</td>
<td>11/40 (27.5%)</td>
<td>1</td>
</tr>
<tr>
<td>CLGY080802#21 male</td>
<td>16/40 (40%)</td>
<td>1</td>
</tr>
<tr>
<td>CLGY080802#21 female</td>
<td>22/35 (63%)</td>
<td>1</td>
</tr>
<tr>
<td>CLGY080802#22</td>
<td>27/40 (67.5%)</td>
<td>1</td>
</tr>
<tr>
<td>CLGY080802#34 male</td>
<td>38/40 (95%)</td>
<td>2</td>
</tr>
<tr>
<td>CLGY080802#34 female</td>
<td>7/40 (17.5%)</td>
<td>1</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------</td>
<td>---</td>
</tr>
<tr>
<td>CLGY080802#38</td>
<td>40/80 (50%)</td>
<td>1</td>
</tr>
<tr>
<td>CLGY080802#39</td>
<td>31/40 (17.5%)</td>
<td>1</td>
</tr>
</tbody>
</table>

**Embryo assay protocol:**
Embryos injected with virus were lysed in GNT-K buffer (50 mM KCl, 1.5 mM MgCl₂, 10 mM Tris-HCl pH 8.5, 0.01% gelatin, 0.45% Nonidet P-40, 0.45% Tween-20, 100 µg/ml proteinase K) at 3-5 dpf days post fertilisation:

15 injected embryos were lysed as three pools of five in 150 µl lysis buffer at 55°C for 2 hours, followed by inactivation of proteinase K at 95°C for 15 minutes. The suspension was spun in a microcentrifuge at max rpm for 1 minute. 2 µl from each sample was used as the template in qPCR.

**Real-time qPCR:** 1X

0.3 µl SFG fwd [10 µM], primer for provirus sequence.
0.3 µl SFG rev [10 µM], primer for provirus sequence.
0.3 µl Wnt-f1 [10 µM], primer for endogenous sequence (internal control).
0.3 µl Wnt-r1 [10 µM], primer for endogenous sequence (internal control).
0.4 µl SFG-FAM [10 µM], TaqMan probe for provirus sequence.
0.4 µl Wnt-YY [10 µM], TaqMan probe for endogenous sequence.
10.0 µl 2X buffer (ABI, CA, USA)
6.0 µl H2O
18.0 µl
2.0 µl template

This qPCR identifies the number of integrated proviruses per cell.

**Primers & probes for real-time PCR (all 5' to 3' orientation):**

- Wnt-f1: gtagctgcaacacaaaagcagca
- Wnt-r1: tggccccgcacgatcctccccgtaa
- Wnt-YY: Yakima Yellow-ccccgtgacgagagcctcctccggc-Darquencher
- SFG fwd*: cgctggaaagacccctctatacaca
- SFG rev*: tcgtgatggagagcctcctttga
- SFG-FAM*: 6-FAM-tgcctgacgaccacctcccaccgc-TAMRA
*(Amsterdam A et al., 1999, Genes & Dev.13:2713-2724.)*

**PCR screen:**

Genomic DNA purification from F1 progeny:

50 embryos were collected 24 hours post fertilisation in an eppendorf tube and excess liquid removed. 200 μl GNT-K buffer was added and incubated at 55°C for minimum 2 hrs. The genomic DNA was then precipitated in 2.5 volumes of absolute EtOH. The precipitation was centrifuged at maximum rpm for 30 minutes, room temperature. The supernatant was removed and the pellet air dried. The pellet was redissolved in 200 μl 0.1 TE buffer at 55°C for 1 hr, with occasional vortexing. A 1 minute spin was performed at max rpm, and the supernatant used as PCR template.

**PCR:**

Primer pair pMV2fwd/MSCVz3rev - 270 bp

PCR program "STAALE/SCR-CLXX"

96° 30s
58° 15s
72° 15s ) 35 x cycles
96° 30s )

1X
0.8 μl fwd primer [10μM]
0.8 μl rev primer [10μM]
1.6 μl dNTP [10mM]
2.0 μl 10X buffer
0.5 μl BioTaq
0.6 μl MgCl2 [50mM]
12.7 μl H2O
19.0 μl
1.0 μl template

**Primers for PCR screen (all 5' to 3' orientation):**

pMV2fwd: GAA TCG TGG TCT CGC TGT TC
MSCV3rev: CGG GTG TTC AGA ACT CGT CAG.
This shows the integrated provirus i.e. whether the animal is transgenic.

**ENHANCER-TRAP SCREENING**

Injected founders (F0 generation) were raised to sexual maturity and outcrossed to wild-type fish. F1 larvae were screened at 24 hours post fertilisation using a TE2000-S inverted microscope (Nikon), with a 500/20 nm excitation filter and a 515 nm BP emission filter for detection of YFP. To establish an enhancer-trap line, progeny from one founder that are positive for YFP, are sorted into separate Petri-dishes and grouped according to YFP expression pattern (Fig. 2).

**GENOME MAPPING OF ACTIVATED INTEGRATIONS**

5 YFP positive larvae, from a given enhancer-trap line, were sorted and raised until 5 dpf. Then genomic DNA was isolated from each individual larva. The precise location of the activated viral integration is determined using a linker mediated PCR (LM-PCR) strategy (Xiaolin W et al., 2003, Science 300: 1749-1751) (Table 2). Genomic DNA is first digested with Msel. Msel (4 base cutter) cuts genomic DNA frequently, thus reducing PCR bias against large fragments and allowing read-through of most genomic DNA fragments in a single sequence pass. Fragments are then ligated to the Msel linker. LM-PCR was performed with one primer specific to the LTR and the other primer specific to the linker. Nested PCR was then performed using a second set of primers. Based on agarose gel analysis of the PCR reaction, the activated integration was identified based on its presence in all YFP positive samples of one specific line. The identified integration was then sequenced after direct cloning into the TOPO TA cloning kit (Invitrogen, Carlsbed, CA).

The precise genomic localisation was subsequently identified by BLAST against the ENSEMBL zebrafish genome sequence (www.ensembl.org). A flanking sequence was judged to be from a genuine activated viral integration if it (i) was present exclusively in YFP positive larvae, (ii) contained both LTR and linker sequences, (iii) matched to a genomic sequence starting within 3 bases after the LTR, (iv) showed 95% or greater identity to genic DNA, (v) matched to only one locus with 95% or greater identity (modified after Wu et al 2003, supra).

Table 3 shows further characterisation of genomic locations of transcriptionally activated provirus.
<table>
<thead>
<tr>
<th>CLGY</th>
<th>GENOMIC POSITION</th>
<th>Gene</th>
<th>putative homologue gene / protein domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1st intron</td>
<td>pax6.2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.5 KB upstream of exon 1</td>
<td>novel</td>
<td>Ser/Thr protein kinase</td>
</tr>
<tr>
<td>8</td>
<td>2nd intron</td>
<td>novel</td>
<td>contains heparan sulfate 6-O-sulfotransferase 2</td>
</tr>
<tr>
<td>11</td>
<td>11th intron</td>
<td>novel</td>
<td>Obscurin</td>
</tr>
<tr>
<td>24</td>
<td>6th intron</td>
<td>novel</td>
<td>neural proliferation, differentiation and control protein-1 (NPDC-1)</td>
</tr>
<tr>
<td>2</td>
<td>nearest gene 25 KB away?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>1st intron</td>
<td>novel</td>
<td>PH (Pleckstrin homology) domain</td>
</tr>
<tr>
<td>21</td>
<td>3.6KB upstream of A, 3.3KB downstream of B</td>
<td>both A and B novel</td>
<td>non known</td>
</tr>
<tr>
<td>3</td>
<td>5 KB upstream of 1st exon</td>
<td>her4</td>
<td></td>
</tr>
<tr>
<td>129</td>
<td>3,4 KB upstream of 1st exon</td>
<td>cyclin D1</td>
<td></td>
</tr>
<tr>
<td>162</td>
<td>3 KB downstream of last exon</td>
<td>apoE</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>640 bp upstream of 1st exon</td>
<td>novel</td>
<td>Histone H2A.G</td>
</tr>
<tr>
<td>30</td>
<td>1st intron</td>
<td>novel</td>
<td>sentrin-specific protease 3</td>
</tr>
<tr>
<td>22</td>
<td>600 bp upstream of 1st exon</td>
<td>hypothetical</td>
<td>lysine-rich region containing protein</td>
</tr>
<tr>
<td>9</td>
<td>1200 bp upstream of 1st exon</td>
<td>hypothetical</td>
<td>non known</td>
</tr>
<tr>
<td>86</td>
<td>2,5 KB upstream of 1st exon</td>
<td>hypothetical</td>
<td>non known</td>
</tr>
<tr>
<td>31</td>
<td>1st intron</td>
<td>hypothetical</td>
<td>non known</td>
</tr>
<tr>
<td>142</td>
<td>1586 bp upstream of 1st exon</td>
<td>novel</td>
<td>tropomyosin family</td>
</tr>
<tr>
<td>41</td>
<td>intron</td>
<td>novel</td>
<td>SL1, shc adaptor</td>
</tr>
<tr>
<td>68</td>
<td>1st intron</td>
<td>novel</td>
<td>non known</td>
</tr>
<tr>
<td>90</td>
<td>811 bp upstream of 1st exon</td>
<td>hoxb4a</td>
<td></td>
</tr>
<tr>
<td>186</td>
<td>2962bp downstream of last exon</td>
<td>novel</td>
<td>Zfp-36</td>
</tr>
<tr>
<td>75</td>
<td>1st intron</td>
<td>hypothetical</td>
<td>phospholipid scramblase 1 or 2</td>
</tr>
<tr>
<td>20</td>
<td>1600bp upstream of 1st exon</td>
<td>hypothetical</td>
<td>non known</td>
</tr>
<tr>
<td>102</td>
<td>intron 2</td>
<td>hypothetical</td>
<td>Srrm1 serine/arginine repetitive matrix 1</td>
</tr>
<tr>
<td>39</td>
<td>last intron</td>
<td>hypothetical</td>
<td>non known</td>
</tr>
<tr>
<td>130</td>
<td>intron 2</td>
<td>efnb2b</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2nd last intron</td>
<td>hypothetical</td>
<td>zinc finger</td>
</tr>
<tr>
<td>----</td>
<td>----------------</td>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>32</td>
<td>1KB downstream of last exon</td>
<td>hypothetical</td>
<td>non known</td>
</tr>
<tr>
<td>128</td>
<td>1KB downstream of last exon in A, 2,6KB upstream of 1st exon in B</td>
<td>A-novel, B-foxd5</td>
<td>A- ambiguous</td>
</tr>
<tr>
<td>47</td>
<td>2nd last intron</td>
<td>hypothetical</td>
<td>GRIN3B</td>
</tr>
<tr>
<td>154</td>
<td>400bp upstream of 1st exon</td>
<td>hypothetical</td>
<td>non known</td>
</tr>
<tr>
<td>77</td>
<td>547bp upstream of 1st exon</td>
<td>gro2</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>700 bp downstream of last exon</td>
<td>hypothetical</td>
<td>non known</td>
</tr>
<tr>
<td>298</td>
<td>1st intron</td>
<td>novel</td>
<td>kelch repeat</td>
</tr>
<tr>
<td>50</td>
<td>last intron</td>
<td>cyp26a1</td>
<td></td>
</tr>
<tr>
<td>189</td>
<td>6 KB upstream of 1st exon</td>
<td>novel</td>
<td>DLGAP1</td>
</tr>
<tr>
<td>239</td>
<td>last intron</td>
<td>novel</td>
<td>dehydrogenase kinase</td>
</tr>
<tr>
<td>318</td>
<td>1st intron</td>
<td>clda7</td>
<td></td>
</tr>
<tr>
<td>299</td>
<td>2nd last intron</td>
<td>hypothetical</td>
<td>NOVA-2</td>
</tr>
</tbody>
</table>

**Viral Insert Mapping Protocol:**

Genomic DNA Preparation of single embryo:

- each single embryo was separated into in a 1,5 ml eppendorf tube
- 100 µl GNT-K buffer was added and incubated for 2 h at 56°C, with vortexing every 30 minutes
- 250µl (2.5 Vol.) of 96% EtOH was added and the mixture mixed
- The mixture was then centrifuged at max. speed for 30 minutes
- The supernatant was aspirated off and the pellet dried for 15-20 minutes at room temperature
- The pellet was redissolved in 25 µl 0,1 TE buffer for 60 min at 56°C with vortexing every 30 minutes
- A 1 minute spin at max. speed was then performed to clear the solution.

Genomic DNA

MseI digestion:

- 2,0 µl DNA
- 1,0 µl MseI (10 U), NEB
- 2,0 µl 10x buffer
- 0,2 µl 100x BSA
and the mixture was incubated for 4 hours at 37°C, then denatured for 20 minutes at 65°C.

Ligation: of MseI digested fragments to the MseI linker. One reaction was set up for the 5' linker and one for the 3' linker:

10 µl digestion
1,3 µl 10x buffer
1,0 µl linker (100 µM)
0,5 µl Ligase T4 (NEB, 400 U)
0,5 µl H₂O
13.3 µl total,

and the mixture was left at 16°C overnight or for 3 h in a water-bath (start temp. at 16°C).

PCR with one primer specific to the LTR and one to the linker:

5': 1st 5LTRMscI + Afl5, 2nd 5LTRnest + Afl5nest
3': 1st 3LTR + AP1, 2nd 3LTRnest + Afl3nest

1x Master Mix

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>10x buffer</td>
<td>2,50 µl</td>
</tr>
<tr>
<td>MgCl₂ (50 mM)</td>
<td>0,75 µl</td>
</tr>
<tr>
<td>fwd primer</td>
<td>0,50 µl</td>
</tr>
<tr>
<td>rev primer</td>
<td>0,50 µl</td>
</tr>
<tr>
<td>dNTP (10 mM)</td>
<td>0,50 µl</td>
</tr>
<tr>
<td>Taq (5U/µl, Invitrogen)</td>
<td>0,25 µl</td>
</tr>
<tr>
<td>H₂O</td>
<td>19,00 µl</td>
</tr>
<tr>
<td>Ligation</td>
<td>1,00 µl</td>
</tr>
</tbody>
</table>

25,00 µl

Take 1 µl from 1st PCR as template for nested PCR

PCR Program: 95°C, 2 min
95°C, 15 sec
55°C, 30 sec  
72°C, 1 min, 24 cycles

1,5 % Metaphor agarose gel: 10 μl loaded + 2 μl loading dye

5

**Direct cloning of identified integration, TOPO TA cloning (Invitrogen, cat. 45-0641):**

1 μl PCR product  
1 μl Salt solution  
3 - 3,5 μl H₂O  
0,5 - 1 μl TOPO vector  
6 μl, incubated at room temperature for 5 min  
4 μl of mix was added to One Shot competent cells E. coli (TOPO 10, cat.4500-01), mixed gently and left for 5 min on ice.  
50 μl was plated on a prepared LB-Amp. plate overnight at 37 C, (LB-Amp plates were prepared by adding 80 μl x-gal (25mg/ml), incubation at 37°C for 30 min before use)

20 5-8 white colonies were picked for 1.5 ml overnight-culture at 37°C

**Mini-Prep:** 1,5 ml of the overnight-culture was used and DNA eluted in 40 μl distilled, deionised H₂O

30 **EcoRI analytic digestion:**

2,0 μl DNA  
2,0 μl 10 x buffer H  
0,2 μl 100x BSA  
1,0 μl EcoRI (10 U, Promega)  
14,8 μl H₂O  
- 1h at 37°C

1,5% Metaphor agarose gel: 10 μl digestion + 2μl loading dye

If inconclusive results were obtained due to weak bands in gel, the reaction was repeated using up to 10 μl of template and up to 3 hrs incubation at 37°C.

**Sequencing:** 1,0 μl Big Dye 3.1 mix (ABI)
1.0 μl Seq. buffer
1.0 μl DNA
0.3 μl M13 fwd primer
6.7 μl H₂O
10.0 μl total

In the case of low yield mini-preps, increased amount of template was necessary.

PCR program: 96°C 10 sec
50°C 5 sec
60°C 4 min, 25 cycles

**Sequences for linkers and primers to amplify genomic DNA next to 3’LTR:**

**Linker upper Strand (name: Afi3-us):**
length: 36 mer
purification: PAGE
5’ GTAATACGACTCACTATAGGGCTCCG CTAAAGGAC 3’

**Linker lower strand (name: Afi3-ls):**
length: 17 mer
special modification: 5’ phosphate, 3’ amino C7.
purification: PAGE
5’-PO4-TAGTCCCTTAAG CGGAG-NH2-3’

3’LTR:
length: 25 mer
5’ GACTTGTTGGCTCGCTGTTCTTG 3’

3’LTRNest:
length: 25 mer
5’ GGTCTCCTCTGATGTGATTGACTACC 3’

AP1: (outside primer on linker)
length: 22 mer
5’ GTAATACGACTCACTATAGGGC 3’
Afl3nest: (nested primer on linker)
Length: 19 mer
5' AGGGCTCCG CTTAAGGGAC 3'

Sequences and primers to amplify genomic sequence next to 5'LTR:
Linker upper Strand (name: Afl5-us):
Length: 35 mer
purification: PAGE
5' GGAATTGCTGGTGCGATACAGGC CTTAAGAGGGAC 3'

Linker lower strand (name: Afl5-ls):
length: 16mer
special modification: 5' phosphate, 3' amino C7
purification: PAGE
5'-PO4-TAGTCCCTCTTAAG GCCT-NH2-3'

Afl5:
Length: 21 mer
5' GGAATTGCTGGTGCGATACA G3'

Afl5nest:
Length: 20 mer
5' AGTACAGGC CTTAAGAGGA 3'

5'LTRMscI:
5' TAGCTTGCCAAACCTACAGGT 3'

5'LTRnest:
5' ACCTACAGGTGGGTCTTTCA 3'
Table 3  Characterisation of genomic location of transcriptionally activated provirus.

<table>
<thead>
<tr>
<th>CLGY #</th>
<th>Genomic Position Relatively to Nearest Gene</th>
<th>Nearest Gene</th>
<th>Orthologue Prediction of Nearest Gene. (to human if not noted otherwise)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Donor splice site, intron1.</td>
<td>pl10</td>
<td>Dead-box protein 3 (DDX3X)</td>
</tr>
<tr>
<td>2</td>
<td>55 582 bp upstream</td>
<td>noz1</td>
<td>Hypothetical protein FLJ14299.</td>
</tr>
<tr>
<td>3</td>
<td>395 bp downstream</td>
<td>novel prediction</td>
<td>(HES5_RAT) [Fugu rubripes]</td>
</tr>
<tr>
<td>4</td>
<td>5 649 bp downstream</td>
<td>apoeb</td>
<td>Apolipoprotein A-I precursor (APOA1)</td>
</tr>
<tr>
<td>5</td>
<td>2nd intron</td>
<td>pax6b</td>
<td>Paired box protein PAX-6 (PAX6)</td>
</tr>
<tr>
<td>6</td>
<td>nd¹</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>530 bp upstream</td>
<td>seph</td>
<td>H3 histone, family 2 (HIST2H3C)</td>
</tr>
<tr>
<td>8</td>
<td>in exon</td>
<td>genescan prediction</td>
<td>Nd</td>
</tr>
<tr>
<td>9</td>
<td>2 961 bp upstream</td>
<td>novel prediction</td>
<td>Unknown (protein for MGC:31065) [Mus musculus]</td>
</tr>
<tr>
<td>10</td>
<td>nd²</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>nd¹</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1 447 bp downstream</td>
<td>novel prediction</td>
<td>Hypothetical protein LOC119504.</td>
</tr>
<tr>
<td>14</td>
<td>6 036 bp downstream</td>
<td>novel prediction</td>
<td>Chromosome 14 open reading frame 115 (C14orf115)</td>
</tr>
<tr>
<td>16</td>
<td>1 056 bp upstream</td>
<td>hsp47</td>
<td>Serpin I2 precursor (SERPINI2)</td>
</tr>
<tr>
<td>19</td>
<td>1st intron</td>
<td>novel prediction</td>
<td>Adenylate cyclase 2 (ADCY2)</td>
</tr>
<tr>
<td>20</td>
<td>6753 bp upstream</td>
<td>genescan prediction³</td>
<td>Dihydropyrimidine dehydrogenase (DPYD)</td>
</tr>
<tr>
<td>21</td>
<td>14041 bp downstream</td>
<td>snai1a</td>
<td>Zinc finger protein SLUG (SNAI2)</td>
</tr>
<tr>
<td>22</td>
<td>70441 bp upstream</td>
<td>sox11b</td>
<td>Transcription factor SOX-11 (SOX11)</td>
</tr>
<tr>
<td>24</td>
<td>4th intron</td>
<td>novel prediction</td>
<td>Neural prolif., diff. and control prot.-1 (NPDC-1)</td>
</tr>
<tr>
<td>30</td>
<td>9951 bp upstream</td>
<td>novel prediction</td>
<td>Hypothetical protein FLJ39237.</td>
</tr>
<tr>
<td>31</td>
<td>1st intron</td>
<td>Q7ZUH0</td>
<td>Hypothetical protein FLJ11712</td>
</tr>
<tr>
<td>32</td>
<td>6th intron</td>
<td>genescan prediction³</td>
<td>novel protein</td>
</tr>
<tr>
<td>33</td>
<td>2nd intron</td>
<td>genescan prediction³</td>
<td>novel protein</td>
</tr>
<tr>
<td>35</td>
<td>4th intron</td>
<td>novel prediction</td>
<td>Neural prolif., diff. and control prot.-1 (NPDC-1)</td>
</tr>
<tr>
<td>38</td>
<td>1970 bp upstream</td>
<td>novel prediction</td>
<td>G protein-coupled receptor 52 (GPR52)</td>
</tr>
<tr>
<td>39</td>
<td>16420 bp downstream</td>
<td>novel prediction</td>
<td>Similar to CG6405 gene product.</td>
</tr>
<tr>
<td>41</td>
<td>5890 upstream</td>
<td>novel prediction</td>
<td>Transducin-like enhancer protein 1 (TLE1)</td>
</tr>
<tr>
<td>47</td>
<td>2nd intron</td>
<td>novel prediction</td>
<td>Glucose-6-phosphatase (G6PC)</td>
</tr>
<tr>
<td>50</td>
<td>3rd intron</td>
<td>genescan prediction³</td>
<td>Hypothetical protein (DKFZP434F011).</td>
</tr>
<tr>
<td>65</td>
<td>132831 bp upstream</td>
<td>sox11b</td>
<td>Transcription factor SOX-11 (SOX11)</td>
</tr>
<tr>
<td>68</td>
<td>2nd intron</td>
<td>genescan prediction³</td>
<td>novel protein</td>
</tr>
<tr>
<td>69</td>
<td>2nd exon</td>
<td>hoxa9a</td>
<td>Homeobox protein HOX-A9 (HOXA9)</td>
</tr>
<tr>
<td>----</td>
<td>-----------------</td>
<td>-------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>75</td>
<td>13 504 bp downstream</td>
<td>sox31</td>
<td>SOX-15 protein (SOX15)</td>
</tr>
<tr>
<td>77</td>
<td>884 bp upstream</td>
<td>gro2</td>
<td>Transducin-like enhancer protein 3 (TLE3)</td>
</tr>
<tr>
<td>86</td>
<td>107 430 bp downstream</td>
<td>novel prediction</td>
<td>Transcription factor SOX-3 (SOX3)</td>
</tr>
<tr>
<td>90</td>
<td>nd$^1$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>97</td>
<td>10 601 bp downstream</td>
<td>novel prediction</td>
<td>Hypothetical gene 2810437L13Rik [Mus musculus].</td>
</tr>
<tr>
<td>101</td>
<td>10 487 bp downstream</td>
<td>Q804R3</td>
<td>ORF4 protein [Rattus norvegicus].</td>
</tr>
<tr>
<td>102</td>
<td>2nd intron</td>
<td>novel prediction</td>
<td>novel protein.</td>
</tr>
<tr>
<td>104</td>
<td>6 574 bp upstream</td>
<td>novel prediction</td>
<td>Lactadherin precursor (SPAG10)</td>
</tr>
<tr>
<td>109</td>
<td>downstream$^4$</td>
<td>novel prediction.</td>
<td>Tubulin-specific chaperon C (TBCC)</td>
</tr>
<tr>
<td>128</td>
<td>1 172 bp downstream</td>
<td>novel prediction</td>
<td>COBW-like protein (NM_018491)</td>
</tr>
<tr>
<td>129</td>
<td>6 063 bp downstream</td>
<td>novel prediction</td>
<td>Calpain 2, large subunit (CAPN2)</td>
</tr>
<tr>
<td>130</td>
<td>2nd intron</td>
<td>Ephrin B2B</td>
<td>Ephrin-B2 precursor (EFNB2)</td>
</tr>
<tr>
<td>131</td>
<td>2 675 bp upstream</td>
<td>ccnd1</td>
<td>G1/S-specific cyclin D1 (CCND1)</td>
</tr>
<tr>
<td>135</td>
<td>10 638 bp upstream</td>
<td>Q7ZZB0</td>
<td>Mitochondrial solute carrier protein (NM_145305)</td>
</tr>
<tr>
<td>142</td>
<td>1 586 bp upstream</td>
<td>novel prediction</td>
<td>Hypothetical protein (FLJ20364)</td>
</tr>
<tr>
<td></td>
<td>2,335 bp upstream</td>
<td>tpbgl</td>
<td>Trophoblast glycoprotein (TPBG)</td>
</tr>
<tr>
<td>---</td>
<td>------------------</td>
<td>-------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>154</td>
<td>11,917 bp downstream</td>
<td>novel prediction</td>
<td>Hypothetical protein (LOC253827)</td>
</tr>
<tr>
<td>155</td>
<td>75,294 bp upstream</td>
<td>eva1</td>
<td>Epithelial V-like antigen 1 precursor (EVA1)</td>
</tr>
<tr>
<td>158</td>
<td>nd&lt;sup&gt;2&lt;/sup&gt;</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>159</td>
<td>18,183 bp downstream</td>
<td>novel prediction</td>
<td>Hypothetical protein FLJ23033</td>
</tr>
<tr>
<td>160</td>
<td>834 bp downstream</td>
<td>Q7ZUG0</td>
<td>Small nuclear ribonucleoprotein E (SNRPE)</td>
</tr>
<tr>
<td>182</td>
<td>5,662 bp downstream</td>
<td>novel prediction</td>
<td>Chromosome 14 open reading frame 115 (C14orf115)</td>
</tr>
<tr>
<td>183</td>
<td>2,899 bp downstream</td>
<td>emx3</td>
<td>nd</td>
</tr>
<tr>
<td>185</td>
<td>4,902 bp upstream</td>
<td>hoxc8a</td>
<td>Homeobox protein HOX-C8 (HOXC8)</td>
</tr>
<tr>
<td>186</td>
<td>2,787 bp downstream</td>
<td>genescan prediction&lt;sup&gt;3&lt;/sup&gt;</td>
<td>Zinc finger protein 36, C3H type-like 1 (ZFP36L1)</td>
</tr>
<tr>
<td>188</td>
<td>nd&lt;sup&gt;1&lt;/sup&gt;</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>189</td>
<td>61,650 bp upstream</td>
<td>genescan prediction&lt;sup&gt;3&lt;/sup&gt;</td>
<td>Adenylate cyclase 5 (ADCY5)</td>
</tr>
<tr>
<td>195</td>
<td>36,639 bp downstream</td>
<td>nrarpa</td>
<td>Notch-reg. ankyrin repeat prot. (Nrarpa) [Mus musculus]</td>
</tr>
<tr>
<td>196</td>
<td>1,711 bp upstream</td>
<td>novel prediction</td>
<td>Transcription factor SOX-5 (SOX5)</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>Distance from Reference</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>------------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>hoxc8a</td>
<td>3,988 bp upstream</td>
<td>Homeobox protein HOX-C8 (HOXC8)</td>
<td></td>
</tr>
<tr>
<td>novel prediction</td>
<td>2nd intron</td>
<td>Mitogen-activated protein kinase 3 (MAP3K3)</td>
<td></td>
</tr>
<tr>
<td>novel prediction</td>
<td>97,704 bp downstream</td>
<td>Putative allantoicase (ALLC)</td>
<td></td>
</tr>
<tr>
<td>genescan prediction</td>
<td>1st intron</td>
<td>novel protein.</td>
<td></td>
</tr>
<tr>
<td>fgfr1</td>
<td>22,852 bp upstream</td>
<td>Fibroblast growth factor receptor 1 (FGFR1)</td>
<td></td>
</tr>
<tr>
<td>nd</td>
<td>2,874 bp upstream</td>
<td>Histone H4 (H4)</td>
<td></td>
</tr>
<tr>
<td>novel prediction</td>
<td>3,970 bp upstream</td>
<td>Myogenic factor MYF-5 (MYF5)</td>
<td></td>
</tr>
<tr>
<td>myf5</td>
<td>116 bp downstream</td>
<td>Homeobox protein HOX-B1 (HOXB1)</td>
<td></td>
</tr>
<tr>
<td>Q7ZTZ7</td>
<td>intron 4</td>
<td>Branched-chain α-ketoacid dehydrogenase kin (BCKD)</td>
<td></td>
</tr>
<tr>
<td>novel prediction</td>
<td>491 bp upstream</td>
<td>Histone 1, H2bn (HIST1H2BN)</td>
<td></td>
</tr>
<tr>
<td>novel prediction</td>
<td>175 bp upstream</td>
<td>Mediator of DNA damage checkpoint 1 (MDC1)</td>
<td></td>
</tr>
<tr>
<td>novel prediction</td>
<td>5th intron</td>
<td>P-selectin precursor (SELP)</td>
<td></td>
</tr>
<tr>
<td>novel prediction</td>
<td>37,223 bp downstream</td>
<td>Melanoma antigen recognized by T cells 2 (MART2)</td>
<td></td>
</tr>
<tr>
<td>novel prediction</td>
<td>283 bp upstream</td>
<td>Enoyl Coenz. A hydratase domain cont. 3 (ECHDC3)</td>
<td></td>
</tr>
<tr>
<td>novel protein</td>
<td>intron 4</td>
<td>novel protein</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sequence Distance</td>
<td>Type</td>
<td>Description</td>
</tr>
<tr>
<td>---</td>
<td>-------------------</td>
<td>-----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>299</td>
<td>12 752 bp downstream</td>
<td>novel prediction</td>
<td>TBC1 domain family member 4 (TBC1D4)</td>
</tr>
<tr>
<td>307</td>
<td>17 877 bp upstream</td>
<td>novel prediction</td>
<td>Tribbles homolog 2 (TRB2)</td>
</tr>
<tr>
<td>308</td>
<td>6 326 bp downstream</td>
<td>novel prediction</td>
<td>Neurotrophic tyrosine kinase, receptor type 2 (NTRK2)</td>
</tr>
<tr>
<td>318</td>
<td>1st intron</td>
<td>cldn7</td>
<td>Claudin-7 (CLDN7)</td>
</tr>
<tr>
<td>368</td>
<td>89 331 bp upstream</td>
<td>sox11b</td>
<td>Transcription factor SOX-11 (SOX11)</td>
</tr>
<tr>
<td>369</td>
<td>5 038 bp downstream</td>
<td>Q7ZVC6</td>
<td>High-mobility group box 1 (HMGB1)</td>
</tr>
<tr>
<td>372</td>
<td>743 bp upstream</td>
<td>ccnd1</td>
<td>G1/S-specific cyclin D1 (CCND1)</td>
</tr>
<tr>
<td>373</td>
<td>6 036 bp downstream</td>
<td>novel prediction</td>
<td>Hypothetical protein FLJ12735</td>
</tr>
<tr>
<td>375</td>
<td>348 bp upstream</td>
<td>ptc1</td>
<td>Patched protein homolog 1 (PTCH)</td>
</tr>
<tr>
<td>376</td>
<td>nd²</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>379</td>
<td>67 976 bp upstream</td>
<td>genescan prediction³</td>
<td>Adenylate cyclase 5 (ADCY5)</td>
</tr>
<tr>
<td>381</td>
<td>13 332 bp upstream</td>
<td>novel prediction</td>
<td>Protein kinase C binding protein 1 (PRKCBP1)</td>
</tr>
<tr>
<td>383</td>
<td>last intron</td>
<td>novel prediction</td>
<td>Hypothetical protein FLJ23403</td>
</tr>
<tr>
<td>385</td>
<td>8th intron</td>
<td>novel prediction</td>
<td>Dynein, cytoplasmic, intermediate polypep. 1 (DNCI1)</td>
</tr>
<tr>
<td>391</td>
<td>23 530 bp downstream</td>
<td>vim</td>
<td>Vimentin (VIM)</td>
</tr>
<tr>
<td>393</td>
<td>4th intron</td>
<td>novel prediction</td>
<td>Gamma-tubulin complex component 2 (TUBGCP2)</td>
</tr>
</tbody>
</table>
Claims

1. A method of enhancer trapping in an organism, comprising introducing a retroviral vector into said organism, wherein said vector comprises a basal promoter and a detectably expressible gene.

2. The method of claim 1 wherein the LTR of said retroviral vector is substantially inactive as a promoter when integrated into the genome of said organism.

3. The method of claim 1 or 2 wherein the LTR of said retroviral vector is active as a promoter in the packaging cells used for retroviral vector production.

4. The method of any preceding claim wherein the LTR of said retroviral vector is active as a promoter in mammalian cells and substantially inactive as a promoter in lower vertebrates or invertebrates.

5. The method of any preceding claim wherein transcription of said detectably expressible gene causes expression of a marker protein which can be detected in a live said organism.

6. The method of claim 5 wherein said detectably expressible gene encodes said marker protein.

7. The method of claim 5 wherein said detectably expressible gene controls the expression of said marker protein.

8. The method of any preceding claim further comprising the step of detecting said detectably expressible gene.

9. The method of claim 8 wherein in said detection step, the cells in which the detectably expressible gene is expressed are identified.

10. The method of any preceding claim wherein said organism is a vertebrate, preferably a lower vertebrate.
11. The method of any preceding claim wherein said organism is zebrafish or medaka fish.

12. The method of any preceding claim wherein said organism is an embryonic organism.

13. The method of any preceding claim wherein said basal promoter is the Gata2 core promoter, the heatshock 70 core promoter or the epithelial keratin 8 promoter.

14. The method of any preceding claim wherein said retroviral vector is based on HIV, MLV or MMLV.

15. The method of any preceding claim wherein the LTR of said retroviral vector is the MLV LTR or the HIV LTR.

16. The method of any one of claims 5 to 15 wherein said marker protein is not normally expressed in said organism.

17. The method of any one of claims 5 to 16 wherein said marker protein is a fluorescent protein, preferably a fluorescent protein selected from the group consisting of GFP, YFP, RFP and eGFP.

18. A retroviral vector for use in enhancer trapping in an organism, said vector comprising a basal promoter and a detectably expressible gene.

19. The retroviral vector of claim 18 having the features as defined in any one of claims 2 to 7 or 13 to 17.

20. A method of identifying an active gene in an organism, said method comprising

   (a) introducing a retroviral vector into said organism, wherein said vector is an enhancer trap vector which comprises

   (i) a basal promoter which is inactive in said organism in the absence of an enhancer, but which becomes active when inserted into the genome of said organism in a position where the promoter may be activated by an enhancer of said active gene; and
(ii) a detectably expressible gene, wherein transcription of said
detectably expressible gene directly or indirectly causes expression of a marker
protein; and
(b) detecting the expression of said marker protein, whereby said
expression is indicative of said active gene.

21. The method of any one of claims 1 to 17 or 20 wherein said retroviral
vector is mammalian and said organism is non-mammalian.

22. The method of claim 21, wherein said mammal is a rodent.

23. The method of claim 21 or 22, wherein said organism is a fish.

24. The method of any one of claims 1 to 17 or 20 to 23 further comprising
the identification, isolation, cloning and/or production of the active gene into or
near which the retroviral vector has been inserted.

25. The method of any one of claims 1 to 17 or 20 to 24 wherein the
introduction of the retroviral vector causes insertional mutagenesis.

26. The method of any one of claims 20 to 24, wherein the retrovirus is as
defined in claim 19.

27. The method of any one of claims 20 to 25 wherein said organism is as
defined in any one of claims 10 to 12.

28. A transgenic organism obtainable by introducing a retroviral vector into
said organism, wherein said vector comprises a basal promoter and a detectably
expressible gene.

29. The transgenic organism of claim 28 wherein said retroviral vector is as
defined in claim 19.

30. The transgenic organism of claim 28 wherein said organism is as defined
in any one of claims 10 to 12.

31. A kit comprising a retroviral vector as defined in claim 19.
32. A kit comprising transgenic organism as defined in claim 28 or 29.

33. A kit as defined in claim 31 or 32 for use in an assay.

34. Use of the method as defined in any one of claims 1 to 17 or 20 to 27 in identifying early genes.

35. Use of a transgenic organism as defined in any one of claims 28 to 30 or a method as defined in any one of claims 1 to 17 or 20 to 27 to identify compounds or conditions which affect gene expression or to identify genes which can be targeted in therapy.
Figure 1
Sequence of construct pCL-GATA-YFP

tagcgcgtctggttggcagtctgtcgtgccgtgacagtctctctgcttcgccagtctggtggcgccgcgtggctgggcatgtgctgggtggggactgtgggggctggggggggtgggggggggtggg

HindIII

coctacatcagctgaactgctggtggggactgtgggggctggggggggtgggggggggtggg

NdeI  NdeI

atatagctgccgctgggtggggactgtgggggctggggggggtgggggggggtggg

KpnI

ctagtcgctgggtggggactgtgggggctggggggggtgggggggggtggg

XmnI  XhoI  BamHI  HindIII

gagctgctgctgggtggggactgtgggggctggggggggtgggggggggtggg

SmaI

agcagcagctgctgctgggtggggactgtgggggctggggggggtgggggggggtggg

DraI

ctagtcgctgggtggggactgtgggggctggggggggtgggggggggtggg

base pairs

1 to 75

76 to 150

226 to 300

301 to 375

451 to 525

526 to 600

601 to 675

751 to 825

826 to 900

901 to 975

976 to 1050

1051 to 1125
Vector construct: **pCL**

**Sequence by Erling**

**Description:** This is the original and functional retrovirus "mother" plasmid.

```
... ... US region 7->R-region

SmaI/KnlI

KpnI

R-region 7->US region

US region end;

SphI

... ... ...
```

**base pairs**

1 to 75

76 to 150

151 to 225

226 to 300

301 to 375

376 to 450

451 to 525

526 to 600
INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12Q1/68 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 7 C12Q 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)
EPO-Internal, BIOSIS, WPI Data, PAJ, MEDLINE

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>US 6 228 639 B1 (GAITANARIS GEORGE A) 8 May 2001 (2001-05-08) columns 2,3; figure 1</td>
<td>18,19, 28-33</td>
</tr>
<tr>
<td>Y</td>
<td>SABLITZKY FRED ET AL: &quot;High frequency expression of integrated proviruses derived from enhancer trap retroviruses&quot; CELL GROWTH AND DIFFERENTIATION, vol. 4, no. 6, 1993, pages 451-459, XP008042760 ISSN: 1044-9523 the whole document</td>
<td>1-35</td>
</tr>
</tbody>
</table>

Further documents are listed in the continuation of box C.

Patent family members are listed in annex.

* Special categories of cited documents:
- "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 cited 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

Date of the actual completion of the international search
9 February 2005

Date of mailing of the international search report
23/02/2005

Name and mailing address of the ISA
European Patent Office, P.B. 5818 Patentlaan 2 NL-2280 HN Hilversum Tel. (+31-70) 340-3040, Tx 31 651 epo nl, Fax (+31-70) 340-3016

Authorized officer
Botz, J
<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>FRIEDRICH G ET AL: &quot;PROMOTER TRAPS IN EMBRYONIC STEM CELLS: A GENETIC SCREEN TO IDENTIFY AND MUTATE DEVELOPMENTAL GENES IN MICE&quot; GENES AND DEVELOPMENT, COLD SPRING HARBOR, NY, US; vol. 5, no. 9, 1 September 1991 (1991-09-01), pages 1513-1523, XP000562802 ISSN: 0890-9369 the whole document</td>
<td>18,19, 28-33</td>
</tr>
<tr>
<td>Y</td>
<td>US 5 627 058 A (RULEY ET AL) 6 May 1997 (1997-05-06) column 3 - column 9</td>
<td>1-35</td>
</tr>
<tr>
<td>Y</td>
<td>WO 98/56902 A (MEDICAL COLLEGE OF GEORGIA RESEARCH INSTITUTE, INC; LIN, SHUO) 17 December 1998 (1998-12-17) pages 9,10, paragraph 9 - page 30</td>
<td>1-35</td>
</tr>
<tr>
<td>Y</td>
<td>WEBER F ET AL: &quot;AN SV-40 ENHANCER TRAP INCORPORATES EXOGENOUS ENHANCERS OR GENERATES ENHANCERS FROM ITS OWN SEQUENCES&quot; CELL, vol. 36, no. 4, 1984, pages 983-992, XP002316872 ISSN: 0092-8674 the whole document</td>
<td>1-35</td>
</tr>
<tr>
<td>Y</td>
<td>EP 0 902 092 A (GSE-FORSCHUNGSZENTRUM FUER UMWELT UND GESUNDHEIT GMBH; CENTRE NATIONAL) 17 March 1999 (1999-03-17) the whole document</td>
<td>1-35</td>
</tr>
</tbody>
</table>

Form PCT/ISA/210 (continuation of second sheet) (January 2004)
Box No. I  Nucleotide and/or amino acid sequence(s) (Continuation of item 1.b of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application and necessary to the claimed invention, the international search was carried out on the basis of:

   a. type of material
      - [x] a sequence listing
      - [ ] table(s) related to the sequence listing

   b. format of material
      - [x] in written format
      - [x] in computer readable form

   c. time of filing/furnishing
      - [ ] contained in the international application as filed
      - [ ] filed together with the international application in computer readable form
      - [x] furnished subsequently to this Authority for the purpose of search

2. [ ] In addition, in the case that more than one version or copy of a sequence listing and/or table relating thereto has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that in the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:
<table>
<thead>
<tr>
<th>Patent document cited in search report</th>
<th>Publication date</th>
<th>Patent family member(s)</th>
<th>Publication date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>US 2002012996 A1</td>
<td>31-01-2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2001056581 A1</td>
<td>27-12-2001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2002083481 A1</td>
<td>27-06-2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU 736267 B2</td>
<td>26-07-2001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU 5807998 A</td>
<td>31-07-1998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA 2276529 A1</td>
<td>09-07-1998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 2001507577 T</td>
<td>12-06-2001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 9829533 A1</td>
<td>09-07-1998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT 223483 T</td>
<td>15-09-2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU 754993 B2</td>
<td>28-11-2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU 7955898 A</td>
<td>30-12-1998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA 2293722 A1</td>
<td>17-12-1998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE 69807687 D1</td>
<td>10-10-2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE 69807687 T2</td>
<td>02-01-2003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DK 986640 T3</td>
<td>06-01-2003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP 1240824 A1</td>
<td>18-09-2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP 0986640 A2</td>
<td>22-03-2000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES 2179512 T3</td>
<td>16-01-2003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 2002504821 T</td>
<td>12-02-2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PT 986640 T</td>
<td>31-12-2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 9856902 A2</td>
<td>17-12-1998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2002178461 A1</td>
<td>28-11-2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP 0902092 A2</td>
<td>17-03-1999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 11187876 A</td>
<td>13-07-1999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 6656735 B1</td>
<td>02-12-2003</td>
</tr>
</tbody>
</table>