Title: A method for simultaneous production of multiple proteins; vectors and cells for use therein

The invention relates to the fields of biochemistry, molecular biology, pharmacology and diagnosis. More specifically the present invention relates to the production of proteins in a host cell. And even more specifically the invention relates to a method for improving expression of two or more proteins in a (host) cell. The method is suited for production of for example recombinant antibodies that can be used in a pharmaceutical preparation or as a diagnostic tool.

Proteins are produced in systems for a wide range of applications in biology and biotechnology. These include research into cellular and molecular function, production of proteins as biopharmaceuticals or diagnostic reagents, and modification of the traits or phenotypes of livestock and crops. Biopharmaceuticals are usually proteins that have an extracellular function, such as antibodies for immunotherapy or hormones or cytokines for eliciting a cellular response. Proteins with extracellular functions exit the cell via the secretory pathway, and undergo post-translational modifications during secretion. The modifications (primarily glycosylation and disulfide bond formation) do not occur in bacteria. Moreover, the specific oligosaccharides attached to proteins by glycosylating enzymes are species and cell-type specific. These considerations often limit the choice of host cells for heterologous protein production to eukaryotic cells (Kaufman, 2000). For expression of human therapeutic proteins, host cells such as bacteria, yeast, or plants may be inappropriate. Even the subtle differences in protein glycosylation between rodents and human, for example, can be sufficient to render proteins produced in rodent cells unacceptable for therapeutic use (Sheeley et al., 1997). The consequences of improper (i.e. non-human) glycosylation include immunogenicity, reduced functional half-life, and loss of activity. This limits the choice of host cells further, to human cell lines or to

cell lines such as Chinese Hamster Ovary (CHO) cells, which may produce glycoproteins with human-like carbohydrate structures (Liu, 1992).

Some proteins of biotechnological interest are functional as multimers, i.e. they consist of two or more, possibly different, polypeptide chains in their biologically and/or biotechnologically active form. Examples include antibodies (Wright & Morrison, 1997), bone morphogenetic proteins (Groeneveld & Burger, 2000), nuclear hormone receptors (Aranda & Pascual, 2001), heterodimeric cell surface receptors (e.g., T cell receptors, (Chan & Mak, 1989)), integrins (Hynes, 1999), and the glycoprotein hormone family (chorionic gonadotrophin, pituitary luteinizing hormone, follicle-stimulating hormone, and thyroid-stimulating hormone, (Thotakura & Blithe, 1995)). Production of such multimeric proteins in heterologous systems is technically difficult due to a number of limitations of current expression systems. These limitations include (1) difficulties in isolating recombinant cells/cell lines that produce the monomer polypeptides at high levels (predictability and yield), (2) difficulties in attaining production of the monomeric polypeptides in stoichiometrically balanced proportions (Kaufman, 2000), and (3) declines in the levels of expression during the industrial production cycle of the proteins (stability). These problems are described in more detail below.

(1) Recombinant proteins such as antibodies that are used as therapeutic compounds need to be produced in large quantities. The host cells used for recombinant protein production must be compatible with the scale of the industrial processes that are employed. Specifically, the transgene (or the gene encoding a protein of interest, the two terms are used interchangeably herein) expression system used for the heterologous protein needs to be retained by the host cells in a stable and active form during the growth phases of scale-up and production. This is achieved by integration of the transgene into the genome of the host cell. However, creation of recombinant cell lines by conventional means is a costly and inefficient process due to the unpredictability of transgene expression among the recombinant host cells.

The unpredictability stems from the high likelihood that the transgene will become inactive due to gene silencing (McBurney et al., 2002). Using conventional technologies, the proportion of recombinant host cells that produce one polypeptide at high levels ranges from 1-2%. In order to construct a cell line that produces two polypeptides at high levels, the two transgenes are generally integrated independently. If the two transgenes are transfected simultaneously on two separate plasmids, the proportion of cells that will produce both polypeptides at high levels will be the arithmetic product of the proportions for single transgenes. Therefore the proportion of such recombinant cell lines ranges from one in 2,500 to one in 10,000. For multimeric proteins with three or more subunits, the proportions decline further. These high-producing cell lines must subsequently be identified and isolated from the rest of the population. The methods required to screen for these rare high-expressing cell lines are time-consuming and expensive.

An alternative to simultaneous transfection of two transgene-bearing plasmids is sequential transfection. In this case the proportion of high-yielding clones will be the sum of the proportions for single transgenes, i.e. 2-4%. Sequential transfection however has (major) drawbacks, including high costs and poor stability. The high costs results from various factors: in particular, the time and resources required for screening for high-expressing cell lines is doubled, since high expression of each subunit must be screened for separately. The poor overall stability of host cells expressing two polypeptides is a consequence of the inherent instability of each of the two transgenes.

(2) Production of multimeric proteins requires balanced levels of transcriptional and translational expression of each of the polypeptide monomers. Imbalanced expression of the monomers is wasteful of the costly resources used in cell cultivation. Moreover, the imbalanced expression of one monomer can have deleterious effects on the cell. These effects include (a) sequestration of cellular factors required for secretion of the recombinant proteins (e.g. chaperones in the endoplasmic reticulum, (Chevet et al., 2001)),

and (b) induction of stress responses that result in reduced rates of growth and protein translation, or even in apoptosis (programmed cell death) (Pahl & Baeuerle, 1997, Patil & Walter, 2001). These deleterious effects lead to losses in productivity and yield and to higher overhead costs.

(3) Silencing of transgene expression during prolonged host cell cultivation is a commonly observed phenomenon. In vertebrate cells it can be caused by formation of heterochromatin at the transgene locus, which prevents transcription of the transgene. Transgene silencing is stochastic; it can occur shortly after integration of the transgene into the genome, or only after a number of cell divisions. This results in heterogeneous cell populations after prolonged cultivation, in which some cells continue to express high levels of recombinant protein while others express low or undetectable levels of the protein (Martin & Whitelaw, 1996, McBurney et al., 2002). A cell line that is used for heterologous protein production is derived from a single cell, yet is often scaled up to, and maintained for long periods at, cell densities in excess of ten million cells per millilitre in cultivators of 1,000 litres or more. These large cell populations (10^{14} - 10^{16} cells) are prone to serious declines in productivity due to transgene silencing (Migliaccio et al., 2000, Strutzenberger et al., 1999).

The instability of expression of recombinant host cells is particularly severe when transgene copy numbers are amplified in an attempt to increase yields. Transgene amplification is achieved by including a selectable marker gene such as dihydrofolate reductase (DHFR) with the transgene during integration. Increased concentrations of the selection agent (in the case of DHFR, the drug methotrexate) select for cells that have amplified the number of DHFR genes in the chromosome. Since the transgene and DHFR are co-localized in the chromosome, the transgene copy number increases too. This is correlated with an increase in the yield of the heterologous protein (Kaufman, 1990). However, the tandem repeats of transgenes that result from amplification are highly susceptible to silencing (Garrick et al., 1998,

Kaufman, 1990, McBurney et al., 2002). Silencing is often due to a decline in transgene copy number after the selection agent is removed (Kaufman, 1990). Removal of the selection agent, however, is routine during industrial biopharmaceutical production, for two reasons. First, cultivation of cells at industrial scales in the presence of selection agents is not economically feasible, as the agents are expensive compounds. Second, and more importantly, concerns for product purity and safety preclude maintaining selection during a production cycle. Purifying a recombinant protein and removing all traces of the selection agent is necessary if the protein is intended for pharmaceutical use. However, it is technically difficult and prohibitively expensive to do so, and demonstrating that this has been achieved is also difficult and expensive. Therefore amplification-based transgenic systems that require continual presence of selections agents are disadvantageous.

Alternatively, silencing can be due to epigenetic effects on the transgene tandem repeats, a phenomenon known as Repeat Induced Gene Silencing (RIGS) (Whitelaw et al., 2001). In these cases the copy number of the transgene is stable, and silencing occurs due to changes in the chromatin structure of the transgenes (McBurney et al., 2002). The presence of a selection agent during cell cultivation may be unable to prevent silencing of the transgene transcription unit because transgene expression is independent of expression of the selectable marker. The lack of a means to prevent RIGS in conventional transgenic systems thus results in costly losses in productivity.

The problems associated with conventional transgene expression technologies for protein production and more specifically for multimeric protein production clearly demonstrate a need in the art for a system that overcomes these problems. The present invention relates to a novel system for creating (host) cells/cell lines that efficiently express two or more proteins, for example two or more polypeptide monomers and optionally produce functional multimeric proteins from them. Important examples of heterologous multimer proteins are recombinant antibodies. In one embodiment the invention takes

advantage of proprietary DNA elements that protect transgenes from silencing, termed STabilizing Anti-Repressor (STAR or STAR™; the terms will be used interchangeably herein) elements, for the production of two or more proteins.

The invention also discloses a novel configuration of transcriptional and translational elements and selectable marker genes. In one embodiment, the invention uses antibiotic resistance genes and protein translation initiation sites with reduced translation efficiency (for example an Internal Ribosome Binding Site, IRES) in novel ways that improve heterologous protein expression. The combination of the STAR elements and these other elements results in a system for obtaining a cell which expresses two or more proteins that (1) predictably produces a high proportion of recombinant cell lines with high yields of heterologous proteins, (2) exhibits balanced and proportional expression of two or more polypeptide monomers which are constituents of a multimeric protein, and (3) creates recombinant cell lines with stable productivity characteristics.

Therefore, the invention provides in one embodiment, a method for obtaining a cell which expresses two or more proteins comprising providing said cell with two or more protein expression units encoding said two or more proteins, characterised in that at least two of said protein expression units comprise at least one STAR sequence.

The terms "cell"/"host cell" and "cell line"/"host cell line" are respectively typically defined as a eukaryotic cell and homogeneous populations thereof that are maintained in cell culture by methods known in the art, and that have the ability to express heterologous proteins.

The term "expression" is typically used to refer to the production of a specific RNA product or products, or a specific protein or proteins, in a cell. In the case of RNA products, it refers to the process of transcription. In the case of protein products, it refers to the processes of transcription, translation and

optionally post-translational modifications. In the case of secreted proteins, it refers to the processes of transcription, translation, and optionally post-translational modification (e.g. glycosylation, disfulfide bond formation, etc.), followed by secretion. In the case of multimeric proteins, it includes assembly of the multimeric structure from the polypeptide monomers. The corresponding verbs of the noun "expression" have an analogous meaning as said noun.

A protein is herein defined as being either (i) a product obtained by the processes of transcription and translation and possibly but not necessarily said product is part of a multimeric protein (for example a subunit) and/or (ii) a product obtained by the processes of transcription, translation and post-translational modification. The term "multimer" or "multimeric protein" is typically defined as a protein that comprises two or more, possibly non-identical, polypeptide chains ("monomers"). The different monomers in a multimeric protein can be present in stoichiometrically equal or unequal numbers. In either case, the proportion of the monomers is usually fixed by the functional structure of the multimeric protein.

The term "protein expression unit" is herein defined as a unit capable of providing protein expression and typically comprises a functional promoter, an open reading frame encoding a protein of interest and a functional terminator, all in operable configuration. A functional promoter is a promoter that is capable of initiating transcription in a particular cell. Suitable promotors for obtaining expression in eukaryotic cells are the CMV-promoter, a mammalian EF1-alpha promoter, a mammalian ubiquitin promoter, or a SV40 promoter. A functional terminator is a terminator that is capable of providing transcription termination. One example of a suitable terminator is an SV40 terminator. The term "an open reading frame encoding a protein of interest (or a transgene)" is typically defined as a fragment of DNA which codes for a specific RNA product or products or a specific protein or proteins, and which is optionally capable of becoming integrated into the genome of a host cell. It includes DNA elements

required for proper transcription and translation of the coding region(s) of the transgene. Said DNA encoding said protein of interest/transgene can either be a DNA encoding a product obtained by the processes of transcription and translation (and possibly but not necessarily said product is part of a multimeric protein, for example a subunit) or a product obtained by the processes of transcription, translation and post-translational modification.

The terms "recombinant cell/host cell" and "recombinant cell line/host cell line" are respectively typically defined as a host cell and homogeneous populations thereof into which a transgene has been introduced for the purpose of producing a heterologous protein or proteins.

A STAR (STabilizing Anti-Repressor) sequence (or STAR element; the terms will be used interchangeably herein) is a naturally occurring DNA element that we have isolated from eukaryotic genomes on the basis of their ability to block transgene repression. Preferably, the STAR elements are recovered from the human genome. A STAR sequence comprises the capacity to influence transcription of genes *in cis* and/or provide a stabilizing and/or an enhancing effect. It has been demonstrated that when STAR elements flank transgenes, the transgene expression level of randomly selected recombinant cell lines can be increased to levels approaching the maximum potential expression of the transgene's promoter. Moreover, the expression level of the transgene is stable over many cell generations, and does not manifest stochastic silencing. Therefore, STAR sequences confer a degree of position-independent expression on transgenes that is not possible with conventional transgenic systems. The position independence means that transgenes that are integrated in genomic locations that would result in transgene silencing are, with the protection of STAR elements, maintained in a transcriptionally active state.

STAR-sequences can be identified (as disclosed for example in example 1 of EP 01202581.3) using a method of detecting, and optionally selecting, a DNA sequence with a gene transcription-modulating quality, comprising

providing a transcription system with a variety of 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. Preferably, 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. Preferably, 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 the herein described method for identification of a DNA sequence with a gene transcription modulating quality, 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. Preferably, 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 the method for identification of a DNA sequence with a gene transcription modulating quality, in spite of the fact that said method does not necessarily measure long term stability of transcription. Preferably, 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 an identified DNA sequence with gene transcription modulating quality 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 for the identification of a DNA sequence with a gene transcription modulating quality 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, 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 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. Preferably, said gene-transcription repression in said vectors results in gene-transcription repressing chromatin. Preferably, 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. Preferably, 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.

DNA sequences with gene transcription modulating quality can be obtained from different sources, for example from a plant or vertebrate, or derivatives thereof, or a synthetic DNA sequence or one constructed by means of genetic engineering. Preferably said DNA sequence comprises a sequence as depicted in Table 3 and/or Figure 6 and/or a functional equivalent and/or a functional fragment thereof.

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 members of the same family of sequences, i.e are likely to share the common feature of the family. Herein, a large number of sequences comprising STAR activity (so-called STAR sequences or STAR elements) 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. A method for detecting the presence of a STAR sequence within a nucleic acid sequence of about 50-5000 base pairs is thus herein provided, 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. Preferably, 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.

As described above, a considerable number of sequences comprising a STAR sequence have been identified herein. 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 patters or combination thereof. Thus, preferably, at least one of said patterns is selected on the basis of optimal discrimination between said at least one sequence comprising a STAR sequence and a control sequence.

Preferably, 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. Even more preferably, 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. Then 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. Particularly preferred, 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

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. Preferably, said control sequence comprises a random sequence comprising a similar AT/CG content as said at least one sequence comprising a STAR sequence. Even more preferably, 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. Preferably, the control sequence comprises between 50% and 150% of the bases of said at least one sequence comprising a STAR sequence. Particularly preferred, 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. Even more preferably, at least one sequence pattern comprises at least 8 bases. Preferably, said at least one sequence pattern comprises a pattern listed in table 4 and/or table 5. 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 herein open 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 herein and test the altered sequence for STAR activity. Such altered sequences are therefore also included herein and can be used in method for obtaining a cell which expresses two or more proteins or in a method for identifying a cell wherein expression of two or more proteins is in a predetermined ratio. 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 4 and table 5 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 displayed 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. Particularly preferred, 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 comprise at least one of the sequences depicted in figure 6. STAR activity is feature shared by the sequences listed in figure 6. 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 discriminant statistics we have identified a set of patterns which performs best in identifying a sequence comprising a STAR element. Preferably, at least one of said sequence patterns for detecting a STAR sequence comprises a sequence pattern GGACCC, CCCTGC, AAGCCC, CCCCCA and/or AGCACC. Preferably, 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, and/or CTGN{4}GCC.

A list of STAR sequences can also be used to determine one or more consensus sequences therein. A consensus sequence for a STAR element is therefore also provided herein. 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. STAR sequences obtainable by a method as described herein are thus provided. 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. Preferably, a method is provided 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 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 described herein. Preferably, said collection of STARs comprises sequence as depicted in figure 6.

Now multiple methods are disclosed for obtaining a STAR sequence, it is clear that we also provide an isolated and/or recombinant nucleic acid sequence comprising a STAR sequence by a method as described herein.

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. Such 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. We further provide 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. The STAR in said different species is thus identified.

Preferably, said species comprises a plant or vertebrate species, preferably a mammalian species. We also provide 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 as disclosed herein 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 as described herein 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 enlarge the set of patterns that can discriminate and so on. This iterative aspect is an important aspect of methods provided herein.

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 "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.

The present invention provides, amongst others, a method for obtaining a cell which expresses two or more proteins, a method for identifying a cell wherein expression of two or more proteins is in a predetermined ratio and a protein expression unit. It is clear that in all these embodiments the above-described obtainable STAR sequences can be used. For example, a STAR sequence from figure 6, table 3, table 4, table 5 or combinations thereof. More

preferably, said STAR sequence is a vertebrate STAR sequence or a plant STAR sequence. Even more preferably, said vertebrate STAR sequence is a human STAR sequence. It is furthermore preferred to use a STAR sequence from a species from which a gene of interest is expressed. For example, when one would like to express two or more proteins and one of the proteins is a human protein, one preferably includes a human STAR sequence for the expression of said human protein.

As outlined above the STAR elements flanking an expression unit are the basis of the stable expression of the monomer transgenes over many cell generations. We have demonstrated that STAR elements can protect individual transgenes from silencing. In the present invention that capability is extended to more than one expression unit introduced (preferentially) independently in a recombinant host cell. Expression units that are not flanked by STAR elements can undergo significant silencing after only 5-10 culture passages, during which time silencing of the STAR protected units is negligible.

The advantages of a method for obtaining a cell which expresses two or more proteins comprising providing said cell with two or more protein expression units encoding said two or more proteins, characterised in that at least two of said protein expression units comprise at least one STAR sequence, are multifold.

The present invention uses STAR sequences for the production of two or more proteins and thereby the invention provides (1) an increased predictability in the creation of recombinant cell lines that efficiently produce the heterologous multimeric proteins of interest, (2) an increased yield of the heterologous multimeric proteins, (3) stable expression of the heterologous multimeric proteins, even during prolonged cultivation in the absence of selection agent and (4) the invention also provides favorable transgene expression characteristics without amplification of the transgene. The increased yield of heterologous proteins provided by the invention may be

obtained at low transgene copy numbers, without selective co-amplification using, for example, the DHFR/methotrexate system. This results in greater stability, since the transgene copy number is low and is not susceptible to decrease due to recombination (McBurney et al., 2002) or repeat-induced gene silencing (Garrick et al., 1998). Fifth, the broad applicability of the method of the invention includes its utility in a wide range of host cell lines. This is for example useful/desirable when a particular multimeric protein is preferably expressed by a particular host cell line (e.g. expression of antibodies from lymphocyte-derived host cell lines).

A method according to the invention therefore provides an improvement of expression of two or more proteins in a (host) cell.

In another embodiment the invention provides a method for identifying a cell wherein expression of two or more proteins is in a predetermined ratio comprising providing

- a collection of cells with two or more protein expression units encoding said two or more proteins,
- selecting cells which express said two or more proteins, and
- identifying from the obtained selection, cells that express said two or more proteins in said predetermined ratio, characterised in that at least two of said protein expression units comprise at least one STAR sequence.

The selection of cells which express said two or more proteins may for example be obtained by performing a SDS-PAGE analysis, a Western blot analysis or an ELISA, which are all techniques which are known by a person skilled in the art and therefore need no further elaboration. The identification of cells that express said two or more proteins in said predetermined ratio can also be performed by these techniques.

The presence of a STAR sequence in at least two of said protein expression units, again, provide the desired predictability, yield, stability and stoichiometrically balanced availability of the two or more proteins.

Especially when polypeptides of a multimeric protein are produced according to a method of the invention it is desirable to provide the required monomers/subunits in a ratio that is relevant for the formation of said multimeric protein. Hence, preferably said monomers/subunits are produced in a biological relevant balanced ratio. If for example, a multimeric protein consists of two subunits A and 1 subunit B it is desired to produce two subunits A for every subunit of B that is produced. Hence, a predetermined ratio is herein defined as the natural occurring ratio (stoichiometry) of the different subunits/monomers/polypeptides which comprise a multimeric protein.

In a more preferred embodiment a cell obtainable according to a method of the invention expresses two proteins. For example, two proteins which together provide a therapeutically advantageous effect. In an even more preferred embodiment the predetermined ratio of the two expressed proteins is 1:1. This is for example useful in the production of multimeric proteins in which the monomers are in a 1:1 ratio. Typical examples are antibodies that comprise two heavy chains and two light chains.

Preferably, the invention provides a method, wherein said two or more protein expression units further encode at least two different selection markers, and wherein said method further comprises a two-step selection marker screening on said cell, wherein said cell is selected in a first step on the presence of a first selection marker and in a second step on the presence of a second selection marker.

In this embodiment of the invention a two-stage antibiotic selection is used which regime results in a high proportion of isolates that express for example transgenes 1 and 2 at high levels; the first stage of selection eliminates cells that do not contain the expression unit or units, and the second stage of selection eliminates colonies that do not transcribe both bicistronic mRNAs at high levels. This regime is one of the aspects for the increased frequency of multimer-expressing recombinant cell lines achieved by

the invention compared to conventional methods. As described herein, it results in an increase in the frequency of expressor lines by more than ten-fold.

In another embodiment the invention provides a method wherein at least one of said protein expression units comprises
a monocistronic gene comprising an open reading frame encoding a protein of interest and wherein said monocistronic gene is under control of a functional promoter.

In yet another embodiment the invention provides a method according to the invention, wherein at least one of said protein expression units comprises
a bicistronic gene comprising an open reading frame encoding a protein of interest, a protein translation initiation site with a reduced translation efficiency, a selection marker and wherein said bicistronic gene is under control of a functional promoter.

In a more preferred embodiment the invention provides a method according to the invention, wherein at least one of said protein expression units comprises
a bicistronic gene comprising an open reading frame encoding a protein of interest, a protein translation initiation site with a reduced translation efficiency, a selection marker and wherein said bicistronic gene is under control of a functional promoter, which protein expression unit further comprises
a monocistronic gene comprising an open reading frame encoding a second selection marker and wherein said monocistronic gene is under control of a functional promoter.

The term "bicistronic gene," is typically defined as a gene capable of providing a RNA molecule that encodes two proteins/polypeptides.

The term "monocistronic gene" is typically defined as a gene capable of providing a RNA molecule that encodes one protein/polypeptide.

The term "selection marker or selectable marker" is typically used to refer to a gene and/or protein whose presence can be detected directly or indirectly in a cell, for example a gene and/or a protein that inactivates a selection agent and protects the host cell from the agent's lethal or growth-inhibitory effects (e.g. an antibiotic resistance gene and/or protein). Another possibility is that said selection marker induces fluorescence or a color deposit (e.g. green fluorescent protein and derivatives, luciferase, or alkaline phosphatase).

The term "selection agent" is typically defined as a chemical compound that is able to kill or retard the growth of host cells (e.g. an antibiotic).

The term "selection" is typically defined as the process of using a selection marker/selectable marker and a selection agent to identify host cells with specific genetic properties (e.g. that the host cell contains a transgene integrated into its genome).

The nouns "clone" and "isolate" typically refer to a recombinant host cell line that has been identified and isolated by means of selection.

The improvements provided by a method according to the invention have three integrated aspects. (1) With existing systems, recombinant cell lines that simultaneously express acceptable quantities of the monomers of multimeric proteins can be created only at very low frequencies; the present invention increases the predictability of creating high-yielding recombinant host cell lines by a factor of ten or more. (2) Existing systems do not provide stoichiometrically balanced and proportional amounts of the subunits of multimeric proteins; the present invention ensures that the expression levels of the subunits will be balanced and proportional. (3) Existing systems do not provide a means of protecting the transgenes that encode the protein subunits from transgene silencing.

FIG 1 provides a, non-limiting, schematic representation of one of the embodiments of this part of the invention. FIG 1A and FIG 1B show two separate protein expression units. This is the configuration of the DNA

elements of the expression units in the plasmid as well as after integration into the genome. Expression unit one is shown in FIG 1A. It contains an open reading frame for a transgene (a reporter gene or subunit 1 of a multimeric (TG S1, transgene subunit 1)). This is upstream of the attenuated EMCV IRES, and of the open reading frame encoding the zeocin resistance selectable marker protein (zeo). This bicistronic transgene is transcribed at high levels from the CMV promoter. Next to this is the neomycin resistance selectable marker (neo; this confers resistance to the antibiotic G418 as well), transcribed as a monocistronic mRNA from the SV40 promoter. These two genes are flanked by STAR elements. In FIG 1B a similar expression unit is depicted. It consists of a second transgene (a second reporter gene or the open reading frame for subunit 2 of a heterodimeric protein (TG S2)) upstream of the attenuated EMCV IRES and the blasticidin selectable marker open reading frame (bsd). This bicistronic transgene is transcribed at high levels from the CMV promoter. Next to this is the neo selectable marker, transcribed as a monocistronic mRNA from the SV40 promoter. The two genes in the second expression unit are flanked by STAR elements as well.

It is clear to a person skilled in the art that the possible combinations of selection markers is numerous. Examples of possible antibiotic combinations are provided above. The one antibiotic that is particularly advantageous is zeocin, because the zeocin-resistance protein (zeocin-R) acts by binding the drug and rendering it harmless. Therefore it is easy to titrate the amount of drug that kills cells with low levels of zeocin-R expression, while allowing the high-expressors to survive. All other antibiotic-resistance proteins in common use are enzymes, and thus act catalytically (not 1:1 with the drug). When a two-step selection is performed it is therefore advantageous to use an antibiotic resistance protein with this 1:1 binding mode of action. Hence, the antibiotic zeocin is a preferred selection marker. For convenience the zeocin antibiotic is in a two-step selection method combined with puromycin-R or blasticidin-R in the second bicistronic gene, and neomycin-R or hygromycin-R

in the monocistronic gene.

It is furthermore clear that it is also possible to combine an antibiotic selection marker with a selection marker which provides induction of fluorescence or which provide a color deposit.

It is also clear to a person skilled in the art that different promoters can be used as long as they are functional in the used cell. The CMV promoter is considered the strongest available, so it is preferably chosen for the bicistronic gene in order to obtain the highest possible product yield. Other examples of suitable promoters are e.g. mammalian promoters for EF1-alpha or ubiquitin. The good expression and stability of the SV40 promoter makes it well suited for expression of the monocistronic gene; enough selection marker protein (for example the antibiotic resistance protein neomycin-R in the example cited herein) is made to confer high expression of said selection marker. Hence, said SV40 promoter is preferentially used as a promoter driving the expression of the selection marker.

In a preferred embodiment the invention provides a method wherein at least one of said protein expression units comprises at least two STAR sequences. In an even more preferred embodiment the invention provides a method wherein said protein expression unit comprising at least two STAR sequences is arranged such that said protein expression unit is flanked on either side by at least one STAR sequence. In yet an even more preferred embodiment the said at least two STAR sequences are essentially identical. Essentially identical STAR sequences are defined herein as STAR sequences which are identical in their important domains, but which may vary within their less important domains (the domains that confer the transcription stabilizing or enhancing quality), for example a pointmutation, deletion or insertion at a less important position within the STAR sequence. Preferentially said essentially identical STAR sequences provide equal amounts of transcription stabilizing or enhancing activity.

The use of STAR sequences to flank at least one protein expression unit is one of the aspects of the balanced and proportional levels of expression of two or more proteins and more specifically for the expression of the monomers of multimeric proteins. The STAR sequences create chromatin domains of definite and stable transcriptional potential. As a result, promoters that drive transcription of each bicistronic mRNA will function at definite, stable levels. A recombinant host cell line created by the method of the invention is readily identified in which these levels result in appropriate proportions of each monomer of the multimeric protein of interest being expressed at high yields.

In another embodiment the protein expression unit contains only the bicistronic gene flanked by STAR elements. The advantages of omitting the monocistronic antibiotic resistance gene are twofold. First, selection of high-expressing recombinant host cells requires the use of only two antibiotics. Second, it prevents repression of the bicistronic and/or monocistronic genes by the phenomena of promoter suppression and transcriptional interference. These phenomena are common problems in conventional transgenic systems in which two or more transcription units are located near each other. Repression by an upstream (5') unit of a downstream (3') unit is termed transcriptional interference, and repression by a downstream unit of an upstream unit is termed promoter suppression (Villemure et al., 2001). Transcriptional interference can result in suppression of adjacent transgenes in all possible arrangements (tandem, divergent, and convergent) (Eszterhas et al., 2002). These phenomena can reduce the efficiency of selection of the IRES-dependent and/or monocistronic antibiotic resistance genes, and reduce the yield of the transgene. Therefore the embodiment of the invention comprising only a bicistronic gene flanked by STAR elements provides an alternative configuration of the components.

In a preferred embodiment the method according to the invention uses a STAR sequence wherein said STAR sequence is depicted in Table 3 and/or figure 6 and/or a functional equivalent and/or a functional fragment thereof.

We have isolated and characterized an extensive collection of STAR sequences using proprietary technology. The strength of these sequences ranges widely. This is manifested by the varying degrees of improvement of transgene expression in recombinant host cells conferred by the STAR elements; some STAR elements provide full protection from silencing, while others only provide partial protection. The range in strength of the STAR elements is also manifested in their varying capacities to improve the predictability of isolating recombinant cell lines that efficiently produce the heterologous proteins of interest. For the present invention we have preferably employed STAR elements that have strong predictability characteristics, in order to have high numbers of efficiently-expressing recombinant cell lines. The STAR elements employed have moderate to strong anti-repressor activity, in order to be able to modulate the levels of recombinant protein production to match the requirements of the product (e.g. balanced and proportional expression of polypeptide monomers). The selected STAR elements also confer significant increases on the stability of expression of the transgenes.

Some STAR elements also display promoter and host cell-type specificity. These characteristics are exploited to create novel transgenic systems to optimize the production of heterologous proteins that require a specific host cell (for example, to achieve a high yield or a pharmaceutically advantageous glycosylation pattern) or a specific mode of expression (for example, the use of an inducible promoter or a constitutive promoter; the use of a promoter with moderate strength or high strength, etc.). Therefore the use of different STAR elements results in different embodiments of the invention that pertain to these types of applications.

A functional equivalent and/or a functional fragment of a sequence depicted in Table 3 and/or figure 6 is defined herein as follows. A functional equivalent of a sequence as depicted in Table 3 and/or figure 6 is a sequence derived with the information given in Table 3 and/or figure 6. For instance, a sequence that can be derived from a sequence in Table 3 and/or figure 6 by

deleting, modifying and/or inserting bases in or from a sequence listed in Table 3 and/or figure 6, wherein said derived sequence comprises the same activity in kind, not necessarily in amount, of a sequence as depicted in Table 3 and/or figure 6. A functional equivalent is further a sequence comprising a part from two or more sequences depicted in Table 3 and/or figure 6. A functional equivalent can also be a synthetic DNA sequence which 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.

Functional sequences of STAR elements can be delineated by various methods known in the art. In one embodiment deletions and/or substitutions are made in STAR sequences. DNA that is modified in such a way is for example tested for activity by 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 STAR sequences enables the elucidation of consensus sequences for elements with a gene transcription modulating and/or a gene transcription repressing quality.

A functional fragment of a STAR sequence as depicted in Table 3 and/or figure 6 can for example be obtained by deletions from the 5' end or the 3' end or from the inside of said sequences or any combination thereof, wherein said derived sequence comprises the same activity in kind, not necessarily in amount.

In a more preferred embodiment said STAR sequence as depicted in Table 3 and/or figure 6 is STAR18 and/or a functional equivalent and/or a functional fragment thereof.

Yet another preferred feature of a method according to the invention is the introduction of a (weak) Internal Ribosome Binding Site (IRES) as an example of a protein translation initiation site with a reduced translation efficiency, between the open reading frame of the protein of interest and the

selection marker open reading frame. In combination with for example the STAR sequence, this component of the present invention comprises a marked improvement in transgenic systems for the expression of two or more proteins.

Internal ribosome binding site (IRES) elements are known from viral and mammalian genes (Martinez-Salas, 1999), and have also been identified in screens of small synthetic oligonucleotides (Venkatesan & Dasgupta, 2001). The IRES from the encephalomyocarditis virus has been analyzed in detail (Mizuguchi et al., 2000). An IRES is an element encoded in DNA that results in a structure in the transcribed RNA at which eukaryotic ribosomes can bind and initiate translation. An IRES permits two or more proteins to be produced from a single RNA molecule (the first protein is translated by ribosomes that bind the RNA at the cap structure of its 5' terminus, (Martinez-Salas, 1999)). Translation of proteins from IRES elements is less efficient than cap-dependent translation: the amount of protein from IRES-dependent open reading frames (ORFs) ranges from less than 20% to 50% of the amount from the first ORF (Mizuguchi et al., 2000). This renders IRES elements undesirable for production of all subunits of a multimeric protein from one messenger RNA (mRNA), since it is not possible to achieve balanced and proportional expression of two or more protein monomers from a bicistronic or multicistronic mRNA. However, the reduced efficiency of IRES-dependent translation provides an advantage that is exploited by the current invention. Furthermore, mutation of IRES elements can attenuate their activity, and lower the expression from the IRES-dependent ORFs to below 10% of the first ORF (Lopez de Quinto & Martinez-Salas, 1998, Rees et al., 1996). The advantage exploited by the invention is as follows: when the IRES-dependent ORF encodes a selectable marker protein, its low relative level of translation means that high absolute levels of transcription must occur in order for the recombinant host cell to be selected. Therefore, selected recombinant host cell isolates will by necessity express high amounts of the transgene mRNA. Since

the recombinant protein is translated from the cap-dependent ORF, it can be produced in abundance resulting in high product yields.

It is clear to a person skilled in the art that changes to the IRES can be made without altering the essence of the function of the IRES (hence, providing a protein translation initiation site with a reduced translation efficiency), resulting in a modified IRES. Use of a modified IRES which is still capable of providing a small percentage of translation (compared to a 5' cap translation) is therefore also included in this invention.

In yet another embodiment the invention provides a method for obtaining a cell which expresses two or more proteins or a method for identifying a cell wherein expression of two or more proteins is in a predetermined ratio, wherein each of said protein expression units resides on a separate DNA-carrier. The present invention preferentially makes use of a separate transcription unit for each protein and/or monomer of a multimeric protein. In each transcription unit the monomer ORF is produced by efficient cap-dependent translation. This feature of the invention contributes that recombinant host cells are isolated which have high yields of each monomer, at levels that are balanced and proportionate to the stoichiometry of the multimeric protein. The increased predictability at which such recombinant host cells are isolated results in an improvement in the efficiency of screening for such isolates by a factor of ten or more. In a preferred embodiment said DNA-carrier is a vector (or plasmid; the terms are used interchangeably herein). In another embodiment said vector is a viral vector and in a more preferred embodiment said viral vector is an adenoviral vector or a retroviral vector. It is clear to person skilled in the art that other viral vectors can also be used in a method according to the invention.

Conventional expression systems are DNA molecules in the form of a recombinant plasmid or a recombinant viral genome. The plasmid or the viral genome is introduced into (mammalian host) cells and integrated into their genomes by methods known in the art. The present invention also uses these

types of DNA molecules to deliver its improved transgene expression system. A preferred embodiment of the invention is the use of plasmid DNA for delivery of the expression system. A plasmid contains a number of components: conventional components, known in the art, are an origin of replication and a selectable marker for propagation of the plasmid in bacterial cells; a selectable marker that functions in eukaryotic cells to identify and isolate host cells that carry an integrated transgene expression system; the protein of interest, whose high-level transcription is brought about by a promoter that is functional in eukaryotic cells (e.g. the human cytomegalovirus major immediate early promoter/enhancer, pCMV (Boshart et al., 1985)); and viral transcriptional terminators for the transgene of interest and the selectable marker (e.g. the SV40 polyadenylation site (Kaufman & Sharp, 1982)).

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 a preferred embodiment the invention provides a method for obtaining a cell which expresses two or more proteins or a method for obtaining a cell wherein expression of two or more proteins is in a predetermined ratio comprising providing two or more protein expression units wherein one of the said protein expression units or said protein(s) of interest encodes an immunoglobulin heavy chain and/or wherein another of the said protein expression units or said protein(s) of interest encodes an

immunoglobulin light chain. According to this embodiment a multimeric protein, an antibody, is obtained. It is clear to a person skilled in the art that it is possible to provide a cell which expresses an immunoglobulin heavy chain from one protein expression unit and an immunoglobulin light chain from another protein expression unit with a third protein expression unit encoding a secretory component or a joining chain. In this way the production of for example sIgA and pentameric IgM is provided.

Preferably, the used host cell secretes the produced multimer. In this way the product is easily isolated from the medium surrounding said host cell.

More preferably, the invention results in the production of a functional multimer. The functionality of the produced multimer is determined with standard procedures. For example, a produced multi subunit enzyme is tested in a corresponding enzymatic assay or binding to an antigen, for example in an ELISA, is used to test the functionality of a produced antibody.

Hence, the selection of a final suitable host cell expressing a multimer involves multiple steps amongst which are the selection for a cell that expresses all the desired subunits of a multimer, followed by a functional analysis of said multimer.

With regard to a multimeric protein high expression levels of the subunits is desired as well as the formation of a functional multimeric protein of said subunits. Surprisingly, the use of a STAR sequence for the production of the subunits of a multimeric protein results in high amount of cells that express the subunits, as compared to control vectors without a STAR sequence. Moreover, the amount of functional multimeric protein is relatively higher when compared to the control.

Production of subunits and the formation of functional multimeric protein from these subunits is in particular of importance for the production of antibodies. When the heavy chain and light chain expression cassette are flanked by a STAR sequence this results in a higher production of functional antibody, as compared to control vectors without a STAR sequence. Hence, the

presence of a STAR sequence results in a higher degree of predictability of functional antibody expression. Preferably, each expression unit comprises at least two STAR sequences which sequences are arranged such that said expression unit is flanked on either side by at least one STAR sequence.

In yet another embodiment a method according to the invention is provided, wherein said protein expression units are introduced simultaneously into said cell.

Preferebly, a functional promoter is a human cytomegalovirus (CMV) promotor, a simian virus (SV40) promoter, a human ubiquitin C promoter or a human elongation factor alpha (EF1-α) promoter.

As disclosed herein within the experimental part, a STAR sequence can confer copy number-dependence on a transgene expression unit, making transgene expression independent of other transgene copies in tandem arrays, and independent of gene-silencing influences at the site of integration. Hence, the invention also provides a method for obtaining a cell which expresses two or more proteins or a method for identifying a cell wherein expression of two or more proteins is in a predetermined ratio in which multiple copies of a protein expression unit encoding a protein of interest is integrated into the genome of said cell (i.e. in which cell, an amplification of the gene of interest is present).

According to this part of the invention, the protein expression units are introduced simultaneously into said (host) cell or collection of cells by methods known in the art. Recombinant host cells are selected by treatment with an appropriate antibiotic, for example G418, using methods known in the art. After formation of individual antibiotic-resistant colonies, another antibiotic or a combination of antibiotics, for example a combination of zeocin and blasticidin, is/are applied, and antibiotic-resistant colonies are identified and isolated. These are tested for the level of expression of transgenes.

In another embodiment the invention provides a protein expression unit comprising

- a bicistronic gene comprising an open reading frame encoding a protein of interest, a protein translation initiation site with a reduced translation efficiency, a selection marker and wherein said bicistronic gene is under control of a functional promoter
- at least one STAR sequence.

In a more preferred embodiment said protein expression unit further comprises
a monocistronic gene comprising an open reading frame encoding a second selection marker and wherein said monocistronic gene is under control of a functional promoter.

In an even more preferred embodiment said protein expression unit comprises at least two STAR sequences which are preferentially arranged such that said protein expression unit is flanked on either side by at least one STAR sequence. Examples of such a protein expression unit are provided within the experimental part of this patent application (for example Figures 1 and 5).

In another embodiment the protein expression unit according to the invention comprises STAR sequences, wherein said STAR sequences are essentially identical.

In a preferred embodiment the invention provides a protein expression unit comprising
- a bicistronic gene comprising an open reading frame encoding a protein of interest, a protein translation initiation site with a reduced translation efficiency, a selection marker and wherein said bicistronic gene is under control of a functional promoter
- at least one STAR sequence, and is optionally provided with a monocistronic gene cassette, wherein said STAR sequence is depicted in Table 3 and/or figure 6 and/or a functional equivalent and/or a functional fragment thereof and even more preferred wherein said STAR sequence is STAR18.

In another embodiment a protein expression unit according to the invention is provided wherein said protein translation initiation site with a

reduced translation efficiency comprises an Internal Ribosome Entry Site (IRES). More preferably a modified, e.g. weaker, IRES is used.

In yet another embodiment a protein expression unit according to the invention is provided wherein said protein expression unit is a vector. In a preferred embodiment said DNA-carrier is a vector (or plasmid; the terms are used interchangeably herein). In another embodiment said vector is a viral vector and in a more preferred embodiment said viral vector is an adenoviral vector or a retroviral vector. It is clear to person skilled in the art that other viral vectors can also be used in a method according to the invention.

In a preferred embodiment a protein expression unit according to the invention is provided, wherein said protein of interest is an immunoglobulin heavy chain. In yet another preferred embodiment a protein expression unit according to the invention is provided, wherein said protein of interest is an immunoglobulin light chain. When these two protein expression units are present within the same (host) cell a multimeric protein and more specifically an antibody is assembled.

The invention includes a cell provided with a protein expression unit comprising a STAR.

The invention also includes a (host) cell comprising at least one protein expression unit according to the invention. Such a (host) cell is then for example used for large-scale production processes.

The invention also includes a cell obtainable according to anyone of the methods as described herein. The invention furthermore includes a protein obtainable from said cell (for example, via the process of protein purification). Preferably, said protein is a multimeric protein and even more preferably said multimeric protein is an antibody. Such an antibody can be used in pharmaceutical and/or diagnostic applications.

The foregoing discussion and the following examples are provided for illustrative purposes, and they are not intended to limit the scope of the

invention as claimed herein. They simply provide some of the preferred embodiments of the invention. Modifications and variations, which may occur to one of ordinary skill in the art, are within the intended scope of this invention. Various other embodiments apply to the present invention, including: other selectable marker genes; other IRES elements or means of attenuating IRES activity; other elements affecting transcription including promoters, enhancers, introns, terminators, and polyadenylation sites; other orders and/or orientations of the monocistronic and bicistronic genes; other anti-repressor elements or parts, derivations, and/or analogues thereof; other vector systems for delivery of the inventive DNA molecules into eukaryotic host cells; and applications of the inventive method to other transgenic systems.

EXAMPLES

Example 1: STAR elements and two-step selection improve the predictability of transgene expression

One object of this invention is to improve transgene expression for heterologous protein production by using a two-step antibiotic selection procedure. The two-step procedure increases the predictability of finding recombinant host cell lines that express the transgene to high levels, thus increasing the yield of the heterologous protein.

Materials and Methods

Plasmid construction

The pSDH-SIB/Z and pSDH-GIB/Z families of plasmids were constructed as follows: The zeocin selectable marker was recovered by polymerase chain reaction amplification (PCR) from plasmid pEM7/zeo (Invitrogen V500-20) using primers E99 and E100 (all PCR primers and mutagenic oligonucleotide sequences are listed in Table 1), and cloned directionally into the *Xba*I and *Not*I sites of multiple cloning site (MCS) B of pIRES (Clontech 6028-1) to create pIRES-zeo. The blasticidin selectable marker was recovered by PCR from plasmid pCMV/bsd (Invitrogen V510-20) using primers E84 and E85, and cloned directionally into the *Xba*I and *Not*I sites MCS-B of pIRES to create pIRES-bsd. The SEAP (secreted alkaline phosphatase) reporter gene was recovered by PCR from plasmid pSEAP2-basic (Clontech 6049-1) using primers F11 and E87, and cloned directionally into MCS-A of pIRES-zeo and pIRES-bsd to create plasmids pIRES-SEAP-zeo and pIRES-SEAP-bsd. The GFP reporter gene was recovered from plasmid phr-GFP-1 (Stratagene 240059) by restriction digestion with *Nhe*I and *Eco*RI, and ligated directionally into MCS-A of pIRES-zeo and pIRES-bsd to create plasmids pIRES-GFP-zeo and pIRES-GFP-bsd. A linker was inserted at the non-methylated *Cla*I site of each of these plasmids (downstream of the

neomycin resistance marker) to introduce an *Age*I site using oligonucleotides F34 and F35.

The pSDH-Tet vector was constructed by PCR of the luciferase open reading frame from plasmid pREP4-HSF-Luc (van der Vlag et al., 2000) using primers C67 and C68, and insertion of the *Sac*II/*Bam*HI fragment into *Sac*II/*Bam*HI-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 multiple cloning sites (MCSI and MCSII). An AscI site was then introduced into MCSI by digestion with *Eco*RI and insertion of a linker (comprised of annealed oligonucleotides D93 and D94). The CMV promoter was amplified from plasmid pCMV-Bsd with primers D90 and D91, and used to replace the Tet-Off promoter in pSDH-Tet by *Sal*I/*Sac*II digestion and ligation to create vector pSDH-CMV. The luciferase open reading frame in this vector was replaced by SEAP as follows: vector pSDH-CMV was digested with *Sac*II and *Bam*HI and made blunt; the SEAP open reading frame was isolated from pSEAP-basic by *Eco*RI/*Sal*I 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 (*Bam*HI site removed) (Promega E1741) digested with *Nco*I/*Xba*I, to create pGL3-puro. pGL3-puro was digested with *Bgl*II/*Sal*I to isolate the SV40-puro resistance gene, which was made blunt and ligated into *Nhe*I digested, blunt-ended pSDH-CS. The resulting vector, pSDH-CSP, is shown in FIG 2. STAR18 was 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 the STAR element 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 1), in the presence of dye terminators (CEQ Dye Terminator Cycle Sequencing Kit, Beckman 608000).

pSDH-CSP plasmids containing STAR elements were modified as follows: for receiving SEAP-IRES-zeo/bsd cassettes, an *Age*I site was introduced at the *Bgl*II site by insertion of a linker, using oligonucleotides F32 and F33; for receiving GFP-IRES-zeo/bsd cassettes, an *Age*I site was introduced at the *Bsu*36I site by insertion of a linker, using oligonucleotides F44 and F45. The SEAP-IRES-zeo/bsd cassettes were inserted into the pSDH-CSP-STAR18 plasmid by replacement of the *Bsu*36I/*Age*I fragment with the corresponding fragments from the pIRES-SEAP-zeo/bsd plasmids. The GFP-IRES-zeo/bsd cassettes were inserted into pSDH-CSP-STAR plasmids by replacement of the *Bgl*II/*Age*I fragment with the corresponding fragments from the pIRES-GFP-zeo/bsd plasmids. The resulting plasmid families, pSDH-SIB/Z and pSDH-GIB/Z, are shown in FIG 3.

All cloning steps were carried out following the instructions provided by the manufacturers of the reagents used, according to methods known in the art (Sambrook et al., 1989).

Transfection and culture of CHO cells

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% CO_2. Cells were transfected with the pSDH-SIZ plasmids 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 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, cells were seeded into fresh culture vessels and 500 micrograms/ml neomycin was added. Neomycin selection was complete within 3-4 days. Fresh medium was then added containing zeocin (100 μg/ml) and cultured further. Individual clones were isolated after 4-5 days and cultured further. Expression of the reporter gene was assessed by measuring SEAP activity approximately 3 weeks after transfection.

Secreted Alkaline Phosphatase (SEAP) assay

SEAP activity (Berger et al., 1988, Henthorn et al., 1988, Kain, 1997, Yang et al., 1997) in the culture media of the 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.

Results

Transfection of the pSDH-SIZ-STAR18 expression vector consistently results in ~10-fold more colonies than transfection of the empty pSDH-SIZ vector, presumably due to the increased proportion of primary transfectants that are able to bring the neomycin resistance gene to expression. The outcome of a typical experiment is shown in Table 2, in which transfection of the empty vector yielded ~100 G418-resistant colonies, and transfection of the STAR18 vector yielded ~1000 colonies.

The expression of the SEAP reporter transgene was compared between the empty pSDH-SIZ vector (hence, without a STAR sequence) and the

STAR18 vector (FIG 4). The populations of G418-resistant isolates were divided into two sets. The first set was cultured with G418 only (one-step selection). For this set, the inclusion of STAR18 to protect the transgene from silencing resulted in higher yield of reporter protein: the maximal level of expression among the 20 clones analyzed was 2-3-fold higher than the maximal expression level of clones without the STAR element. The inclusion of STAR18 also led increased predictability: more than 25% of the STAR18 clones had expression levels greater than or equal to the maximum expression level observed in the STARless clones. In this population of STAR18 clones, 70% had expression above the background level, while only 50% of the STARless clones had expression above the background level.

The performance of STAR18 was even better when used in a two-step selection. The second set of G418-resistant isolates was treated with zeocin. Clones that survived the two-step selection regime were assayed for expression of the SEAP reporter transgene. In this case too, the STAR18 element increased the yield compared to the STARless clones by approximately three-fold. The predictability was also increased by inclusion of STAR18: ~80% of the population had expression levels greater than the highest-expressing STARless clone.

When the one-step selection is compared with the two-step selection, it can be seen that the latter is superior in terms of both yield and predictability. In fact with two-step selection, no clones appear with background levels of expression. This is due to the requirement imposed on clones that survive zeocin selection that they have high levels of transcription of the bicistronic SEAP-zeocin gene. As indicated in Table 2, the elimination of low-producing clones by the second antibiotic selection step increases the predictability of finding high-producing clones; when STAR18 is included in the expression unit, this increased predictability is improved from three-fold to thirty-fold. In summary, when STAR elements are used in combination with two-step antibiotic selection, the predictability of finding clones with high yields of a

transgene is dramatically improved. Application of this increased predictability to two or more transgenes simultaneously will significantly increase the likelihood of finding clones that have high yields of multimeric proteins.

Example 2: Simultaneous expression of two proteins is improved by two-step selection and STAR elements

A second object of this invention is to improve the expression of heterologous multimeric proteins such as antibodies. This example demonstrates that the combination of STAR elements and two-step antibiotic selection improves the predictability of establishing recombinant host cell lines that express balanced and proportional amounts of two heterologous polypeptides at high yields. This method of the invention is applicable in practice to multimeric proteins such as antibodies. It is demonstrated in this example using two reporter proteins, secreted alkaline phosphatase (SEAP) and green fluorescent protein (GFP).

Materials and Methods

Plasmids

The pSDH-SIB/Z and pSDH-GIB/Z families of plasmids described in Example 1 are used. Cloning of STAR elements x and y, transfection and culture of host cells, and SEAP assay are described in Example 1. The assay for GFP is performed according to the manufacture's instructions.

Results

Results show an increased number of clones wherein the two reporter proteins are both expressed. Moreover, expression was balanced in many of such clones.

Example 3: General-purpose vectors for simultaneous expression of multiple polypeptides

The expression system tested and validated in Example 1 has been modified to facilitate its application to any polypeptide that is preferably co-expressed with another polypeptide or polypeptides in a host cell, for example the heavy and light chains of recombinant antibodies. It is designed for easy and rapid construction of the expression units. This improved system is described in this example.

Materials and Methods

Plasmids

The construction of the plasmids PP1 to PP5 is described below, and their map is shown in FIG 5. Plasmid pd2EGFP (Clontech 6010-1) was modified by insertion of a linker at the *Bsi*WI site to yield pd2EGFP-link. The linker (made by annealing oligonucleotides F25 and F26) introduces sites for the *Pac*I, *Bgl*II, and *Eco*RV restriction endonucleases. This creates the multiple cloning site MCSII for insertion of STAR elements. Then primers F23 and F24 were used to amplify a region of 0.37 kb from pd2EGFP, which was inserted into the *Bgl*II site of pIRES (Clontech 6028-1) to yield pIRES-stuf. This introduces sites for the *Asc*I and *Swa*I restriction endonucleases at MCSI, and acts as a "stuffer fragment" to avoid potential interference between STAR elements and adjacent promoters. pIRES-stuf was digested with *Bgl*II and *Fsp*I to liberate a DNA fragment composed of the stuffer fragment, the CMV promoter, the IRES element (flanked by multiple cloning sites MCS A and MCS B), and the SV40 polyadenylation signal. This fragment was ligated with the vector backbone of pd2EGFP-link produced by digestion with *Bam*HI and *Stu*I, to yield pd2IRES-link.

The open reading frames of the zeocin-, neomycin, or puromycin-resistance genes were inserted into the *Bam*HI/*Not*I sites of MCS B in pd2IRES-link as follows: the zeocin-resistance ORF was amplified by PCR

with primers F18 and E100 from plasmid pEM7/zeo, digested with *Bam*HI and *Not*I, and ligated with *Bam*HI/*Not*I-digested pd2IRES-link to yield pd2IRES-link-zeo. The neomycin-resistance ORF was amplified by PCR with primers F19 and F20 from pIRES, digested with *Bam*HI and *Not*I, and ligated with *Bam*HI/*Not*I-digested pd2IRES-link to yield pd2IRES-link-neo. The puromycin-resistance ORF was amplified by PCR with primers F21 and F22 from plasmid pBabe-Puro (Morgenstern & Land, 1990), digested with *Bam*HI and *Not*I, and ligated with *Bam*HI/*Not*I-digested pd2IRES-link to yield pd2IRES-link-puro.

The GFP reporter ORF was introduced into pd2IRES-link-puro by amplification of phr-GFP-1 with primers F16 and F17, and insertion of the *Eco*RI-digested GFP cassette into the *Eco*RI site in MCS A of the pd2IRES-link-puro plasmid, to yield plasmid PP1 (FIG 5A). Correct orientation was verified by restriction mapping. The SEAP reporter ORF was introduced into pd2IRES-link-zeo and pd2IRES-link-neo by PCR amplification of pSEAP2-basic with primers F14 and F15, and insertion of the *Eco*RI-digested SEAP cassette into the *Eco*RI sites in MCS A of the plasmids pd2IRES-link-zeo (to yield plasmid PP2, FIG 5B) and pd2IRES-link-neo (to yield plasmid PP3, FIG 5C). Correct orientation was verified by restriction mapping.

Plasmids PP1, PP2 and PP3 contain a bicistronic gene for expression of a reporter protein and and antibiotic resistance marker. In order to carry out two-step antibiotic selection with separate antibiotics, a monocistronic resistance marker was introduced as follows: pIRES-stuf was digested with *Cla*I, made blunt with Klenow enzyme, and digested further with *Bgl*II. This liberated a DNA fragment composed of the stuffer fragment, the CMV promoter, the IRES element (flanked by multiple cloning sites MCS A and MCS B), the SV40 polyadenylation signal, and the neomycin resistance marker under control of the SV40 promoter. This fragment was ligated with the vector backbone of pd2EGFP-link produced by digestion with *Bam*HI and *Stu*I, to yield pd2IRES-link-neo. Then as described above the GFP and puro cassettes

were introduced to yield PP4 (FIG 5D), and the SEAP and zeo cassettes were introduced to yield PP5 (FIG 5E).

Example 4: 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 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 1), and insertion of the *Sac*II/*Bam*HI fragment into *Sac*II/*Bam*HI-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 *Asc*I site was then introduced into MCSI by digestion with *Eco*RI 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 *Sal*I/*Sac*II 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 *Sac*II and *Bam*HI and made blunt; the SEAP open reading frame was isolated from pSEAP-basic (Clontech 6037-1) by *Eco*RI/*Sal*I 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 (*Bam*HI site removed) (Promega E1741) digested with *Nco*I/*Xba*I, to create pGL3-puro. pGL3-puro was digested with *Bgl*II/*Sal*I to isolate the SV40-puro resistance gene, which was made blunt and ligated into *Nhe*I 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 1), 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% CO_2. 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 5: 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 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 4. 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 4. 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 proteins is possible without incurring the costs of the antibiotic or of difficult downstream processing

Example 6: Minimal essential sequences of STAR elements

STAR elements are isolated from the genetic screen as described herein. 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 *Bam*HI site, and transfected into U-2 OS/Tet-Off/LexA-HP1 cells. 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). 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 7: 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 4 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 8: 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 4. pSDH-SV40 is, amongst others, derived from pSelect-SV40-zeo. The selection vector for STAR elements, pSelect-SV40-zeo 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.

pSDH-SV40 was constructed by PCR amplification of the SV40 promoter (primers D41 and D42) from plasmid pSelect-SV40-Zeo, followed by digestion of the PCR product with *Sac*II and *Sal*I. The pSDH-CMV vector was digested with *Sac*II and *Sal*I 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 4. 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 4.

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 4 and 7, 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 9: 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 4. U-2 OS cells were co-transfected with plasmids pSDH-Tet-STAR66-native and pSDH-Tet-STAR66-opposite, and cultivated as described in Example 4. 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 10: 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 *Pvu*II and *Xho*I 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 *Bam*HI/*Sac*II-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 11: 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 4). 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 12: 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 20; 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 *Sau*3AI 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 20, 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 20, 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 20, 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 20, 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 13: 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 herein and in EP 01202581.3, 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 *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 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 *Bam*HI-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 1), 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 21 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 9 (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 4). 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 4 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 5. 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 4) or dyad (Table 5) 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 (Table 4); those from dyad analysis are numbers 73, 194, 419, and 497 (Table 5).

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 22). 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 9. 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 9 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 14; 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 *Mbo*I. 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 94°C for 1 minute, 54°C for 1minute, and 72°C 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 10; 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 23) 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.

Example 15: STAR-shielded genes that reside on multiple vectors are expressed simultaneously in CHO cells

STAR elements function to block the effect of transcriptional repression influences on transgene expression units. One of the benefits of STAR elements for heterologous protein production is the increased predictability of finding high-expressing primary recombinant host cells. This feature allows for the simultaneous expression of different genes that reside on multiple, distinct vectors. In this example we use two different STAR7-shielded genes, GFP and RED, which are located on two different vectors. When these two vectors are transfected simultaneously to Chinese hamster ovary (CHO) cells, both are expressed, whereas the corresponding, but unprotected GFP and RED genes, show hardly such simultaneous expression.

Material and Methods

The STAR7 element is tested in the ppGIZ-STAR7 and ppRIP-STAR7 vectors (FIG 24). The construction of the pPlug&Play (ppGIZ and ppRIP) vectors is described below. Plasmid pGFP (Clontech 6010-1) is modified by insertion of a linker at the *Bsi*WI site to yield pGFP-link. The linker (made by annealing oligonucleotides 5'GTACGGATATCAGATCTTTAATTAAG3' and 5'GTACCTTAATTAAAGATCTGATATCC3') introduces sites for the *Pac*I, *Bgl*II, and *Eco*RV restriction endonucleases. This creates the multiple cloning site MCSII for insertion of STAR elements. Then primers (5'GATCAGATCTGGCGCGCCATTTAAATCGTCTCGCGCGTTTCGGTGATGACGG3') and (5'AGGCGGATCCGAATGTATTTAGAAAAATAAACAAATAGGGG3') are used to amplify a region of 0.37 kb from pGFP, which is inserted into the *Bgl*II site of pIRES (Clontech 6028-1) to yield pIRES-stuf. This introduces sites for the

*Asc*I and *Swa*I restriction endonucleases at MCSI, and acts as a "stuffer fragment" to avoid potential interference between STAR elements and adjacent promoters. pIRES-stuf is digested with *Bgl*II and *Fsp*I to liberate a DNA fragment composed of the stuffer fragment, the CMV promoter, the IRES element (flanked by multiple cloning sites MCS A and MCS B), and the SV40 polyadenylation signal. This fragment is ligated with the vector backbone of pGFP-link produced by digestion with *Bam*HI and *Stu*I, to yield pIRES-link.

The open reading frames of the zeocin-resistance gene is inserted into the *Bam*HI/*Not*I sites of MCS B in pIRES-link as follows: the zeocin-resistance ORF is amplified by PCR with primers 5'GATCGGATCCTTCGAAATGGCCAAGTTGACCAGTGC3' and 5'AGGCGCGGCCGCAATTCTCAGTCCTGCTCCTC3' from plasmid pEM7/zeo, digested with *Bam*HI and *Not*I, and ligated with *Bam*HI/*Not*I-digested pIRES-link to yield pIRES-link-zeo. The GFP reporter ORF is introduced into pIRES-link-zeo by amplification of phr-GFP-1 with primers 5'GATCGAATTCTCGCGAATGGTGAGCAAGCAGATCCTGAAG3' and 5'AGGCGAATTCACCGGTGTTTAAACTTACACCCACTCGTGCAGGCTGCCC AGG3', and insertion of the *Eco*RI-digested GFP cassette into the *Eco*RI site in MCS A of the pIRES-link-zeo plasmid. This creates the ppGIZ (for ppGFP-IRES-zeo) 5' STAR7 is cloned into the *Sal*I site and 3' STAR7 is cloned into the *Pac*I site.

The puromycin-resistance ORF is amplified by PCR with primers 5'GATCGGATCCTTCGAAATGACCGAGTACAAGCCCACG3' and 5'AGGCGCGGCCGCTCAGGCACCGGGCTTGCGGGTC3' from plasmid pBabe-Puro (Morgenstern & Land, 1990), digested with *Bam*HI and *Not*I, and ligated with *Bam*HI/*Not*I-digested pIRES-link to yield pIRES-link-puro. The RED gene is amplified by PCR with primers 5'GATCTCTAGATCGCGAATGGCCTCCTCCGAGAACGTCATC3' and 5'AGGCACGCGTTCGCGACTACAGGAACAGGTGGTGGCG3' from plasmid pDsRed2 (Clontech 6943-1), digested with *Xba*I and *Mlu*I and ligated to *Nhe*I-

*Mlu*I digested pIRES-link-puro to yield ppRIP (for ppRED-IRES-puro). 5' STAR7 is cloned into the *Sal*I site and 3' STAR7 is cloned into the *Pac*I site.

Transfection and culture of CHO cells

The Chinese Hamster Ovary cell line CHO-K1 (ATCC CCL-61) is cultured in HAMS-F12 medium + 10% Fetal Calf Serum containing 2 mM glutamine, 100 U/ml penicillin, and 100 micrograms/ml streptomycin at 37° C/5% CO_2. Cells are transfected with the plasmids using Lipofectamine 2000 (Invitrogen) as described by the manufacturer. Briefly, cells are seeded to culture vessels and grown overnight to 70-90% confluence. Lipofectamine reagent is combined with plasmid DNA at a ratio of 7.5 microliters per 3 microgram (e.g. for a 10 cm Petri dish, 20 micrograms DNA and 120 microliters Lipofectamine) and added after a 30 minutes incubation at 25^{0}C to the cells. After a 6 hour incubation the transfection mixture is replaced with fresh medium, and the transfected cells are incubated further. After overnight cultivation, cells are trypsinized and seeded into fresh petri dishes with fresh medium with zeocin added to a concentration of 100 μg/ml and the cells are cultured further. When individual colonies become visible (approximately ten days after transfection) medium is removed and replaced with fresh medium (puromycin)

Individual colonies are isolated and transferred to 24-well plates in medium with zeocin. Expression of the GFP and RED reporter genes is assessed approximately 3 weeks after transfection.

One tested construct consists of a monocistronic gene with the GFP gene, an IRES and the Zeocin resistance gene under control of the CMV promoter, but either with or without STAR7 element to flank the entire construct (FIG 24). The other construct consists of a monocistronic gene with the RED gene, an IRES and the puromycin resistance gene under control of the CMV promoter, but either with or without STAR7 element to flank the entire construct (FIG 24).

The constructs are transfected to CHO-K1 cells. Stable colonies that are resistant for both zeocin and puromycin are expanded before the GFP and RED signals are determined on a XL-MCL Beckman Coulter flowcytometer. The percentage of cells in one colony that are double positive for both GFP and RED signals is taken as measure for simultaneous expression of both proteins and this is plotted in FIG 24.

Results

FIG 24 shows that simultaneous expression in independent zeocin and puromycin resistant CHO colonies of GFP and a RED reporter genes that are flanked by a STAR element results in a higher number of cells that express both GFP and RED proteins, as compared to the control vectors without STAR7 element. The STAR7 element therefore conveys a higher degree of predictability of transgene expression in CHO cells. In the STAR-less colonies at most 9 out of 20 colonies contain double GFP/RED positive cells. The percentage of double positive cells ranges between 10 and 40%. The remaining 11 out of 20 colonies have less than 10% GFP/RED positive cells. In contrast, in 19 out of 20 colonies that contain the STAR-shielded GFP and RED genes, the percentage GFP/RED double positive cells ranges between 25 and 75%. In 15 out of these 19 double positive colonies the percentage GFP/RED double positive cells is higher than 40%. This result shows that it is more likely that simultaneous expression of two genes is achieved when these genes are flanked with STAR elements.

Example 16: Expression of a functional antibody from two separate plasmids is easier obtained when STAR elements flank the genes encoding the heavy and light chains.

Due to the ability of STAR elements to convey higher predictability to protein expression two genes can be expressed simultaneously from distinct vectors. This is shown in example 15 for two reporter genes, GFP and RED.

Now the simultaneous expression of a light and a heavy antibody chain is tested. In example 16, STAR7-shielded light and heavy antibody cDNAs that reside on distinct vectors are simultaneously transfected to Chinese hamster ovary cells. This results in the production of functional antibody, indicating that both heavy and light chains are expressed simultaneously. In contrast, the simultaneous transfection of unprotected light and heavy antibody cDNAs shows hardly expression of functional antibody.

Materials and Methods

The tested constructs are the same as described in Example 15, except that the GFP gene is replaced by the gene encoding the light chain of the RING1 antibody (Hamer et al., 2002) and the RED gene is replaced by the gene encoding the heavy chain of the RING1 antibody. The light chain is amplified from the RING1 hybridoma (Hamer et al., 2002) by RT-PCR using the primers 5'CAAGAATTCAATGGATTTTCAAGTGCAG3' and 5'CAAGCGG CCGCTTTGTCTCTAACACTCATTCC3'. The PCR product is cloned into pcDNA3 after restriction digestion with *Eco*RI and *Not*I and sequenced to detect potential frame shifts in the sequence. The cDNA is excised with *Eco*RI and *Not*I, blunted and cloned in ppGIZ plasmid. The heavy chain is amplified from the RING1 hybridoma (Hamer et al., 2002) by RT-PCR using the primers 5'ACAGAATTCTTACCATGGATTTTGGGCTG3' and 5'ACAGCGGCCGCTCATTTACCAGGAGAGTGGG3'. The PCR product is cloned into pcDNA3 after restriction digestion with *Eco*RI and *Not*I and sequenced to detect potential frame shifts in the sequence. The cDNA is excised with *Eco*RI and *Not*I, blunted and cloned in ppRIP plasmid.

Results

CHO colonies are simultaneously transfected with the RING1 Light Chain (LC) and RING1 Heavy Chain (HC) cDNAs that reside on two distinct vectors. The Light Chain is coupled to the zeocin resistance gene through an

IRES, the Heavy Chain is coupled to the puromycin resistance gene through an IRES. FIG 25 shows that simultaneous transfection to CHO cells of the heavy and light chain encoding cDNAs results in the establishment of independent zeocin and puromycin resistant colonies. When the constructs are flanked by the STAR7 element this results in a higher production of functional RING1 antibody, as compared to the control vectors without STAR7 element. The STAR7 element therefore conveys a higher degree of predictability of antibody expression in CHO cells.

In the STAR-less colonies only 1 out of 12 colonies express detectable antibody. In contrast, in 7 out of 12 colonies that contain the STAR-shielded Light and Heavy Chain genes, produce functional RING1 antibody that detects the RING1 antigen in an ELISA assay. Significantly, all these 7 colonies produce higher levels of RING1 antibody than the highest control colony (arbitrarily set at 100%). This result shows that it is more likely that simultaneous expression of two genes encoding two antibody chains is achieved when these genes are flanked with STAR elements.

Table 1. Oligonucleotides used for polymerase chain reactions (PCR primers) or DNA mutagenesis

Number	Sequence
C65	AACAAGCTTGATATCAGATCTGCTAGCTTGGTCGAGCTGATACTTCCC
C66	AAACTCGAGCGGCCGCGAATTCGTCGACTTTACCACTCCCTATCAGTGATAGAG
C67	AAACCGCGGCATGGAAGACGCCAAAAACATAAAGAAAGG
C68	TATGGATCCTAGAATTACACGGCGATCTTTCC
C81	AAACCATGGCCGAGTACAAGCCCACGGTGCGCC
C82	AAATCTAGATCAGGCACCGGGCTTGCGGGTCATGC
C85	CATTTCCCCGAAAAGTGCCACC
D30	TCACTGCTAGCGAGTGGTAAACTC
D41	GAAGTCGACGAGGCAGGCAGAAGTATGC
D42	GAGCCGCGGTTTAGTTCCTCACCTTGTCG
D51	TCTGGAAGCTTTGCTGAAGAAAC
D58	CCAAGTTGACCAGTGCC
D70	TACAAGCCAACCACGGCCT
D71	CGGAAGTGCTTGACATTGGG
D80	GTTCGTGGACACGACCTCCG
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
E84	GATCTCTAGAATGGCCAAGCCTTTGTCTCAAG
E85	AGGCGCGGCCGCTTAGCCCTCCCACACATAACCAGAG
E87	AGGCACGCGTTCATGTCTGCTCGAAGCGGCC

E92	AGGCGCTAGCACGCGTTCTACTCTTTTCCTACTCTG
E93	GATCAAGCTTACGCGTCTAAAGGCATTTTATATAG
E94	AGGCGCTAGCACGCGTTCAGAGTTAGTGATCCAGG
E95	GATCAAGCTTACGCGTCAGTAAAGGTTTCGTATGG
E96	AGGCGCTAGCACGCGTTCTACTCTTTCATTACTCTG
E97	CGAGGAAGCTGGAGAAGGAGAAGCTG
E98	CAAGGGCCGCAGCTTACACATGTTC
E99	GATCACTAGTATGGCCAAGTTGACCAGTGC
E100	AGGCGCGGCCGCAATTCTCAGTCCTGCTCCTC
F11	GATCGCTAGCAATCGCGACTTCGCCCACCATGC
F14	GATCGAATTCTCGCGACTTCGCCCACCATGC
F15	AGGCGAATTCACCGGTGTTTAAACTCATGTCTGCTCGAAGCGGCCGG
F16	GATCGAATTCTCGCGAATGGTGAGCAAGCAGATCCTGAAG
F17	AGGCGAATTCACCGGTGTTTAAACTTACACCCACTCGTGCAGGCTGCCCAGG
F18	GATCGGATCCTTCGAAATGGCCAAGTTGACCAGTGC
F19	GATCGGATCCTTCGAAATGATTGAACAAGATGGATTGC
F20	AGGCGCGGCCGCTCAGAAGAACTCGTCAAGAAGGCG
F21	GATCGGATCCTTCGAAATGACCGAGTACAAGCCCACG
F22	AGGCGCGGCCGCTCAGGCACCGGGCTTGCGGGTC
F23	GATCAGATCTGGCGCGCCATTTAAATCGTCTCGCGCGTTTCGGTGATGACGG
F24	AGGCGGATCCGAATGTATTTAGAAAAATAAACAAATAGGGG
F25	GTACGGATATCAGATCTTTAATTAAG
F26	GTACCTTAATTAAAGATCTGATATCC
F32	GATCGAGGTACCGGTGTGT
F33	GATCACACACCGGTACCTC
F34	CGGAGGTACCGGTGTGT
F35	CGACACACCGGTACCTC
F44	TGAGAGGTACCGGTGTGT
F45	TCAACACACCGGTACCTC

Table 2. STAR elements and two-step selection increase the predictability of transgene expression

	without STAR	fold improvement	with STAR
(carry out first antibiotic selection)			
Number colonies[1]	~100	10-fold	~1000
High producers			
percent	5%	3-fold	15%
number	5		150
(characterize 20 colonies)	(20% of population)		(2% of population)
High producers	1	3-fold[2]	3
Low producers	19		17
(carry out second antibiotic selection, killing low producers)			
Survivors to characterize	5	30-fold[3]	150

[1]Colonies per microgram plasmid DNA.
[2]Manifesting the three-fold improvement due to the presence of STARs in the percent of high producers in the original population of colonies resistant to the first antibiotic.
[3]Manifesting the arithmetic product of the fold improvement in the number of colonies and the increased percentage of high producers due to the presence of STARs.

Table 3. Sequences of various STAR elements

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
GGTAAAGGCAGTGAGACCACGTAGGGGTCATTGCAGTAATCCAGGCTGGA
GATGATGGTGGTTCAGTTGGAATAGCAGTGCATGTGCTGTAACAACCTCA
GCTGGGAAGCAGTATATGTGGCGTTATGACCTCAGCTGGAACAGCAATGC
ATGTGGTGGTGTAATGACCCCAGCTGGGTAGGGTGCATGTGATGGAACAA
CCTCAGCTGGGTAGCAGTGTACTTGATAAAATGTTGGCATACTCTACATTT
GTTATGAGGGTAGTGCCATTAAATTTCTCCACAAATTGGTTGTCACGTATG
AGTGAAAAGAGGAAGTGATGGAAGACTTCAGTGCTTTTGGCCTGAATAAA
TAGAAGACGTCATTTTCAGTAATGGAGACAGGGAAGACTAANGNAGGGT
GGATTCAGTAGAGCAGGTGTTCAGTTTTGAATATGATGAACTCTGAGAGA
GGAAAAACTTTTTCTACCTCTTAGTTTTTGNGNCTGGACTTAANATTAAAG
GACATANGACNGAGANCAGACCAAATNTGCGANGTTTTTATATTTTACTT
GCNGAGGGAATTTNCAAGAAAAAGAAGACCCAANANCCATTGGTCAAAA
CTATNTGCCTTTTAANAAAAAGANAATTACAATGGANANANAAGTGTTGN
CTNGGCAAAAATTGGG

STAR4 reverse

GGATTNGAGCTAGCGGCGCGCCAAGCTTGGATCTTAGAAGGACAGAGTG
GGGCATGGAAATGCACCACCAGGGCAGTGCAGCTTGGTCACTGCCAGCTC
CNCTCATGGGCAGAGGGCTGGCCTCTTGCAGCCGACCAGGCACTGAGCG
CCATCCCAGGGCCCTCGCCAGCCCTCAGCAGGGCCAGGACACACAAGCCT
TTGACTTCCTCCTGTCACTGCTGCTGCCATTCCTGTTTTGTGGTCATCACT
CCTTCCCTGTCCTCAGACTGCCCAGCACTCAAGGATGTCCTGTGGTGGCA
TCAGACCATATGCCCCTGAANAGGAGTGAGTTGGTGTTTTTTGCCGCGCC

CANAGAGCTGCTGTCCCCTGAAAGATGCAAGTGGGAATGATGATGNTCAC
CATCNTCTGACACCAAGCCCTTTGGATAGAGGCCCCAACAGTGAGGATGG
GGCTGCACTGCATTGCCAAGGCAACTCTGTNNTGACTGCTACANGACANT
CCCAGGACCTGNGAAGNNCTATANATNTGATGCNAGGCACCT

STAR6 forward

CCACCACAGACATCCCCTCTGGCCTCCTGAGTGGTTTCTTCAGCACAGCTT
CCAGAGCCAAATTAAACGTTCACTCTATGTCTATAGACAAAAAGGGTTTTG
ACTAAACTCTGTGTTTTAGAGAGGGAGTTAAATGCTGTTAACTTTTTAGGG
GTGGGCGAGAGGAATGACAAATAACAACTTGTCTGAATGTTTTACATTTCT
CCCCACTGCCTCAAGAAGGTTCACAACGAGGTCATCCATGATAAGGAGTA
AGACCTCCCAGCCGGACTGTCCCTCGGCCCCCAGAGGACACTCCACAGAG
ATATGCTAACTGGACTTGGAGACTGGCTCACACTCCAGAGAAAAGCATGG
AGCACGAGCGCACAGAGCANGGGCCAAGGTCCCAGGGACNGAATGTCTA
GGAGGGAGATTGGGGTGAGGGTANTCTGATGCAATTACTGNGCAGCTCA
ACATTCAAGGGAGGGGAAGAAAGAAACNGTCCCTGTAAGTAAGTTGTNCA
NCAGAGATGGTAAGCTCCAAATTTNAACTTTGGCTGCTGGAAAGTTTNNG
GGCCNANANAANAAACANAAANATTTGAGGTTTANACCCACTAACCCNT
ATNANTANTTATTAATACCCCTAATTANACCTTGGATANCCTTAAAATATC
NTNTNAAACGGAACCCTCNTTCCCNTTTNNAAATNNNAAAGGCCATTNN
GNNCNAGTAAAAATCTNNNTTAAGNNNTGGGCCCNAACAAACNTNTTCC
NAGACACNTTTTTTNTCCNGGNATTTNTAATTTATTTCTAANCC

STAR6 reverse

ATCGTGTCCTTTCCAGGGACATGGATGAAGCTGGAAGCCATCATCCTCAG
CAAACTAACACAGGAACAGAAAACCAAATACCACATGTTCTCACTCATAA
GTGGGAGCTGAACAGTGAGAACACATGGACACAGGGAGGGGAACATCAC
ACACCAAGGCCTGTCTGGTGTGGGGAGGGGAGGGAGAGCATCAGGACAA
ATAGCTAATGCATGTGGGGCTTAAACCTAGATGACGGGTTGATAGGTGCA

GCAATCCACTATGGACACATATACCTATGTAACAACCCNACCTTNTTGAC
ATGTATCCCAGAACTTAAAGGAAAATAAAAATTAAAAAAAATTNCCCTGG
AATAAAAAAGAGTGTGGACTTTGGTGAGATN

STAR8 forward

GGATCACCTCGAAGAGAGTCTAACGTCCGTAGGAACGCTCTCGGGTTCAC
AAGGATTGACCGAACCCCAGGATACGTCGCTCTCCATCTGAGGCTTGNTC
CAAATGGCCCTCCACTATTCCAGGCACGTGGGTGTCTCCCCTAACTCTCC
CTGCTCTCCTGAGCCCATGCTGCCTATCACCCATCGGTGCAGGTCCTTTCT
GAANAGCTCGGGTGGATTCTCTCCATCCCACTTCCTTTCCCAAGAAAGAA
GCCACCGTTCCAAGACACCCAATGGGACATTCCCNTTCCACCTCCTTNTC
NAAAGTTNGCCCAGGTGTTCNTAACAGGTTAGGGAGAGAANCCCCCAGG
TTTNAGTTNCAAGGCATAGGACGCTGGCTTGAACACACACACACNCTC

STAR8 reverse

GGATCCCGACTCTGCACCGCAAACTCTACGGCGCCCTGCAGGACGGCGGC
CTCCTGCCGCTTGGACGCCAGNCAGGAGCTCCCCGGCAGCAGCAGAGCA
GAAAGAAGGATGGCCCCGCCCCACTTCGCCTCCCGGCGGTCTCCCTCCCG
CCGGCTCACGGACATAGATGGCTGCCTAGCTCCGGAAGCCTAGCTCTTGT
TCCGGGCATCCTAAGGAAGACACGGTTTTTCCTCCCGGGGCCTCACCACA
TCTGGGACTTTGACGACTCGGACCTCTCTCCATTGAATGGTTGCGCGTTCT
CTGGGAAAG

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. 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 5. 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 6. STAR elements, 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

STAR	Location[1]	Length[2]
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
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

STAR	Location[1]	Length[2]
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

STAR	Location[1]	Length[2]
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 of EP 01202581.3.

[4]The STARs with these numbers in Tables 2 and 4 of EP 01202581.3 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"of EP 01202581.3.

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. 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 4.

[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. 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 10. 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

DESCRIPTION OF FIGURES

The drawings show representative versions of the DNA molecules of the invention. These portions of DNA, referred to as (a) protein expression unit(s), is/are created and manipulated in vectors such as recombinant plasmid molecules and/or recombinant viral genomes. The protein expression units are integrated into host cell genomes as part of the method of the invention, and the schematic drawings represent the configuration of the DNA elements in the expression units in both the vector molecules and the host cell genome.

FIG 1. Schematic diagram of the invention.

FIG 1A shows the first expression unit. It is flanked by STAR elements, and comprises a bicistronic gene containing (from 5' to 3') a transgene (encoding for example a reporter gene or one subunit of a multimeric protein; TG S1, "transgene subunit 1"), an IRES, and a selectable marker (zeo, conferring zeocin resistance) under control of the CMV promoter. A monocistronic selectable marker (neo, conferring G418 resistance) under control of the SV40 promoter is included. Both genes have the SV40 transcriptional terminator at their 3' ends (t).

FIG 1B shows the second expression unit. It is flanked by STAR elements, and contains a bicistronic gene containing (from 5' to 3') a transgene (encoding for example a different reporter gene or another subunit of a multimeric protein; TG S2), an IRES, and a selectable marker (bsd, conferring blasticidin resistance) under control of the CMV promoter. A monocistronic selectable marker (neo, conferring G418 resistance) under control of the SV40 promoter is included. Both genes have the SV40 transcriptional terminator at their 3' ends.

FIG 2. The pSDH-CSP plasmid.

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 (amp^R) for propagation in *Escherichia coli*.

FIG 3. The pSDH-SIB/Z and pSDH-GIB/Z families of plasmids.

These plasmids are derived from the pSDH-CSP plasmid (FIG 2), by replacement of the monocistronic SEAP and puro genes with a bicistronic gene under control of the CMV promoter and a monocistronic neomycin resistance selectable marker gene (neo) under control of the SV40 promoter.

Panel A, pSDH-SIB/Z in which the bicistronic gene encodes secreted alkaline phosphatase (SEAP) in the 5' position and blasticidin (bsd) or zeocin (zeo) resistance selectable markers in the 3' position, relative to the internal ribosome binding site (IRES).

Panel B, pSDH-GIB/Z in which the bicistronic gene encodes green fluorescent protein (GFP) in the 5' position and blasticidin (bsd) or zeocin (zeo) resistance selectable markers in the 3' position, relative to the internal ribosome binding site (IRES).

FIG 4. Comparison of the consequences of one-step and two-step antibiotic selection on the predictability of transgene expression.

Recombinant CHO cell isolates containing plasmid pSDH-SIZ or plasmid pSDH-SIZ-STAR18 were selected on G418 (panel A) or sequentially on G418 and zeocin (panel B) and assayed for SEAP activity.

FIG 5. The PP (Plug and Play) family of plasmids.

These plasmids contain a bicistronic expression unit (containing an internal ribosome binding site, IRES) between multiple cloning sites (MCS) for

insertion of STAR elements. MCSI, SbfI-SalI-XbaI-AscI-SwaI; MCSII, BsiWI-EcoRV-BglII-PacI.

Panel A, the bicistronic gene encodes green fluorescent protein (GFP) and the puromycin resistance marker (puro).

Panel B, the bicistronic gene encodes secreted alkaline phosphatase (SEAP) and the zeocin resistance marker (zeo).

Panel C, the bicistronic gene encodes SEAP and the neocin resistance marker (neo).

Panel D, the bicistronic gene encodes GFP and puro, and an adjacent monocistronic gene encodes neo.

Panel E, the bicistronic gene encodes SEAP and zeo, and an adjacent monocistronic gene encodes neo.

Bicistronic genes are under control of the CMV promoter (pCMV) and the monocistronic gene is under control of the SV40 promoter (pSV40). A stuffer fragment of 0.37 kb (St) separates MCSI from pCMV. Both the bicistronic and monocistronic genes have the SV40 polyadenylation site at their 3' ends.

FIG 6. STAR sequences

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).

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. STAR element orientation

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. 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 21.

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

FIG 22. 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 23. RT-PCR assay of Arabidopsis STAR strength

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.

FIG. 24. STAR elements allow efficient and simultaneous expression of two genes from two distinct vectors.

The ppGIZ, ppGIZ-STAR7, ppRIP and ppRIP-STAR7 vectors used for testing simultaneous expression of respectively GFP and RED are shown. The expression unit comprises (from 5' to 3') genes encoding the GFP or RED proteins, an IRES, and a selectable marker (zeo, conferring zeocin resistance or respectively puro, puromycin resistance gene) under control of the CMV promoter. The expression unit has the SV40 transcriptional terminator at its 3' end (t). The cassettes with the GFP and RED expression units are either flanked by STAR7 elements (STAR7-shielded) or not (Control). The two control constructs or the two STAR7-shielded vectors are simultaneously transfected to CHO-K1 cells. Stable colonies that are resistant to both zeocin and puromycin are expanded and the GFP and RED signals are determined on a XL-MCL Beckman Coulter flowcytometer. The percentage of cells in one colony that are double positive for both GFP and RED signals is taken as measure for simultaneous expression of both proteins and this is plotted in FIG 24.

FIG. 25. STAR elements improve expression of a functional antibody in CHO cells.

The different vectors containing the Light and Heavy Chain of the RING1 antibody are shown in FIG 25. The constructs are simultaneously transfected to CHO cells. Stable colonies that are resistant to both zeocin and puromycin are expanded. The cell culture medium of these colonies is tested for the detection of functional RING1 antibody in an ELISA with RING1 protein as antigen. The values are dividing by the number of cells in the colony. The highest value detected in the STAR-less control is arbitrarily set at 100%.

REFERENCES

Aranda, A, and Pascual, A. (2001) Nuclear hormone receptors and gene expression *Physiol Rev 81*, 1269-304.

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.

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

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.

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.

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.

Chan, A, and Mak, TW. (1989) Genomic organization of the T cell receptor *Cancer Detect Prev 14*, 261-7.

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.

Chevet, E, Cameron, PH, Pelletier, MF, Thomas, DY, and Bergeron, JJ. (2001) The endoplasmic reticulum: integration of protein folding, quality control, signaling and degradation *Curr Opin Struct Biol 11*, 120-4.

Das, GC, Niyogi, SK, and Salzman, NP. (1985) SV40 promoters and their regulation *Prog Nucleic Acid Res Mol Biol 32*, 217-36.

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.

Eszterhas, SK, Bouhassira, EE, Martin, DI, and Fiering, S. (2002) Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position *Mol Cell Biol 22*, 469-79.

European patent application 01202581.3

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.

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

Gill, DR, Smyth, SE, Goddard, CA, Pringle, IA, Higgins, CF, Colledge, WH, and Hyde, SC. (2001) Increased persistence of lung gene expression using plasmids containing the ubiquitin C or elongation factor 1alpha promoter *Gene Ther 8*, 1539-46.

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*, 5547-51.

Groeneveld, EH, and Burger, EH. (2000) Bone morphogenetic proteins in human bone regeneration *Eur J Endocrinol 142*, 9-21.

Hamer, CM, Sewalt, RGAB, Den Blaauwen, JL, Hendrix, M, Satijn, DPE, and Otte, AP. (2002). A panel of monoclonal antibodies against human Polycomb group proteins. Hybridoma and Hybridomics 21, 245-52.

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.

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.

Hynes, RO. (1999) Cell adhesion: old and new questions *Trends Cell Biol 9*, M33-7.

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.

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

Kaufman, RJ. (2000) Overview of vector design for mammalian gene expression *Mol Biotechnol 16*, 151-60.

Kaufman, RJ. (1990) Selection and coamplification of heterologous genes in mammalian cells *Methods in Enzymology 185*, 536-566.

Kaufman, RJ, and Sharp, PA. (1982) Construction of a modular dihydrofolate reductase cDNA gene: analysis of signals utilized for efficient expression *Mol Cell Biol 2*, 1304-19.

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.

Liu, DT. (1992) Glycoprotein pharmaceuticals: scientific and regulatory considerations, and the US Orphan Drug Act *Trends Biotechnol 10*, 114-20.

Lopez de Quinto, S, and Martinez-Salas, E. (1998) Parameters influencing translational efficiency in aphthovirus IRES- based bicistronic expression vectors *Gene 217*, 51-6.

Martin, DI, and Whitelaw, E. (1996) The vagaries of variegating transgenes *Bioessays 18*, 919-23.

Martinez-Salas, E. (1999) Internal ribosome entry site biology and its use in expression vectors *Curr Opin Biotechnol 10*, 458-64.

McBurney, MW, Mai, T, Yang, X, and Jardine, K. (2002) Evidence for repeat-induced gene silencing in cultured Mammalian cells: inactivation of tandem repeats of transfected genes *Exp Cell Res 274*, 1-8.

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

Migliaccio, AR, Bengra, C, Ling, J, Pi, W, Li, C, Zeng, S, Keskintepe, M, Whitney, B, Sanchez, M, Migliaccio, G, and Tuan, D. (2000) Stable and unstable transgene integration sites in the human genome: extinction of the Green Fluorescent Protein transgene in K562 cells *Gene 256*, 197-214.

Mizuguchi, H, Xu, Z, Ishii-Watabe, A, Uchida, E, and Hayakawa, T. (2000) IRES-dependent second gene expression is significantly lower than cap-dependent first gene expression in a bicistronic vector *Mol Ther 1*, 376-82.

Morgenstern, JP, 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*, 3587-96.

Pahl, HL, and Baeuerle, PA. (1997) The ER-overload response: activation of NF-kappa B *Trends Biochem Sci 22*, 63-7.

Patil, C, and Walter, P. (2001) Intracellular signaling from the endoplasmic reticulum to the nucleus: the unfolded protein response in yeast and mammals *Curr Opin Cell Biol 13*, 349-55.

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.

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

Rees, S, Coote, J, Stables, J, Goodson, S, Harris, S, and Lee, MG. (1996) Bicistronic vector for the creation of stable mammalian cell lines that predisposes all antibiotic-resistant cells to express recombinant protein *Biotechniques 20*, 102-4, 106, 108-10.

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.

Sambrook, J, Fritsch, EF, and Maniatis, T (1989) *Molecular Cloning: A Laboratory Manual*, Second ed., Cold Spring Harbor Laboratory Press, Plainview NY.

Sanger, F, Nicklen, S, and Coulson, AR. (1977) DNA sequencing with chain-terminating inhibitors *Proc Natl Acad Sci U S A 74*, 5463-7.

Schorpp, M, Jager, R, Schellander, K, Schenkel, J, Wagner, EF, Weiher, H, and Angel, P. (1996) The human ubiquitin C promoter directs high ubiquitous expression of transgenes in mice *Nucleic Acids Res 24*, 1787-8.

Sheeley, DM, Merrill, BM, and Taylor, LC. (1997) Characterization of monoclonal antibody glycosylation: comparison of expression systems and identification of terminal alpha-linked galactose *Anal Biochem 247*, 102-10.

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.

Strutzenberger, K, Borth, N, Kunert, R, Steinfellner, W, and Katinger, H. (1999) Changes during subclone development and ageing of human antibody-producing recombinant CHO cells *J Biotechnol 69*, 215-26.

Thotakura, NR, and Blithe, DL. (1995) Glycoprotein hormones: glycobiology of gonadotrophins, thyrotrophin and free alpha subunit *Glycobiology 5*, 3-10.

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, JL, Sewalt, RG, van Driel, R, and Otte, AP. (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.

Venkatesan, A, and Dasgupta, A. (2001) Novel fluorescence-based screen to identify small synthetic internal ribosome entry site elements *Mol Cell Biol 21*, 2826-37.

Villemure, JF, Savard, N, and Belmaaza, A. (2001) Promoter Suppression in Cultured Mammalian Cells can be Blocked by the Chicken beta-Globin Chromatin Insulator 5'HS4 and Matrix/Scaffold Attachment Regions *J Mol Biol 312*, 963-74.

Whitelaw, E, Sutherland, H, Kearns, M, Morgan, H, Weaving, L, and Garrick, D. (2001) Epigenetic effects on transgene expression *Methods Mol Biol 158*, 351-68.

Wright, A, and Morrison, SL. (1997) Effect of glycosylation on antibody function: implications for genetic engineering *Trends Biotechnol 15*, 26-32.

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.

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.

Claims

1. A cell comprising two polypeptide expression units each encoding at least one polypeptide of interest, characterized in that said polypeptide expression units each comprise at least one sequence having the capacity to at least in part block chromatin-associated repression, wherein said sequence having the capacity to at least in part block chromatin-associated repression is chosen from the group consisting of: (a) any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6; (b) a functional equivalent of any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6; and (c) a functional fragment of any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6.

2. A cell according to claim 1, wherein said two polypeptide expression units each further encode a different selection marker.

3. A cell according to claim 1 or 2, wherein at least one of said polypeptide expression units comprises a monocistronic gene comprising an open reading frame encoding a polypeptide of interest and wherein said monocistronic gene is under control of a functional promoter.

4. A cell according to claim 1 or 2, wherein at least one of said polypeptide expression units comprises a bicistronic gene comprising in the following order: (i) an open reading frame encoding a polypeptide of interest, (ii) an Internal Ribosome Entry Site (IRES), and (iii) a selection marker, and wherein said bicistronic gene is under control of a functional promoter.

5. A cell according to any one of the preceding claims, wherein at least one of said polypeptide expression units comprises at least two of said sequences having the capacity to at least in part block chromatin-associated repression, arranged such that said polypeptide expression unit is flanked on both sides by at least one of said sequences having the capacity to at least in part block chromatin-associated repression.

6. A cell according to claim 5, wherein said at least two sequences having the capacity to at least in part block chromatin-associated repression are essentially identical.

7. A cell according to any one of the preceding claims, wherein at least one polypeptide of interest comprises an immunoglobulin heavy chain, or an immunoglobulin light chain.

8. A cell according to claim 7, wherein at least one polypeptide of interest comprises an immunoglobulin heavy chain and the other polypeptide of interest comprises an immunoglobulin light chain, wherein said heavy and light chain can form a functional antibody.

9. A cell according to any one of the preceding claims, wherein said sequence having the capacity to at least in part block chromatin-associated repression is chosen from the group consisting of: (a) SEQ ID: 7 of Figure 6; (b) a functional equivalent of SEQ ID: 7 of Figure 6; and (c) a functional fragment of SEQ ID: 7 of Figure 6.

10. A method for expressing at least two polypeptides of interest in a cell, said method comprising culturing a cell according to any one of the preceding claims under conditions wherein said polypeptide expression units are expressed.

11. A polypeptide expression unit comprising:
- a bicistronic gene comprising in the following order: (i) an open reading frame encoding a polypeptide of interest, (ii) an Internal Ribosome Entry Site (IRES), and (iii) a selection marker, and wherein said bicistronic gene is under control of a functional promoter; and
- at least one sequence having the capacity to at least in part block chromatin-associated repression, wherein said sequence having the capacity to at least in part block chromatin-associated repression is chosen from the group consisting of: (a) any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6; (b) a functional equivalent of any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6; and (c) a functional fragment of any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6.

12. A polypeptide expression unit according to claim 11, comprising at least two of said sequences having the capacity to at least in part block chromatin-associated repression, arranged such that said polypeptide expression unit is flanked on both sides by at least one of said sequences having the capacity to at least in part block chromatin-associated repression.

13. A polypeptide expression unit according to claim 11 or 12, wherein said polypeptide of interest comprises an immunoglobulin heavy chain or an immunoglobulin light chain.

14. A polypeptide expression unit according to any one of claims 11 to 13, wherein sequence having the capacity to at least in part block chromatin-associated repression is chosen from the group consisting of: (a) SEQ ID: 7 of Figure 6; (b) a functional equivalent of SEQ ID: 7 of Figure 6; and (c) a functional fragment of SEQ ID: 7 of Figure 6.

15. A polypeptide expression unit according to any one of claims 11 to 14, wherein said selection marker encodes the zeocin-resistance protein.

16. A method for obtaining a host cell expressing two polypeptides of interest, the method comprising:
a) providing host cells comprising:
(i) a first polypeptide expression unit comprising a bicistronic gene comprising a promoter functionally linked to a sequence encoding a first polypeptide of interest and a first selectable marker gene, and
(ii) a second polypeptide expression unit comprising a bicistronic gene comprising a promoter functionally linked to a sequence encoding a second polypeptide of interest and a second selectable marker gene,
wherein said second selectable marker gene is different from said first selectable marker gene, and

wherein said first polypeptide expression unit, or said second polypeptide expression unit, or each of said first and said second polypeptide expression units comprise at least one sequence having the capacity to at least in part block chromatin-associated repression, wherein said sequence having the capacity to at least in part block chromatin-associated repression is chosen from the group consisting of: (a) any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6; (b) a functional equivalent of any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6; and (c) a functional fragment of any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6; and

b) selecting a host cell by selecting for expression of said first and second selectable marker genes.

17. A method for expressing two polypeptides of interest, the method comprising culturing a host cell selected by a method according to claim 16, to express said first and second polypeptides.

18. A method according to claim 17, wherein the two polypeptides of interest form part of a multimeric protein.

19. A method according to claim 17 or 18, wherein at least one of said two polypeptides of interest comprises an immunoglobulin heavy chain, or an immunoglobulin light chain.

20. A method according to claim 19, wherein at least one of said two polypeptides of interest comprises an immunoglobulin heavy chain and the other polypeptide of interest comprises an immunoglobulin light chain, wherein said heavy and light chain can form a functional antibody.

21. A method according to any one of claims 16-20, wherein at least one of said bicistronic genes comprises an internal ribosome entry site (IRES).

22. A set of two polypeptide expression units, said set comprising:

(i) a first polypeptide expression unit comprising a bicistronic gene comprising a promoter functionally linked to a sequence encoding a first polypeptide of interest and a first selectable marker gene, and

(ii) a second polypeptide expression unit comprising a bicistronic gene comprising a promoter functionally linked to a sequence encoding a second polypeptide of interest and a second selectable marker gene,

wherein said second selectable marker gene is different from said first selectable marker gene, and wherein

said first polypeptide expression unit, or said second polypeptide expression unit, or both said first and said second polypeptide expression units comprise at least one sequence having the capacity to at least in part block chromatin-associated repression,

wherein said sequence having the capacity to at least in part block chromatin-associated repression is chosen from the group consisting of: (a) any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6; (b) a functional equivalent of any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6; and (c) a functional fragment of any one of SEQ ID: 1 through SEQ ID: 65 of Figure 6.