Title: DNA sequences comprising gene transcription regulatory qualities and methods for detecting and using such DNA sequences.

The invention relates to the fields of medicine and cellular biology. The invention in particular relates to means and methods for regulation of gene transcription. The invention further relates to means and methods for determining whether a DNA sequence comprises a gene transcription modulating quality and/or a gene transcription repressing quality.

With the progression of the various genome projects, sequences of entire organismal genomes have become available. The flood of data has raised the interest of many investigators. One of the more noticeable discoveries was the observation that the human genome does not code for significantly more genes than the genome of simple organisms like the fruit fly. The focus of many investigators is now shifting from the identification of genes to the determination of gene expression and gene function. Examples of such technologies are DNA microarrays, functional genomics applications and proteomics. These technologies have in common that they are centered around the function and expression of coding sequences. However, while our knowledge of genes increases dramatically, the understanding of how the expression of the genes is regulated is limiting the ability to apply this rapidly increasing knowledge. This is for instance the case in the generation of transgenic plants and animals and in human gene therapy. In these applications foreign nucleic acid is typically introduced into cells to obtain expression of coding sequences. Often integration of the foreign nucleic acid into the genome of the cell is required for prolonged function of the introduced sequences. However, integration of sequences into the genome leads to unpredictability of expression because the surrounding DNA influences the transcription of the integrated sequences. This unpredictability is in part due to the fact that introduced sequences cannot be provided yet with sufficient

genetic information to functionally isolate the integrated sequences from the transcription influencing effects of the surrounding DNA. In another part this is due to the fact that not enough is known about the transcription influencing effects of surrounding DNA.

The present invention is concerned with DNA sequences that comprise a capacity to influence transcription of genes *in cis*. Typically, although not necessarily, the investigated sequences do not code by themselves for a functional protein. Various sequence elements with the capacity to affect gene transcription *in cis*, have been identified. These elements range from promoters, enhancers, and silencers to boundary elements and matrix attachment regions.

The fact that so many different types of regulatory sequences have been discovered gives the impression that it is very easy to design effective expression cassettes. However, quite the contrary is true. The designing of expression cassettes is still often driven by trial and error. It is quite often possible to obtain some kind of expression of a foreign gene in a target cell or its progeny. However, very often it is difficult to predict with any kind of accuracy the level of expression or the persistence of expression that an expression cassette can display in a target cell.

The present invention provides among others means and methods for detecting and isolating new transcription regulatory elements. A method of detecting, and optionally selecting, a DNA sequence with a gene transcription-modulating quality is provided, comprising providing a transcription system with a variety of a fragment-comprising vectors, said vectors comprising i) an element with a gene-transcription repressing quality, and ii) a promoter directing transcription of a reporter gene, the method further comprising performing a selection step in said transcription system in order to identify said DNA sequence with said gene transcription modulating quality. In a

preferred embodiment said fragments are located between i) said element with a gene-transcription repressing quality, and ii) said promoter directing transcription of said reporter gene. RNA polymerase initiates the transcription process after binding to a specific sequence, called the promoter, that signals where RNA synthesis should begin. A modulating quality can enhance transcription from said promoter *in cis*, in a given cell type and/or a given promoter. The same DNA sequence can comprise an enhancing quality in one type of cell or with one type of promoter, whereas it can comprise another or no gene transcription modulating quality in another cell or with another type of promoter. Transcription can be influenced through a direct effect of the regulatory element (or the protein(s) binding to it) on the transcription of a particular promoter. Transcription can however, also be influenced by an indirect effect, for instance because the regulatory element affects the function of one or more other regulatory elements. A gene transcription modulating quality can also comprise a stable gene transcription quality. With stable is meant that the observed transcription level is not significantly changed over at least 30 cell divisions. A stable quality is useful in situations wherein expression characteristics should be predictable over many cell divisions. Typical examples are cell lines transfected with foreign genes. Other examples are transgenic animals and plants and gene therapies. Very often, introduced expression cassettes function differently after increasing numbers of cell divisions or plant or animal generations. In a preferred embodiment a stable quality comprises a capacity to maintain gene transcription in subsequent generations of a transgenic plant or animal. Of course in case expression is inducible said quality comprises the quality to maintain inducibility of expression in subsequent generations of a transgenic plant or animal. Frequently, expression levels drop dramatically with increasing numbers of cell divisions. With a method of the invention it is possible to detect and optionally select a DNA sequence that is capable of at least in part preventing the dramatic drop in transcription levels with increasing numbers of cell

divisions. Thus in a preferred embodiment said gene transcription modulating quality comprises a stable gene transcription quality. Strikingly, fragments comprising a DNA sequence with said stable gene transcription quality can be detected and optionally selected with a method of the invention, in spite of the fact that said method does not necessarily measure long term stability of transcription. In a preferred embodiment of the invention said gene transcription modulating quality comprises a stable gene transcription enhancing quality. It has been observed that incorporation of a DNA sequence with a gene transcription modulating quality in an expression vector with a gene of interest, results in a higher level of transcription of said gene of interest, upon integration of the expression vector in the genome of a cell and moreover that said higher gene expression level is also more stable than in the absence of said DNA sequence with a gene transcription modulating quality.

In experiments designed to introduce a gene of interest into the genome of a cell and to obtain expression of said gene of interest, the following has been observed. If together with said gene of interest also a DNA sequence with a gene transcription modulating quality was introduced, more clones could be detected that expressed more than a certain amount of gene product of said gene of interest, than when said DNA sequence was not introduced together with said gene of interest. Thus the present invention also provides a method for increasing the number of cells expressing a more than a certain level of a gene product of a gene of interest upon providing said gene of interest to the genome of said cells, comprising providing said cell with a DNA sequence comprising a gene transcription modulating quality together with said gene of interest.

The chances of detecting a fragment with a gene transcription-modulating quality vary with the source from which the fragments are derived. Typically, there is no prior knowledge of the presence or absence of fragments with said quality. In those situations many fragments will not comprise a DNA sequence with a gene transcription-modulating quality. In

these situations a formal selection step for DNA sequences with said quality is introduced. This is done by selection vectors comprising said sequence on the basis of a feature of a product of said reporter gene, that can be selected for or against. For instance, said gene product may induce fluorescence or a color deposit (e.g. green fluorescent protein and derivatives, luciferase, or alkaline phosphatase) or confer antibiotic resistance or induce apoptosis and cell death.

A method of the invention is particularly suited for detecting and optionally selecting a DNA sequence comprising a gene transcription-enhancing quality. It has been observed that at least some of the selected DNA sequences, when incorporated into an expression vector comprising a gene of interest, can dramatically increase gene transcription of said gene of interest in a host cell even when the vector does not comprise an element with a gene-transcription repressing quality. This gene transcription enhancing quality is very useful in cell lines transfected with foreign genes or in transgenic animals and plants.

Said transcription system can be a cell free *in vitro* transcription system. With the current expertise in automation such cell free systems can be accurate and quick. However, for the present invention said transcription system preferably comprises host cells. Using host cells warrants that fragments are detected and optionally selected with activity in cells.

An element with a gene transcription repressing quality will, in a method of the invention, repress transcription from a promoter in the transcription system used. Said repression does not have to lead to undetectable expression levels. Important is that the difference in expression levels in the absence or presence of repression is detectable and optionally selectable. In a preferred embodiment gene-transcription repression in said vectors results in gene-transcription repressing chromatin. In this preferred embodiment DNA sequences can be detected, and optionally selected that are capable of at least in part counteracting the formation of gene-transcription repressing chromatin. In one aspect a DNA sequence capable of at least in part

counteracting the formation of gene-transcription repressing chromatin comprises a stable gene transcription quality. In a preferred embodiment the DNA sequence involved in gene-transcription repression is a DNA sequence that is recognized by a protein complex and wherein said transcription system comprises said complex. Preferably said complex comprises a heterochromatin-binding protein comprising HP1, a Polycomb-group (Pc-G) protein, a histone deacetylase activity or MeCP2 (methyl-CpG-binding protein). Many organisms comprise one or more of these proteins. These proteins frequently exhibit activity in other species as well. Said complex can thus also comprise proteins from two or more species. The mentioned set of known chromatin-associated protein complexes are able to convey long-range repression over many base pairs. The complexes are also involved in stably transferring the repressed status of genes to daughter cells upon cell division. Sequences selected in this way are able to convey long-range anti-repression over many base pairs (van der Vlag et al., 2000).

The vector used can be any vector that is suitable for cloning DNA and that can be used in a transcription system. When host cells are used it is preferred that the vector is an episomally replicating vector. In this way, effects due to different sites of integration of the vector are avoided. DNA elements flanking the vector at the site of integration can have effects on the level of transcription of the promoter and thereby mimic effects of fragments comprising DNA sequences with a gene transcription modulating quality. In a preferred embodiment said vector comprises a replication origin from the Epstein-Barr virus (EBV), OriP, and a nuclear antigen (EBNA-1). Such vectors are capable of replicating in many types of eukaryotic cells and assemble into chromatin under appropriate conditions.

In another aspect the invention provides a DNA sequence comprising i) a DNA sequence isolated from a plant or vertebrate, or derivatives thereof, or ii) a synthetic DNA sequence or one constructed by means of genetic

engineering, which DNA sequence is a repression-inhibiting sequence which, by the method according to the present invention can be detected, selected and optionally cloned. In another aspect the invention provides a DNA sequence comprising i) a DNA sequence isolated from a plant or vertebrate, or derivatives thereof, or ii) a synthetic DNA sequence or one constructed by means of genetic engineering, which DNA sequence is detected, selected and optionally cloned by means of the method of the invention. Preferably said DNA sequence comprises a sequence as depicted in Table 4A or a functional homologue thereof. A functional homologue of a sequence as depicted in Table 4 is a sequence derived with the information given in Table 4 (be it table 4A or table 4B). For instance, a sequence that can be derived from a sequence in Table 4 by deleting, modifying and/or inserting bases in or from a sequence listed in Table 4, wherein said derived sequence comprises the same activity in kind, not necessarily in amount, of a sequence as depicted in Table 4. A functional homologue is further a sequence comprising a part from two or more sequences depicted in Table 4. A synthetic DNA sequence is a sequence that is not derived directly or indirectly from a sequence present in an organism. For instance a sequence comprising a drosophila scs or scs' sequence is not a synthetic sequence, even when the scs or scs' sequence was artificially generated.

In one aspect the invention is concerned with increasing knowledge of higher order gene regulation and with means and methods for utilizing this knowledge. Whereas elements, such as classical promoters and enhancers, have been characterized that direct and regulate transcription of single genes, higher order regulatory elements that govern the gene transcription capabilities of entire chromosome regions have as yet received little attention. Much of our knowledge regarding such higher order elements comes from the study of embryogenesis. During embryogenesis cells become committed to

different developmental pathways. Once committed, cells rarely change their fates, even after many cell divisions.

It has become increasingly clear that the stable transmission of cell type specific gene transcription patterns is not dependent on the context of a promoter, but is instead mediated by changes in the structure of the DNA and associated proteins, termed chromatin. Gene regulation at the chromosomal level involves modifications of DNA (e.g. methylation), histones, (e.g. acetylation and/or methylation), and long-range interactions between distant chromosomal elements.

The chromatin template is a highly condensed complex of DNA, histones, and non-histone proteins, which is able to package the entire genome into the nucleus and simultaneously allow the appropriate transcription of specific genes. The eukaryotic chromosome is not a uniform template for the activation of gene transcription. Different types of chromatin and chromatin regions can be distinguished, which differentially affect gene transcription. The so-called heterochromatin regions identify 'closed' chromatin structures whereas euchromatin is associated with a more diffuse and 'open' chromatin structure. The euchromatin region can be subject to structural changes, resulting in more or less condensed structures, referred to as facultative heterochromatin and euchromatin. The formation of facultative euchromatin or heterochromatin is believed to represent the underlying mechanism of chromatin-mediated gene regulation, keeping genes in an active or a repressed state, in a cell type specific manner.

In all eukaryotes several chromatin-associated protein complexes have been identified that are involved in the maintenance of cell type specificity, one of which is the Polycomb group (PcG) complex. The PcG complex is involved in the stable repression of genes, in which changes in chromatin structure are believed to play an important role. Similarly, a second class of proteins, named the trithorax group (TrG), has been identified that counteracts the action of the PcG proteins. TrG proteins are involved in the

maintenance of gene gene transcription. Based on their respective modes of action, PcG and TrG proteins therefore represent a cellular memory system that is important for the heritable transmission of gene transcription patterns.

How PcG and TrG complexes are associated with their target genes is still unclear. Genetic studies have characterized cis-acting regulatory sequences that maintain transcriptionally inactive states of genes. The silencing mediated by these cis-acting regulatory sequences is dependent on the presence of functional PcG proteins, and hence these sequences have been termed PcG response elements (PREs). Sequences have been identified that are involved in PcG mediated repression of chromatin. As yet however, (in both vertebrates or plants) complete PREs comprising all sequence information required to mediate repression of chromatin have not been found.

In addition, it has, as yet, not been possible to study sequences with long range repression capabilities in a coherent way. This is for a large part due to the inability to systematically screen for such long-range acting sequences. In one aspect the invention provides means and methods for systematically detecting such sequences in DNA. In one embodiment the invention provides a method for identifying a DNA sequence with a gene transcription repressing quality comprising,

- providing a collection of test nucleic acids,
- generating a collection of expression vectors comprising test nucleic acids and a first reporter gene under transcriptional control of a promoter,
- providing cells with said collection of expression vectors,
- selecting a cell or vector-containing progeny thereof, wherein transcription of said first reporter gene is repressed, and
- identifying said test nucleic acid in said cell. Said identified test nucleic acid comprises the capacity to repress said promoter function and thus comprises a gene transcription repressing quality. Preferably, said identified test nucleic acid is also retreived and cloned. Said quality comprises at least in part, the capacity to reduce the level of transcription from said promoter when

physically linked to said promoter, compared to the level in the absence of the DNA sequence with said quality. In a preferred embodiment said gene transcription repressing quality comprises a gene transcription repressing chromatin quality. I.e. wherein said reduction of the level of transcription is the result of chromatin having a gene transcription repressing configuration. This configuration preferably encompasses said promoter. However, it is also possible that said configuration encompasses an enhancer or the like thereby at least in part inactivating the transcription enhancing effect of said enhancer on said promoter. In a particularly preferred embodiment said DNA sequence with a gene transcription repressing chromatin quality comprises a polycomb-group-like responsive element.

Using the above mentioned method it is possible to retrieve several nucleic acid sequences comprising the capacity to reduce the level of transcription from a promoter and thus nucleic acid sequences comprising a gene transcription repressing quality. Sequences with analogous function can be compared with each other for sequence similarities such that one or more consensus sequences for elements with a gene transcription repressing quality such as polycomb-group-like responsive elements, can be deduced. Moreover, considering that entire sequences of organismal genomes are known and more will follow shortly, it is possible to screen these genomes or parts thereof, and predict the occurrence of these sequences in the genome. Knowledge of the occurrence and localization of DNA sequences comprising gene transcription-modulating qualities and/or gene transcription repressing qualities in the genome will greatly increase our knowledge of higher order regulation of gene transcription in the genome.

A Polycomb-group response element is an element that is capable of repressing the transcription from a promoter in response to the direct and/or indirect interaction of one or more Polycomb group proteins with said element.

A polycomb-group-like response element is a Polycomb-group response element or alternatively it is an element capable of repressing the transcription of a promoter upon the direct and/or indirect interaction of one or more proteins with said element, wherein said one or more proteins do not belong to the Polycomb-group, but wherein as a result of said interaction gene transcription repressing chromatin is formed. Examples of such proteins are chromatin-associated proteins such as heterochromatin protein1 (HP1) (Eisenberg et al., 1990) Another chromatin-associated protein that represses gene activity is methyl-CpG-binding protein, MeCP2 (Nan et al., 1997). In a preferred embodiment a polycomb-group-like responsive element of the invention comprises the capacity to repress transcription of a promoter over long distances, preferably over more than 2000 base pairs (van der Vlag et al., 2000).

A collection of test nucleic acids can be generated in many ways. Using artificial sequences as test nucleic acids it is possible to obtain a consensus sequence for a gene transcription repressing quality. Different qualities can comprise different consensus sequences. Preferably said collection is generated from chromosomal DNA. In this way a gene transcription repressing quality comprising a sequence occurring naturally in the chromosome is found. This has the advantage that the location of these qualities in the chromosome can be determined whereupon their influence on higher order gene transcription in said location can be resolved.

A reporter gene is a gene encoding an expression product of which the presence can be detected directly or indirectly in a cell. In methods for detecting a gene transcription repressing quality, the transfer of an expression vector into a cell will lead to expression of said reporter gene. However, in case test nucleic acid comprises a gene transcription repressing quality, such as a polycomb-group-like responsive element, expression will be repressed in said cell thereby leading to an at least in part reduced expression of said reporter gene. The presence or absence of a nucleic acid capable of repressing

transcription of said promoter can thus be detected by detecting said expression product in said cell, whereby reduced detection or no detection indicates the presence of a gene transcription repressing quality. A reporter gene can encode a fluorescent reporter protein. Reduced expression can then be detected by fluorometric means for instance in a flow cytometer. Cells displaying no or low fluorescence can be sorted using a fluorescence activated cell sorter and the expression vector and/or the test nucleic acid isolated, for instance by means of an amplification reaction. Preferably, said first reporter gene comprises a selectable reporter gene, the expression of which directly or indirectly provides said cell with at least a growth disadvantage over cells not expressing or expressing low levels of said first reporter gene. In cases where DNA sequences with a gene-transcription repressing quality are screened for, expression of said first reporter gene is, preferably, directly or indirectly toxic to said cell. Non-limiting examples of such a toxic expression product is ricin or toxic variants thereof . In another example, said first reporter gene encodes an apoptosis inducing gene product. Preferably said apoptosis inducing gene product comprises adenovirus 13S E1A or a functional equivalent thereof (Breckenridge and Shore, 2000). In another embodiment said apoptosis inducing gene product comprises apoptin or a functional equivalent thereof (Pietersen and Noteborn, 2000).

Another example is a gene encoding a so-called suicide product such as the herpes simplex virus thymidine kinase (HSV-tk). Addition of gancyclovir to cultures of cells expressing HSV-tk will result in the formation of a toxic substance in these cells and therefore kill these cells. In a particularly preferred embodiment said suicide gene comprises cytosine deaminase. Cytosine deaminase converts cytosine to uracil. This enzyme activity is found in prokaryotes and lower eukaryotes, but is absent in higher eukaryotes. The gene is used as a metabolic suicide gene in combination with the prodrug 5-fluorocytosine (5-FC). Cytosine deaminase is able to convert the non-toxic 5-FC

into 5-fluorouracil, which kills cells by disrupting DNA synthesis, thereby triggering apoptosis (Mullen et al., 1992; Wei and Huber, 1996).

A promoter controlling transcription of said first reporter gene can be any promoter that is active, or can be activated in said cells. By selecting a particular promoter, it is possible to select a gene transcription repressing quality such as a polycomb-group-like responsive element capable of repressing transcription of said particular promoter. In this way it is possible to select qualities that specifically repress the class of promoters said particular promoter belongs to. In a preferred embodiment said promoter comprises a promoter of which the activity can be induced upon providing a cell comprising said promoter with a signal. Such an inducible promoter preferably comprises a tetracycline responsive promoter. The signal here being tetracycline, doxycycline and equivalent compounds. Such promoters can also be adapted for tetracycline responsiveness in eukaryotic cells (Yin et al., 1996). Promoters and transacting molecules are available that either induce or repress expression of a gene, upon the addition of tetracycline or equivalents thereof.

Cells transfected with an expression vector of the invention can, with a typically low frequency and for reasons not associated with the presence of a DNA sequence with a gene transcription repressing quality, not express detectable amounts of expression product of said first reporter gene. This can for instance be due to a recombination event disrupting the coding sequence of said first reporter gene. In a preferred embodiment of the invention said collection of expression vectors further comprises a second reporter gene. Expression of said second reporter gene is preferably under the control of a second promoter. Methods for the detection of expression of an expression product of said second reporter gene can be used to confirm the expression repressing activity of said test nucleic acid, thereby at least in part reducing the number of cells falsely not expressing said first reporter gene. In a

preferred embodiment said second reporter gene is used to select for cells comprising an expression cassette. In this way, cells not comprising said expression cassette can easily be disregarded. To this end said expression product of said second reporter gene preferably comprises a positive dominant selectable reporter gene. Preferably, said positive dominant selectable reporter gene encodes an expression product capable of conferring resistance to an otherwise toxic compound. Non-limiting examples are G418 resistance and hygromycin resistance.

Considering that a gene transcription repressing quality can suppress transcription it is preferred that in this embodiment an expression vector further comprises at least one DNA sequence with a gene transcription modulating quality, capable of counteracting transcription repressing effects of a DNA sequence with a gene transcription repressing quality. The placement of said transcription counteracting element in the expression vector is preferably such that it effectively interferes with a reducing effect of said gene transcription repressing quality could have on the level of transcription of said second reporter gene. In a preferred embodiment said DNA sequence with a gene transcription modulating quality functionally separates the expression cassettes comprising said first and said second reporter gene. Preferably, said second reporter gene (and promoter controlling transcription of said second reporter gene) is flanked by DNA sequences with a gene transcription modulating quality. Examples of DNA sequences with a gene transcription modulating quality are the so-called STAR elements listed in Tables 1 and 2.

Methods of the invention result in the cloning and identification of a number of elements comprising a gene transcription modulating and/or a gene transcription repressing quality. Such an element may contain irrelevant nucleic acid that is not instrumental in performing said quality, for instance not involved in the formation of gene-transcription repressing chromatin.

Functional sequences in such elements can be delineated by various methods known in the art. In one embodiment deletions and/or substitutions are made in a DNA sequence with a gene transcription modulating or a gene transcription repressing quality. DNA that is modified in such a way, is tested for activity in a method of the invention. This can be done using a single modified nucleic acid or by generating a collection of test nucleic acids comprising said modified nucleic acid. Elucidation of functional sequences within DNA sequences of the invention enables the elucidation of consensus sequences for elements with a gene transcription modulating and/or a gene transcription repressing quality. Considering that there are several polycomb-group-like complexes that each comprise different functionalities and expression patterns, it is anticipated that more than one type of consensus sequence is found with a method of the invention. Analogously it is anticipated that more than one type of consensus sequence is found for an element comprising a gene transcription modulating quality. The invention thus further provides a library of isolated and/or recombinant nucleic acids comprising gene transcription modulating and/or gene transcription repressing qualities such as polycomb-group-like response elements. In one embodiment said library comprises isolated and/or recombinant nucleic acids comprising the same consensus sequence. In a preferred embodiment said library comprises more than one type of consensus sequence. Said library can be used for instance for determining whether a given DNA molecule comprises a DNA modulating quality. In a preferred embodiment said library comprises essentially all elements with a gene transcription enhancing function, elements comprising a stable gene transcription quality and/or elements with a gene transcription repressing quality such as polycomb-group-like response elements, of a chromosome. Together with knowledge on the location of these elements on a chromosome this allows a person skilled in the art to generate a prediction for higher order regulation of gene expression of genes naturally present on said chromosome and for genes (foreign nucleic acid) introduced

into said chromosome by recombinant means. Such a prediction can be used for example to select a suitable candidate location on said chromosome for the insertion of foreign DNA. A suitable location can be a location expected to be specifically expressed in a certain cell, cell type and/or tissue. Preferably, said chromosome comprises chromosome 21 or chromosome 22. In a particularly preferred embodiment all DNA sequences comprising a gene transcription modulating or a gene transcription repressing quality in a cell, are in the library. In this embodiment the entire genome can be used for the prediction of a suitable candidate location. In one embodiment said library has been generated in different cell lines of species ranging from plants to human. In different cell lines and/or species different proteins (or protein complexes) capable of interacting with DNA sequences with a gene transcription repressing quality, will be expressed, resulting in different DNA elements with a gene transcription repressing quality. Similarly different proteins that interact directly or indirectly with DNA sequences comprising a gene transcription modulating quality will be expressed. Therefore the make-up of the library is cell-type dependent and dependent on the presence of the relevant proteins. This is also the case with polycomb-group-like response elements. If HP1 is expressed in cell type one, elements depending on HP1 will be detected by method of invention. If HP1 is not expressed in cell type two, method of invention will not detect the element that has been retrieved from cell type one.

In one aspect of the invention said library comprises at least one element capable of at least in part counteracting the formation of gene-transcription repressing chromatin. Together with knowledge of the location of DNA sequences with a gene transcription repressing quality on a chromosome or genome, knowledge of the location of such counteracting elements allows more accurate prediction of higher order regulation of gene transcription of (inserted) genes in said chromosome or genome. Preferably said library further comprises other transcription regulatory elements such as enhancers and

silencers. Although such sequences have limited influence on higher order gene regulation, information on the location of such other sequences further increases the accuracy of the prediction on suitable locations in the genome for the expression of foreign sequences introduced therein. Preferably, said library comprises essentially all DNA sequences comprising a gene transcription modulating quality and/or all other regulatory sequences of a chromosome.

Considering that already a chromosome typically consists of several tens of millions of bases, it is preferred that the information that the library can give on higher order gene regulation is incorporated into an at least partially automated system.

Another use of a library of the invention is the generation of a prediction on transcription of genes upon targeted modification of sequences on a chromosome such that "higher order" regulatory sequences are mutated. For instance, one or more polycomb-group-like responsive elements of the invention, and/or other regulatory elements on said chromosome can be mutated. This is expected to change the transcription levels of the genes that are in the vicinity of the polycomb-group-like responsive elements and/or other expression modulating elements.

Yet another use of a library or system of the invention is the prediction of gene expression resulting from mutations in the genome. In cases where a mutation results in altered gene transcription, detection of such altered gene transcription can indicate the presence of said naturally occurring mutation. This approach is useful for instance in limiting the number of sequences or proteins to be tested in a diagnostic assay. This is particularly important in microarray approaches because in these approaches the number of expressed sequences to be tested for, is limited by the number of sequences that an array can maximally hold. With means and methods of the invention it is possible to limit the number of sequences to be tested in microarray approaches.

Yet another use of a system or library of the invention is the discovery of drug targets. Regulatory elements, be they "higher order" elements or not, function because of the protein (complexes) that can bind to them. A system of the invention can be used to determine whether targeting of drugs to interfere with the binding or function of a particular protein (complex) holds promise for the alteration of expression of a particular gene.

The invention further provides a DNA sequence comprising a gene-transcription repressing quality obtainable by a method of the invention. In a preferred embodiment said DNA sequence comprising a gene-transcription repressing quality is derived from a vertebrate of a plant. More preferably, said DNA sequence comprising a gene-transcription repressing quality comprises a sequence according to table 4 B, or a functional homologue thereof. It is also possible to provide a DNA construct with a DNA sequence of the invention, or to modify such a DNA sequence. In a preferred embodiment a DNA construct is providid comprising a promoter operably linked with a nucleic acid of interest. Preferably, the amount of activity of a quality of said DNA sequence with a gene transcription modulating and/or repressing quality, is dependent on the orientation of said DNA sequence in said construct, compared to said promoter. Preferably said gene transcription modulating and/or repressing quality is dependent on the presence of a signal. Preferably, said signal comprises a DNA binding protein. Preferably, said signal comprises a human immuno-deficiency virus TAT protein.

One of the uses of a DNA sequence comprising a gene transcription modulating quality or a gene transcription repressing quality is of course the regulation of transcription of a gene of interest. Transcription of a gene of interest can be altered by altering sequences in the vicinity of said gene such that a DNA sequence with said quality is provided or removed. Specific

expression characteristics can be designed by combining (parts of) DNA sequences with a gene transcription modulating and/or a gene transcription repressing quality. For instance, duplication of a sequence with a stable gene transcription quality in an expression vector will lead to improved stability of expression in a target cell or progeny upon introduction of said vector in said target cell. By combining DNA sequences with gene transcription modulating qualities altered gene transcription modulating qualities can be generated either in kind or amount or both.

It is also possible to design DNA sequences with a desired gene transcription modulating and/or a gene transcription repressing quality. DNA binding proteins together with other proteins and DNA sequences determine qualities of the DNA sequence. It is possible to insert one or more other protein binding DNA sequences into a DNA sequence with a quality. By allowing binding of the binding protein(s) it is possible to interfere with, or direct, the quality, thus allowing the generation of DNA sequences with designer qualities. It is of course also possible to remove protein binding sites from a DNA sequence with a particular gene transcription modulating and/or a gene transcription repressing quality thereby altering the quality of the resulting DNA sequences. The combination of addition and removal is also possible. Particular gene transcription modulating and/or gene transcription repressing qualities can be selected for by tuning detection methods described in the present invention. It is for instance possible to synthesize DNA sequences with inducible gene transcription modulating and/or gene transcription repressing qualities. By for instance including TAT-binding elements in a DNA sequence comprising a gene-transcription repressing quality, it is possible to at least in part inactivate the gene-transcription repressing quality in a cell comprising TAT. Similarly there are DNA binding proteins available that only bind to their target sequence in the absence or presence of a signal. Non-limiting examples of such proteins are the TET-repressor and the various mutations

thereof, the lac-repressor, steroid hormone receptors, the retinoic acid receptor, and derivatives. It is possible for instance to design a DNA sequence with a cell type specific gene transcription modulating and/or gene transcription repressing quality. For instance, in case of the above mentioned TAT example. The referred to DNA sequence can be made specific for HIV infected cells that express TAT. Alternatively, The DNA sequence can be made specific for a protien complex that is expressed in a cell type specific fashion.

Expression constructs comprising a DNA sequence comprising a gene transcription modulating and/or gene transcription repressing quality are suitable for obtaining expression from said construct in cells comprising more than one copy of said expression construct. Also when the expression construct is present in the genome of said cell and, also when the expression cassette is present in more than one copy in said cell. Moreover, they even work when integrated into the same position in more than one copy.

In a preferred embodiment of the invention said DNA sequence with a gene transcription modulating quality comprises a so-called STAR (Stabilizing Anti-Repression) sequence. A STAR sequence as used herein refers to a DNA sequence comprising one or more of the mentioned gene transcription modulating qualities.

Several methods are available in the art to extract sequence identifiers from a family of DNA sequences sharing a certain common feature. Such sequence identifiers can subsequently be used to identify sequences that share one or more identifiers. Sequences sharing such one or more identifiers are likely to be a member of the same family of sequences, i.e are likely to share the common feature of the family. In the present application a large number of sequences comprising STAR activity (so-called STAR sequences) were used to obtain sequence identifiers (patterns) which are characteristic for sequences

comprising STAR activity. These patterns can be used to determine whether a test sequence is likely to contain STAR activity. In one aspect the invention thus provides a method for detecting the presence of a STAR sequence within a nucleic acid sequence of about 50-5000 base pairs, comprising determining the frequency of occurrence in said sequence of at least one sequence pattern and determining that said frequency of occurrence is representative of the frequency of occurrence of said at least one sequence pattern in at least one sequence comprising a STAR sequence. In principle any method is suited for determining whether a sequence pattern is representative of a STAR sequence. Many different methods are available in the art. In a preferred embodiment of the invention the step of determining that said occurrence is representative of the frequency of occurrence of said at least one sequence pattern in at least one sequence comprising a STAR sequence comprises, determining that the frequency of occurrence of said at least one sequence pattern significantly differs between said at least one STAR sequence and at least one control sequence. In principle any significant difference is discriminative for the presence of a STAR sequence. However, in a particularly preferred embodiment the frequency of occurrence of said at least one sequence pattern is significantly higher in said at least one sequence comprising a STAR sequence compared to said at least one control sequence.

A considerable number of sequences comprising a STAR sequence have been identified in the present invention. It is possible to use these sequences to test how efficient a pattern is in discriminating between a control sequence and a sequence comprising a STAR sequence. Using so-called discriminant analysis it is possible to determine on the basis of any set of STAR sequences in a species, the most optimal discriminative sequence patterns or combination thereof. Thus, preferably, at least one of said patterns is selected on the basis of a desired, and preferably an optimal discrimination between said at least one sequence comprising a STAR sequence and a control sequence. A desired

discrimination can be a certain significance factor associated with the pattern through bioinformatics.

In a preferred embodiment the frequency of occurrence of a sequence pattern in a test nucleic acid is compared with the frequency of occurrence in a sequence known to contain a STAR sequence. In this case a pattern is considered representative for a sequence comprising a STAR sequence if the frequencies of occurrence are similar. In a preferred embodiment another criterion is used. The frequency of occurrence of a pattern in a sequence comprising a STAR sequence is compared to the frequency of occurrence of said pattern in a control sequence. By comparing the two frequencies it is possible to determine for each pattern thus analysed, whether the frequency in the sequence comprising the STAR sequence is significantly different from the frequency in the control sequence. In this embodiment a sequence pattern is considered to be representative of a sequence comprising a STAR sequence, if the frequency of occurrence of the pattern in at least one sequence comprising a STAR sequence is significantly different from the frequency of occurrence of the same pattern in a control sequence. By using larger numbers of sequences comprising a STAR sequence the number of patterns for which a statistical difference can be established increases, thus enlarging the number of patterns for which the frequency of occurrence is representative for a sequence comprising a STAR sequence. Preferably said frequency of occurrence is representative of the frequency of occurrence of said at least one sequence pattern in at least 2 sequences comprising a STAR sequence, more preferably in at least 5 sequences comprising a STAR sequence. More preferably in at least 10 sequences comprising a STAR sequence. More preferably, said frequency of occurrence is representative of the frequency of occurrence of said at least one sequence pattern in at least 20 sequences comprising a STAR sequence. In a particularly preferred embodiment said frequency of occurrence is representative of the frequency of occurrence of said at least one sequence pattern in at least 50 sequences comprising a STAR sequence.

The patterns that are indicative for a sequence comprising a STAR sequence are also dependent on the type of control nucleic acid used. The type of control sequence used is preferably selected on the basis of the sequence in which the presence of a STAR sequence is to be detected. In a preferred embodiment said control sequence comprises a random sequence comprising a similar AT/CG content as said at least one sequence comprising a STAR sequence. In another preferred embodiment the control sequence is derived from the same species as said sequence comprising said STAR sequence. For instance, if a test sequence is scrutinized for the presence of a STAR sequence, active in a plant cell, then preferably the control sequence is also derived from a plant cell. Similarly, for testing for STAR activity in a human cell, the control nucleic acid is preferably also derived from a human genome. In a preferred embodiment the control sequence comprises between 50% and 150% of the bases of said at least one sequence comprising a STAR sequence. In a particularly preferred embodiment said control sequence comprises between 90% and 110% of the bases of said at least one sequence comprising a STAR sequence. More preferably, between 95% and 105%.

A pattern can comprise any number of bases larger than two. Preferably, at least one sequence pattern comprises at least 5, more preferably at least 6 bases. In another embodiment at least one sequence pattern comprises at least 8 bases. In a preferred embodiment said at least one sequence pattern comprises a pattern listed in table 9 and/or table 10. A pattern may consist of a consecutive list of bases. However, the pattern may also comprise bases that are interrupted one or more times by a number of bases that are not or only partly discriminative. A partly discriminative base is for instance indicated as a purine.

Preferably, the presence of STAR activity is verified using a functional assay. Several methods are presented herein to determine whether a sequence comprises STAR activity. STAR activity is confirmed if the sequence is capable of performing at least one of the following functions: (i) at least in part

inhibiting the effect of sequence comprising a gene transcription repressing element of the invention, (ii) at least in part blocking chromatin-associated repression, (iii) at least in part blocking activity of an enhancer, (iv) conferring upon an operably linked nucleic acid encoding a transcription unit compared to the same nucleic acid alone. (iv-a) a higher predictability of transcription, (iv-b) a higher transcription, and/or (iv-c) a higher stability of transcription over time.

The large number of sequences comprising STAR activity identified in the present invention opens up a wide variety of possibilities to generate and identify sequences comprising the same activity in kind not necessarily in amount. For instance, it is well within the reach of a skilled person to alter the sequences identified in the present invention and test the altered sequences for STAR activity. Such altered sequences are therefore also part of the present invention. Alteration can include deletion, insertion and mutation of one or more bases in the sequences.

Sequences comprising STAR activity were identified in stretches of 400 bases. However, it is expected that not all of these 400 bases are required to retain STAR activity. Methods to delimit the sequences that confer a certain property to a fragment of between 400 and 5000 bases are well known. The minimal sequence length of a fragment comprising STAR activity is estimated to be about 50 bases.

Table 9 and table 10 list patterns of 6 bases that have been found to be over represented in nucleic acid molecules comprising STAR activity. This over representation is considered to be representative for a STAR sequence. The tables were generated for a family of 65 STAR sequences. Similar tables can be generated starting from a different set of STAR sequences, or from a smaller or larger set of STAR sequences. A pattern is representative for a STAR sequence if it is over represented in said STAR sequence compared to a sequence not comprising a STAR element. This can be a random sequence. However, to exclude a non relevant bias, the sequence comprising a STAR sequence is

preferably compared to a genome or a significant part thereof. Preferably a genome of a vertebrate or plant, more preferably a human genome. A significant part of a genome is for instance a chromosome. Preferably the sequence comprising a STAR sequence and said control sequence are derived from nucleic acid of the same species.

The more STAR sequences are used for the determination of the frequency of occurrence of sequence patterns, the more representative for STARs the patterns are that are over- or under-represented. Considering that many of the functional features that can be expressed by nucleic acids, are mediated by proteinaceous molecules binding to it, it is preferred that the representative pattern is over-represented in the STAR sequences. Such over-represented pattern can be, part of, a binding site for such a proteinaceous molecule. Preferably said frequency of occurrence is representative of the frequency of occurrence of said at least one sequence pattern in at least 2 sequences comprising a STAR sequence, more preferably in at least 5 sequences comprising a STAR sequence. More preferably in at least 10 sequences comprising a STAR sequence. More preferably, said frequency of occurrence is representative of the frequency of occurrence of said at least one sequence pattern in at least 20 sequences comprising a STAR sequence. In a particularly preferred embodiment said frequency of occurrence is representative of the frequency of occurrence of said at least one sequence pattern in at least 50 sequences comprising a STAR. Preferably, said sequences comprising a STAR sequence comprises at least one of the sequences depicted in figure 26.

STAR activity is feature shared by the sequences listed in figure 26. However, this does not mean that they must all share the same identifier sequence. It is very well possible that different identifiers exist. Identifiers may confer this common feature onto a fragment containing it, though this is not necessarily so.

By using more sequences comprising STAR activity for determining the frequency of occurrence of a sequence pattern or patterns, it is possible to select patterns that are more often than others present or absent in such a STAR sequence. In this way it is possible to find patterns that are very frequently over or under represented in STAR sequences. Frequently over or under represented patterns are more likely to identify candidate STAR sequences in test sets. Another way of using a set of over or under represented patterns is to determine which pattern or combination of patterns is best suited to identify a STAR in a sequence. Using so-called discriminative statistics we have identified a set of patterns which performs best in identifying a sequence comprising a STAR element. In a preferred embodiment at least one of said sequence patterns for detecting a STAR sequence comprises a sequence pattern GGACCC, CCCTGC, AAGCCC, CCCCCA and/or AGCACC. In another embodiment at least one of said sequence patterns for detecting a STAR sequence comprises a sequence pattern CCCN{16}AGC, GGCN{9}GAC, CACN{13}AGG, CTGN{4}GCC.

A list of STAR sequences can also be used to determine one or more consensus sequences therein. The invention therefore also provides a consensus sequence for a STAR element. This consensus sequence can of course be used to identify candidate STAR elements in a test sequence.

Moreover, once a sequence comprising a STAR element has been identified in a vertebrate it can be used by means of sequence homology to identify sequences comprising a STAR element in other species belonging to vertebrate. Preferably a mammalian STAR sequence is used to screen for STAR sequences in other mammalian species. Similarly, once a STAR sequence has been identified in a plant species it can be used to screen for homologous sequences with similar function in other plant species. The invention in one aspect provides a STAR sequence obtainable by a method according to the invention. Further provided is a collection of STAR sequences. Preferably said STAR sequence is a vertebrate or plant STAR sequence. More preferably, said STAR

sequence is a mammalian STAR sequence or an angiosperm (monocot, such as rice or dicot such as Arabidopsis). More preferably, said STAR sequence is a primate and/or human STAR sequence.

A list of sequences comprising STAR activity can be used to determine whether a test sequence comprises a STAR element. There are, as mentioned above, many different methods for using such a list for this purpose. In a preferred embodiment the invention provides a method for determining whether a nucleic acid sequence of about 50-5000 base pairs comprises a STAR sequence said method comprising, generating a first table of sequence patterns comprising the frequency of occurrence of said patterns in a collection of STAR sequences of the invention, generating a second table of said patterns comprising the frequency of occurrence of said patterns in at least one reference sequence, selecting at least one pattern of which said frequency of occurrence differs between the two tables,
determining, within said nucleic acid sequence of about 50-5000 base pairs, the frequency of occurrence of at least one of said selected patterns, and determining whether the occurrence in said test nucleic acid is representative of the occurrence of said selected pattern in said collection of STAR sequences. Alternatively, said determining comprises determining whether the frequency of occurrence in said test nucleic acid is representative of the frequency of occurrence of said selected pattern in said collection of STAR sequences. Preferably said method further comprises determining whether said candidate STAR comprises a gene transcription modulating quality using a method of the invention. Preferably, said collection of STARS comprises sequence as depicted in figure 26.

In another aspect the invention provides an isolated and/or recombinant nucleic acid sequence comprising a STAR sequence obtainable by a method of the invention.

As mentioned above, a STAR sequence can exert its activity in a directional way, i.e. more to one side of the fragment containing it than to the other. Moreover, STAR activity can be amplified in amount by multiplying the number of STAR elements. The latter suggests that a STAR element may comprise one or more elements comprising STAR activity. Another way of identifying a sequence capable of conferring STAR activity on a fragment containing it comprises selecting from a vertebrate or plant sequence, a sequence comprising STAR activity and identifying whether sequences flanking the selected sequence are conserved in another species. Such conserved flanking sequences are likely to be functional sequences. In one aspect the invention therefore provides a method for identifying a sequence comprising a STAR element comprising selecting a sequence of about 50 to 5000 base pairs from a vertebrate or plant species comprising a STAR element and identifying whether sequences flanking said selected sequence in said species are conserved in at least one other species. The invention therefore further provides a method for detecting the presence of a STAR sequence within a nucleic acid sequence of about 50-5000 base pairs, comprising identifying a sequence comprising a STAR sequence in a part of a chromosome of a cell of a species and detecting significant homology between said sequence and a sequence of a chromosome of a different species. Preferably, said species comprises a plant or vertebrate species, preferably a mammalian species. The invention also provides a method for detecting the presence of a STAR element within a nucleic acid sequence of about 50-5000 base pairs of a vertebrate or plant species, comprising identifying whether a flanking sequence of said nucleic acid sequence is conserved in at least one other species.

It is important to note that methods of the invention for detecting the presence of a sequence comprising a STAR sequence using bioinformatical information are iterative in nature. The more sequences comprising a STAR

sequence are identified with a method of the invention the more patterns are found to be discriminative between a sequence comprising a STAR sequence and a control sequence. Using these newly found discriminative patterns more sequences comprising a STAR sequence can be identified which in turn enlarges the set of patterns that can discriminate and so on. This iterative aspect is an important aspect of methods provided in the present invention.

The term quality in relation to a sequence refers to an activity of said sequence. The term STAR, STAR sequence or STAR element, as used herein, refers to a DNA sequence comprising one or more of the mentioned gene transcription modulating qualities. The term SINC or SINC element as listed below refers to a DNA sequence comprising one or more of the mentioned gene transcription repressing qualities. The term "DNA sequence" as used herein does, unless otherwise specified, not refer to a listing of specific ordering of bases but rather to a physical piece of DNA. A transcription quality with reference to a DNA sequence refers to an effect that said DNA sequence has on transcription of a gene of interest. "Quality" as used herein refers to detectable properties or attributes of a nucleic acid or protein in a transcription system.

Examples

Example 1. Methods to isolate STAR and SINC elements

Materials and methods

Plasmids and strains. The selection vector for STAR elements, pSelect-SV40-zeo ("pSelect", Figure 1) is constructed as follows: the pREP4 vector (Invitrogen V004-50) is used as the plasmid backbone. It provides the Epstein Barr oriP origin of replication and EBNA-1 nuclear antigen for high-copy episomal replication in primate cell lines; the hygromycin resistance gene with the thymidine kinase promoter and polyadenylation site, for selection in mammalian cells; and the ampicillin resistance gene and colE1 origin of

replication for maintenance in *Escherichia coli*. The vector contains four consecutive LexA operator sites between *Xba*I and *Nhe*I restriction sites (Bunker and Kingston, 1994). Embedded between the LexA operators and the *Nhe*I site is a polylinker consisting of the following restriction sites: *Hin*dIII-*Asc*I-*Bam*HI-*Asc*I-*Hin*dIII. Between the *Nhe*I site and a *Sal*I site is the zeocin resistance gene with the SV40 promoter and polyadenylation site, derived from pSV40/Zeo (Invitrogen V502-20); this is the selectable marker for the STAR screen.

The pSDH vector (Figure 2) is constructed as follows: The luciferase reporter gene from pGL3-Control (Promega E1741) is amplified by PCR and inserted into *Sac*II/*Bam*HI-digested pUHD10-3 (Gossen and Bujard, 1992). This places luciferase under control of the Tet-Off promoter, and upstream of the SV40 polyadenylation signal. Multiple cloning sites are introduced by PCR, upstream of the Tet-Off promoter (MCSI, *Xho*I-*Not*I-*Eco*RI-*Sal*I) and downstream of the polyadenylation signal (MCSII, *Nhe*I-*Bgl*II-*Eco*RV-*Hin*dIII). Gene libraries are constructed by *Sau*3AI digestion of human genomic DNA, either purified from placenta (Clontech 6550-1) or carried in bacterial/P1 (BAC/PAC) artificial chromosomes. The BAC/PAC clones contain genomic DNA from the 1q12 cytogenetic region (clones RP1154H19 and RP3328E19) or from the HOX cluster of homeotic genes (clones RP1167F23, RP1170019, and RP11387A1). The DNAs are size-fractionated, and the 0.5 - 2 kb size fraction is ligated into *Bam*HI-digested pSelect vector, by standard techniques (Sambrook et al., 1989).

The construction of the host strains has been described (van der Vlag et al., 2000). Briefly, they are based on the U-2 OS human osteosarcoma cell line (American Type Culture Collection HTB-96). U-2 OS is stably transfected with the pTet-Off plasmid (Clontech K1620-A), encoding a protein chimera consisting of the Tet-repressor DNA binding domain and the VP16 transactivation domain. The cell line is subsequently stably transfected with fusion protein genes containing the LexA DNA binding domain, and the coding

regions of either HP1 or HPC2 (two Drosophila Polycomb group proteins that repress gene expression when tethered to DNA). The LexA-repressor genes are under control of the Tet-Off transcriptional regulatory system (Gossen and Bujard, 1992).

Library screening and STAR element characterization. The gene libraries in pSelect are transfected into the U-2 OS/Tet-Off/LexA-repressor cell line by calcium phosphate precipitation (Graham and van der Eb, 1973; Wigler et al., 1978) as recommended by the supplier of the transfection reagent (Life Technologies). Transfected cells are cultured under hygromycin selection (25 μg/ml) and tetracycline repression (doxycycline, 10 ng/ml) for 1 week (50% confluence). Then the doxycycline concentration is reduced to 0.1 ng/ml to induce the LexA-repressor genes, and after 2 days zeocin is added to 250 μg/ml. The cells are cultured for a further 4-5 weeks, until the control cultures (transfected with empty pSelect) are killed by the zeocin.

Zeocin-resistant colonies from the library transfection are propagated, and plasmid DNA is isolated and rescued into *E. coli* by standard techniques (Sambrook et al., 1989). The candidate STAR elements in the rescued DNA are analyzed by restriction endonuclease mapping (Sambrook et al., 1989), DNA sequence analysis (Sanger et al., 1977), and for STAR activity (zeocin resistance) after re-transfection to U-2 OS/Tet-Off/LexA-repressor and lowering the doxycycline concentration.

Candidate STAR elements that have DNA sequence corresponding to known sequence in the human genome are identified by BLAST searches (Altschul et al., 1990) of the human genome database (http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html 20 June 2001). The chromosomal locations of the elements are recorded, along with the proportion of repetitive DNA and the identity of adjacent genes.

Those candidates that show STAR activity upon re-transfection are characterized further by subcloning the STAR fragment into the pSDH plasmid and stable integration in U-2 OS chromosomal DNA. pSDH plasmids

are co-transfected into U-2 OS cells with pBABE-puro (Morgenstern and Land, 1990), and selected for puromycin resistance. Per STAR element, populations of approximately 30 individual clones are isolated and cultured. The clones are periodically assayed for luciferase activity according to the manufacturer's instructions (Roche 1669893).

Results

STAR element functional characterization. The screens of human genomic DNA and of the HOX and 1q12 loci yielded 17 bona fide STAR elements. The criteria are that (1) the elements displayed STAR activity upon re-transfection of the pSelect-based clones into the host U-2 OS human osteosarcoma cell line (indicating that the anti-repressor activity expressed in the initial screen is plasmid-specific and not due to artefactual changes in the host cells); (2) the elements contain DNA sequence that matches sequence in the human genome sequence database (indicating that the clone does not contain contaminating DNA sequence, from e.g. bacterial or vector sources).

The STAR elements are sub-cloned into the pSDH plasmid and integrated into the host cell genome. Expression of the reporter genes is assayed in populations of stable transfectants to demonstrate the ability of the STAR elements to protect reporter genes from silencing after integration at random into the genome. This provides information (1) on the proportion of clones which display high expression, and (2) on the degree of over-expression elicited by the STAR elements.

Expression of the luciferase reporter gene by a clone is considered significant if it is two-fold above the average level for the plasmids containing no STAR elements (the reference level). For all plasmids a distribution in expression level is observed among the clones: from no expression to expression significantly above the reference level, and from few over-expressers to many over-expressers. Superior STAR activity is manifested by plasmids that result in many over-expressing clones, including some highly over-expressing clones.

Results from a representative experiment are shown in Table 1, and in Figures 3 - 5:

The results indicate that the human STAR elements which are tested yield a much higher proportion of over-expressing clones than the unprotected reporter gene, or the reporter gene protected by the *Drosophila* SCS element (Kellum and Schedl, 1992). Furthermore, the degree of over-expression by these plasmids is much greater from the STAR-protected reporter gene than the unprotected or SCS-protected reporter.

STAR element sequence and genomic position data. Table 2 lists the chromosomal locations of each of the 17 STAR elements, as well as the identity of nearby genes and the repetitive DNA content of the elements. The STAR elements are distributed over a number of chromosomes. They are diverse in their actual DNA sequence and repetitive DNA content, and display various degrees of association with neighboring genes.

SINC element screen

Materials and methods

The plasmid for the SINC screen, pSINC-Select ("pSS", Figure 6) is constructed as follows: the pREP4 vector (Invitrogen V004-50) is used as the plasmid backbone. It provides the Epstein Barr oriP origin of replication and EBNA-1 nuclear antigen for high-copy episomal replication in primate cell lines; the hygromycin resistance gene with the thymidine kinase promoter and polyadenylation site, for selection in mammalian cells; and the ampicillin resistance gene and colE1 origin of replication for maintenance in *Escherichia coli*. The vector contains a Tet-Off promoter, consisting of tandem Tet Responsive Elements (TREs) from plasmid pUDH10-3 (Gossen and Bujard, 1992), for regulation by the Tet-Off transcriptional regulatory system. The TREs regulate expression of the codA::upp gene encoding a fusion protein (cyotsine deaminase/uracil phosphoribosyltransferase; Invivogen porf-codaupp). This is a so-called "suicide gene"; the activity of the codA::upp enzyme converts a pro-drug 5-fluorocytosine (5-FC) to a toxic drug, 5-

fluorouracil (5-FU), thereby causing apoptosis and cell death (Mullen et al., 1992; Tiraby et al., 1998; Wei and Huber, 1996). Upstream from the Tet-Off promoter is a *Bgl*II restriction site for cloning *Sau*3AI-digested genomic DNA for screening. The pREP4 DNA is separated from the genomic DNA and suicide gene by STAR elements in order to prevent silencing of essential plasmid elements in the pREP4 component by cloned SINC elements.

Genomic DNA from a library of BAC clones comprising human chromosome 22 (Invitrogen/Research Genetics 96010-22) is partially digested with *Sau*3AI and ligated into *Bgl*II-digested pSS (Sambrook et al., 1989). The library of recombinant plasmids is transfected into the U-2 OS/Tet-Off cell line by calcium phosphate precipitation (Graham and van der Eb, 1973; Wigler et al., 1978) as recommended by the supplier of the transfection reagent (Life Technologies). Transfected cells are cultured under hygromycin selection (25 μg/ml) and tetracycline repression (doxycycline, 10 ng/ml) for 3 weeks. Then 5-FC is added to a concentration of 1 μg/ml and the cells are cultured for a further 3 weeks to select for SINC elements.

Candidate SINC-containing colonies are harvested and used in a polymerase chain reaction with primers PCR1 and PCR2 (Figure 6); the PCR products are digested with *Hin*dIII and *Xho*I restriction endonucleases and cloned into pBluescript II SK(+) (Stratagene 212207) by conventional techniques (Sambrook et al., 1989). The DNA sequences of the candidate SINC elements are determined (Sanger et al., 1977), and corresponding sequences in the human genome are identified by BLAST searches (Altschul et al., 1990) of the human genome database (http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html 20 June 2001). The chromosomal locations of the elements are recorded, along with the proportion of repetitive DNA and the identity of adjacent genes.

Results

At the end of the selection period no colonies are evident in the control cultures (empty pSS), and a number of colonies are evident in the cultures containing

pSS with genomic DNA. These surviving clones contain candidate SINC elements. The elements are recovered by PCR and subcloned into a standard cloning vector, pBluescript. The DNA sequences of the elements are determined, and compared with the human genome sequence (Table 3). In all cases, the sequenced elements are found on chromosome 22, as expected.

Example 2. Expression characteristics of the transgene that are due to the STAR, the SINC, or the combined STAR/SINC.

Background: site-specific recombination is used to precisely remove heterologous DNAs from their chromosomal locations. This is routinely carried out by one of two systems: the cre recombinase and loxP target of bacteriophage P1 (Feng et al., 1999), or the FLP recombinase and FRT (FLP recombinase target) of yeast (Wigley et al., 1994). In these systems, a DNA region (usually containing a reporter gene and/or a selectable marker) is flanked in the chromosome by the loxP or FRT target. The activity of the recombinase then catalyzes the precise excision of the DNA region from the chromosome. The recombinase resolves its two recognition sequences to a single site, deleting the sequence between them. Thus, a span of DNA must be flanked by target sites to be subsequently deleted in vivo upon introduction or activation of recombinase (Schwenk et al., 1995; Dymecki, 1996). The Cre and Flp recombinases catalyze recombination between two 13-base-pair inverted repeats, separated by a spacer with a minimum of 6 (loxP) or 8 (FRT) base pairs (Senecoff et al., 1985). The loxP sequence is ATAACTTCGTATA and the FRT sequence is GAAGTTCCTATAC.

Protocol: Using conventional DNA cloning (Sambrook et al., 1989), a reporter gene (encoding a reporter protein, for example green fluorescent protein (GFP) (Bierhuizen et al., 1997) or luciferase (Himes and Shannon, 2000) is constructed that is flanked in a plasmid by a pair of STAR elements, by a pair

of SINC elements, or by a pair of STAR/SINC combination elements. In each case, the elements are themselves flanked by recombinase target sites. One element is flanked by a pair of loxP sites, and the other is flanked by a pair of FRT sites (Figure 1). Upon transfection the plasmid integrates into the host chromosome in a small percentage of cells, and the integrants are selected by antibiotic resistance. Similar constructs are made for each of the three test elements (STAR, SINC, STAR/SINC).

Using conventional techniques, ("SuperFect Transfection Reagent Handbook," Qiagen, November, 1997) these plasmids are transfected into the U-2 OS human osteosarcoma cell line, and selected for hygromycin resistance. Hygromycin-resistant isolates have the plasmid stably integrated in the genome of the cell line. Individual isolates are propagated in cell culture medium, and the expression of the transgenic reporter gene is assayed, for example by flow cytometry (Stull et al., 2000).

Then using conventional techniques (transfection, or hormone stimulation), the stable isolates from above are treated so as to introduce or activate recombinase activity. This is done sequentially, such that for example the cre recombinase activity catalyzes excision of STAR1, and subsequently FLP recombinase activity catalyzes excision of STAR2. The level of expression of the reporter gene in these cells is assayed and the value compared with the reference value of the parental, STAR-containing isolate.

Example 3. Sequence analysis of STARs; determination of minimal essential sequence for element function; sequence conservation among elements; and properties of tandem and multiple elements

Background: DNA fragments containing STAR or SINC elements are isolated by genetic selection using the pSelect (Figure 1) or pSS (Figure 6) plasmids, respectively. This section describes the approach to characterize the DNA sequences within those fragments that have STAR or SINC activity.

Protocols:

DNA sequence: Oligonucleotides are designed based on the sequence of the pSelect and pSS selection plasmids for sequencing the DNA fragments. The fragments are sequenced using the dideoxy chain termination technique (Sanger et al., 1977). DNA squences are then localized to chromosome position using the public human genome sequence database (http://www.ncbi.nlm.nih.gov:80/cgi-bin/Entrez/hum_srch?chr=hum_chr.inf&query). Genes and gene density in the vicinity of the fragment sequence are recorded from the genome sequence annotation. Transcriptional activity of those genes is determined from public databases of DNA microarray (http://arrays.rockefeller.edu/xenopus/links.html) and SAGE (Serial Analysis of Gene Expression; http://bioinfo.amc.uva.nl/HTM-bin/index.cgi) data. Once positional information on STAR and SINC sequences is compiled, the data are analysed in terms of underlying consensus sequences. Consensus sequences or trends (understood by this are local areas rich in particular nucleotide combinations, e.g. rich in C and G bases) are detected using similarity search algorithms such as clustalw (Higgins et al., 1996) and blosum similarity scoring (Altschul and Gish, 1996). Any underlying consensuses or trends found are then used to identify other potential STARs on a genome scale by performing BLAST searches (Altschul et al., 1990). Previous research has identified transcriptional regulatory proteins that bind to known insulators and boundary elements (Gaszner et al., 1999; Gerasimova and Corces, 1998). In the described examples, the protein binding sites coincide with DNase I hypersensitive sites which are essential for insulator or boundary function. The hypothesis that STAR elements are also bound by known regulatory proteins is examined by searching the TRANSFAC database of transcription factors (http://transfac.gbf.de/TRANSFAC/) for sequence motifs that occur in the STAR elements. Sequence motifs that are common among the

members of the STAR or SINC collections are indicators that the corresponding transcription factor binds to that element.

Minimal essential sequence: Using this sequence knowledge STAR (or SINC) elements are truncated and tested for functionality. This is done using the polymerase chain reaction (PCR) to clone sub-fragments of the STAR- or SINC-containing fragments into pSelect or pSS by standard techniques (Sambrook et al., 1989). The plasmids containing the sub-fragments are transfected into U-2 OS cells and tested for functionality by assaying for antibiotic resistance (STAR elements) or pro-drug resistance (SINC elements).

Directionality: The STAR and SINC elements are tested for their directionality using the pSelect and pSS plasmids, respectively. For example, the direction of STAR elements isolated by the pSelect screen is referred to as 5'3' orientation. The orientation of the element is reversed by conventional recombinant DNA techniques (Sambrook et al., 1989). The resulting plasmids are transfected into the U-2 OS cell line and expression of the reporter gene is assayed (Bierhuizen et al., 1997; Himes and Shannon, 2000). The level of expression from the plasmid with the reverse-orientation element is compared to that with the 5'3' orientation. If the reverse-orientation plasmid has similar expression levels, then the STAR element does not display directionality.

Combinations and multiples of elements: To determine whether STAR elements are able to function in mixed pairs, different elements are combined and tested. The analysis is performed in the pSDH plasmid by inserting one STAR element in MCSI and a different STAR in MCSII by recombinant DNA techniques (Sambrook et al., 1989). The resulting plasmids are transfected, and the expression of the reporter gene is assayed (Bierhuizen et al., 1997; Himes and Shannon, 2000); the results are compared with the expression from plasmids containing the same element at MCSI and MCSII; if the expression

is similar for the two types of plasmids, then it is concluded that different STAR elements do not interfere with each other.

The strength of single STAR or SINC elements is compared with tandem repeats of elements. This is done by concatamerization of the STAR or SINC elements of interest with DNA ligase and insertion of the ligation product into the pSDH or pSS plasmids by recombinant DNA techniques (Sambrook et al., 1989). The resulting plasmids are transfected into U-2 OS cells, and the expression of the reporter gene is assayed (Bierhuizen et al., 1997; Himes and Shannon, 2000); the results are compared with the expression from plasmids containing single STAR or SINC elements.

Example 4. Determination of the distance over which a STAR, a SINC, or a combination thereof functions.

Background: STAR elements are used to optimize expression of single and multiple transgenes. To determine if a single pair of STAR elements can protect large or multiple transgenes from silencing it is necessary to determine the range over which STAR elements act. Similar information is determined for SINC elements and STAR/SINC combinations.

Protocol: STAR and SINC elements are tested for their functionality over distance using derivative plasmids based on pSelect or pSS respectively, as follows. A library of random DNA fragments from 500bp to 10kb is assembled by standard DNA cloning techniques (Sambrook et al., 1989). Fragments are selected from this library that do not possess STAR or SINC activity, by testing in the pSelect and pSS plasmids as described above. For STAR elements and STAR/SINC combinations, these fragments are inserted between the cloning site and the promoter of the reporter gene in the appropriate pSelect plasmid (Figure 1). This set of plasmids is transfected into the U-2 OS cell line, and expression measured as described above. The strength of reporter

gene expression is correlated with the length of the random DNA fragment separating the STAR element from the promoter. SINC elements are assessed in an analogous fashion: random DNA fragments are inserted between the SINC element and the promoter of the appropriate pSS plasmid, and the degree of repression of the reporter gene is correlated with the length of the random DNA fragment.

Example 5. (a) Use of a naturally occurring SINC element in the genetic selection for STAR elements.

Background: The current screens for STAR elements use chimeric lexA-PcG proteins to provide repression of the selectable marker in the selection plasmid. By repeating the selection using naturally occurring SINC elements, STAR elements are identified that are specific to the repressive activity due to these naturally occurring SINC elements.

The SINC element screen is based on the ability of genetic selection to identify randomly generated fragments of genomic DNA that are able to silence a "tet-off" promoter and block the expression of the codA::upp suicide gene. The SINC elements recovered from this selection represent a random sampling of genomic silencing elements, and different classes of elements are recovered. For this protocol, these diverse SINC elements are used to recover different classes of STAR elements than those recovered in the aforementioned lexA-PcG based selections.

Protocol: SINC elements from the current selection are characterized and sorted into classes on the basis of of functional and DNA sequence features (functional features include strength of repression; sequence features include identifiable conserved motifs; see example 3). Representative elements from each class are used to replace the lexA binding sites in the pSelect plasmid via standard DNA cloning techniques (Sambrook et al., 1989). A gene bank is

made with each of these new plasmids, and used to identify new, SINC-specific STAR elements as described (van der Vlag et al., 2000). This is done with whole genomic DNA, and with DNA from the BAC clone that also contains the SINC element being used.

Example 5 (b) Determination of the maximal length of STAR and SINC elements

Background: STAR elements are cloned as fragments of DNA recovered using the pSelect plasmid, which is done with genomic DNA fragments less than 2 kb. However, these might be portions of a more extended STAR element. Extended STAR activity is examined by the following experiments.
Protocol: STAR elements cloned in pSelect are mapped to the human genome sequence. In order to determine if they are portions of a more extended STAR element, regions of 4 kb that encompass the clones are amplified by PCR and cloned into the pSelect and/or pSDH plasmid by standard recombinant DNA techniques (Sambrook et al., 1989). The resulting plasmids are transfected into U-2 OS cells and assayed for reporter gene expression as described above; plasmids containing the original 2 kb STAR element are included as a control. Three possible results can be expected: (1) similar expression by the control and extended STAR isolates, demonstrating that the STAR element is confined to the original 2 kb fragment; (2) lower expression by the extended STAR isolates, suggesting that the STAR element is contained within the 2 kb fragment and does not act effectively over a distance or that the extended fragment contains a SINC element; (3) higher expression by the extended STAR isolates, suggesting that the extended region contains a more complete STAR element. In the case of result (3), the exercise is reiterated with a larger PCR fragment of 6 kb.
A STAR element may also be a composite of sites to which various proteins bind. Therefore large DNA fragments with STAR activity could be divisible

into smaller fragments with STAR activity (see example 3). Elements that are greater than 2 kb are recognized as STAR elements if they still display STAR activity after truncation to less than 2 kb (including by internal deletion).

Example 6. Methylation and histone acetylation states of STAR elements, SINC elements, or combinations thereof and of the adjacent transgenes.

Background: The regulatory properties of STAR and SINC elements are associated with the local chromatin structure, which is determined by the DNA itself and by DNA-associated proteins. Changes in chromatin structure that are associated with changes in gene expression are often produced by secondary modifications of the macromolecules, especially methylation of DNA or acetylation of histone proteins. Identifying the secondary modifications occurring at STAR and SINC elements and at adjacent transgenes provides hallmarks for these elements.

Protocol: DNA methylation: STAR or SINC elements or combinations thereof are cloned into the pSelect plasmid by standard techniques (Sambrook et al., 1989). U-2 OS cells are stably transfected with these plasmids, and with pSelect lacking a STAR or SINC element as a control to determine basal DNA methylation at the reporter gene. Cells are harvested and the chromatin purified by standard procedures (Thomas, 1998). The DNA is digested with the *Hpa*II and *Msp*I restriction endonucleases in separate reactions (Sambrook et al., 1989). Both of these restriction enzymes are able to cut the non-methylated sequence CCGG. When the external C is methylated, both *Msp*I and *Hpa*II cannot cleave. However, unlike *Hpa*II, *Msp*I can cleave the sequence when the internal C is methylated. The DNA is subjected to Southern blotting and the blot is analyzed by indirect end-labeling (Pazin and Kadonaga, 1998). As a control, the corresponding pSelect plasmid as naked, unmethylated DNA, is also cut with the described enzymes and subjected to Southern blotting.

Comparison of the different sizes of the DNA fragments reveals whether the DNA is methylated in vivo or not.

Histone acetylation: The same transfected cell lines used for DNA methylation analysis are used for these experiments. The method described below yields a high resolution map of the histone acetylation pattern on the STAR and SINC elements and the reporter gene (Litt et al., 2001). Micrococcal nuclease digests of nuclei are fractionated on sucrose gradients, and purified nucleosome monomers and dimers are enriched for acetylated histones by immunoprecipitation with anti-acetylhistone antibodies. The nucleosome fraction and immunoprecipitates are subjected to analysis, for example by real-time PCR (Jung et al., 2000) using primers and a Taqman probe that anneal to the reporter gene or to the STAR or SINC element to yield 0.2 kb products, with a moving window of 0.1 kb. The rate of increase of the Taqman probe fluorescent signal during the PCR (which is proportional to the abundance of the template DNA in the sample) is then measured. The ratio of the abundance of the template DNA in the nucleosome fraction and the immunoprecipitates provides a fine-map of the pattern of histone acetylation for each 0.1 kb on the reporter gene and STAR or SINC element (or on the reporter gene in the absence of an element).

Example 7. In vivo Nucleosome positioning and DNAse I hypersensitive sites

Background: Chromatin is comprised of DNA, histones, and non-histone proteins. The histones form a core particle that is wrapped by ~150 bp of DNA to make a nucleosome. Nucleosomes are separated by 50-75 bp of linker DNA. Stably positioned nucleosomes on chromosomal DNA repress gene expression, and factors that exclude nucleosomes or otherwise remodel chromatin can overcome this repression. The positioning of nucleosomes in a chromosomal region is analyzed by micrococcal nuclease (MNase) assay; MNase cuts chromatin preferentially in the linker DNA. Similarly, some areas of DNA are

constitutively exposed to non-histone proteins, and these are frequently regulatory regions, i.e. sites where cis-acting regulatory factors bind. Experimentally, these site are hypersensitive to digestion by the enzyme DNase I.

Protocol: To determine the position of nucleosomes on the reporter gene and on either the STAR or SINC elements, MNase is used (Saluz and Jost, 1993). Nuclei are purified from cultured U-2 OS cells and digested with MNase as described above (histone acetylation). To search for DNase I hypersensitive sites in the STAR and SINC elements or the reporter gene, purified nuclei are treated with DNase I at an appropriate concentration (e.g. 100 μg/ml genomic DNA and 20-100 U/ml DNaseI), as described (Wallrath et al., 1998). Naked DNA is digested with DNase I as a control. For both techniques, the reporter gene and STAR or SINC elements are fine-mapped using primer extension or indirect end-labelling and Southern blotting, as described (Tanaka et al., 1996; van der Vlag et al., 2000). The MNase assay reveals a ladder of discrete bands on an autoradiogram corresponding to the positions of nucleosomes on the STAR or SINC elements or the reporter gene. DNase I hypersensitive sites are manifested as discrete bands in the resulting autoradiogram that are absent or less prominent in the naked DNA control.

Example 8. Cell-type, tissue dependence, and promoter dependence of STAR and SINC elements.

Background: It has been reported that some insulators or boundary elements may display tissue specificity (Takada et al., 2000). STAR elements have many features in common with insulators and boundary elements. Both promiscuous and tissue-specific STAR and SINC elements have biotechnological value in transgenic applications. The assay described below is performed to assess cell-type dependence. Cell and tissue specificity of the elements are examined

further by examining the expression of genes in the vicinity of the elements in the human genome, using public databases of DNA microarray (http://arrays.rockefeller.edu/xenopus/links.html) and SAGE (Serial Analysis of Gene Expression; http://bioinfo.amc.uva.nl/HTM-bin/index.cgi) data.

Protocol: STAR elements are tested in the pSDH plasmid, and SINC elements in the pSS plasmid. Three cell lines are transfected using standard protocols: the human U-2 OS osteosarcoma cell line (Heldin et al., 1986), the Vero cell line from African green monkey kidney (Simizu et al., 1967), and the CHO cell line from Chinese hamster ovary (Kao and Puck, 1968). Elements able to function in all three cell types are catagorized as promiscuous. Those only displaying activity in one or two of the cell-lines are catagorized as restricted in their cell-type functionality.

Promoter specficity: STAR and SINC elements are currently selected and tested in the context of function with two promoters, the entire cytomegalovirus (CMV) promoter or the Tetracycline Response Element and minimal CMV promoter (in combination with the tTA transcriptional activator). To assess promoter specificity, STAR and SINC function are tested with other commonly used viral promoters, namely the simian virus type 40 (SV40) early and late promoters, the adenoviral E1A and major late promoters, and the Rous sarcoma virus (RSV) long terminal repeat (Doll et al., 1996; Smith et al., 2000; Weaver and Kadan, 2000; Xu et al., 1995). Each of these promoters are cloned separately into the pSelect and pSS plasmids by standard techniques (Sambrook et al., 1989) along with STAR or SINC elements, respectively. The resulting plasmids are transfected into the U-2 OS cell line and assayed for reporter gene expression, as described above. The ability of SINC elements to silence these promoters, or STAR elements to protect against silencing, is determined by comparison with plasmids lacking STAR or SINC elements.

Example 9. Methods for improvement of STAR and SINC elements

Background: Improved STAR and SINC elements are developed. Improvements yield increased strength of anti-repressive or repressive activity, and elements with inducible and tissue-specific activity. These improvements are made by a combination of techniques.

Protocols

Forced evolution: Error prone PCR (Cherry et al., 1999; Henke and Bornscheuer, 1999) is used to introduce an average of one to two point mutations per element. The mutagenized elements are screened using pSelect (or pSS) plasmids containing reporter-selectable marker fusion proteins by for example fluorescence activated cell sorting and antibiotic resistance (Bennett et al., 1998). Subsequent rounds of error prone PCR and selection are carried out to derive elements with further improvements in activity.

Tandem and heterologous combinations: As described above, tandem and heterologous combinations of elements are tested for activity in comparison with the single elements (example 3).

The relative dominance of STAR and SINC elements is tested on a case by case basis. It is used to test the strength of an element; for example, if a new STAR element is dominant to a known, strong SINC element, then the STAR is classified as very strong. The possibility that the dominance relationship between a STAR and a SINC is cell type-, tissue-, or promoter-specific is also considered (example 8). The dominance test utilizes the pSelect plasmid, with individual SINC elements placed upstream of individual STAR elements by standard recombinant DNA techniques (Sambrook et al., 1989). The plasmids are transfected to U-2 OS cells and reporter gene expression is assayed. SINC dominance is manifested by lower expression than the plasmid with only a STAR element, while STAR dominance is manifested by higher expression than the plasmid with only a SINC element.

Introduction of binding sites for other DNA-binding proteins to STAR and SINC elements to add novel characteristics (e.g. inducibility, tissue specificity)
Background: Regulatable STAR and SINC elements are created by combining them with binding sites for signal-dependent DNA binding proteins. In one example this would involve juxtaposition of a STAR or SINC or STAR/SINC combination and a glucocorticoid response element (GRE). In the absence of glucocorticoid stimulation the STAR or SINC element would function as described. Upon stimulation, the naturally occurring glucocorticoid receptor binds to the GRE and interferes with STAR or SINC function.
Protocol: Using conventional DNA cloning (Sambrook et al., 1989), a GRE is introduced into the pSelect or pSS vector adjacent to STAR or SINC elements, respectively. The plasmid is transfected into U-2 OS cells as described above. Cells are divided into two cultures; one is treated with glucocorticoid (10 μM). Expression of the reporter gene is measured and compared between the two cultures. Differences in expression demonstrate the ability to regulate STAR and SINC function by action of a signal-dependent DNA-binding protein.
Promiscuous STAR and SINC elements: Testing or enhancing these characteristics involves cultivation in different cell lines, and long term cultivation without antibiotic selection (examples 8 and 10).

Example 10. STAR and SINC elements obviate the need for continuous selection for maintenance of the transgene.
Background: In transgenesis, reliance on selection markers has two drawbacks: the selection agent is usually expensive and carries a metabolic cost to the cells, and there are regulatory and ethical objections to including selectable markers in transgenic applications, especially if the transgene itself is in the product (e.g. crop plants, gene therapy vectors). STAR and SINC elements reduce or eliminate the need to maintain selection after establishing the transgenic isolate. Consequently, the resistance gene can be removed from

the transgenic genome by site-specific recombination with diminished loss of transgene expression.

Protocol: Stably transfected U-2 OS cell lines containing chromosomally-integrated STAR elements flanking reporter genes are produced by co-transfection of the pSDH plasmid with a trans-acting antibiotic resistance plasmid as described above. The experiment involves testing the stability of the reporter gene expression level in these cell lines during prolonged (3-6 month) cultivation in the absence of selection. This is tested with STAR elements flanking the luciferase or GFP reporter genes in pSDH plasmids. The antibiotic resistance gene is removed by constructing an expression plasmid (based on pSDH) in which the antibiotic selection marker is flanked by recombinase target sites. The selectable marker is subsequently excised by recombinase activity, as described above (example 2).

Example 11

Predictability and yield are improved by application of STAR elements in expression systems

STAR elements function to block the effect of transcriptional repression influences on transgene expression units. These repression influences can be due to heterochromatin ("position effects", (Boivin & Dura, 1998)) or to adjacent copies of the transgene ("repeat-induced gene silencing", (Garrick et al., 1998)). Two of the benefits of STAR elements for heterologous protein production are increased predictability of finding high-expressing primary recombinant host cells, and increased yield during production cycles. These benefits are illustrated in this example.

Materials and Methods

Construction of the pSDH vectors and STAR-containing derivatives: The pSDH-Tet vector was constructed by polymerase chain reaction amplification

(PCR) of the luciferase open reading frame from plasmid pREP4-HSF-Luc (van der Vlag et al., 2000) using primers C67 and C68 (all PCR primers and mutagenic oligonucleotides are listed in Table 5), and insertion of the SacII/BamHI fragment into SacII/BamHI-digested pUHD10-3 (Gossen & Bujard, 1992). The luciferase expression unit was re-amplified with primers C65 and C66, and re-inserted into pUHD10-3 in order to flank it with two multiple cloning sites (MCSI and MCSII). An AscI site was then introduced into MCSI by digestion with EcoRI and insertion of a linker (comprised of annealed oligonucleotides D93 and D94). The CMV promoter was amplified from plasmid pCMV-Bsd (Invitrogen K510-01) with primers D90 and D91, and used to replace the Tet-Off promoter in pSDH-Tet by SalI/SacII digestion and ligation to create vector pSDH-CMV. The luciferase open reading frame in this vector was replaced by SEAP (Secreted Alkaline Phosphatase) as follows: vector pSDH-CMV was digested with SacII and BamHI and made blunt; the SEAP open reading frame was isolated from pSEAP-basic (Clontech 6037-1) by EcoRI/SalI digestion, made blunt and ligated into pSDH-CMV to create vector pSDH-CS. The puromycin resistance gene under control of the SV40 promoter was isolated from plasmid pBabe-Puro (Morgenstern & Land, 1990) by PCR, using primers C81 and C82. This was ligated into vector pGL3-control (BamHI site removed) (Promega E1741) digested with NcoI/XbaI, to create pGL3-puro. pGL3-puro was digested with BglII/SalI to isolate the SV40-puro resistance gene, which was made blunt and ligated into NheI digested, blunt-ended pSDH-CS. The resulting vector, pSDH-CSP, is shown in FIG 7. All cloning steps were carried out following the instructions provided by the manufacturers of the reagents, according to methods known in the art (Sambrook et al., 1989).

STAR elements were inserted into MCSI and MCSII in two steps, by digestion of the STAR element and the pSDH-CSP vector with an appropriate restriction enzyme, followed by ligation. The orientation of STAR elements in recombinant pSDH vectors was determined by restriction mapping. The

identity and orientation of the inserts were verified by DNA sequence analysis. Sequencing was performed by the dideoxy method (Sanger et al., 1977) using a Beckman CEQ2000 automated DNA sequencer, according to the manufacturer's instructions. Briefly, DNA was purified from E. coli using QIAprep Spin Miniprep and Plasmid Midi Kits (QIAGEN 27106 and 12145, respectively). Cycle sequencing was carried out using custom oligonucleotides C85, E25, and E42 (Table 5), in the presence of dye terminators (CEQ Dye Terminator Cycle Sequencing Kit, Beckman 608000).

Transfection and culture of CHO cells with pSDH plasmids: The Chinese Hamster Ovary cell line CHO-K1 (ATCC CCL-61) was cultured in HAMS-F12 medium + 10% Fetal Calf Serum containing 2 mM glutamine, 100 U/ml penicillin, and 100 micrograms/ml streptomcyin at 37 C/5% CO2. Cells were transfected with the pSDH-CSP vector, and its derivatives containing STAR6 or STAR49 in MCSI and MCSII, using SuperFect (QIAGEN) as described by the manufacturer. Briefly, cells were seeded to culture vessels and grown overnight to 70-90% confluence. SuperFect reagent was combined with plasmid DNA (linearized in this example by digestion with PvuI) at a ratio of 6 microliters per microgram (e.g. for a 10 cm Petri dish, 20 micrograms DNA and 120 microliters SuperFect) and added to the cells. After overnight incubation the transfection mixture was replaced with fresh medium, and the transfected cells were incubated further. After overnight cultivation, 5 micrograms/ml puromycin was added. Puromycin selection was complete in 2 weeks, after which time individual puromycin resistant CHO/pSDH-CSP clones were isolated at random and cultured further.

Secreted Alkaline Phosphatase (SEAP) assay: SEAP activity (Berger et al., 1988, Henthorn et al., 1988, Kain, 1997, Yang et al., 1997) in the culture medium of CHO/pSDH-CSP clones was determined as described by the manufacturer (Clontech Great EscAPe kit #K2041). Briefly, an aliquot of

medium was heat inactivated at 65 C, then combined with assay buffer and CSPD chemiluminescent substrate and incubated at room temperature for 10 minutes. The rate of substrate conversion was then determined in a luminometer (Turner 20/20TD). Cell density was determined by counting trypsinized cells in a Coulter ACT10 cell counter.

Transfection and culture of U-2 OS cells with pSDH plasmids: The human osteosarcoma U-2 OS cell line (ATCC #HTB-96) was cultured in Dulbecco's Modified Eagle Medium + 10% Fetal Calf Serum containing glutamine, penicillin, and streptomycin (supra) at 37 C/5% CO2. Cells were co-transfected with the pSDH-CMV vector, and its derivatives containing STAR6 or STAR8 in MCSI and MCSII, (along with plasmid pBabe-Puro) using SuperFect (supra). Puromycin selection was complete in 2 weeks, after which time individual puromycin resistant U-2 OS/pSDH-CMV clones were isolated at random and cultured further.

Luciferase assay: Luciferase activity (Himes & Shannon, 2000) was assayed in resuspended cells according to the instructions of the assay kit manufacturer (Roche 1669893), using a luminometer (Turner 20/20TD). Total cellular protein concentration was determined by the bicinchoninic acid method according to the manufacturer's instructions (Sigma B-9643), and used to normalize the luciferase data.

Results

Recombinant CHO cell clones containing the pSDH-CSP vector, or pSDH-CSP plasmids containing STAR6 or STAR49 (Table 6), were cultured for 3 weeks. The SEAP activity in the culture supernatants was then determined, and is expressed on the basis of cell number (FIG 8). As can be seen, clones with STAR elements in the expression units were isolated that express 2-3 fold higher SEAP activity than clones whose expression units do not include STAR elements. Furthermore, the number of STAR-containing clones that express

SEAP activity at or above the maximal activity of the STAR-less clones is quite high: 25% to 40% of the STAR clone populations exceed the highest SEAP expression of the pSDH-CSP clones.

Recombinant U-2 OS cell clones containing the pSDH-CMV vector, or pSDH-CMV plasmids containing STAR6 or STAR8 (Table 6), were cultured for 3 weeks. The luciferase activity in the host cells was then determined, and is expressed as relative luciferase units (FIG 9), normalized to total cell protein. The recombinant U-2 OS clones with STAR elements flanking the expression units had higher yields than the STAR-less clones: the highest expression observed from STAR8 clones was 2-3 fold higher than the expression from STAR-less clones. STAR6 clones had maximal expression levels 5 fold higher than the STAR-less clones. The STAR elements conferred greater predictability as well: for both STAR elements, 15 to 20% of the clones displayed luciferase expression at levels comparable to or greater than the STAR-less clone with the highest expression level.

These results demonstrate that, when used with the strong CMV promoter, STAR elements increase the yield of heterologous proteins (luciferase and SEAP). All three of the STAR elements introduced in this example provide elevated yields. The increased predictability conferred by the STAR elements is manifested by the large proportion of the clones with yields equal to or greater than the highest yields displayed by the STAR-less clones.

Example 12

STAR elements improve the stability of transgene expression

During cultivation of recombinant host cells, it is common practice to maintain antibiotic selection. This is intended to prevent transcriptional silencing of the transgene, or loss of the transgene from the genome by processes such as

recombination. However it is undesirable for production of heterologous proteins, for a number of reasons. First, the antibiotics that are used are quite expensive, and contribute significantly to the unit cost of the product. Second, for biopharmaceutical use, the protein must be demonstrably pure, with no traces of the antibiotic in the product. One advantage of STAR elements for heterologous protein production is that they confer stable expression on transgenes during prolonged cultivation, even in the absence of antibiotic selection; this property is demonstrated in this example.

Materials and Methods

The U-2 OS cell line was transfected with the plasmid pSDH-Tet-STAR6 and cultivated as described in Example 11. Individual puromycin-resistant clones were isolated and cultivated further in the absence of doxycycline. At weekly intervals the cells were transferred to fresh culture vessels at a dilution of 1:20. Luciferase activity was measured at periodic intervals as described in Example 11. After 15 weeks the cultures were divided into two replicates; one replicate continued to receive puromycin, while the other replicate received no antibiotic for the remainder of the experiment (25 weeks total).

Results

Table 7 presents the data on luciferase expression by an expression unit flanked with STAR6 during prolonged growth with or without antibiotic. As can be seen, the expression of the reporter transgene, luciferase, remains stable in the U-2 OS host cells for the duration of the experiment. After the cultures were divided into two treatments (plus antibiotic and without antibiotic) the expression of luciferase was essentially stable in the absence of antibiotic selection. This demonstrates the ability of STAR elements to protect transgenes from silencing or loss during prolonged cultivation. It also demonstrates that this property is independent of antibiotic selection.

Therefore production of heterologous proteins is possible without incurring the costs of the antibiotic or of difficult downstream processing

Example 13

Minimal essential sequences of STAR elements

STAR elements are isolated from the genetic screen described in Example 1. The screen uses libraries constructed with human genomic DNA that was size-fractionated to approximately 0.5 – 2 kilobases (supra). The STAR elements range from 500 to 2361 base pairs (Table 6). It is likely that, for many of the STAR elements that have been isolated, STAR activity is conferred by a smaller DNA fragment than the initially isolated clone. It is useful to determine these minimum fragment sizes that are essential for STAR activity, for two reasons. First, smaller functional STAR elements would be advantageous in the design of compact expression vectors, since smaller vectors transfect host cells with higher efficiency. Second, determining minimum essential STAR sequences permits the modification of those sequences for enhanced functionality. Two STAR elements have been fine-mapped to determine their minimal essential sequences.

Materials and Methods:

STAR10 (1167 base pairs) and STAR27 (1520 base pairs) have been fine-mapped. They have been amplified by PCR to yield sub-fragments of approximately equal length (FIG 10 legend). For initial testing, these have been cloned into the pSelect vector at the BamHI site, and transfected into U-2 OS/Tet-Off/LexA-HP1 cells as described in Example 1. After selection for hygromycin resistance, LexA-HP1 was induced by lowering the doxycycline concentration. Transfected cells were then incubated with zeocin to test the ability of the STAR fragments to protect the SV40-Zeo expression unit from repression due to LexA-HP1 binding.

Results

In this experiment STAR10 and STAR 27 confer good protection against gene silencing, as expected (FIG 10). This is manifested by robust growth in the presence of zeocin.

Of the 3 STAR10 sub-fragments, 10A (~400 base pairs) confers on transfected cells vigorous growth in the presence of zeocin, exceeding that of the full-length STAR element. Cells transfected with pSelect constructs containing the other 2 sub-fragments do not grow in the presence of zeocin. These results identify the ~400 base pair 10A fragment as encompassing the DNA sequence responsible for the anti-repression activity of STAR10.

STAR27 confers moderate growth in zeocin to transfected cells in this experiment (FIG 10). One of the sub-fragments of this STAR, 27B (~500 base pairs), permits weak growth of the host cells in zeocin-containing medium. This suggests that the anti-repression activity of this STAR is partially localized on sub-fragment 27B, but full activity requires sequences from 27A and/or 27C (each ~ 500 base pairs) as well.

Example 14

STAR elements function in diverse strains of cultured mammalian cells

The choice of host cell line for heterologous protein expression is a critical parameter for the quality, yield, and unit cost of the protein. Considerations such as post-translational modifications, secretory pathway capacity, and cell line immortality dictate the appropriate cell line for a particular biopharmaceutical production system. For this reason, the advantages provided by STAR elements in terms of yield, predictability, and stability should be obtainable in diverse cell lines. This was tested by comparing the function of STAR6 in the human U-2 OS cell line in which it was originally cloned, and the CHO cell line which is widely applied in biotechnology.

Materials and Methods:

The experiments of Example 11 are referred to.

Results

The expression of the SEAP reporter gene in CHO cells is presented in FIG 8; the expression of the luciferase reporter gene in U-2 OS cells is presented in FIG 9. By comparison of the results of these two experiments, it is apparent that the STAR6 element is functional in both cell lines: reporter gene expression was more predictable in both of them, and clones of each cell line displayed higher yields, when the reporter gene was shielded from position effects by STAR6. These two cell lines are derived from different species (human and hamster) and different tissue types (bone and ovary), reflecting the broad range of host cells in which this STAR element can be utilized in improving heterologous protein expression.

Example 15

STAR elements function in the context of various transcriptional promoters

Transgene transcription is achieved by placing the transgene open reading frame under control of an exogenous promoter. The choice of promoter is influenced by the nature of the heterologous protein and the production system. In most cases, strong constitutive promoters are preferred because of the high yields they can provide. Some viral promoters have these properties; the promoter/enhancer of the cytomegalovirus immediate early gene ("CMV promoter") is generally regarded as the strongest promoter in common biotechnological use (Boshart et al., 1985, Doll et al., 1996, Foecking & Hofstetter, 1986). The simian virus SV40 promoter is also moderately strong (Boshart et al., 1985, Foecking & Hofstetter, 1986) and is frequently used for

ectopic expression in mammalian cell vectors. The Tet-Off promoter is inducible: the promoter is repressed in the presence of tetracycline or related antibiotics (doxycycline is commonly used) in cell-lines which express the tTA plasmid (Clontech K1620-A), and removal of the antibiotic results in transcriptional induction (Deuschle et al., 1995, Gossen & Bujard, 1992, Izumi & Gilbert, 1999, Umana et al., 1999).

Materials and Methods:

The construction of the pSDH-Tet and pSDH-CMV vectors is described in Example 11. pSDH-SV40 was constructed by PCR amplification of the SV40 promoter (primers D41 and D42) from plasmid pSelect-SV40-Zeo (Example 1), followed by digestion of the PCR product with SacII and SalI. The pSDH-CMV vector was digested with SacII and SalI to remove the CMV promoter, and the vector and SV40 fragment were ligated together to create pSDH-SV40. STAR6 was cloned into MCSI and MCSII as described in Example 11. The plasmids pSDH-Tet, pSDH-Tet-STAR6, pSDH-Tet-STAR7, pSDH-SV40 and pSDH-SV40-STAR6 were co-transfected with pBabe-Puro into U-2 OS using SuperFect as described by the manufacturer. Cell cultivation, puromycin selection, and luciferase assays were carried out as described in Example 11.

Results

FIGS 9, 11, and 12 compare the expression of the luciferase reporter gene from 3 different promoters: two strong and constitutive viral promoters (CMV and SV40), and the inducible Tet-Off promoter. All three promoters were tested in the context of the STAR6 element in U-2 OS cells. The results demonstrate that the yield and predictability from all 3 promoters are increased by STAR6. As described in Examples 11 and 14, STAR6 is beneficial in the context of the CMV promoter (FIG 9). Similar improvements are seen in the context of the

SV40 promoter (FIG 11): the yield from the highest-expressing STAR6 clone is 2-3 fold greater than the best pSDH-SV40 clones, and 6 STAR clones (20% of the population) have yields higher than the best STAR-less clones. In the context of the Tet-Off promoter under inducing (low doxycycline) concentrations, STAR6 also improves the yield and predictability of transgene expression (FIG 12): the highest-expressing STAR6 clone has a 20-fold higher yield than the best pSDH-Tet clone, and 9 STAR6 clones (35% of the population) have yields higher than the best STAR-less clone. It is concluded that this STAR element is versatile in its transgene-protecting properties, since it functions in the context of various biotechnologically useful promoters of transcription.

Example 16

STAR element function can be directional

While short nucleic acid sequences can be symmetrical (e.g. palindromic), longer naturally-occurring sequences are typically asymmetrical. As a result, the information content of nucleic acid sequences is directional, and the sequences themselves can be described with respect to their 5' and 3' ends. The directionality of nucleic acid sequence information affects the arrangement in which recombinant DNA molecules are assembled using standard cloning techniques known in the art (Sambrook et al., 1989). STAR elements are long, asymmetrical DNA sequences, and have a directionality based on the orientation in which they were originally cloned in the pSelect vector. In the examples given above,using two STAR elements in pSDH vectors, this directionality was preserved. This orientation is described as the native or 5'-3' orientation, relative to the zeocin resistance gene (see FIG 13). In this example the importance of directionality for STAR function is tested in the pSDH-Tet vector. Since the reporter genes in the pSDH vectors are flanked on both sides by copies of the STAR element of interest, the orientation of each STAR copy

must be considered. This example compares the native orientation with the opposite orientation (FIG 13).

Materials and Methods:

The STAR66 element was cloned into pSDH-Tet as described in Example 11. U-2 OS cells were co-transfected with plasmids pSDH-Tet-STAR66-native and pSDH-Tet-STAR66-opposite, and cultivated as described in Example 11. Individual clones were isolated and cultivated; the level of luciferase expression was determined as described (supra).

Results

The results of the comparison of STAR66 activity in the native orientation and the opposite orientation are shown in FIG 14. When STAR66 is in the opposite orientation, the yield of only one clone is reasonably high (60 luciferase units). In contrast, the yield of the highest-expressing clone when STAR66 is in the native orientation is considerably higher (100 luciferase units), and the predictability is much higher as well: 7 clones of the native-orientation population (30%) express luciferase above the level of the highest-expressing clone from the opposite-orientation population, and 15 of the clones in the native-orientation population (60%) express luciferase above 10 relative luciferase units. Therefore it is demonstrated that STAR66 function is directional.

Example 17

Transgene expression in the context of STAR elements is copy number-dependent

Transgene expression units for heterologous protein expression are generally integrated into the genome of the host cell to ensure stable retention during cell division. Integration can result in one or multiple copies of the expression

unit being inserted into the genome; multiple copies may or may not be present as tandem arrays. The increased yield demonstrated for transgenes protected by STAR elements (supra) suggests that STAR elements are able to permit the transgene expression units to function independently of influences on transcription associated with the site of integration in the genome (independence from position effects (Boivin & Dura, 1998)). It suggests further that the STAR elements permit each expression unit to function independently of neighboring copies of the expression unit when they are integrated as a tandem array (independence from repeat-induced gene silencing (Garrick et al., 1998)). Copy number-dependence is determined from the relationship between transgene expression levels and copy number, as described in the example below.

Materials and Methods:

U-2 OS cells were co-transfected with pSDH-Tet-STAR10 and cultivated under puromycin selection as described (supra). Eight individual clones were isolated and cultivated further. Then cells were harvested, and one portion was assayed for luciferase activity as described (supra). The remaining cells were lysed and the genomic DNA purified using the DNeasy Tissue Kit (QIAGEN 69504) as described by the manufacturer. DNA samples were quantitated by UV spectrophotometry. Three micrograms of each genomic DNA sample were digested with PvuII and XhoI overnight as described by the manufacturer (New England Biolabs), and resolved by agarose gel electrophoresis. DNA fragments were transferred to a nylon membrane as described (Sambrook et al., 1989), and hybridized with a radioactively labelled probe to the luciferase gene (isolated from BamHI/SacII-digested pSDH-Tet). The blot was washed as described (Sambrook et al., 1989) and exposed to a phosphorimager screen (Personal F/X, BioRad). The resulting autoradiogram (FIG 15) was analyzed by densitometry to determine the relative strength of the luciferase DNA bands, which represents the transgene copy number.

Results

The enzyme activities and copy numbers (DNA band intensities) of luciferase in the clones from the pSDH-Tet-STAR10 clone population is shown in FIG 16. The transgene copy number is highly correlated with the level of luciferase expression in these pSDH-Tet-STAR10 clones (r = 0.86). This suggests that STAR10 confers copy number-dependence on the transgene expression units, making transgene expression independent of other transgene copies in tandem arrays, and independent of gene-silencing influences at the site of integration.

Example 18

STAR elements function as enhancer blockers but not enhancers

Gene promoters are subject to both positive and negative influences on their ability to initiate transcription. An important class of elements that exert positive influences are enhancers. Enhancers are characteristically able to affect promoters even when they are located far away (many kilobase pairs) from the promoter. Negative influences that act by heterochromatin formation (e.g. Polycomb group proteins) have been described above, and these are the target of STAR activity. The biochemical basis for enhancer function and for heterochromatin formation is fundamentally similar, since they both involve binding of proteins to DNA. Therefore it is important to determine whether STAR elements are able to block positive influences as well as negative influences, in other words, to shield transgenes from genomic enhancers in the vicinity of the site of integration. The ability to shield transgenes from enhancer activity ensures stable and predictable performance of transgenes in biotechnological applications. This example examines the performance of STAR elements in an enhancer-blocking assay.

Another feature of STAR activity that is important to their function is the increased yield they confer on transgenes (Example 11). STARs are isolated on

the basis of their ability to maintain high levels of zeocin expression when heterochromatin-forming proteins are bound adjacent to the candidate STAR elements. High expression is predicted to occur because STARs are anticipated to block the spread of heterochromatin into the zeocin expression unit. However, a second scenario is that the DNA fragments in zeocin-resistant clones contain enhancers. Enhancers have been demonstrated to have the ability to overcome the repressive effects of Polycomb-group proteins such as those used in the method of the STAR screen (Zink & Paro, 1995). Enhancers isolated by this phenomenon would be considered false positives, since enhancers do not have the properties claimed here for STARs. In order to demonstrate that STAR elements are not enhancers, they have been tested in an enhancer assay.

The enhancer-blocking assay and the enhancer assay are methodologically and conceptually similar. The assays are shown schematically in FIG 17. The ability of STAR elements to block enhancers is performed using the E47/E-box enhancer system. The E47 protein is able to activate transcription by promoters when it is bound to an E-box DNA sequence located in the vicinity of those promoters (Quong et al., 2002). E47 is normally involved in regulation of B and T lymphocyte differentiation (Quong et al., 2002), but it is able to function in diverse cell types when expressed ectopically (Petersson et al., 2002). The E-box is a palindromic DNA sequence, CANNTG (Knofler et al., 2002). In the enhancer-blocking assay, an E-box is placed upstream of a luciferase reporter gene (including a minimal promoter) in an expression vector. A cloning site for STAR elements is placed between the E-box and the promoter. The E47 protein is encoded on a second plasmid. The assay is performed by transfecting both the E47 plasmid and the luciferase expression vector into cells; the E47 protein is expressed and binds to the E-box, and the E47/E-box complex is able to act as an enhancer. When the luciferase expression vector does not contain a STAR element, the E47/E-box complex enhances luciferase expression (FIG 17A, situation 1). When STAR elements

are inserted between the E-box and the promoter, their ability to block the enhancer is demonstrated by reduced expression of luciferase activity (FIG 17A, situation 2); if STARs cannot block enhancers, luciferase expression is activated (FIG 17A, situation 3).

The ability of STAR elements to act as enhancers utilizes the same luciferase expression vector. In the absence of E47, the E-box itself does not affect transcription. Instead, enhancer behaviour by STAR elements will result in activation of luciferase transcription. The assay is performed by transfecting the luciferase expression vector without the E47 plasmid. When the expression vector does not contain STAR elements, luciferase expression is low (FIG 17B, situation 1). If STAR elements do not have enhancer properties, luciferase expression is low when a STAR element is present in the vector (FIG 17B, situation 2). If STAR elements do have enhancer properties, luciferase expression will be activated in the STAR-containing vectors (FIG 17B, situation 3).

Materials and Methods:

The luciferase expression vector was constructed by inserting the E-box and a human alkaline phosphatase minimal promoter from plasmid mu-E5+E2x6-cat(x) (Ruezinsky et al., 1991) upstream of the luciferase gene in plasmid pGL3-basic (Promega E1751), to create pGL3-E-box-luciferase (gift of W. Romanow). The E47 expression plasmid contains the E47 open reading frame under control of a beta-actin promoter in the pHBAPr-1-neo plasmid; E47 in constitutively expressed from this plasmid (gift of W. Romanow).

STAR elements 1, 2, 3, 6, 10, 11, 18, and 27 have been cloned into the luciferase expression vector. Clones containing the Drosophila scs element and the chicken beta-globin HS4-6x core ("HS4") element have been included as positive controls (they are known to block enhancers, and to have no intrinsic enhancer properties (Chung et al., 1993, Kellum & Schedl, 1992)), and the empty luciferase expression vector has been included as a negative control. All

assays were performed using the U-2 OS cell line. In the enhancer-blocking assay, the E47 plasmid was co-transfected with the luciferase expression vectors (empty vector, or containing STAR or positive-control elements). In the enhancer assay, the E47 plasmid was co-transfected with STARless luciferase expression vector as a positive control for enhancer activity; all other samples received a mock plasmid during co-transfection. The transiently transfected cells were assayed for luciferase activity 48 hours after plasmid transfection (supra). The luciferase activity expressed from a plasmid containing no E-box or STAR/control elements was subtracted, and the luciferase activities were normalized to protein content as described (supra).

Results

FIG 18 shows the results of the enhancer-blocking assay. In the absence of STAR elements (or the known enhancer-blocking elements scs and HS4), the E47/E-box enhancer complex activates expression of luciferase ("vector"); this enhanced level of expression has been normalized to 100. Enhancer activity is blocked by all STAR elements tested. Enhancer activity is also blocked by the HS4 and scs elements, as expected (Bell et al., 2001, Gerasimova & Corces, 2001). These results demonstrate that in addition to their ability to block the spreading of transcriptional silencing (negative influences), STAR elements are able to block the action of enhancers (positive influences).

FIG 19 shows the results of the enhancer assay. The level of luciferase expression due to enhancement by the E47/E-box complex is set at 100 ("E47"). By comparison, none of the STAR elements bring about significant activation of luciferase expression. As expected, the scs and HS4 elements also do not bring about activation of the reporter gene. Therefore it is concluded that at least the tested STAR elements do not possess enhancer properties.

Example 19

Characterization of a Silencing Inducing Chromatin (SINC) element

Materials and Methods

The general features of the SINC screen have been described in Example 1, and some aspects of it are recapitulated here. One version of the pSS vector used for screening SINC elements in genomic DNA is pSS-codA::upp (FIG 20). It consists of the suicide gene expression unit, flanked by two STAR6 elements. The expression unit, consisting of the codA::upp suicide gene under control of the Tet-Off promoter, is downstream of a BglII restriction site. A second version of the pSS vector, pSS-hrGFP (FIG 21), was created by replacement of one STAR6 element with STAR8, and replacement of the suicide gene with the hrGFP gene, encoding green fluorescent protein (Stratagene 240059). Human genomic DNA from chromosome 22 (Research Genetics 96010-22) was partially digested with Sau3AI and size fractionated. The 0.5 – 10 kilobase-pair fraction was ligated into the BglII site of pSS-codA::upp. This library represented ~20,000 independent clones with an average insert size of 1.2 kilobase pairs. The library was amplified in Escherichia coli. The purified DNA from the amplified library was transfected into U-2 OS/Tet-Off cells (van der Vlag et al., 2000) by standard techniques (calcium phosphate; Life Technologies 18306-019). A control transfection was carried out using empty pSS-codA::upp vector DNA, yielding 2400 hygromycin resistant colonies. Transfected cells were selected for hygromycin resistance (25 mg/ml) over a 3 week period at high doxycycline (10 ng/ml), and 1800 hygromycin-resistant colonies were recovered from the library transfection. These colonies were then incubated with the pro-drug 5-fluorocytosine (5-FC) at 1 mg/ml, with a boost of 5 mg/ml for 4 days, at a doxycycline concentration of 10 ng/ml. After 3 weeks all but 3 weakly-growing control colonies (transfected with empty pSS-codA::upp) had died; 58 of the library-transfected colonies survived. These colonies were allowed to recover from pro-drug treatment, and cultivated further. The 5-FC-resistant isolates were harvested, the cells were lysed, and a portion of the DNA was subjected to PCR amplification using primers D30 and

D51 to recover the SINC elements. The PCR products from six 5-FC-resistant colonies were cloned between the HindIII and XhoI sites of pBluescript II SK(+) plasmid (Stratagene 212207) by conventional methods (Sambrook et al., 1989). The DNA sequences of the candidate SINC elements were determined as described (supra) using commercially available primers for the pBluescript vector (Stratagene 300301 and 300302). The sequences of these SINC elements are presented in Table 4B.

The 6 candidate SINC elements were cloned into plasmid pSS-hrGFP in their native orientations, and the resulting plasmids were transfected into U-2 OS/Tet-Off cells. After selection for hygromycin resistance, the populations of pSS-hrGFP-SINC transfectants were cultivated further at a high doxycycline concentration (10 ng/ml). Total cellular RNA was extracted using the RNeasy Mini Kit (QIAGEN 74104) as described by the manufacturer. Northern blot analysis of GFP mRNA abundance in these populations was assessed using standard techniques (Sambrook et al., 1989). The GFP probe was a BamHI-EcoRI fragment encompassing base pairs 690 to 1419 in phrGFP-1. The blots were also probed for hygromycin mRNA as a control for PSS-hrGFP-derived plasmid copy number, and for beta-actin as a control for genomically-encoded mRNA quantity. The hygromycin probe was an SfuI-Sal I fragment extending from 8219-10144 in pREP4 (Invitrogen), and the beta-actin probe was from Clontech, #9800-1. After hybridization and washing, the blots were exposed to phosphorimager screens, and the radioactive signals were visualized and quantitated using a BioRad Personal F/X phosphorimager.

Results

SINC elements cloned adjacent to the GFP reporter gene will induce silencing of reporter gene transcription, but will not affect transcription of other genes. Accurate measurement of SINC activity takes advantage of this fact by determining the expression of GFP relative to the expression of two reference genes, rather than simply measuring absolute GFP expression. One reference

gene is the hygromycin resistance gene on the pSS-hrGFP plasmid (outside the domain defined by the STAR elements; FIG 21), and the other is the genomic beta-actin gene. SINC activity is quantified by RNA blot analysis as a reduction in the ratio of the GFP signal to the hygromycin and beta-actin signals. Among the candidate SINC elements that have been characterized, some display a significant relative reduction in GFP transcription, indicating that these DNAs are able to induce formation of silent chromatin. The SINC35 element (labelled PSINKS35 in Table 4B) has the strongest activity of these candidates. It brings about a 69% reduction in the GFP/hygromycin ratio, and a 75% reduction in the GFP/beta-actin signal. The strength of SINC activity in the other 5 candidates described in the original application, and in a number of other candidate SINC elements that have been isolated and characterized subsequent to submission of that application, is less. Therefore SINC35 has superior performance as a potent genetic element for induction of silent chromatin in diverse biotechnological applications.

Example 20

STAR elements are conserved between mouse and human

BLAT analysis of the STAR DNA sequence against the human genome database (http://genome.ucsc.edu/cgi-bin/hgGateway) reveals that some of these sequences have high sequence conservation with other regions of the human genome. These duplicated regions are candidate STAR elements; if they do show STAR activity, they would be considered paralogs of the cloned STARs (two genes or genetic elements are said to be paralogous if they are derived from a duplication event (Li, 1997)).

BLAST analysis of the human STARs against the mouse genome (http://www.ensembl.org/Mus_musculus/blastview) also reveals regions of high sequence conservation between mouse and human. This sequence conservation has been shown for fragments of 15 out of the 65 human STAR elements. The

conservation ranges from 64% to 89%, over lengths of 141 base pairs to 909 base pairs (Table 8). These degrees of sequence conservation are remarkable and suggest that these DNA sequences may confer STAR activity within the mouse genome as well. Some of the sequences from the mouse and human genomes in Table 8 could be strictly defined as orthologs (two genes or genetic elements are said to be orthologous if they are derived from a speciation event (Li, 1997)). For example, STAR6 is between the SLC8A1 and HAAO genes in both the human and mouse genomes. In other cases, a cloned human STAR has a paralog within the human genome, and its ortholog has been identified in the mouse genome. For example, STAR3a is a fragment of the 15q11.2 region of human chromosome 15. This region is 96.9% identical (paralogous) with a DNA fragment at 5q33.3 on human chromosome 5, which is near the IL12B interleukin gene. These human DNAs share approximately 80% identity with a fragment of the 11B2 region on mouse chromosome 11. The 11B2 fragment is also near the (mouse) IL12B interleukin gene. Therefore STAR3a and the mouse 11B2 fragment can be strictly defined as paralogs. In order to test the hypothesis that STAR activity is shared between regions of high sequence conservation in the mouse and human genome, one of the human STARs with a conserved sequence in mouse, STAR18, has been analyzed in greater detail. The sequence conservation in the mouse genome detected with the original STAR18 clone extends leftward on human chromosome 2 for about 500 base pairs (FIG 22; left and right relate to the standard description of the arms of chromosome 2). In this example we examine whether the region of sequence conservation defines a "naturally occurring" STAR element in human that is more extensive in length than the original clone. We also examine whether the STAR function of this STAR element is conserved between mouse and human.

Materials and Methods

The region of mouse/human sequence conservation around STAR 18 was recovered from human BAC clone RP11-387A1 by PCR amplification, in three fragments: the entire region (primers E93 and E94), the leftward half (primers E93 and E92), and the rightward half (primers E57 and E94). The corresponding fragments from the homologous mouse region were recovered from BAC clone RP23-400H17 in the same fashion (primers E95 and E98, E95 and E96, and E97 and E98, respectively). All fragments were cloned into the pSelect vector and transfected into a U-2 OS/Tet-Off/LexA-HP1 cell line (supra). Following transfection, hygromycin selection was carried out to select for transfected cells. The LexA-HP1 protein was induced by lowering the doxycycline concentration, and the ability of the transfected cells to withstand the antibiotic zeocin (a measure of STAR activity) was assessed by monitoring cell growth.

Results

The original STAR18 clone was isolated from Sau3AI digested human DNA ligated into the pSelect vector on the basis of its ability to prevent silencing of a zeocin resistance gene. Alignment of the human STAR18 clone (497 base pairs) with the mouse genome revealed high sequence similarity (72%) between the orthologous human and mouse STAR18 regions. It also uncovered high similarity (73%) in the region extending for 488 base pairs immediately leftwards of the Sau3AI site that defines the left end of the cloned region (FIG 22). Outside these regions the sequence similarity between human and mouse DNA drops below 60%.

As indicated in FIG 22, both the human and the mouse STAR18 elements confer survival on zeocin to host cells expressing the lexA-HP1 repressor protein. The original 497 base pair STAR18 clone and its mouse ortholog both confer the ability to grow (FIG 22, a and d). The adjacent 488 base pair regions of high similarity from both genomes also confer the ability to grow, and in fact their growth phenotype is more vigorous than that of the original STAR18

clone (FIG 22, b and e). When the entire region of sequence similarity was tested, these DNAs from both mouse and human confer growth, and the growth phenotype is more vigorous than the two sub-fragments (FIG 22, c and f). These results demonstrate that the STAR activity of human STAR18 is conserved in its ortholog from mouse. The high sequence conservation between these orthologous regions is particularly noteworthy because they are not protein-coding sequences, leading to the conclusion that they have some regulatory function that has prevented their evolutionary divergence through mutation.

This analysis demonstrates that cloned STAR elements identified by the original screening program may in some cases represent partial STAR elements, and that analysis of the genomic DNA in which they are embedded can identify sequences with stronger STAR activity.

Example 21

STAR elements contain characteristic DNA sequence motifs

STAR elements are isolated on the basis of their anti-repression phenotype with respect to transgene expression. This anti-repression phenotype reflects underlying biochemical processes that regulate chromatin formation which are associated with the STAR elements. These processes are typically sequence-specific and result from protein binding or DNA structure. This suggests that STAR elements will share DNA sequence similarity. Identification of sequence similarity among STAR elements will provide sequence motifs that are characteristic of the elements that have already been identified by functional screens and tests. The sequence motifs will also be useful to recognize and claim new STAR elements whose functions conform to the claims of this patent. The functions include improved yield and stability of transgenes expressed in eukaryotic host cells.

Other benefits of identifying sequence motifs that characterize STAR elements include: (1) provision of search motifs for prediction and identification of new STAR elements in genome databases, (2) provision of a rationale for modification of the elements, and (3) provision of information for functional analysis of STAR activity. Using bio-informatics, sequence similarities among STAR elements have been identified; the results are presented in this example.

Bio-informatic and Statistical Background. Regulatory DNA elements typically function via interaction with sequence-specific DNA-binding proteins. Bio-informatic analysis of DNA elements such as STAR elements whose regulatory properties have been identified, but whose interacting proteins are unknown, requires a statistical approach for identification of sequence motifs. This can be achieved by a method that detects short DNA sequence patterns that are over-represented in a set of regulatory DNA elements (e.g. the STAR elements) compared to a reference sequence (e.g. the complete human genome). The method determines the number of observed and expected occurrences of the patterns in each regulatory element. The number of expected occurrences is calculated from the number of observed occurrences of each pattern in the reference sequence.

The DNA sequence patterns can be oligonucleotides of a given length, e.g. six base pairs. In the simplest analysis, for a 6 base pair oligonucleotide (hexamer) composed of the four nucleotides (A, C, G, and T) there are 4^6 = 4096 distinct oligonucleotides (all combinations from AAAAAA to TTTTTT). If the regulatory and reference sequences were completely random and had equal proportions of the A, C, G, and T nucleotides, then the expected frequency of each hexamer would be 1/4096 (~0.00024). However, the actual frequency of each hexamer in the reference sequence is typically different than this due to biases in the content of G:C base pairs, etc. Therefore the frequency of each oligonucleotide in the reference sequence is determined empirically by counting, to create a "frequency table" for the patterns.

The pattern frequency table of the reference sequence is then used to calculate the expected frequency of occurrence of each pattern in the regulatory element set. The expected frequencies are compared with the observed frequencies of occurrence of the patterns. Patterns that are "over-represented" in the set are identified; for example, if the hexamer ACGTGA is expected to occur 5 times in 20 kilobase pairs of sequence, but is observed to occur 15 times, then it is three-fold over-represented. Ten of the 15 occurrences of that hexameric sequence pattern would not be expected in the regulatory elements if the elements had the same hexamer composition as the entire genome. Once the over-represented patterns are identified, a statistical test is applied to determine whether their over-representation is significant, or may be due to chance. For this test, a significance index, "sig", is calculated for each pattern. The significance index is derived from the probability of occurrence of each pattern, which is estimated by a binomial distribution. The probability takes into account the number of possible patterns (4096 for hexamers). The highest sig values corespond to the most overrepresented oligonucleotides (van Helden et al., 1998). In practical terms, oligonucleotides with sig >= 0 are considered as over-represented. A pattern with sig >= 0 is likely to be over-represented due to chance once (=10^0) in the set of regulatory element sequences. However, at sig >= 1 a pattern is expected to be over-represented once in ten (=10^1) sequence sets, sig >= 2 once in 100 (=10^2) sequence sets, etc.
The patterns that are significantly over-represented in the regulatory element set are used to develop a model for classification and prediction of regulatory element sequences. This employs Discriminant Analysis, a so-called "supervised" method of statistical classification known to one of ordinary skill in the art (Huberty, 1994). In Discriminant Analysis, sets of known or classified items (e.g. STAR elements) are used to "train" a model to recognize those items on the basis of specific variables (e.g. sequence patterns such as hexamers). The trained model is then used to predict whether other items should be classified as belonging to the set of known items (e.g. is a DNA

sequence a STAR element). In this example, the known items in the training set are STAR elements (positive training set). They are contrasted with sequences that are randomly selected from the genome (negative training set) which have the same length as the STAR elements. Discriminant Analysis establishes criteria for discriminating positives from negatives based on a set of variables that distinguish the positives; in this example, the variables are the significantly over-represented patterns (e.g. hexamers).

When the number of over-represented patterns is high compared to the size of the training set, the model could become biased due to over-training. Over-training is circumvented by applying a forward stepwise selection of variables (Huberty, 1994). The goal of Stepwise Discriminant Analysis is to select the minimum number of variables that provides maximum discrimination between the positives and negatives. The model is trained by evaluating variables one-by-one for their ability to properly classify the items in the positive and negative training sets. This is done until addition of new variables to the model does not significantly increase the model's predictive power (i.e. until the classification error rate is minimized). This optimized model is then used for testing, in order to predict whether "new" items are positives or negatives (Huberty, 1994).

It is inherent in classification statistics that for complex items such as DNA sequences, some elements of the positive training set will be classified as negatives (false negatives), and some members of the negative training set will be classified as positives (false positives). When a trained model is applied to testing new items, the same types of misclassifications are expected to occur.

In the bio-informatic method described here, the first step, pattern frequency analysis, reduces a large set of sequence patterns (e.g. all 4096 hexamers) to a smaller set of significantly over-represented patterns (e.g. 100 hexamers); in the second step, Stepwise Discrimant Analysis reduces the set of over-represented patterns to the subset of those patterns that have maximal discriminative power (e.g. 5-10 hexamers). Therefore this approach provides

simple and robust criteria for identifying regulatory DNA elements such as STAR elements.

DNA-binding proteins can be distinguished on the basis of the type of binding site they occupy. Some recognize contiguous sequences; for this type of protein, patterns that are oligonucleotides of length 6 base pairs (hexamers) are fruitful for bio-informatic analysis (van Helden et al., 1998). Other proteins bind to sequence dyads: contact is made between pairs of highly conserved trinucleotides separated by a non-conserved region of fixed width (van Helden et al., 2000). In order to identify sequences in STAR elements that may be bound by dyad-binding proteins, frequency analysis was also conducted for this type of pattern, where the spacing between the two trinucleotides was varied from 0 to 20 (i.e. XXXN{0-20}XXX where X's are specific nucleotides composing the trinucleotides, and N's are random nucleotides from 0 to 20 base pairs in length). The results of dyad frequency analysis are also used for Linear Discriminant Analysis as described above.

Materials and Methods

Using the genetic screen described in the original patent application, sixty-six (66) STAR elements were initially isolated from human genomic DNA and characterized in detail (Table 6). The screen was performed on gene libraries constructed by Sau3AI digestion of human genomic DNA, either purified from placenta (Clontech 6550-1) or carried in bacterial/P1 (BAC/PAC) artificial chromosomes. The BAC/PAC clones contain genomic DNA from regions of chromosome 1 (clones RP1154H19 and RP3328E19), from the HOX cluster of homeotic genes (clones RP1167F23, RP1170019, and RP11387A1), or from human chromosome 22 (Research Genetics 96010-22). The DNAs were size-fractionated, and the 0.5 - 2 kb size fraction was ligated into BamHI-digested pSelect vector, by standard techniques (Sambrook et al., 1989). pSelect plasmids containing human genomic DNA that conferred resistance to zeocin at low doxycycline concentrations were isolated and propagated in Escherichia

coli. The screens that yielded the STAR elements of Table 6 have assayed approximately 1-2% of the human genome.

The human genomic DNA inserts in these 66 plasmids were sequenced by the dideoxy method (Sanger et al., 1977) using a Beckman CEQ2000 automated DNA sequencer, using the manufacturer's instructions. Briefly, DNA was purified from E. coli using QIAprep Spin Miniprep and Plasmid Midi Kits (QIAGEN 27106 and 12145, respectively). Cycle sequencing was carried out using custom oligonucleotides corresponding to the pSelect vector (primers D89 and D95, Table 5), in the presence of dye terminators (CEQ Dye Terminator Cycle Sequencing Kit, Beckman 608000). Assembled STAR DNA sequences were located in the human genome (database builds August and December 2001) using BLAT (Basic Local Alignment Tool (Kent, 2002); http://genome.ucsc.edu/cgi-bin/hgGateway ; Table 6). In aggregate, the combined STAR sequences comprise 85.6 kilobase pairs, with an average length of 1.3 kilobase pairs.

Sequence motifs that distinguish STAR elements within human genomic DNA were identified by bio-informatic analysis using a two-step procedure, as follows (see FIG 23 for a schematic diagram). The analysis has two input datasets: (1) the DNA sequences of the STAR elements (STAR1 – STAR65 were used; Table 6); and (2) the DNA sequence of the human genome (except for chromosome 1, which was not feasible to include due to its large size; for dyad analysis a random subset of human genomic DNA sequence (~27 Mb) was used).

Pattern Frequency Analysis. The first step in the analysis uses RSA-Tools software (Regulatory Sequence Analysis Tools; http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools/ ; references (van Helden et al., 1998, van Helden et al., 2000, van Helden et al., 2000)) to determine the following information: (1) the frequencies of all dyads and hexameric oligonucleotides in the human genome; (2) the frequencies of the oligonucleotides and dyads in the 65 STAR elements; and (3) the significance

indices of those oligonucleotides and dyads that are over-represented in the STAR elements compared to the genome. A control analysis was done with 65 sequences that were selected at random from the human genome (i.e. from 2689 x 10^3 kilobase pairs) that match the length of the STAR elements of Table 6.

Discriminant Analysis. The over-represented oligonucleotides and dyads were used to train models for prediction of STAR elements by Linear Discriminant Analysis (Huberty, 1994). A pre-selection of variables was performed by selecting the 50 patterns with the highest individual dicriminatory power from the over-represented oligos or dyads of the frequency analyses. These pre-selected variables were then used for model training in a Stepwise Linear Discriminant Analysis to select the most discriminant combination of variables (Huberty, 1994). Variable selection was based on minimizing the classification error rate (percentage of false negative classifications). In addition, the expected error rate was estimated by applying the same discriminant approach to the control set of random sequences (minimizing the percentage of false positive classifications).

The predictive models from the training phase of Discriminant Analysis were tested in two ways. First, the STAR elements and random sequences that were used to generate the model (the training sets) were classified. Second, sequences in a collection of 19 candidate STAR elements (recently cloned by zeocin selection as described above) were classified. These candidate STAR elements are listed in Table 11 (SEQ ID:67-84).

Results

Pattern frequency analysis was performed with RSA-Tools on 65 STAR elements, using the human genome as the reference sequence. One hundred sixty-six (166) hexameric oligonucleotides were found to be over-represented in the set of STAR elements (sig >= 0) compared to the entire genome (Table 9). The most significantly over-represented oligonucleotide, CCCCAC, occurs 107

times among the 65 STAR elements, but is expected to occur only 49 times. It has a significance coefficient of 8.76; in other words, the probability that its over-representation is due to random chance is 1 / 10^8.76, i.e. less than one in 500 million.

Ninety-five of the oligonucleotides have a significance coefficient greater than 1, and are therefore highly over-represented in the STAR elements. Among the over-represented oligonucleotides, their observed and expected occurrences, respectively, range from 6 and 1 (for oligo 163, CGCGAA, sig = 0.02) to 133 and 95 (for oligo 120, CCCAGG, sig = 0.49). The differences in expected occurrences reflect factors such as the G:C content of the human genome. Therefore the differences among the oligonucleotides in their number of occurrences is less important than their over-representation; for example, oligo 2 (CAGCGG) is 36 / 9 = 4-fold over-represented, which has a probability of being due to random chance of one in fifty million (sig = 7.75).

Table 9 also presents the number of STAR elements in which each over-represented oligonucleotide is found. For example, the most significant oligonucleotide, oligo 1 (CCCCAC), occurs 107 times, but is found in only 51 STARs, i.e. on average it occurs as two copies per STAR. The least abundant oligonucleotide, number 166 (AATCGG), occurs on average as a single copy per STAR (thirteen occurrences on eleven STARs); single-copy oligonucleotides occur frequently, especially for the lower-abundance oligos. At the other extreme, oligo 4 (CAGCCC) occurs on average 3 times in those STARs in which it is found (37 STARs). The most widespread oligonucleotide is number 120 (CCCAGG), which occurs on 58 STARs (on average twice per STAR), and the least widespread oligonucleotide is number 114 (CGTCGC), which occurs on only 6 STARs (and on average only once per STAR).

Results of dyad frequency analysis are given in Table 10. Seven hundred thirty (730) dyads were found to be over-represented in the set of STAR elements (sig >= 0) compared to the reference sequence. The most significantly over-represented dyad, CCCN{2}CGG, occurs 36 times among the 65 STAR

elements, but is expected to occur only 7 times. It has a significance coefficient of 9.31; in other words, the probability that its over-representation is due to chance is 1 / 10^9.31, i.e. less than one in 2 billion.

Three hundred ninety-seven of the dyads have a significance coefficient greater than 1, and are therefore highly over-represented in the STAR elements. Among the over-represented dyads, their observed and expected occurrences, respectively, range from 9 and 1 (for five dyads (numbers 380, 435, 493, 640, and 665)) to 118 and 63 (for number 30 (AGGN{2}GGG), sig = 4.44).

The oligonucleotides and dyads found to be over-represented in STAR elements by pattern frequency analysis were tested for their discriminative power by Linear Discriminant Analysis. Discriminant models were trained by step-wise selection of the best combination among the 50 most discriminant oligonucleotide (Table 9) or dyad (Table 10) patterns. The models achieved optimal error rates after incorporation of 4 (dyad) or 5 variables. The discriminative variables from oligo analysis are numbers 11, 30, 94, 122, and 160 (Tables 9); those from dyad analysis are numbers 73, 194, 419, and 497 (Table 10).

The discriminant models were then used to classify the 65 STAR elements in the training set and their associated random sequences. The model using oligonucleotide variables classifies 46 of the 65 STAR elements as STAR elements (true positives); the dyad model classifies 49 of the STAR elements as true positives. In combination, the models classify 59 of the 65 STAR elements as STAR elements (91%; FIG 24). The false positive rates (random sequences classified as STARs) were 7 for the dyad model, 8 for the oligonucleotide model, and 13 for the combined predictions of the two models (20%). The STAR elements of Table 6 that were not classified as STARs by LDA are STARs 7, 22, 35, 44, 46, and 65. These elements display stabilizing anti-repressor activity in functional assays, so the fact that they are not classified as STARs

by LDA suggests that they represent another class (or classes) of STAR elements.

The models were then used to classify the 19 candidate STAR elements in the testing set listed in Table 11. The dyad model classifies 12 of these candidate STARs as STAR elements, and the oligonucleotide model classifies 14 as STARs. The combined number of the candidates that are classified as STAR elements is 15 (79%). This is a lower rate of classification than obtained with the training set of 65 STARs; this is expected for two reasons. First, the discriminant models were trained with the 65 STARs of Table 6, and discriminative variables based on this training set may be less well represented in the testing set. Second, the candidate STAR sequences in the testing set have not yet been fully characterized in terms of in vivo function, and may include elements with only weak anti-repression properties.

This analysis demonstrates the power of a statistical approach to bio-informatic classification of STAR elements. The STAR sequences contain a number of dyad and hexameric oligonucleotide patterns that are significantly over-represented in comparison with the human genome as a whole. These patterns may represent binding sites for proteins that confer STAR activity; in any case they form a set of sequence motifs that can be used to recognize STAR element sequences.

Using these patterns to recognize STAR elements by Discriminant Analysis, a high proportion of the elements obtained by the genetic screen of the invention are in fact classified as STARs. This reflects underlying sequence and functional similarities among these elements. An important aspect of the method described here (pattern frequency analysis followed by Discriminant Analysis) is that it can be reiterated; for example, by including the 19 candidate STAR elements of Table 11 with the 66 STAR elements of Table 6 into one training set, an improved discriminant model can be trained. This improved model can then be used to classify other candidate regulatory elements as STARs. Large-scale in vivo screening of genomic sequences using

the method of the invention, combined with reiteration of the bio-informatic analysis, will provide a means of discriminating STAR elements that asymptotically approaches 100% recognition and prediction of elements as the genome is screened in its entirety. These stringent and comprehensive predictions of STAR function will ensure that all human STAR elements are recognized, and are available for use in improving transgene expression.

Example 22

Cloning and characterization of STAR elements from Arabidopsis thaliana

Transgene silencing occurs in transgenic plants at both the transcriptional and post-transcriptional levels (Meyer, 2000, Vance & Vaucheret, 2001). In either case, the desired result of transgene expression can be compromised by silencing; the low expression and instability of the transgene results in poor expression of desirable traits (e.g. pest resistance) or low yields of recombinant proteins. It also results in poor predictability: the proportion of transgenic plants that express the transgene at biotechnologically useful levels is low, which necessitates laborious and expensive screening of transformed individuals for those with beneficial expression characteristics. This example describes the isolation of STAR elements from the genome of the dicot plant Arabidopsis thaliana for use in preventing transcriptional transgene silencing in transgenic plants. Arabidopsis was chosen for this example because it is a well-studied model organism: it has a compact genome, it is amenable to genetic and recombinant DNA manipulations, and its genome has been sequenced (Bevan et al., 2001, Initiative, 2000, Meinke et al., 1998).

Materials and Methods:

Genomic DNA was isolated from Arabidopsis thaliana ecotype Columbia as described (Stam et al., 1998) and partially digested with MboI. The digested DNA was size-fractionated to 0.5 – 2 kilbase pairs by agarose gel electrophoresis and purification from the gel (QIAquick Gel Extraction Kit,

QIAGEN 28706), followed by ligation into the pSelect vector (supra). Transfection into the U-2 OS/Tet-Off/LexA-HP1 cell line and selection for zeocin resistance at low doxycycline concentration was performed as described (supra). Plasmids were isolated from zeocin resistant colonies and re-transfected into the U-2 OS/Tet-Off/LexA-HP1 cell line.

Sequencing of Arabidopsis genomic DNA fragments that conferred zeocin resistance upon re-transfection was performed as described (supra). The DNA sequences were compared to the sequence of the Arabidopsis genome by BLAST analysis ((Altschul et al., 1990); URL http://www.ncbi.nlm.nih.gov/blast/Blast).

STAR activity was tested further by measuring mRNA levels for the hygromycin- and zeocin-resistance genes in recombinant host cells by reverse transcription PCR (RT-PCR). Cells of the U-2 OS/Tet-Off/lexA-HP1 cell line were transfected with pSelect plasmids containing Arabidopsis STAR elements, the Drosophila scs element, or containing no insert (supra). These were cultivated on hygromycin for 2 weeks at high doxycycline concentration, then the doxycycline concentration was lowered to 0.1 ng/ml to induce the lexA-HP1 repressor protein. After 10 days, total RNA was isolated by the RNeasy mini kit (QIAGEN 74104) as described by the manufacturer. First-strand cDNA synthesis was carried out using the RevertAid First Strand cDNA Synthesis kit (MBI Fermentas 1622) using oligo(dT)18 primer as described by the manufacturer. An aliquot of the cDNA was used as the template in a PCR reaction using primers D58 and D80 (for the zeocin marker), and D70 and D71 (for the hygromycin marker), and Taq DNA polymerase (Promega M2661). The reaction conditions were 15-20 cycles of 94C for 1 minute, 54C for 1minute, and 72C for 90 seconds. These conditions result in a linear relationship between input RNA and PCR product DNA. The PCR products were resolved by agarose gel electrophoresis, and the zeocin and hygromycin bands were detected by Southern blotting as described (Sambrook et al., 1989), using PCR products produced as above with purified pSelect

plasmid as template. The ratio of the zeocin and hygromycin signals corresponds to the normalized expression level of the zeocin gene.

Results

The library of Arabidopsis genomic DNA in the pSelect vector comprised 69,000 primary clones in E. coli, 80% of which carried inserts. The average insert size was approximately 1000 base pairs; the library therefore represents approximately 40% of the Arabidopsis genome.

A portion of this library (representing approximately 16% of the Arabidopsis genome) was transfected into the U-2 OS/Tet-Off/LexA-HP1 cell line. Hygromycin selection was imposed to isolate transfectants, which resulted in 27,000 surviving colonies. These were then subjected to zeocin selection at low doxycycline concentration. Putative STAR-containing plasmids from 56 zeocin-resistant colonies were rescued into E. coli and re-transfected into U-2 OS/Tet-Off/LexA-HP1 cells. Forty-four of these plasmids (79% of the plasmids tested) conferred zeocin resistance on the host cells at low doxycycline concentrations, demonstrating that the plasmids carried STAR elements. This indicates that the pSelect screen in human U-2 OS cells is highly efficient at detection of STAR elements from plant genomic DNA.

The DNA sequences of these 44 candidate STAR elements were determined. Thirty-five of them were identified as single loci in the database of Arabidopsis nuclear genomic sequence (Table 12; SEQ ID:85 – SEQ ID:119). Four others were identified as coming from the chloroplast genome, four were chimeras of DNA fragments from two loci, and one was not found in the Arabidopsis genome database.

The strength of the cloned Arabidopsis STAR elements was tested by assessing their ability to prevent transcriptional repression of the zeocin-resistance gene, using an RT-PCR assay. As a control for RNA input among the samples, the transcript levels of the hygromycin-resistance gene for each STAR transfection were assessed too. This analysis has been performed for 12 of the Arabidopsis

STAR elements. The results (FIG 25) demonstrate that the Arabidopsis STAR elements are superior to the Drosophila scs element (positive control) and the empty vector ("SV40"; negative control) in their ability to protect the zeocin-resistance gene from transcriptional repression. In particular, STAR-A28 and STAR-A30 enable 2-fold higher levels of zeocin-resistance gene expression than the scs element (normalized to the internal control of hygromycin-resistance gene mRNA) when the lexA-HP1 repressor is expressed.

These results demonstrate that the method of the invention can be successfully applied to recovery of STAR elements from genomes of other species than human. Its successful application to STAR elements from a plant genome is particularly significant because it demonstrates the wide taxonomic range over which the method of the invention is applicable, and because plants are an important target of biotechnological development.

Brief description of the drawings

Figure 1. The pSelect family of plasmids for selecting and characterizing STAR elements. A resistance marker (zeocin or puromycin) or reporter gene (GFP or luciferase) under control of the promiscuous SV40 promoter is adjacent to a BamHI cloning site flanked by AscI and HindIII sites. Upstream of the cloning site are lexA operators to which lexA protein can bind. Binding of chimeric lexA-Polycomb group proteins to the operators causes repression of the marker or reporter gene. DNA fragments inserted at the cloning site that block repression are identified by the persistent expression of the marker or reporter gene. The plasmid replicates episomally in cultured mammalian cells due to the oriP seqeunce.

Figure 2. The pSDH family of plasmids for testing STAR elements. Two multiple cloning sites (MCSI and MCSII) flank a reporter gene (GFP or luciferase) whose expression is driven by an upstream promoter (CMV, Tet-off, or SV40). STAR elements to be tested are inserted at MCSI and MCSII. These contain unique restriction sites (MCSI: XhoI, NotI, EcoRI, and SalI; MCSII, HindIII, EcoRV, BglII, and NheI). The plasmid replicates after integrating at random in the genome of mammalian cells.

Figure 3. Proportion of clones over-expressing luciferase. U-2 OS human osteosarcoma cells were stably transfected with pSDH plasmids (containing the luciferase reporter gene under control of the tet-off promoter), and individual transfected clones were isolated and cultivated. Luciferase expression was measured enzymatically. The average luciferase expression by clones containing the STARless pSDH ("reference level") was determined. Clones from the sets for all plasmids were scored as "over-expressing" if their luciferase activity was more than 2-fold higher than the reference level. The percentage of over-expressing clones in each plasmid set in plotted.

Figure 4. Fold over-expression by over-expressing clones. The range of over-expression in STAR-containing pSDH plasmids integrated into genomic DNA was determined by dividing the luciferase activities of each clone by the reference level. For those displaying significant expression (more than 2-fold above the reference level), the actual fold increases were noted; the minimum and median values of these data are plotted for each plasmid.

Figure 5. Fold over-expression by over-expressing clones. The range of over-expression in STAR-containing pSDH plasmids integrated into genomic DNA was determined by dividing the luciferase activities of each clone by the reference level. For those displaying significant expression (more than 2-fold above the reference level), the actual fold increases were noted; the maximum values of these data are plotted for each plasmid.

Figure 6. The pSS (SINC-Select) plasmid for selecting and characterizing SINC elements. The codA::upp suicide gene encodes a protein that converts the pro-drug 5-fluorocytosine to the toxic drug 5-fluorouracil. Upon induction by lowered tetracycline concentration, host cells become sensitive to the pro-drug. Genomic DNA fragments inserted at the cloning site (BglII-XhoI) that have silencing activity will prevent expression of the suicide gene and allow formation of pro-drug resistant colonies. STAR elements flank the selection components to prevent spreading of silenced chromatin to the functional components of the plasmid. The plasmid replicates episomally in cultured mammalian cells due to the oriP seqeunce.

FIG 7. The pSDH-CSP plasmid used for testing STAR activity. The Secreted Alkaline Phosphatase (SEAP) reporter gene is under control of the CMV promoter, and the puromycin resistance selectable marker (puro) is under control of the SV40 promoter. Flanking these two genes are multiple cloning

sites into which STAR elements can be cloned. The plasmid also has an origin of replication (ori) and ampicillin resistance gene (ampR) for propagation in Escherichia coli

.

FIG 8. STAR6 and STAR49 improve predictability and yield of transgene expression. Expression of SEAP from the CMV promoter by CHO cells transfected with pSDH-CSP, pSDH-CSP-STAR6, or pSDH-CSP-STAR49 was determined. The STAR-containing constructs confer greater predictability and elevated yield relative to the pSDH-CSP construct alone.

FIG 9. STAR6 and STAR8 improve predictability and yield of transgene expression. Expression of luciferase from the CMV promoter by U-2 OS cells transfected with pSDH-CMV, pSDH-CMV-STAR6, or pSDH-CMV-STAR8 was determined. The STAR-containing constructs confer greater predictability and elevated yield relative to the pSDH-CMV construct alone.

FIG 10. Minimal essential sequences of STAR10 and STAR27. Portions of the STAR elements were amplified by PCR: STAR10 was amplified with primers E23 and E12 to yield fragment 10A, E13 and E14 to yield fragment 10B, and E15 and E16 to yield fragment 10C. STAR27 was amplified with primers E17 and E18 to yield fragment 27A, E19 and E20 to yield fragment 27B, and E21 and E22 to yield fragment 27C. These sub-fragments were cloned into the pSelect vector. After transfection into U-2 OS/Tet-Off/LexA-HP1 cells, the growth of the cultures in the presence of zeocin was monitored. Growth rates varied from vigorous (+++) to poor (+/-), while some cultures failed to survive zeocin treatment (-) due to absence of STAR activity in the DNA fragment tested.

FIG 11. STAR element function in the context of the SV40 promoter. pSDH-SV40 and pSDH-SV40-STAR6 were transfected into the human osteosarcoma

U-2 OS cell line, and expression of luciferase was assayed with or without protection from gene silencing by STAR6 in puromycin-resistant clones.

FIG 12. STAR element function in the context of the Tet-Off promoter. pSDH-Tet and pSDH-Tet-STAR6 were transfected into the human osteosarcoma U-2 OS cell line, and expression of luciferase was assayed with or without protection from gene silencing by STAR6 in puromycin-resistant clones.

FIG 13. Schematic diagram of the orientation of STAR elements as they are cloned in the pSelect vector (panel A), as they are cloned into pSDH vectors to preserve their native orientation (panel B), and as they are cloned into pSDH vector in the opposite orientation (panel C).

FIG 14. Directionality of STAR66 function. The STAR66 element was cloned into pSDH-Tet in either the native (STAR66 native) or the opposite orientation (STAR66 opposite), and transfected into U-2 OS cells. Luciferase activity was assayed in puromycin resistant clones.

FIG 15. Copy number-dependence of STAR function. Southern blot of luciferase expression units in pSDH-Tet-STAR10, integrated into U-2 OS genomic DNA. Radioactive luciferase DNA probe was used to detect the amount of transgene DNA in the genome of each clone, which was then quantified with a phosphorimager.

FIG 16. Copy number-dependence of STAR function. The copy number of pSDH-Tet-STAR10 expression units in each clone was determined by phosphorimagery, and compared with the activity of the luciferase reporter enzyme expressed by each clone.

FIG 17. Enhancer-blocking and enhancer assays. The luciferase expression vectors used for testing STARs for enhancer-blocking and enhancer activity are shown schematically. The E-box binding site for the E47 enhancer protein is upstream of a cloning site for STAR elements. Downstream of the STAR cloning site is the luciferase gene under control of a human alkaline phosphatase minimal promoter (mp). The histograms indicate the expected outcomes for the three possible experimental situations (see text). Panel A: Enhancer-blocking assay. Panel B: Enhancer assay.

FIG 18. Enhancer-blocking assay. Luciferase expression from a minimal promoter is activated by the E47/E-box enhancer in the empty vector (vector). Insertion of enhancer-blockers (scs, HS4) or STAR elements (STAR elements 1, 2, 3, 6, 10, 11, 18, and 27) block luciferase activation by the E47/E-box enhancer.

FIG 19. Enhancer assay. Luciferase expression from a minimal promoter is activated by the E47/E-box enhancer in the empty vector (E47). Insertion of the scs and HS4 elements or various STAR elements (STARs 1, 2, 3, 6, 10, 11, 18, and 27) do not activate transcription of the reporter gene.

FIG 20. The pSS-codA::upp vector used for isolation of SINC elements. The codA::upp suicide gene encodes a protein that converts the pro-drug 5-fluorocytosine to the toxic drug 5-fluorouracil. Upon induction by lowered doxycycline concentration, host cells become sensitive to the pro-drug. Genomic DNA fragments inserted at the BglII cloning site that have silencing activity will prevent expression of the suicide gene and allow formation of pro-drug resistant colonies. STAR elements flank the selection components to prevent spreading of silenced chromatin to the functional components of the plasmid. The plasmid is selected after transfection into mammalian cells with the hygromycin-resistance gene, and after transformation into E. coli with the

ampicillin-resistance gene. It replicates episomally in cultured mammalian cells due to the oriP and EBNA-1 sequences, and in E. coli due to the ori sequence.

FIG 21. The pSS-hrGFP plasmid is identical to the pSS-codA::upp plasmid, except for replacement of the suicide gene with hrGFP (encoding green fluorescent protein) and of STAR6 with STAR8 downstream of the GFP reporter gene.

FIG 22. STAR18 sequence conservation between mouse and human. The region of the human genome containing 497 base pair STAR18 is shown (black boxes); the element occurs between the HOXD8 and HOXD4 homeobox genes on human chromosome 2. It is aligned with a region in mouse chromosome 2 that shares 72% sequence identity. The region of human chromosome 2 immediately to the left of STAR18 is also highly conserved with mouse chromosome 2 (73% identity; gray boxes); beyond these region, the identity drops below 60%. The ability of these regions from human and mouse, either separately or in combination, to confer growth on zeocin is indicated: -, no growth; +, moderate growth; ++, vigorous growth; +++, rapid growth.

FIG 23. Schematic diagram of bio-informatic analysis workflow. For details, see text.

FIG 24. Results of discriminant analysis on classification of the training set of 65 STAR elements. STAR elements that are correctly classified as STARs by Stepwise Linear Discriminant Analysis (LDA) are shown in a Venn diagram. The variables for LDA were selected from frequency analysis results for hexameric oligonucleotides ("oligos") and for dyads. The diagram indicates the concordance of the two sets of variables in correctly classifying STARs.

FIG 25. U-2 OS/Tet-Off/lexA-HP1 cells were transfected with candidate Arabidopsis STAR elements and cultivated at low doxycycline concentrations. Total RNA was isolated and subjected to RT-PCR; the bands corresponding to the zeocin and hygromycin resistance mRNAs were detected by Southern blotting and quantified with a phosphorimager. The ratio of the zeocin to hygromycin signals is shown for transfectants containing zeocin expression units flanked by 12 different Arabidopsis STAR elements, the Drosophila scs element, or no flanking element.

Figure 26

Sequences comprising STAR1 – STAR65 (SEQ ID:1 – 65)

Sequences comprising STAR66 and testing set (SEQ ID:66 – 84),

Sequences comprising Arabidopsis STAR A1-A35 (SEQ ID:85-119).

References

Altschul, S.F. and Gish, W. (1996) Local alignment statistics. Methods Enzymol, 266, 460-480.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J Mol Biol, 215, 403-410.

Bell, AC, West, AG, and Felsenfeld, G. (2001) Insulators and boundaries: versatile regulatory elements in the eukaryotic genome Science 291, 447-50.

Bennett, R.P., Cox, C.A. and Hoeffler, J.P. (1998) Fusion of green fluorescent protein with the Zeocin-resistance marker allows visual screening and drug selection of transfected eukaryotic cells. Biotechniques, 24, 478-482.

Berger, J, Hauber, J, Hauber, R, Geiger, R, and Cullen, BR. (1988) Secreted placental alkaline phosphatase: a powerful new quantitative indicator of gene expression in eukaryotic cells Gene 66, 1-10.

Bevan, M, Mayer, K, White, O, Eisen, JA, Preuss, D, Bureau, T, Salzberg, SL, and Mewes, HW. (2001) Sequence and analysis of the Arabidopsis genome Curr Opin Plant Biol 4, 105-10.

Bierhuizen, M.F., Westerman, Y., Visser, T.P., Wognum, A.W. and Wagemaker, G. (1997) Green fluorescent protein variants as markers of retroviral-mediated gene transfer in primary hematopoietic cells and cell lines. Biochem Biophys Res Commun, 234, 371-375.

Boivin, A, and Dura, JM. (1998) In vivo chromatin accessibility correlates with gene silencing in Drosophila Genetics 150, 1539-49.

Boshart, M, Weber, F, Jahn, G, Dorsch-Hasler, K, Fleckenstein, B, and Schaffner, W. (1985) A very strong enhancer is located upstream of an immediate early gene of human cytomegalovirus Cell 41, 521-30.

Breckenridge, D.G. and Shore, G.C. (2000) Regulation of apoptosis by E1A and Myc oncoproteins. Crit Rev Eukaryot Gene Expr 10, 273-280.

Bunker, C.A. and Kingston, R.E. (1994) Transcriptional repression by Drosophila and mammalian Polycomb group proteins in transfected mammalian cells. Mol Cell Biol, 14, 1721-1732.

Cherry, J.R., Lamsa, M.H., Schneider, P., Vind, J., Svendsen, A., Jones, A. and Pedersen, A.H. (1999) Directed evolution of a fungal peroxidase. Nat Biotechnol, 17, 379-384.

Chung, JH, Whiteley, M, and Felsenfeld, G. (1993) A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila Cell 74, 505-14.

Deuschle, U, Meyer, WK, and Thiesen, HJ. (1995) Tetracycline-reversible silencing of eukaryotic promoters Mol Cell Biol 15, 1907-14.

Doll, R.F., Crandall, J.E., Dyer, C.A., Aucoin, J.M. and Smith, F.I. (1996) Comparison of promoter strengths on gene delivery into mammalian brain cells using AAV vectors. Gene Ther, 3, 437-447.

Eissenberg, J.C., James T.C., Foster-Hartnett D.M., Hartnett T., Ngan V., and Elgin S.C.R. (1990) Mutation in a heterochromatin-specific chromosomal protein is associated with suppression of position-effect variegation in *Drosophila melanogaster*. Proc Natl Acad Sci (USA) 87: 9923-9927.

Feng, Y.Q., Seibler, J., Alami, R., Eisen, A., Westerman, K.A., Leboulch, P., Fiering, S. and Bouhassira, E.E. (1999) Site-specific chromosomal integration in mammalian cells: highly efficient CRE recombinase-mediated cassette exchange. J Mol Biol, 292, 779-785.

Foecking, MK, and Hofstetter, H. (1986) Powerful and versatile enhancer-promoter unit for mammalian expression vectors Gene 45, 101-5.

Garrick, D, Fiering, S, Martin, DI, and Whitelaw, E. (1998) Repeat-induced gene silencing in mammals Nat Genet 18, 56-9.

Gaszner, M., Vazquez, J. and Schedl, P. (1999) The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction. Genes Dev, 13, 2098-2107.

Gerasimova, T.I. and Corces, V.G. (1998) Polycomb and trithorax group proteins mediate the function of a chromatin insulator. Cell, 92, 511-521.

Gerasimova, TI, and Corces, VG. (2001) Chromatin insulators and boundaries: effects on transcription and nuclear organization Annu Rev Genet 35, 193-208.

Gossen, M. and Bujard, H. (1992) Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. Proc Natl Acad Sci U S A, 89, 5547-5551.

Graham, F.L. and van der Eb, A.J. (1973) Transformation of rat cells by DNA of human adenovirus 5. Virology, 54, 536-539.

Heldin, C.H., Johnsson, A., Wennergren, S., Wernstedt, C., Betsholtz, C. and Westermark, B. (1986) A human osteosarcoma cell line secretes a growth factor structurally related to a homodimer of PDGF A-chains. Nature, 319, 511-514.

Henke, E. and Bornscheuer, U.T. (1999) Directed evolution of an esterase from Pseudomonas fluorescens. Random mutagenesis by error-prone PCR or a mutator strain and identification of mutants showing enhanced enantioselectivity by a resorufin-based fluorescence assay. Biol Chem, 380, 1029-1033.

Henthorn, P, Zervos, P, Raducha, M, Harris, H, and Kadesch, T. (1988) Expression of a human placental alkaline phosphatase gene in transfected cells: use as a reporter for studies of gene expression Proc Natl Acad Sci U S A 85, 6342-6.

Higgins, D.G., Thompson, J.D. and Gibson, T.J. (1996) Using CLUSTAL for multiple sequence alignments. Methods Enzymol, 266, 383-402.

Himes, S.R. and Shannon, M.F. (2000) Assays for transcriptional activity based on the luciferase reporter gene. Methods Mol Biol, 130, 165-174.

Huberty, CJ (1994) Applied discriminant analysis, Wiley and Sons, New York.

Initiative, AG. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana Nature 408, 796-815.

Izumi, M, and Gilbert, DM. (1999) Homogeneous tetracycline-regulatable gene expression in mammalian fibroblasts J Cell Biochem 76, 280-9.

Jung, R., Soondrum, K. and Neumaier, M. (2000) Quantitative PCR. Clin Chem Lab Med, 38, 833-836.

Kain, SR. (1997) Use of secreted alkaline phosphatase as a reporter of gene expression in mammalian cells Methods Mol Biol 63, 49-60.

Kao, F.T. and Puck, T.T. (1968) Genetics of somatic mammalian cells, VII. Induction and isolation of nutritional mutants in Chinese hamster cells. Proc Natl Acad Sci U S A, 60, 1275-81.

Kellum, R. and Schedl, P. (1992) A group of scs elements function as domain boundaries in an enhancer-blocking assay. Mol Cell Biol, 12, 2424-2431.

Kent, WJ. (2002) BLAT--the BLAST-like alignment tool Genome Res 12, 656-64.

Knofler, M, Meinhardt, G, Bauer, S, Loregger, T, Vasicek, R, Bloor, DJ, Kimber, SJ, and Husslein, P. (2002) Human Hand1 basic helix-loop-helix (bHLH) protein: extra-embryonic expression pattern, interaction partners and identification of its transcriptional repressor domains Biochem J 361, 641-51.

Li, W-H (1997) Molecular Evolution, Sinauer Associates, Sunderland MA., Meinke, DW, Cherry, JM, Dean, C, Rounsley, SD, and Koornneef, M. (1998) Arabidopsis thaliana: a model plant for genome analysis Science 282, 662, 679-82.

Litt, M.D., Simpson, M., Recillas-Targa, F., Prioleau, M.N. and Felsenfeld, G. (2001) Transitions in histone acetylation reveal boundaries of three separately regulated neighboring loci. EMBO J, 20, 2224-2235.

Meyer, P. (2000) Transcriptional transgene silencing and chromatin components Plant Mol Biol 43, 221-34.

Morgenstern, J.P. and Land, H. (1990) Advanced mammalian gene transfer: high titre retroviral vectors with multiple drug selection markers and a complementary helper-free packaging cell line. Nucleic Acids Res, 18, 3587-3596.

Mullen, C.A., Kilstrup, M. and Blaese, R.M. (1992) Transfer of the bacterial gene for cytosine deaminase to mammalian cells confers lethal sensitivity to 5-fluorocytosine: a negative selection system. Proc Natl Acad Sci U S A, 89, 33-37.

Nan, X., Javier Campoy, F., and Bird A. (1997) MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. Cell 88, 471-481.

Petersson, K, Ivars, F, and Sigvardsson, M. (2002) The pT alpha promoter and enhancer are direct targets for transactivation by E box-binding proteins Eur J Immunol 32, 911-20.

Pazin, M.J. and Kadonaga, J.T. (1998) Transcriptional and structural analysis of chromatin assembled in vitro. In Gould, H. (ed.) Chromatin: A Practical Approach. Oxford University Press, Oxford, pp. 172-194.

Pietersen, A. and H.M. Noteborn. (2000) Apoptin. Adv Exp Med Biol 465, 153-161.

Quong, MW, Romanow, WJ, and Murre, C. (2002) E protein function in lymphocyte development Annu Rev Immunol 20, 301-22.

Ruezinsky, D, Beckmann, H, and Kadesch, T. (1991) Modulation of the IgH enhancer's cell type specificity through a genetic switch Genes Dev 5, 29-37.

Saluz, H.P. and Jost, J.P. (1993) Approaches to characterize protein-DNA interactions in vivo. Crit Rev Eukaryot Gene Expr, 3, 1-29.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Plainview NY.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A, 74, 5463-5467.

Simizu, B., Rhim, J.S. and Wiebenga, N.H. (1967) Characterization of the Tacaribe group of arboviruses. I. Propagation and plaque assay of Tacaribe virus in a line of African green monkey kidney cells (Vero). Proc Soc Exp Biol Med, 125, 119-123.

Smith, R.L., Traul, D.L., Schaack, J., Clayton, G.H., Staley, K.J. and Wilcox, C.L. (2000) Characterization of promoter function and cell-type-specific expression from viral vectors in the nervous system. J Virol, 74, 11254-11261.

Stam, M, Viterbo, A, Mol, JN, and Kooter, JM. (1998) Position-dependent methylation and transcriptional silencing of transgenes in inverted T-DNA repeats: implications for posttranscriptional silencing of homologous host genes in plants Mol Cell Biol 18, 6165-77.

Stull, R.A., Hyun, W.C. and Pallavicini, M.G. (2000) Simultaneous flow cytometric analyses of enhanced green and yellow fluorescent proteins and cell surface antigens in doubly transduced immature hematopoietic cell populations. Cytometry, 40, 126-134.

Takada, T., Iida, K., Akasaka, K., Yasue, H., Torii, R., Tsujimoto, G., Taira, M. and Kimura, H. (2000) Evaluation of heterologous insulator function with regard to chromosomal position effect in the mouse blastocyst and fetus. Mol Reprod Dev, 57, 232-237.

Tanaka, S., Livingstone-Zatchej, M. and Thoma, F. (1996) Chromatin structure of the yeast URA3 gene at high resolution provides insight into structure and positioning of nucleosomes in the chromosomal context. J Mol Biol, 257, 919-934.

Thomas, J.O. (1998) Isolation and fractionation of chromatin and linker histones. In Gould, H. (ed.) Chromatin: A Practical Approach. Oxford University Press, Oxford, pp. 1-34.

Tiraby, M., Cazaux, C., Baron, M., Drocourt, D., Reynes, J.P. and Tiraby, G. (1998) Concomitant expression of E. coli cytosine deaminase and uracil phosphoribosyltransferase improves the cytotoxicity of 5-fluorocytosine. FEMS Microbiol Lett, 167, 41-49.

Umana, P, Jean-Mairet, J, and Bailey, JE. (1999) Tetracycline-regulated overexpression of glycosyltransferases in Chinese hamster ovary cells Biotechnol Bioeng 65, 542-9.

van der Vlag, J., den Blaauwen, J.L., Sewalt, R.G., van Driel, R. and Otte, A.P. (2000) Transcriptional repression mediated by polycomb group proteins and other chromatin-associated repressors is selectively blocked by insulators. J Biol Chem, 275, 697-704.

van Helden, J, Andre, B, and Collado-Vides, J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies J Mol Biol 281, 827-42.

van Helden, J, Andre, B, and Collado-Vides, J. (2000) A web site for the computational analysis of yeast regulatory sequences Yeast 16, 177-87.

van Helden, J, Rios, AF, and Collado-Vides, J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads Nucleic Acids Res 28, 1808-18.

Vance, V, and Vaucheret, H. (2001) RNA silencing in plants--defense and counterdefense Science 292, 2277-80.

Wallrath, L.L., Swede, M.J. and Elgin, S.C.R. (1998) Mapping chromatin structure in Drosophila. In Gould, H. (ed.) Chromatin: A Practical Approach. Oxford University Press, Oxford, pp. 59-77.

Weaver, L.S. and Kadan, M.J. (2000) Evaluation of adenoviral vectors by flow cytometry. Methods, 21, 297-312.

Wei, K. and Huber, B.E. (1996) Cytosine deaminase gene as a positive selection marker. J Biol Chem 271, 3812-3816.

Wigler, M, Pellicer, A., Silverstein, S. and Axel, R. (1978). Biochemical transfer of single-copy eukaryotic genes using total cellular DNA as donor. Cell 14, 725-731.

Wigley, P., Becker, C., Beltrame, J., Blake, T., Crocker, L., Harrison, S., Lyons, I., McKenzie, Z., Tearle, R., Crawford, R. and et al. (1994) Site-specific transgene insertion: an approach. Reprod Fertil Dev, 6, 585-588.

Xu, Z.Z., Krougliak, V., Prevec, L., Graham, F.L. and Both, G.W. (1995) Investigation of promoter function in human and animal cells infected with human recombinant adenovirus expressing rotavirus antigen VP7sc. J Gen Virol, 76, 1971-1980.

Yang, TT, Sinai, P, Kitts, PA, and Kain, SR. (1997) Quantification of gene expression with a secreted alkaline phosphatase reporter system Biotechniques 23, 1110-4.

Yin, D.X., Zhu, L., and Schimke R.T. (1996) Tetracyclin-controlled gene expression system achieves high-level and quantitative control of gene expression. Analyt Biochem 235, 195-201.

Zink, D, and Paro, R. (1995) Drosophila Polycomb-group regulated chromatin inhibits the accessibility of a trans-activator to its target DNA Embo J 14, 5660-71.

Table 1. STAR elements improve transgene expression.

Plasmid	Over-expressing clones, %	Fold over-expression (range)	Number of clones
Empty	12	3-11	25
SCS (positive control)	24	3-160	21
STAR-6	62	2-200	26
STAR-3	39	5-820	23
STAR-8	63	7-315	19
STAR-4	31	25-1500	13
STAR-1	57	5-80	23

Expression of the luciferase reporter gene is measured in cell lines containing integrated pSDH plasmids, without ("empty," the negative control) or containing STAR elements (including the positivie control element, SCS from *Drosophila*). The mean expression level of the negative control is defined as the reference level, and clones are considered over-expressing if their expression level is >2-fold above the reference level. The percentage of over-expressing clones for each plasmid and the fold over-expression is reported, along with the number of clones analyzed for each plasmid.

Table 2. Cloned STAR element.

Clone	Chromosomal location[1]	Adjacent genes[2]	Repeat sequence
STAR-1	N.d.		
STAR-2	N.d.		
STAR-3	For 5q33.3 Rev 10q22.2	Chr10 part in Histone. Acetyltransferase gene	
STAR-4	For 1p31.1 Rev 14q24.1	No genes within 10 kb Intron of Regulator of G-protein signalling	83% repetitive LINE2 & LTR ERV_Class1
STAR-5	For 3q13.1 Rev 10q22.1*		
STAR-6	2p21	L5 kb Unknown putative kinase R 20 kb Microtuble associated protein	19% SINE (MIR) 29% LINE
STAR-7	1q32.2		12% Alu 4% MIR (SINE) LINE1 2.5% L31CR1 11.5% MER1 7% Low complex 2%
STAR-8	9q32	ZFP KRAB box containing Zinc Finger Protein	35% ERV_ClassI (LTR) 2% simple repeat
STAR-9	See STAR4		
STAR-10	N.d.		
STAR-11	2p25.1	R 15 kb unknown DNA binding protein inhibitor (Myc type)	12% Alu (SINE) 26% MalRs (LINE)
STAR-12	5q35.3	R 15 kb unknown ADAM TS2 family metallo proteinase	3% Low complexity
STAR-13	See STAR4 and 9		
STAR-14	F N.d. R 20q13.33		
STAR-15	1p36.36	L 6 kb Voltage-gated K channel subunit R 4 kb unknown	14% LTR (MaLRs)
STAR-16	F 8p23.1 R 8p22 etc.		No repeat on sequenced parts
STAR-17	2q31.1	L 6 kb BTEB1 transcription factor R 40 kb HNRNP	10% simple and low complexity

[1]Chromosomal location is determined by BLAST search of DNA sequence data from the STAR clones against the human genome database. The location is given according to standard

nomenclature referring to the cytogenetic ideogram of each chromosome; e.g. 1p2.3 is the third cytogenetic sub-band of the second cytogenetic band of the short arm of chromosome 1 (http://www.ncbi.nlm.nih.gov/Class/MLACourse/Genetics/chrombanding.html). F, forward sequencing reaction result; R, reverse sequencing reaction result.N.d., not yet determined.
[2]Based on Human Genome Map View Build 22 (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch?chr=hum_chr.inf&query April 2001). L, left; R, right.
*Position ambiguous, several hits

Table 3. SINC elements recovered from human chromosome 22 by selection in the pSS vector.

SINC	Length (nt)	Chromosomal location[1]	Remarks
psinks 9	700	22q11.21	Contains LTR; nearest gene ZNF 74, an RNA binding protein. LTR very repetitive.
psinks 12	750	22q12.3	Located in intron of acetylglucosaminyl-transferase-like protein (664 kb) implicated in tumour formation.
psinks 19	600	22q13.1	Located in intron of calcium channel, almost exclusively expressed in brain.
psinks 28	950	22q13.31	Located in intron of kidney protein of unknown function. Contains SINEelement
psinks 30	700	22q13.33	Contains part of SINE.
psinks 35	650	22q11.21	Covers exon for solute carrier. (Nuclear gene for mitochondrion).

[1]Chromosomal location is determined by BLAST search of DNA sequence data from the STAR clones against the human genome database. The location is given according to standard nomenclature referring to the cytogenetic ideogram of each chromosome; e.g. 1p2.3 is the third cytogenetic sub-band of the second cytogenetic band of the short arm of chromosome 1 (http://www.ncbi.nlm.nih.gov/Class/MLACourse/Genetics/chrombanding.html).

Table 4 A: Sequence of various star elements in one strand (forward) or the opposite strand (reverse).

STAR3 forward

ACGTNCTAAGNAAACCATTATTATCATGACATTAACCTATAAAAATAGGC
GTATCACGAGGCCCTTTCGTCTTCACTCGAGCGGCCAGCTTGGATCTCGA
GTACTGAAATAGGAGTAAATCTGAAGAGCAAATAAGATGAGCCAGAAAAC
CATGAAAAGAACAGGGACTACCAGTTGATTCCACAAGGACATTCCCAAGG
TGAGAAGGCCATATACCTCCACTACCTGAACCAATTCTCTGTATGCAGATT
TAGCAAGGTTATAAGGTAGCAAAAGATTAGACCCAAGAAAATAGAGAACT
TCCAATCCAGTAAAAATCATAGCAAATTTATTGATGATAACAATTGTCTCC
AAAGGAACCAGGCAGAGTCGTGCTAGCAGAGGAAGCACGTGAGCTGAAA
ACAGCCAAATCTGCTTTGTTTTCATGACACAGGAGCATAAAGTACACACCA
CCAACTGACCTATTAAGGCTGTGGTAAACCGATTCATAGAGAGAGGTTCT
AAATACATTGGTCCCTCATAGGCAAACCGCAGTTCACTCCGAACGTAGTC
CCTGGAAATTTGATGTCCAGNATAGAAAAGCANAGCAGNCNNNNNNTAT
ANATNNNGNTGANCCANATGNTNNCTGNNC

STAR3 reverse

GAGCTAGCGGCGCGCCAAGCTTGGATCCCGCCCCGCCCCCTCCGCCCTCG
AGCCCCGCCCCTTGCCCTAGAGGCCCTGCCGAGGGGCGGGGCCTGTCCC
TCCTCCCCTTTCCCCCGCCCCCTACCGTCACGCTCAGGGGCAGCCTGACC
CCGAGCGGCCCCGCGGTGACCCTCGCGCAGAGGCCTGTGGGAGGGGCGT
CGCAAGCCCCTGAATCCCCCCCCGTCTGTTCCCCCCTCCCGCCCAGTCTC
CTCCCCCTGGGAACGCGCGGGGTGGGTGACAGACCTGGCTGCGCGCCAC
CGCCACCGCGCCTGCCGGGGGCGCTGCCGCTGCCTGAGAAACTGCGGCT
GCCGCCTGGAGGAGGTGCCGTCGCCTCCGCCACCGCTGCCGCCGCCGCC
AGGGGTAGGAGCTAAGCCGCCGCCATTTTGTGTCCCCCTGTTGTTGTCGT
TGACATGAATCCGACATGACACTGATTACAGCCCAATGGAGTCTCATTAA
ACCCGAGTCGCGGTCCCGCCCCGCCGCTGCTCCATTGGAGGAGACCAAAG

ACACTTAAGGCCACCCGTTGGCCTACGGGTCTGTCTGTCACCCACTCACT
AACCACTCTGCAGCCCATTGGGGCAGGTTCCTGCCGGTCATNTCGCTTCC
AATAAACACACCCCTTCGACCCCATNATTCCCCCCCTTCGGGAACCACCC
CCGGGGGAGGGGTCCACTGGNCAATACCAATTNAANAGAACCGCTNGGG
TCCGCCTNTTTNCGGGCNCCCTATTGGGTT

STAR4 forward

GGGGAGGATTCTTTTGGCTGCTGAGTTGAGATTAGGTTGAGGGTAGTGAA
GGTAAAGGCAGTGAGACCACGTAGGGGTCATTGCAGTAATCCAGGCTGG
AGATGATGGTGGTTCAGTTGGAATAGCAGTGCATGTGCTGTAACAACCTC
AGCTGGGAAGCAGTATATGTGGCGTTATGACCTCAGCTGGAACAGCAATG
CATGTGGTGGTGTAATGACCCCAGCTGGGTAGGGTGCATGTGATGGAACA
ACCTCAGCTGGGTAGCAGTGTACTTGATAAAATGTTGGCATACTCTACATT
TGTTATGAGGGTAGTGCCATTAAATTTCTCCACAAATTGGTTGTCACGTAT
GAGTGAAAAGAGGAAGTGATGGAAGACTTCAGTGCTTTTGGCCTGAATAA
ATAGAAGACGTCATTTTCAGTAATGGAGACAGGGAAGACTAANGNAGGGT
GGATTCAGTAGAGCAGGTGTTCAGTTTTGAATATGATGAACTCTGAGAGA
GGAAAAACTTTTTCTACCTCTTAGTTTTTGNGNCTGGACTTAANATTAAAG
GACATANGACNGAGANCAGACCAAATNTGCGANGTTTTTATATTTTACTT
GCNGAGGGAATTTNCAAGAAAAAGAAGACCCAANANCCATTGGTCAAAA
CTATNTGCCTTTTAANAAAAAGANAATTACAATGGANANANAAGTGTTGN
CTNGGCAAAAATTGGG

STAR4 reverse

GGATTNGAGCTAGCGGCGCGCCAAGCTTGGATCTTAGAAGGACAGAGTG
GGGCATGGAAATGCACCACCAGGGCAGTGCAGCTTGGTCACTGCCAGCTC
CNCTCATGGGCAGAGGGCTGGCCTCTTGCAGCCGACCAGGCACTGAGCG
CCATCCCAGGGCCCTCGCCAGCCCTCAGCAGGGCCAGGACACACAAGCCT
TTGACTTCCTCCTGTCACTGCTGCTGCCATTCCTGTTTTGTGGTCATCACT
CCTTCCCTGTCCTCAGACTGCCCAGCACTCAAGGATGTCCTGTGGTGGCA

TCAGACCATATGCCCCTGAANAGGAGTGAGTTGGTGTTTTTTGCCGCGCC
CANAGAGCTGCTGTCCCCTGAAAGATGCAAGTGGGAATGATGATGNTCAC
CATCNTCTGACACCAAGCCCTTTGGATAGAGGCCCCAACAGTGAGGATGG
GGCTGCACTGCATTGCCAAGGCAACTCTGTNNTGACTGCTACANGACANT
CCCAGGACCTGNGAAGNNCTATANATNTGATGCNAGGCACCT

STAR6 forward

CCACCACAGACATCCCCTCTGGCCTCCTGAGTGGTTTCTTCAGCACAGCTT
CCAGAGCCAAATTAAACGTTCACTCTATGTCTATAGACAAAAAGGGTTTTG
ACTAAACTCTGTGTTTTAGAGAGGGAGTTAAATGCTGTTAACTTTTTAGGG
GTGGGCGAGAGGAATGACAAATAACAACTTGTCTGAATGTTTTACATTTC
TCCCCACTGCCTCAAGAAGGTTCACAACGAGGTCATCCATGATAAGGAGT
AAGACCTCCCAGCCGGACTGTCCCTCGGCCCCCAGAGGACACTCCACAGA
GATATGCTAACTGGACTTGGAGACTGGCTCACACTCCAGAGAAAAGCATG
GAGCACGAGCGCACAGAGCANGGGCCAAGGTCCCAGGGACNGAATGTCT
AGGAGGGAGATTGGGGTGAGGGTANTCTGATGCAATTACTGNGCAGCTC
AACATTCAAGGGAGGGGAAGAAAGAAACNGTCCCTGTAAGTAAGTTGTNC
ANCAGAGATGGTAAGCTCCAAATTTNAACTTTGGCTGCTGGAAAGTTTNN
GGGCCNANANAANAAACANAAANATTTGAGGTTTANACCCACTAACCCN
TATNANTANTTATTAATACCCCTAATTANACCTTGGATANCCTTAAAATAT
CNTNTNAAACGGAACCCTCNTTCCCNTTTNNAAATNNNAAAGGCCATTN
NGNNCNAGTAAAAATCTNNNTTAAGNNNTGGGCCCNAACAAACNTNTTC
CNAGACACNTTTTTTNTCCNGGNATTTNTAATTTATTTCTAANCC

STAR6 reverse

ATCGTGTCCTTTCCAGGGACATGGATGAAGCTGGAAGCCATCATCCTCAG
CAAACTAACACAGGAACAGAAAACCAAATACCACATGTTCTCACTCATAAG
TGGGAGCTGAACAGTGAGAACACATGGACACAGGGAGGGGAACATCACA
CACCAAGGCCTGTCTGGTGTGGGGAGGGGAGGGAGAGCATCAGGACAAA

TAGCTAATGCATGTGGGGCTTAAACCTAGATGACGGGTTGATAGGTGCAG
CAATCCACTATGGACACATATACCTATGTAACAACCCNACCTTNTTGACAT
GTATCCCAGAACTTAAAGGAAAATAAAAATTAAAAAAAATTNCCCTGGAA
TAAAAAAGAGTGTGGACTTTGGTGAGATN

STAR8 forward

GGATCACCTCGAAGAGAGTCTAACGTCCGTAGGAACGCTCTCGGGTTCAC
AAGGATTGACCGAACCCCAGGATACGTCGCTCTCCATCTGAGGCTTGNTC
CAAATGGCCCTCCACTATTCCAGGCACGTGGGTGTCTCCCCTAACTCTCC
CTGCTCTCCTGAGCCCATGCTGCCTATCACCCATCGGTGCAGGTCCTTTCT
GAANAGCTCGGGTGGATTCTCTCCATCCCACTTCCTTTCCCAAGAAAGAA
GCCACCGTTCCAAGACACCCAATGGGACATTCCCNTTCCACCTCCTTNTC
NAAAGTTNGCCCAGGTGTTCNTAACAGGTTAGGGAGAGAANCCCCCAGG
TTTNAGTTNCAAGGCATAGGACGCTGGCTTGAACACACACACACNCTC

STAR8 reverse

GGATCCCGACTCTGCACCGCAAACTCTACGGCGCCCTGCAGGACGGCGGC
CTCCTGCCGCTTGGACGCCAGNCAGGAGCTCCCCGGCAGCAGCAGAGCA
GAAAGAAGGATGGCCCCGCCCCACTTCGCCTCCCGGCGGTCTCCCTCCCG
CCGGCTCACGGACATAGATGGCTGCCTAGCTCCGGAAGCCTAGCTCTTGT
TCCGGGCATCCTAAGGAAGACACGGTTTTTCCTCCCGGGGCCTCACCACA
TCTGGGACTTTGACGACTCGGACCTCTCTCCATTGAATGGTTGCGCGTTC
TCTGGGAAAG

STAR18 forward

TGGATCCTGCCGCTCGCGTCTTAGTGTTTCTCCCTCAAGACTTTCCTTCTG
TTTTGTTGTCTTGTGCAGTATTTTACAGCCCCTCTTGTGTTTTTCTTTATTT
CTCGTACACACACGCAGTTTTAAGGGTGATGTGTGTATAATTAAAAGGAC
CCTTGGCCCATACTTTCCTAATTCTTTAGGGACTGGGATTGGGTTTGACTG
AAATATGTTTTGGTGGGGATGGGACGGTGGACTTCCATTCTCCCTAAACT

GGAGTTTTGGTCGGTAATCAAAACTAAAAGAAACCTCTGGGAGACTGGAA
ACCTGATTGGAGCACTGAGGAACAAGGGAATGAAAAGGCAGACTCTCTGA
ACGTTTGATGAAATGGACTCTTGTGAAAATTAACAGTGAATATTCACTGTT
GCACTGTACGAAGTCTCTGAAATGTAATTAAAAGTTTTTATTGAGCCCCCG
AGCTTTGGCTTGCGCGTATTTTTCCGGTCGCGGACATCCCACCGCGCAGA
GCCTCGCCTCCCCGCTGNCCTCAGCTCCGATGACTTCCCCGCCCCCGCCC
TGCTCGGTGACAGACGTTCTACTGCTTCCAATCGGAGGCACCCTTCGCGG

STAR18 reverse

TGGATCCTGCCGCTCGCGTCTTAGTGTTTCTCCCTCAAGACTTTCCTTCTG
TTTTGTTGTCTTGTGCAGTATTTTACAGCCCCTCTTGTGTTTTTCTTTATTT
CTCGTACACACACGCAGTTTTAAGGGTGATGTGTGTATAATTAAAAGGAC
CCTTGGCCCATACTTTCCTAATTCTTTAGGGACTGGGATTGGGTTTGACTG
AAATATGTTTTGGTGGGGATGGGACGGTGGACTTCCATTCTCCCTAAACT
GGAGTTTTGGTCGGTAATCAAAACTAAAAGAAACCTCTGGGAGACTGGAA
ACCTGATTGGAGCACTGAGGAACAAGGGAATGAAAAGGCAGACTCTCTGA
ACGTTTGATGAAATGGACTCTTGTGAAAATTAACAGTGAATATTCACTGTT
GCACTGTACGAAGTCTCTGAAATGTAATTAAAAGTTTTTATTGAGCCCCCG
AGCTTTGGC

Table 4 B: the sequence of various sinc elements

PSINKS 9

GATCAGGA TAATAAGTAC GCTGGGAAGA CAACAAAATG
ATTTAAATCT TAGACAAGTC ATTCTAGGTG TCTCCACTGT
TTCAGTTCTT GCATTCATTC TTGTGGTATC TTTTCCCTTT
TACCAATAAA AAAGCTCCCT GACATCACAT TGTGGCAGTC
CCCATGGTTT GCCGCAGTTA CTGCGGGACT GAACGAAGGA
GGACGAATGA AGAAATGAAA ACCAAGGAAA AAAGGAGCTG
TTTAAAGAAG GGTCCAGGGA AGAAGAAGAG GGCTCCCAGC
TTCTAGTGAG CAAGGGCAGC AGCCCTGAGC TTCTACAGCC
CTTCATATTT ATTGAGTAGA AAGAGCAGGG AGCAGGAGGT
AATGATTGGT CAGCTTCTCA ATTGATCACA GGTTCACATT
ATTGCTAACA GATTTCACAT GTGCCTAATC TCAAGAAACG
CCGCGCCTGG GGCATGACTG CCCTCAGCAT TCCCTCTGGG
TGGCAGACGC AGTTTGCCAA CATTCTGCAT TCATGAGAAC
AGTTTACTGT TTACTCATAT AACCTCCAGT GGTACACCGA
GTTGATC

PSINKS 12

GATCTAA TTTCTCTGTA TTTAATTCCC ATGTCTATTT
TGTCTATTTT CAAGATTGAT TTACATTGCA GGTTCCGATG
CAACCACTGA CTTACATTGC AGGTTCTAAT GTAACCACTG
TCCTTAACGA GTACATAGAT TTGTTTCCTT CTCTCCAGGA
GCATGAGATT TGTTGCCTCC AGGAAAGGCA ACAAATCTAC
TATTCCTTA AGGACAGTGG TTCTCAAAGG ATTGTCCTGG
GAACAGCAGC ATCACCTACA CAGTAGTTAG AAATGCACAT
TCTGAGGCCT CCCAAGACCT GCTAACTCAG ACACTTGGGG

AGAAGAAGGG GTTCCAACAA GCCTTCTAGG TCATTCTGAT
GCATGCTGGA GTTTGAGAAT CGATGCTCTA GGAAAAACAC
CAGTACTAT CTACCATCAA CTTGACCACT CAAGTGTCAC
CATTCACTGA AGTTTAACTA CAATGTCCAG AGAATTAATT
GTGTACCAGG CACTATGCGG AAGGCTGAAT GCTGCCTCAC
AATCCANAGT GGTATGTGTG TAAATGACTA AATAAAATGC
AAAATGGGAT GACATG

PSINKS 19

G ATCCTCCATC TGCTCCACCC ACTTCCATGT AAGTGATCCT
GGGCTGATCA CTTCCTCTCT CTAGACTTCG TTTCTTTTTT
TTCTTTTTTA GACCGAGTCT CACTCTGTCA CCCAGGCTGG
AGTGCAGTGG TGAGATCTTG GCTCACTGCA ACCTCCACCT
CCTGGGTTCA AGCAATTCTC CTGTCTCAGC CTCCTGAGTA
GATAGGACTA TAGGTGCACA CCACCATACC TAGCTAATTT
TTGTTTTTTT AGTAGAGATG GGGTTTCACC ATATTGGTCA
GCCTGTTCTC AAACTCCTGA CCTCAGGTGA TCCACCCACC
TCAGCCTCCC AAAGTGCTGG GATTACAGGT GTGAGCCACC
GAGCCGGGCT GCCCTTCTCT GGACTTTGAT TTCCTCATCT
ATAAAACAGA CAACAATCCC TACTATGACC ATCCAGAAGG
GTTAATCTAT GCTTCATTGC AATCCTAATC AAAAATCCCA
ACATTTTGGC CGTGGAGCCT GCCCAGATGG TTCTAGGATT
TATTTGGATG GGAAAATAGT CAAGACAAGC TT

PSINKS 28

GAT CATGGAGGGA GAGAACAACC AACCACACAC
TGACTGGTCA CCCCTGAAGT TCACAGCCAC TACCCTGTAG
AGGCCCCGAG GTTGCCGGCA AGCCCAGTAT ACTTCCATCT

AAACTCCCCT TGCACCTGCT CCTCCTGTTC CAGACAATGA
GCTGTAACAC GCACATCCAC ACCACACATC ACCCACAGCA
GGGGCAGGAG GCAGCTAAGC ATGGGCTTCA GAGTCCTCCC
ACCAGCAGCG CCTACCAGCT ACAAGCCTGA CGTCTCTGTG
TGTGTGTGTA AATTTCACTA AATATTTCTT CCTTTGTTTT
TTAAAAATTT ACATGAAATG CACATTTTTG CTGTGACAGA
AGCATGTAAC TGTGATCCTA ACACACCTAC TCCTCCGCCT
TTTACTGCCG TCTGCTTCCC TCTCTTCTCC ACGCCCACTC
GACTGCAGTA TCGATGCCAA CAACATGATG TGTGTCCTTC
CATGTTTCCC TGCTCATGCA TTCGCATGTA AGCCACCGCA
CATGTCACTG TATGTACACA CAGGGGATTC TGAGGCCAAT
GTTTTACAAG GATTACGTTA TACACCCTTT TCTGCAGTGA
GTTTTTCCCA GGCAACCTCC CAGGCCCCAT GGTGTAGCTC
TGGGTCAATC CTTTTTTTTT TTTTTGGAGA CAGAGTCTCA
CTCTGTCGCC CAGGCTGGAG TGCAGTGGTG CAATTTGGGC
TCACTGCAAC CTCCGCCTCC CGGGTTCAAG CGATTCTCCT
GCCTCAGCCT CCTGAGTAGC TGACATTACA AGCGCGCACT
ACCACACCCG GCTAATTTTT GTATTTTTAG TAGAGATACA
GTTTCACCAT GTTGGTCAGG CTGGTCTTGA ACTCCTGAGC
TCGTGATC

PSINKS 30

GATCC ACCCGCCTCG GACTCCCAAA GTGCTGGGAT
TACAGGTGTG AGCCACTGTG CTTGGCCCGA ATCAGGAATA
ATTCTGATGG CTAAGGAAGA CAGCTTCCGA GAGAGTAGGA
GAAAGGGCAC AGGATTCCAG GCAGAAGGCC CATCTAGGGC
AAAGGCGAAG GTGTGGCTCA GCCTGCCTCC TTTGGGGAAT
GGCGAGTGTG TTCTGGGCTC AGGGTTCTTG GTAAGGGACA
GAGAAGACTC GGGAAAGATC AGTTGAGCTG GAATGTGCAG

GCTCTTGAGT ACCCTGCTCA GGAGCTGGAG GTGGGCTACC
CTGCAAACTC CAGGCCATGA AGCCCAGGAA GATGTCAGGC
TGGTCTTCCC ATGCCCTTGT GTATCTGAGA CCAACTGTCA
CTAAATGTTT CCTTTACGCC CTGGACACAC AGCTAGACTC
TACTTCTCAG ATTCTCTTGA AATACAAGTC TTTAGCCAGA
GGGTGTGGAG GGAAATGCTG TGTATCACTT TGAGGTTGAG
GCCATCAAAG CCTCCCACAG GTGGCCCCCT CTTTCTCTCC
CCACGTACTT ATGATGTTGA TGCCCAAGGC AGCTTGAGTA
CTACCTGCTG AAGGCAGGGC CTCTGTCACC ATAGATC

PSINKS 35

GATCCAC CTGCCTCGGC CTCCCAAAGT GCTGGGATTA
CAGGCATGAG CCACCATGCC TGGCCAAAAA CTTCTACCTG
CTTGGAAAGT TGACTGGTCA CACAGCCTAG CAAATGAGGT
TGGGATGTGG GATGTGCCTG GTTCCAATCC CAGCCCTTTA
CTGTTCCCAT AGGAGGTGGG GACAGGCCTC ACCCAGGCGT
CCAGCATCCT GCAGCTGAAT CTTGAGCATT TCCATGGGAC
AGGTCACCAC GACCTGGCAC ATCCCAGCCC CACACCCGGC
AAGCATCTCC ATCTTCAGGT TCCGCTGCAT CCTATGGGAA
CAGGCGTCAG GCTCCTTCAG CCGCAGGCCA CAGGCCTGCC
CTGGTGCAGC TGCCCTCTTG TGAGAGGGGG ACTTTCCCTG
GATGGCACCC GTGGCTGCCA CTCACCCAGC TGGTCAAGTC
ATCAGCTAGC CCTTAGGTGT GGTCTCTGTA CGGACAGGGG
ACTAAGTTTA AAACAAAGCC TGCTAGGGAG GTAGCACCGC
ATGGAAGCTG AAACAGTGAC AGAGAAAACT ACCCAGACCA
GGCGTTGTCC TTGATC

Table 5. Oligonucleotides used for polymerase chain reactions (PCR primers) or DNA mutagenesis

Number	Sequence
C65	AACAAGCTTGATATCAGATCTGCTAGCTTGGTCGAGCTGATACTTCCC
C66	AAACTCGAGCGGCCGCGAATTCGTCGACTTTACCACTCCCTATCAGTGATAGAG
C67	AAACCGCGGCATGGAAGACGCCAAAAAACATAAAGAAAGG
C68	TATGGATCCTAGAATTACACGGCGATCTTTCC
C81	AAACCATGGCCGAGTACAAGCCCACGGTGCGCC
C82	AAATCTAGATCAGGCACCGGGCTTGCGGGTCATGC
C85	CATTTCCCCGAAAAGTGCCACC
D30	TCACTGCTAGCGAGTGGTAAACTC
D41	GAAGTCGACGAGGCAGGCAGAAGTATGC
D42	GAGCCGCGGTTTAGTTCCTCACCTTGTCG
D51	TCTGGAAGCTTTGCTGAAGAAAC
D89	GGGCAAGATGTCGTAGTCAGG
D90	AGGCCCATGGTCACCTCCATCGCTACTGTG
D91	CTAATCACTCACTGTGTAAT
D93	AATTACAGGCGCGCC
D94	AATTGGCGCGCCTGT
D95	TGCTTTGCATACTTCTGCCTGCCTC
E12	TAGGGGGGATCCAAATGTTC
E13	CCTAAAAGAAGATCTTTAGC
E14	AAGTGTTGGATCCACTTTGG
E15	TTTGAAGATCTACCAAATGG
E16	GTTCGGGATCCACCTGGCCG
E17	TAGGCAAGATCTTGGCCCTC
E18	CCTCTCTAGGGATCCGACCC
E19	CTAGAGAGATCTTCCAGTAT
E20	AGAGTTCCGGATCCGCCTGG
E21	CCAGGCAGACTCGGAACTCT
E22	TGGTGAAACCGGATCCCTAC
E23	AGGTCAGGAGATCTAGACCA
E25	CCATTTTCGCTTCCTTAGCTCC

E42	CGATGTAACCCACTCGTGCACC
E57	AGAGATCTAGGATAATTTCG
E92	AGGCGCTAGCACGCGTTCTACTCTTTTCCTACTCTG
E93	GATCAAGCTTACGCGTCTAAAGGCATTTTATATAG
E94	AGGCGCTAGCACGCGTTCAGAGTTAGTGATCCAGG
E95	GATCAAGCTTACGCGTCAGTAAAGGTTTCGTATGG
E96	AGGCGCTAGCACGCGTTCTACTCTTTCATTACTCTG
E97	CGAGGAAGCTGGAGAAGGAGAAGCTG
E98	CAAGGGCCGCAGCTTACACATGTTC
D58	CCAAGTTGACCAGTGCC
D80	GTTCGTGGACACGACCTCCG
D70	TACAAGCCAACCACGGCCT
D71	CGGAAGTGCTTGACATTGGG

Table 6. STAR elements of the invention, including genomic location and length

STAR	Location[1]	Length[2]
1	2q31.1	750
2	7p15.2	916
3[3]	15q11.2 and 10q22.2	2132
4	1p31.1 and 14q24.1	1625
5[4]	20q13.32	1571
6	2p21	1173
7	1q34	2101
8	9q32	1839
9[4]	10p15.3	1936
10	Xp11.3	1167
11	2p25.1	1377
12	5q35.3	1051
13[4]	9q34.3	1291
14[4]	22q11.22	732
15	1p36.31	1881
16	1p21.2	1282
17	2q31.1	793
18	2q31.3	497
19	6p22.1	1840
20	8p13.3	780
21	6q24.2	620
22	2q12.2	1380
23	6p22.1	1246
24	1q21.2	948
25[5]	1q21.3	1067
26	1q21.1	540
27	1q23.1	1520
28	22q11.23	961
29	2q13.31	2253
30	22q12.3	1851
31	9q34.11 and 22q11.21	1165
32	21q22.2	771

STAR	Location[1]	Length[2]
33	21q22.2	1368
34	9q34.14	755
35	7q22.3	1211
36	21q22.2	1712
37	22q11.23	1331
38	22q11.1 and 22q11.1	~1000
39	22q12.3	2331
40	22q11.21	1071
41	22q11.21	1144
42	22q11.1	735
43	14q24.3	1231
44	22q11.1	1591
45	22q11.21	1991
46	22q11.23	1871
47	22q11.21	1082
48	22q11.22	1242
49	Chr 12 random clone, and 3q26.32	1015
50	6p21.31	2361
51	5q21.3	2289
52	7p15.2	1200
53	Xp11.3	1431
54	4q21.1	981
55	15q13.1	501
56	includes 3p25.3	741
57	4q35.2	1371
58	21q11.2	1401
59	17 random clone	872
60	4p16.1 and 6q27	2068
61	7p14.3 and 11q25	1482
62	14q24.3	1011
63	22q13.3	1421
64	17q11.2	1414
65	7q21.11=28.4	1310
66	20q13.33 and 6q14.1	~2800

[1]Chromosomal location is determined by BLAST search of DNA sequence data from the STAR elements against the human genome database. The location is given according to standard nomenclature referring to the cytogenetic ideogram of each chromosome; e.g. 1p2.3 is the third cytogenetic sub-band of the second cytogenetic band of the short arm of chromosome 1 (http://www.ncbi.nlm.nih.gov/Class/MLACourse/Genetics/chrombanding.html). In cases where the forward and reverse sequencing reaction identified DNAs from different genomic loci, both loci are shown.

[2]Precise lengths are determined by DNA sequence analysis; approximate lengths are determined by restriction mapping.

[3]Sequence and location of STAR3 has been refined since assembly of Tables 2 and 4.

[4]The STARs with these numbers in Tables 2 and 4 have been set aside (hereafter referred to as "oldSTAR5" etc.) and their numbers assigned to the STAR elements shown in the DNA sequence appendix. In the case of oldSTAR5, oldSTAR14, and oldSTAR16, the cloned DNAs were chimeras from more than two chromosomal locations; in the case of oldSTAR9 and oldSTAR13, the cloned DNAs were identical to STAR4.

[5]Identical to Table 4 "STAR18".

Table 7. STAR elements convey stability over time on transgene expression[1]

	Cell Divisions[2]	Luciferase Expression[3]
STAR6 plus puromycin	42	18,000
	60	23,000
	84	20,000
	108	16,000
STAR6 without puromycin[4]	84	12,000
	108	15,000
	144	12,000

[1]Plasmid pSDH-Tet-STAR6 was transfected into U-2 OS cells, and clones were isolated and cultivated in doxycycline-free medium as described in Example 1. Cells were transferred to fresh culture vessels weekly at a dilution of 1:20.

[2]The number of cell divisions is based on the estimation that in one week the culture reaches cell confluence, which represents ~6 cell divisions.

[3]Luciferase was assayed as described in Example 1.

[4]After 60 cell divisions the cells were transferred to two culture vessels; one was supplied with culture medium that contained puromycin, as for the first 60 cell divisions, and the second was supplied with culture medium lacking antibiotic.

Table 8. Human STAR elements and their putative mouse orthologs and paralogs

SEQ:ID	STAR	Human[1]	Mouse[2]	Similarity[3]
1	1	2q31.1	2D	600 bp 69%
2	2	7p15.2	6B3	909 bp 89%
3	3a	5q33.3	11B2	248 bp 83%
4	3b	10q22.2	14B	1. 363 bp 89% 2. 163 bp 86%
5	6	2p21	17E4	437 bp 78%
6	12	5q35.3	11b1.3	796 bp 66%
7	13	9q34.3	2A3	753 bp 77%
8	18	2q31.3	2E1	497 bp 72%
9	36	21q22.2	16C4	166 bp 79%
10	40	22q11.1	6F1	1. 270 bp 75% 2. 309 bp 70%
11	50	6p21.31	17B1	1. 451 bp 72% 2. 188 bp 80% 3. 142 bp 64%
12	52	7p15.2	6B3	1. 846 bp 74% 2. 195 bp 71%
13	53	Xp11.3	XA2	364 bp 64%
14	54	4q21.1	5E3	1. 174 bp 80% 2. 240 bp 73% 3. 141 bp 67% 4. 144 bp 68%
15	61a	7p14.3	6B3	188 bp 68%

[1]Cytogenetic location of STAR element in the human genome.

[2]Cytogenetic location of STAR element ortholog in the mouse genome.

[3]Length of region(s) displaying high sequence similarity, and percentage similarity. In some cases more than one block of high similarity occurs; in those cases, each block is described separately. Similarity <60% is not considered significant.

Table 9. Oligonucleotide patterns (6 base pairs) over-represented in STAR elements.

The patterns are ranked according to significance coefficient. These were determined using RSA-Tools with the sequence of the human genome as reference. Patterns that comprise the most discriminant variables in Linear Discriminant Analysis are indicated with an asterisk.

Number	Oligonucleotide sequence	Observed occurrences	Expected occurrences	Significance coefficient	Number of matching STARs
1	CCCCAC	107	49	8.76	51
2	CAGCGG	36	9	7.75	23
3	GGCCCC	74	31	7.21	34
4	CAGCCC	103	50	7.18	37
5	GCCCCC	70	29	6.97	34
6	CGGGGC	40	12	6.95	18
7	CCCCGC	43	13	6.79	22
8	CGGCAG	35	9	6.64	18
9	AGCCCC	83	38	6.54	40
10	CCAGGG	107	54	6.52	43
11	GGACCC *	58	23	6.04	35
12	GCGGAC	20	3	5.94	14
13	CCAGCG	34	10	5.9	24
14	GCAGCC	92	45	5.84	43
15	CCGGCA	28	7	5.61	16
16	AGCGGC	27	7	5.45	17
17	CAGGGG	86	43	5.09	43
18	CCGCCC	43	15	5.02	18
19	CCCCCG	35	11	4.91	20
20	GCCGCC	34	10	4.88	18
21	GCCGGC	22	5	4.7	16
22	CGGACC	19	4	4.68	14
23	CGCCCC	35	11	4.64	19
24	CGCCAG	28	8	4.31	19
25	CGCAGC	29	8	4.29	20
26	CAGCCG	32	10	4	24
27	CCCACG	33	11	3.97	26
28	GCTGCC	78	40	3.9	43
29	CCCTCC	106	60	3.87	48
30	CCCTGC *	92	50	3.83	42
31	CACCCC	77	40	3.75	40
32	GCGCCA	30	10	3.58	23
33	AGGGGC	70	35	3.55	34

34	GAGGGC	66	32	3.5	40
35	GCGAAC	14	2	3.37	13
36	CCGGCG	17	4	3.33	12
37	AGCCGG	34	12	3.29	25
38	GGAGCC	67	34	3.27	40
39	CCCCAG	103	60	3.23	51
40	CCGCTC	24	7	3.19	19
41	CCCCTC	81	44	3.19	43
42	CACCGC	33	12	3.14	22
43	CTGCCC	96	55	3.01	42
44	GGGCCA	68	35	2.99	39
45	CGCTGC	28	9	2.88	22
46	CAGCGC	25	8	2.77	19
47	CGGCCC	28	10	2.73	19
48	CCGCCG	19	5	2.56	9
49	CCCCGG	30	11	2.41	17
50	AGCCGC	23	7	2.34	17
51	GCACCC	55	27	2.31	38
52	AGGACC	54	27	2.22	33
53	AGGGCG	24	8	2.2	18
54	CAGGGC	81	47	2.18	42
55	CCCGCC	45	21	2.15	20
56	GCCAGC	66	36	2.09	39
57	AGCGCC	21	6	2.09	18
58	AGGCCC	64	34	2.08	32
59	CCCACC	101	62	2.05	54
60	CGCTCA	21	6	2.03	17
61	AACGCG	9	1	1.96	9
62	GCGGCA	21	7	1.92	14
63	AGGTCC	49	24	1.87	36
64	CCGTCA	19	6	1.78	14
65	CAGAGG	107	68	1.77	47
66	CCCGAG	33	14	1.77	22
67	CCGAGG	36	16	1.76	25
68	CGCGGA	11	2	1.75	8
69	CCACCC	87	53	1.71	45
70	CCTCGC	23	8	1.71	20
71	CAAGCC	59	32	1.69	40
72	TCCGCA	18	5	1.68	17
73	CGCCGC	18	5	1.67	9
74	GGGAAC	55	29	1.63	39
75	CCAGAG	93	58	1.57	49

76	CGTTCC	19	6	1.53	16
77	CGAGGA	23	8	1.5	19
78	GGGACC	48	24	1.48	31
79	CCGCGA	10	2	1.48	8
80	CCTGCG	24	9	1.45	17
81	CTGCGC	23	8	1.32	14
82	GACCCC	47	24	1.31	33
83	GCTCCA	66	38	1.25	39
84	CGCCAC	33	15	1.19	21
85	GCGGGA	23	9	1.17	18
86	CTGCGA	18	6	1.15	15
87	CTGCTC	80	49	1.14	50
88	CAGACG	23	9	1.13	19
89	CGAGAG	21	8	1.09	17
90	CGGTGC	18	6	1.06	16
91	CTCCCC	84	53	1.05	47
92	GCGGCC	22	8	1.04	14
93	CGGCGC	14	4	1.04	13
94	AAGCCC *	60	34	1.03	42
95	CCGCAG	24	9	1.03	17
96	GCCCAC	59	34	0.95	35
97	CACCCA	92	60	0.93	49
98	GCGCCC	27	11	0.93	18
99	ACCGGC	15	4	0.92	13
100	CTCGCA	16	5	0.89	14
101	ACGCTC	16	5	0.88	12
102	CTGGAC	58	33	0.88	32
103	GCCCCA	67	40	0.87	38
104	ACCGTC	15	4	0.86	11
105	CCCTCG	21	8	0.8	18
106	AGCCCG	22	8	0.79	14
107	ACCCGA	16	5	0.78	13
108	AGCAGC	79	50	0.75	41
109	ACCGCG	14	4	0.69	7
110	CGAGGC	29	13	0.69	24
111	AGCTGC	70	43	0.64	36
112	GGGGAC	49	27	0.64	34
113	CCGCAA	16	5	0.64	12
114	CGTCGC	8	1	0.62	6
115	CGTGAC	17	6	0.57	15
116	CGCCCA	33	16	0.56	22
117	CTCTGC	97	65	0.54	47

118	AGCGGG	21	8	0.52	17
119	ACCGCT	15	5	0.5	11
120	CCCAGG	133	95	0.49	58
121	CCCTCA	71	45	0.49	39
122	CCCCCA *	77	49	0.49	42
123	GGCGAA	16	5	0.48	14
124	CGGCTC	29	13	0.47	19
125	CTCGCC	20	8	0.46	17
126	CGGAGA	20	8	0.45	14
127	TCCCCA	95	64	0.43	52
128	GACACC	44	24	0.42	33
129	CTCCGA	17	6	0.42	13
130	CTCGTC	17	6	0.42	14
131	CGACCA	13	4	0.39	11
132	ATGACG	17	6	0.37	12
133	CCATCG	17	6	0.37	13
134	AGGGGA	78	51	0.36	44
135	GCTGCA	77	50	0.35	43
136	ACCCCA	76	49	0.33	40
137	CGGAGC	21	9	0.33	16
138	CCTCCG	28	13	0.32	19
139	CGGGAC	16	6	0.3	10
140	CCTGGA	88	59	0.3	45
141	AGGCGA	18	7	0.29	17
142	ACCCCT	54	32	0.28	36
143	GCTCCC	56	34	0.27	36
144	CGTCAC	16	6	0.27	15
145	AGCGCA	16	6	0.26	11
146	GAAGCC	62	38	0.25	39
147	GAGGCC	79	52	0.22	42
148	ACCCTC	54	32	0.22	33
149	CCCGGC	37	20	0.21	21
150	CGAGAA	20	8	0.2	17
151	CCACCG	29	14	0.18	20
152	ACTTCG	16	6	0.17	14
153	GATGAC	48	28	0.17	35
154	ACGAGG	23	10	0.16	18
155	CCGGAG	20	8	0.15	18
156	ACCCAC	60	37	0.12	41
157	CTGGGC	105	74	0.11	50
158	CCACGG	23	10	0.09	19
159	CGGTCC	13	4	0.09	12

160	AGCACC *	54	33	0.09	40
161	ACACCC	53	32	0.08	38
162	AGGGCC	54	33	0.08	30
163	CGCGAA	6	1	0.02	6
164	GAGCCC	58	36	0.02	36
165	CTGAGC	71	46	0.02	45
166	AATCGG	13	4	0.02	11

Table 10. Dyad patterns over-represented in STAR elements.
The patterns are ranked according to significance coefficient. These were determined using RSA-Tools with the random sequence from the human genome as reference. Patterns that comprise the most discriminant variables in Linear Discriminant Analysis are indicated with an asterisk.

Number	Dyad sequence	Observed occurrences	Expected occurrences	Significance coefficient
1	CCCN{2}CGG	36	7	9.31
2	CCGN{6}CCC	40	10	7.3
3	CAGN{0}CGG	36	8	7.13
4	CGCN{15}CCC	34	8	6.88
5	CGGN{9}GCC	33	7	6.82
6	CCCN{9}CGC	35	8	6.72
7	CCCN{1}GCG	34	8	6.64
8	CCCN{0}CAC	103	48	6.61
9	AGCN{16}CCG	29	6	5.96
10	CCCN{4}CGC	34	8	5.8
11	CGCN{13}GGA	26	5	5.77
12	GCGN{16}CCC	30	7	5.74
13	CGCN{5}GCA	25	5	5.49
14	CCCN{14}CCC	101	49	5.43
15	CTGN{4}CGC	34	9	5.41
16	CCAN{12}GCG	28	6	5.37
17	CGGN{11}CAG	36	10	5.25
18	CCCN{5}GCC	75	33	4.87
19	GCCN{0}CCC	64	26	4.81
20	CGCN{4}GAC	19	3	4.78
21	CGGN{0}CAG	33	9	4.76
22	CCCN{3}CGC	32	8	4.67
23	CGCN{1}GAC	20	3	4.58

24	GCGN{2}GCC	29	7	4.54
25	CCCN{4}GCC	76	34	4.53
26	CCCN{1}CCC	103	52	4.53
27	CCGN{13}CAG	33	9	4.5
28	GCCN{4}GGA	64	27	4.48
29	CCGN{3}GGA	26	6	4.46
30	AGGN{2}GGG	118	63	4.44
31	CACN{5}GCG	22	4	4.42
32	CGCN{17}CCA	27	6	4.39
33	CCCN{9}GGC	69	30	4.38
34	CCTN{5}GCG	28	7	4.37
35	GCGN{0}GAC	19	3	4.32
36	GCCN{0}GGC	40	7	4.28
37	GCGN{2}CCC	26	6	4.27
38	CCGN{11}CCC	32	9	4.17
39	CCCN{8}TCG	23	5	4.12
40	CCGN{17}GCC	30	8	4.12
41	GGGN{5}GGA	101	52	4.11
42	GGCN{6}GGA	71	32	4.1
43	CCAN{4}CCC	96	48	4.1
44	CCTN{14}CCG	32	9	4.09
45	GACN{12}GGC	45	16	4.07
46	CGCN{13}CCC	30	8	4.04
47	CAGN{16}CCC	92	46	4.02
48	AGCN{10}GGG	75	35	3.94
49	CGGN{13}GGC	30	8	3.93
50	CGGN{1}GCC	30	8	3.92
51	AGCN{0}GGC	26	6	3.9
52	CCCN{16}GGC	64	28	3.89
53	GCTN{19}CCC	67	29	3.87
54	CCCN{16}GGG	88	31	3.81
55	CCCN{9}CGG	30	8	3.77
56	CCCN{10}CGG	30	8	3.76
57	CCAN{0}GCG	32	9	3.75
58	GCCN{17}CGC	26	6	3.74
59	CCTN{6}CGC	27	7	3.73
60	GGAN{1}CCC	63	27	3.71
61	CGCN{18}CAC	24	5	3.7
62	CGCN{20}CCG	21	4	3.69
63	CCGN{0}GCA	26	6	3.69
64	CGCN{20}CCC	28	7	3.69
65	AGCN{15}CCC	67	30	3.65

66	CCTN{7}GGC	69	31	3.63
67	GCCN{5}CGC	32	9	3.61
68	GCCN{14}CGC	28	7	3.59
69	CAGN{11}CCC	89	45	3.58
70	GGGN{16}GAC	53	21	3.57
71	CCCN{15}GCG	25	6	3.57
72	CCCN{0}CGC	37	12	3.54
73	CCCN{16}AGC *	67	30	3.54
74	AGGN{9}GGG	96	50	3.52
75	CGCN{12}CTC	28	7	3.46
76	CACN{8}CGC	23	5	3.43
77	CCAN{7}CCG	31	9	3.42
78	CGGN{1}GCA	25	6	3.41
79	CGCN{14}CCC	29	8	3.4
80	AGCN{0}CCC	76	36	3.4
81	CGCN{13}GTC	18	3	3.37
82	GCGN{3}GCA	26	7	3.35
83	CGGN{0}GGC	34	11	3.35
84	GCCN{14}CCC	68	31	3.33
85	ACCN{7}CGC	21	4	3.32
86	AGGN{7}CGG	33	10	3.31
87	CCCN{16}CGA	22	5	3.3
88	CGCN{6}CAG	31	9	3.29
89	CAGN{11}GCG	29	8	3.29
90	CCGN{12}CCG	19	4	3.26
91	CGCN{18}CAG	27	7	3.24
92	CAGN{1}GGG	80	39	3.21
93	CGCN{0}CCC	32	10	3.2
94	GCGN{18}GCC	26	7	3.18
95	CGGN{15}GGC	27	7	3.15
96	CCCN{15}AGG	72	34	3.14
97	AGGN{20}GCG	26	7	3.14
98	CGGN{5}CTC	26	7	3.13
99	TCCN{17}CGA	23	5	3.12
100	GCGN{4}CCC	30	9	3.08
101	CCCN{2}CGC	30	9	3.07
102	CGTN{3}CAG	28	8	3.06
103	CCGN{13}GAG	27	7	3.05
104	CTCN{6}CGC	28	8	3.04
105	CGCN{4}GAG	21	5	3.03
106	GCGN{5}GGA	24	6	3.03
107	CCGN{1}CAG	27	7	3.01

108	CGCN{11}CCG	18	3	2.99
109	GCGN{19}CCC	26	7	2.98
110	CGCN{18}GAA	21	5	2.98
111	GGGN{19}GGA	78	39	2.95
112	CCAN{1}CGG	24	6	2.94
113	CCCN{7}GCG	25	6	2.94
114	AGGN{10}CCC	84	43	2.92
115	CCAN{0}GGG	97	52	2.88
116	CAGN{10}CCC	82	41	2.87
117	CCGN{18}CCG	19	4	2.86
118	CCGN{18}GGC	26	7	2.85
119	CCCN{2}GCG	24	6	2.84
120	CGCN{1}GGC	25	7	2.83
121	CCGN{5}GAC	19	4	2.81
122	GGAN{0}CCC	52	22	2.8
123	CCCN{1}CCG	29	9	2.78
124	CCCN{15}ACG	23	6	2.75
125	AGCN{8}CCC	66	31	2.73
126	CCCN{3}GGC	60	27	2.71
127	AGGN{9}CGG	31	10	2.7
128	CCCN{14}CGC	27	8	2.7
129	CCGN{0}CCG	19	4	2.7
130	CGCN{8}AGC	23	6	2.69
131	CGCN{19}ACC	21	5	2.68
132	GCGN{17}GAC	17	3	2.66
133	AGCN{1}GCG	24	6	2.63
134	CCGN{11}GGC	31	10	2.63
135	CGGN{4}AGA	26	7	2.63
136	CGCN{14}CCG	17	3	2.62
137	CCTN{20}GCG	24	6	2.62
138	CCAN{10}CGC	26	7	2.61
139	CCCN{20}CAC	69	33	2.6
140	CCGN{11}GCC	27	8	2.6
141	CGCN{18}CCC	26	7	2.59
142	CGGN{15}CGC	16	3	2.57
143	CGCN{16}GCC	24	6	2.55
144	CGCN{20}GGC	23	6	2.54
145	CGCN{19}CCG	18	4	2.52
146	CGGN{10}CCA	28	8	2.51
147	CGCN{17}CCC	26	7	2.51
148	CGCN{11}ACA	23	6	2.51
149	CGGN{0}ACC	17	3	2.5

150	GCGN{10}GCC	24	6	2.49
151	GCGN{8}GAC	17	3	2.49
152	CCCN{15}GGG	84	32	2.44
153	CGGN{16}GGC	27	8	2.44
154	CGCN{16}CCA	23	6	2.42
155	GCCN{3}CCC	73	36	2.4
156	CAGN{4}GGG	94	51	2.4
157	CCCN{6}GCG	23	6	2.38
158	CCGN{16}CGC	17	3	2.38
159	CCCN{17}GCA	61	28	2.37
160	CGCN{13}TCC	24	6	2.37
161	GCCN{1}CGC	29	9	2.36
162	CCGN{19}GAG	26	7	2.35
163	GGGN{10}GGA	89	48	2.35
164	CAGN{5}CCG	32	11	2.35
165	CGCN{3}AGA	19	4	2.32
166	GCCN{0}GCC	29	9	2.32
167	CCCN{8}GGC	61	28	2.31
168	CCTN{6}GCG	22	6	2.29
169	GACN{6}CCC	48	20	2.29
170	CGGN{1}CCC	26	8	2.27
171	CCCN{15}CCG	30	10	2.27
172	CAGN{9}CCC	84	44	2.26
173	CGGN{10}GGC	27	8	2.26
174	CGAN{10}ACG	10	1	2.26
175	GCGN{3}TCC	21	5	2.26
176	CCCN{3}GCC	75	38	2.24
177	GCGN{1}ACC	17	3	2.24
178	CCGN{9}AGG	27	8	2.23
179	CGCN{16}CAG	26	8	2.23
180	GGCN{0}CCC	62	29	2.22
181	AGGN{12}CCG	26	8	2.19
182	CCGN{0}GCG	16	3	2.19
183	CCGN{2}GCC	30	10	2.18
184	CCGN{11}GTC	19	4	2.17
185	CAGN{0}CCC	88	47	2.17
186	CCCN{5}CCG	32	11	2.17
187	GCCN{20}CCC	66	32	2.15
188	GACN{2}CGC	18	4	2.14
189	CGCN{6}CAC	23	6	2.13
190	AGGN{14}GCG	25	7	2.1
191	GACN{5}CGC	17	3	2.1

192	CCTN{19}CCG	29	9	2.1
193	CCGN{12}GGA	24	7	2.08
194	GGCN{9}GAC *	44	18	2.08
195	AGGN{10}GGG	94	52	2.07
196	CCGN{10}GAG	25	7	2.07
197	CGCN{6}GGA	20	5	2.06
198	CGCN{7}AGC	23	6	2.04
199	CCAN{13}CGG	26	8	2.03
200	CGGN{6}GGA	25	7	2.03
201	CGCN{19}GCC	24	7	2.03
202	CCAN{12}CGC	24	7	2.02
203	CGGN{1}GGC	41	16	2.02
204	GCGN{3}CCA	25	7	2.01
205	AGGN{1}CGC	21	5	2
206	CTCN{5}CGC	24	7	1.98
207	CCCN{0}ACG	30	10	1.97
208	CAGN{17}CCG	29	9	1.96
209	GGCN{4}CCC	62	30	1.96
210	AGGN{8}GCG	26	8	1.96
211	CTGN{1}CCC	88	48	1.94
212	CCCN{16}CAG	85	46	1.94
213	CGCN{9}GAC	16	3	1.93
214	CAGN{6}CCG	29	9	1.92
215	CGTN{12}CGC	11	1	1.92
216	CTCN{7}GCC	69	35	1.92
217	CGCN{19}TCC	22	6	1.92
218	CCCN{7}GCC	67	33	1.91
219	CAGN{13}CGG	30	10	1.9
220	CGCN{1}GCC	27	8	1.9
221	CGCN{17}CCG	17	4	1.89
222	AGGN{4}CCC	63	31	1.89
223	AGCN{10}CGC	21	5	1.89
224	CCCN{11}CGG	30	10	1.88
225	CCCN{8}GCC	75	39	1.86
226	CCGN{1}CGG	22	3	1.86
227	CCCN{1}ACC	71	36	1.85
228	CGCN{0}CAG	25	7	1.85
229	CCGN{19}TGC	23	6	1.82
230	GCGN{4}CGA	12	2	1.82
231	CCGN{19}GCC	30	10	1.82
232	CCAN{10}CCC	85	46	1.81
233	CAGN{13}GGG	91	51	1.81

234	AGCN{18}CGG	23	6	1.81
235	CGAN{8}CGC	11	1	1.81
236	AGCN{4}CCC	63	31	1.8
237	GGAN{6}CCC	61	30	1.8
238	CGGN{13}AAG	23	6	1.8
239	ACCN{11}CGC	19	5	1.79
240	CCGN{12}CAG	28	9	1.78
241	CCCN{12}GGG	76	29	1.77
242	CACN{17}ACG	22	6	1.76
243	CAGN{18}CCC	82	44	1.76
244	CGTN{10}GTC	19	5	1.75
245	CCCN{13}GCG	23	6	1.75
246	GCAN{1}CGC	20	5	1.73
247	AGAN{4}CCG	24	7	1.73
248	GCGN{10}AGC	22	6	1.72
249	CGCN{0}GGA	12	2	1.72
250	CGGN{4}GAC	17	4	1.69
251	CCCN{12}CGC	26	8	1.68
252	GCCN{15}CCC	65	33	1.68
253	GCGN{6}TCC	20	5	1.66
254	CGGN{3}CAG	33	12	1.65
255	CCCN{3}CCA	88	49	1.65
256	AGCN{3}CCC	59	28	1.65
257	GGGN{16}GCA	65	33	1.65
258	AGGN{8}CCG	28	9	1.64
259	CCCN{0}CCG	29	10	1.64
260	GCGN{5}GAC	16	3	1.64
261	CCCN{9}ACC	60	29	1.64
262	CTGN{5}CGC	25	8	1.64
263	CGCN{14}CTC	23	7	1.64
264	CGGN{14}GCA	23	7	1.63
265	CCGN{8}GCC	26	8	1.62
266	CCGN{7}CAC	23	7	1.62
267	AGCN{8}GCG	21	6	1.61
268	CGGN{16}GGA	29	10	1.61
269	CCAN{12}CCG	26	8	1.61
270	CGGN{2}CCC	26	8	1.6
271	CCAN{13}GGG	71	37	1.6
272	CGGN{15}GCA	21	6	1.6
273	CGCN{9}GCA	20	5	1.58
274	CGGN{19}CCA	26	8	1.58
275	GGGN{15}CGA	20	5	1.57

276	CCCN{10}CGC	26	8	1.57
277	CTCN{14}CGC	26	8	1.55
278	CACN{11}GCG	20	5	1.55
279	CCGN{2}GGC	24	7	1.55
280	CTGN{18}CCC	85	47	1.54
281	GGGN{13}CAC	58	28	1.54
282	CCTN{15}GGC	62	31	1.54
283	CCCN{20}CGA	20	5	1.54
284	CCCN{8}CGA	20	5	1.53
285	GAGN{7}CCC	61	30	1.53
286	CGCN{2}CCG	22	6	1.53
287	CCCN{0}TCC	98	57	1.52
288	AGCN{0}GCC	21	6	1.52
289	CCCN{2}TCC	82	45	1.52
290	CCGN{5}CCC	30	10	1.52
291	CGCN{13}CGC	16	3	1.51
292	CCCN{1}CGC	28	9	1.51
293	GCCN{16}GCA	53	25	1.51
294	CCCN{16}CCA	84	46	1.5
295	CCGN{13}CGC	19	5	1.5
296	CCGN{17}CAG	28	9	1.49
297	CGGN{18}GGC	26	8	1.49
298	CCGN{14}AGG	23	7	1.49
299	CCCN{5}CGG	26	8	1.49
300	CCCN{6}GGA	58	28	1.49
301	ACGN{2}CCC	20	5	1.49
302	CCAN{9}CCG	27	9	1.48
303	CCCN{19}CCA	78	42	1.48
304	CAGN{0}GGG	77	41	1.48
305	AGCN{1}CCC	58	28	1.47
306	GCGN{7}TCC	27	9	1.46
307	ACGN{18}CCA	25	8	1.46
308	GCTN{14}CCC	61	30	1.46
309	GCGN{14}CCC	23	7	1.46
310	GCGN{19}AGC	20	5	1.45
311	CCGN{8}CAG	29	10	1.45
312	GCGN{6}GCC	22	6	1.45
313	GCGN{10}GCA	20	5	1.44
314	CCTN{7}GCC	69	36	1.44
315	GCCN{13}GCC	54	26	1.42
316	CCCN{14}GCC	63	32	1.42
317	CCCN{15}CGG	26	8	1.42

318	CCAN{13}CGC	23	7	1.42
319	AGCN{11}GGG	67	35	1.41
320	GGAN{0}GCC	64	32	1.4
321	GCCN{3}TCC	61	30	1.4
322	CCTN{5}GCC	69	36	1.39
323	CGGN{18}CCC	25	8	1.39
324	CCTN{3}GGC	59	29	1.38
325	CCGN{0}CTC	22	6	1.38
326	AGCN{17}GCG	19	5	1.37
327	ACGN{14}GGG	20	5	1.37
328	CGAN{12}GGC	19	5	1.37
329	CCCN{20}CGC	24	7	1.37
330	ACGN{12}CTG	24	7	1.36
331	CCGN{0}CCC	36	14	1.36
332	CCGN{10}GGA	23	7	1.36
333	CCCN{3}GCG	21	6	1.36
334	GCGN{14}CGC	22	3	1.35
335	CCGN{8}CGC	16	4	1.35
336	CGCN{10}ACA	22	6	1.34
337	CCCN{19}CCG	28	10	1.33
338	CACN{14}CGC	20	5	1.32
339	GACN{3}GGC	46	21	1.32
340	GAAN{7}CGC	19	5	1.32
341	CGCN{16}GGC	21	6	1.31
342	GGCN{9}CCC	64	33	1.31
343	CCCN{9}GCC	64	33	1.31
344	CGCN{0}TGC	26	9	1.3
345	CCTN{8}GGC	67	35	1.3
346	CCAN{8}CCC	82	46	1.29
347	GACN{2}CCC	42	18	1.28
348	GGCN{1}CCC	54	26	1.27
349	CGCN{0}AGC	24	7	1.26
350	AGGN{4}GCG	28	10	1.26
351	CGGN{6}TCC	22	6	1.25
352	ACGN{19}GGC	20	5	1.25
353	CCCN{8}ACG	21	6	1.24
354	CCCN{18}GCC	62	31	1.24
355	GCCN{2}CGA	19	5	1.24
356	CCCN{8}GCG	28	10	1.23
357	CCCN{0}CTC	76	41	1.23
358	GCCN{11}CGC	27	9	1.22
359	AGCN{9}CCC	59	29	1.22

360	GCTN{0}GCC	71	38	1.21
361	CGCN{3}CCC	26	9	1.21
362	CCCN{2}CCC	117	72	1.19
363	GCCN{9}CGC	23	7	1.19
364	GCAN{19}CGC	19	5	1.19
365	CAGN{4}CGG	32	12	1.18
366	CAGN{2}GGG	80	44	1.17
367	GCCN{16}CCC	67	35	1.16
368	GAGN{5}CCC	60	30	1.16
369	CCTN{16}TCG	20	6	1.16
370	CCCN{2}GGC	62	32	1.15
371	GCGN{13}GGA	24	8	1.15
372	GCCN{17}GGC	66	25	1.15
373	CCCN{14}GGC	58	29	1.14
374	AGGN{3}CCG	31	12	1.14
375	CACN{0}CGC	32	12	1.14
376	CGGN{18}CAG	28	10	1.14
377	AGCN{1}GCC	57	28	1.13
378	CGCN{18}GGC	23	7	1.13
379	CCCN{5}AGG	64	33	1.11
380	AACN{0}GCG	9	1	1.11
381	CCCN{10}CCA	88	50	1.09
382	CGCN{13}GAG	20	6	1.09
383	CGCN{7}GCC	25	8	1.08
384	CCCN{9}CCG	28	10	1.07
385	CGCN{16}CCC	24	8	1.05
386	GAAN{13}CGC	18	5	1.05
387	GGCN{3}CCC	49	23	1.03
388	TCCN{11}CCA	87	50	1.03
389	CACN{0}CCC	70	38	1.02
390	CGCN{16}CCG	15	3	1.02
391	CGGN{15}AGC	21	6	1.02
392	CCCN{12}GCG	21	6	1.02
393	CCCN{9}GAG	59	30	1.01
394	CCGN{20}TCC	24	8	1.01
395	CGCN{0}CGC	17	4	1.01
396	ATGN{7}CGG	20	6	1
397	GGGN{20}GCA	59	30	1
398	CGGN{4}GGC	26	9	0.99
399	CGGN{16}AGC	22	7	0.99
400	CGGN{5}GGC	25	8	0.99
401	GCGN{0}GGA	25	8	0.98

402	GGCN{20}CAC	52	25	0.98
403	CCCN{9}CCC	97	58	0.97
404	ACCN{17}GGC	44	20	0.97
405	CCCN{6}CGA	18	5	0.96
406	AAGN{10}CGG	26	9	0.96
407	CGCN{17}CAC	21	6	0.95
408	CCCN{16}CGG	25	8	0.94
409	GACN{18}GGC	39	17	0.94
410	GGGN{15}GAC	47	22	0.92
411	GCCN{4}TCC	66	35	0.92
412	GGCN{15}CCC	56	28	0.92
413	CAGN{12}CGC	24	8	0.92
414	CCAN{3}GCG	22	7	0.91
415	CCGN{16}GAG	22	7	0.9
416	AGCN{2}CGC	24	8	0.89
417	GAGN{4}CCC	54	27	0.89
418	AGGN{3}CGC	23	7	0.88
419	CACN{13}AGG *	67	36	0.88
420	CCCN{4}CAG	88	51	0.88
421	CCCN{2}GAA	63	33	0.87
422	CGCN{19}GAG	21	6	0.87
423	ACGN{18}GGG	21	6	0.87
424	CCCN{4}GGC	62	32	0.87
425	CGGN{9}GAG	28	10	0.86
426	CCCN{3}GGG	66	26	0.86
427	GAGN{4}GGC	66	35	0.85
428	CGCN{5}GAG	18	5	0.84
429	CCGN{20}AGG	24	8	0.84
430	CCCN{15}CCC	88	51	0.83
431	AGGN{17}CCG	25	8	0.82
432	AGGN{6}GGG	89	52	0.82
433	GGCN{20}CCC	57	29	0.82
434	GCAN{17}CGC	19	5	0.82
435	CGAN{11}ACG	9	1	0.81
436	CGCN{2}GGA	19	5	0.81
437	CTGN{5}CCC	79	45	0.8
438	TCCN{20}CCA	77	43	0.8
439	CCAN{2}GGG	59	30	0.8
440	CCGN{15}GCG	14	3	0.8
441	CCAN{5}GGG	69	38	0.79
442	CGGN{1}TGC	24	8	0.79
443	CCCN{14}GCG	21	6	0.79

444	CAGN{0}CCG	27	10	0.79
445	GCCN{9}TCC	60	31	0.78
446	AGGN{20}CGC	22	7	0.78
447	CCCN{6}GAC	42	19	0.77
448	CGGN{11}CCA	23	7	0.76
449	GGGN{14}CAC	57	29	0.75
450	GCAN{15}CGC	19	5	0.74
451	CGCN{2}ACA	20	6	0.74
452	ACCN{9}CCC	57	29	0.73
453	GCGN{9}CGC	20	3	0.73
454	CAGN{15}GCG	23	7	0.73
455	CCCN{18}GTC	45	21	0.72
456	GCGN{3}CCC	24	8	0.72
457	CGGN{11}GCC	23	8	0.72
458	CCCN{1}CGG	24	8	0.71
459	GCCN{4}CCA	70	38	0.71
460	CCCN{4}CCG	30	12	0.7
461	CGTN{2}GCA	21	6	0.7
462	AGCN{7}TCG	18	5	0.69
463	CCGN{15}GAA	20	6	0.69
464	ACCN{5}CCC	62	33	0.69
465	CGCN{14}GAG	19	5	0.68
466	CCCN{7}CGC	30	12	0.68
467	GAGN{12}CGC	21	6	0.68
468	GGCN{17}CCC	58	30	0.67
469	ACGN{11}CTC	21	7	0.65
470	ACAN{9}CGG	24	8	0.65
471	CTGN{7}CCC	82	47	0.65
472	CCCN{2}GCC	72	40	0.65
473	CGGN{2}GCA	24	8	0.64
474	CCCN{0}TGC	83	48	0.64
475	CGCN{7}ACC	18	5	0.63
476	GCAN{2}GCC	54	27	0.63
477	GCGN{8}CCA	20	6	0.63
478	AGCN{0}CGC	22	7	0.63
479	GCGN{2}GCA	18	5	0.63
480	CCGN{2}GTC	18	5	0.62
481	CCGN{3}ACA	21	7	0.62
482	ACGN{13}TGG	21	7	0.62
483	CCAN{8}CGC	23	8	0.62
484	CCGN{9}GGC	23	8	0.61
485	CCAN{5}CCG	25	9	0.61

486	AGGN{3}GGG	97	59	0.61
487	CAGN{2}GGC	78	45	0.61
488	CCCN{8}CAG	81	47	0.61
489	AGCN{5}CAG	80	46	0.6
490	CGGN{16}GCC	22	7	0.6
491	GCGN{15}CCC	23	8	0.6
492	CCCN{11}GCC	59	31	0.59
493	CGAN{2}ACG	9	1	0.59
494	CGGN{4}GCC	22	7	0.59
495	CACN{6}CGC	19	6	0.59
496	CGGN{5}ACG	11	2	0.59
497	CTGN{4}GCC *	66	36	0.59
498	GGGN{18}CGA	18	5	0.59
499	CCTN{8}CGC	22	7	0.59
500	GCCN{4}CCC	67	37	0.58
501	CGGN{10}GCC	22	7	0.58
502	GCCN{5}GGA	54	27	0.57
503	ACCN{7}GCG	15	4	0.57
504	CCCN{8}CGC	24	8	0.57
505	CAGN{5}CCC	77	44	0.56
506	CACN{14}GGA	63	34	0.56
507	CCCN{1}GCC	94	57	0.55
508	CCCN{5}AGC	67	37	0.55
509	GGCN{5}GGA	59	31	0.55
510	CGAN{17}GAG	19	6	0.55
511	CGCN{7}ACA	18	5	0.54
512	CCAN{13}CCC	87	52	0.54
513	CGGN{20}GGC	24	8	0.54
514	CCCN{17}GCC	58	30	0.53
515	CCTN{10}CCG	30	12	0.53
516	CCCN{8}CCG	27	10	0.53
517	CGCN{3}GAG	18	5	0.52
518	CGCN{7}AAG	17	5	0.51
519	CGGN{11}GGA	23	8	0.51
520	CCGN{15}CCG	15	4	0.51
521	CCCN{3}GCA	57	30	0.51
522	CGGN{2}CAG	24	8	0.5
523	AGGN{2}CCG	24	8	0.5
524	CCCN{4}CAC	69	38	0.5
525	GGAN{19}CCC	56	29	0.49
526	CCCN{8}CAC	68	38	0.49
527	ACCN{6}CCG	18	5	0.49

528	CCCN{6}GGC	54	28	0.49
529	CCCN{6}CCG	29	11	0.48
530	CGCN{14}GCC	26	9	0.47
531	CCGN{5}TCC	25	9	0.46
532	GCCN{6}GCC	55	28	0.46
533	CGGN{7}GGA	24	8	0.45
534	GGGN{6}GGA	87	52	0.44
535	GCCN{12}TCC	60	32	0.44
536	AGTN{16}CCG	17	5	0.44
537	GGCN{19}GCC	68	29	0.44
538	CCGN{3}CCG	22	7	0.44
539	CCCN{8}ACC	58	31	0.44
540	CAGN{15}GCC	77	44	0.44
541	CCCN{17}CGG	24	8	0.44
542	GCGN{1}CCA	22	7	0.44
543	CCCN{14}CAG	79	46	0.44
544	CCCN{8}CCC	89	53	0.44
545	ACAN{12}GCG	23	8	0.43
546	AGGN{4}CCG	23	8	0.43
547	CGCN{13}GCC	23	8	0.43
548	GAGN{2}CGC	23	8	0.42
549	CCCN{9}GCG	21	7	0.42
550	CGCN{17}ACA	17	5	0.42
551	GCGN{17}CCA	23	8	0.42
552	AAGN{18}CCG	20	6	0.42
553	CGCN{1}GGA	18	5	0.41
554	CCAN{1}CCC	90	54	0.41
555	CGTN{18}TGC	20	6	0.41
556	TCCN{14}CGA	17	5	0.41
557	CACN{5}GGG	56	29	0.4
558	CCGN{12}GCA	21	7	0.4
559	CTGN{6}CCC	77	44	0.4
560	CGGN{8}GGC	32	13	0.4
561	CCAN{11}GGG	68	38	0.4
562	ACGN{19}CAA	21	7	0.39
563	GGGN{20}CCC	72	31	0.39
564	CGCN{3}CAG	23	8	0.39
565	AGCN{17}GGG	58	31	0.37
566	CACN{20}CCG	21	7	0.37
567	ACGN{17}CAG	24	8	0.37
568	AGGN{1}CCC	60	32	0.37
569	CGTN{12}CAC	20	6	0.37

570	CGGN{9}GGC	23	8	0.37
571	CGCN{10}GCG	18	3	0.37
572	CCCN{6}CTC	80	47	0.36
573	CCGN{10}AGG	23	8	0.36
574	CCCN{18}CAG	79	46	0.36
575	AGCN{17}CCG	21	7	0.36
576	AGCN{9}GCG	18	5	0.36
577	CCAN{3}GGC	62	34	0.36
578	CCCN{11}GGC	57	30	0.35
579	ACGN{5}GCA	23	8	0.35
580	CCCN{14}CGG	23	8	0.35
581	CCCN{5}CCA	91	55	0.35
582	CCGN{1}AGG	22	7	0.34
583	GGGN{10}GAC	45	22	0.34
584	CGCN{15}CCA	20	6	0.34
585	CCTN{19}CGC	22	7	0.34
586	CGTN{3}CGC	10	2	0.33
587	AGCN{14}CCG	21	7	0.33
588	GGCN{2}CGA	17	5	0.33
589	CAGN{8}CCC	79	46	0.33
590	CCGN{2}GAC	16	4	0.33
591	AGCN{19}AGG	70	40	0.32
592	CCTN{4}GGC	64	35	0.32
593	CCGN{11}AGC	22	7	0.32
594	CACN{4}CGC	18	5	0.32
595	CCGN{1}CCC	30	12	0.31
596	CTGN{13}GGC	73	42	0.31
597	CGCN{16}ACC	15	4	0.31
598	CACN{18}CAG	79	46	0.31
599	GGCN{8}GCC	68	29	0.29
600	GGGN{15}GGA	78	46	0.29
601	CCGN{16}GCC	22	7	0.29
602	CCGN{20}ACC	18	5	0.29
603	CGAN{7}CCC	17	5	0.28
604	CCGN{6}CTC	23	8	0.28
605	CGGN{10}CTC	22	7	0.28
606	CAGN{16}CGC	23	8	0.28
607	CCAN{3}AGG	77	45	0.27
608	GCCN{18}GCC	52	27	0.27
609	CGCN{18}GGA	19	6	0.26
610	CCGN{20}GGC	22	7	0.26
611	ACAN{10}GCG	17	5	0.26

612	CGGN{5}CCC	25	9	0.25
613	CCCN{7}TCC	75	43	0.25
614	ACGN{10}CGC	10	2	0.25
615	CCCN{3}TCC	81	48	0.25
616	CCGN{8}CGG	20	3	0.24
617	CCAN{15}CGG	22	7	0.24
618	CCGN{6}CCG	17	5	0.24
619	CAGN{3}GCG	25	9	0.24
620	GAGN{1}CCC	62	34	0.24
621	CCGN{18}TGC	22	7	0.23
622	CCCN{7}CCA	85	51	0.23
623	CGGN{3}CCA	24	9	0.23
624	ACGN{1}CCC	18	5	0.23
625	CGGN{13}TGA	21	7	0.22
626	CTCN{6}GGC	53	28	0.22
627	GCGN{2}GAC	15	4	0.22
628	GGGN{11}ACC	49	25	0.22
629	CGCN{4}GGA	17	5	0.22
630	CCCN{11}CCG	27	10	0.22
631	CCGN{19}GCA	20	6	0.22
632	GCGN{0}GCA	20	6	0.21
633	AGAN{7}CCC	61	33	0.21
634	CGGN{2}CCA	21	7	0.21
635	CCCN{7}CCC	89	54	0.21
636	ACCN{4}GCG	15	4	0.2
637	CCTN{15}CGC	20	6	0.2
638	AGCN{9}GTC	44	21	0.2
639	CCCN{18}CTC	74	43	0.2
640	CGCN{18}CGA	9	1	0.19
641	CCCN{15}GCC	62	34	0.18
642	ACCN{11}GGC	45	22	0.18
643	AGGN{15}CGC	29	12	0.18
644	GCGN{0}CCA	27	10	0.18
645	GCGN{9}AGC	18	5	0.17
646	GGGN{18}GCA	59	32	0.17
647	CCCN{17}CAG	77	45	0.17
648	CCAN{8}CGG	22	8	0.16
649	CCGN{10}GGC	21	7	0.16
650	GCAN{0}GCC	76	44	0.16
651	CAGN{2}CGC	20	6	0.16
652	CGCN{8}GGC	19	6	0.16
653	CTGN{17}GGC	65	36	0.16

654	GGGN{14}ACC	46	23	0.16
655	CCGN{1}TGC	20	6	0.16
656	CAGN{8}CGC	22	8	0.15
657	AAGN{11}CGC	17	5	0.15
658	CCGN{6}TCC	22	8	0.14
659	CCAN{18}CCC	72	42	0.14
660	CCAN{0}CCC	84	51	0.14
661	GAGN{6}CCC	53	28	0.14
662	AGCN{20}GGC	52	27	0.14
663	CAGN{0}CGC	21	7	0.14
664	CCGN{12}CTC	22	8	0.14
665	CGCN{15}ACG	9	1	0.13
666	GGCN{17}CGA	15	4	0.13
667	CCGN{16}AAG	19	6	0.13
668	CGCN{14}TCC	19	6	0.12
669	AGGN{7}CGC	20	7	0.12
670	CGGN{7}CCC	22	8	0.12
671	CGCN{4}GCC	34	15	0.12
672	CGAN{6}CCC	17	5	0.12
673	CCCN{19}GGA	60	33	0.11
674	CCCN{16}GCG	28	11	0.11
675	CCAN{7}CGC	20	7	0.11
676	CCCN{6}GCC	80	48	0.11
677	GCCN{14}TCC	55	29	0.11
678	AGGN{14}GCC	64	36	0.1
679	CGCN{11}GCC	20	7	0.1
680	TCCN{0}GCA	17	5	0.09
681	GCGN{8}CCC	27	11	0.09
682	CCAN{11}GCG	19	6	0.09
683	CACN{4}GGG	51	26	0.09
684	CGGN{7}TCC	20	7	0.09
685	GCGN{5}GCC	20	7	0.09
686	ACGN{12}CAG	26	10	0.09
687	CCGN{19}CGC	14	4	0.08
688	CGGN{8}TGC	18	5	0.08
689	CCCN{1}GAG	65	37	0.07
690	GCGN{19}TGA	18	6	0.07
691	GGCN{15}GCC	70	31	0.07
692	CCGN{7}CCC	27	11	0.07
693	ACAN{19}CCC	63	35	0.07
694	ACCN{16}GGG	47	24	0.07
695	AGAN{1}GGC	64	36	0.07

696	GGGN{17}TGA	64	36	0.06
697	CAGN{5}GGG	83	50	0.06
698	GCCN{13}CGC	22	8	0.06
699	GCGN{7}GGA	19	6	0.06
700	CAGN{14}CCA	94	58	0.06
701	CCGN{4}GTC	16	4	0.06
702	CCCN{13}CGC	22	8	0.06
703	GCGN{14}ACC	15	4	0.05
704	CAGN{20}GGG	81	49	0.05
705	CCGN{4}CCC	27	11	0.05
706	CGCN{5}GGC	18	6	0.05
707	CCTN{6}GGC	57	31	0.05
708	AGGN{3}GGC	67	38	0.05
709	CGGN{11}CGC	14	4	0.05
710	CTGN{18}GGA	77	46	0.04
711	CACN{17}CCA	74	43	0.04
712	CGGN{3}GAG	22	8	0.04
713	CCCN{9}CCA	82	49	0.03
714	CCCN{1}ACG	18	6	0.03
715	CAGN{1}GCC	72	42	0.03
716	AGGN{6}CCG	23	8	0.03
717	AGCN{9}GGG	57	31	0.03
718	CCCN{7}GGC	54	29	0.02
719	CCTN{13}CCC	88	54	0.02
720	CCGN{19}TTC	20	7	0.02
721	CCCN{7}CCG	27	11	0.02
722	CGAN{6}GGC	17	5	0.01
723	CGGN{4}CTC	21	7	0.01
724	CGGN{0}CGC	13	3	0.01
725	CCTN{13}ACG	19	6	0.01
726	GGGN{6}CAC	53	28	0.01
727	CCCN{16}CGC	21	7	0.01
728	CCCN{10}CTC	76	45	0
729	CCCN{0}CAG	92	57	0
730	GCCN{5}CCC	65	37	0

Table 11. Candidate STAR elements tested by Linear Discriminant Analysis

Candidate STAR	Location[1]	Length
T2 F	20q13.33	~2800
T2 R	6q14.1	~2800
T3 F	15q12	~2900
T3 R	7q31.2	~2900
T5 F	9q34.13	ND[2]
T5 R	9q34.13	ND
T7	22q12.3	~1200
T9 F	21q22.2	~1600
T9 R	22q11.22	~1600
T10 F	7q22.2	~1300
T10 R	6q14.1	~1300
T11 F	17q23.3	~2000
T11 R	16q23.1	~2000
T12	4p15.1	~2100
T13 F	20p13	~1700
T13 R	1p13.3	~1700
T14 R	11q25	~1500
T17	2q31.3	ND
T18	2q31.1	ND

[1]Chromosomal location is determined by BLAT search of DNA sequence data from the STAR elements against the human genome database. The location is given according to standard nomenclature referring to the cytogenetic ideogram of each chromosome; e.g. 1p2.3 is the third cytogenetic sub-band of the second cytogenetic band of the short arm of chromosome 1 (http://www.ncbi.nlm.nih.gov/Class/MLACourse/Genetics/chrombanding.html). F, forward sequencing reaction result; R, reverse sequencing reaction result. When the forward and reverse sequencing results mapped to different genomic locations, each sequence was extended to the full length of the original clone (as determined by restriction mapping) based on sequence information from the human genome database.

[2]ND: Not Determined.

Table 12. Arabidopsis STAR elements of the invention, including chromosome location and length

STAR	Chromosome	Length, kb
A1	I	1.2
A2	I	0.9
A3	I	0.9
A4	I	0.8
A5	I	1.3
A6	I	1.4
A7	II	1.2
A8	II	0.8
A9	II	0.9
A10	II	1.7
A11	II	1.9
A12	II	1.4
A13	II	1.2
A14	II	2.1
A15	II	1.4
A16	II	0.7
A17	II	1.5
A18	III	1.5
A19	III	0.7
A20	III	2.0
A21	IV	1.8
A22	IV	0.8
A23	IV	0.6
A24	IV	0.5
A25	V	0.9
A26	V	1.9
A27	V	1.1
A28	V	1.6
A29	V	0.9
A30	V	2.0
A31	V	2.0
A32	V	1.3

A33	V	0.9
A34	I	0.9
A35	II	1.1

Claims

1. A recombinant DNA construct comprising a promoter operably linked with a nucleic acid sequence of interest,
characterized in that said construct further comprises a sequence having anti-repressor activity,
wherein said sequence having anti-repressor activity is selected from the group consisting of:

(a) a sequence selected from the group consisting of SEQ.ID: 1–65 in Figure 26;
(b) a fragment of a sequence selected from the group consisting of SEQ ID: 1-65 in Figure 26;
(c) a functional homologue or derivative of a sequence selected from the group consisting of SEQ ID: 1–65 in Figure 26;
(d) a paralog or ortholog of a sequence selected from the group consisting of SEQ ID: 1–65 in Figure 26,

and wherein said sequence having anti-repressor activity confers to a human U-2 OS osteosarcoma cell the ability to grow after 4-5 weeks of culturing in the presence of 250 µg/ml zeocin and 0.1 ng/ml doxycyclin, when said cell comprises a LexA-repressor fusion protein containing the LexA DNA binding domain and a coding region of HP1 or HPC2 under control of the Tet-Off transcriptional regulatory system, when said said sequence having anti-repressor activity is cloned into a polylinker sequence in a plasmid, said polylinker being positioned between four LexA operator sites and the SV40 promoter that controls the zeocin resistance gene, when the plasmid is present in said cell.

2. A DNA construct according to claim 1, wherein said nucleic acid sequence of interest is a transgene open reading frame.

3. A DNA construct according to claim 2, wherein said promoter is an exogenous promoter.

4. A DNA construct according to any one of claims 1-3, wherein said promoter is a

strong constitutive promoter, preferably a viral promoter, or an inducible promoter.

5. A DNA construct according to any one of claims 1-4, wherein said promoter is a CMV promoter, a SV40 promoter or a Tet-Off promoter.

6. A DNA construct according to any one of claims 1-5, comprising in the following order:
(i) a sequence having anti-repressor activity as defined in claim 1,
(ii) the promoter operably linked with the nucleic acid sequence of interest, and
(iii) a sequence having anti-repressor activity as defined in claim 1.

7. A method for obtaining a host cell, comprising the step of transfecting the host cell with a DNA construct according to any one of claims 1-6.

8. A cell comprising a DNA construct according to any one of claims 1-6.

9. A cell according to claim 8, comprising multiple copies of said DNA construct.

10. A cell according to claim 8 or 9, which is a CHO cell.

11. A method for producing a gene product in a cell comprising providing an expression cassette comprising:
i) a transgene encoding said gene product, and
ii) a sequence having anti-repressor activity as defined in claim 1,
and allowing transcription of said expression cassette in a cell.

12. A method according to claim 11, wherein said cell is a CHO cell.

13. A method according to claim 11 or 12, wherein the expression cassette comprises in the following order:
(i) a sequence having anti-repressor activity as defined in claim 1,

(ii) the transgene comprising a promoter operably linked to an open reading frame encoding said gene product, and
(iii) a sequence having anti-repressor activity as defined in claim 1.

14. A method according to any one of claims 11-13, wherein multiple of said expression cassettes are provided and transcription thereof is allowed in said cell, and expression of said transgene is copy number-dependent.

15. Use of a sequence having anti-repressor activity as defined in claim 1, for regulating transcription of a nucleic acid of interest.