# ASSESSING THE EFFICIENCY OF NOVEL GENE TRAP VECTORS IN MURINE EMBRYONIC STEM CELLS

**ANESTIS TSAKIRIDIS**

**PhD**
**The University of Edinburgh**
**2006**

I declare that the work presented in this thesis is my own, except where otherwise stated, and it has not been submitted for any other degree or professional qualification

Anestis Tsakiridis

To my family

## ACKNOWLEDGEMENTS

I would like to thank my supervisors Lesley Forrester and Josh Brickman for their advice, patience and support throughout the course of my PhD. Many thanks also go to Valerie Wilson and John Ansell for their valuable feedback as members of my PhD committee. Additionally, I would like to thank all my "comrades" at the John Hughes Bennett Laboratory, and especially Kay Samuel, Richard Axton, Alistair Watt, Anu Bashamboo, Melany Jakcson and Bernard Ramsahoye as well as Paul Perry (HGU MRC) for their advice and technical assistance in some of the experiments. A big thank you goes to Rosalind for her love, support and understanding.

# ABSTRACT

Gene trapping is a random insertional mutagenesis strategy that aims to identify novel genes and analyse their function. It usually involves the introduction into embryonic stem (ES) cells of promoterless reporter/selector gene constructs whose expression can be activated only after integration downstream of a gene's regulatory elements. Gene trap insertions result in production of fusion transcripts consisting of the reporter and endogenous sequences and the mutated genes can be readily identified using PCR-based methods such as RACE. Furthermore the biological consequences of the integration event can be assessed after germ-line transmission. One limitation of conventional gene trapping is that it can only target genes expressed in ES cells since selection of insertional events relies on the endogenous promoter's activity to drive expression of the selectable marker and to circumvent this problem a new class of gene trap vectors called poly(A) trap vectors was developed. These constructs contain a 3′ selectable marker whose expression is driven by a constitutively active internal promoter relaxing the requirement for endogenous gene expression. The selectable marker lacks a polyA signal but incorporates a splice donor (SD) signal so only integrations upstream of an endogenous gene's splice acceptor (SA) and polyA sequences can be selected thus eliminating intergenic background insertions. However, it has been recently demonstrated that poly(A) trap vectors are biased towards integrations into the 3′most-intron of their target genes due to the action of an mRNA-surveillance mechanism called nonsense-mediated mRNA decay (NMD).

The aim of the study presented here was to assess the efficiency of a series of gene trap vectors that incorporate two novel features in their design: (i) the presence of an ATG-less, 5′ triple fusion between egfp, β-galactosidase and neomycin/hygromycin resistance genes to function as a reporter/selector of the trapped gene's expression state and (ii) a 3′ poly(A) trap cassette that contains the previously uncharacterized rabbit β-globin exon 2/intron 2 SD junction and an AU-rich element (ARE) derived from the human GM-CSF gene. Our results provide evidence that the triple fusion functions properly and can be potentially used as a reporter of trapped locus activity. We also show that the presence of the ARE appears to improve the performance of the rabbit β-globin SD sequence in the context of poly(A) trapping. More importantly, preliminary data suggest that our vectors may be resistant to NMD and thus potentially unbiased in their insertional preference.

# CONTENTS

# CHAPTER 1

# INTRODUCTION

## 1.1 The mouse as a model for studying embryonic development

Embryonic development is an intricate, tightly regulated, process. The identification of the genes which are involved in the control of this process is one of the major goals of developmental biology – for this reason, a variety of model organisms have been employed. Vertebrate development is routinely studied through the use of *Xenopus*, chicken, zebrafish and mouse. The principal invertebrate models include the fruit fly *Drosophila* and the nematode worm *Caenorhabditis elegans*. Although each of these species has its advantages and disadvantages as a developmental model, the premier experimental system for dissecting the events behind mammalian embryogenesis as well as modelling human disease is the mouse. This is mainly due to the high degree of similarity shared between mice and humans at a genomic (99% of mouse genes have a direct human counterpart; Waterston et al., 2002), anatomical and physiological level. Additional advantages include its small body size, short generation time, knowledge of its genome sequence and amenability to genetic manipulation. Moreover, the existence of a large number of well-characterised inbred mouse strains allows investigators in different laboratories to design experiments on the same defined genetic background.

### 1.1.1 An overview of mouse embryogenesis

Fertilisation of the egg in the oviduct marks the initiation of early mouse embryogenesis. This leads to the formation of the zygote which then undergoes consecutive rounds of mitotic divisions to form a cluster of cells called the morula. At the late morula stage, after establishment of cell polarity and morula "compaction" (during which the cells of morula are flattened and cell outlines are not clearly distinguishable), the conceptus enters the uterine lumen and gives rise to the blastocyst. The latter consists of a cavity (blastocoel) and two distinct cell populations, the inner cells mass (ICM) and the trophectoderm (TE). Blastocyst maturation (4.5 days post coitum or 4.5 dpc) then takes place through shedding of the zona pellucida (the outer glycoprotein-based shell that surrounds oocytes/preimplantation embryos) and further differentiation of the ICM into a pluripotent epithelial layer (the epiblast or primitive ectoderm) and the primitive endoderm (Figure 1.1). Each of the components of the mature blastocyst possesses a predefined developmental destiny: the trophectoderm gives rise to extra-embryonic structures such as the placenta; the primitive endoderm contributes to the formation of the visceral and parietal endoderm that lines the yolk sac and the epiblast generates the embryo proper as well as some extra-embryonic membranes (Figure 1.1a). Blastocyst maturation is then followed by implantation of the embryo into the uterine wall which is facilitated by the attachment of the trophectoderm into the uterine lining.

Shortly after implantation, a cylindrical structure known as the egg cylinder appears (6 dpc) containing both the now elongated epiblast as well as the trophectoderm-derived extra-embryonic tissue. The anterior visceral endoderm (AVE) then initiates induction of the anterior regions of the embryo (Beddington and Robertson, 1998) and gastrulation begins at 6.5 dpc with the formation of the primitive streak. During this stage pluripotent

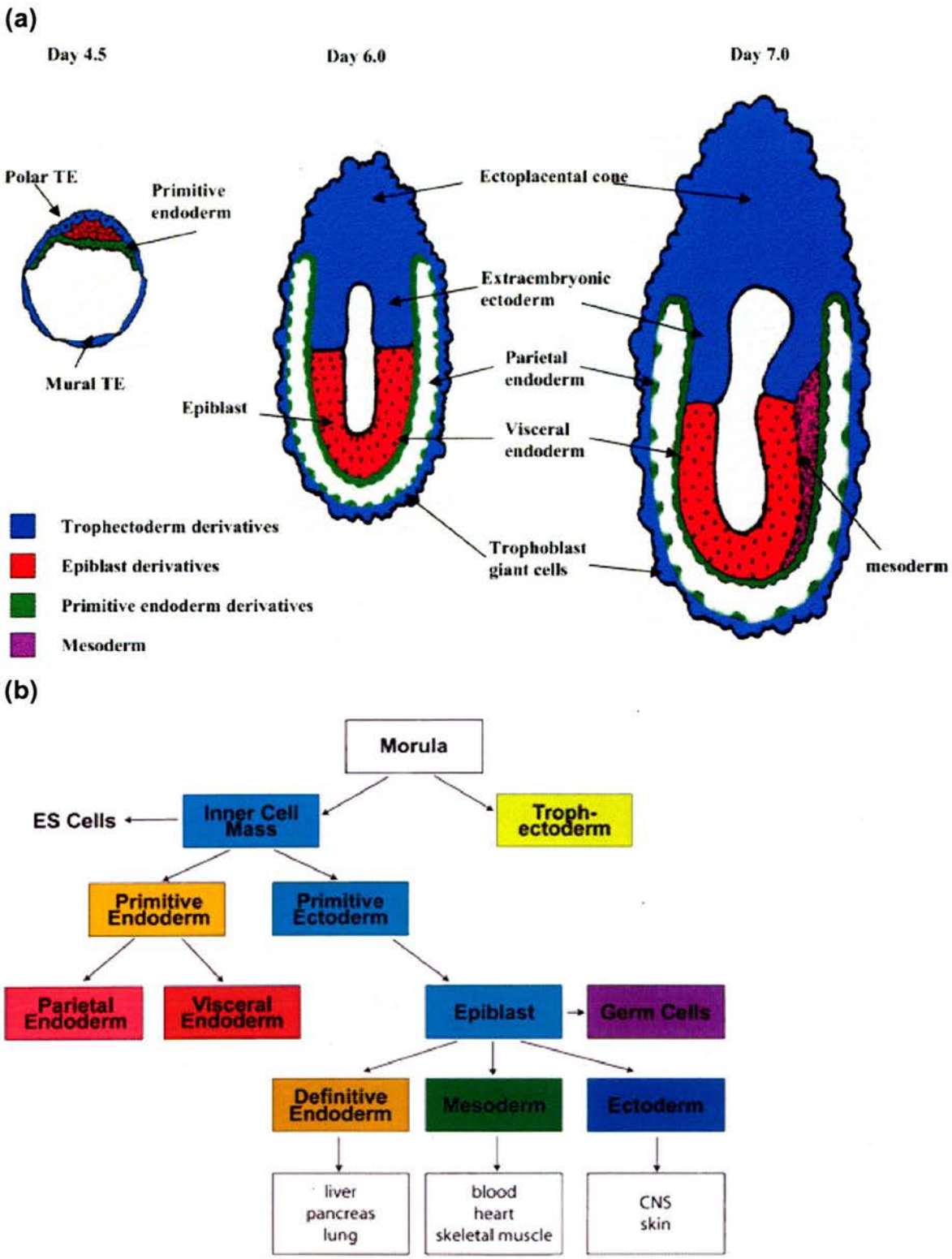



**Figure 1.1** Overview of early mouse development **(a)** Lineage descendants of the three lineages of the blastocyst up to the time of gastrulation (from Rossant, 2004). **(b)** Scheme of early mouse development showing the link between early cell populations and primary germ layers (from Keller, 2005).

epiblast cells are allocated to the three primary germ layers of the embryo (ectoderm, mesoderm and definitive endoderm) and the extraembryonic mesoderm of the yolk sac and amnion. The primary germ layers will eventually produce all foetal tissue lineages (Figure 1.1b). This temporally and spatially regulated process involves the migration of epiblast cells to the primitive streak followed by an epithelial to mesenchymal transition that results in the formation of mesoderm and definitive endoderm. The most posterior allocated epiblast cells of the primitive streak constitute the basis of mesoderm which, in turn, gives rise to haematopoietic and endothelial lineages. The epiblast cells which colonise the primitive streak later and are located in a more anterior position form cardiac mesoderm, head mesenchyme, and paraxial mesoderm. The most anterior-positioned within the primitive streak epiblast cells produce endoderm and axial mesoderm, whereas anterior, non-primitive streak epiblast cells generate the ectoderm. By the end of gastrulation the embryo has a distinct head and developing forelimb buds and organogenesis commences with the formation of heart, the cranial neural folds, and the appearance of the somites.

### 1.1.2 Molecular control of mouse embryogenesis

An important aspect of mouse embryonic development is that it is orchestrated by unique gene expression programs. Egg fertilisation triggers the degradation of maternal, oocyte-derived transcripts and global zygotic genome activation (ZGA) occurs, as shown by microarray studies, between the 2- and 4-cell stages leading to genetic reprogramming (Hamatani et al., 2004) (Figure 1.2a) This is followed by a second transcriptional wave (mid-preimplantation gene activation; MGA) which peaks at the 8-cell stage prior to the morula-to-blastocyst transition (Hamatani et al., 2004) (Figure 1.2a). Subsequent differentiation into the ICM and trophectoderm lineages is

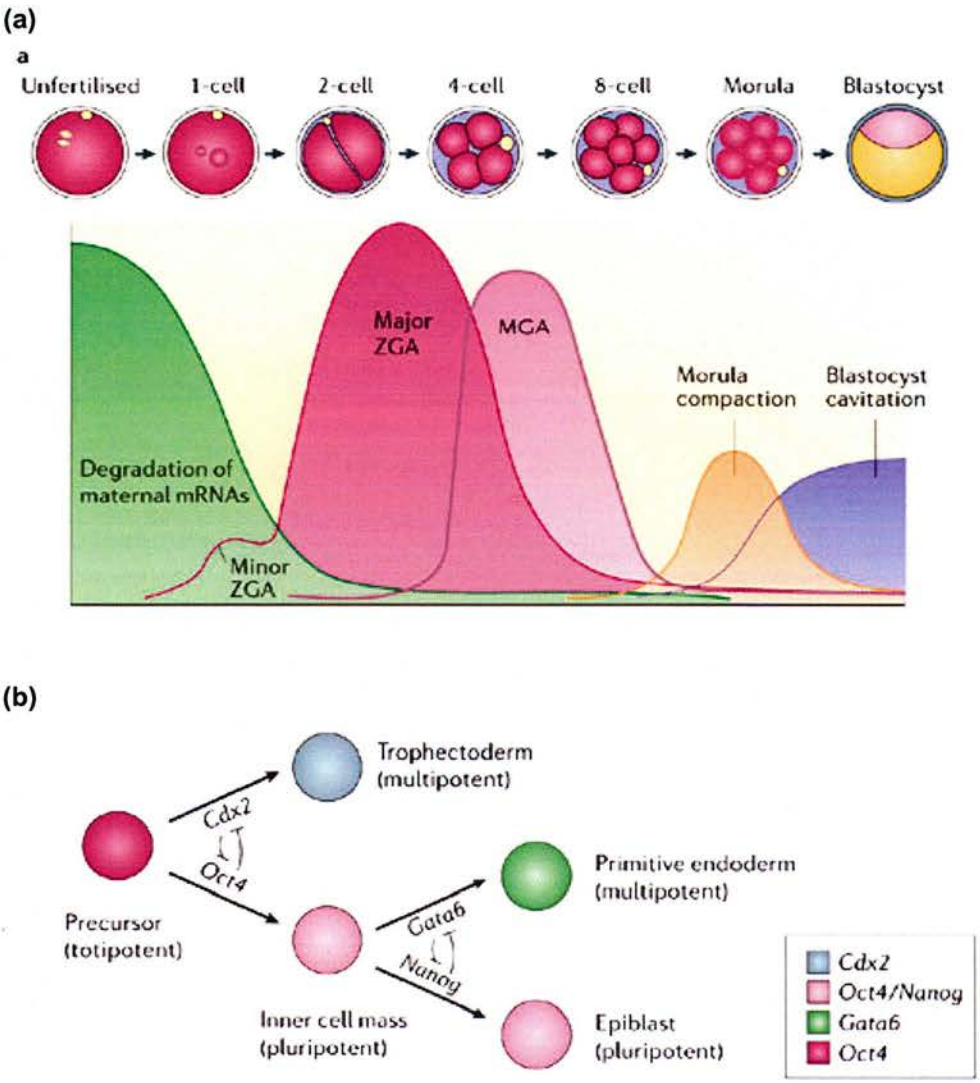



**Figure 1.2** Genes governing the development of the preimplantation mouse embryo. **(a)** Gene expression during preimplantation embryo development. The diagram depicts the waves of gene expression that occur in preimplantation embryos, based on microarray studies. ZGA, zygotic genome activation; MGA, mid-preimplantation gene activation. **(b)** Genetic model of lineage decision. *Oct4* (dark pink) is expressed throughout the embryo before the late morula stage. The expression of *Nanog* (light pink) is specifically induced in the inside cells of late morulae. *Cdx2* (blue) is expressed in the outer layer of cells in late morulae and is required for the repression of *Oct4* and *Nanog* in the trophectoderm (Tr) of the blastocyst. *Oct4* is crucial for inner cell mass (ICM) formation. *Gata6* (green) is expressed in the primitive endoderm of the late blastocyst, where *Oct4* and *Nanog* are repressed. *Oct4* represses *Cdx2* expression, which in turn represses *Oct4* expression to allow segregation of the ICM and Tr lineages of the blastocyst. An antagonism between *Nanog* and *Gata6* segregates epiblast and primitive endoderm within the ICM. (Figure taken from Wang and Dey, 2006).

further governed by the expression of several genes, mainly transcription factors, which it has been argued may act by "co-occupying a substantial fraction of their target genes and collaborate to form a circuitry of autoregulatory and feed-forward loops" (Wang and Dey, 2006; Boyer et al., 2005). ICM specification, for example, is instructed by OCT4 (encoded by the *Pou5f1* gene), SOX2 and NANOG, which prevent the formation of extraembryonic lineages in a co-operative fashion (Nichols, et al., 1998; Niwa et al., 2000; Avilion et al., 2003; Mitsui et al., 2003; Chambers et al., 2003) (Figure 1.2b). Conversely, trophectoderm development is mainly controlled by the caudal-type homeodomain protein CDX2 (Strumpf et al., 2005) and the T-box transcription factor *Eomes* (Russ et al., 2000) (Figure 1.1b) while the protein products of the *Gata4* and *Gata6* genes appear to be important in defining primitive endoderm fate (Fujikura et al., 2002).

Tissue lineage specification is conducted through the activation of lineage-specific gene expression programs and parallel restriction of antagonistic developmental pathways ("reciprocal repressive circuitry among the different molecules"; Rossant, 2004). It is also strongly influenced by position-specific cell interactions and signalling. It is speculated that lineage specification might be initiated by allocation of the appropriate factors to respective precursor cell populations which in turn triggers the activation of regulatory molecular networks linked to a lineage-specific transcriptional profile and leads to mature lineage development (Rossant, 2004). Gene deletion/overexpression studies and expression profiling have provided some valuable insights into the molecular mechanisms that drive these processes. Although the whole picture is far from complete, these studies have revealed the important role played by members of the TGFβ (e.g. BMP4 and nodal), Wnt and FGF families. Figure 1.3 summarises some of

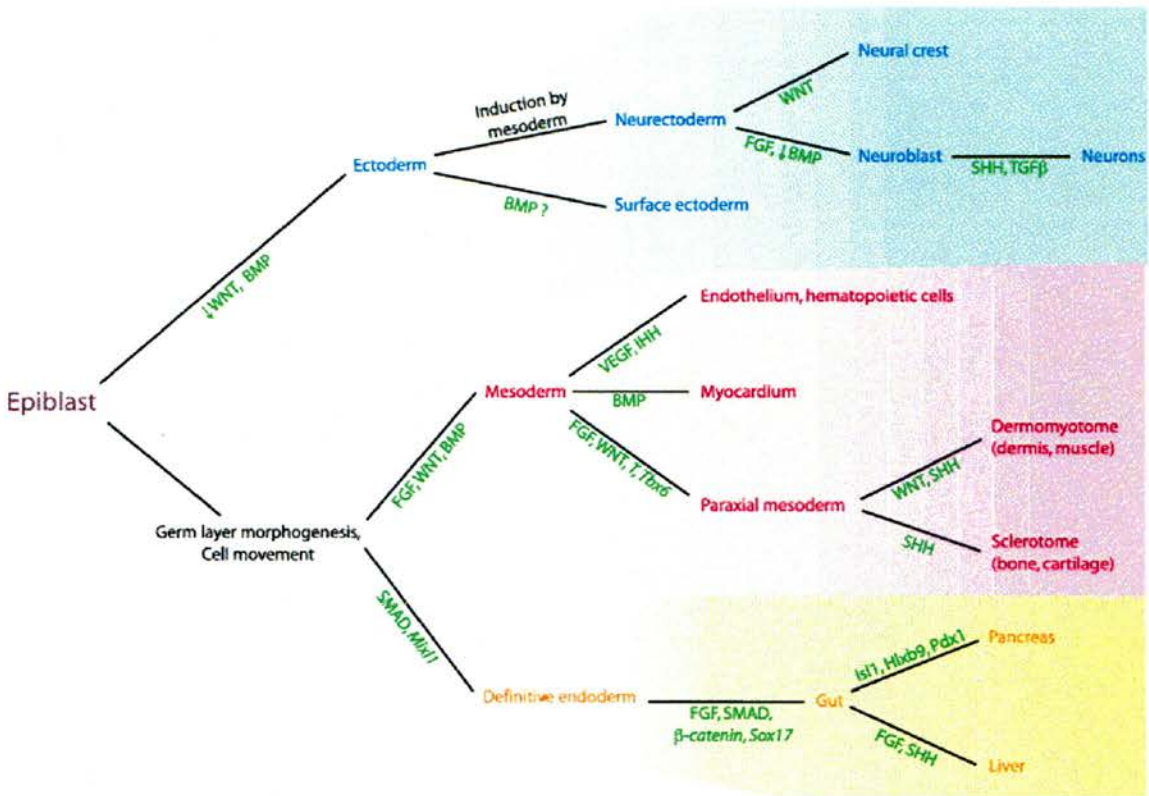



**Figure 1.3** Specification and differentiation of germ layer derivatives in mouse embryos. Blue, ectodermal derivatives; pink, mesodermal derivatives; orange, definitive endoderm derivatives; green, signalling pathways and transcription factors (from Loebel et al., 2003).

the main determinants of germ layer specification/differentiation. Table 1.1 offers some examples of germ layer and lineage-specific genes.

Embryonic differentiation events also appear to be controlled at an epigenetic level especially through alterations in chromatin structure, nuclear dynamics and histone modification status. Chromatin structure is an important regulator of gene function as it influences genome accessibility. Studies on embryonic stem (ES) cells (see next section) demonstrated that an undifferentiated cellular state is characterised by an abundance in euchromatic, gene transcription-permissive regions while the induction of differentiation is accompanied by an increase in the fraction of highly condensed, transcription-restrictive heterochromatic foci, (Francastel et al., 2000; Arney and Fisher, 2004). It is possible that the implementation of gene expression programs associated with specific developmental stages is carried out through the "targeted, physical segregation of genes into active and inactive chromatin domains" (Meshorer and Mistelli, 2006). ATP-dependent chromatin remodelling factors such as BRG1, SNF5 and SSRP1 probably play a critical role in this process and gene targeting experiments have already indicated their importance in embryonic development (Bultman et al., 2000; Klochendler-Yeivin et al., 2000; Cao et al., 2003). The correlation between gene expression and chromatin structure/nuclear dynamics has been recently demonstrated in the mouse embryo for the *Hoxb* gene cluster during gastrulation; *Hoxb1* expression in the posterior primitive streak was found to be accompanied by chromatin decondensation and subsequent "looping out" from its chromosome territory (Chambeyron et al., 2005).

| Gene | Developmental Role | Reference(s) |
|---|---|---|
| *Bmp4* | Mesoderm formation | Winnier et al., 1995 |
| *Brachyury* | Early mesoderm induction | Wilson et al., 1995 |
| *Smad4* | Epiblast proliferation, gastrulation and mesoderm formation | Sirard et al., 1998 |
| *Wnt-3a* | Mesoderm formation | Yoshikawa et al., 1997 |
| *Fgfr1* | Mesoderm differentiation | Ciruna and Rossant, 2001 |
| *Scl* | Haematopoietic/endothelial specific | Elefanty et al., 1997; Endoh et al., 2002 |
| *Vegf* | Haematopoietic/endothelial specific | Carmeliet et al., 1996 |
| *Flk1* | Haematopoietic/endothelial specific | Shalaby et al., 1995; Shalaby et al., 1997 |
| *Nkx2-5* | Cardiac differentiation | Lien et al., 2002 |
| *MyoD* | Myogenesis | Ridgeway et al., 2000 |
| *Sox1* | Neuroectoderm formation | Pevny et al., 1998 |
| Chordin | BMP antagonist, anterior neural differentiation | Bachiller et al., 2000 |
| Noggin | BMP antagonist, anterior neural differentiation | Bachiller et al., 2000 |
| *Nurr1* | Neuronal differentiation | Kim et al., 2002 |
| *Shh* | Neuronal differentiation | Ye et al., 1998 |
| *Sox17* | Definitive endoderm differentiation | Kanai-Azuma et al., 2002 |
| *Mixl1* | Definitive endoderm differentiation | Hart et al., 2002 |
| *β-catenin* | Definitive endoderm specification | Lickert et al., 2002 |
| *Gtar* | Hepatocyte lineage specification | Watt et al., 2001 |

**Table 1.1** Examples of genes that are involved in the specification of primary germ layers and/or resulting lineages. Blue, ectoderm-specific; pink, mesoderm-specific; orange, definitive endoderm-specific;

### 1.1.3 The use of ES cells as a model for embryonic development

ES cells are pluripotent cells derived from the inner cell mass of blastocyst-stage embryos (Figure 1.1b) (Evans and Kaufman, 1981; Martin, 1981). They possess two key properties that distinguish them from all other organ-specific stem cells identified to date. First, they can be maintained and expanded as pure populations of karyotypically normal, undifferentiated cells for extended periods of time. Second, they are pluripotent and hence have the capacity to generate every cell type in the body, both *in vivo* and *in vitro*.

The ability of ES cells to differentiate into derivatives equivalent to all three embryonic germ layers makes them an attractive developmental model that is easier to manipulate compared to the mouse embryo. ES cell differentiation *in vitro* is usually achieved, in a directed manner, through three alternative approaches: embryoid body (EB) formation in which ES cells are allowed to aggregate and form three dimensional structures (Doetschman et al., 1985; Keller, 1995); co-culture with differentiation-promoting stromal cell lines (e.g. OP9; Nakano et al. 1994); culture on extracellular matrix proteins (Nishikawa et al., 1998). Figure 1.4 shows a model comparing, based on our existing knowledge, the early stages of embryonic and ES/B development.

Another attractive feature of ES cells is their amenability to genetic manipulation. This characteristic has been greatly exploited for dissecting gene function, especially in a developmental context. Genetic changes can be introduced in culture in a directed or random manner and the resulting phenotype can be assessed both *in vitro* and, more importantly, *in vivo*. The developmental "pluripotentiality" of ES cells can be employed for generating mutant mice since they contribute to the formation of somatic tissues of a

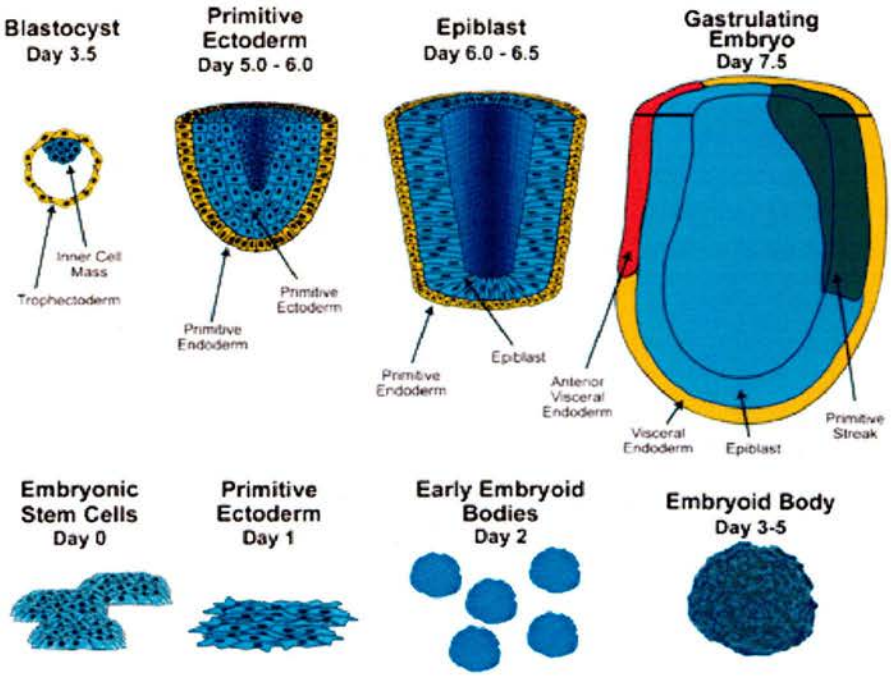



**Figure 1.4** ES/EB development as a tool for modelling early mouse embryonic development (figure obtained from Keller, 2005).

resulting chimaeric animal, after placement back into an embryonic environment through host blastocyst injection. Furthermore, they contribute to functional germ cells enabling the transmission of genetic alterations through the germline (Bradley et al., 1984).

## 1.2 Mutagenesis strategies in the mouse

The classic route for characterising the function of genes, including those which are critical in various aspects of mouse development, involves the use of mutagenesis experiments. One of the main post-genomic goals of the mouse genetics community is the production of at least one heritable mutation, in either ES cells or mice, within every gene of the mouse genome ("genome saturation") and the identification of each consequent phenotype. For this reason, an International Mouse Mutagenesis Consortium (IMMC) has been established combining the efforts of research groups from all around the world both from the public and private domains (Nadeau et al., 2001). Mutagenesis strategies are generally divided into two categories: forward genetic approaches in which the phenotype is the starting point towards identifying the affected gene sequence and reverse genetic approaches which are characterised by the introduction of a change into a predefined gene sequence followed by assessment of the resulting mutant phenotype.

### 1.2.1 Forward genetic strategies

#### 1.2.1.1 ENU mutagenesis

The majority of phenotype-driven mutagenesis regimes utilise the alkylating agent N-ethyl-N-nitrosourea (ENU) to introduce point mutations in spermatogonial stem cells (Russell et al., 1979). The resulting animals are then screened through the use of appropriately designed mating schemes

and/or under a specific set of conditions according to the biological questions of the study. Commonly employed screens can often be region-specific, focusing on mutations that lie on a particular genomic interval or genome-wide which are considered more useful for dissecting the genetic basis of a specific biological process (Kile and Hilton, 2005). ENU is a very efficient mutagen: it induces mutations at single loci in approximately one sperm in 1,000 (Hitotsumachi et al., 1985; Lyon and Morris, 1966), a rate which is approximately 100-fold higher than the spontaneous mutation rate (Kile and Hilton, 2005).

Large-scale ENU-based mutagenesis screens are being performed in several centres world-wide. For example, the German ENU mutagenesis project involves mating of ENU-treated male mice with wild-type females to generate F1 founders which are then screened for dominant/semi-dominant mutant phenotypes or bred further to subsequently study recessive phenotypes (Hrabe de Angelis et al., 2000; http://www.gsf.de/ieg/groups/enu-mouse.html). Similar phenotype-driven projects are being carried out by the Medical Research Centre at Harwell in UK (http://www.mgu.har.mrc.ac.uk/research), the RIKEN Institute in Japan (http://www.gsc.riken.go.jp/Mouse/), and the Centre for Modelling Human Disease in Canada (http://cmhd.mshri.on.ca/enu_mutagenesis/index.html). However, despite the fact that mutant generation through ENU is almost effortless, the identification of the genes affected through this approach by positional cloning or by a candidate-gene approach remains cumbersome. Moreover, some loci have been shown to be "immutable" by ENU (Rinchik and Carpenter, 1999) while the breeding schemes employed for the phenotypic analysis of the resulting mutants are often costly and time-consuming.

### 1.2.2 Reverse genetic strategies

#### 1.2.2.1 Gene targeting

Gene targeting is the most widely-used genotype-based mutagenesis approach in mice. It involves the genetic modification of ES cells through the introduction of exogenous DNA (targeting vector) that contains two homology arms, specific to the locus of interest and consequent homologous recombination which facilitates the replacement of the targeted allele with the exogenous sequence. The modified ES cells can then be used to generate, after blastocyst injection, chimaeric mice so that mutant phenotypes can be studied *in vivo* on a defined genetic background. Gene targeting proved an extremely useful and popular mutagenesis tool for elucidating gene function *in vivo* and providing valuable insights into key biological processes. To date, this method has facilitated the production of approximately 3,600 mutants corresponding to a 15% coverage of the entire mouse genome (Skarnes, 2005). However, the need for screening a large number of ES clones in order to identify a small number of correctly targeted events renders gene targeting a labour-intensive and time-consuming approach (Skarnes, 2005) and thus unsuitable for high-throughput experiments. Furthermore, despite recent technical advances (e.g. with the development of recombineering), the engineering of targeting vectors is technically challenging, targeting efficiencies are low, and vectors often integrate at random (Carlson and Largaespada, 2005).

#### 1.2.2.2 RNAi

RNA interference (RNAi) is used as a means to induce the sequence-specific down-regulation of target mRNA transcripts through the employment of a conserved post-transcriptional gene regulatory mechanism. It can be triggered either by transiently transfected, synthetic 21-23 nt RNAs

with 2-nt 3'overhangs (small interfering RNAs or siRNAs) or by longer RNA transcripts (shRNAs or pri-miRNAs) stably expressed from PolII or PolIII-driven vectors. RNAi is a convenient tool for fast, high-throughput mutagenesis studies and has been successfully utilised for the construction of large-scale-arrayed, sequence-verified libraries targeting most known and predicted genes in the human and mouse genomes (Silva et al., 2005). However, unlike traditional gene targeting systems it mediates a knockdown rather than a knockout of target gene expression and hence it might not be the most appropriate means of generating loss-of-function phenotypes. Furthermore, there are still some caveats and unresolved issues concerning RNAi technology itself. The most important issue to be addressed is non-specific suppressive activity including the induction of off-target knockdown silencing and the triggering of the interferon response system. The off-target effect involves the sequence-specific silencing of non-targeted genes: it has been shown that expression of a non-targeted transcript with as few as 11 consecutive nucleotide matches with a siRNA sequence can be downregulated (Jackson et al., 2003). The interferon response is linked to the non-specific, sequence-independent effect mediated by an innate, immune response against any exogenous dsRNA (Stark et al., 1998; Grandvaux et al., 2002). Although siRNAs were generally assumed to be too short for triggering an interferon response in mammalian cells (Caplen et al., 2001; Elbashir et al., 2001), two recent reports demonstrated that expression of shRNAs in mammalian cells can activate the interferon pathway (Bridge et al., 2003; Sledz et al., 2003).

### 1.2.3 Random insertional mutagenesis

#### 1.2.3.1 Insertional mutagenesis using DNA microinjection/retroviruses

Insertional mutagenesis, conceptually, combines the advantages of both forward and reverse genetic approaches. It involves the random insertion of mutagenic exogenous DNA into a target genome simultaneously providing a molecular tag for cloning of the mutated gene. It was first introduced in 1976 with the report of retroviral DNA insertion into the mouse germline (Jaenisch, 1976). In 1981 microinjecting exogenous DNA into fertilized oocytes yielded the first transgenic mice (Constantini and Lacy, 1981; Gordon and Ruddle, 1981; Harbers et al., 1981; Wagner et al., 1981). However, these initial attempts were characterised by a low probability (5%) of producing a phenotype (Jaenisch, 1988; Spence et al., 1989; Weiher et al., 1990) and DNA rearrangements occurring at the mutated locus (Gridley et al., 1987). For these reasons alternative insertional mutagenesis approaches were pursued.

#### 1.2.3.2 Transposon-based insertional mutagenesis

Transposon-mediated insertional mutagenesis has attracted much attention lately due to the development of elements which can be employed in vertebrate systems. Transposons are naturally occurring mobile genetic elements that move around within the genome of an organism. In the mouse, various transposable elements have been identified, including retrotransposons and DNA transposons. The LINE-1 (L1) family of retrotransposons (mobile elements that transpose via a 'copy-and-paste' mechanism which involves reverse transcription of an RNA intermediate) have been shown to mobilize in cultured mammalian cells (Moran et al., 1996) and in the mouse germline (Ostertag et al., 2002). This finding renders

them a potential candidate for use in *in vivo* insertional mutagenesis. However, the frequency of their germline transposition was found to be low (Farley et al., 2004) while their tendency for 5'truncations (Farley et al., 2004) and promotion of deletions at the site of integration (Gilbert et al., 2002) complicates their use as effective mutagenic agents.

DNA transposons which mobilize by a 'cut-and-paste' mechanism, have, so far, proved more efficient as insertional mutagens. The most successful systems utilize the *piggyBac* transposon derived from the cabbage looper moth (Ding et al., 2005) and *Sleeping Beauty* (SB), a synthetic element made from defective copies of an ancestral Tc1/mariner fish transposon (Ivics et al., 1997). *piggyBack* has been demonstrated to possess the potential to carry as much as 14 kb of exogenous DNA and it can mobilize efficiently within human and mouse somatic cells as well as the mouse germline (Ding et al., 2005). SB has been shown to be active in mouse ES cells (Luo et al., 1998) and, most importantly, SB transposition can be achieved in the germline of mice that are transgenic for transposon and SB transposase (Dupuy et al., 2001; Fischer et al., 2001; Horie et al., 2001). Furthermore, gene-trap transposons have been successfully mobilized to mutate genes *in vivo* (Carlson et al., 2003; Horie et al., 2003). A Mouse Transposon Insertion Database (MTID) has been established offering to the scientific community a significant number of mice carrying germline SB insertions into genes or chromosomal regions of interest (Roberg-Perez et al., 2003; http://mouse.ccgb.umn.edu/transposon/).

Transposon-based strategies are an attractive tool for performing rapid forward genetic screens, especially *in vivo*, since the induction via this approach of gene mutations and the recovery of resulting mutant phenotypes occur almost simultaneously (Carlson and Largaespada, 2005).

However, genome-wide mutagenesis using any transposon system is not currently feasible due to the relatively low germline transposition frequency (Carlson and Largaespada, 2005). Furthermore, SB-related transposons are subject to 'local hopping', a phenomenon in which new integration sites are linked to the donor locus in 50-80% of the cases (Carlson et al., 2003; Horie et al., 2003) while they also exhibit a small but significant integration bias towards genes and their upstream regulatory sequences, although the bias is much less than that observed with retroviruses (Yant et al., 2005).

## 1.3 Gene trap mutagenesis

Gene entrapment is one of the most powerful and popular insertional mutagenesis approaches. It was first employed in Drosophila (O'Kane and Gehring, 1987) and then adopted as a tool for identifying and mutating developmentally regulated genes in the mouse (Gossler et al., 1989). It is based on the random introduction into ES cells (via electroporation or retroviral infection) of a DNA vector that is designed to signal its presence via the activation of a reporter gene. The latter mimics the expression of the endogenous gene (or the activity of a disrupted enhancer/promoter) and potentially mutates the locus. The "trapped" ES cells can then be selected *in vitro* and subsequent germline transmission enables the analysis of the insertion's *in vivo* phenotypic consequences. Huge libraries of ES cell clones bearing random integrations can therefore be rapidly generated and stored indefinitely enabling the implementation of high-throughput approaches. Additionally, the sequence of the "trapped" gene/site of integration can be easily identified using PCR-based techniques such as RACE (Rapid Amplification of cDNA Ends, Frohman et al. 1988) and inverse PCR (von Melchner et al., 1990) or plasmid rescue (Hicks et al., 1997).

### 1.3.1 Basic entrapment approaches

#### 1.3.1.1 Enhancer trapping

An enhancer-trap vector (Figure 1.5) usually contains a minimal promoter that requires the vector to insert near a *cis*-acting enhancer element in order to activate a reporter fused to the minimal promoter. Enhancer trapping has been successfully used in Drosophila as a means of detecting transcriptional regulatory elements (O'Kane and Gehring, 1987) and its principles have also been applied in the mouse (Gossler et al. 1989; Korn et al. 1992; Marikawa et al., 2004); data based on reporter expression patterns both in ES cells and chimaeric embryos (Gossler et al., 1989; Korn et al., 1992; Marikawa et al., 2004) as well as molecular cloning of integration sites (Korn et al., 1992; Marikawa et al., 2004) have provided evidence that enhancer trap vectors possess the potential to tag cellular enhancers active during embryonic development. However, this type of entrapment has not been widely exploited mainly due to its low mutagenicity (Korn et al., 1992; Marikawa et al., 2004). Furthermore, reporter expression in the case of enhancer trap vectors is likely to be affected by the specific promoter employed, without necessarily reflecting the activity of the trapped, endogenous enhancer elements (Marikawa et al., 2004).

#### 1.3.1.2 Promoter trapping

Promoter-trap vectors (von Melchner and Ruley, 1989; Friedrich and Soriano, 1991; Reddy et al. 1991; von Melchner et al., 1992) consist of a promoterless reporter gene/selectable marker whose activity is dependent on vector insertion in the correct orientation and translational frame into an exon (those vectors can also be classified as exon traps) of an active gene. This results in the generation of a fusion transcript that comprises upstream endogenous exonic sequence and the reporter gene (Figure 1.6). The first

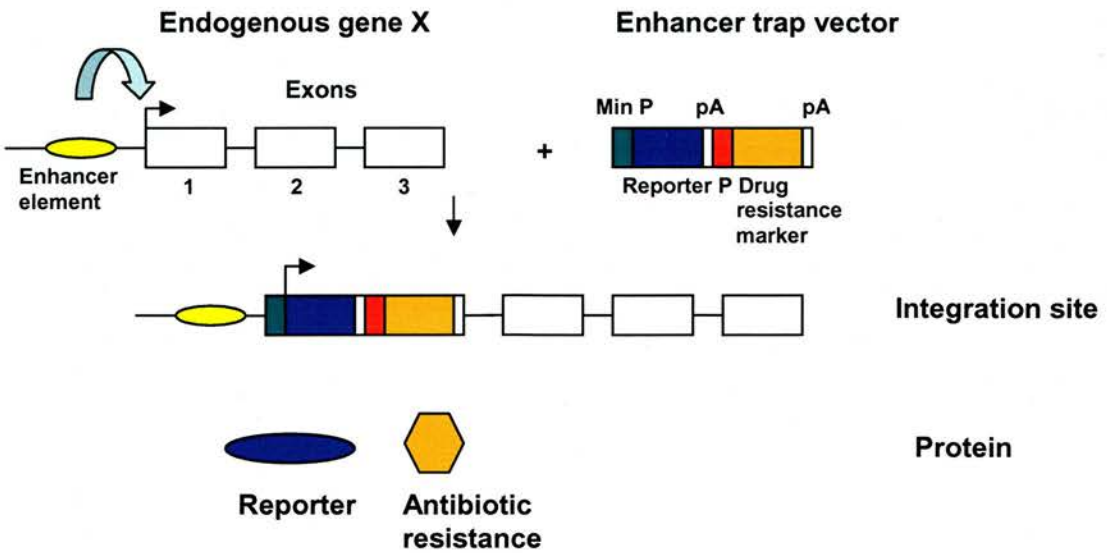



**Figure 1.5** Enhancer trapping. The enhancer-trap vector contains a minimal promoter upstream of a reporter gene. Insertion of the vector close to the enhancer of gene X leads to the transcription and translation of the reporter when gene X is active. It should be noted that in the example given here the vector also contains a constitutive promoter that drives the expression of an antibiotic resistance gene for selection of integration events. P, promoter; Min P, minimal promoter; pA, polyadenylation signal site.

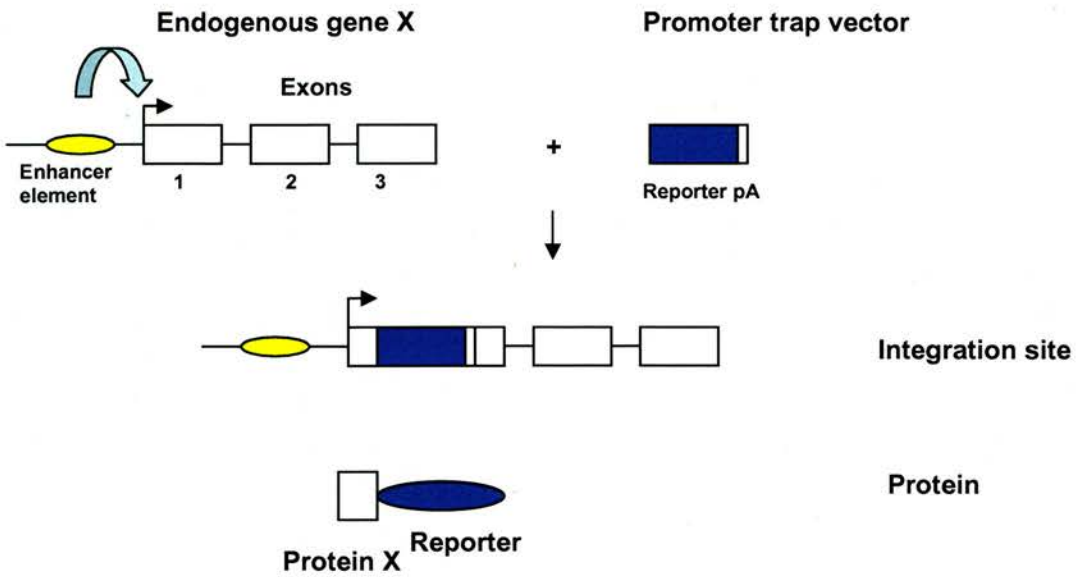



**Figure 1.6** Promoter trapping. The promoter-trap vector consists only of a reporter/selectable marker. Insertion of the vector into the coding sequence of a transcriptionally active gene X will activate the reporter/selector resulting in the generation of a fusion transcript and protein between trapped gene X and the vector's reporter/selector gene. pA, polyadenylation signal site.

promoter trap vectors (U3Neo and U3His) to be employed successfully in ES cells contained a selectable marker (*neo* or *his* conferring resistance to neomycin and L-histidinol respectively) inserted into the U3 region of the 3'long terminal repeat (3'LTR) of a replication-defective Moloney murine leukaemia virus (von Melchner et al., 1992). It was shown that drug selection of infected ES cells leads to the isolation of insertional events that represent as predicted unique fusions between *neo/his* and upstream sequences of trapped genes (von Melchner et al., 1992). Locus disruption in two cases was associated with a homozygous embryonic lethal phenotype after germ-line transmission. Both affected genes (*hnRNP C* and *fug1*) were later found to play a critical role in postimplantation mouse development demonstrating the ability of this type of vectors to capture developmentally important genes (DeGregori et al., 1994; Williamson et al., 2000). The use in a subsequent study of a similar vector (U3βgeo) consisting of the fusion between the β-galactosidase and *neo* genes (*βgeo*; see next section) resulted in the generation of mutant mice which carried an insertion within the *Prmt1* gene (Scherer et al., 1996). In this case homozygous embryos failed to develop beyond E6.5 indicating that the mutated locus is essential for embryonic development (Pawlak et al., 2000). In general, promoter trap vectors seem to be highly mutagenic presumably due to the fact that they integrate directly into exons; for example 78% of mutant mice carrying U3βgeo insertions were shown to exhibit an obvious phenotype (Hansen et al., 2003). They can also be used for trapping single exon genes, a type of genes that cannot be mutated through the use of SA (splice acceptor)-containing gene trap vectors (see next section) (Hansen et al., 2003).

The above studies suggest that promoter trapping is an efficient mutagenesis strategy in the mouse. For this reason U3 promoter trap vectors

have been employed as tools for the large-scale construction of trapped ES cell libraries (U3Neo, Hicks et al., 1997; U3βgeo, Hansen et al., 2003). However, this type of vectors appears to preferentially integrate towards the 5'UTR of their target genes (Hansen et al., 2003) and therefore trapping in this case is not entirely unbiased. Moreover, it has been recently demonstrated that the majority (95%) of U3Neo vector insertions occur within introns and disrupted endogenous transcripts tend to splice in-frame to a cryptic 3' SA site within the vector's *neo* gene (Osipovich et al., 2004). It therefore appears that the mechanism of entrapment by this specific vector is more similar to the one employed by gene trap constructs (see next section).

#### 1.3.1.3 Gene trapping

Gene trap vectors (Gossler et al., 1989) (Figure 1.7) contain a splice acceptor site (SA) immediately upstream of a promoterless reporter gene which is followed by a poly(A) signal. The integration of the vector into the intron of an expressed gene in the correct reading frame results in splicing between the reporter gene and the endogenous upstream exon leading to the generation of a fusion transcript and an active protein that incorporates the N-terminal portion of the mutated native protein fused to the reporter protein (Skarnes et al., 1992). Hence reporter expression is controlled by the regulatory elements of the disrupted locus and should theoretically reflect the trapped gene's expression status (Skarnes et al., 1992; Friendrich and Soriano, 1991). This implies that the tagged integration can serve as a tissue-specific or cell lineage marker. More importantly, the insertional event generates a truncated version of the disrupted gene's protein product and is therefore likely to lead to a loss-of-function phenotype. The rest of this chapter focuses on specific aspects of gene trapping and discusses theoretical and practical issues associated with this type of mutagenesis.

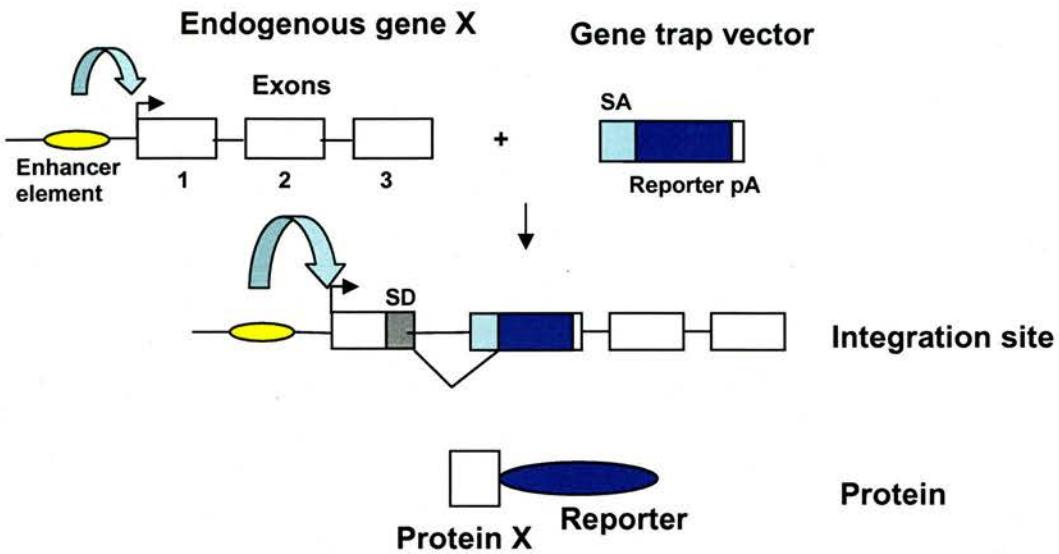



**Figure 1.7** Gene trapping. The gene-trap vector contains a splice acceptor upstream of the reporter/selector gene. Integration in an intron leads after splicing between an upstream endogenous SD and the vector's SA to the generation of a fusion transcript and protein between the upstream exon of gene X and the reporter if gene X is active. pA, polyadenylation signal site; SD, splice donor; SA, splice acceptor.

### 1.3.2 Vector Designs

#### 1.3.2.1 Basic vector designs

The prototype plasmid gene trap vector (termed pGT4.5) contained a SA (from the mouse *engrailed-2* gene) followed by the ATG-less *lacZ* gene which served as a tag of the trapped gene's activity and a poly(A) signal (Gossler et al., 1989). The vector also included an internal promoter-driven *neo* gene for selection of integration events (Gossler et al., 1989). The *lacZ* gene was chosen as a reporter because the activity of its protein product, β-galactosidase, can be easily assayed in ES cells and embryos using the chromogenic substrate X-gal which yields a blue colour while it can also tolerate large N-terminal fusions (Friedrich and Soriano, 1991). The use of this vector provided the first insights into the principles underlying gene trapping and demonstrated, through the production of abnormal *in vivo* phenotypes, the value of this approach as a mutagenesis tool (Skarnes et al., 1992). A similar design was subsequently adopted in the construction of the PT1-ATG vector (Hill and Wurst, 1993). In this case, the inclusion of an ATG upstream of the *lacZ* gene resulted in a three-fold increase in the number of β-gal-expressing clones, probably reflecting the detection of out-of-frame splicing events (Hill and Wurst, 1993). The inclusion in this class of vectors of an autonomous promoter driving selection, theoretically offers the potential to trap genes that are not expressed in undifferentiated ES cells. However this feature also results in an increased background due to selection of intergenic integrations and is associated with low trapping efficiency (0.2-5%) (Niwa et al., 1993; Wurst et al., 1995; Forrester et al., 1996).

The need for a reporter that allows the direct selection of insertion events combined with the parallel monitoring of trapped locus activity led to the development of the *βgeo* gene (Friedrich and Soriano, 1992). Its

construction was based on the in-frame insertion of *neo* into the 3'end of *lacZ* (Friedrich and Soriano, 1992). The retroviral vector ROSA*βgeo* and its plasmid equivalent pSA*βgeo* were the first gene trap constructs to incorporate *βgeo* as a reporter (Friedrich and Soriano, 1992). A wide variety of *βgeo*-containing vectors, both retroviral and plasmid, were subsequently designed and employed in a large number of gene trapping experiments mainly due to their high trapping and mutagenic efficiency. Their use revealed that resistance to G418 is a more sensitive means of isolating gene trap clones compared to X-gal staining since a fraction of $neo^R$ clones are always negative for β-galactosidase activity (Friedrich and Soriano, 1992; Skarnes et al., 1995). The correction of a point mutation that causes reduced enzyme activity and was found present within the *neo* component of the original *βgeo* reporter opened up the possibility to access genes expressed at low levels (Skarnes et al., 1995).

More recent versions of *βgeo*-based vectors often include an internal ribosome entry site (IRES) from the encephalomyocarditis virus between the SA site and the *βgeo* sequence (termed IRES*βgeo*; Chowdhury *et al.* 1997) to facilitate bi-cistronic translation (i.e. the translation of two proteins from one mRNA). Inclusion of this element enables the reporter gene to be translated even when it is not fused in-frame to the endogenous, trapped gene. Another strategy aiming to address the issue of out-of-frame translational fusions involves the combined, parallel use of pGT1, 2 and 3 gene trap vectors (SA*βgeo*-pA). In this case the splice acceptor's exon sequence was modified to produce three vector derivatives (pGT1, 2 and 3) which can generate fusion proteins in each of the three reading frames (Wilson et al., 1995; Tate et al., 1998; Sutherland et al., 2001; Tzouanacou et al., 2005).

It has been shown that when a conventional *βgeo* gene trap vector integrates within genes that contain an upstream secretory signal sequence, the resulting *βgeo* fusion proteins are translocated (providing the insertion takes place upstream of any endogenous transmembrane domain-encoding sequence) in the lumen of the endoplasmic reticulum and β-galactosidase is abolished (Skarnes et al., 1995). Tagging of this type of genes cannot therefore be achieved using conventional gene trap vectors and this limitation was overcome through the design and employment of a vector (pGT1.8TM) that includes the transmembrane domain of the rat CD4 gene upstream of the *βgeo* reporter ("secretory trap"; Skarnes et al., 1995) (Figure 1.8). This novel approach proved to be effective in recovering insertional mutations within genes that encode secreted and type I membrane proteins (Skarnes et al., 1995). Vector pGT1.8TM and its modified versions were subsequently used in a large-scale screen aiming to identify developmental important genes that belong to this specific class (Mitchell et al., 2001). A similar strategy employing retroviral gene trap vectors that contain a reporter fusion between the human CD2 receptor gene which provides a type II transmembrane domain and the neomycin resistance gene (*Ceo*) has been recently reported (De-Zolt et al., 2006).

A substantial number of mouse mutants generated through the use secretory and classic *βgeo*-type gene trap constructs have been described and some examples are given in Table 1.2. These studies clearly demonstrate the value of gene trapping in aiding the identification and functional annotation of genes that are critical in early mouse development.

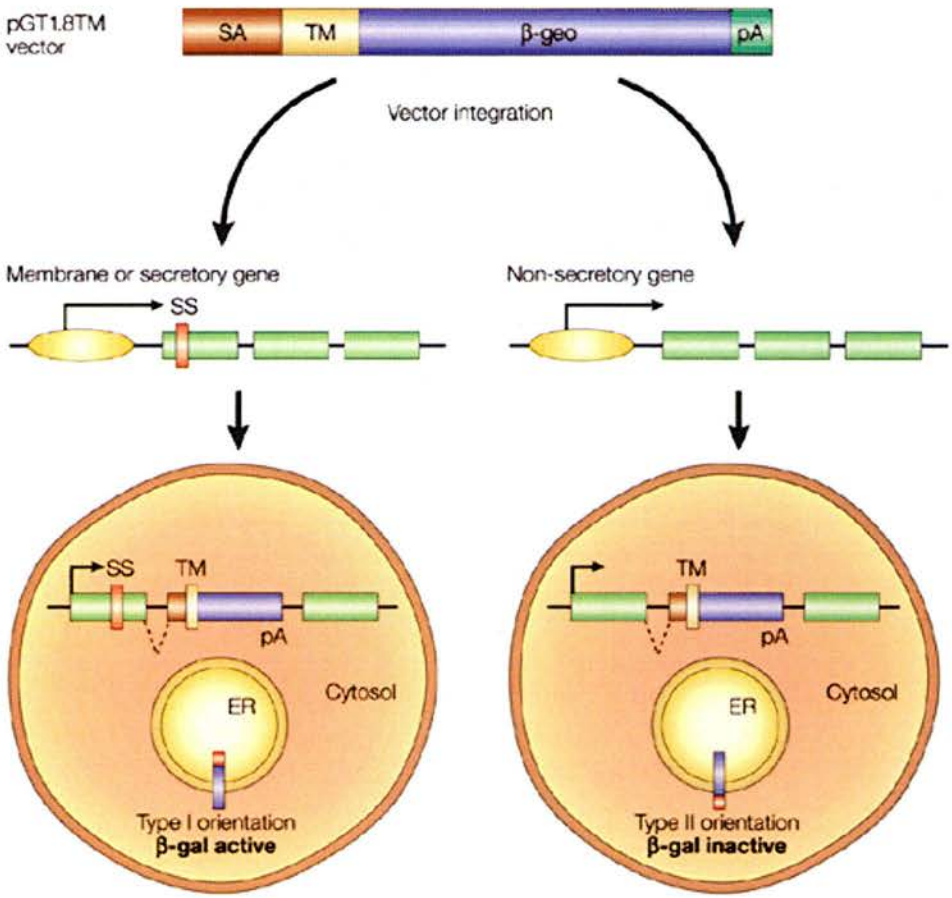



**Figure 1.8** The secretory-trap vector. The secretory-trap vector uses protein sorting and the fact that β-galactosidase (β-gal) activity is abolished in the endoplasmic reticulum (ER) specifically to trap genes that encode secreted and transmembrane proteins that are expressed in embryonic stem (ES) cells. The pGT1.8TM secretory-trap vector contains a transmembrane (TM) domain immediately downstream of a splice acceptor (SA) site, followed by the β-*geo* reporter with its own polyA site. The TM domain of the secretory-trap vector sequesters gene-trap fusion proteins that do not have a signal sequence into the ER lumen, thereby extinguishing β-gal activity. When a fusion protein contains a secretory signal (SS) sequence, it is translocated into the cytosol where β-gal activity can be assayed. So, the secretory-gene-trap vector enriches for insertions into genes that encode secreted or transmembrane proteins by using a modification of blue–white selection (from Stanford et al., 2001).

| Vector | Disrupted locus | Phenotype | Reference |
|---|---|---|---|
| ROSAβgeo | *Arkadia* | Recessive lethal, failure to maintain anterior embryonic structures | Episkopou et al., 2001 |
| ROSAβgeo | *BTF3* | Early postimplantation lethality | Deng and Behringer, 1995 |
| ROSAβgeo | *Ctbp2* | Defects in axial patterning, embryonic lethality 10.5 dpc | Hildebrand and Soriano, 2002 |
| pGT1.8geo | *β-E- catenin* | Embryonic lethality 6.5 dpc, defective implantation | Torres et al., 1997 |
| pGT1.8geo | *Neuropilin-2* | Axon guidance and lymphatic vessel defects | Chen et al., 2000; Yuan et al., 2002 |
| pGT1.8geo | *Taube Nuss* | Preimplantation lethality | Voss et al., 2000 |
| pGT1.8TM | *Netrin-1* | Defects in commissural axon guidance, neonatal lehtality | Serafini et al., 1996 |
| PLAP | *ADAM19* | Neonatal lethality, cardiac defects | Zhou et al., 2004 |
| IRESβgeo | *Apaf1* | Embryonic lethality 16.5 dpc | Cecconi et al., 1998 |
| IRESβgeo | *Map1β* | Neonatal lethality, neuronal developmental defects | Gonzalez-Billault et al. 2000 |
| pT1βgeo | *Birc6* | Embryonic lethality 11.5-16.5 dpc | Ren et al., 2005 |

**Table 1.2** Examples of characterised mouse mutants generated through the use of different, representative gene trap vectors and carry insertions within developmentally critical and tissue-specific genes.

#### 1.3.2.2 Alternative gene trap vector components

Gene trap constructs that utilize reporter/selection systems other than the traditional *lacZ/neo* combinations have also been developed. Hygromycin and phleomycin are two examples of alternative to *neo* drug resistance markers that have been used by gene trappers in their effort to construct novel vectors (Natarajan and Boulter, 1995; Camus et al. 1996). Alkaline phosphatase (AP) has been employed as an alternative to β-galactosidase to monitor expression (Xiong et al. 1998; Leighton et al., 2001). Like β-galactosidase, AP activity can also be monitored during *in vitro* differentiation protocols as well as in *vivo* hence providing an excellent spatiotemporal tag of endogenous gene expression (Xiong et al. 1998). Furthermore, the property of the human placental AP to label axons, when expressed transgenically in neurons, has been exploited in the construction of the PLAP secretory trap vector which can be utilised as a tool for identifying genes involved in neuronal axon guidance (Leighton et al., 2001).

Reporter systems that allow negative selection have also been described. Such an example is the *galtek* fusion between *lacZ* and the Herpes Simplex virus thymidine kinase (HSV-*tk*) (Cannon et al., 1999) whose inclusion in a gene trap vector (ROSA*galtek*) enables the elimination via FIAU (1-2-deoxy-2-fluoro-β-D-arabinofuranosyl-5-iodouracil) selection of *galtek*-marked cells during *in vitro* differentiation and hence the function of expressing cells (and consequently the trapped gene's role) can be elucidated without requiring any prior knowledge of their identity or anatomical position (Cannon et al., 1999).

More recent vector designs incorporate fluorescent protein-encoding genes such as GFP or eGFP as reporter tags of trapped gene expression (Ishida and Leder, 1999; Lai et al., 2002; Taniwaki et al., 2005; Xin et al., 2005; Zheng and Hughes, 1999). The use of fluorescent reporters offers the

advantage of allowing the non-invasive monitoring of expression in living cells and embryos. Additionally, flow cytometry provides an extra means to monitor fluorescence easily, in an automated manner making it a useful system for high-throughput approaches. Some groups have also reported the successful employment of entrapment vectors that include bipartite reporter fusions that combine eGFP with selection marker genes such as nitroreductase (Medico et al., 2001), *neo* (Chen and Chen, 2004), *hygro* (Cobellis et al., 2005) and *puro* (De Palma et al. 2005). Finally, a recently described gene trap system uses a β-lactamase reporter enzyme to cleave a non-toxic fluorogenic substrate, CCF2/AM, which results in a shift in the emission wavelength from 530 nm (green) to 460 nm (blue), as expression of the trapped gene increases (Whitney et al., 1998; Scheel et al., 2005). This enzyme and fluorophore-based combination allows "measurement of small changes in transcriptional activity through enzymatic amplification of the transcriptional signal and quantitative ratiometric analysis of gene expression, making the system less susceptible to variations due to background fluorescence and cell density" (Zlokarnik et al., 1998; Scheel et al., 2005).

#### 1.3.2.3 Microarray-coupled gene trap mutagenesis

The ROSAβgeo vector constituted the basis for the development of a novel microarray-linked gene trap strategy (Chen et al., 2004a). In this case the ROSAβgeo vector was modified to include a 3'poly(A) trap cassette (see section 1.3.6) that contained a PGK promoter driving the expression of the hygromycin resistance gene. The latter lacked a poly(A) signal but incorporated a splice donor (SD) sequence. Gene trap clones were isolated by G418 selection using the 5' *βgeo* component while the poly(A) trap module served as an artificial exon for the cloning of the trapped genes by 3'RACE

PCR (Chen et al., 2004a). This vector (named ROSAFARY) was used for the construction of a library consisting of 2,880 individual gene-trapped ES clones. Trapped transcripts were then employed for the construction of a cDNA microarray which was used for the expression-based identification of trapped genes that are induced or repressed by the platelet-derived growth factor (PDGF) (Chen et al., 2004a).

#### 1.3.2.4 Conditional gene trapping

The current trend in the design of gene trap vectors involves the generation of constructs that include recombination sites allowing recombinase-mediated, post-insertional modifications of the gene-trap locus (Thorey et al. 1998; Araki et al. 1999; Hardouin and Nagy, 2000; Schnutgen et al., 2005; Xin et al., 2005; Cobellis et al., 2005; Taniwaki et al., 2005). This allows the utilisation of a trapped gene's promoter elements to drive the expression of a knocked-in transgene for use in rescue or cell-labelling experiments as well as the spatiotemporal control of the induced mutations. These recombinase-mediated cassette exchange (RMCE) strategies have mainly utilised the Cre/loxP system although recent developments in this area incorporate other site-specific DNA recombining enzymes such as FLP recombinase (Cobellis et al., 2005; Schnutgen et al., 2005; Xin et al., 2005).

The Cre/loxP technology was also adopted in the design of a novel inducible gene trapping strategy that uses an ES cell line modified to respond to transient Cre expression by permanent drug resistance switching (Chen et al., 2004b). In the un-induced state the responder ES cells are blasticidin S resistant while Cre-catalysed recombination using loxP sites engineered within the responder ES cells renders them resistant to puromycin. When the responder cells are infected by a retroviral gene trap vector that carries a promoterless (linked to a splice acceptor) *cre* gene as a reporter, then the drug

resistance status of the trapped responder cells reflects the expression state of the trapped gene, since Cre expression is regulated by the trapped gene's transcriptional activity. The employment of the above system by Chen et al. (2004b) enabled the enrichment of integration events that occur within developmentally regulated genes (85% of the cases) and demonstrated that the *cre* reporter is extremely sensitive, a property that can be exploited for "accessing chromosomal regions that are not detectable by other reporters" (Chen et al., 2004b).

#### 1.3.2.6 Targeted trapping

A novel approach, termed 'targeted trapping', which combines the power of gene targeting with the simplicity of gene trapping has recently been described (Friedel et al., 2005) (Figure 1.9). Targeted trapping is based on the construction of targeting vectors which consist of a promoterless gene trap cassette flanked with genomic sequences that serve as homology arms facilitating via homologous recombination insertion into an intron of a target gene (Friedel et al., 2005). The vector's reporter gene (*βgeo*) is expressed only if correct splicing takes place between an upstream, endogenous splice donor and the construct's splice acceptor in a manner similar to random gene trapping (Figure 1.9). Hence antibiotic resistance is dependent on the activity of the 'trapped/targeted' endogenous promoter and this should theoretically result in the elimination of most non-specific insertions. A further degree of enrichment is obtained through the employment of gene trap cassettes that lack a translational initiation codon to promote the selection of in-frame integrations. Friedel et al. (2005) demonstrated that a higher targeting efficiency (50%) compared to conventional gene targeting approaches can be achieved using this strategy

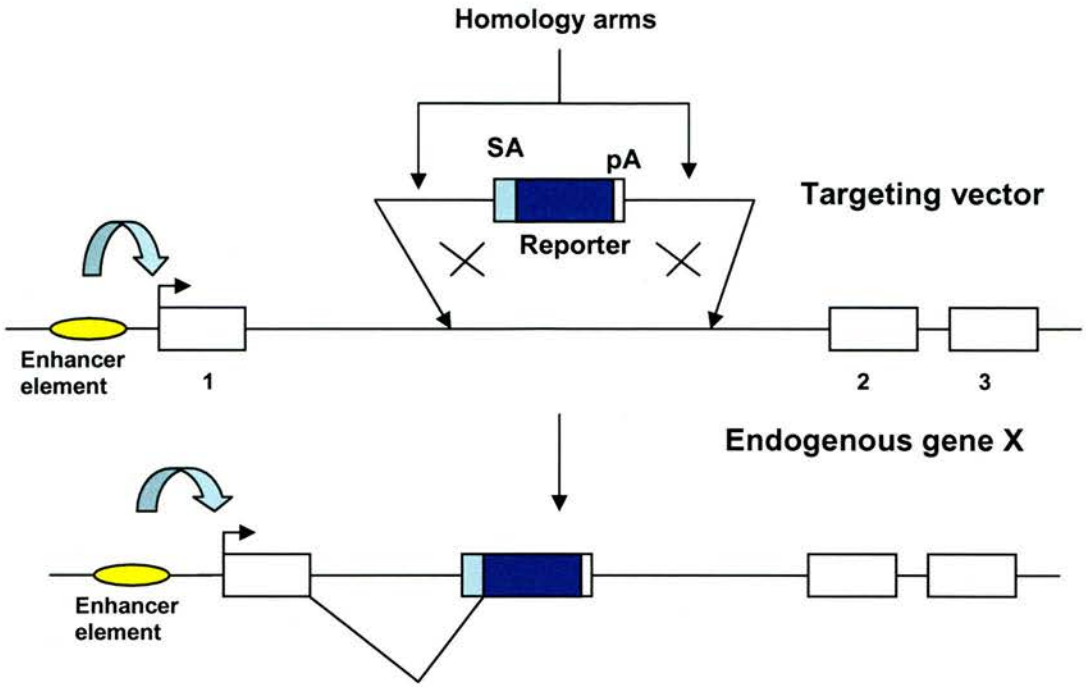



**Figure 1.9** Targeted trapping. Targeted trapping relies on homologous recombination to introduce a promoterless gene trap cassette into the targeted locus. The trapping cassette is flanked by genomic sequences of the target locus that act as homology arms. Exons are depicted by open boxes. SA, splice acceptor; pA, polyadenylation signal. (Figure adapted from Skarnes, 2005).

provided that targeted loci are expressed above a certain threshold (1% of the transferrin receptor gene expression level for the specific vector they used).

### 1.3.3 Technical issues and limitations

Analysis of mutant mice resulting from gene trapping experiments has revealed that the expression of a gene trap vector's reporter gene may not faithfully recapitulate that of the trapped gene in all cases (e.g. Deng and Behringer, 1995; Pall et al., 2004). A lack of correlation could be caused, for example, by the insertional disruption of intronic regulatory elements that are essential for gene activity, often, in a tissue-specific manner (Lothian and Lendahl, 1997; Haerry and Gehring, 1997) and confirmation of the trapped gene's expression profile for example by *in situ* hybridisation is always a good practice.

Obviously the mutagenic outcome of a gene trap event depends on the location at which a vector integrates within a transcriptional unit. Some insertions, particularly those taking place at the 3'end of genes are more likely to generate hypomorphic rather than null mutations. Although this type of genetic lesion might still be informative, it usually does not provide an insight into the trapped gene's function because of the lack of a phenotype. Another reason for a lack of phenotype is "splicing around" events. These occur when a gene trap vector's splice acceptor is ignored by the endogenous splicing machinery resulting in the generation of a wild-type transcript and lack of a phenotype (Skarnes et al., 1992; Sam et al., 1998; Faisst and Gruss, 1998; Voss et al., 1998). It is therefore important to confirm (e.g. by RT PCR or Northern blotting) the existence of a fusion transcript predicted to result from a gene trap event.

The method of delivery (electroporation or retroviral infection) can be a critical determinant of a vector's integrational behaviour. Introduction via

retroviruses might compromise the theoretically random nature of gene trapping as they tend to insert into a target gene's 5'portion (including the 5' UTR and first intron) (Harbers et al., 1984; Vijaya et al., 1986; Soriano et al., 1987). However, this feature might also render them more suitable for the generation of null mutations. Moreover, the major advantage of retroviral infection is that it ensures the integration of a single copy of the entire vector. On the other hand, electroporation-based delivery strategies are considered to be more random but they often result in tandem vector integrations, chromosomal deletions and rearrangements and electroporated vectors are more sensitive to digestion by exonucleases (Friendrich and Soriano, 1991; Niwa et al., 1993; Forrester et al., 1996; Neilan and Barsh, 1999).

### 1.3.4 Directed trapping

Gene trapping is mainly employed by investigators whose research revolves around the genetic dissection of specific developmental pathways. *In vivo* gene trapping has been used successfully in the past to identify novel genes that are expressed either within specific tissues or in spatiotemporal patterns that may be of interest to the investigator (Gossler et al. 1989; Friedrich and Soriano, 1991; Skarnes et al. 1992; Wurst et al. 1995). In this case, "the developmental function of a trapped gene can be elucidated by breeding transgenic animals that are heterozygous for the gene trap and analysing homozygous offspring for morphological or physiological defects" (Baker et al., 1997). However, this strategy is expensive both in terms of resources and time (since it involves the generation of hundreds of chimeric embryos from a large 'pool' of trapped ES clones). Thus it is not suitable for smaller laboratories with specific research interests. An alternative approach involves the *in vitro* pre-selection of gene trap events prior to generating transgenic animals on the basis of a defined set of parameters tailored to an

investigator's specific biological questions (Baker et al., 1997). Such pre-selection strategies demonstrate the potential of gene trapping as both a reverse and forward genetic mutagenesis strategy, as they can be either phenotype or genotype-driven.

#### 1.3.4.1 *In vitro* phenotype-based screens

This type of screens relies on the isolation of insertional events that give rise to interesting reporter expression patterns *in vitro*. For example, an important selection criterion frequently employed in such 'directed' gene trap screens is the subcellular localisation of the trapped/reporter protein fusion, which is also theoretically an indicator of the trapped protein's function. This concept has been successfully implemented in gene trap studies aiming to identify genes that encode proteins confined to nuclear compartments (Tate et al., 1998; Sutherland et al., 2001). Furthermore, the capacity of ES cells to differentiate into various cell types *in vitro* has been exploited widely in expression and induction gene trap screens which aim to identify developmentally important and lineage-specific genes; trapped ES cell clones are driven to differentiate *in vitro* (see Section 1.1.3) and the ones that exhibit reporter expression at the desired developmental stage are selected for further analysis *in vivo*.

Expression trap screens have been employed successfully to identify and mutate genes that are expressed in haematopoietic (Stanford et al. 1998; Muth et al. 1998; Hidaka et al. 2000) and endothelial lineages (Stanford et al. 1998; Hirashima et al. 2004), cardiomyocytes (Baker et al. 1997), chondrocytes (Baker et al. 1997), cardiomyocytes (Chen and Chen, 2004), and neurons (Shirai et al. 1996). Induction screens, (which involve the directed administration of specific developmental cues) have identified and mutated genes regulated by retinoic acid (Forrester et al. 1996; McClive et al. 1998;

Sam et al. 1998; Gajovic et al. 1998; Komada et al. 2000), engrailed homeobox proteins (Mainguy et al. 2000), WNT-3A proteins (Yamaguchi et al., 2005), BMP2 (Kluppel et al. 2002) and FGF1 (Tateossian et al. 2004).

An alternative approach aiming to characterise genes of a predefined developmental and functional profile involves the use of gene trapping in cell lines other than ES cells. Some examples include the isolation of: genes activated during programmed cell death by applying a Cre/loxP gene trap system to interleukin-3 (IL-3)-dependent FDCP1 haematopoietic cells (Russ et al., 1996; Wempe et al., 2001); genes modulated by the granulocyte-macrophage colony-stimulating factor (GM-CSF) by employing the GM-CSF/IL-3 dependent human premyeloid TF-1 cell line (Baghdoyan et al., 2000); and genes regulated by germ cells and nerve growth factor (NGF) by using the Sertoli differentiated cell line 15P-1 (Vidal et al., 2001).

#### 1.3.4.2 Genotype-driven pre-selection

Genotype-driven pre-selection of gene trap events is based on the prior sequence identification of the disrupted locus. These screens were greatly assisted by the progressive optimisation of RACE protocols in a high-throughput context (Townley et al., 1997) and the completion of the mouse genome sequencing. All large-scale gene trap projects predominantly employ this approach; different gene trap vectors are introduced into ES cells to construct huge libraries which are then 'screened' at a genomic level through the generation of RACE tags unique for each gene trap clone. These RACE sequences which are usually publicly accessible can then be utilised for the further isolation and analysis (*in vitro* or *in vivo*) of gene trap clones that carry an insertion at a locus of interest.

### 1.3.5 The quest for genome saturation

The International Gene Trap Consortium (Nord et al., 2006; website: http://www.igtc.org/) was established as a subgroup of the International Mouse Mutagenesis Consortium aiming to generate an international resource of ES cells with gene trap insertions in every, or most genes in the mouse genome ("genome saturation"). The Consortium brings together all major centres that perform large-scale, high-throughput gene trap mutagenesis such as The Centre for Modelling Human Disease (To *et al.* 2004; http://www.cmhd.ca/sub/genetrap.asp), the German Gene Trap Consortium (Hansen et al., 2003; http://genetrap.de), the Sanger Institute (http://www.sanger.ac.uk/PostGenomics/genetrap/), and BayGenomics (Stryke *et al.* 2003; http://baygenomics.ucsf.edu/) as well as smaller groups, which employ gene trapping to address more specific biological questions. An overview of the major gene trap groups is given in Table 1.3. To date, the IGTC has attained 40% genome coverage in approximately 45,000 gene trap cell lines (Nord et al., 2006). These clones are freely available to the scientific community. A similar high-throughput, gene trap mutagenesis-based project is being carried out independently by the private company Lexicon Genetics yielding the development of the largest to date sequence-tagged, gene trap library (Omnibank; www.lexicon-genetics.com/omnibank/) of >270,000 mouse embryonic stem cell clones. These represent mutations in approximately 60% of mouse genes (Zambrowicz et al., 2003).

The analysis of data obtained from these high-throughput projects employing a variety of different vectors (both plasmid and retroviral) has provided some valuable insights into the nature of gene trap mutagenesis. There is now considerable evidence showing that gene trapping is not as random as was initially believed. Hansen et al. (2003) (German Gene Trap Consortium) demonstrated that different vectors have different site

| Research Group | Website | Comments |
|---|---|---|
| The Sanger Gene Trap Resource | http://www.sanger.ac.uk/genetrap | Focus on high-throughput generation of conditional gene trap clones |
| BayGenomics Gene Trap Project | http://www.baygenomics.ucsf.edu | Large-scale screen using secretory trap vectors |
| Centre for Modeling Human Disease (CMHD) | http://www.cmhd.ca | Mainly employ poly(A) trap vectors-also perform expression trap screens |
| German Gene Trap Project (GGTC) | http://www.genetrap.gsf.de | Mainly based on the use of U3- and ROSAβgeo-based vectors |
| Fred Hutchinson Cancer Research Center (FHCRS) | http://www.fhcrc.org/ labs/soriano/trap.html | Mainly employ ROSAβgeo-based vectors |
| Manitoba Institute of Cell Biology | http://www.escells.ca | Gene trap library constructed by using U3-based promoter trap vectors |
| IRBM-TIGEM | http://genetrap.tigem.it | Use of loxP/FTR containing gene trap vectors |
| EGTC | http://egtc.jp | Focus on conditional gene trapping |
| NAISTrap | http://bsw3.naist.jp/kawaichi/naistrap.html | Use of RET-based poly(A) trap vectors |
| Lexicon Genetics | www.lexicon-genetics.com/omnibank/ | Private company – established a trapped ES clone libarary corresponding to 60% genome coverage |

**Table 1.3** An overview of all major gene trap groups that carry out large-scale gene trap projects.

preferences for integration within the mouse genome. Some of these "hot spots" were found to be vector-independent whereas some others were specific to different vectors. A consensus has emerged that the probability of successfully trapping a gene is influenced by its size and its expression level although secondary chromatin structure might also be a critical determinant (Hansen et al., 2003).

Results obtained from the large-scale gene trap mutagenesis project conducted by Lexicon indicate that the rate of trapping new genes is not linear but declines after a specific point (Zambrowicz et al., 2003; Skarnes et al., 2004) (Figure 1.10). This is corroborated by the public gene trap resources (Figure 1.10); the efficiency of their trapping has now dropped to approximately 10% (i.e. one novel gene mutated per 10 trapped clones isolated) (Skarnes, 2005). However, the initial trapping rate was higher for IGTC compared to Lexicon (Figure 1.10). This may reflect the fact that the IGTC trapping efforts are based on the employment of a combination of different gene trap constructs both retroviral and plasmid. Lexicon, on the other hand, has adopted a less diverse, standardised experimental protocol. It is now widely accepted that the best route to achieve saturation is only through the parallel use of different vector types (Hansen et al., 2003; Skarnes et al., 2004). Interestingly, a similar large-scale project that aimed to identify and mutate genes essential for embryonic development in zebrafish through the use of a retroviral gene trap vector yielded the generation of mutants that correspond to only 25% genome coverage (Amsterdam et al., 2004). This can be probably attributed to the fact that only a single trapping approach was employed, thus providing a further piece of evidence that the use of a combination of vectors and methodologies is the optimal strategy to achieve saturation. This implies that the employment of alternative vector designs,

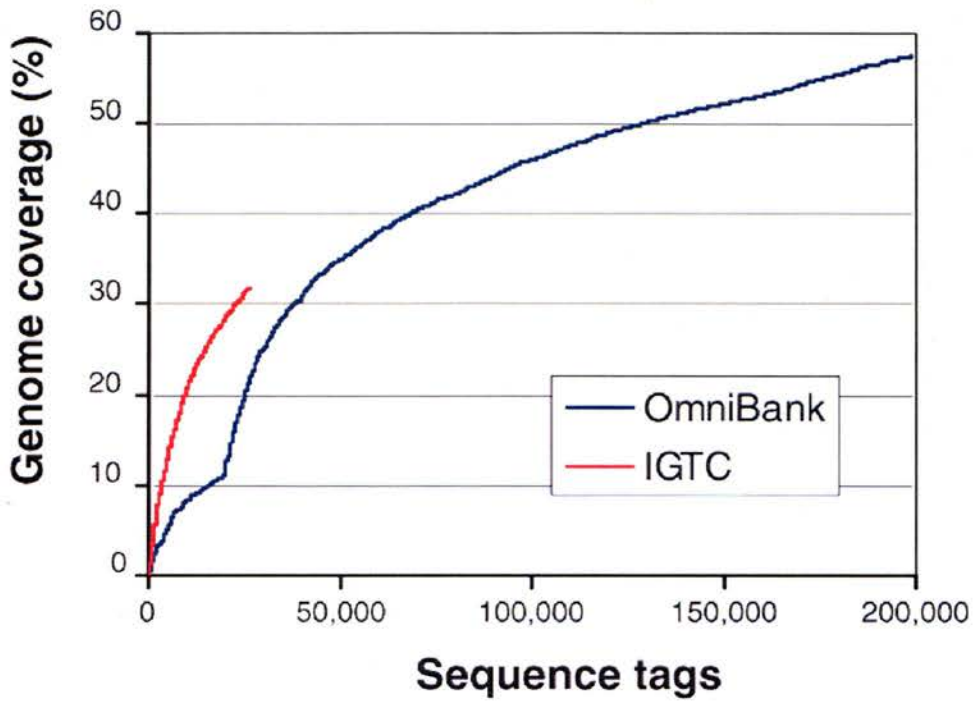



**Figure 1.10** Comparison of the rates of trapping of the IGTC and OmniBank (Lexicon) resources (from Skarnes et al., 2004).

which target fractions of the mouse genome inaccessible to conventional gene trap vectors would be beneficial.

### 1.3.6 Poly(A) trapping

Trapping genes that are not expressed in undifferentiated ES cells is a major challenge for vector designers. Traditional *βgeo* type vectors can lead to the entrapment of even weakly expressed in ES cells genes since selection on the basis of resistance to neomycin is more sensitive than X-gal staining (see Section 1.3.2.1). Hence resistant clones that are βgal negative are likely to carry integrations within genes expressed at low levels but still some minimal degree of transcription is a prerequisite for trapping (Friedrich and Soriano, 1991; Skarnes et al., 1995). This limitation results in the under-representation of mutations in many developmentally important genes which are transcriptionally silent in ES cells; a database search using the IGTC server shows that many genes required for germ layer/lineage specification such as Wnt3a, Scl, MyoD, Nkx2-5, Sox1 and Sox17 have not been trapped to date by any of the gene trap vectors employed by the IGTC members. It is estimated that approximately 60-70% of all mouse genes are accessible to entrapment through the use of *βgeo* constructs (Skarnes et al., 2004; Schnutgen et al., 2005). It is therefore evident that in order to achieve genome saturation using gene trap mutagenesis, alternative vector designs, targeting this 'untrappable' 30-40% of the mouse genome, are desirable. For this reason, a new generation of constructs, termed poly(A) trap vectors, were developed.

A typical poly(A) trap vector is characterized by the presence of a constitutive promoter that drives the expression of a drug resistance marker (Figure 1.11) thus alleviating the need for endogenous trapped gene expression to select for vector insertions and enhancing the vector's potential

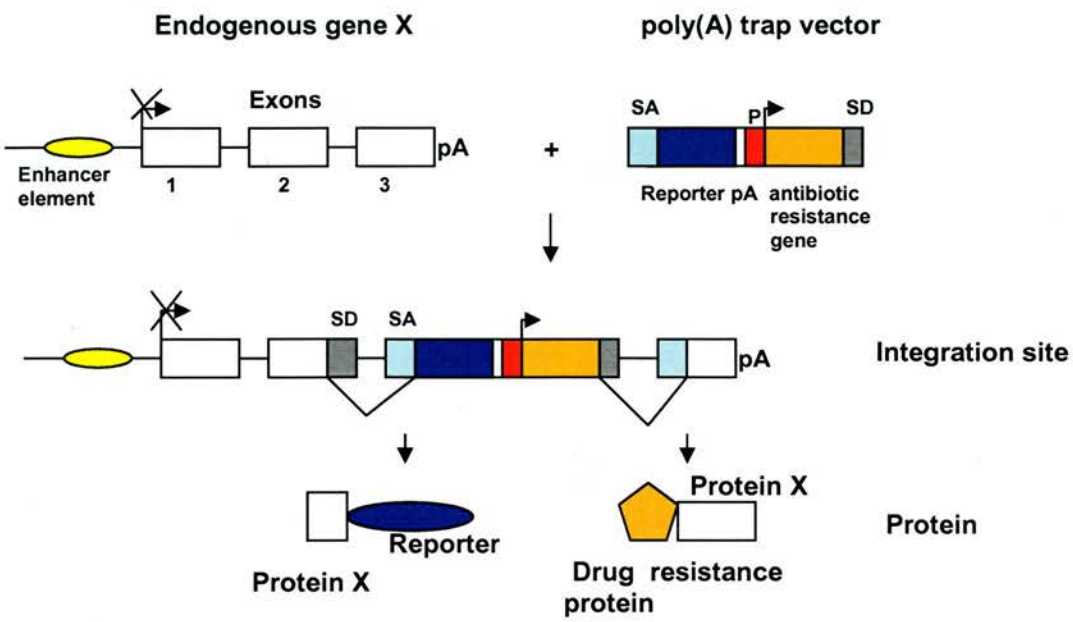



**Figure 1.11** Poly(A) trapping. Schematic representation of poly(A) trapping. PolyA trap vectors contain a promoter driven selectable marker (e.g. *neo*) lacking a polyadenylation signal downstream of a reporter gene (e.g. *lacZ*). A stable fusion *neo* transcript and protein that incorporates downstream endogenous sequences is generated only after insertion into a gene whose endogenous polyA signal is acquired by the vector via a splice donor site. The upstream reporter gene acts as a gene trap and allows the monitoring of endogenous gene expression. P, promoter; pA, polyadenylation signal site; SD, splice donor; SA, splice acceptor.

to trap genes not expressed in undifferentiated ES cells. Furthermore, the selectable marker lacks a poly(A) signal, but includes a splice donor (SD) sequence and hence a spliced polyA signal (derived via splicing) from an endogenous gene is required to generate stable selectable marker's mRNA and, in turn, resistant clones. This feature should theoretically favour the selection of insertions within transcriptional units and the elimination of 'background' intergenic integrations. In addition, the presence of several termination codons following the selectable marker prevents the expression of the 3' trapped exons. The trapped gene's sequence can be determined through the use of 3'RACE PCR which is more straightforward in comparison to the technically challenging 5'RACE protocols. It should be noted that poly(A) trap vectors still employ a 5'SA-reporter-pA component similar to a traditional βgeo-type construct (Figure 1.11) but in this case the 5'component acts exclusively as a means of monitoring the trapped gene's expression status and not as a selector of the gene trap event.

#### 1.3.6.1 Evolution of poly(A) trap vector designs

The first polyA-trap vectors (U1 and U2) to be developed (Niwa et al., 1993) contained a phosphoglycerate kinase (PGK) promoter to drive the expression of the neomycin resistance gene which lacked a polyA addition signal. A *lacZ* reporter was located upstream of PGK promoter-*neo* and downstream of a SA allowing for endogenous gene expression to be monitored. However, the use of these vectors was characterised by large deletions or rearrangements spanning more than 10kb in the 3'-flanking region of the vector (Niwa et al., 1993). An improved poly(A) trap vector (pPAT) was subsequently designed. It incorporated a SD sequence (derived from the mouse *fyn* gene exon 3/intron 3 splice junction) which was placed downstream of the *neo* gene leading to more efficient poly(A) trapping

(Yoshida et al., 1995). Although the integrational behaviour of this vector at the molecular level was not further investigated in depth (only 12 clones were successfully analysed by 3'RACE PCR), its design constituted the basis for the construction of the next generation of poly(A) trap vectors. A modified version of the pPAT vector, which was subsequently employed in a gene trap expression screen designed to enrich for neuron-specific genes, was shown to exhibit a high propensity for concatemer formation (Carroll et al., 2001).

The first evidence indicating the capacity of poly(A) trap vectors to target developmentally regulated genes came from the use of a vector (IRESβgalNeo(-pA)) that consisted of a 5' SA-IRES-βgal-pA cassette and a poly(A) trap component which included the *β-actin* promoter driving the expression of the neomycin resistance gene and a SD from the mouse *Pax-2* gene (Salminen et al., 1998). This vector appeared to integrate within transcriptional units that are inactive in undifferentiated ES cells since the number of the resulting β-galactosidase positive gene trap clones increased after *in vitro* differentiation, suggesting an induction in the expression of trapped lineage-specific genes (Salminen et al., 1998). However, employment of the IRESβgalNeo(-pA) vector in combination with an expression screening strategy showed that the vector's *Pax-2* SD does not always function properly (Hirashima et al., 2004).

The large-scale gene trap project carried out by Lexicon Genetics was initially based on the use of poly(A) trap vectors (Zambrowicz et al., 1998). In this case a retroviral poly(A) trap vector (VICTR20) containing a mutagenic component (SA-IRES-βgeo-pA) and a poly(A) trap module (PGK promoter-puro-SD) was successfully employed to generate a library of 2,000 gene trap ES cell lines (Omnibank; Zambrowicz et al., 1998). Most importantly, this

specific study showed, through RT-PCR-based expression analysis of trapped genes, that several trapped clones carried vector integrations within genes that are not expressed in undifferentiated ES cells (Zambrowicz et al., 1998). However, the VICTR20 poly(A) trap vector was later found to possess a low trapping efficiency (only 10% of the integrations occurred within functional genes; Brian Zambrowicz, personal communication). Consequently, Lexicon focused on the use of a SA-type construct as a tool for 'building' their Omnibank library (http://www.lexgen.com/omnibank/; Zambrowicz et al., 2003). The new construct they used is similar in design to the ROSAFARY vector (section 1.3.2.3). It relies on a promoterless selectable marker (*βgeo*) to facilitate gene entrapment but it also includes a 3'trapping component containing a PGK promoter-driven artificial exon (instead of a selectable marker) and a SD sequence, both derived from the murine *Bruton's tyrosine kinase* (*Btk*) gene (Zambrowicz et al., 2003). The poly(A) trap cassette in this vector serves as a means of acquiring more informative sequence (compared to 5'RACE-obtained sequence) from the 3' end of a trapped gene through 3' RACE.

The most recently developed poly(A) trap vector designs incorporate (as in the case of modern SA-type vectors) loxP and FRT sites for use in RMCE-based approaches (Ishida and Ledder; Cobellis et al., 2005; Xin et al., 2005; Osipovich et al., 2005). This feature is particularly useful for testing and identifying optimal sequence elements thus bypassing the need to construct new vectors (Osipovich et al., 2005).

#### 1.3.6.2 Large-scale projects employing poly(A) trapping

A high-throughput gene trap project that employs poly(A) trapping is being carried out by William Stanford's group in the Centre for Modelling Human Disease (CMHD) (Table 1.3) (To et al., 2004). Their efforts involve the

utilisation of a variety of vectors that contain combinations of different functional components within their 5'reporter and poly(A) trap modules (www.cmhd.ca/genetrap/vectors). Moreover, most of these vectors include loxP sites thus allowing the engineering, through the action of *Cre* recombinase, of sequences inserted in individual trapped clones.

Another RMCE-based, poly(A) trapping strategy that has been used on a large-scale gene trap project (NAISTrap database; website: http://bsw3.naist.jp/kawaichi/naistrap.html) (Table 1.3) is based on the employment of a retroviral polyA trap vector called RET (removable exon trap; Ishida and Leder, 1999) (Figure 1.12). The RET vector employs GFP as the promoterless reporter of disrupted gene expression. It also includes the MC1 promoter-driven HSV-*tk* gene for negative selection and a 3'poly(A) trap component consisting of the RNA polymerase II promoter, *neo* and a SD from the mouse *hprt* gene (exon 8/intron 8 splice junction) (Figure 1.12). Ishida and Leder (1999) tested successfully the functional components of this novel vector and 3'RACE analysis of GFP-negative, trapped clones showed that these clones carry integrations within genes that are not expressed or are very weakly expressed. In addition, they demonstrated that integrated RET proviruses can be efficiently removed from the genome of infected cells using Cre-mediated homologous recombination (Ishida and Leder, 1999). However, poly(A) trapping using the RET vector was subsequently found to be non-random due to the vector's tendency for integrations in the last intron of its target genes (Shigeoka et al., 2005) (see section 1.6.3.4). It is worth noting that the RET vector has been used to create a gene trap clone library from human female ES cells and led to the successful development of a model for studying X chromosome inactivation through the employment of

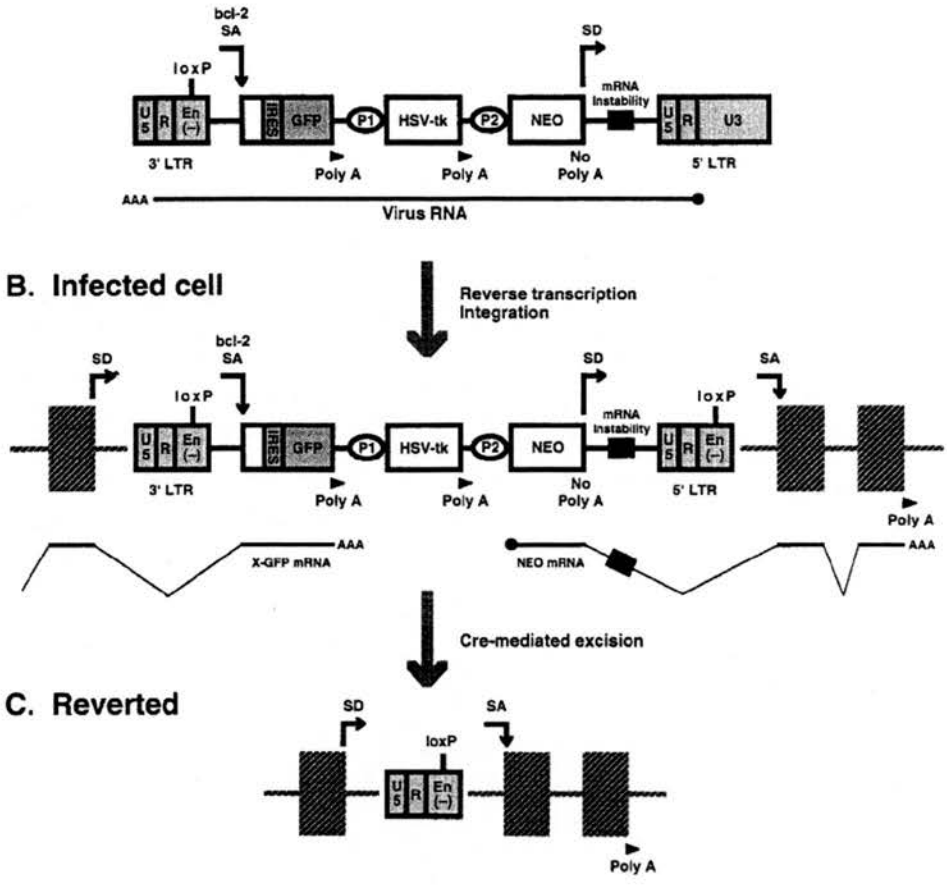



**Figure 1.12** The RET gene-trap vector. (**A**) Structure of the RET vector in a packaging cell line. (**B**) Structure of the RET provirus in infected cells. (**C**) Structure of residual provirus elements following Cre-mediated excision. Diagonally striped black rectangles represent exons of a trapped gene. En (-): enhancer deletion; P1: MC1 promoter; P2: RNA polymerase II promoter (short form); X-GFP mRNA: fusion transcript for the 5'-half of the trapped gene X and the GFP cDNA (from Ishida and Ledder, 1999).

a trapped clone that contained an integration in the X chromosome (Dhara and Benvenisty, 2004).

The same group (NAISTrap) has also reported the efficient coupling of poly(A) trapping with microarray technology; a modified version of the RET vector was used for the generation of a library of 900 trapped ES cell clones and cDNA fragments derived from the disrupted genes were then utilised for the construction of DNA arrays to facilitate profiling of the expression patterns of the trapped loci (Matsuda et al., 2004). This "synergistic coupling" of poly(A) trapping with DNA arrays enables the rapid screening of a large number of clones and the consequent identification of mutated transcriptional units that exhibit expression patterns of interest (Matsuda et al., 2004).

#### 1.3.6.3 AU-rich elements and poly(A) trapping

AU-rich elements (AREs) are *cis*-sequence element regulators of mRNA turnover that reside in the 3′ untranslated region (3′UTR) of many cytokine and oncogene transcripts (Mitchell and Tollervey, 2000a; Zhang et al., 2002). They are composed of variable numbers of the AUUUA pentamer or UUAUUUAUU nonamer and are classified into three categories based on the number and the distribution of the AUUUA pentamer they contain. Class I AREs (e.g. *c-Fos*) contain three to five scattered pentamers coupled with a nearby U-rich region; class II AREs (e.g. GM-CSF, TNF-α, COX-2) include multiple AUUUA motifs with some overlapping; and class III elements (e.g. *c-Jun*) do not contain any pentamers but possess U-rich regions (Zhang et al., 2002).

AREs have been identified by their ability to mediate decay of the host mRNA via deadenylation and subsequent degradation (Brewer and Ross, 1988; Wilson and Treisman, 1988; Shyu et al., 1989; Shyu et al., 1991; Xu et al.,

1997; Ford et al.,1999). However, AREs can also lead to stabilization of an mRNA transcript. The stability status of ARE-containing mRNAs depends on the identity and relative levels of ARE-binding proteins which exert their effects in a redundant/additive or antagonistic fashion (for a review see Barreau et al., 2005). Examples of proteins that promote ARE-mRNA degradation include AUF1 (Loflin et al., 1999), TTP (Lai et al., 1999), and BRF1 (Stoecklin et al., 2002). Furthermore, it has been recently demonstrated that the enzyme Dicer is involved, *in vitro*, in the degradation of a reporter RNA molecule containing the GM-CSF ARE (Takahashi et al., 2006). Examples of ARE-binding proteins that facilitate ARE-mRNA stabilization include HuR (Fan and Steitz, 1998; Peng et al., 1998) and NF-90 (Shim et al., 2002).

The destabilising action of AREs was first demonstrated through the use of a β-globin reporter mRNA. When the ARE from the human GM-CSF was inserted into the 3'UTR of the rabbit β-globin gene, the otherwise stable β-globin transcript was destabilised and reduced to approximately 3% of the wild-type levels (Shaw and Kamen, 1986). The first and rate-limiting step in the degradation of ARE-mRNAs is deadenylation, which can proceed either synchronously (class I and III AREs) or asynchronously (class II AREs) (Chen and Shyu, 1994; Chen et al., 1994; Chen et al., 1995). mRNA deadenylation is followed by 5'-3' (in yeast) or 3'-5' (in mammals) degradation which is carried out *in vitro* by a large, highly conserved amongst eukaryotes, multiprotein complex called the exosome. The yeast exosome consists of at least ten subunits all of which exhibit exonucleolytic and/or RNA binding activities. Both yeast and human exosome components exhibit sequence similarity to *E. coli* RNase PH, RNase D, or RNase R (Allmang et al., 1999; Mitchell and Tollervey, 2000b). It has been shown that certain ARE-binding

proteins (e.g. TTP) can physically associate with the exosome and facilitate its recruitment (Brewer, 1998; Chen et al., 2001) to the target mRNA. RHAU, a putative DExH RNA helicase that also interacts with the exosome is considered to be an additional key player in this process (Tran et al., 2004). 5′-3′ degradation of ARE-mRNAs might also occur in mammalian cells (Kedersha and Anderson, 2002; Gao et al., 2001) and it is linked to cellular entities such as P- or GW bodies (Sheth and Parker, 2003; Kedersha et al., 2005) which contain mRNA decapping (Dcp1 and Dcp2) and degradation (the 5′-3′ exonuclease Xrn1) factors (Kedersha et al., 2005; Ingelfinger et al., 2002; Eystathioy et al., 2003).

The RNA-destabilising property of AREs was exploited in the design of poly(A) trap vectors as a means for reducing background due to intergenic insertions and vector read-through events (Ishida and Ledder, 1999; Matsuda et al., 2004; Osipovich et al., 2005). For example the RET vector was engineered to include an ARE from the human GM-CSF gene located downstream of the vector's SD (Figure 1.12) (Ishida and Ledder, 1999). This should theoretically result in the destabilisation of *neo* mRNA transcripts that arise from direct read-through into the vector's SD intron or target genomic regions. Conversely, proper splicing events between the vector's SD and an endogenous SA should result in the removal of the ARE and the consequent stabilisation of the *neo* fusion transcripts, provided that 'trapping' of a functional gene's poly(A) signal occurred. Initial data from Ishida and Ledder suggest that this approach has a positive effect in enhancing the efficiency of poly(A) trapping: when HAT resistant, RET vector (+ or –ARE)-infected NIH 3T3 tk(-) cells (HAT selection was used to isolate all infection events) were selected in neomycin, the percentage of G418 resistant colonies that contained integrations with the ARE-containing version of the RET

vector was found to be 13% compared to 26% which was obtained after infection with the non-ARE vector (Ishida and Ledder, 1999). Although this reduction is likely to reflect an enrichment in the fraction of desired intragenic integrations, the comparison between the +ARE and the -ARE versions of the RET vector was not accompanied by the corresponding 3'RACE PCR data. Therefore further investigation is required in order to define the role (especially in molecular terms) of an ARE as an enhancer of a poly(A) trap vector's performance.

#### 1.3.6.4 Limitations of poly(A) trapping

Despite the advantages that poly(A) trapping theoretically offers, the performance of constructs that utilise this strategy has so far been quite inconsistent. Some of the limitations associated with the use of poly(A) trap vectors include inefficient 3'splicing due to poor SD performance (Hirashima et al., 2004; Zambrowicz et al., 2003) as well as high background due to a tendency for integrations outside functional genes (Zambrowicz et al., 2003) probably through promiscuous employment of poly(A) signal sites. Several investigators have tried to address these issues by testing various combinations of different internal promoters, selectable markers and splice donor sequences (Zambrowicz et al., 2003; Matsuda et al., 2004; Shigeoka et al., 2005; Osipovich et al., 2005).

The most serious caveat of poly(A) trapping is the recent demonstration that poly(A) trap vectors tend to integrate into the last (3'most) intron of their target genes, a bias that jeopardises the theoretically random nature of poly(A) trapping and, most importantly, its mutagenic potential (Shigeoka et al., 2005). Shigeoka et al., observed that the 88% of gene trap insertions, generated by their large scale gene trap project employing the RET poly(A) vector, exhibited this 3'-most intron bias. The

same tendency was demonstrated for the poly(A) trap constructs employed by the CMHD gene trap group (Shigeoka et al., 2005). This bias is believed to be caused by an mRNA surveillance mechanism called nonsense-mediated mRNA decay (NMD) (Shigeoka et al., 2005). The role of NMD consists of preventing the production of truncated proteins that could function in a dominant-negative fashion by eliminating abnormal transcripts that prematurely terminate translation (for a review see Maquat, 2004). NMD induces the degradation of transcripts in which a termination codon (TC) is placed more than 55-60 nt 5′ to the last exon-exon junction and hence is recognised by the NMD surveillance machinery as a premature TC (PTC) (Shigeoka et al., 2005).

It was hypothesised that the TC of the poly(A) trap cassette's *neo* resistance gene is recognised as a PTC when vector insertion occurs into one of the upstream introns (other than the last one) of a trapped gene (Shigeoka et al., 2005). This subsequently leads to the degradation of the resulting neo-trapped gene fusion transcript due to the action of NMD (Shigeoka et al., 2005) (Figure 1.13a). In contrast, vector insertion into a gene's last intron would result in the generation of an NMD-immune fusion transcript because the distance between the *neo* TC and the last exon-exon junction is probably too short to efficiently elicit an NMD response (Shigeoka et al., 2005). This has been proposed to account for the over-representation of poly(A) trap insertions in the last intron of target genes (Figure 1.13b). Consequently, an improved poly(A) trap vector, termed UPATrap, was developed (Shigeoka et al., 2005). Its structure is similar to that of the RET vector apart from the fact that it incorporates an IRES sequence which is flanked by two loxP sites and is positioned between the *neo* cassette's TC and the SD sequence (Shigeoka et al., 2005). This novel poly(A) trapping strategy was found to resist NMD of

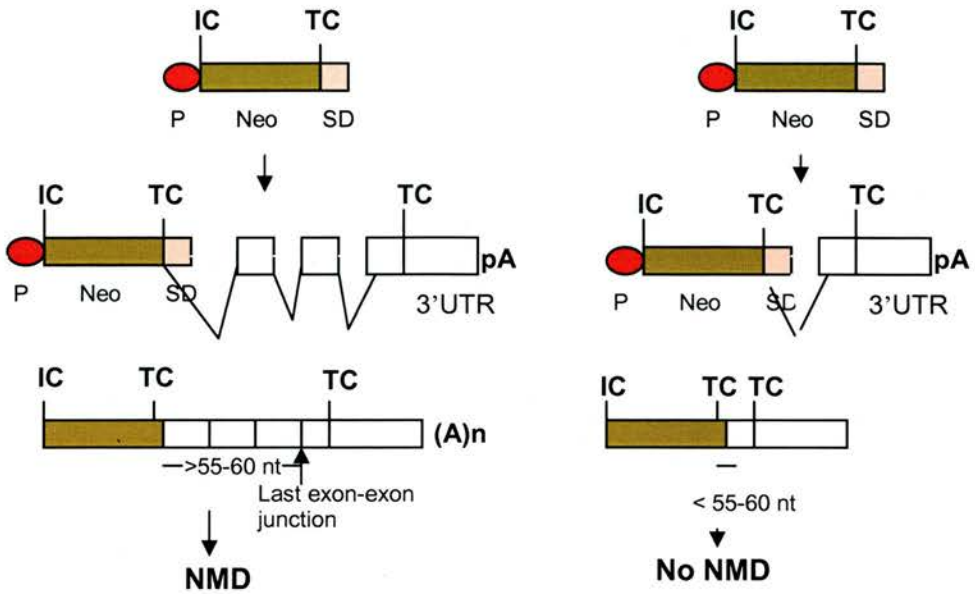



**Figure 1.13** Proposed model for the biased selection of the vector integration sites in poly(A) trapping. **(a)** Vector integration within an intron other than the last one places the termination codon (TC) of the vector's *neo* cassette at a distance greater than 60 nt from the last exon-exon junction causing the degradation of the resulting fusion transcript through NMD since the TC of the *neo* coding sequence is recognised as a PTC. **(b)** Vector integration within a target gene's last intron results in resistance to NMD since the distance between the *neo* TC and the last exon-exon junction is shorter than 60 nt. It should be noted that only the vector's poly(A) trap component is shown. TC, termination codon; IC, initiation codon; P, promoter; pA, polyadenylation signal; SD, splice donor; NMD, nonsense mediated mRNA decay. (Adapted from Shigeoka et al., 2005).

the selectable-marker mRNA and permit the trapping of transcriptionally silent genes without a bias in the vector-integration site (Shigeoka et al., 2005). The authors postulated that the presence of the IRES probably blocks the NMD initiation stages either by "displacing the exon–exon junction complexes (EJCs) from downstream exon–exon junctions" or by promoting the formation of a complex secondary structure at the RNA level (Shigeoka et al., 2005).

The biased nature of poly(A) trapping is likely to contribute to the fact that the *in vivo* mutagenic capacity of poly(A) trap vectors has proved so far unconvincing. The tendency of this type of vectors to integrate within the last intron of genes is likely to result in the generation of hypomorphic mutations that do not generate a phenotype. The few published examples of mutant strains generated through the employment of poly(A) trap vectors seem to support this prediction; there is only one reported example of a poly(A) trap vector-generated mutation that is linked to an obvious (not lethal) phenotype (Tarrant et al., 2002; vector insertion in this case took place within an intron closer to the 5'end of the gene) while the majority of the reported mutants showed no abnormalities (Tsukahara et al., 2000; Tsukahara et al., 2001; Carroll et al., 2001; Hirashima et al., 2004).

## 1.4 Project overview

This thesis reports the characterization of a series of novel gene trap vector constructs. The novel features of these vectors include: (a) a 5'tripartite reporter fusion between the eGFP, β-galactosidase and neomycin or hygromycin resistance genes and (b) a 3'poly(A) trap component that contains a SD sequence derived from the rabbit *βglobin* gene and an ARE from the human GM-CSF gene. The project's experimental design aimed to assess the function of each of these components. We provide evidence that

the triple fusion functions properly and can be potentially used as a reporter of trapped locus activity. We also show that the presence of the ARE appears to improve the performance of the rabbit *βglobin* SD sequence in the context of poly(A) trapping. More importantly, preliminary data suggest that our vectors may be resistant to NMD and thus potentially unbiased in their insertional preference.

# CHAPTER 2

# METHODS AND MATERIALS

## 2.1 Molecular Biology Methods

### 2.1.1 Plasmid Vector Construction

Plasmid gene trap vectors pEGeo2, pEGeo2puroSD2 and pEHygro2neoSD2 (Figure 2.1) were constructed by Dr. Joshua Brickman and Dr. Elena Tzouanakou at the Institute of Stem Cell Research (ISCR). Vector pEGeo2 consists of the mouse En-2 splice acceptor and a triple fusion between egfp, β-galactosidase and neomycin resistance genes followed by the SV40 polyA signal site. Vector pEGeo2puroSD2consists of the mouse En-2 splice acceptor and a triple fusion between egfp, β-galactosidase and neomycin resistance genes followed by the SV40 polyA signal site and a 3′ cassette that contains the human β-actin promoter driving the constitutive expression of the puromycin resistance gene and the rabbit β-globin exon 2/intron 2 splice donor junction. Vector pEHygro2neoSD2 consists of the mouse En-2 splice acceptor and a triple fusion between egfp, β-galactosidase and hygromycin B resistance genes followed by the SV40 polyA sequence site and a 3′ cassette containing the constitutive human β-actin promoter, the neomycin resistance gene and the rabbit β-globin exon 2/intron 2 splice donor junction. It should be noted that in all cases a glycine hinge domain (16 glycine codons) was inserted at both ends of the egfp sequence in order to confer flexibility to the tripartite fusion protein (Bronchain et al., 1999) while the ATG translation initiation codon was removed from egfp.

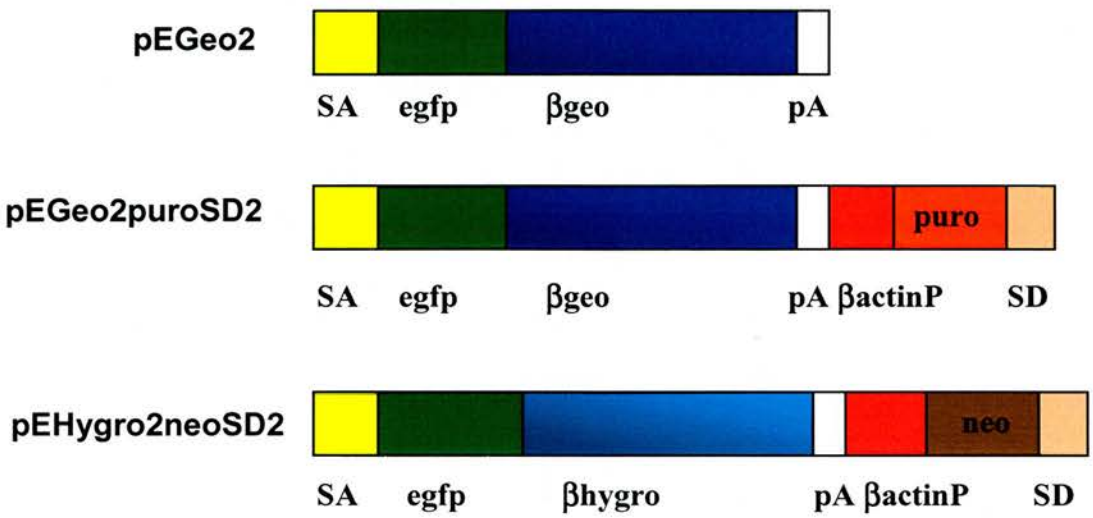



**Figure 2.1** Gene trap vectors employed. Schematic representation of the linearised versions of plasmid gene trap vectors pEGeo2, pEGeo2puroSD2and pEHygro2neoSD2 . *En-2*, engrailed-2; pA, polyA; SA, splice acceptor; SD, splice donor; P, promoter; neo, neomycin resistance gene; βgeo, fusion between β-galactosidase and neomycin resistance genes; βhygro, fusion between β-galactosidase and hygromycin resistance genes.

### 2.1.2 Construction of pEHygro2neoSD2 (+ARE) gene trap vector

Gene trap vector pEHygro2neoSD2 (+ARE) (Figure 2.2) is identical to plasmid pEHygro2neoSD2 but it also includes a 50 bp AU-rich RNA destabilising element (ARE) (TAATATTTATATATTTATATTTTAAAATATTTATTTATTTATTTATTTAA ) from the human GM-CSF gene (Xu et al. 1997). The map of the plasmid is given in Appendix 1. The construction of the vector involved the following steps:

(i) Plasmid pBKnSDARE (30 µg) which contains the constitutive human β-actin promoter, the neomycin resistance gene, the rabbit β-globin IVS2 splice donor (exon 2/intron 2 splice donor junction) and the ARE was digested at 37°C with restriction endonucleases AscI and PacI (New England Biolabs) (70 U of each) and the resulting 2404 bp fragment predicted to include the β-actin P/neo/SD cassette was excised and gel-purified using the Qiaquick® Gel Extraction kit (Qiagen) following the manufacturer's guidelines.

(ii) Gene trap vector pEHygro2neoSD2 (Figure 2.1) was also doubly digested with AscI/PacI and the resulting 10085 bp fragment (SA/egfp/βhygro/pA cassette) was excised and gel-purified.

(iii) The purified 10085 bp fragment was dephosporylated by incubating together with 2 µl of Shrimp Alkaline Phosphatase (USB Corporation) at 37°C for 15 minutes and then at 65°C for 15 minutes and then ethanol-precipitated by adding 1/10 of the total reaction volume (1/10V=3 µl) of 3M sodium acetate (pH 5.2) and 2 volumes (=60 µl) of 100% ethanol, freezing for 5 minutes at –140 °C and spinning at 13,000 rpm for 20 minutes in a tabletop microcentrifuge (Biofuge 13, Heraeus, Sepatech). The precipitated pellet was washed twice with

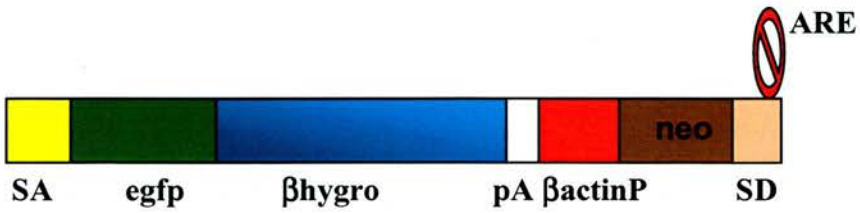



**Figure 2.2** Schematic representation of the linearised version of plasmid gene trap vector pEHygro2neoSD2 +ARE. *En-2*, engrailed-2; pA, polyA; SA, splice acceptor; SD, splice donor; P, promoter; neo, neomycin resistance gene; βhygro, fusion between β-galactosidase and hygromycin resistance genes.

70% ethanol and after air-drying in a laminar flow sterile hood for 5-10 minutes it was resuspended in 20 µl of distilled $H_2O$.

(iv) The 2404 bp fragment generated from the double AscI/PacI digestion of plasmid pBKnSDARE and the 10085 bp fragment produced from the double AscI/PacI digestion of plasmid vector pEHygro2neoSD2 were mixed together at various different molar ratios (also included vector- and insert-only controls) in the presence of 2 U of T4 DNA ligase (Roche) and 2 µl 10x ligation buffer (660 mM Tris-HCl, 50 mM $MgCl_2$, 50 mM DTT, 10 mM ATP, pH 7.5 at 20°C-final reaction volume=20 µl). The ligation reactions were left at 16°C overnight.

(v) 3 µl of each ligation reaction were added into 40 µl aliquots of MAX Efficiency ® DH5α™ chemically competent cells (Invitrogen) and the mixtures were left on ice for 40 minutes followed by heat-shocking at 42°C for 45 s. After heat-shock the cells were transferred onto ice and left for 10 minutes. S.O.C medium (2% Tryptone; 0.5% Yeast Extract; 10 mM NaCl; 2.5 mM KCl; 10 mM $MgCl_2$; 10 mM $MgSO_4$; 20 mM glucose-Invitrogen) was then added and the tubes were shaken horizontally (200 rpm) at 37 $^0$C for 1 hour and 20 minutes. The transformants were spread on pre-warmed ampicillin-containing (100 µg/ml) LB agar plates, which were then incubated overnight at 37 $^0$C. Approximately 30 ampicillin-resistant clones were obtained and plasmid DNA was isolated from each of them using the QIAprep Spin Miniprep Kit (Qiagen). The success of the ligation reactions was determined by a combination of diagnostic restriction enzyme digests, PCR and sequencing of the resulting DNA minipreps.

### 2.1.3 Nucleic Acid Manipulation and Cloning

#### 2.1.3.1 Transformation of Bacterial Cells

1µl of plasmid DNA was added into a 75 µl aliquot of chemically competent, subcloning efficiency DH5α bacterial cells. After mixing gently, cells were incubated on ice for 15 minutes and then heat-shocked at 42°C for 2 minutes. The transformation mixture was then immediately transferred onto ice and after a further 5-10 minute incubation, 1ml of pre-warmed Luria Broth (LB) culture medium was added and the tube containing the mixture was left shaking (200 rpm) horizontally at 37°C for 1 hour in an orbital shaker. Finally, transformed cells were spun at 3,000 rpm for 3 minutes in a table-top microcentrifuge (Biofuge 13, Heraeus, Sepatech) and 50 µl of the resuspended cell pellet (volume of approximately 100 µl) were spread on a prewarmed LB agar plate with 100 µg/ml ampicillin as selection. The plate was left at 37°C overnight until visible colonies appeared. Glycerol stocks of transformed cells were made by inoculating 5 ml of selective LB medium (100 µg/ml Ampicillin) with a single transformed colony picked with a sterile yellow Gilson tip from a fresh plate. The culture was grown at 37°C for about 8 hours with shaking to reach log phase and then cells were centrifuged at 4,000 rpm for 5 minutes and the pellet was resuspended in LB containing 15% sterile glycerol. Cells were then dispensed into 500 µl aliquots in pre-chilled screw capped eppendorf tubes and stored at –80°C.

#### 2.1.3.1 TA Cloning of PCR Products

Purified PCR products were cloned using the TOPO TA Cloning® Kit for Sequencing (Invitrogen) following the supplier's guidelines. Briefly, direct insertion of the purified PCR products into the linearised pCR®4-TOPO® plasmid vector (Invitrogen), which contains overhanging 3′ deoxythymidine (T) residues, was performed by incubating 4 µl of purified

PCR product with 1 µl pCR®4-TOPO® vector and 1 µl salt solution (200 mM NaCl; 10 mM $MgCl_2$) provided by the kit at room temperature for approximately 5 minutes. 2 µl of the TOPO® Cloning reaction were then added into a vial of One Shot® TOP10 Chemically Competent E. coli cells (Invitrogen) and cells were incubated on ice for approximately 10 minutes. The ice incubation was followed by heat-shocking of the cells for 30 seconds at 42$^0$C after which the tubes were immediately transferred onto ice and 250 µl of room temperature S.O.C medium (2% Tryptone; 0.5% Yeast Extract; 10 mM NaCl; 2.5 mM KCl; 10 mM $MgCl_2$; 10 mM $MgSO_4$; 20 mM glucose-Invitrogen) was added. Tubes were capped tightly and shaken horizontally (200 rpm) at 37$^0$C for 1 hour. 50 µl from each transformation were spread on a pre-warmed ampicillin-containing (100 µg/ml) LB agar plate, which was then incubated overnight at 37$^0$C. The resulting colonies were picked for subsequent plasmid DNA isolation and sequencing (see below).

### 2.1.3.2 DNA Isolation

#### 2.1.3.3.1 DNA plasmid minipreps

A single colony was picked with a sterile yellow Gilson tip and used for inoculation of 5 ml of ampicillin-containing (100 µg/ml) LB medium. The culture was incubated at 37 $^0$C overnight with shaking. Bacterial cells were harvested by centrifugation at 4,000 rpm for 5 minutes (CENTRA-3, IEC). Plasmid DNA extraction was then carried out using the QIAprep Spin Miniprep Kit (Qiagen) according to the manufacturer's instructions. The bacterial pellet was resuspended in 250 µl of Buffer P1 (Resuspension Buffer-50 mM Tris-Cl, pH 8.0; 10 mM EDTA; 100 µg/ml RNase A) and transferred to an eppendorf microcentrifuge tube. Buffers P2 (Lysis Buffer-200 mM NaOH; 1% SDS) and N3 (high salt-containing buffer) were then added sequentially

(250 µl and 350 µl respectively) to facilitate alkaline lysis of the cells, lysate neutralization and precipitation of proteins and chromosomal DNA. The tube was centrifuged for 10 minutes at 13,000 rpm in a table-top microcentrifuge (Biofuge 13, Heraeus, Sepatech) and the supernatant was applied by pipetting to a QIAprep spin column (contains a silica membrane for selective adsorption of plasmid DNA in the presence of the high-salt buffer N3) which was then centrifuged for 1 minute at 13,000 rpm. After discarding the flow-through the column was washed first by adding 0.5ml Buffer PB and centrifuging for 1 minute at 13,000 rpm and then by adding 0.75 ml Buffer PE (ethanol-containing) and centrifuging twice for 1 minute at 13,000 rpm. Elution of DNA was performed by adding 35 µl of buffer EB to the centre of the column (placed in a clean 1.5 ml eppendorf microcentrifuge tube), letting it stand for 1 minute and centrifuging at 13,000 rpm for 1 minute.

#### 2.1.3.3.2 DNA plasmid maxipreps

For the preparation of higher amounts of plasmid DNA the HiSpeed Plasmid Maxi Kit (Qiagen) was employed following the manufacturer's instructions. A single colony was picked with a sterile yellow Gilson tip and used for inoculation of 150 ml of ampicillin-containing (100 µg/ml) LB medium. The culture was incubated at 37$^0$C overnight with shaking (200 rpm). Bacterial cells were harvested by centrifugation at 6000 rpm for 15 minutes at 4$^0$C (SLA-1500 rotor, RC5C, Sorvall Instruments, Du Pont). All buffers were provided in the kit unless otherwise stated. Briefly, the pellet was resuspended in 10 ml buffer P1 and cells were lysed in 10 ml of buffer P2. The lysate was then neutralized by addition of chilled buffer P3 (3 M potassium acetate, pH 5.5) and poured into the barrel of a QIAfilter Cartridge

where it was left for 10 minutes at room temperature. After debris precipitation the cleared lysate was loaded onto a HiSpeed Maxi Tip (an anion-exchange resin that was pre-equilibrated by applying 10 ml of buffer QBT-750 mM NaCl; 50 mM MOPS, pH 7.0; 15% isopropanol; 0.15% Triton® X-100) by gravity flow. The HiSpeed Maxi Tip was then washed with 60 ml of medium-salt buffer QC (1 M NaCl; 50 mM MOPS, pH 7.0; 15% isopropanol). The plasmid DNA was eluted from the HiSpeed Maxi Tip using 15 ml of high-salt buffer QF (1.25 M NaCl; 50 mM Tris-Cl, pH 8.5; 15% isopropanol) and precipitated using 10.5 ml of isopropanol. The precipitated plasmid DNA was then applied to the QIAprecipitator Module using the syringe provided in the kit and washed with 2 ml of 70% ethanol. The final step involved the elution of the plasmid DNA into a microcentrifuge tube using 500 µl of buffer TE (10 Mm Tris-Cl, pH 8.0; 1 mM EDTA). Plasmid DNA concentration for both minipreps and maxipreps was determined by UV spectrophotometry at 260 nm.

#### 2.1.3.3.3 Preparation of genomic DNA for Southern Blotting

Gene trap ES cell clones were plated on a 96-well plate and grown to confluency. Each well was washed, after aspirating the medium, with 100 µl of PBS and 50 µl of Lysis Buffer (10 mM Tris, pH7.5; 10 mM EDTA; 10 mM NaCl; 0.5% Sarcosyl; 1mg/ml Proteinase K) was added. The plate was then sealed with parafilm and left overnight at 55°C. After the incubation, 200 µl of 95% ethanol/75 mM NaOH was added to each well and the plate was left at room temperature overnight. The ethanol/NaOH mixture was then removed and the wells were washed three times with 70% ethanol and left to dry at room temperature. 30 µl of XbaI (10 U, Roche) restriction digest cocktail (also containing 1XBuffer H, 1 mM Spermidine, 100 µg/ml BSA, 100

µg/ml RNase A) was added to each well followed by an overnight incubation at 37°C. The next day DNA loading buffer was added directly to the wells and the samples were ran overnight at 25 V on a 0.7% TAE agarose gel.

### 2.1.3.4 Vector preparation prior to electroporation

Plasmid vectors pEGeo2, pEGeo2puroSD2and pEHygro2neoSD2 were digested using the restriction endonuclease HindIII (New England Biolabs) for 4 hours by incubation at 37°C. Approximately 180µg of each vector were digested using 10 units of enzyme per µg of plasmid and 1X the supplier's recommended optimal buffer (NEBuffer 2). Once complete linearisation was confirmed by gel electrophoresis plasmids were precipitated by adding 1/10 of the total reaction volume (1/10 V) of 3M sodium acetate (pH 5.2) and 2 volumes of 100% ethanol, freezing for 5 minutes at –140 °C and spinning at 13,000 rpm for 20 minutes in a tabletop microcentrifuge (Biofuge 13, Heraeus, Sepatech). The precipitated pellet was washed twice with 70% ethanol and after air-drying in a laminar flow sterile hood for 5-10 minutes it was resuspended in PBS to obtain a final concentration of 1 µg/µl.

The modified version of vector pEHygro2neoSD2 (+/-ARE) without the cryptic SA was generated by double digestion of approximately 40 µg of DNA with restriction endonucleases HindIII and MfeI (New England Biolabs) using 1X the supplier's recommended buffer for digestion with MfeI (NEBuffer 4), 1 U of MfeI/µg of plasmid and 2 U of HindIII/µg of plasmid all present in the same reaction mixture. Digestion was performed at 37°C overnight. The largest restriction fragment (size of 9501 or 9551 bp depending on the vector used) was then excised and purified from the gel slice using the QIAEXII gel purification kit (QIAGEN) following the

manufacturer's instructions. In brief, the gel slice was transferred into a microcentrifuge tube and weighed. Three volumes of buffer QX1 (high salt-containing, solubilisation buffer), two volumes of $H_2O$ and 70 µl of QIAEXII suspension (consists of silica-gel particles) were then added to the tube containing the gel slice and the mixture was incubated at 50°C for 10 minutes. The sample was centrifuged for 30 s and the supernatant was removed by pipetting. The pellet was then washed once with 500 µl of buffer QX1 to remove residual agarose and three times with ethanol-containing buffer PE to remove salt contaminants and air-dried in a laminar flow sterile hood for 15 minutes until it became white. DNA elution was performed by adding 20 µl of PBS (pH 7), incubating at 50°C for 5 minutes and centrifuging for 30 s.

#### 2.1.3.5 RNA Isolation

Total RNA was isolated from gene trap embryonic stem (ES) cell clones grown either on 6-well plates or 25 $cm^2$ flasks using the mono-phasic phenol/guanidium isothiocyanate TRIZOL® reagent (Invitrogen) according to the manufacturer's instructions. In summary, cells were lysed by adding 1 ml (6-well plates) or 2.5 ml (25 $cm^2$ flasks) of TRIZOL reagent, transferred into a 1.5 ml RNase-free Eppendorf microcentrifuge tube and then incubated for 5 minutes at room temperature. 0.2 ml of chloroform per 1ml of TRIZOL reagent was then added and the tube was shaken vigorously by hand for 15 s and incubated at room temperature for 2 to 3 minutes. After incubation the sample was centrifuged at 11,000 rpm for 15 minutes at 4°C (Centrifuge 5415 C, Eppendorf). Following centrifugation the RNA-containing upper aqueous phase was transferred to a fresh tube and RNA was precipitated using 0.5 ml of isopropanol per 1ml of TRIZOL reagent used for the original

homogenization. The sample was then incubated for 10 minutes at room temperature and centrifuged at 11,000 rpm for 10 minutes at 4°C. After removal of the supernatant by decanting, the gel-like RNA pellet was washed with 75% ethanol by adding 1 ml of 75% ethanol per 1ml of TRIZOL reagent used for the original homogenization, mixing by vortexing and centrifuging at 9,000 rpm at 4°C. Finally, the RNA pellet was air-dried for 5-10 minutes, resuspended in 40 or 80 µl of RNase-free water (Ambion) and incubated for 10 minutes at 55°C. The RNA sample concentration was determined by spectrophotometry and RNA integrity was assessed by agarose gel electrophoresis. RNA samples were stored at -80°C.

#### 2.1.3.6 DNase-treatment of total RNA samples

Contaminating genomic DNA was removed from total RNA samples by employing the TURBO DNA-free™ kit (Ambion) in accordance with the supplier's guidelines. All reagents were provided in the kit unless otherwise stated. In brief, 3 U of TURBO DNase and 0.1 volume 10X TURBO DNase buffer were added to an RNA preparation followed by incubation at 37°C in a water bath for 30 minutes, a further addition of 3 U of TURBO DNase and a further 30 minute incubation at 37°C. Removal of the DNase as well as divalent cations was performed at the end of the second incubation by adding 0.2 volumes of DNase Inactivation Reagent and incubation for 2 minutes at room temperature during which the sample was mixed occasionally. RNA was recovered in the supernatant of the reaction after centrifugation at 13,000 rpm for 1.5 minutes (Biofuge 13, Heraeus, Sepatech), transferred into a fresh, RNase-free microcentrifuge tube and stored at -80°C.

### 2.1.3.7 polyA+ RNA Isolation from total RNA

polyA+ mRNA was isolated from total RNA preparations using the Oligotex mRNA Spin-Column kit (Qiagen) following the supplier's guidelines. Approximately 10 µg of total RNA was transferred into an RNase-free 1.5 ml microcentrifuge tube and the volume was adjusted with RNase-free water to 250 µl. 250 µl Buffer OBB (20 mM Tris:Cl, pH 7.5; 1 M NaCl; 2 mM EDTA; 0.2% SDS-pre-warmed at 37$^0$C) and 15 µl of Oligotex Suspension (10% Oligotex particles in 10 mM Tris-Cl, pH 7.5; 500 mM NaCl; 1mM EDTA; 0.1% SDS; 0.1% $NaN_3$) were then added to the RNA preparation and the sample was incubated for 3 minutes at 70$^0$C in a heating block and then at room temperature for 10 minutes. It should be noted that the Oligotex suspension consists of polystyrene-latex particles of uniform size that are covalently linked to $dC_{10}T_{30}$ oligonucleotides designed to promote binding of polyadenylic acids. The Oligotex:RNA complex was pelleted by centrifugation for 2 minutes at 13,000 rpm (Biofuge 13, Heraeus, Sepatech) and, after removal of the supernatant by pipetting, resuspended in 400 µl of wash buffer OW2 (10 mM Tris:Cl, pH 7.5; 150 mM NaCl; 1 mM EDTA). It was then transferred onto an RNase-free small spin column placed in a 1.5 ml microcentrifuge tube and centrifuged for 1 minute at 13,000 rpm. After centrifugation the spin column was transferred into a new 1.5 ml microcentrifuge tube, washed for a second time with 400 µl of buffer OW2 and centrifuged for one minute at maximum speed. Finally, the spin column was transferred into a new 1.5 ml microcentrifuge tube and polyA+ mRNA was eluted by applying 20 µl of buffer OEB (5 mM Tris:Cl, pH 7.5-pre-heated at 70$^0$C) and centrifuging for one minute at 13,000 rpm. A second elution step with another 20 µl of buffer OEB and a centrifugation was performed. The resulting samples were stored at 80$^0$C.

### 2.1.4 Polymerase Chain Reaction (PCR)

All PCR reactions were performed using the Peltier Thermal Cycler PTC-200 (MJ Research). dNTPs were supplied by ABgene (Advanced Biotechnologies). All primers were obtained by MWG Biotech unless otherwise stated.

#### 2.1.4.1 5' Rapid Amplification of cDNA ends-PCR (5' RACE-PCR)

5'RACE-PCR was performed by employing the GeneRacer™ kit (Invitrogen) following the manufacturer's guidelines. An overview of the strategy employed is shown in Figure 2.3. In brief, polyA+ mRNA samples were treated with Calf Intestinal Phosphatase to eliminate truncated mRNA and non-mRNA species, then treated with Tobacco Acid Pyrophosphatase to remove the 5'cap structure from intact, full length mRNA and finally 5'ligated to the GeneRacer™ RNA Oligo. cDNA synthesis then followed using a gene trap vector-specific primer. Finally, the first-strand cDNA was amplified using a reverse gene trap vector-specific primer and a primer homologous to the GeneRacer™ RNA Oligo (the GeneRacer™ 5' Primer). A second round of PCR using nested primers was also performed. The resulting 5'RACE-PCR products were gel purified and cloned using the TOPO TA Cloning® Kit as described in Section 2.1.3.2. All reagents used were provided in the kit unless otherwise stated.

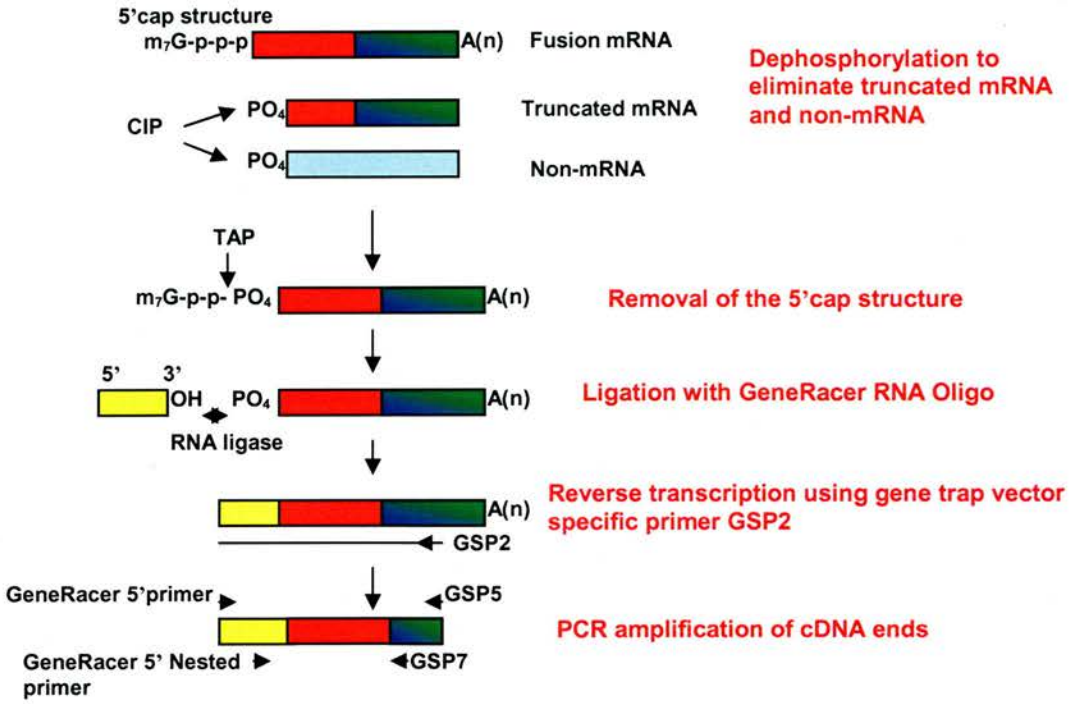



**Figure 2.3** Overview of the 5'RACE PCR strategy employed. Red box, trapped gene's sequence; Blue/green box, gene trap vector's reporter gene sequence; Yellow box, GeneRacer RNA Oligo; CIP, calf intestinal phosphatase; TAP, tobacco acid pyrophosphatase.

*(i) Dephosphorylation*

7 µl (50-250 ng) of polyA+ mRNA was dephosphorylated using 1 µl Calf Intestinal Phosphatase (10 U/µl in 25 mM Tris-HCl, pH 7.6; 1 mM $MgCl_2$; 0.1 mM $ZnCl_2$; 50% glycerol), 1 µl 10XCIP Buffer (0.5 M Tris-Cl, pH 8.5; 1 mM EDTA) and 1 µl RNaseOut™ (40 U/µl in 25 mM Tris-Cl, pH 8; 50 mM KCl; 0.5 mM EDTA; 8 mM DTT; 50% glycerol). The 10 µl reaction was set up in a 1.5 ml microcentrifuge tube and after mixing gently its components by pipetting it was incubated at 50°C in a heat block for 1 hour. After incubation the tube was centrifuged briefly and placed on ice. Dephosphorylation was followed by RNA precipitation. 90 µl RNase-free water and 100 µl phenol:chloroform:isoamyl alcohol (25:24:1) were added and the reaction was vortexed vigorously for approximately 30 seconds followed by centrifugation for 5 minutes at 13,000 rpm at room temperature (Biofuge 13, Heraeus, Sepatech). The top aqueous phase was then transferred to a new sterile microcentrifuge tube and 2 µl of mussel glycogen (10 mg/ml), 10 µl of Sodium Acetate (3M, pH 5.2) and 220 µl of 95% ethanol were added sequentially to it. The mixture was frozen on dry ice or at –140°C for 10 minutes and then centrifuged at 13,000 rpm for 20 minutes at 4°C. The resulting RNA pellet was washed by the addition of 500 µl 70% ethanol and centrifugation at 13,000 rpm for 2 minutes at 4°C. After removal of the ethanol by pipetting the pellet was air-dried for 5 minutes and resuspended in 7 µl of RNase-free water.

*(ii) Removal of the mRNA cap structure*

The dephosphorylated RNA (7 µl) was then treated with 1µl tobacco acid pyrophosphate (0.5 U/µl in 10 mM Tris-HCl, pH 7.5; 0.1 M NaCl; 0.1 mM EDTA; 1 mM DTT; 0.01% Triton® X-100; 50% glycerol) to remove the 5′ cap structure from intact, full-length mRNA. The reaction mixture also

included 1 µl 10XTAP Buffer (0.5 M sodium acetate, pH 6.0; 10 mM EDTA; 1% β-mercaptoethanol; 0.1% Triton® X-100) and 1 µl RNaseOut™. Incubation took place at 37°C in a water bath for an hour.

*(iii) Ligation to the RNA Oligo*

The now decapped RNA was precipitated as described before and the resuspended RNA pellet (7 µl) was transferred to a microcentrifuge tube containing the pre-aliquoted, lyophilized GeneRacer™ RNA Oligo (0.25 µg, provided in the kit), which was resuspended by pipetting up and down several times. The tube was placed at 65°C in a heat block for five minutes and then chilled on ice for 2 minutes. The ligation reaction mixture was then set up by adding 1 µl of 10XLigase Buffer (330 mM Tris-Acetate, pH 7.8; 660 mM potassium acetate; 100 mM magnesium acetate; 5 mM DTT), 1 µl ATP (10 mM), 1 µl RNaseOut™ and 1 µl T4 RNA Ligase (5 U/µl in 50 mM Tris-HCl, pH 7.5; 0.1 M NaCl; 0.1 mM EDTA; 1 mM DTT; 0.1% Triton® X-100; 50% glycerol) to the tube containing the RNA sample and the GeneRacer™ RNA Oligo. Incubation took place at 37°C in a water bath for one hour. Finally, RNA was again precipitated as before and the resulting pellet was resuspended in 10 µl of RNase-free water.

*(iv) cDNA synthesis*

10 µl of the now dephosphorylated, decapped and ligated mRNA was mixed in a microcentrifuge tube with 1 µl of primer GSP2 or lacZRT2, 1 µl of dNTP mix (10 mM) and 1 µl of RNase-free water and the tube was incubated at 65°C in a heat block for 5 minutes and then chilled on ice for a minute. The reaction volume was then made up to 20 µl by the addition of 4 µl 5X First Strand Buffer (250 mM Tris-HCl, pH 8.3; 375 mM KCl; 15 mM $MgCl_2$), 1 µl DTT (0.1 M), 1 µl RNaseOut™ and 1 µl Superscript™ III Reverse

Transcriptase (200 U/µl in 20 mM Tris-HCl, pH 7.5; 100 mM NaCl; 0.1 mM EDTA; 1 mM DTT; 0.01% Nonidet P-40; 50% glycerol). The reagents were mixed by retro-pipetting and then cDNA synthesis took place at 55$^0$C for 45 minutes. The reverse transcription reaction was inactivated by heating at 70$^0$C for 15 minutes followed by the addition of 1 µl of RNase H (2 U/µl in 20 mM Tris-HCl, pH 7.5; 100 mM KCl; 10 mM $MgCl^2$; 0.1 mM EDTA; 0.1 mM DTT; 50 µg/ml BSA; 50% glycerol) to the reaction mix and incubation at 37$^0$C for 20 minutes.

*(v) Amplification of cDNA ends*

The 1$^{st}$ round PCR mix included 1 µl of cDNA as the template together with 3 µl of the GeneRacer™ 5′ Primer (10 µM-homologous to the GeneRacer™ RNA Oligo, provided in the kit), 1 µl of the reverse gene trap vector-specific primer GSP5 or lacZ-RACE-1 (12.5µM), 1 µl of dNTP mix (10mM), 1 µl $MgCl_2$ (50mM), 5 µl of 10X PCR Buffer (200 mM Tris-HCl, pH 8.4; 500 mM KCl), 0.5 µl of Platinum™ *Taq* DNA polymerase (Invitrogen-5 U/µl in 20 mM Tris-HCl, pH 8.0; 40 mM NaCl; 2 mM Sodium Phosphate; 0.1 mM EDTA; 1 mM DTT; 50% glycerol) and 37 µl of $H_2O$. The reaction took place in a 0.5 ml tube. The thermal cycling parameters employed were: 94$^0$C for 2 minutes; 94 $^0$C for 30s, 72 $^0$C for 2 minutes (X5); 94$^0$C for 30s, 70$^0$C for 2 minutes (X5); 94 $^0$C for 30s, 68 $^0$C for 30s and 72 $^0$C for 2 minutes (X20). Nested PCR was then performed using 1 µl of the 1$^{st}$ round PCR product (undiluted or diluted X50) as a DNA template and 3 µl of the GeneRacer™ 5′ Nested Primer (10 µM, provided in the kit) and 1 µl of the vector specific primer GSP5 (if the lacZ-RACE-1 primer was used for the 1$^{st}$ round PCR) or GSP7 (12.5 µM). The rest of the reaction reagents as well as the cycling parameters were identical to the ones used in the first round PCR. The resulting PCR products were analysed by agarose gel electrophoresis, and

bands over 300 bp were excised, transferred to a S.N.A.P™ column placed on a sterile microcentrifuge tube and centrifuged at 13,000 rpm for 1 minute at room temperature. The purified PCR products were then either cloned using the TOPO TA Cloning® Kit (see Section 2.1.3.2) or stored at –20°C overnight.

*(vi) Primer sequences (5'to 3')*

**1) GeneRacer™ RNA Oligo:**

CGACUGGAGCACGAGGACACUGACAUGGACUGAAGGAGUAGAAA

**2) GeneRacer™ 5′ Primer:**

CGACTGGAGCACGAGGACACTGA

**3) GeneRacer™ 5′ Nested Primer:**

GGACACTGACATGGACTGAAGGAGTA

**4) GSP2:**

CGCGCTTCTCGTTGGGGTCTTTGCTC

**5) GSP5:**

GGACGTAGCCTTCGGGCATGGCGGAC

**6) GSP7:**

GGCCAGGGCACGGGCAGCTTGCCGGT

### 2.1.4.2 3' Rapid Amplification of cDNA ends-PCR (3' RACE-PCR)

3′ RACE-PCR was also performed using the GeneRacer™ kit according to the manufacturer's instructions. An overview of the strategy employed is shown in Figure 2.4. In brief, polyA+ mRNA was reverse transcribed at 50°C as described above using the GeneRacer™ Oligo dT primer (provided with the kit). Two rounds of PCR were then performed

using the same reagents and conditions as in 5'RACE apart from the DNA template and the sets of primers. The first round employed the cDNA synthesized with the GeneRacer™ Oligo dT primer as a template and primers NeoA (12.5 µM-specific to the gene trap vector) and GeneRacer™ 3' Primer (10 µM-homologous to the GeneRacer™ Oligo dT primer, provided in the kit). For the second round PCR the nested primers NeoB (12.5 µM, vector-specific) and GeneRacer™ 3' Nested Primer (10 µM, provided in the kit) were used in combination with 1 µl of undiluted 1st round PCR product. The PCR cycling parameters employed were identical to the ones used for 5'RACE-PCR. The PCR products were visualized by agarose electrophoresis and prepared for sequencing using the PCR Product Pre-Sequencing kit (USB Corporation) according to the manufacturer's instructions. In brief, 5 µl of PCR amplification mixture was mixed with 1 µl of Exonuclease I (10 U/µl) and 1 µl of Shrimp Alkaline Phosphatase (2 U/µl) and incubated at 37°C for 15 minutes followed by incubation at 80°C for 15 minutes. Incubations were performed using a thermocycler.

*Primer sequences (5'to 3')*

**1) GeneRacer™ Oligo dT Primer:**

GCTGTCAACGATACGCTACGTAACGGCATGACAGTG$(T)_{24}$

**2) NeoA:**

CAATGCGGCGGCTGCATACGCTTGAT

**3) NeoB:**

AAGCCGGTCTTGTCGATCAGGATGAT

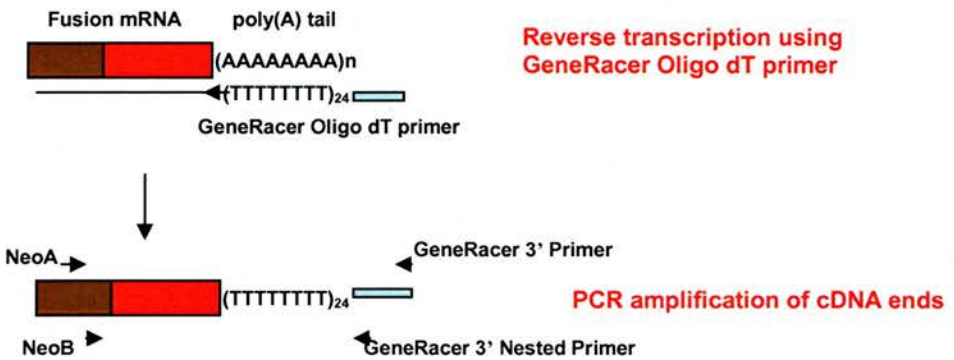



**Figure 2.4** Overview of the 3'RACE PCR strategy employed. Red box, trapped gene's sequence; Brown box, gene trap vector's neomycin resistance gene sequence; Light blue box, GeneRacer oligo dT primer.

**4) GeneRacer™ 3′ Primer**

GCTGTCAACGATACGCTACGTAACG

**5) GeneRacer™ 3′ Nested Primer**

CGCTACGTAACGGCATGACAGTG

### 2.1.4.3 Reverse Transcriptase PCR (RT-PCR)

cDNA synthesis was performed as described above using either primer GSP2 or GeneRacer™ Oligo dT primer depending on the sample. To 2 µl of the first strand cDNA reaction was added: 5 µl 10X PCR buffer (200 mM Tris-HCl, pH 8.4; 500 mM KCl); 1.5 µl $MgCl_2$ (50mM); 1 µl primer X (12.5µM): 1 µl primer Y (12.5 µM); 1 µl dNTP mix (10 mM); 0.5 µl Platinum™ *Taq* DNA polymerase (Invitrogen-5 U/µl in 20 mM Tris-HCl, pH 8.0; 40 mM NaCl; 2 mM Sodium Phosphate; 0.1 mM EDTA; 1 mM DTT; 50% glycerol) with water added to 50 µl final volume. PCR amplification was carried out using one cycle of 94$^0$C for 2 minutes and 30-35 cycles of 94$^0$C for 30 s; 55$^0$C or 58$^0$C for 30 s; 72$^0$C for 2 minutes. All sequences of the primers used are given in Appendix 2.

## 2.1.5 Sequencing

Purified 5′RACE PCR and RT PCR products were TOPO-TA cloned and then the ligations were used for subsequent transformation of TOP10 chemically competent cells as described in Section 2.1.3.2. Plasmid DNA was isolated from the transformants (about 10 clones analysed per PCR product, see Sections 2.1.3.2 and 2.1.3.3) and 5 µl were used as a template for the sequencing reaction together with 1 µl of primers T3 or T7 (0.1 µg/µl in TE buffer, pH 8). 3′RACE PCR products were directly sequenced without TA cloning and after treatment with Exonuclease I and Shrimp Alkaline

Phosphatase (described in Section 2.1.4.2). Again 5 µl of the final PCR mixture (after the second, nested 3'RACE PCR round) were used as a template for the sequencing reaction together with 1 µl of primers T2 or SDEX. Plasmid pEHygro2neoSD2 was sequenced in detail using 1 µl of plasmid maxi-prep as a template and a variety of different primers listed in Appendix 2. The sequencing reactions were carried out using the BigDye Terminator Cycle Sequencing ready Reaction Kit (Perkin Elmer) either by the I.C.A.P.B. Sequencing Facility Core (Ashworth Laboratories, University of Edinburgh) or by the sequencing service in MRC, Human Genetics Unit.

*Primer sequences (5'to 3')*

**1) T3 primer:**

ATTAACCCTCACTAAAGGGA

**2) T7 primer:**

TAATACGACTCACTATAGGG

**3) T2 primer:**

GCAGTGCAAATCCGTCGGCATCCA

**4) SDEX primer:**

CTTTGGTCCCGGATCCTGAGAACTTCA

### 2.1.6 Southern Blotting

*(i) Probe preparation*

20 µg of plasmid pd1EGFP-N1 (Clonetech) were doubly digested using 200 units of restriction endonuclease HindIII (Roche) and 100 of restriction endonuclease NcoI (Roche) in the presence of 1xSure/Cut buffer H (from a 10x stock-500 mM Tris-HCl, 1 M NaCI, 100 mM $MgCl_2$, 10 mM

Dithioerythritol, pH 7.5 at 37°C) and by incubating at 37°C overnight to release a 716 bp eGFP fragment. The latter was gel-purified after agarose gel electrophoresis using the Qiaquick® Gel Extraction kit (Qiagen) following the manufacturer's instructions and DNA quantification was carried out by UV spectrophotometry at 260 nm and agarose gel electrophoresis. Approximately 50-100 ng of probe (in 11 µl of $H_2O$) were heated at 100°C for 5 minutes and then mixed with 4 µl of High Prime (Roche) and 5 µl of redivue $^{32}P$-, $^{33}P$-Nucleotides (Amersham Biosciences) on ice. The mixture was incubated at 37°C for 1 hour. Removal of unincorporated radioactive precursors was performed by column chromatography using Sephadex G-50 (see Current Protocols in Molecular Biology, Vol 1). The now purified radioactive probe was incubated at 100°C for 5 minutes.

*(ii) DNA Transfer*

The agarose gel with the digested genomic DNA samples (see Section 2.1.3.3.3) was placed in a clean glass dish containing 250 ml of 0.25 M HCl and left shaking for 15 minutes. HCl was then replaced by 250 ml of denaturation solution (0.5 M NaOH, 1.5 M NaCl) and the gel was left shaking in it for 45 minutes. Finally the gel was transferred into 250 ml of 10xSSC and shaken for a further 15 minutes. DNA transfer onto a Hybond-N+ membrane (Amersham Biosciences) was carried out overnight by upward capillary action using the Whatman 3MM filter paper wick method (see Current Protocols in Molecular Biology, Vol. 1). The next day the membrane was washed briefly in 2xSSC and air-dried at room temperature after which it was wrapped in cling film and exposed to UV light on a UV transilluminator (Ultra Violet Products) for 2 minutes to immobilise the DNA by UV crosslinking.

*(iii) Hybridisation*

The membrane was rolled and placed in a prewarmed hybridisation cylinder glass bottle (DNA side facing inwards) and 10 ml of ExpressHyb (Clontech) solution (pre-warmed at 65°C) were added. The bottle was left rolling at 65°C in an oven for 30 minutes and then the purified radioactive probe was added into it. Hybridisation took place overnight with rolling at 65°C.

*(iv) Washes*

After hybridization and removal of the ExpressHyb solution the hybridization bottle was rinsed with 50 ml of 2XSSC/1% SDS solution which was subsequently replaced with prewarmed 2XSSC/1%SDS and the blot was left rolling for 15 minutes at 65°C. This wash was repeated and followed by two washes with 0.2XSSC/1%SDS for 30 minutes at 65°C. Finally the membrane was wrapped in cling film, placed in a cassette with X-ray film (Kodak) and left to expose for 2 days at -80°C. The film was developed using an automatic processor (SRX-101A, Konica).

## 2.2 Bioinformatics

DNA sequence traces were handled using the Chromas Lite software (Technelysium Pty Ltd.). RACE tags were determined by aligning 5′ and 3′ RACE PCR sequences against the respective gene trap vector sequences using the bl2seq program (NCBI, http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi). Vector integrations were determined by performing BLAST analysis of the RACE tags using the NCBI (blastn, http://www.ncbi.nlm.nih.gov/BLAST/), Ensembl Mouse Blast View v.36 Dec 2005 (http://www.ensembl.org/Mus_musculus/blastview) and UCSC (http://genome.ucsc.edu/index.html?org=Mouse) databases. Analysis

of sequences and data mining also employed the International Gene Trap Consortium (IGTC) databases (http://www.genetrap.org/), databases from other gene trap groups (accessed through the IGTC website), Expasy Proteomics Server (http://ca.expasy.org/) and the JavaScript DNA Translator 1.1 (http://www.bioinformatics.vg/bioinformatics_tools/JVT.shtml). Primers were designed using the Oligonucleotide Properties Calculator software (http://www.basic.northwestern.edu/biotools/oligocalc.html). Restriction enzyme target sites within DNA sequences were determined using Restriction Mapper version 3 (http://www.restrictionmapper.org/).

## 2.3 ES Cell Culture and Manipulation

The feeder-independent mouse ES cell line E14 derived from the inbred mouse strain 129/Ola (Hooper et al., 1987) was used in all experiments. All ES cell manipulations were performed in laminar flow sterile hoods under strictly sterile conditions that included wiping of all surfaces and spraying of all items entering the hoods with 70% ethanol. ES cells were incubated at 5% $CO_2$ at 37$^0$C in a humidified incubator (Galaxy S, Wolf Laboratories). Cells were maintained in ES cell medium consisting of 1X Glasgow Minimal Essential Medium (Gibco) supplemented with 20% fetal calf serum; 0.25% sodium bicarbonate (Gibco); 0.1% non-essential aminoacids (Gibco); 2mM L-glutamine (Gibco); 1mM sodium pyruvate (Gibco); 0.1 mM β-mercaptoethanol (Sigma); 100 U/ml Leukaemia Inhibitory Factor (LIF). LIF was obtained as culture supernatant from COS-7 cells transiently transfected with a murine LIF expression plasmid (pDR10) (Smith, 1991). Serial dilutions of the supernatant were tested on CP1 ES cells (Bradley et al., 1984) and 100x the minimum concentration required to keep the cells undifferentiated was generally used as the working concentration.

All flasks and plates were gelatinized (5 minutes with 0.1% gelatin in PBS) prior to addition of ES cells. All stock solutions were prepared by Aileen Leask at the JHBL.

### 2.3.1 Thawing ES Cells

Frozen ES cell cryovials were thawed at 37$^0$C and the cell suspension was transferred using a plugged pasteur pipette into a 15 ml tube containing pre-warmed ES cell medium to a final volume of 10 ml. Cell were then centrifuged at 1200 rpm for 3 minutes (Mistral 1000 centrifuge, MSE) and the resulting cell pellet was resuspended in 10 ml of ES cell medium. Cells were finally transferred into a 25cm$^2$ flask and fresh medium was added after 24 hours to remove cell debris and traces of freezing medium.

### 2.3.2 Passage and Expansion of ES Cells

Cells were normally passaged every two days. Culture medium was aspirated off the culture flask and the cell monolayer was washed once by adding 5 ml PBS. Removal of PBS was then followed by the addition of 1 ml of trypsin solution (1% trypsin; 1% chick serum; EDTA in PBS) and cells were placed at 37$^0$C for 5 minutes. 7 ml of ES cell medium was added to the resulting single cell suspension and cells were centrifuged at 1000 rpm for 5 minutes. The medium/trypsin were then aspirated and the cell pellet was resuspended in 8 ml of fresh culture media. The cell number was determined using a haemocytometer and 1X10$^6$ cells in 8 ml of ES cell medium were transferred into a new gelatinized 25cm$^2$ flask.

### 2.3.3 Freezing ES Cells

ES cells were frozen at 1 cryovial per 25cm$^2$. Cells were trypsinised as described above and centrifuged at 1000 rpm for 5 minutes. The resulting cell pellet was then resuspended in 2 ml of freezing medium consisting of 10% dimethyl sulphoxide (DMSO-Sigma) in ES cell medium, transferred into a cryovial and placed at -80$^0$C and finally at -140$^0$C for long term storage.

### 2.3.4 Generation of gene trap clones

ES cells grown in 225 cm$^2$ flasks were harvested and after centrifugation the resulting pellet was resuspended in 20 ml of PBS. Cells were then counted using a haemocytometer, centrifuged and pre-chilled PBS was added to give a concentration of 10$^7$cells/ ml PBS. 0.8 ml (i.e. 8,000,000 cells) of the cell suspension were transferred into an electroporation cuvette (0.4 cm, Bio-Rad) where they were mixed with 8 µg of the vector plasmid DNA to be electroporated and left on ice for 5 minutes. The cuvettes were then placed on the GenePulser™ (Bio-Rad) and electroporation was performed by applying a single pulse at 250 V and capacity of 500 µF aiming for a time constant between 7 and 10. After electroporation the cuvettes were placed onto ice and after 10 minutes cells from each cuvette were transferred into pre-chilled 15 ml Corning tubes containing approximately 4.5 ml of ES medium and incubated on ice for a further 10 minutes. Finally, cells were plated on gelatinized 10 cm$^2$ plates (the contents of one 15 ml Corning tube split into two plates) containing 8 ml of pre-warmed ES medium and placed into the incubator. The next day the medium was replaced by fresh medium into which the appropriate antibiotic was added (hygromycin B obtained from Roche at a concentration of 150 µg/ml or G418 obtained from Gibco at

180 μg/ml). Antibiotic selection lasted for approximately 5-7 days for hygromycin and 8-10 days for G418) after which individual resistant colonies were picked (after replacing ES medium with PBS) using a yellow Gilson tip and transferred onto gelatinized 96-well or 24-well plates. Once the clones reached confluency they were replica plated in 24-well plates: one set of plates was frozen to serve as a backup (the ES medium was replaced by 200 μl of freezing medium and plates were sealed and transferred first at -20$^0$C and eventually at -80$^0$C), one set was employed for FACS analysis of eGFP expression, a third set was used for analysis of *lacZ* (β-galactosidase) expression by X-gal staining, and another set was further expanded until cells were consequently plated on 6-well plates or 25 cm$^2$ flasks that were utilised for RNA isolation (see Section 2.1.3.5).

### 2.3.5 Analysis of lacZ expression by X-gal staining

β-galactosidase expression was detected (both in the presence of LIF and after RA induction) by staining with X-gal (5-bromo-4-chloro-3-indonyl-β-D-galactopyranoside-Sigma) in 24-well plates. Briefly, cells were washed twice in PBS and then fixed in 0.2 % gluteraldehyde, 5 mM EGTA, 2 mM $MgCl_2$, and 100 mM $Na_2HPO_4$ for 5 minutes. After fixation, cells were washed 3 times for 5 minutes each in wash buffer (2 mM $MgCl_2$, 0.02% NP-40 and 100mM $Na_2HPO_4$). X-gal staining solution contained 5 mM potassium ferricyanide, 5 mM potassium ferrocyanide, 2mM $MgCl_2$, 0.02% NP-40, and 1 mg/ml X-gal. Stocks were made in dimethyl formamide (DMF) and stored in the dark at –20$^0$C. X-gal staining was carried out overnight at 37$^0$C and then staining solution was replaced with wash buffer and plates were kept at 4$^0$C.

## 2.4 Analysis of eGFP expression by flow cytometry

All flow cytometry analyses were performed by Kay Samuel (John Hughes Bennett Lab). Briefly, cells were harvested, washed twice in PBS (supplemented with 0.1% BSA and 0.1% sodium azide), counted and finally resuspended at 2 x $10^6$ cells/ ml for use. Where possible data for 40,000 cells was acquired using a FACSCaliber bench top cytometer equipped with a 488 nm laser (Becton Dickinson) without compensation. eGFP expression was detected in fluorescence 1 channel (FL1). Data was also acquired in the fluorescence 2 channel (FL2) to allow detection of changes in autofluorescence. Samples were only considered positive for eGFP expression if there was an increase in fluorescence in FL1 over that detected in controls and no change in FL2. Data was analysed using Cellquest software (Becton Dickinson). Controls included were mock transfected cells and non-electroporated wild-type cells. Relative eGFP expression between positive samples was quantified by comparison of peak channel valves derived from histogram plots. An electronic gate was drawn around eGFP positive events on the FL1 / FL2 dot plot. Data for these events was imported into a FL1 histogram plot to generate peak channel value data (Figure 2.5).

## 2.5 Microscopy

Cells were routinely examined with a Leitz Labovert light microscope. For analysis of eGFP expression cells from the gene trap clone to be examined were normally plated at a low density on a gelatinized 10 $cm^2$ plate and once distinct colonies appeared the ES medium was replaced by PBS and fluorescence was visualized using the Zeiss Axiovert 25 fluorescence microscope with UV light through Zeiss filter set 44. Cells were photographed with an AxioCam digital camera using AxioVision 3.1 software.

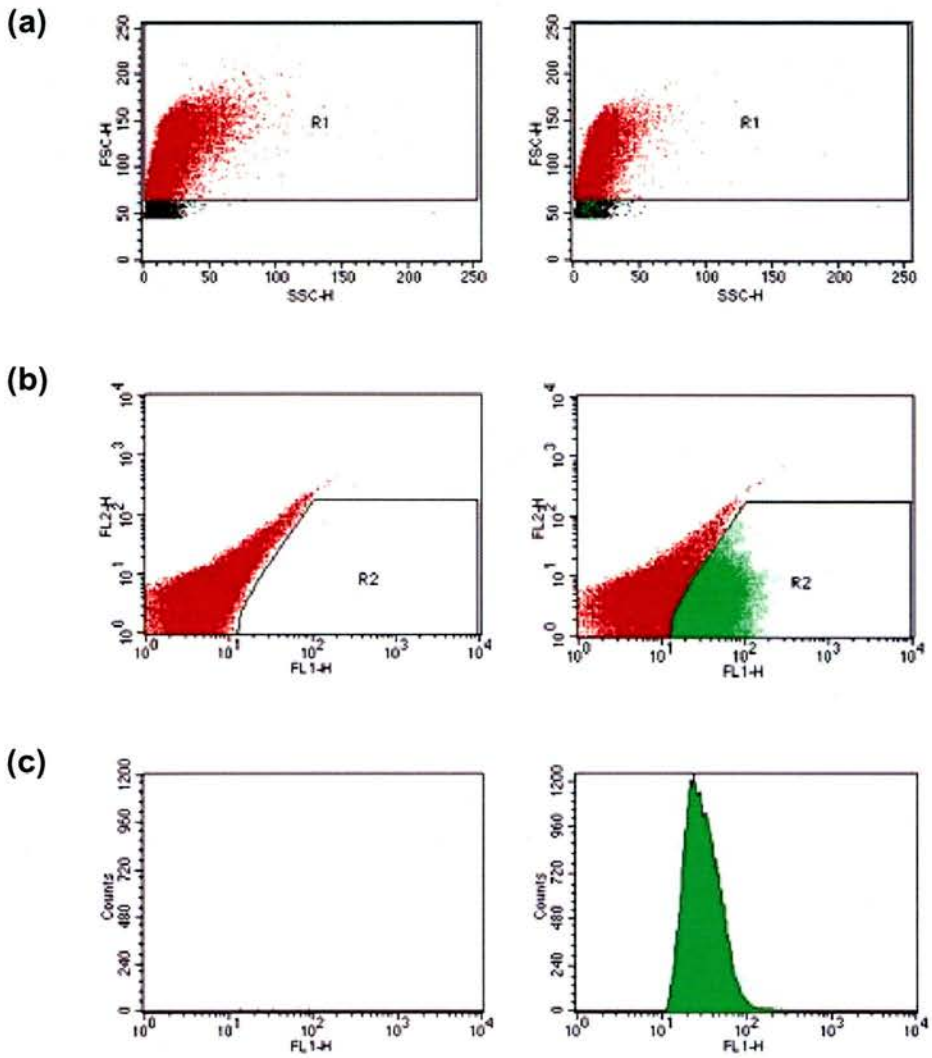



**Figure 2.5** Analysis of eGFP expression by flow cytometry. **(a)** Dot plot diagram depicting the forward and side scatter profiles used to exclude dead cells from acquisition obtained by the employment of an electronic gate. **(b)** Dot plot representation of flow cytometry data showing the electronic gate around eGFP positive events. Data from eGFP positive events was imported into histogram plots **(c)** to generate peak channel value data.

# CHAPTER 3

# RESULTS

## 3.1 Vector design and objectives

A series of plasmid gene trap vectors were constructed as tools to entrap developmentally regulated genes (Figure 3.1). Their novel features include:

- an ATG-less tripartite reporter fusion combining the *egfp*, *lacZ* and either the neomycin (*βgeo*) or hygromycin (*βhygro*) resistance genes present within a 5'SA module – a stretch of glycine (Gly) codons was inserted upstream of the *egfp* gene (Figure 3.2) in order to provide a flexible hinge between eGFP and upstream sequences and prevent steric hindrance or folding interference (Bronchain et al., 1999);
- the rabbit *β-globin* exon 2/intron 2 splice donor (SD) junction which is contained within a 3'poly(A) trap cassette that also includes the human *β-actin* promoter driving the constitutive expression of the *neo* gene – this poly(A) trap module is part of the *egfpβhygro*-containing vector;
- an ARE from the human GM-CSF gene (Xu et al., 1997) cloned into the *β-globin* intron, approximately 120 nt downstream of the SD exon/intron junction.

A series of experiments were performed aiming to characterise each of these novel components and assess their potential for use in a gene trapping context.

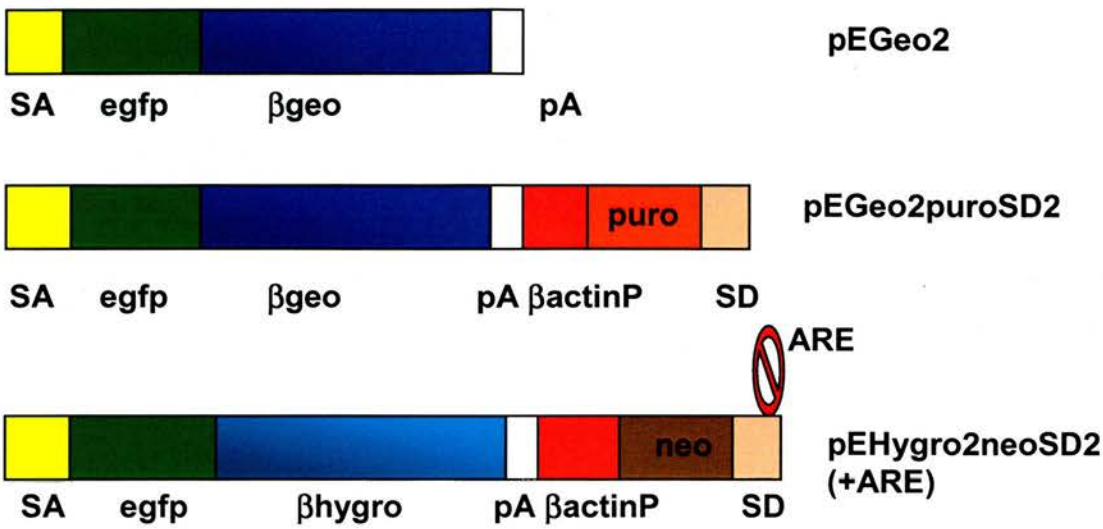



**Figure 3.1** Schematic representation of the gene trap constructs employed. *En-2*, engrailed-2; pA, polyA; SA, splice acceptor; SD, splice donor; P, promoter; ARE, AU-rich element.

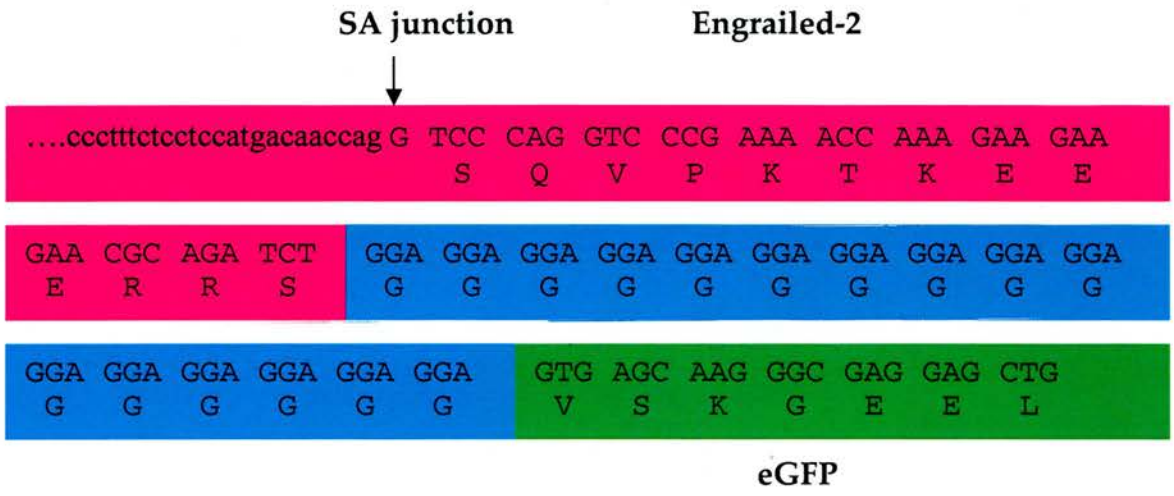



**Figure 3.2** Nucleotide and predicted amino acid sequences of the *En-2* SA-gly-egfp junction present in both the *egfpβgeo* and *egfpβhygro* triple fusions. The mouse Engrailed-2 SA is shown in pink with the intron in lower case letters and exon in uppercase letters. The glycine bridge is shown in blue and the beginning of eGFP in green. *En-2*. engrailed; SA, splice acceptor.

## 3.2 Characterisation of the triple reporter fusion

### 3.2.1 Experimental strategy

To test the function of the triple reporter fusion, linearised vectors pEGeo2, pEGeo2puroSD2 and pEHygro2neoSD2 (equivalent to vector pEHygro2neoSD2 (+ARE) shown in Figure 3.1 but without including the ARE) were introduced independently by electroporation into ES cells (Figure 3.3); the *egfpβgeo* fusion was tested in the context of gene trap vectors pEGeo2 and pEGeo2puroSD2 whereas the *egfpβhygro* fusion was characterized as the 5'component of vector pEHygro2neoSD2. After neomycin or hygromycin selection, depending on the construct used, individual gene trap clones were isolated, expanded and replicated into three sets of 24-well plates (Figure 3.3). One set was frozen, another was used for analysis of *egfp* expression by flow cytometry and fluorescence microscopy, and the last set of replica plates was X-gal stained in order to determine β-galactosidase expression levels. Some selected clones were further expanded 5'RACE-PCR analysis. The latter was employed as a means of characterising the molecular nature of the gene trap integrations.

### 3.2.2 Reporter expression analysis provides evidence that the triple fusion is functioning

Electroporations with the two different *egfp/βgeo*-containing constructs yielded similar numbers of neomycin resistant colonies per electroporation plate. 65 pEGeo2-electroporated and 159 pEGeo2puroSD2-electroporated $neo^R$ gene trap ES clones were picked and analysed for reporter expression (Table 3.1). 49 % and 39 % of the total number of $neo^R$ clones tested by X-gal staining were found *lacZ* positive for vectors pEGeo2 and pEGeo2puroSD2

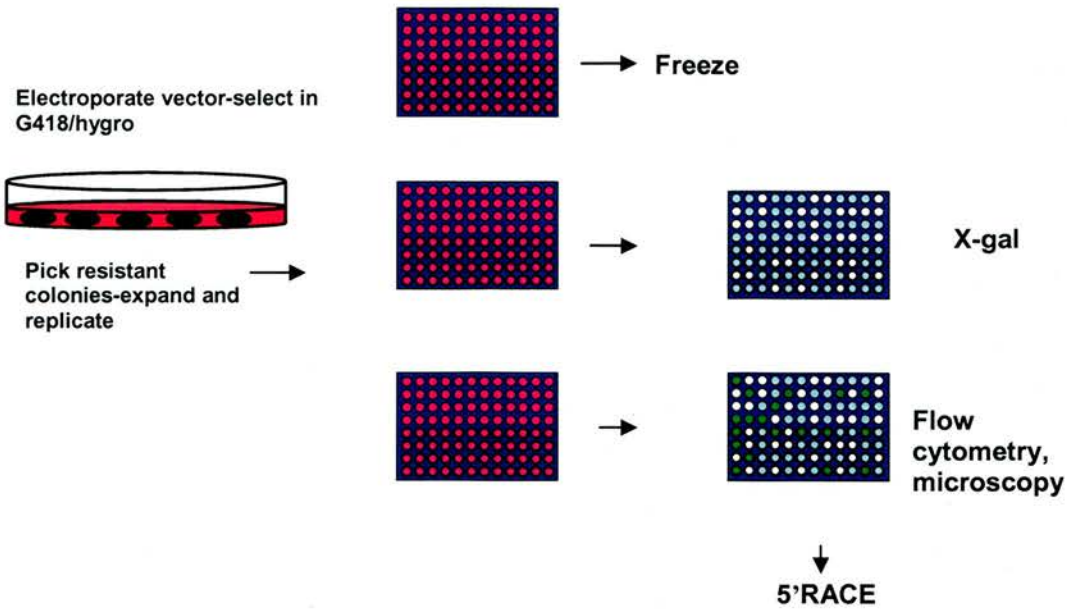



**Figure 3.3** Schematic representation of the experimental strategy adopted for the characterisation of the triple reporter fusion.

| Vector | Selection | No. clones tested | lacZ(+) | lacZ (+) eGFP (+) | lacZ (+) eGFP (-) | lacZ (-) eGFP (+) |
|---|---|---|---|---|---|---|
| pEGeo2 | Neo | 65 (100%) | 32 (49%) | 15 (23%) | 17 (26%) | - (0%) |
| pEGeo2puroSD2 | Neo | 159 (100%) | 62 (39%) | 31 (21%) | 29 (18%) | 1 (0.6%) |
| pEHygro2neoSD2 | Hygro | 32 (100%) | 18 (56%) | 10 (31%) | 8 (25%) | 1 (3%) |

**Table 3.1** Summary of the results obtained after analysis of gene trap clones for β-galactosidase (X-gal staining) and eGFP (flow cytometry) expression. Numbers in parentheses indicate percentages relative to the total number of clones analysed.

respectively. These percentages are within the range of values representing the proportion of β-galactosidase positive, G418-resistant gene trap clones that have been reported by investigators employing similar *βgeo*-containing constructs e.g. 24-37% (Bonaldo et al., 1998) or 60% (Skarnes et al., 1995). The number (n=32) of hygromycin resistant gene trap clones generated after electroporation with the pEHygro2neoSD2 construct was considerably lower probably due to the stringent nature of hygromycin selection (Table 3.1). 56% of these were found to express β-galactosidase (Table 3.1).

The resulting *lacZ* expression patterns varied both in terms of staining intensity and cellular distribution (Figures 3.4). Some of the trapped clones that we tested were characterized by restricted β-galactosidase activity which was confined to a few isolated cells (Figure 3.4a) or larger cell subpopulations (Figure 3.4b). Others showed a more widespread cellular distribution of *lacZ* expression which was often localised in speckles within most cells (Figure 3.4h), or it appeared as diffuse, intense staining present within the majority of cells (Figure 3.4e). Different, distinct patterns of subcellular localization e.g. nuclear or cytoplasmic were also observed (Figures 3.4i and k respectively). This variation in β-galactosidase activity indicates that a wide range of genes expressed at varying levels in ES cells (as reflected by the diversity of the expression patterns observed) are accessible to entrapment by our triple fusion-containing vectors.

eGFP expression analysis of the clones by flow cytometry was carried out taking into account two parameters: (i) the percentage of fluorescent cells within the clone and (ii) the fluorescence intensity of the positive cell subpopulations as estimated by peak channel value data relative to untransfected/mock-transfected controls. Based on these criteria 23% and 21.6% of the neomycin resistant clones carrying pEGeo2 and

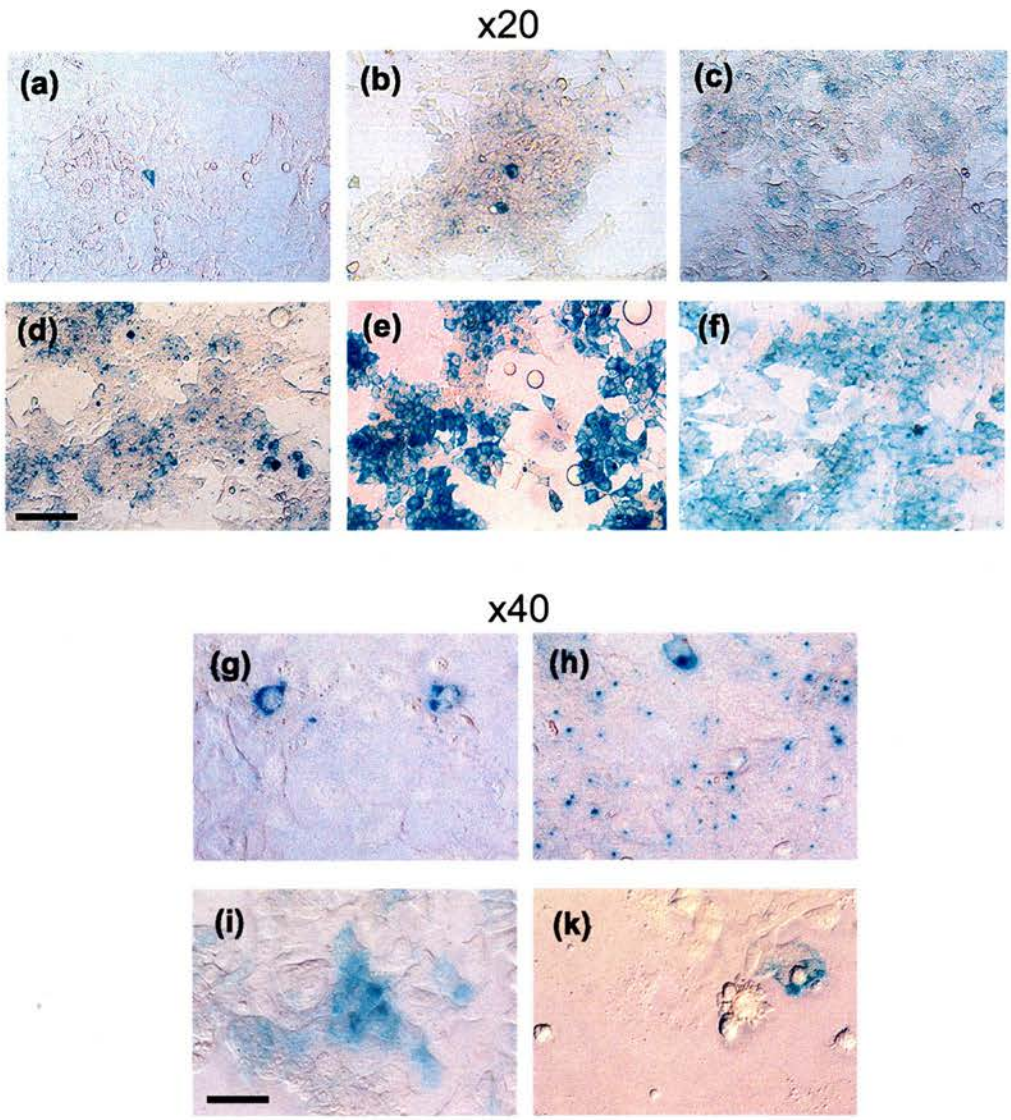



**Figure 3.4** Representative β-galactosidase expression patterns of hygromycin and neomycin resistant gene trap clones. Various different X-gal staining patterns were obtained e.g restricted (a-b), faint/diffuse (c), widespread/intense (e), spotty (h). Also note the different subcellular localisations of the lacZ reporter protein e.g. nuclear (i) or cytoplasmic (g and k). Objective x20, scale bar=80 μm (a-f) and x40, scale bar= 30 μm (g-k).

pEGeo2puroSD2 vector insertions respectively were found to be eGFP positive by flow cytometry (Table 3.1). This percentage was slightly higher (34%) for the pEHygro2neoSD2-electroporated, hygro$^R$ gene trap clones (Table 3.1). Some examples of eGFP positive clones as determined by flow cytometry are given in Figures 3.5 and 3.6.

Gene trap clones expressing eGFP were also detected by fluorescence microscopy (Figures 3.7 and 3.8). However, this was possible only in the case of clones that were found by flow cytometry to be high eGFP expressors. It was found that the higher the percentage of eGFP-expressing cells within a gene trap clone (as determined by flow cytometry) the higher the probability of detecting the presence of eGFP within the same clone by fluorescence microscopy. Moreover, we were unable to detect by microscopy less bright clones that were defined as eGFP positive based on their flow cytometric profile. This finding probably reflects a 'lowest sensitivity limit' for the eGFP detection potential of the specific microscope that was employed during this study.

### 3.2.3 Assessing the correlation between the eGFP and lacZ proteins

Approximately half of the β-galactosidase positive clones resulting from all three electroporations were found to be fluorescent (Table 3.1). To get an idea of the extent of correlation between the eGFP and β-galactosidase reporter proteins, all gene trap clones were categorized into four groups on the basis of their β-galactosidase expression profile:

- (-) group, this includes clones that showed no *lacZ* expression; (+/-) in which clones exhibited a small degree of *lacZ* expression mainly restricted in isolated cells (Figure 3.9);

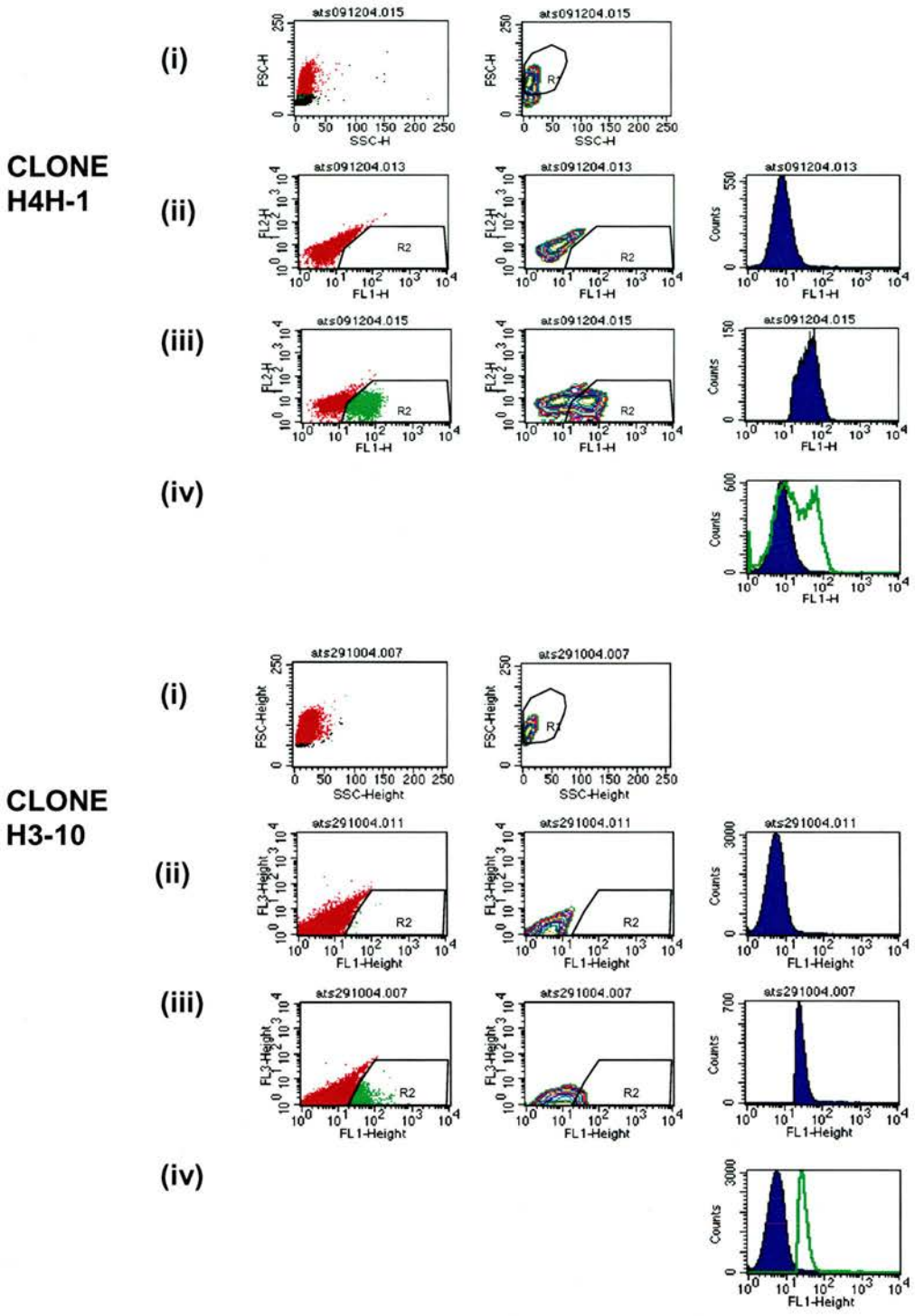



**Figure 3.5** Flow cytometry profiles of two representative eGFP expressing, hygromycin resistant clones. **(i)** Dot plot of forward and side scatter profiles of each clone used to gate "live" events (R1) before further analysis **(ii)** Flow cytometry data from analysis of E14 control sample **(iii)** Plots showing gating of fluorescent positive events (R2) **(iv)** Histogram plot showing the log increase of the peak channel log shift of eGFP expressing cells (green line) relative to the control cells (filled bar).

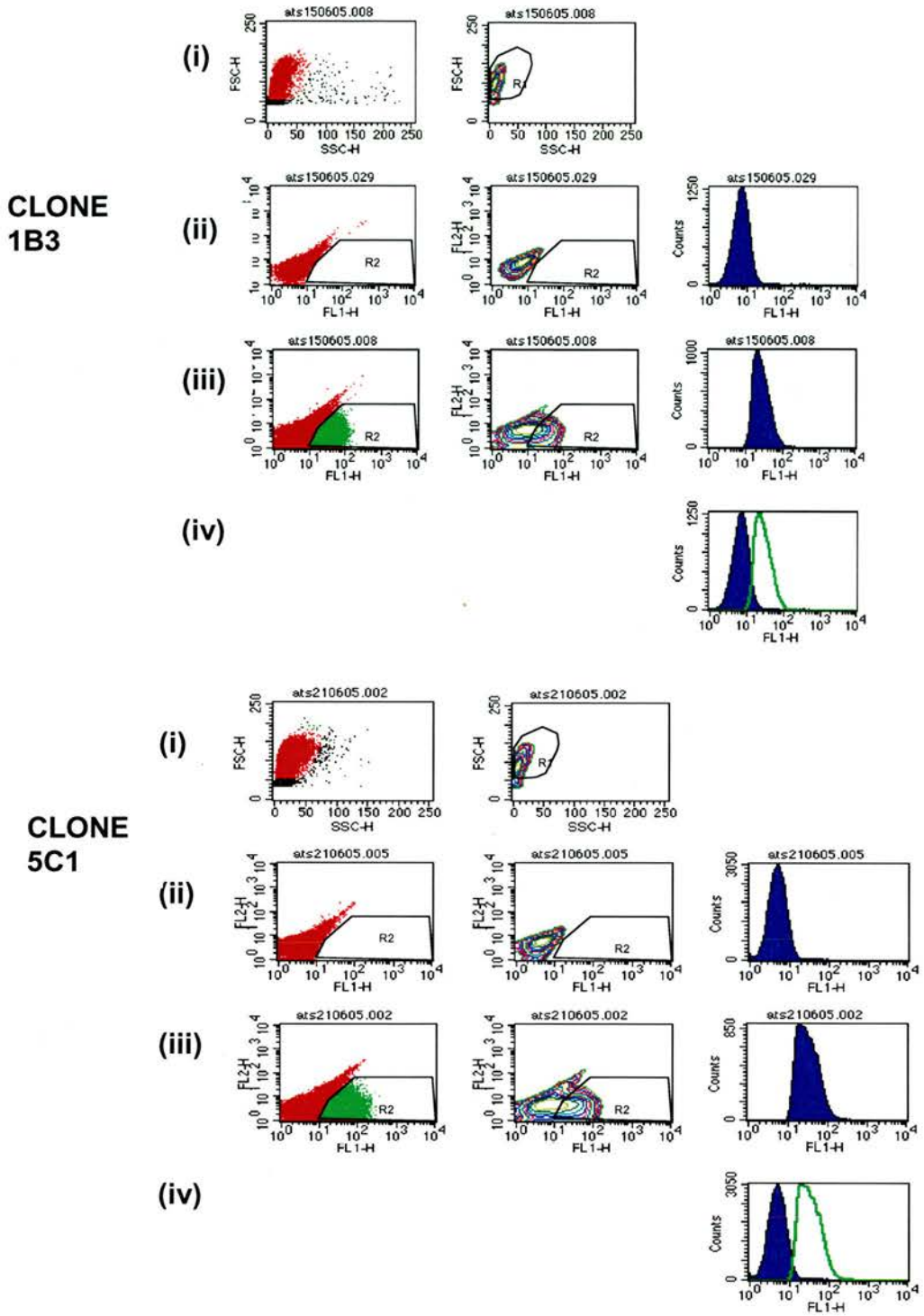



**Figure 3.6** Flow cytometry profiles of two representative eGFP expressing, neomycin resistant clones. **(i)** Dot plot of forward and side scatter profiles of each clone used to gate "live" events (R1) before further analysis **(ii)** Flow cytometry data from analysis of E14 control sample **(iii)** Plots showing gating of fluorescent positive events (R2) **(iv)** Histogram plot showing the log increase of the peak channel log shift of eGFP expressing cells (green line) relative to the control cells (filled bar).

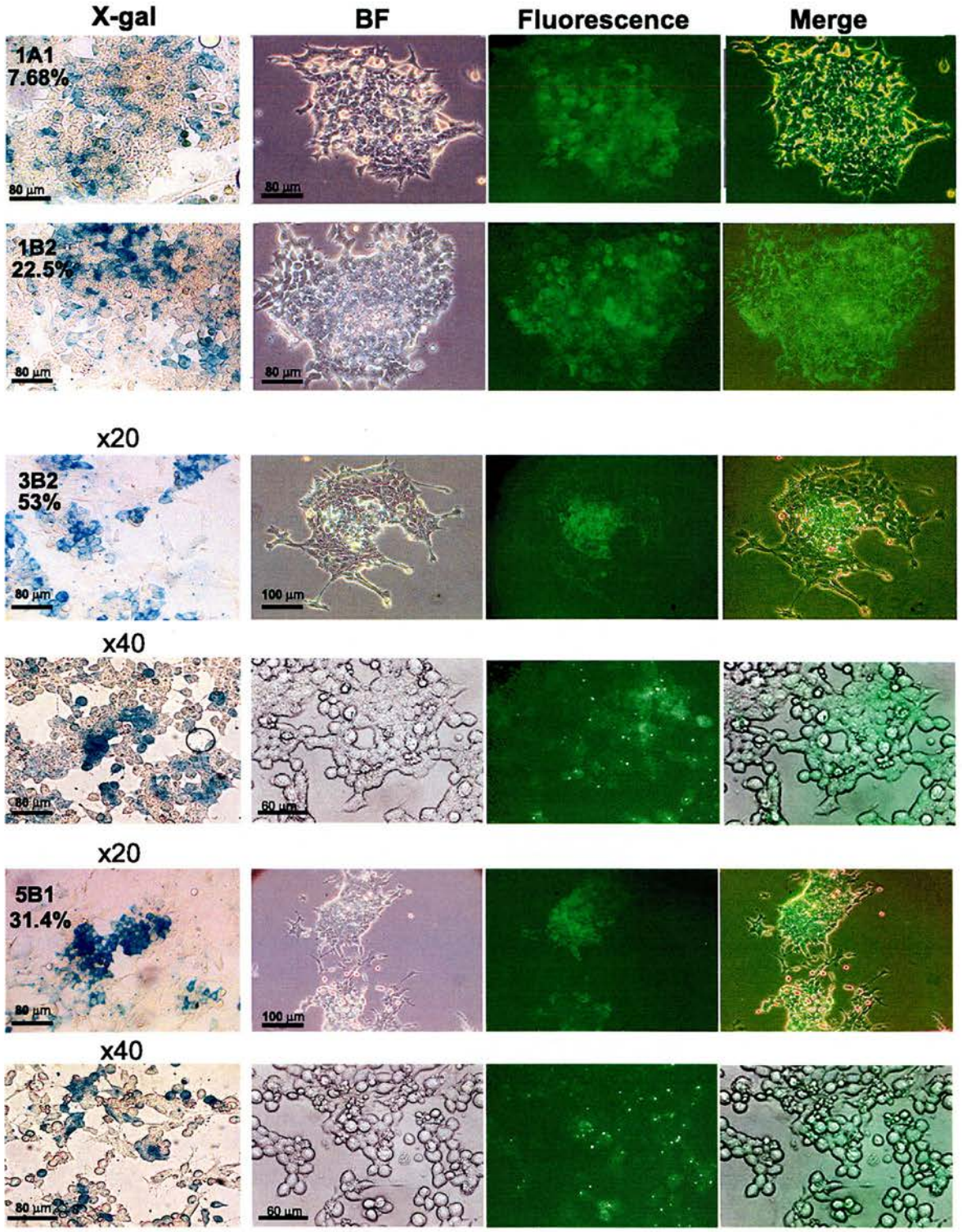



**Figure 3.7** eGFP expression of representative neomycin resistant clones. The names of the clones and their lacZ expression profiles are also shown. Percentages indicate the fraction of fluorescent cells within each clone as determined by flow cytometry. Objective x20 (clones 1A1, 1B2 and 3B2, 5B1-top panel) and x40 (3B2, 5B1-bottom panel). BF, brightfield.

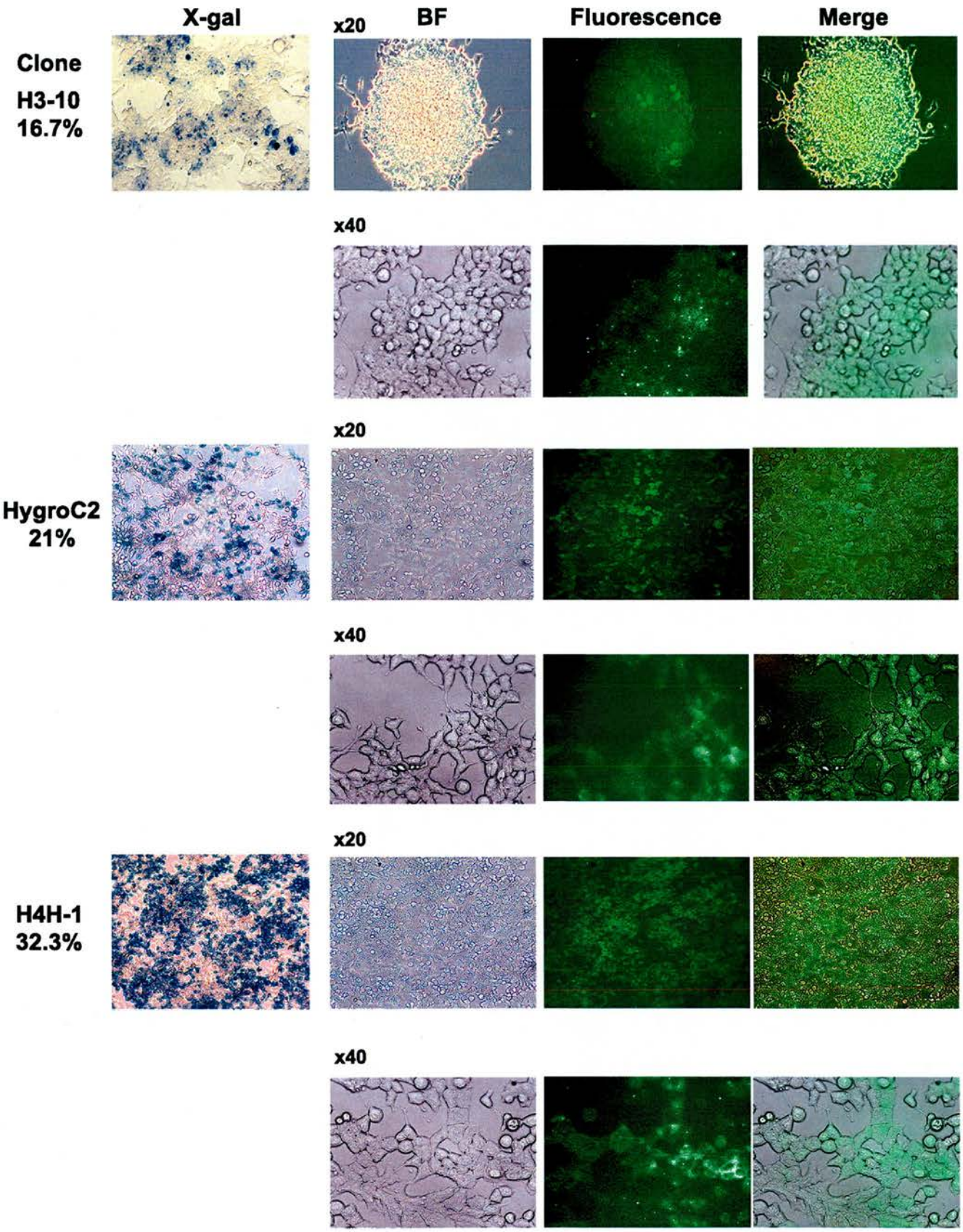



**Figure 3.8** eGFP expression of representative hygromycin resistant clones. Corresponding lacZ expression profiles are also shown. Percentages represent the fraction of fluorescent cells within each clone as determined by flow cytometry.

- (+), which consists of clones that showed widely distributed β-galactosidase (Figure 3.9);
- (++), which contains clones that are strongly β-galactosidase positive in terms of intensity and cellular distribution (Figure 3.9).

This grouping was empirical based on subjective observational criteria. The percentage of the eGFP expressing cells (as determined by flow cytometry) present within each clone was then plotted against its corresponding *lacZ* expression group (Figure 3.10).

Predictably, the clones with the highest levels of β-galactosidase activity were clones that also possessed the highest eGFP positive cell populations (Figure 3.10). This correlation is obvious for all three electroporations with the members of the (++) *lacZ* expression group containing the highest fractions of fluorescent cell populations (values up to 88.2% of the total in the case of vector pEGeo2). It should be noted that some clones belonging to the (+/-) and (+) lacZ expression groups and possessed relatively low to minimal fluorescent cell populations (0.48-5%) were classified as eGFP positive because they were found by flow cytometry to be considerably brighter than the control samples. The employment of fluorescence microscopy revealed that the expression pattern of eGFP, in general, correlates well with that of β-galactosidase both in terms of cellular distribution and intensity (Figure 3.11).

Interestingly, flow cytometry also revealed the presence of two lacZ negative clones (clones 3D4 and H3-17 from electroporations with vectors pEGeo2puroSD2 and pEHygro2neoSD2 respectively) that possessed surprisingly high populations of fluorescent cells (depicted as red data points in Figure 3.10) (Figure 3.12). The pEHygro2neoSD2-electroporated, lacZ (-),

+/- + ++

**Figure 3.9** β-galactosidase expression patterns indicative of the three *lacZ* expression (+/-, +, ++) groups used to categorize gene trap clones (objective x20).

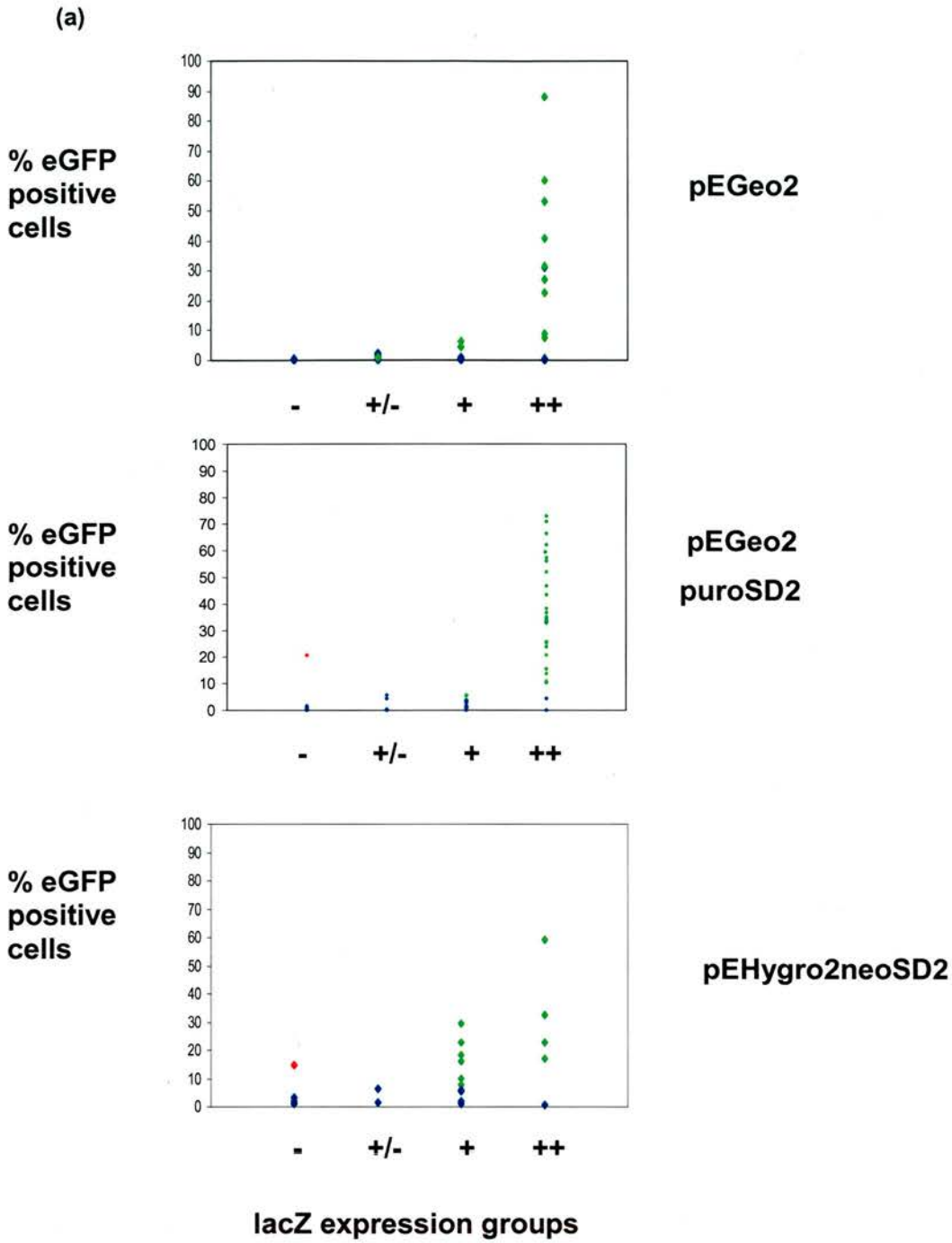



**Figure 3.10** Analysis of the degree of correlation between eGFP and β-galactosidase reporter proteins. Percentages of fluorescent cells as determined for each clone by flow cytometry were plotted against corresponding lacZ expression groups for gene trap clones resulting from electroporations with all three vector constructs. Green data points correspond to clones that are defined as eGFP positive. Red data points correspond to lacZ (-) clones that were found by flow cytometry to contain high numbers of eGFP expressing cells.

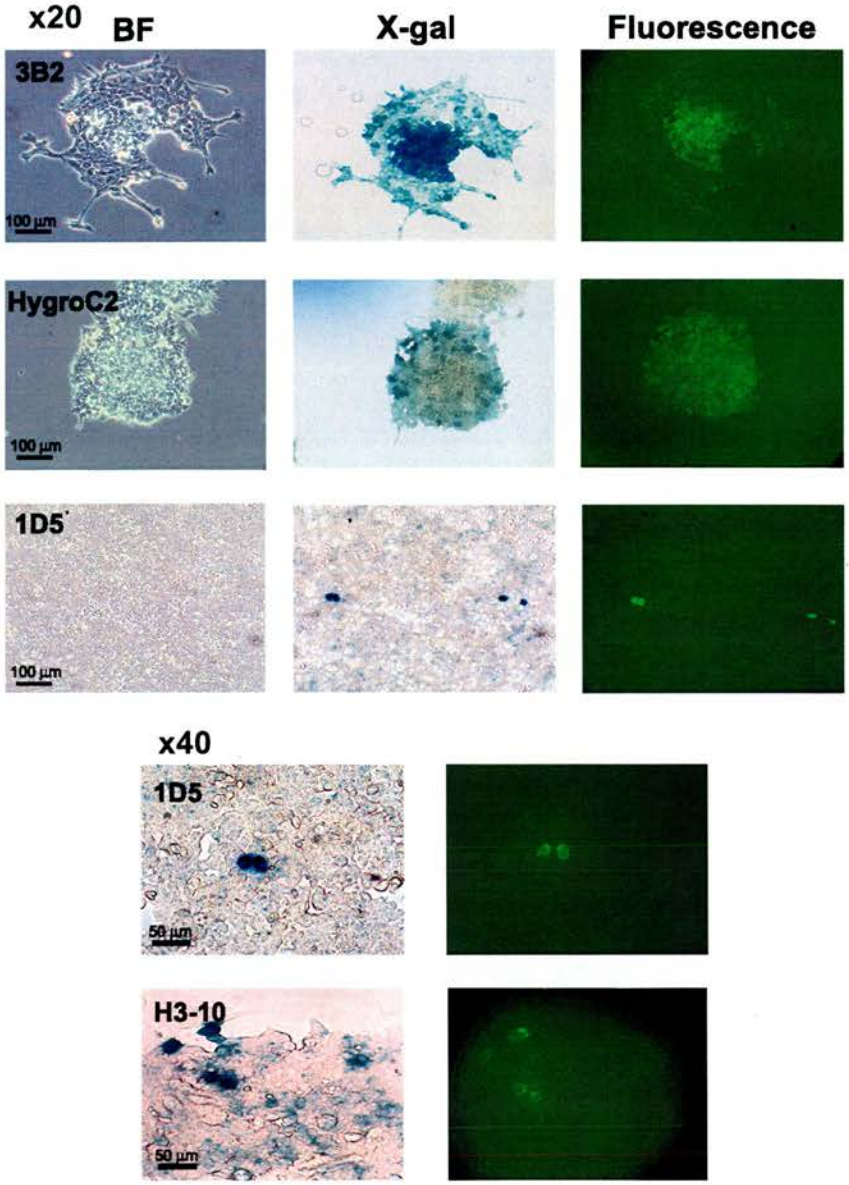



**Figure 3.11** Examples of correlation in the expression of β–galactosidase and eGFP proteins. Note the similar expression patterns of the two reporters in terms of intensity and cellular distribution.Both neomycin (3B2, 1D5) and hygromycin (HygroC2, H3-10) resistant clones are shown.BF, Brightfield.

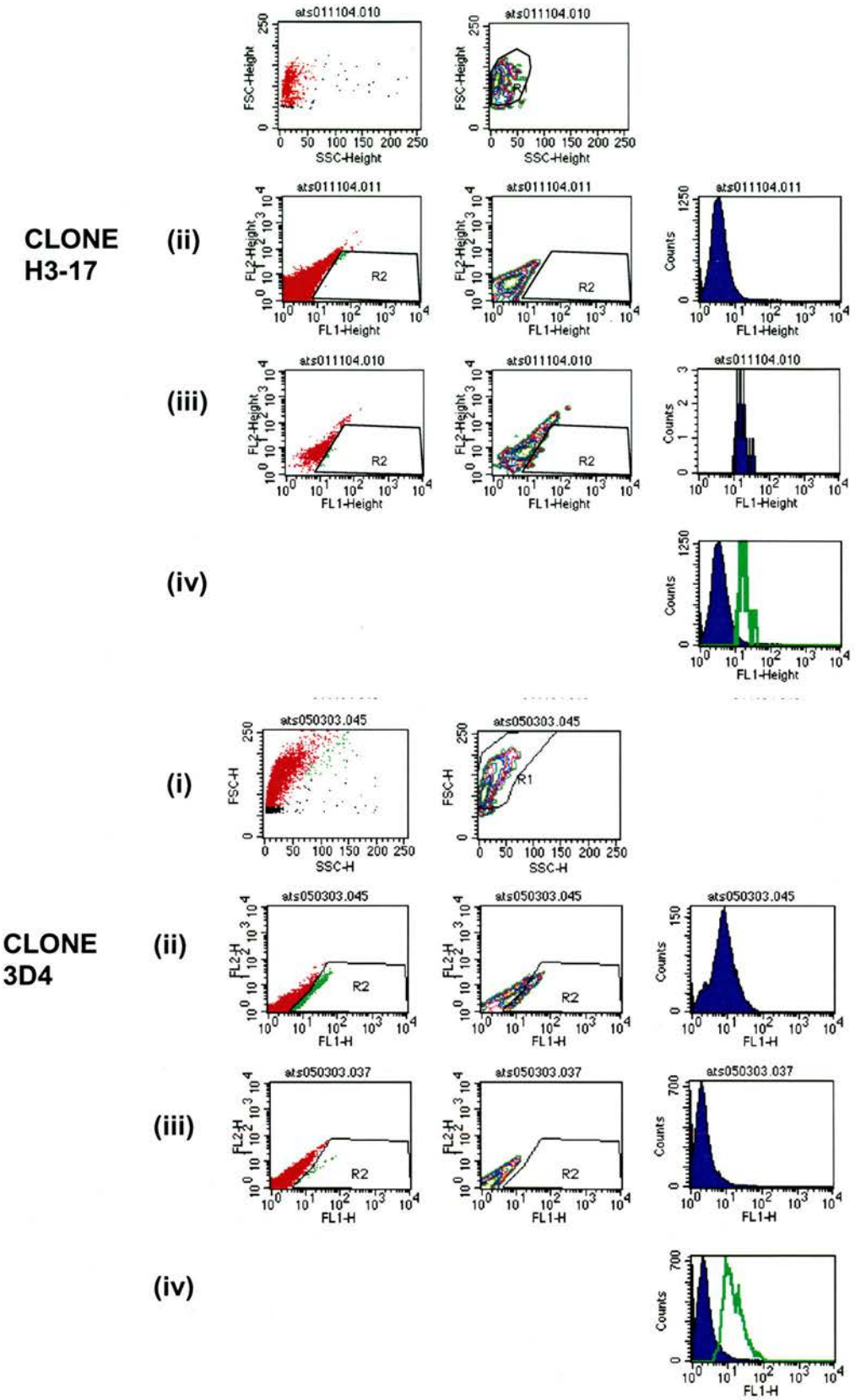



**Figure 3.12** Flow cytometry profiles of lacZ-/eGFP+ clones 3D4 and H3-17.

| RACE tag ID | Vector | Chromosome | Transcript BLAST hit |
|---|---|---|---|
| H3-10-1 | pEHygro2 neoSD2 | 18 | *Fbxo15* |
| H3-10-2 | pEHygro2 neoSD2 | 18 | *Fbxo15* (5'UTR) |
| H3-1 | pEHygro2 neoSD2 | 2 | *Dido1* (5'UTR) |
| H3-17 | pEHygro2 neoSD2 | X/Y pseudo autosomal region | *Erdr1* |
| H4H-1 | pEHygro2 neoSD2 | ? | GAPDH-related pseudogene |
| 5B1 | pEGeo2 | 4 | *Rnu17d* (non-protein coding) |
| 1A1 | pEGeo2 | 1 | *Gas5* (non-protein coding) |
| 5C1 | pEGeo2 | 5 | *Ubc* (5'UTR) |
| 1B3 | pEGeo2 | 1 | *Bat2d* (EST transcript) |
| 1A6 | pEGeo2 | 3 | EST transcript ENSMUSESTT0000002342 |
| 1B2 | pEGeo2 | 5 | Expressed sequence AI506816 (EST transcript) |

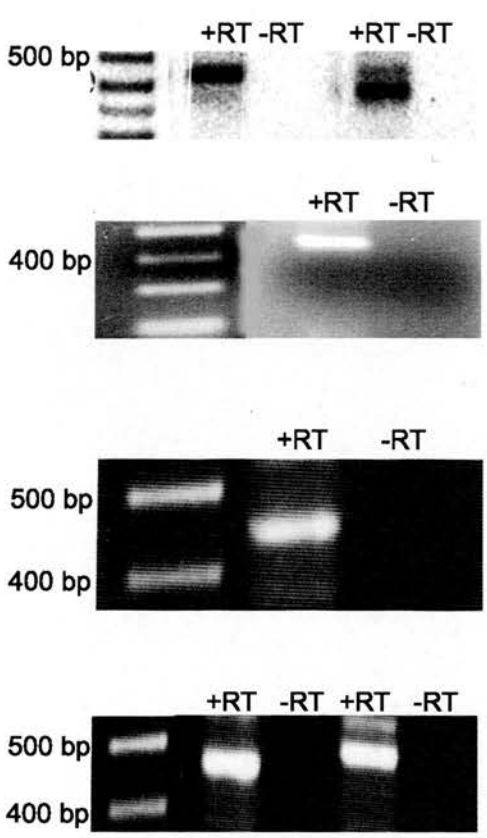
**Table 3.2** Results of BLAST analysis of the 5' RACE sequence tags generated from clones electroporated with vectors pEGeo2 and pEHygro2neoSD2.

Some selected fusion transcripts were confirmed by RT-PCR using RACE product- and vector-specific primers (Figure 3.13). In all the cases examined the *En-2* splice acceptor junction was utilized by the splicing machinery in the predicted manner to give rise to a fusion transcript consisting of endogenous trapped sequence and the vector's triple fusion coding sequence. All integrations occurred within distinct loci. The sequences and detailed analysis of all the 5′ RACE tags are given in Appendix 3.

### 3.2.4.1 Overview of integrations

#### 3.2.4.1.1 *egfpβhygro* vector integrations

5′ RACE analysis of clone H3-10 generated two different, properly spliced transcripts termed H3-10-1 and H3-10-2. Both products were found by BLAST analysis to be homologous to adjacent mouse genomic regions on chromosome 18. H3-10-1 transcript was found to be identical to the last 137 bp of the first exon (157 bp in total) of *Fbxo15* gene (Ensembl gene ID: ENSMUSG00000034391) (Figure 3.14). H3-10-2 product showed homology to a genomic area approximately 380 bp upstream of H3-10-1 that has not been reported to be part of an exon. The existence of both fusion transcripts was confirmed by RT-PCR analysis (Figure 3.13). To eliminate the possibility that H3-10-2 represents a PCR artifact, the RT-PCR product that was generated using a vector-specific *egfp* primer and a primer designed on the H3-10-2 sequence and corresponded to the expected 400 bp size (Figure 3.13) was cloned and sequenced and it was found to be identical to transcript H3-10-2.
The above data indicates the trapping of two alternatively spliced transcripts and suggests that integration of vector pEHygro2neoSD2 took place in the first intron of *Fbxo15* which consists of 10 exons in total (Tokuzawa et al., 2003) (Figure 3.12). *Fbxo15* encodes an F-box containing protein and was




| RACE transcript | Primers | Expected product |
|---|---|---|
| H3-10-1 | Fbxo-ex1/GSP5 | 542 bp |
| | Fbxo-ex2/GSP5 | 458 bp |
| H3-10-2 | H3-10/GSP5 | 400 bp |
| 5B1 | 5B1/GSP5 | 430 bp |
| 1B3 | 1B3/GSP5 | 431 bp |
| | BAT2D/GSP5 | 463 bp |

**Figure 3.13** RT PCR confirmation of some of the fusion transcripts cloned after 5' RACE PCR analysis of gene trap clones. In all cases a vector-specific (egfp) primer called GSP5 was used in conjunction with primers that were complementary to the trapped sequence.

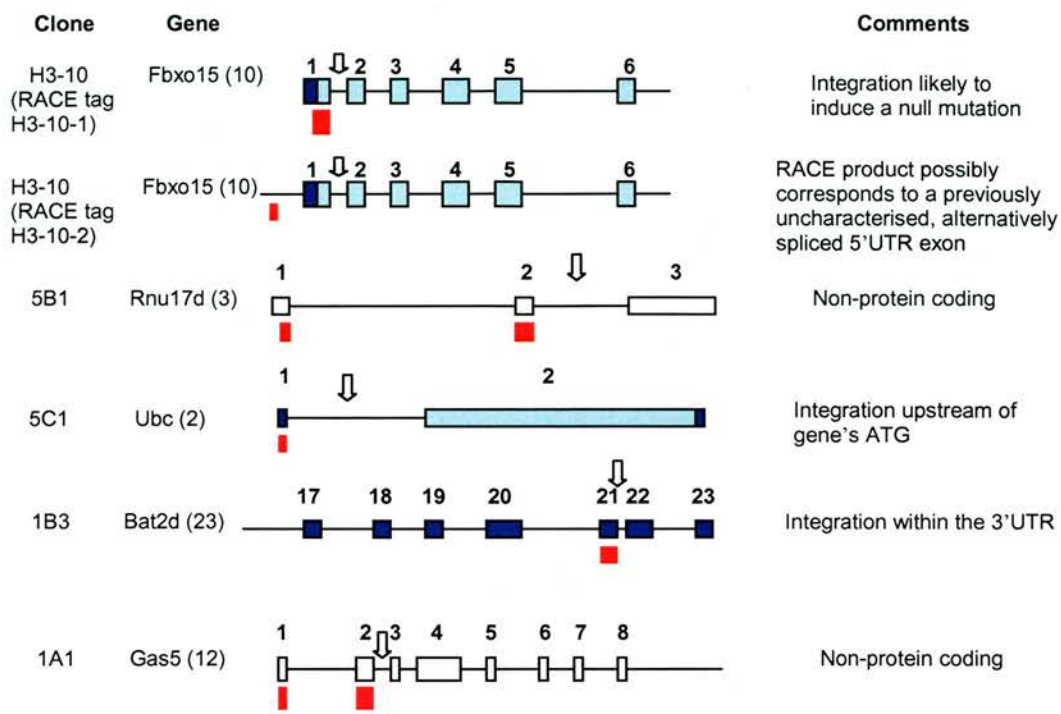


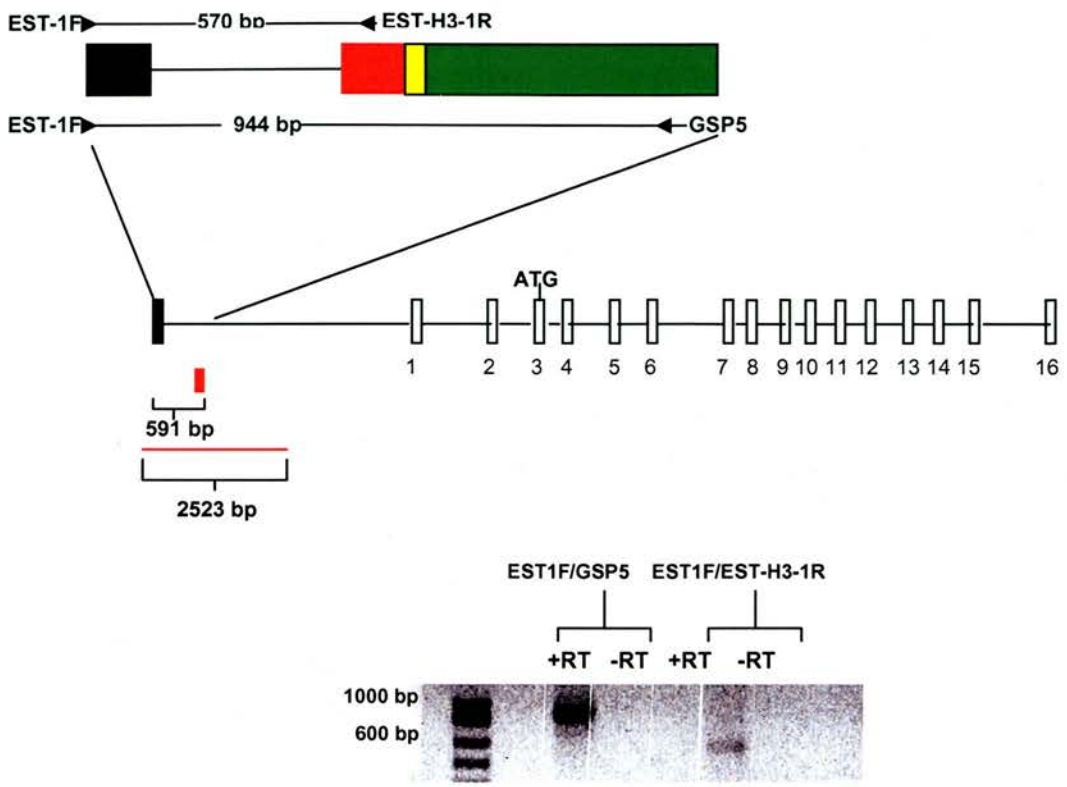

**Figure 3.14** Overview of some of the gene trap insertions predicted by 5'RACE PCR analysis of *egfpβhygro* and *egfpβgeo* triple fusion-containing gene trap clones. White arrows indicate the intron at which vector insertion occurred. Numbers in parentheses refer to the total number of reported exons present within the trapped loci. Note that the F-box domain in the case of clone H3-10 is encoded by exons 3-5. Dark blue boxes, 5' or 3' UTR exons; Light blue boxes, protein coding exons; red boxes, regions of homology to corresponding cloned RACE products. Only exons adjacent to the site of integration are shown.

found to be a target of Oct3/4 (Tokuzawa et al., 2003). The gene trap event represented by RACE product H3-10-1 is an in-frame fusion between the *egfpβhygro* reporter sequence and the *Fbxo15* coding sequence. The resulting chimaeric protein is predicted to incorporate the first 8 N-terminal aminoacids of the FBXO15 protein and is translated from the trapped gene's ATG (Appendix 3). This integration is likely to induce a null mutation in the *Fbxo15* gene due to the severity of the truncation that is introduced (Figure 3.14).

BLAST analysis of the 84 bp RACE product H3-1 indicates that vector insertion in this case took place upstream of the first 5'UTR reported exon (Futterer et al., 2005) of the *Dido* (*Death inducer-obliterator*) locus on chromosome 2. *Dido* gives rise to at least three splice variants with distinct roles in apoptosis (Garcia-Domingo et al., 1999) and tumour suppression (Futterer et al., 2005). The H3-1 transcript is homologous to a region that is adjacent to a predicted *Dido* exon (http://www.ensembl.org/Mus_musculus/exonview?db=core;transcript =ENSMUST00000037764). The RACE product's homology with two EST sequences (Accession numbers: AK051426 and AK036185) which also include the predicted *Dido* exon, the existence of identical 5'RACE tags generated by other groups (gene trap cell lines M029F04 and W068C04) and RT-PCR-based evidence (Figure 3.15a) indicate that the trapped H3-1 transcript is probably derived from the predicted exon. This finding is also supported by the fact that the sequence of the H3-1 RACE product is highly conserved between mouse and rat (3.15b).

The transcript generated from clone H3-17 was found to be highly homologous to an EST (Accession number: BC058113) that is related to a

**(a)**




**(b)**

```
Mouse  CTGGATAGCTTCGGGTAGAGCTATGGAGTTAAG AGACT
       ||   | |||||  || ||||||||||||||||   |-|||||----------
Rat    CTAAACAGCTTTGGTTAGAGCTATGGAGTT  GCAGACTTAGCGCGTACA

Mouse  GCCACTGTGGTCATTGTGTCTGACAGATCTGCTAAACTGGTCTTA
       |||||||||||| |||||||  |||||||||| |||||||  ||||
Rat    GCCACTGTGGTCGTTGTGTCCAACAGATCTGCAAAACTGGCCTTA
```

**Figure 3.15** **(a)** Genomic organisation of Dido locus and RT-PCR strategy adopted for analysis of RACE transcript H3-1. Reported *Dido* exons are shown as open boxes. The predicted *Dido* exon is depicted by a black box. The region homologous to EST AK051426 is depicted by a red line. The primer combinations employed include EST1F which was designed within the predicted *Dido* exon and vector-specific GSP5 (expected product size=944 bp) or primers EST1F and EST-H3-1R which is a RACE product-specific primer (expected product size=570 bp). Note that the RACE product is homologous to a region within the EST sequence. Yellow box, *En-2* SA; Green box, *egfp* coding sequence; Red box, RACE transcript H3-1. **(b)** The H3-1 RACE transcript sequence is conserved among mouse (chromosome 1) and rat (chromosome 3). The RACE product's sequence is shown in red. The rat sequence is shown in black letters. (Information obtained from the UCSC genome server).

recently identified gene called erythroid differentiation regulator or *erdr1* (Dormer et al., 2004). The latter has been involved in the induction of haemoglobin synthesis and it is speculated to have a more general role in cell survival and growth control (Dormer et al., 2004). The EST has an ORF that theoretically gives rise to a 145 aminoacid polypeptide. The integration event is predicted to give rise to an ATG-containing fusion protein between 51 aminoacids of the trapped gene and the vector's reporter protein (Appendix 3). However, the frame of translation of the fusion protein is different to the one employed for the generation of the endogenous protein. It should be noted that H3-17 is the only one of the clones analysed by 5'RACE PCR that was found to be lacZ(-)/eGFP (+).

Analysis of the RACE product isolated from clone H4H-1 indicates that in this case the vector landed within a GAPDH-related pseudogene. BLAST analysis of transcript H4H-1 yielded multiple high homology hits that corresponded to ESTs and genomic DNA regions located at various different chromosomes throughout the mouse genome. All hits were related to a sequence that is present in pseudogenes belonging to the glyceraldehyde 3-phosphate (GAPDH) multigene family. The latter includes a single functional gene that encodes the glycolysis enzyme GAPDH and more than 300 retroprocessed pseudogenes which are dispersed throughout the mouse genome (Garcia-Meunier et al., 1993).

#### 3.2.4.1.2 *egfpβgeo* vector integrations

The neomycin resistant clone 5B1 carries an integration within the *Rnu17d* locus, which is located on chromosome 4. *Rnu17d* is a non-protein coding gene that consists of three exons. Its transcript gives rise to intron-located small nucleolar RNAs (snoRNAs) of the H/ACA-box class (Pelczar and Filipowicz, 1998). BLAST analysis revealed that vector insertion took

place within the second intron of *Rnu17d* (Figure 3.14). The generation of a fusion transcript incorporating *Rnu17d* and vector reporter sequences was confirmed by RT PCR (Figure 3.13). Similarly, clone 1A1 was found to contain a vector insertion within another snoRNA host gene called *gas5* (Smith and Steitz, 1998). Like Rnu17d, *gas5* has little protein-coding potential (Smith and Steitz, 1998). The 1A1 RACE transcript was found to be homologous to *gas5* exons 1 and 2 (12 in total) indicating that the pEGeo2 vector landed within intron 3 (Figure 3.14).

Molecular analysis of clone 5C1 revealed that the affected locus in this case is the polyubiquitin C gene (*Ubc*) on chromosome 5. The latter consists of a short 5'UTR exon and a long exon containing tandem ubiquitin coding regions (Perelygin et al., 2002). The gene trap construct inserted into the only intron of the gene hence trapping the short 5'UTR exon (Figure 3.14).

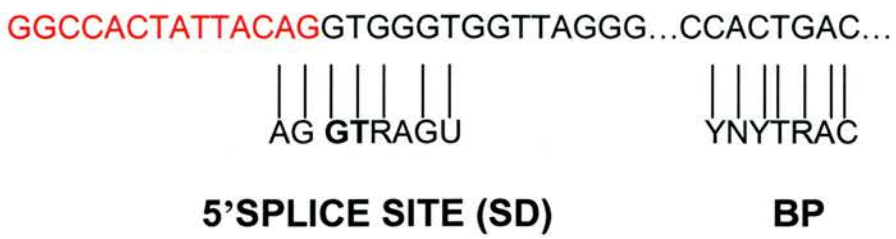
BLAST analysis of the RACE transcript from clone 1B3 suggests that vector insertion occurred within an EST gene (Bat2d; Ensembl transcript ID: ENSMUST00000046646) that is positioned on chromosome 1 and is predicted to encode a protein containing a BAT2 domain. The predicted genomic model of the gene includes 23 exons. The gene trap vector integrated within the 3'UTR between exons 21 and 22 since the 1B3 RACE product's sequence was found to be identical to the last 76 bp of exon 21 (Figure 3.14). The existence of the fusion transcript within clone 1B3 and consequently the vector's site of integration were confirmed by RT PCR (Figure 3.13).

Clone 1B2 carries a vector insertion within another EST gene which is called "expressed sequence AI506816" and is located on chromosome 5. The integration in this case took place upstream of the transcript's ATG which is positioned at the start of the longest ORF. However, *in silico* translation of RACE transcript 1B2 indicates that a translational event initiating at an ATG

present within the cloned RACE tag theoretically yields a fusion protein that is composed of 12 aminoacids encoded by the trapped sequence and the triple reporter fusion (Appendix 3). This finding suggests that reporter protein translation in this case occurred at a different translational frame from the one employed for the production of the disrupted predicted protein.

The 5'RACE product which was derived from clone 1A6 was found to be 100% homologous to a genomic region that lies on the 4$^{th}$ intron of an EST transcript ENSMUSESTT00000023421 (transcript's predicted genomic model was retrieved from the Ensembl genome server) which is located on chromosome 3. The EST transcript is predicted to consist of five exons and vector integration probably took place within intron 4. The translated RACE product sequence suggests the presence of an ORF compatible with reporter protein expression (Appendix 3) although no ATG was present either within the trapped, cloned sequence or directly upstream of it. It is likely that transcript 1A6 corresponds to an exon that has not been identified using *in silico* approaches. Evidence in support of this hypothesis comes from the fact that the boundary between the 3' end of the trapped sequence and the downstream genomic sequence on chromosome 3 exhibits sequence characteristics typical of a consensus splice donor (Fig. 3.16a). Furthermore, the RACE product's sequence exhibited considerable similarity to an EST (Accession: BY727029) and a RACE sequence tag obtained from another gene trap group (gene trap cell line ID: AD0233). It was also found to be highly conserved between mouse and rat (3.16b).

**(a)**




**(b)**

```
Mouse GGTCCCTGGAGGAGAGAGCAAGGAAGGAGAGGAAGAGTTTTTTTACCATG
      |||| |||||  ||||||||||||||||||||  ||| ||| | || ||
Rat   GGTCTCTGGAATAGAGAGCAAGGAAGGAGAGGAGAAGTGTTTCTGCCCTG

Mouse CTTTGGAGGGAGAAGAACCCGGCAGCCATGTGAGGTCTCTGGAGGAGCTG
      ||||||| |||||||||| | ||||||||||||||||||| |||| || ||
Rat   CTTTGGAAGGAGAAGAACACAGCAGCCATGTGAGGTCTCAGGAGAAGATG

Mouse GAGCCTGTGGCCACTATTACAG
      | | |||||||||||| |||||
Rat   GGG-CTGTGGCCACTACTACAG
```

**Figure 3.16** **(a)** The boundary between RACE transcript 1A6 and the downstream genomic sequence on chromosome 3 exhibits sequence characteristics of a splice donor site. The consensus splice donor sequence (nearly invariant nucleotides in bold) is also given and compared with the genomic sequence at the site of vector integration in clone 1A6, along with a putative branch point sequence present within an intron between a 5'and a 3'splice site. The 3'end of the trapped (and potentially exonic) sequence is shown in red. Sequence was retrieved from the Ensembl mouse genome server: http://www.ensembl.org/Multi/blastview. (b) The 1A6 RACE transcript sequence is conserved among mouse and rat. The RACE product's sequence is shown in red. The rat sequence is shown in black letters. (Information obtained from the UCSC genome server). SD, splice donor; BP, branch point; R, purine; Y, pyrimidine; N, any nucleotide

### 3.2.4.2 Southern blot analysis reveals the presence of multiple vector copies

A common characteristic shared by a considerable fraction (5/11 or 45%) of the cloned RACE sequence tags is the presence of stop codons in the same translational frame as the triple fusion's coding sequence (transcripts H3-10-2, H3-1, 5B1, 1B3, 1A1; see Appendix 3). The cloning of RACE transcripts that are incompatible with the translation of the vector's reporter is surprising because the gene trap clones from which those transcripts were obtained exhibited strong reporter expression despite the fact that no ATG was present in the *egfp* component of the triple fusion. RT PCR analysis of 4/5 of these RACE transcripts (H3-10-2, H3-1, 5B1 and 1B3) indicates that the stop codon-containing RACE products were genuine, constituting parts of actively transcribed mRNA species (Figure 3.13) and in the case of transcript H3-10-2 the resulting RT PCR fragment's sequence was determined and found to be identical to that of H3-10-2 thus confirming the authenticity of the specific RACE transcript.

It is possible that in the case of clone H3-10 reporter expression is facilitated by transcript H3-10-1 and the generation of H3-10-2 is the result of alternative splicing. It is also likely that the clones from which the stop codon-containing RACE products were isolated contain more than one vector copies and hence reporter expression is enabled by unidentified fusion transcripts originating from different integration events within the same clones. Indeed, Southern blot analysis (using an EGFP-specific probe) of four reporter positive gene trap clones electroporated with vector pEHygro2neoSD2 suggests that multiple vector insertion (2-3 copies) occurred in all cases examined (Figure 3.17). Two of the clones analysed (H3-1, H3-10) belonged to the group of clones that yielded stop codon-containing RACE products. Furthermore, three (H3-1, H3-10 and H4H-1) out of four

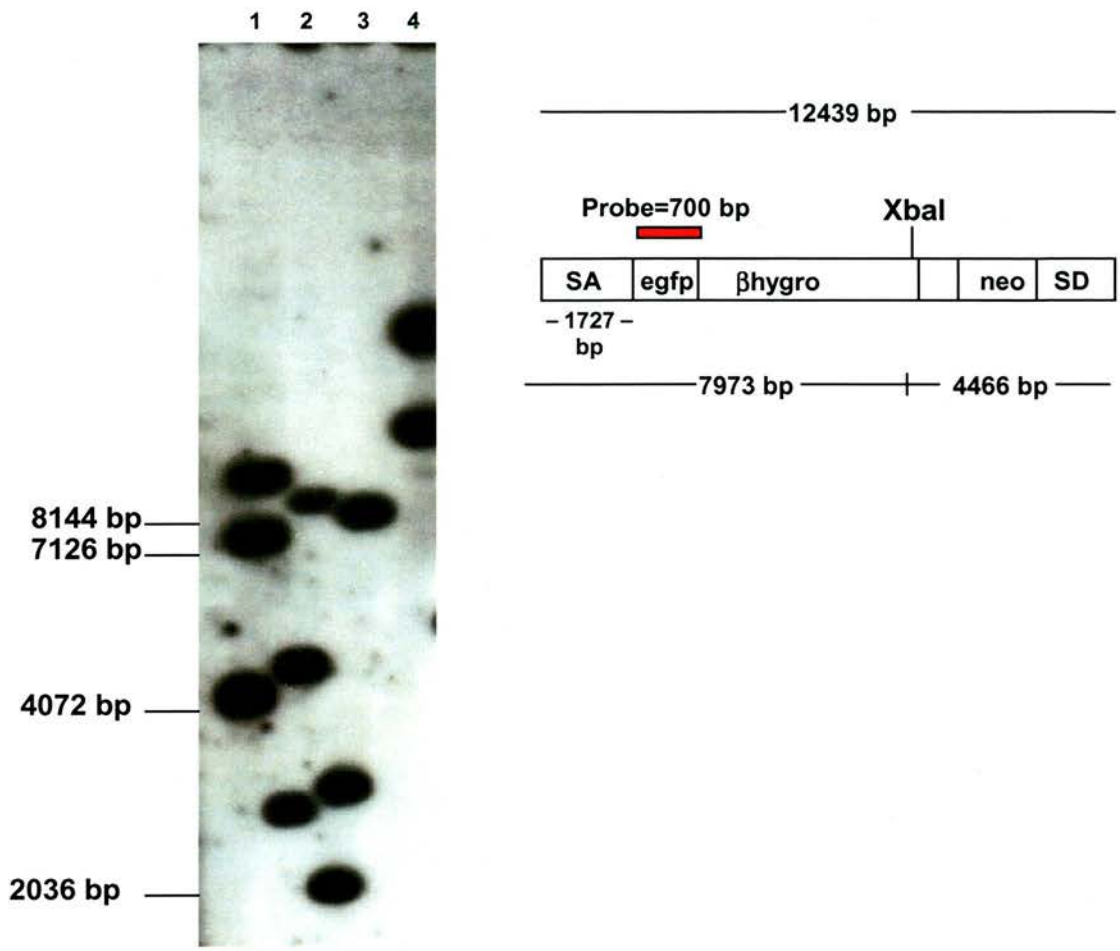



**Figure 3.17** Southern blot analysis of hygromycin resistant clones H3-1, H3-10, H4H-1 and HygroC2 using an EGFP-specific probe. Genomic DNA samples were digested with XbaI which cuts within vector pEHygro2neoSD2 only once at position 7973 nt. Theoretically probing of digested DNA should reveal a fragment or fragments of a minimum size around 7973 bp provided that an intact vector insertion occurred directly downstream of an endogenous XbaI restriction site. Lane 1, clone H3-1; lane 2, clone H3-10; lane 3, clone H4H-1; lane 4, clone HygroC2. A wild type E14 negative control was also included (not shown). SA, splice acceptor; SD, splice donor.

clones appear to contain degraded vector copies as they gave rise to fragments of considerably smaller size than the minimum expected one (approximately 7973 bp) which corresponds to an intact vector insertion event (Figure 3.17). These truncations appear to affect both the 5′ and 3′ integrity of the vector as some of the resulting bands were found to be as small as 2000 bp; the distance of the *egfp* gene from the vector's linearised 5′ end is 1776 bp. Hence this is the largest 5′ vector portion that can theoretically be degraded allowing at the same time the detection of the intact *egfp* gene (Figure 3.17). It is also likely that more complex integration events associated with vector/flanking DNA rearrangements as well as concatemer formation might have occurred. For example, the insertion of two intact vector copies in tandem and in a head-to-tail orientation would theoretically yield a band of 12439 bp (the total size of the vector); the occurrence of such an event could potentially explain the existence of the top band resulting from analysis of clone H3-1 (Lane 1 in Figure 3.17) whose size is approximately 12000 bp.

## 3.3 Characterisation of the poly(A) trap cassette

### 3.3.1 Experimental strategy

To assess the performance of the novel poly(A) trap cassette we employed plasmid vectors pEHygro2neoSD2 and pEHygro2neoSD2 (+ARE). In brief, linearised vectors were introduced by electroporation into ES cells and, after selection in G418, resistant clones were picked into 24-well plates and replicated. One set of replica plates was frozen and the other expanded for RNA isolation and subsequent analysis by 3′RACE PCR.

### 3.3.2 3'RACE analysis of pEHygro2neoSD2-electroporated clones reveals the presence of a cryptic SA site and problematic SD function

In a pilot experiment, electroporation of the HindIII-linearised pEHygro2neoSD2 vector into wild type E14 ES cells gave rise to an average of 70 neomycin resistant colonies per electroporation plate. We picked 100 colonies of which 50 were further propagated for subsequent 3'RACE PCR using *neo*-specific primers. Sequences of fusion transcripts were successfully generated from 43 G418 resistant ES cells clones (43/50, 86%).

Ideally, the proper function of the *β-globin* SD in conjunction with an endogenous, downstream SA and poly(A) signal should yield a 3'RACE product whose sequence is homologous to the splice donor's exonic sequence followed by endogenous mouse exonic sequence since the splice donor's *β-globin* intron 2 should, theoretically, be removed upon splicing (Figure 3.18). However, analysis of the resulting transcripts revealed that no proper SD function took place in approximately half (21/43, 49%) of the clones analysed. The sequence of 9 (9/43, 21% of the total) RACE products consisted of the *neo* coding sequence followed directly by mouse genomic sequence of different length and chromosomal origin in each case. The remaining 12 (12/43, 28%) transcripts contained unspliced SD sequence indicating that SD read-through took place in these cases (Table 3.3).

In 22 (51%) of the obtained RACE sequences the *βglobin* exon 2/intron 2 junction was utilised in the predicted manner since the splice donor's *βglobin* exonic sequence was followed by a properly spliced sequence. However, the spliced sequence originated from the vector's backbone and it was located directly downstream of the *β-globin* SD intron 2 suggesting that cryptic splicing within the vector occurred (Figure 3.19a) (Table 3.3). This intra-vector splicing event employed the poly(A) trap's SD and the cryptic

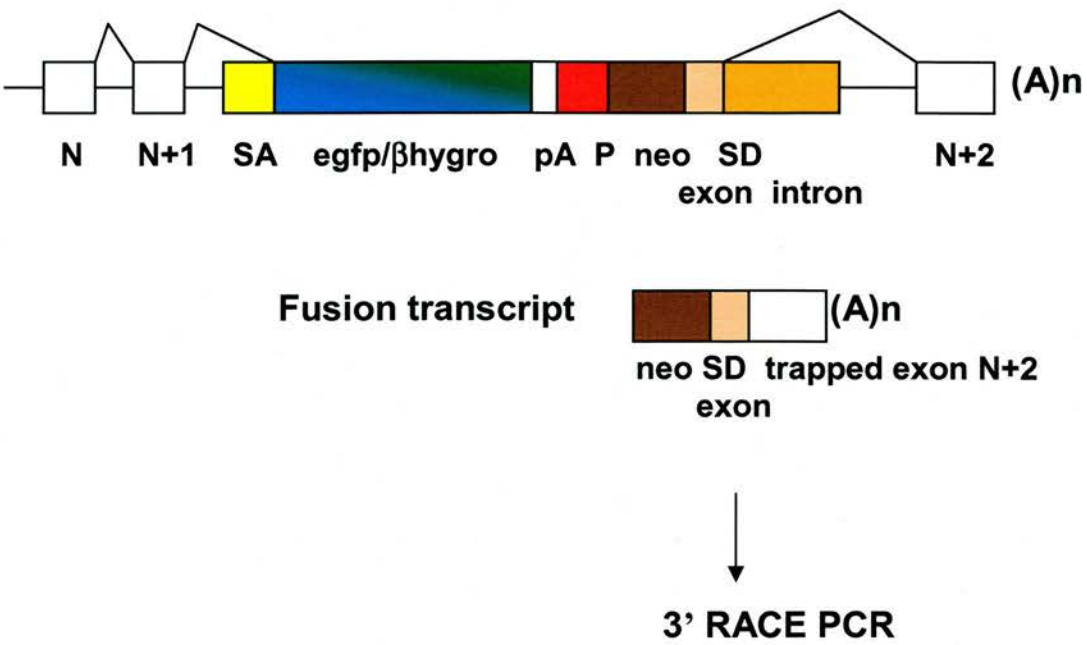



**Figure 3.18** Poly(A) trapping with the pEHygro2neoSD2 vector. Efficient SD usage should result after splicing into an endogenous exonic sequence in the generation of a fusion transcript that consists of neo, *βglobin* SD exon and trapped exon N+2. Transcript stabilisation should occur through the employment of the trapped gene's poly(A) signal. Open boxes, trapped gene's exons.

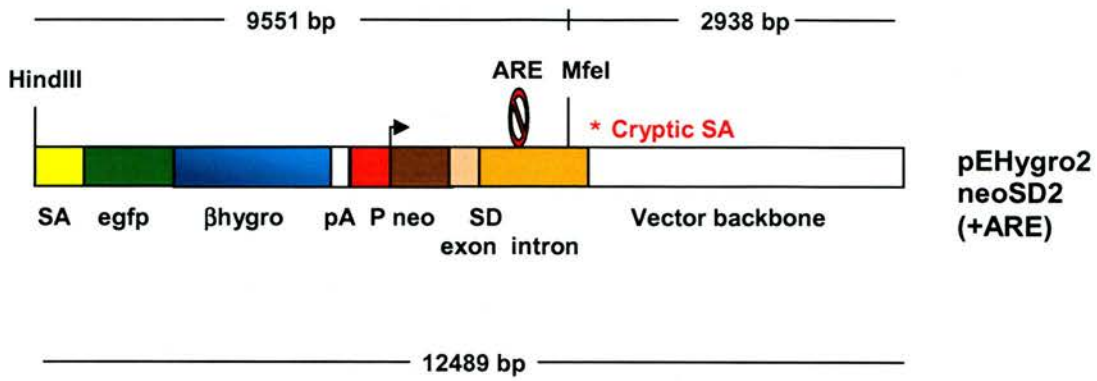
SA site present within the vector backbone resulting in the generation of a fusion transcript between *neo*, the SD's exonic sequence and vector backbone sequence. The "trapped" sequence originating from the vector backbone had a constant length of 90 nt in all cases and consisted of a 5′ 53 bp-segment, which was found by BLAST analysis to be homologous to the beginning of the rabbit *β-globin* exon 3, and an additional 37 bp stretch of downstream vector backbone sequence followed by a stretch of A's. This stretch of A's found at the end of the transcripts indicates that a proper polyadenylation event took place.

### 3.3.3 The inclusion of an ARE improves the vector's SD function

To address the intra-vector splicing problem associated with the use of the HindIII-linearised vector we adopted a strategy that involved the construct's double digestion with the restriction enzymes HindIII and MfeI. A unique MfeI target site is present within the 3'end of the splice donor's intron and upstream of the backbone's cryptic splice acceptor (Figure 3.19a). Hence this strategy is predicted to result in the removal of the vector backbone including the cryptic SA. Both pEHygro2neoSD2 and pEHygro2neoSD2 (+ARE) plasmid vectors were digested with HindIII and MfeI giving rise to a 9501 bp or 9551 (in the case of the +ARE construct) fragment that corresponds to the functional vector and a 2938 bp fragment that corresponds to the cryptic SA-containing vector backbone (Figure 3.19b). The "functional vector" fragment was then gel-purified following agarose gel electrophoresis of the digestion mixture and used for electroporation into wild type ES cells.

G418 selection following transfection with the pEHygro2neoSD2 (+ARE) vector resulted in approximately 1.7-fold reduction in the number of

**(a)**




**(b)**

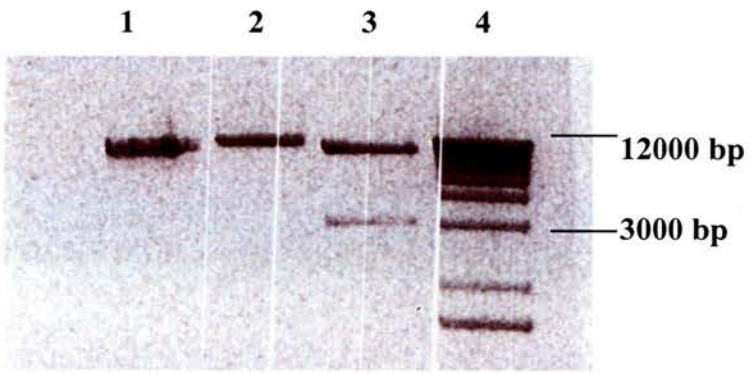



**Figure 3.19** **(a)** Schematic representation of vector pEHygro2neoSD2 (+ARE) showing the target sites of HindIII and MfeI restriction enzymes relative to the cryptic SA site within the vector backbone. Vector pEHygro2neoSD2 has the same structure but lacks the ARE. **(b)** Agarose gel analysis of the products generated after double digestion of vector pEHygro2neoSD2 (+ARE) with HindIII and MfeI (lane 3). Lane 1: gel-purified fragment (9551 bp) that corresponds to the functional vector after the double digestion. Lane 2: HindIII-linearised vector (12489 bp). Lane 4: DNA size markers.

neomycin resistant clones (40 resistant clones per electroporation plate) compared to the number obtained when the ARE-less pEHygro2neoSD2 construct was used (67 resistant clones per electroporation plate). This result might indicate a reduction in the fraction of clones that acquired neomycin resistance through read-through events and employment of cryptic poly(A) signal sites and therefore a decrease in the "background".

45 pEHygro2neoSD2- and 43 pEHygro2neoSD2 (+ARE)-electroporated gene trap clones were analysed by 3′ RACE PCR. Our double-digestion approach appeared to be successful in addressing the intra-vector splicing issue; the cryptic SA was used in only six pEHygro2neoSD2-containing clones (6/45, 13%) and in none of pEHygro2neoSD2 (+ARE)-containing clones (Table 3.3). We postulate that this small fraction of intra-vector splicing events observed in pEHygro2neoSD2 clones is likely to represent contamination by HindIII-only digested vector during the gel purification process.

The *βglobin* SD functioned efficiently in most of the pEHygro2neoSD2 (+ARE) clones (33/43, 77%). This is evidenced by the succesfull generation of fusion transcripts which incorporated endogenous mouse sequences correctly spliced into the splice donor's exon while the *βglobin* intron 2 was successfully removed during the splicing process (Table 3.3). Seven (7/43, 16%) of the RACE transcripts cloned were indicative of read-through into the splice donor's intron (Table 3.3) while in two cases (2/43, 7%), RACE product sequences consisted of *neo* coding sequence followed by endogenous genomic sequence. Only 5 3′RACE transcripts (5/45, 11%) isolated from pEHygro2neoSD2 clones were indicative of proper SD function (Table 3.3). The SD was not used appropriately in the majority (34/45, 76%) of the pEHygro2neoSD2 3′ RACE products as SD read-through occurred in most of

| Construct | Efficient SD use (%) | Intra-vector splicing (%) | SD read-through (%) |
|---|---|---|---|
| pEHygro2neoSD2 (HindIII) | 0 | 51 | 28 |
| pEHygro2neoSD2 (HindIII/MfeI) | 11 | 13 | 69 |
| pEHygro2neoSD2 (+ARE) (HindIII/MfeI) | 77 | 0 | 16 |

**Table 3.3** The ARE enhances the *βglobin* splice donor's performance while the vector's double HindIII/MfeI digestion reduces the incidence of splicing events that employ the vector's backbone cryptic SA (defined as intra-vector splicing). Percentages corresponding to RACE products consisting of *neo* sequence followed by endogenous mouse genomic sequences are not shown. N=43 for pEHygro2neoSD2 (HindIII); N=45 for pEHygro2neoSD2 (HindIII/MfeI); N=43 for pEHygro2neoSD2 (+ARE) (HindIII/MfeI).

the cases (Table 3.3). These data suggests that the presence of an ARE has a positive effect on the *βglobin* SD performance.

### 3.3.4 Analysis of trapped transcripts

The properly spliced trapped sequences generated from 3′ RACE PCR analysis of the clones were examined using the BLAST (or BLAT) algorithm and the NCBI (http://www.ncbi.nlm.nih.gov/BLAST/), Ensembl (http://www.ensembl.org/Mus_musculus/index.html) and UCSC (http://genome.ucsc.edu/index.html?org=Mouse) mouse databases. All sequence tags (resulting from electroporations with both + and -ARE vectors) and details of their BLAST analysis are given in Appendix 4. Results are summarized in Tables 3.4 and 3.5.

Depending on the quality of the generated sequence and the site of integration, fusion transcripts were often found to incorporate a poly(A) signal site and a terminal $(A)_n$ tail. This indicates that the vector's poly(A) trap cassette functioned in the predicted manner by capturing endogenous poly(A) signals for stabilisation of the vector's *neo* transcript. Gene trap events were distributed among 14 chromosomes with chromosomes 4 and 6 containing the highest number (n=4) of vector insertions (Figure 3.20).

Of the 38 properly spliced RACE transcripts isolated from both pEHygro2neoSD2 and pEHygro2neoSD2 (+ARE) clones, 15 (40%) matched reported transcribed sequences (Figure 3.21). Of these, 8 (22%) matched exons of known genes, and 7 (18%) were found to be homologous to EST transcripts (Tables 3.4 and 3.5; Figure 3.21). One of the RACE transcripts (from clone 2) that matched exons of known genes was found to be homologous to the reverse strand of the trapped gene (*Rfx4*).

A substantial fraction (8/38, 22%) of the RACE products did not align with any known/EST transcripts but overlapped with regions that were

| Clone | Chromosome | Details |
|---|---|---|
| N2 | ? | Mouse ETn LTR |
| N19 | 7 | Phlda2 (exon) |
| N23 | 6 | Unknown |
| N29 | ? | Mouse ETn LTR |
| N34 | 5 | Ccdc18 (intron) |

**Table 3.4** Identities of trapped sequences obtained after 3'RACE PCR analysis of clones electroporated with the HindIII/MfeI pEHygro2neoSD2 vector and in which the rabbit *βglobin* splice donor functioned in the predicted manner.

| Clone | Chromosome | Details |
|---|---|---|
| 1, 6, 7, 30, 32, 33, 35 | ? | Mouse ETn LTR |
| 2 | 10 | Rfx4 (exon) |
| 4 | 18 | GENSCAN000000402318 |
| 5 | 12 | Ylpm1 (exon) |
| 8 | 4 | 45S pre-ribosomal RNA gene (Repeat) |
| 11 | 19 | EST (Accesion no: AI506879) |
| 12 | 11 | EST (Accesion no: AK155239) |
| 13 | 4 | GENSCAN00000391247 |
| 16 | 4 | Pnrc2 (exon) |
| 19 | 8 | GENSCAN00000375112 |
| 22 | 3 | D3Wsu161e gene (intron) |
| 23 | 12 | GENSCAN00000387185 |
| 28 | 4 | Tmem57 (exon) |
| 29 | 3 | Repeat |
| 34 | 14 | GENSCAN00000376976 |
| 44 | 6 | GENSCAN00000381797 |
| 46 | 7 | EST (Accesion no: AK140938) |
| 2D4 | 12 | EST (Accession no: XM_903470) |
| 2A4 | 6 | EST (Accesion no: BF019192) |
| 2C1 | 13 | GENSCAN00000370450 |
| 1C6 | 18 | Tcerg1 (exon) |
| 3C4 | 2 | Cds2 (exon) |
| 2C4 | 13 | Hist1h2bh (exon) |
| 2C6 | 7 | EST (Accesion no: BY432161) |
| 2A4 | 10 | EST (Accesion no: DV656129) |
| 1A4 | 6 | Repeat |
| 1D6 | 14 | GENSCAN00000376976 |

**Table 3.5** Identities of trapped sequences obtained after 3'RACE PCR analysis of clones electroporated with the HindIII/MfeI pEHygro2neoSD2 (+ARE) vector and in which the rabbit *βglobin* splice donor functioned in the predicted manner.

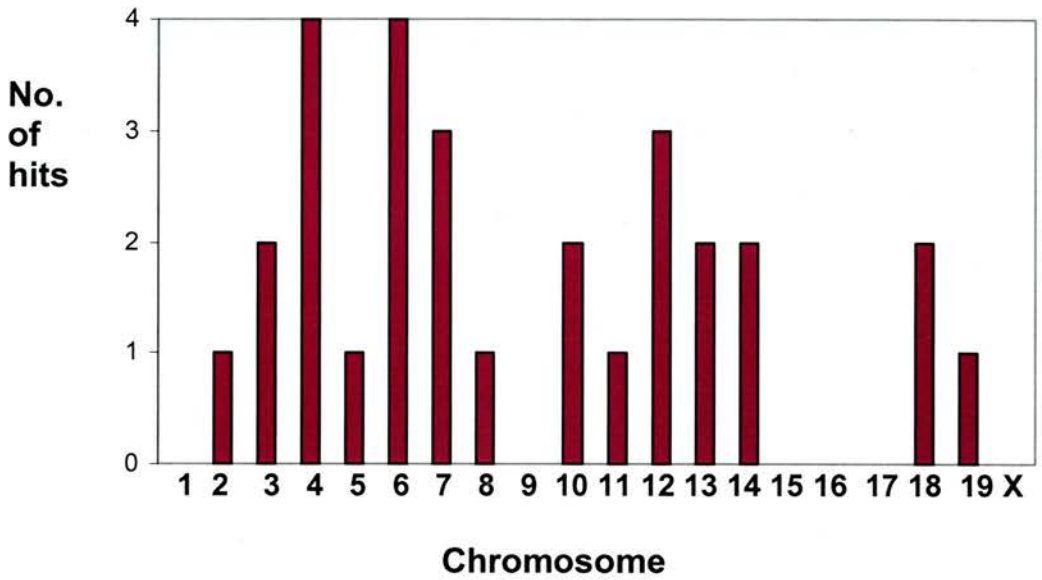



**Figure 3.20** Chromosomal distribution of poly(A) trap events resulting from analysis of properly spliced 3'RACE transcripts. RACE products were isolated from neomycin resistant clones electroporated with both pEHygro2neoSD2 and pEHygro2neoSD2 (+ARE) vectors.

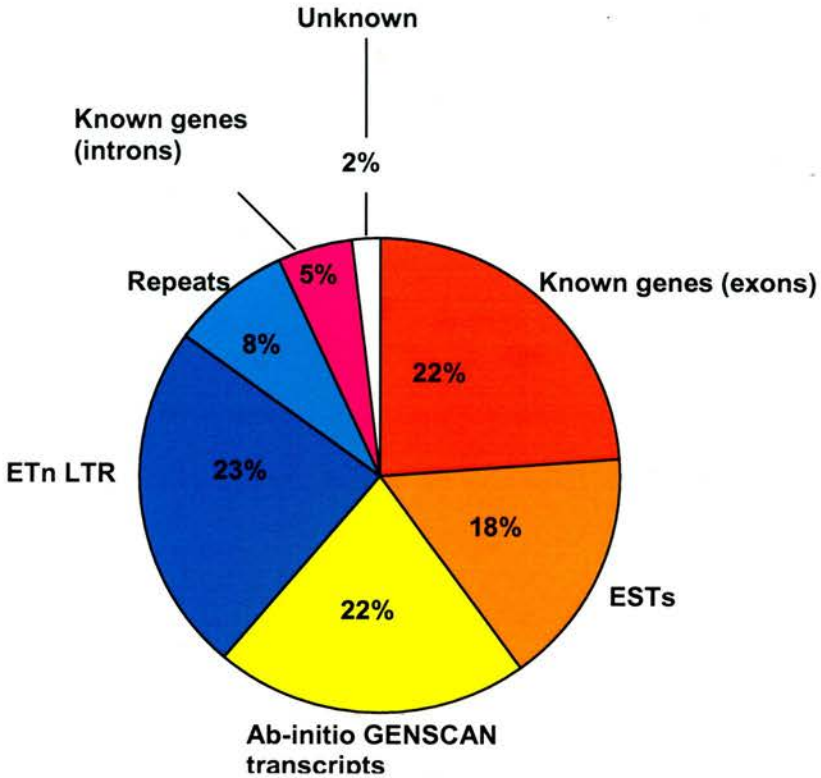



**Figure 3.21** BLAST hit distribution of 3'RACE transcripts derived from both pEHygro2neoSD2 (+ARE) and pEHygro2neoSD2-electroporated, neomycin resistant clones.

located within ab-initio predicted transcripts determined using the GENSCAN program (Table 3.5; Figure 3.21; information obtained from the Ensembl mouse genome server). The latter is designed to identify exon/intron structures of vertebrate genes with an accuracy of 75-80% (Burge and Karlin, 1997). Some of the RACE transcripts that were found homologous to GENSCAN transcripts also exhibited splicing patterns i.e. different segments of their sequence were found to be homologous to different adjacent genomic regions indicating the trapping of different exons (e.g. clone 19; see Appendix 4). Many of the GENSCAN transcript-homologous 3'RACE sequences were also found to be highly conserved between different organisms (e.g. clones 4, 13, 19, 23 and 44: Appendix 4). These observations suggest that some of the RACE products are likely to represent insertions in *bona fide* novel genes. Two (5%) RACE transcripts (from clones N34 and 22) matched the reverse strands of sequences located in introns of known genes and one (2%) transcript matched a region without any known or predicted genes in the vicinity but had a splicing pattern (Appendix 4). Again, some of these sequences might indicate the existence of genes not predicted by other means.

A significant percentage (23%) of the 3'RACE products corresponded to a sequence from the 3' LTR (long terminal repeat) present in the repetitive ETn (early transposon) family of sequences which are highly dispersed within the mouse genome (200 copies; Tanaka and Ishihara, 2001). The same ETn sequence has also been trapped repeatedly by other vectors that include a poly(A) trap cassette (Yoshida et al., 1995; Chen et al., 2004a). Finally, three clones (8, 29 and 1A4; 8%) yielded RACE transcripts that matched other repetitive sequences on specific chromosomal locations (Table 3.5; Appendix 4). In general, the nature and the percentages of BLAST matches obtained for

our 3'RACE transcripts reflect findings of other studies employing poly(A) trap constructs (Matsuda et al., 2004; Osipovich et al., 2005).

### 3.3.5 Our poly(A) trap vectors do not appear to exhibit a bias in their integration site preference

It has been recently demonstrated that one of the major limitations associated with poly(A) trapping is the tendency shown by this type of vectors to insert into the last intron of their target genes presumably due to the action of nonsense-mediated mRNA decay (NMD) (Shigeoka et al., 2005; see also relevant section in the Introduction). Examination of some of our 3'RACE products that match exons of transcripts whose genomic structure is well defined indicates that our poly(A) trap vectors do not appear to suffer from this bias (Table 3.6). None of the RACE transcripts analysed was indicative of vector integration into a gene's last intron (Table 3.6). In 4/7 cases vector insertion took place within the first intron, two approximately in the middle (closer to the 3'end) of the gene and one in the intron before the last (Table 3.6). These RACE products appear to represent poly(A) trapping events that evaded NMD.

Interestingly, the termination codon (TC) of the pEHygro2neoSD2 vector's *neo* coding sequence is located 362 bp upstream of the splice donor's splice junction. This distance significantly exceeds the "60 nt from the last exon-exon junction" limit for evasion of NMD. Hence any fusion transcript between *neo* and downstream trapped endogenous sequences should theoretically constitute a default target for NMD irrespective of the site of the vector's integration since in any case the *neo's* termination codon is always further than 60 nt from a target gene's last exon-exon junction and would therefore be recognised as a potentially premature termination codon. However, the fact that we successfully generated through 3'RACE PCR *bona*

| Gene Symbol | Accession number | Total number of exons | Number of trapped exons |
|---|---|---|---|
| Ylpm1 | NM_178363 | 22 | 2 |
| Rfx4 | AY342003 | 18 | 8 |
| Pnrc2 | AK077403 | 3 | 2 |
| Tmem57 | BC037192 | 11 | 10 |
| Tcerg1 | BC040284 | 22 | 8 |
| Cds2 | AK170888 | 13 | 12 |
| ENSMUSESTT00003795807 | AK140938 | 5 | 4 |

**Table 3.6** Overview of vector insertion sites within trapped genes. Only RACE sequences that were homologous to exons of well-defined transcripts were used for the analysis. Genes consisting of two exons were excluded. The number of trapped exons corresponds to the number of exons downstream of the vector integration site.

*fide neo* fusion transcripts derived from clones electroporated with vectors pEHygro2neoSD2 and pEHygro2neoSD2 (+ARE) suggests that somehow our vector is immune to NMD and this might also explain the absence of a bias towards integrations into the 3'most intron of the vector's target genes.

### 3.3.6 Overview of disrupted genes

Many of the known genes trapped by our vectors have been shown to possess distinct roles in mouse embryonic development and exhibit restricted, developmentally regulated expression patterns. For example, the *Phlda2* gene which is trapped by vector pEHygro2neoSD2 in clone N19 exhibits restricted expression in the lateral mesoderm and the most posterior extent of the primitive streak as well as in extra-embryonic tissues (Dunwoodie and Beddington, 2002). Knockout studies suggest that this locus probably plays a role in regulating placental growth (Frank et al., 2002). The *Rfx4* gene (clone 2) encodes a 735 aa winged helix transcription factor which appears to be essential for early brain development (Blackshear et al., 2003). Another example of a developmentally regulated gene that has been trapped by our vectors is *Tmem57* (or C61; clone 28), a neuronal-specific gene whose expression is upregulated in the telencephalic cortical plate and mitral cells in the olfactory bulb after day E14.5 of embryonic development (Kuvbachieva et al., 2005).

Interestingly, one of the clones (2C4) gave rise to a fusion transcript that was homologous to the histone gene *Hist1h2bh*; histone-encoding genes are known to lack introns and a poly(A) tail ending instead in a highly-conserved between metazoans stem-loop sequence (Marzluff et al., 2002). The same *Hist1h2bh*-related sequence has also been captured by other groups (CMHD cell lines CMHD-GT_150D12-3 and CMHD-GT_238E5-3 and

FHCRC cell line FHCRC-GT-S19-5A1; information obtained from www.igtc.org) using constructs that incorporate a poly(A) trap component. It is likely that in this case the trapped gene's single exon contains a cryptic SA site that was activated upon the vector's integration upstream of the gene giving rise after cryptic splicing to a fusion *neo/Hist1h2bh* transcript that is stabilised in the same manner as the native *Hist1h2bh* transcript.

A combination of direct (RT-PCR; Figure 3.22) and indirect evidence (published literature, EST databases) shows that all the known genes trapped by the pEHygro2neoSD2 (+/-ARE) vectors are expressed at various levels in undifferentiated ES cells. A search of the IGTC database revealed that 7/8 of these genes have been trapped by conventional SA-type gene trap constructs which is another indication that these genes are expressed in ES cells. Interestingly, the *Phlda2* gene (clone N19) has only been trapped by poly(A) trapping (CMHD cell line CMHD-GT_222A12-3) despite the fact that is highly expressed in ES cells (3.21; Dunwoodie and Beddington, 2002). This could suggest that poly(A) trap vectors target a unique set of genes that are not accessible to entrapment by other vector types.

Since our sample size is quite small for assessing the above speculation we expanded our analysis to gene targets of other poly(A) trap constructs. We examined all IGTC genes that have been trapped by poly(A) trap vectors employed by the CMHD. We found that, to date, poly(A) trap vectors have trapped 1025 genes which correspond to 6.2% mouse genome coverage (the total number of IGTC trapped ENSEMBL genes, to date, is 6644 representing 40.36% of genome coverage). Half of these genes (513 or 50%) have only been trapped by means of poly(A) trapping while the rest (512) have been trapped by both poly(A) trap and promoter/SA-type vectors (Figure 3.23a). Interestingly, 27 of 513 the genes mutated exclusively by

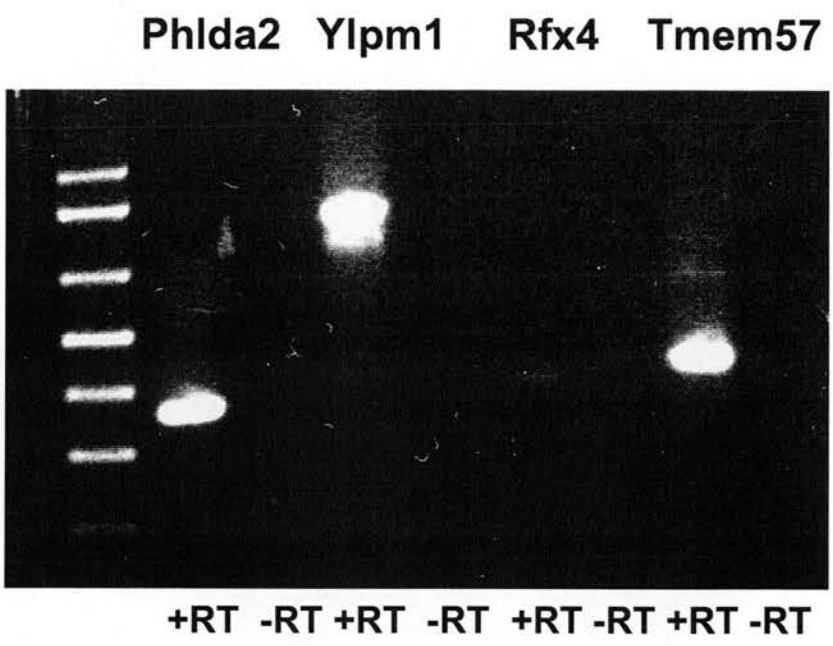



**Figure 3.22** RT-PCR analysis of wild type undifferentiated E14 ES cells using primer combinations specific for genes Phlda2, Ylpm1, Rfx4 and Tmem57. Minus RT controls are also shown.

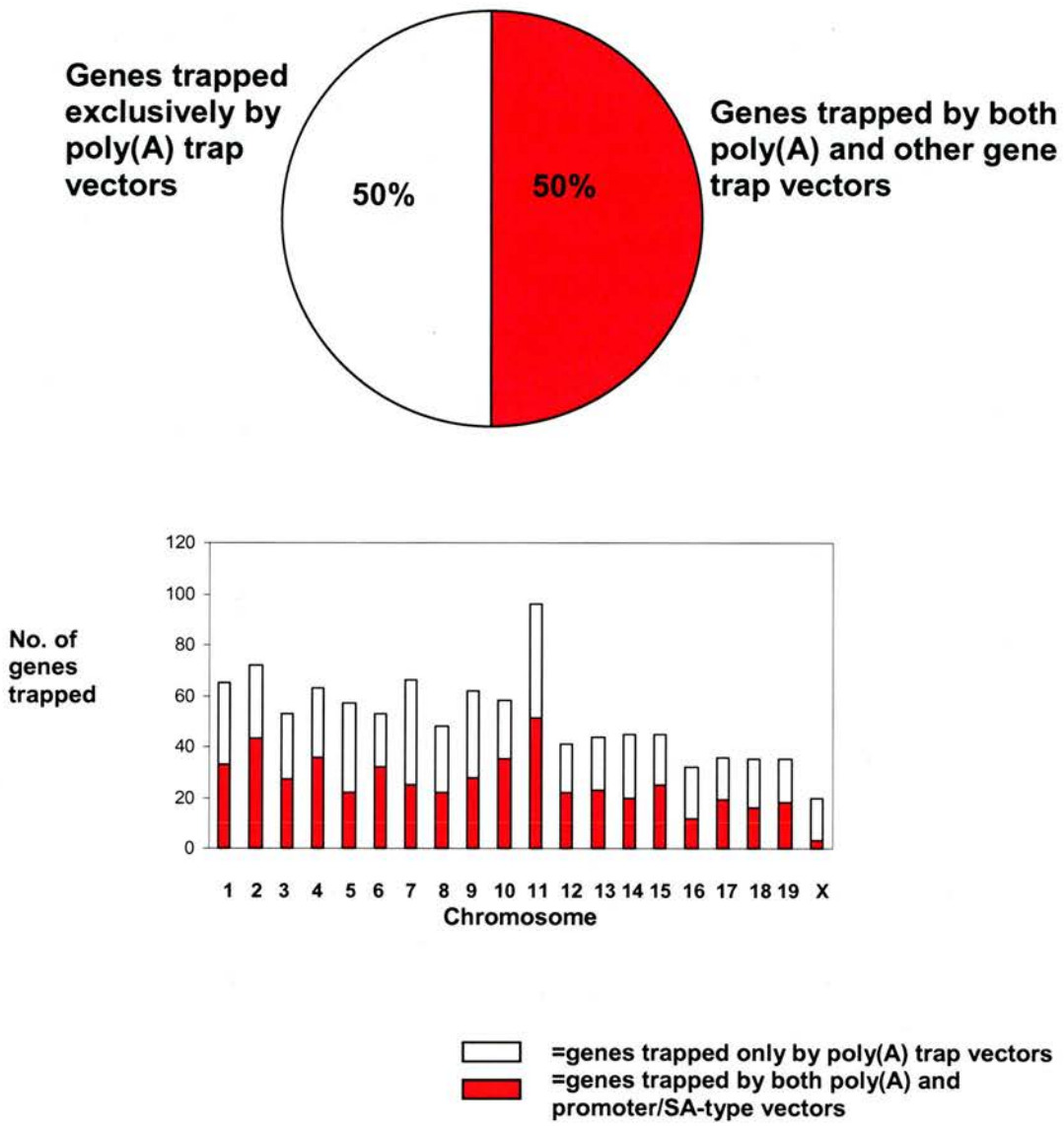



**Figure 3.23** **(a)** Chart depicting the "accessibility" of 1025 genes targeted by poly(A) trap vectors to entrapment by conventional SA-type constructs. **(b)** Chromosomal distribution of the above genes (N=1025).

poly(A) trapping are expressed on undifferentiated ES cells based on existing EST data.

The chromosomal distribution of the poly(A)-trapped genes is shown in Figure 3.23b. The highest number of poly(A)-trapped genes (96) were located on chromosome 11. In general, the number of genes present in a chromosome appeared to correlate with the number of genes trapped by poly(A) trap vectors suggesting that poly(A) trapping events occur more frequently in chromosomes with a high gene density (Figure 3.24). A similar preference has also been reported for SA and promoter entrapment vectors (Hansen et al., 2003). Furthermore, the highest density of exclusive poly(A) trapping gene targets was found on chromosomes 7 and X; in these chromosomes the number of genes trapped only by poly(A) trap vectors was considerably higher (2- and 5.5-fold for chromosomes 7 and X respectively) (Figure 3.23b). We also observed the existence of insertional "hot spots" that were specific to the poly(A) trap vectors employed; some loci were trapped up to 20 times (*Pacrg*) by different poly(A) trap constructs but not by any other class of vectors (Table 3.7). Analysis of the vector integration sites for 180 of the 513 poly(A) trapping-specific genes that contain more than two exons and whose exon/intron structures are well defined revealed that the majority of vector insertions (approximately 75%) occurred within the 3'-most intron of the genes (Figure 3.25). This observation is in agreement with the findings of Shigeoka et al., (2005) and demonstrates the biased nature of entrapment using poly(A) trap vectors.

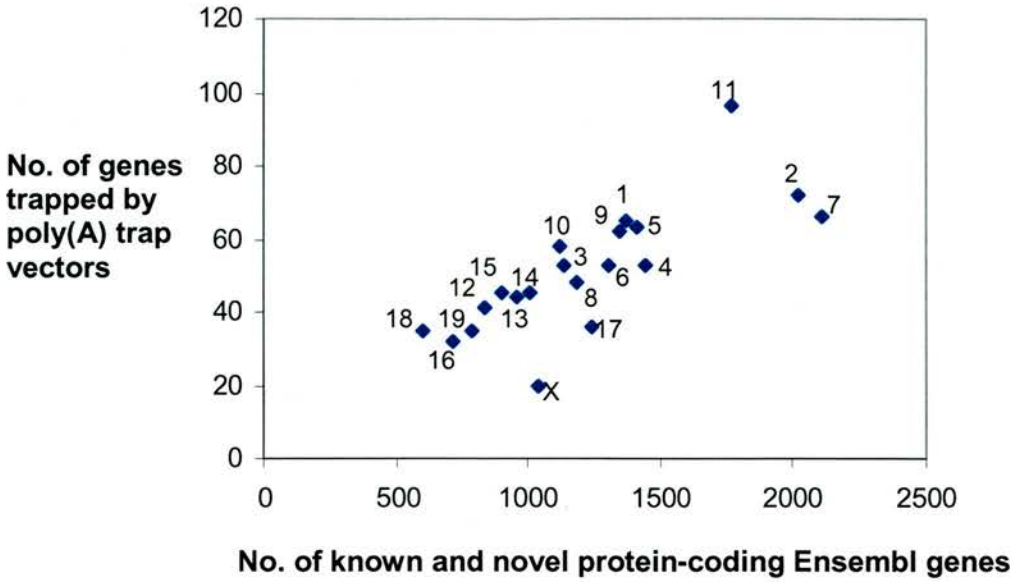



**Figure 3.24** Correlation between the number of poly(A) trapped genes and the number of known/novel Ensembl genes per chromosome.

| Gene | Chromosome | No. of insertions |
|---|---|---|
| Plekha4 | 7 | 7 |
| Wdr20 | 12 | 9 |
| A930008G19Rik | 7 | 4 |
| Rbks | 5 | 18 |
| Ppp1r15b | 1 | 8 |
| Pacrg | 17 | 20 |
| Ccdc57 | 11 | 11 |
| Apod | 16 | 9 |

**Table 3.7** Genes that have been multiply disrupted exclusively by poly(A) trap vectors and hence represent potential poly(A) trapping "hot spots".

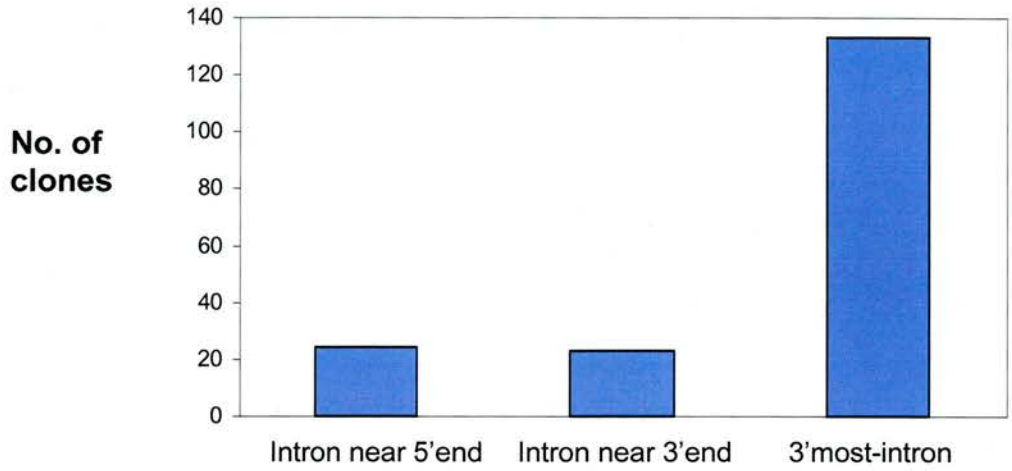



**Figure 3.25** Distribution of vector insertion sites within poly(A) trapping-specific genes. To identify the insertion sites, 180 genes that contain more than two exons and whose genomic structure is well defined were analysed. Introns near the 5'and 3'ends were defined as being located 5'and 3' to the middle exon or intron of a gene, respectively.

### 3.3.7 Reporter expression of trapped clones

To examine whether the 5′ triple reporter fusion within our poly(A) trap vectors recapitulates the expression of the poly(A)-trapped genes we determined the β-galactosidase activity of 100 undifferentiated (+LIF) $neo^R$ including ones that were found by 3′RACE PCR to contain vector insertions within genes that are expressed in ES cells and therefore should theoretically be X-gal positive (Section 3.3.6). 15 of these 100 clones were also subjected to hygromycin selection for analysis of the reporter's hygromycin resistance component. Unexpectedly, none of the clones examined appeared to be lacZ positive or hygromycin resistant.

We hypothesised that the absence of reporter gene expression in pEHygro2neoSD2 (+ARE)-electroporated $neo^R$ clones may be associated with an increased sensitivity to deletions in the vector's 5′ functional segment induced upon electroporation while the poly(A) cassette remained unaffected resulting in the generation of neomycin resistant clones. We analysed by Southern blotting four $neo^R$ clones (1A4, 2A4, 3C4, 2C4) electroporated with vector pEHygro2neoSD2 (+ARE) using a reporter-specific probe (EGFP). Two of these clones (3C4, 2C4) yielded 3′RACE products that are indicative of a vector insertion within genes that are expressed in ES cells. Our results demonstrate that at least one copy of the triple fusion's eGFP component is present within all clones (Figure 3.26). 2/4 clones (1A4, 2A4) were found to contain more than one vector copies (Figure 3.26). Furthermore, all clones examined appeared to contain degraded vector fragments as they gave rise to bands of a smaller size than the minimum expected one (approximately 7973 bp). The extent of deletions does not appear to be as great as in the case of the hygromycin selected clones

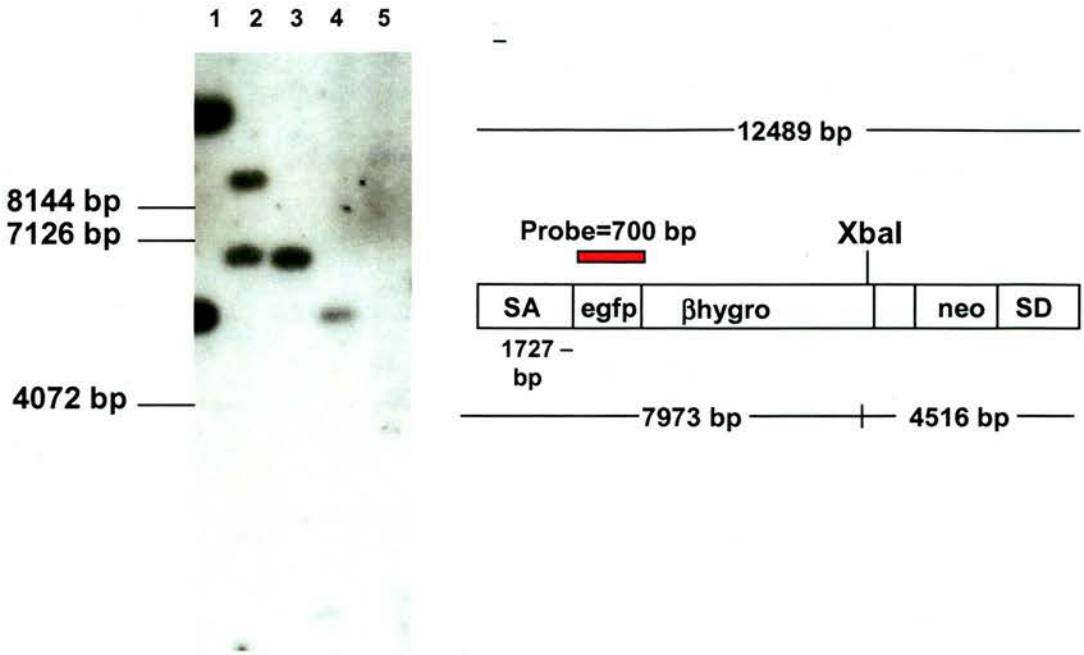



**Figure 3.26** Southern blot analysis of neomycin resistant clones 1A4, 2A4, 3C4 and 2C4 using an EGFP-specific probe. Genomic DNA samples were digested with XbaI which cuts within vector PEHygro2neoSD2 (+ARE) only once at position 7973 nt. Theoretically probing of digested DNA should reveal a fragment or fragments of a minimum size of approximately 7973 bp provided that an intact vector insertion occurred directly downstream of an endogenous XbaI restriction site. Lane 1, clone 1A4; lane 2, clone 2A4; lane 3, clone 3C4; lane 4, clone 2C4; lane 5; wild-type E14 ES cells.SA, splice acceptor; SD, splice donor.

analysed by Southern blotting (Figure 3.17). In the cases of clones 2A4 (lower band) and 3C4 the degradation-indicative bands (size=approximately 6000 bp) could correspond to 5′ truncations that have only affected the *En-2* splice acceptor and this fact could explain the absence of reporter expression in the case of clone 3C4 (Figure 3.26). Clone 2C4 yielded a smaller fragment of approximately 5 kb which suggests that in this case the vector was degraded at both 5′ and 3′ ends (Figure 3.26). However, clone 2C4 should contain a functional 3′ poly(A) trap cassette as it was found to be *neo* resistant. It is possible that this clone also includes a second vector integration that lacks completely the 5′ *egfp* component (and hence cannot be detected by Southern blotting using an EGFP-specific probe) but has retained an intact 3′ poly(A) trap module. Unfortunately, efforts to analyse the same clones using a *neo*-specific probe were not successful.

# CHAPTER 4

# DISCUSSION

## 4.1 Characterisation of novel gene trap vector components

This Thesis describes experiments aiming to test a series of gene trap vectors that were developed as tools for trapping developmentally regulated genes. One of the novel features of these vectors is the presence of a 5′ triple reporter fusion that incorporates the *egfp*, *lacZ* and neomycin or hygromycin resistance genes (*egfpβgeo* and *egfpβhygro* respectively) placed downstream of the *En-2* SA. The inclusion of eGFP in this reporter system provides an extra molecular tag directly reflecting the protein levels of the trapped gene in addition to the enzymatic amplification-based *lacZ* marker. Moreover, the presence of eGFP offers a noninvasive means of monitoring trapped gene expression in living cell populations and permits the rapid screening of a trapped cell library by flow cytometry thus offering the potential for more high-throughput approaches. The combination of antibiotic resistance (provided either by *βgeo* or *βhygro*) and eGFP expression can be also exploited for enriching reporter-expressing cells during *in vitro* differentiation and, consequently, facilitating lineage selection (Li et al., 1998).

The other novel feature of our vectors is the presence of a 3′poly(A) trap cassette that includes the previously uncharacterized rabbit *β-globin* exon 2/intron 2 splice donor (SD) junction and a 50 bp mRNA instability signal (AU-rich element) (ARE) which is derived from the human GM-CSF

gene (Xu et al., 1997) and is cloned into the *β-globin* intron, approximately 120 nt downstream of the SD exon/intron junction. The presence of the ARE directly downstream of the poly(A) trap's SD should potentially eliminate any neomycin resistant clones that arise from the stabilisation of the *neo* mRNA transcript (via SD read-through) by cryptic poly(A) signal sites that reside within the splice donor's *β-globin* intron or intergenic mouse genomic regions thus resulting in the reduction of 'background' vector integrations.

### 4.1.2 Characterisation of the 5'triple reporter fusion

#### 4.1.2.1 The triple fusion proteins possess the potential to be efficient reporters of endogenous gene expression

Reporter expression analysis of neomycin and hygromycin resistant gene trap clones containing the *egfpβgeo* and *egfpβhygro* respectively suggests that the individual components of the fusion proteins can function properly. The percentage of reporter expressing, hygro$^R$ clones was found to be higher than the one obtained for the neo$^R$ clones (Table 3.1). This is likely to reflect the fact that a higher level of endogenous trapped gene expression is required for activation of the *egfpβhygro* gene and subsequent hygromycin resistance compared to *egfpβgeo* and hence a greater number of hygro$^R$ clones are expected to be positive for reporter expression.

A wide variety of *lacZ/egfp* expression profiles of differing intensity and cell distribution were observed in both neomycin and hygromycin resistant clones indicating that our vectors can disrupt distinct loci of different transcriptional activity states (Figures 3.4-3.8). Overall, flow cytometry appears to be a more sensitive means of detecting fluorescence (both as a function of fluorescence distribution and intensity) compared to microscopy (Figure 3.10). Detection of eGFP expression by fluorescence microscopy was possible only in the case of clones that were also found by

flow cytometry to express high levels of eGFP. However, even the highest eGFP expressors appeared to be only moderately bright under microscopy and the use of a more efficient fluorescent marker (e.g. "Venus"; Nagai et al., 2002) as an alternative to eGFP might be the best option for future vector designs.

In general, we observed a correlation between eGFP expression and β-galactosidase activity; clones characterized by high *lacZ* expression were also found by flow cytometry/fluorescence microscopy to contain the highest fractions of fluorescent cell populations (Figure 3.10). X-gal staining of eGFP positive clones and subsequent analysis by microscopy demonstrated that β-galactosidase activity correlates well with eGFP expression both in terms of cellular distribution and intensity (Figure 3.11).

Approximately half of the X-gal positive clones were found to be negative for eGFP expression (Table 3.1). In the case of clones characterized by lower levels of *lacZ* expression the lack of fluorescence could be attributed to the fact that the eGFP reporter is less sensitive compared to *lacZ* as it is detected without enzymatic amplification and therefore requires higher expression for detection (>30,000 molecules/cell for the bright mutants; Whitney et al., 1998). However, some of the eGFP negative clones were found to express high levels of β-galactosidase and should theoretically possess detectable fluorescent cell populations.

A more likely explanation for the absence of eGFP expression in *lacZ* positive clones might be the loss of fluorescence due to protein misfolding or steric hindrance. It should be noted that we attempted to address this problem by inserting a glycine-based linker on both termini of the eGFP protein. However, the large size of the triple fusion protein combined with any N-terminal additions of endogenous trapped protein may still promote

susceptibility to structural obstructions. Alternatively the integrity of the protein's structure may be affected by the conditions of its local microenvironment such as pH and hence certain localizations of the fusion protein within a cell (potentially facilitated by 'trapped' localization signals) would result in reduction or loss of fluorescence. This 'structural alteration' model as a causative agent for the loss of fluorescence activity is supported by previous reports showing that variations in eGFP fluorescence are due to difference in structure rather than protein levels (Li et al., 1997). The development of a folding-enhanced GFP variant, called 'superfolder' GFP has been recently reported (Pedelacq et al., 2006). This shows increased resistance to fusion partner misfolding and chemical denaturants while exhibiting improved folding kinetics. 'Superfolder' GFP might prove an ideal candidate reporter for future gene trap vector design studies.

Interestingly, we also identified two clones that were *lacZ* positive but possessed a surprisingly high proportion of fluorescent cells (Figure 3.12). These clones might represent vector integrations that give rise to fusion proteins which have a detrimental effect on the enzymatic activity of β-galactosidase. It has been demonstrated that genes which encode secretory and membrane-spanning proteins cannot be tagged by conventional *βgeo* gene trap vectors (Skarnes et al., 1995). This led to the development of specialized vector designs targeting this specific class of genes (Skarnes et al., 1995; De-Zolt et al., 2006). 5'RACE PCR analysis of the *lacZ* negative, eGFP positive clone H3-17 revealed that vector integration occurred within the *erdr1* locus which encodes a protein that was shown to localize in the inner side of the cytoplasmic membrane and is likely to be secreted (Dormer et al. 2004). Hence gene trap clone H3-17 probably represents a gene trap event in which the trapped gene's expression is being detected through the

expression of the *egfp* but not the *lacZ* component of the reporter triple fusion. This finding supports the idea that the *egfp*-containing triple reporter fusion within our vectors enables the identification of gene trap events into genes whose entrapment cannot be detected with βgal-based constructs. To test this hypothesis further examination of clone H3-17 is required. This could be done, for example, through the use of an anti-βgal antibody and subsequent analysis by immunofluorescence which would enable the detection of the subcellular localization of the *egfpβhygro* fusion protein.

### 4.1.2.2 Molecular analysis of gene trap clones provides evidence for the occurrence of complex vector integration events

A small number of gene trap clones containing both the *egfpβgeo* and *egfpβhygro* fusion were analysed by 5'RACE analysis. The results of this preliminary 'molecular screen' suggest that the *En-2* splice acceptor junction included within the vectors was utilized in the predicted manner. In three cases (clones H3-10, H3-17 and 1B2) vector insertion occurred within transcriptional units (*Fbxo15, erdr1* and "expressed sequence AI506816") in the right orientation and is predicted to give rise to a chimaeric protein that is translated from an ATG present in the trapped sequence and consists of an endogenous N-terminal portion and the vector's reporter protein. Of the three gene trap events, the most likely to induce a null mutation is the disruption of the *Fbxo15* locus. The latter is a target of the pluripotency transcription factor *Oct3/4*, and its F-box-containing protein product is predominantly expressed in undifferentiated ES cells (Tokuzawa et al., 2003). The vector in this instance (as indicated by RACE product H3-10-1) integrated within the gene's first intron and the resulting fusion protein is theoretically translated in the same frame as the endogenous protein containing only 8 of the 517 aminoacids encoded by the affected gene.

We also identified an alternatively spliced *Fbxo15* fusion transcript (H3-10-2) which probably represents a previously unidentified 5'UTR exon located upstream of the gene's reported first exon. This transcript has been trapped by other gene trap groups (SIGTR gene trap cell line RRF222) and its authenticity was confirmed by RT-PCR. The regions homologous to transcripts H3-10-1 and H3-10-2 were located 547 and 109 bp respectively downstream of an 18-bp enhancer element which is activated by *Oct3/4* and *Sox2* and is required for ES cell-specific expression of *Fbxo15* (Tokuzawa et al., 2003). Targeting of the *Fbxo15* exons 3-7 which code for the F-box was not associated with any overt developmental defects (Tokuzawa et al., 2003). However, we speculate that the severity of the *Fbxo15* mutation introduced by the gene trap event in clone H3-10 is likely to be greater than the lesion generated by Tokuzawa et al. since a larger portion of the FBXO15 protein is disrupted and vector integration might also interfere with the function of the *Fbxo15* enhancer. This hypothesis can be tested by generating mutant mice using gene trap ES cell clone H3-10.

The majority of the rest of the cloned 5'RACE products were indicative of vector insertions within 5' and 3'UTR regions of genes (H3-1, *Dido*; 5C1, *Ubc;* 1B3, *Bat2d*) and non-protein coding loci (5B1, *Rnu17d*; 1A1, *Gas5*). The common feature shared by these RACE transcripts was the lack of ATG codons within the trapped sequences combined with the presence of stop codons in the same frame as the triple fusion's ATG-less coding sequence (Appendix 3). Hence these RACE transcripts are incompatible with the translation of the vector's reporter protein despite the fact that they were all obtained from clones which were strongly positive for *lacZ* and *egfp* expression. Southern blot analysis of four representative hygromycin resistant clones suggests that multiple vector integration took place in all

cases (Figure 3.17). It is therefore likely that reporter expression in clones that gave rise to stop codon-containing 5'RACE products might be facilitated by independent, translation-compatible gene trap events that occurred within the same clones. Indeed, two of the clones examined by Southern blotting (H3-1 and H3-10) were also found to generate stop-codon containing 5'RACE products.

The results of the Southern blot analysis might also be indicative of more complex insertional events such as vector/flanking DNA rearrangements and concatemer formation. These are commonly occurring events during electroporation (Friendrich and Soriano, 1991; Niwa et al., 1993; Forrester et al., 1996; Neilan and Barsh, 1999) and further optimisation of the electroporation conditions will be required to address this issue. Moreover, Southern blot analysis of the clones using probes that are specific to different segments of the vector and restriction enzymes that cut at defined sites within mouse genomic sequences flanking the site of vector integration would be useful for elucidating the exact nature of the integration events. Additionally, analysis of the gene trap clones using fluorescence in situ hybridization (FISH) together with a gene trap vector-specific probe could be employed for mapping the vector insertions and consequently reveal whether multiple vector integrations took place within distinct chromosomes or as concatemers within the same locus.

A less likely explanation for the isolation of stop codon-containing 5'RACE transcripts from reporter-expressing gene trap clones involves the initiation of reporter protein translation downstream of the stop codons and at a codon other than ATG. The existence of mRNAs whose translation starts exclusively at a non-AUG codon has been previously reported. The translational regulator p97 which initiates solely at a GUG codon is such an

example (Imataka et al., 1997). A GUG codon is present in our vector's reporter sequence and is the first codon of the *egfp* gene directly downstream of the stretch of glycines (Figure 3.1). This codon could theoretically serve as a translational initiation signal in a fashion similar to p97. Other codons have also been demonstrated to facilitate initiation of protein synthesis. The AGG and GUC codons, for example, have been shown to successfully initiate protein synthesis *in vitro* (Drabkin and Rajbhandary, 1998). Both of these codons are present in our stop codon-containing cloned 5'RACE products; GTC is the third codon in the translated splice acceptor exonic sequence (Figure 3.1) while an AGG codon was generated in all cases from the in frame splicing of the last two bases (AG) of the upstream endogenous trapped sequences and the first G base present into the vector's splice acceptor coding leader sequence (see Appendix 3). It should be noted that similar 5′ RACE products that include in frame stop codons have also been obtained by other gene trap groups that employ ATG-less, *βgeo*-type gene trap vectors and therefore identification/selection of the corresponding clones was, theoretically, dependent on active translation of the *βgeo* reporter (e.g. SIGTR cell lines AJ0170, AJ0182, AM0414, AR0145). Many of these transcripts are indicative of vector insertions within non-protein-coding transcriptional units (e.g. Rnu22, Rnu87, Mmu-mir-350 and 28S rRNA).

### 4.1.3 Characterisation of the 3'poly (A) trap cassette

#### 4.1.3.1 3' RACE analysis reveals a high proportion of SD read-through events

Processing of pre-messenger RNAs (pre-mRNAs) transcribed by RNA polymerase II (Pol II) begins with the capping of their 5′ end. This involves the removal of the 5′ triphosphate of the pre-mRNA and the addition of

guanosine monophosphate which is subsequently methylated to produce m7GpppN (for a review see Shatkin and Manley, 2000). The 5′ cap serves as a binding site for the heterodimeric nuclear cap binding complex (CBC) (Visa et al., 1996). Both the 5′ cap and CBC mark the 5′ end of the transcript's first exon. Capping is followed by splicing during which introns are removed and exons are ligated together through the action of a multi-component ribonucleoprotein (RNP) complex called the spliceosome. A 5′ splice site is usually characterised by the consensus sequence $^{-2}$AG*GUPuAGU$^{+6}$ (the star indicates the exon-intron junction) and defines the 5′ boundary of most introns (Goldstrohm et al., 2001). The 3′ end of introns is defined by the branch point/polypyrimidine tract which are located several basepairs downstream of the 5′ splice site and close to the 3′ splice site ($^{-4}$NPyAG*PuN$^{+2}$). The 3′ splice site, in turn, marks the upstream boundary of the next exon in the transcript. Internal exons are located upstream of 5′splice sites while 3′ terminal exons are characterised by the presence of a polyadenylation signal.

The spliceosome's components "associate in a step-wise manner with pre-mRNA" (Tazi et al., 2005); 5′site recognition occurs through base-pairing with U1snRNP and is influenced and stabilised by several General Splicing Factors (GSFs) that are members of the SR (serine-arginine rich) family of proteins (Kohtz et al., 1994). In 5′exons 5′splice site recognition is enhanced by the interaction of the splicing machinery with the CBC at the cap structure. This interaction defines the first exon of the transcript (Figure 4.1a) (Robberson et al., 1990). Splicing Factor 1 (SF1) then binds to the branch point (Arning et al., 1996; Berglund et al., 1997) within the adjacent intron and interacts with the 65 kDa subunit of the heterodimeric factor U2AF which also binds the polypyrimidine tract (Berglund et al., 1998). The 35 kDa

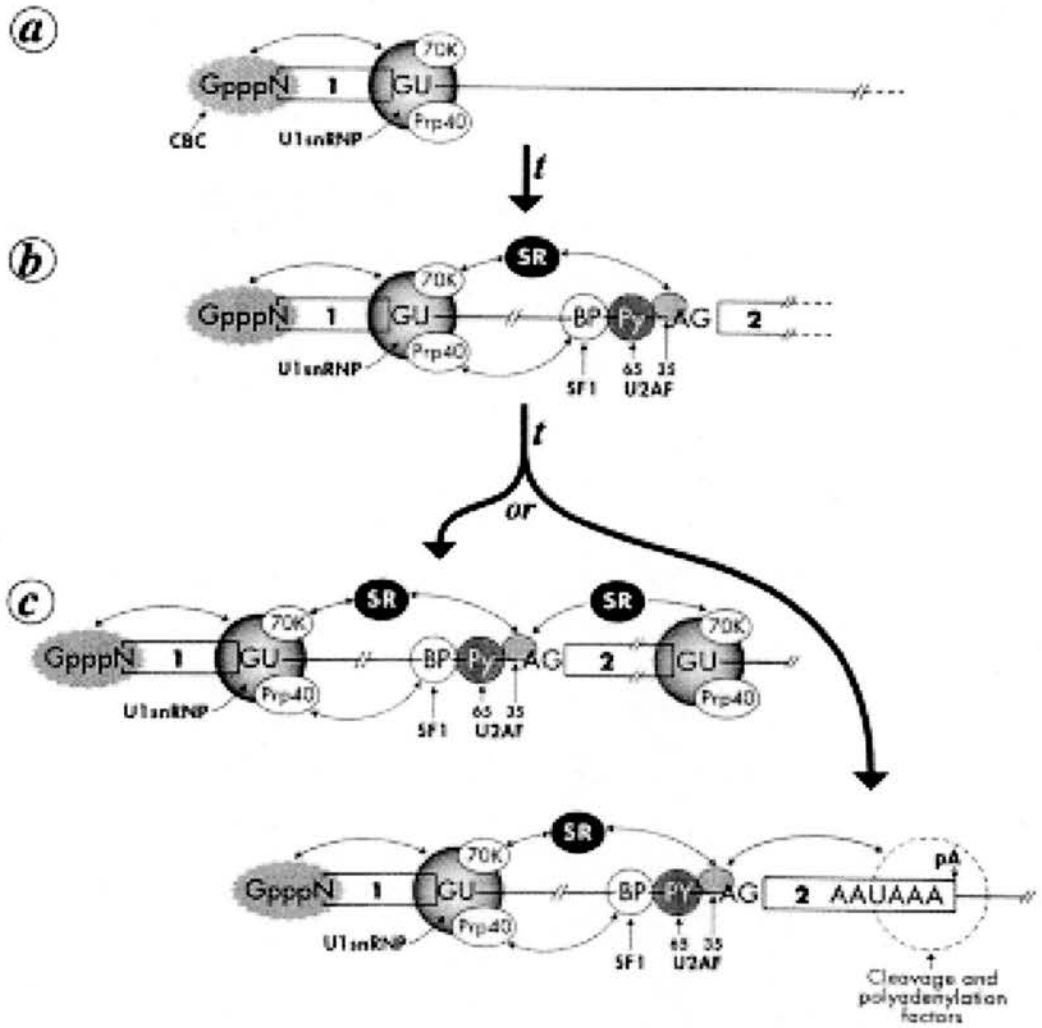



**Figure 4.1** Protein–protein interactions across exons and introns mark the splice sites in pre-messenger RNAs. **(a)** Early in the life of a nascent pre-messenger RNA, the cap-binding complex (CBC) interacts with factors assembled on the 50 splice site (indicated by the highly conserved GU dinucleotide).**(b)** At a later time (t), once the 30 splice site has emerged from the elongating RNAPII, cross intron interactions can be seen. U1 snRNP components, U1-70K protein (70K) and Prp40/FBP11, can interact with SF1 and U2AF on the branch point (BP), polypyrimidine tract (Py) and 30 splice site (indicated by the highly conserved AG dinucleotide). **(c)** At an even later time one of two scenarios are possible: a new downstream 50 splice site defines an internal exon or a downstream polyadenylation signal (pA) defines a terminal exon. In both cases cross-exon interactions are noted (Figure obtained from Goldstrohm et al., 2001).

subunit of U2AF binds the invariable AG dinucleotide at the intron's 3' end (Guth et al., 1999; Wu et al., 1999). Hence the U1snRNP complex at the 5'splice site and the SF1-U2AF complex at the branch point/polypyrimidine tract/3'splice site mark the borders of the intron. Furthermore, these complexes interact with each other either directly (Bedford et al., 1998) or indirectly via SR proteins (Wu and Maniatis, 1993) (Figure 4.1b). The SF-1-U2AF complex also interacts via SR proteins (Wu and Maniatis, 1993) with the U1snRNP/GSF complex at the next downstream 5'splice site (Figure 4.1c) and this association defines the intervening exon as an internal one (Robberson et al., 1990). Additionally, exon definition is influenced by the interaction of SR proteins with exonic splicing enhancers (ESEs) which are cis-acting sequence elements that facilitate the inclusion of an exon (Blencowe, 2000). 3'terminal exons are defined by the interaction of the 3'splice site factors with the protein complexes that recognise and bind to the cleavage and poly(A) signals (Figure 4.1c) (Niwa et al., 1990; Niwa and Berget, 1991; Vagner, 2000).

Splice site recognition is then followed by the ATP-dependent binding of U2snRNP to the branch point and spliceosome assembly is completed through the employment of the tri-RNP U4-U5-U6 (Tazi et al., 2005). Dissociation of the U1 and U4 snRNPs through the action of DExD/H box proteins renders the spliceosome catalytically active. This state is accompanied by several spliceosome rearrangements that involve the U2, U5 and U6 snRNPs and which eventually trigger via two *trans*-esterification steps intron excision and ligation of the adjacent exons (Tazi et al., 2005; Goldstrohm et al., 2001).

A high proportion of neomycin resistant clones carrying predominantly insertions with the HindIII/MfeI-digested pEHygro2neoSD2

(-ARE) construct gave rise to 3′ RACE products in which no SD use took place. The majority of these products consisted of unspliced *βglobin* SD sequence that included various portions of the *βglobin* intron 2 and a cryptic poly(A) signal followed by a stretch of A's. The generation of such transcripts provides an interesting insight into the mechanisms of splice site recognition and exon definition; in these cases the consensus *βglobin* exon 2/intron 2 junction is ignored by the splicing machinery and the retained intronic segment is recognized as part of a terminal exon as the transcript's 3′end formation occurs via the employment of cryptic poly(A) signal sites residing within the intron. This might be due to the fact that the sequence elements (branch point, polypyrimidine tract and 3′splice site) that define the intron's 3′ boundary are removed upon MfeI digestion. This would, in turn, limit the access of the 3′splice site factors (SF-1, U2AF and SR proteins) required for the recognition of the *βglobin* intron's 5′and 3′ boundaries and lead to increased levels of impairment of the consensus exon/intron splice junction. It should be noted proper SD use still occurs as evidenced by the generation of a small number of properly spliced 3′RACE transcripts (Table 3.3) but at a lower level. In this case the efficient employment of the *βglobin* SD might depend on the sequence context of endogenous intronic sequences that are present downstream of the vector's integration site; these sequences might potentially function as recognition sites for docking of the appropriate factors that facilitate 3′splice site recognition.

#### 4.1.3.2 Inclusion of the ARE enhances the performance of the *βglobin* SD

Comparison of the 3′RACE results obtained from clones electroporated with the + and -ARE versions of vector pEHygro2neoSD2 (N=45 and 43 respectively) indicates that the inclusion of the ARE improves

the efficiency of the *βglobin* SD. We observed a reduction (approximately 1.7-fold) in the number of neomycin resistant clones compared to the number obtained when the construct without the ARE was used. This result could potentially reflect a reduction in the fraction of clones that acquired neomycin resistance through read-through events and employment of cryptic poly(A) signal sites and therefore a decrease in the 'background'. 3'RACE PCR analysis further supports this hypothesis; the percentage of 3'RACE products from pEHygro2neoSD2 (+ARE)-containing clones that are indicative of vector read-through was considerably reduced (approximately 4.3-fold; Table 3.3) compared to RACE transcripts from –ARE clones. This reduction in 'background' was accompanied with an enhancement in the performance of the vector's rabbit *βglobin* splice donor as in the 77% (7-fold increase compared to the –ARE vector) of the clones examined the splice donor's splice junction was used in the predicted manner (Table 3.3).

We postulate that the presence of the ARE within the splice donor's *βglobin* intron results in the degradation of mRNA species arising from read-through events. In the majority of the –ARE vector-containing clones these read-through transcripts probably become stabilised through the employment of cryptic poly(A) signal sites present within the SD intron (see also previous section 4.1.3.1). Indeed, the presence of such poly(A) addition motifs was observed in most of the RACE products that represented read-through events and these transcripts were also found to include a poly(A) tail. A further clue comes from the fact that the 3'RACE products obtained from pEHygro2neoSD2+ARE-electroporated clones and in which read-through was observed contained only a small portion of the ARE sequence, probably as a result of the element's destabilisation occurring in these cases. This implies that degradation of the ARE sequence is linked to an increase in

the incidence of read-through events. An experimental setting which also includes electroporation with a vector which is identical to vector pEHygro2neoSD2 (+ARE) but contains a scrambled sequence cloned into the *βglobin* SD instead of the ARE would be an ideal control for assessing whether the improvement in the SD function is indeed specifically linked to the presence of the destabilising element.

It would also be interesting to test whether this positive effect exerted by the ARE is SD sequence-specific i.e. restricted only to an enhancement in the performance of the specific *βglobin* splice donor used in this study or whether it could also improve the performance of other splice donor sequences. A study involving the testing of the RET poly(A) trap vector which incorporates a similar ARE in conjunction with the mouse *hprt* exon 8/intron 8 SD also indicated that this sequence element might have a positive effect on a poly(A) trap vector's performance; it was shown that a reduction (2-fold) in the number of drug-selected, +ARE vector-containing clones occurred compared to those containing the –ARE vector (Ishida and Ledder, 1999). However, the above comparison was not combined with 3'RACE data and hence our study is to our knowledge the first demonstration of the fact that an mRNA destabilising element can enhance the efficiency of a poly(A) trap vector's SD.

#### 4.1.3.3 Gene trapping efficiency of pEHygro2neoSD2 (+/-ARE) vectors is similar to other poly(A) trap vectors

A big proportion of the properly spliced 3'RACE products derived from clones electroporated with vectors pEHygro2neoSD2 and pEHygro2neoSD2 (+ARE) contained a poly(A) signal site and a poly(A) tail. This suggests that our vectors functioned appropriately as poly(A) trap vectors. BLAST analysis of the transcripts shows that our vectors target genes

at a comparable rate to other poly(A) trap vectors; 40% of the trapped sequences matched exons of known genes and EST transcripts (this percentage was found to be 49.5% by Matsuda et al., 2004 and 48% by Osipovich et al., 2005) (Figure 3.21).

22% of the RACE products were found to be homologous to ab-initio predicted GENSCAN transcripts and many of them had a splicing pattern i.e. different segments of their sequence were found to be homologous to different adjacent genomic regions indicating the trapping of different exons. Moreover, one transcript (2%) did not align with any known/predicted transcript but also exhibited a splicing pattern. Some of these RACE products could represent artefacts/splicing to sequences capable of functioning as cryptic 3'exons. However, it is likely that at least a fraction of them corresponds to genuine, novel functional genes. This speculation is supported by the fact that many of the sequences belonging to this group were found to be highly conserved between different species. Furthermore, Matsuda et al (2004) employed similar poly(A) trapped sequences (i.e. not found to be homologous to any known genes/ESTs) for the construction of DNA arrays and showed that some of them were indeed expressed. Further assessment of these transcripts is required (for example by analysing their expression by RT-PCR/*in situ* hybridisation in different tissues/developmental stages) in order to elucidate whether they correspond to previously unidentified genes.

A considerable proportion (30%) of the 3'RACE products corresponded to a sequence from the 3'LTR of the ETn family of retrotransposon elements. This sequence appears to be a trapping target for other poly (A) trap vectors too. ETn elements were first discovered and characterised in 1980s (Brulet et al., 1983; Brulet et al., 1985). They represent a

family of middle repetitive sequences which are transcribed during early mouse embryogenesis in ES and embryonic carcinoma (EC) cell lines (Brulet et al., 1985) and possess retrotransposon-like properties (Mager and Freeman, 2000; Baust et al., 2003). We speculate that the high fraction of vector integrations within these elements might be attributed to the presence of a potent splice acceptor and poly(A) signal site in the 3'LTR of this highly dispersed in the mouse genome (200 copies; Tanaka and Ishihara, 2001) family of sequences, which facilitates splicing events between the vector's splice donor and the ETn LTR's splice acceptor. Interestingly, a significant number of germline and somatic cell mutations in mouse cell lines have been reported to be caused by insertions of ETn elements into genes (Baust et al., 2002; Baust et al., 2003) and in many cases the loss of gene function was due to aberrant transcripts as a result of ETn-induced alternative splicing. Furthermore, some functional genes have been shown to contain ETn LTR sequences within their 3'UTR (an example was reported by Tsuiji et al., 2002) and hence a proportion of the ETn LTR integrations might represent insertional events within transcriptional units that have incorporated these repetitive elements in their structure by transposition.

#### 4.1.3.4 Poly(A) trapping using the pEHygro2neoSD2 constructs appears to be immune to NMD

Examination of the vector insertion sites within transcripts with well documented exon/intron structure indicates that our vector does not appear to suffer from the bias exhibited by other poly(A) trap vectors to integrate within the last intron of their target genes. However, a greater number of clones must be analysed in order to confirm this observation. It is also essential to demonstrate by Southern blotting and the use of a neo-specific probe that the 3'RACE transcripts we obtained result from single vector

integrations as it is possible that neomycin resistance of the clones analysed arises from different, independent insertion events if more than one vector copies are present within the same clones. Data from a preliminary Southern blot analysis (using an EGFP-specific probe) of four neomycin resistant clones suggest that in some cases multiple vector integrations might have occurred (Figure 3.26 and Section 4.1.3.6). However, analysis of the neomycin resistant clone 3C4 yielded a single band, a finding that could be indicative of a single vector insertion. 3'RACE analysis of the same clone resulted in an NMD-resistant transcript that represents an insertion event into the first intron of the *Cds2* gene (13 exons in total) (Tables 3.5 and 3.6). This suggests that the neomycin resistance of this clone is generated by an "unbiased splicing event" stemming from a single vector integration within the Cds2 locus.

Another piece of evidence that indicates that poly(A) trapping using the pEHygro2neoSD2 vectors is not subject to NMD is the fact that we successfully generated through 3'RACE PCR *bona fide neo* fusion transcripts derived from clones electroporated with vectors pEHygro2neoSD2 and pEHygro2neoSD2 (+ARE); these fusion products should, theoretically, have been degraded by NMD since the termination codon (TC) of the pEHygro2neoSD2 vector's *neo* coding sequence is located 362 bp upstream of the splice donor's splice junction, a distance that significantly exceeds the "60 nt from the last exon-exon junction" limit for evasion of NMD and would therefore be recognised as a potentially premature termination codon.

The reasons behind the potentially unbiased insertional behaviour of our vectors are unknown. We speculate that the 'distance-relative-to-a-downstream-exon-exon-junction' rule for eliciting NMD might be broken due to the presence of a sequence or sequences located downstream of the

*neo's* termination codon (TC) and within the rabbit *βglobin* SD used. The existence of such stabilising elements that confer resistance to NMD has been reported for the general control norepressible (GCN4) and yeast AP-1 (YAP-1) mRNAs of *S. cerevisiae* (Maquat et al., 2004; Ruiz-Echevarria and Peltz, 2000). More interestingly, nonsense mutations engineered to introduce premature termination codons (PTCs) within the human *βglobin* exon 1 have been shown to unexpectedly fail to elicit NMD, even though intron 1 is more than 55-60 nt downstream (Romao et al., 2000). Alternatively, additional, unidentified *cis*-acting sequence determinants that facilitate NMD might exist and their presence in the human *βglobin* gene has been speculated (Danckwardt et al., 2002). It is likely that such sequence elements reside in the portion of the rabbit *βglobin* exon 2 that was deleted in our vector's splice donor since the latter contains only the last 29 nt of exon 2. A further piece of evidence that points to a potential link between the specific rabbit *β-globin* splice donor used in our study and NMD immunity comes from the fact that some cases of β-thalassemia are associated with nonsense mutations that introduce PTCs within the human *βglobin* gene but do not activate an NMD response resulting in the accumulation of abnormal, stable NMD-resistant transcripts (Lapoumeroulie et al., 1986; Hall and Thein, 1994). It might be useful to test the insertional behaviour of different mutated versions of our *βglobin* SD in order to determine whether the presence or absence of any sequence elements is associated with resistance to NMD. Furthermore, the employment of different splice donor controls in a poly(A) trapping context could provide an insight into whether immunity to NMD is indeed linked to the specific *βglobin* SD incorporated in our poly(A) trap vector.

#### 4.1.3.5 The pEHygro2neoSD2 (+/-ARE) poly (A) trap vectors possess the potential to capture developmentally regulated genes

Many of the known genes trapped by our poly(A) trap vectors have been reported to be developmentally regulated with different functions and expression profiles, indicating that the pEHygro2neoSD2 constructs could be potentially used as tools for entraping this class of genes. RT-PCR expression analysis and data from published reports/EST databases showed that all of the known trapped genes are expressed in undifferentiated ES cells. Clearly, a larger number of clones need to be analysed in order to determine whether our poly(A) trap vectors disrupt genes that are not expressed in undifferentiated ES cells. 7/8 of our trapped genes have already been captured by SA-type vectors indicating that both poly(A)- and promoter-type gene trap strategies roughly target the same fraction of the mouse genome. However, one of the pEHygro2neoSD2-trapped genes (*Phlda2*) has only been previously trapped through the use of a poly(A) trap vector employed by another gene trap group (CMHD).

#### 4.1.3.6 Poly(A) trap vectors can disrupt genes that are not accessible to entrapment by other gene trap vectors

To further test whether poly(A) trap vectors target a set of genes that cannot be trapped by SA-based gene trap constructs we examined all gene targets of poly(A) trap vectors within the IGTC database. We found that, to date, poly(A) trap vectors have trapped 1025 genes which correspond to 6.2% mouse genome coverage (the total number of IGTC-trapped ENSEMBL genes, to date, is 6644 representing 40.36% of genome coverage). Half of these genes (513 or 50%) have been trapped exclusively by means of poly(A) trapping (Figure 3.23a). However, analysis of the vector integration sites for

180 of the 513 poly(A) trapping-specific genes that contain more than two exons and whose exon/intron structures are well defined revealed that the majority of vector insertions (approximately 75%) occurred within the 3'-most intron of the genes (Figure 3.25). Many of the poly(A) trap-specific genes are lineage- and tissue-restricted and have not been shown to be expressed in ES cells (based on expression data obtained from the Mouse Genome Informatics website: http://www.informatics.jax.org/). Interestingly, a proportion (27/513) of the genes mutated exclusively by poly(A) trapping (e.g. *Phlda2, Wdr74, Rhox6, Rab3gap1*) appear to be expressed in undifferentiated ES cells. These observations suggest that a fraction of mouse genes can only be disrupted through the use of poly(A) trap constructs. The differences in the integrational preferences between poly(A) trap and conventional gene trap vectors are likely to reflect the differences in their basic design; drug selection in poly(A) trap vectors is driven by a constitutive promoter and depends on the proper functioning of a SD sequence and 'capture' of a downstream poly(A) signal while the activation of the reporter/selector gene in SA-type vectors relies on the function of a SA sequence and the activity of upstream endogenous regulatory/promoter elements. These differences also indicate that the combined employment of both trapping systems might be beneficial for mutating different, complementary fractions of the mouse genome.

#### 4.1.3.7 Clones trapped by the pEHygro2neoSD2 vectors do not show reporter expression

No β-galactosidase positive or hygromycin resistant gene trap cell lines were detected after analysis of reporter expression by X-gal staining or switch to hygromycin selection of $neo^r$, pEHygro2neoSD2 (+ARE)-electroporated clones. This finding does not agree with previous studies

demonstrating that a fraction of poly(A)-trapped clones (2.8-22.7% depending on the study) always exhibit β-galactosidase activity in the presence of LIF (Yoshida et al., 1995; Salminen et al., 1998; Hirashima et al., 2004). Southern blot analysis of four neomycin resistant clones using an EGFP-specific probe indicates that complex vector insertional events might have taken events (Figure 3.26). These events might be associated with vector/DNA rearrangements as well as vector deletions and they are likely to negatively affect the functionality of the 5'reporter cassette and therefore be the cause for the lack of reporter expression.

The generation of 3'RACE products consisting of a sequence that is homologous to the *neo* coding sequence followed by mouse genomic sequences of different length and chromosomal origin (Sections 3.3.2 and 3.3.3) might also indicate a propensity towards 3'vector deletions. It is likely that these 3'RACE sequences are derived from clones in which the introduced plasmid vector was subjected to degradation by exonucleases resulting in the deletion of the poly(A) trap component's SD while the integrity of the upstream *neo* gene remained unaffected. We postulate that this increased sensitivity to degradation exhibited by our vectors is linked to a reduced structural integrity induced by their double digestion with HindIII/MfeI prior to electroporation into ES cells. Further optimisation of the electroporation protocol might be beneficial in addressing the issue of multiple and complex insertional events. Furthermore, the construction of a retroviral version of vector pEHygro2neoSD2 (+ARE) is also underway since retroviral delivery ensures the integration of an intact single copy of the vector.

## 4.2 Future directions

Our overview of gene targets of poly(A) trap vectors indicates that poly(A) trapping may enable the disruption of genes that cannot be trapped by other means. However, the biased insertional behaviour of poly(A) trap constructs compromises their mutagenic capacity.

Our results suggest that the pEHygro2neoSD2 (+/-ARE) vectors are not biased towards integrations into the last introns of their target genes and therefore can be the ideal vehicles for poly(A) trapping experiments. They could also constitute the basis for the design of the 'optimal' poly(A) trap vector. The 5'cassette of such a vector would ideally include a strong consensus SA sequence such as the widely used mouse *En-2* SA, which is present in our constructs, or the intron 2/exon 3 splice junction from the human *Bcl-2* gene. The latter is considered to be relatively resistant to alternative splicing and it has been shown to function efficiently in a poly(A) trapping context (Ishida and Ledder, 1999). The 5'cassette would also include an easily assayable and sensitive 5'reporter gene such as a fluorescent marker (e.g. Venus) to allow monitoring of trapped gene activity in living cells. The reporter could be fused, for example, to a drug resistance marker such as puromycin or hygromycin. This coupling would be particularly useful for lineage selection and purification of trapped cell populations. The inclusion of an IRES sequence upstream of the vector's reporter/selector gene would also be beneficial as it would alleviate the requirement for an in-frame translational fusion between the reporter and the trapped endogenous polypeptide. An alternative route to achieve multicistronic expression of different reporter proteins (e.g. Venus and puromycin/hygromycin) could involve the employment of viral "2A peptides" such as the T2A peptide from the insect virus *Thosea asigna* (TaV) (Donnelly et al., 2001) that would link the reporter and selector genes (Szymczak et al., 2004; Bill Skarnes, personal

communication). Preliminary data indicate that this approach increases reporter sensitivity and efficiency in a gene/targeted trapping context (Bill Skarnes, personal communication).

The 3'cassette of the 'optimal' poly(A) would consist of a promoter-driven neomycin resistance gene linked to a downstream efficient SD. Such a cassette should be characterized by an unbiased insertional behaviour and the *β-actin* promoter-*neo*-βglobin exon 2/intron 2 SD (+ARE) cassette incorporated in our vectors could potentially meet this criterion. Alternatively, the poly(A) trap cassette (RNA polymerase II promoter-*neo*-IRES-*hprt* exon 8/intron8 SD) of the UPATrap poly(A) trap vector could be employed as it has been shown to facilitate unbiased poly(A) trapping that is not subject to NMD (Shigeoka et al., 2005). The best method for introducing this 'optimal' poly(A) trap vector into ES cells would be via retroviral infection since this approach ensures the insertion of single, intact vector copies.

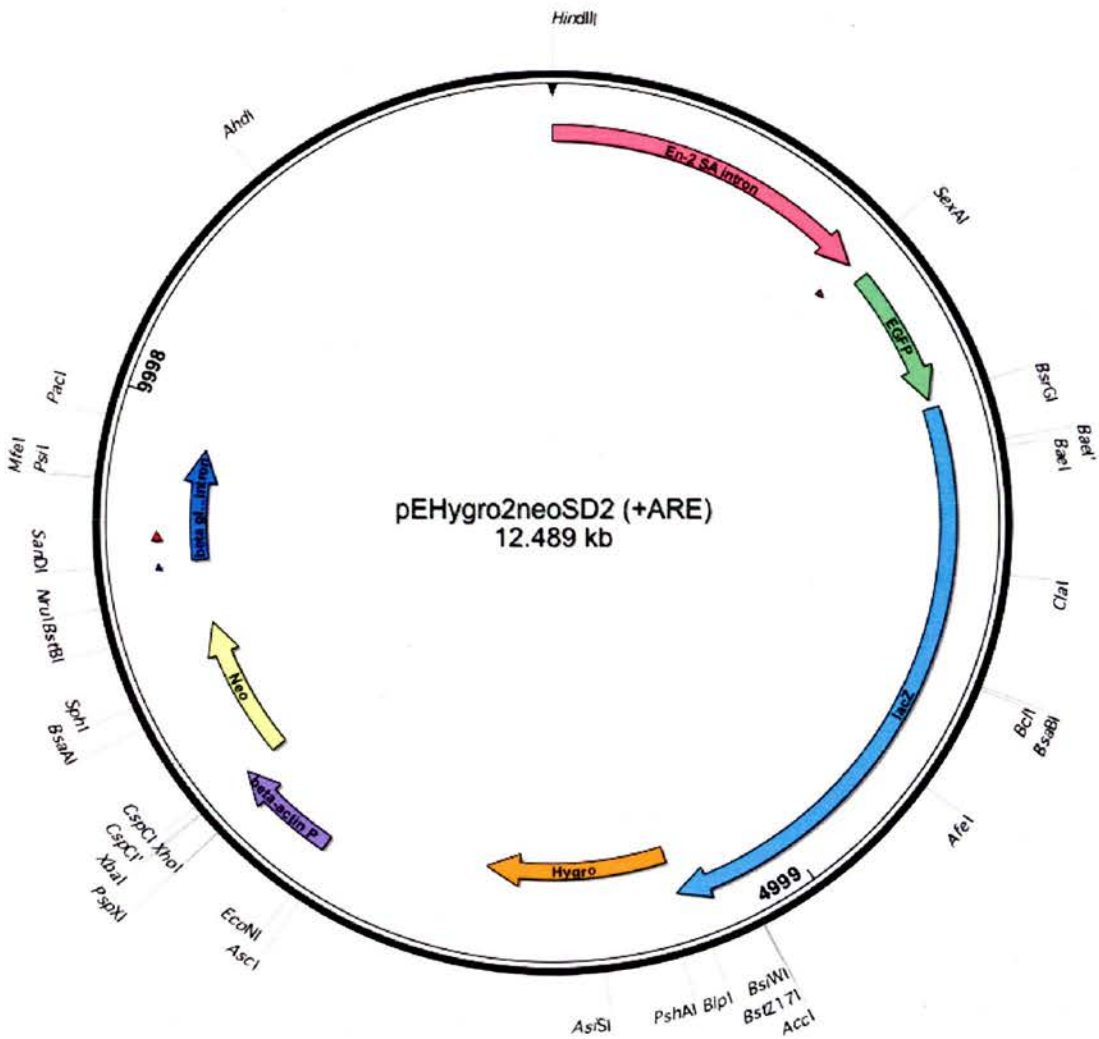
A highly efficient poly(A) trap vector could be then employed in large-scale mutagenesis projects and in conjunction with conventional SA- and promoter-type vectors. This combination of trapping strategies might prove to be the best route for improving the trapping rate of new loci (poly(A) trapping can be potentially used for the disruption of non-protein coding transcriptional units such as microRNAs, an important class of genes that cannot be captured through the use of SA-based constructs) and eventually achieving genome saturation. Moreover, the 'optimal' poly(A) trap vector would be the ideal tool for conducting expression/induction gene trap screens (e.g. through the use of the OP9 co-culture system for induction of haematopoietic differentiation) to identify developmentally regulated genes whose expression is induced upon *in vitro* differentiation. Another

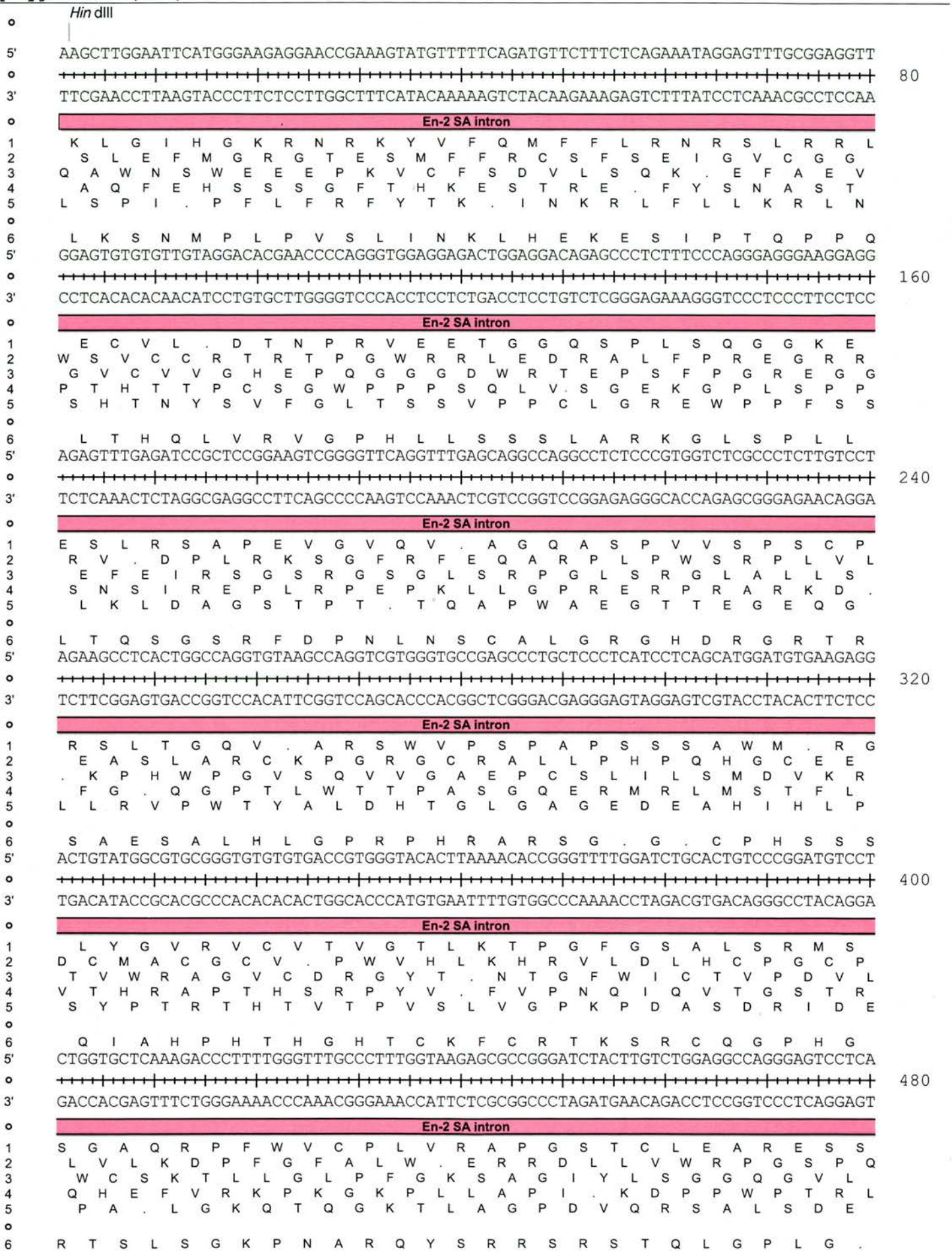
potential use could involve the adoption of the poly(A) trap vector in a "targeted trapping" context through the addition of homology arms (Friedel et al., 2005); this approach would theoretically allow the targeting of loci that are expressed at levels below the threshold (1% of the transferrin receptor gene) required for disruption through the use of a SA-type vector. However, it is important to establish that poly(A) trap vectors can function as effective mutagens and hence a thorough characterisation of their mutagenic capacity through the generation of transgenic mice from "trapped" ES cell clones is still required.

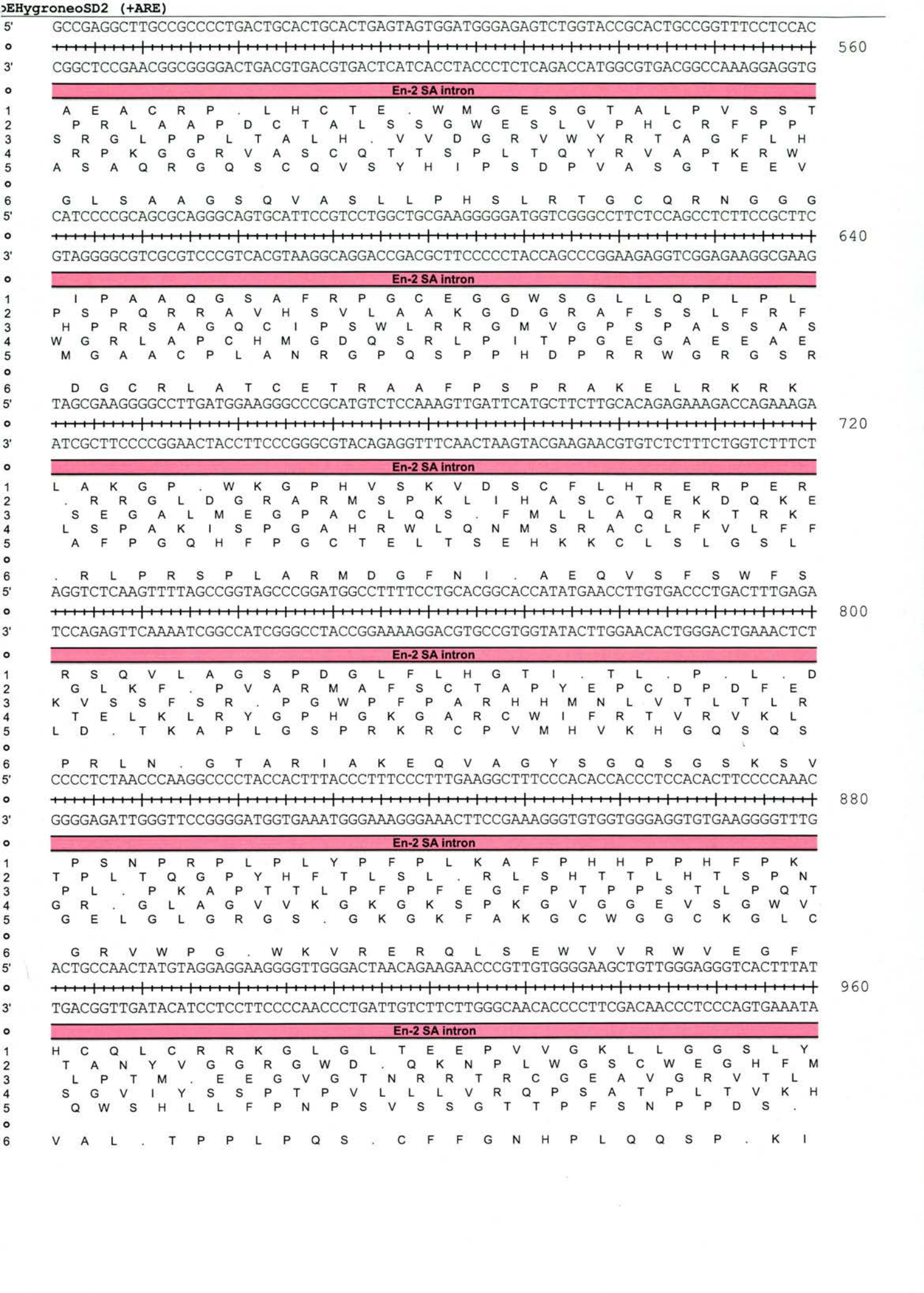
# APPENDICES

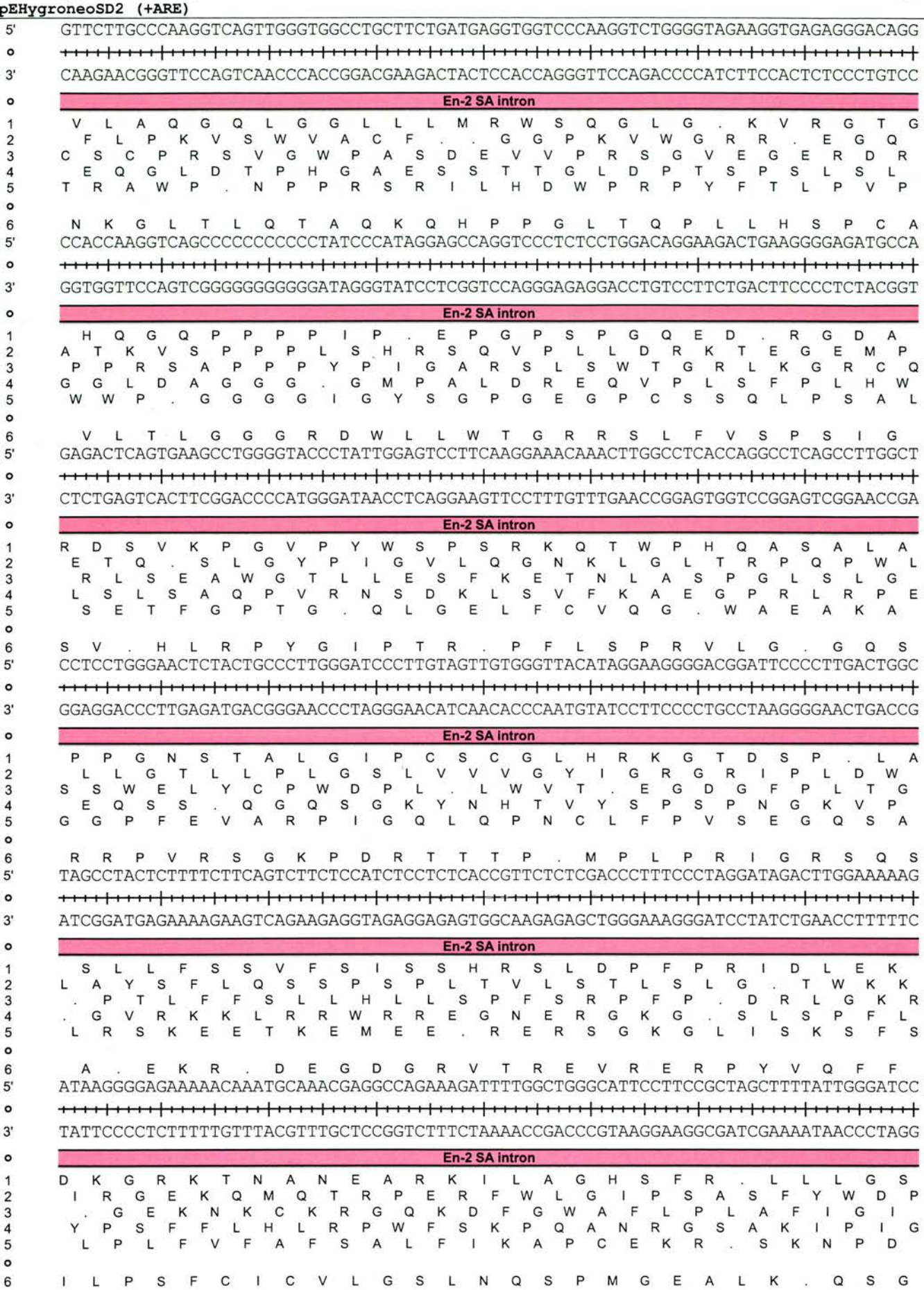
# APPENDIX 1

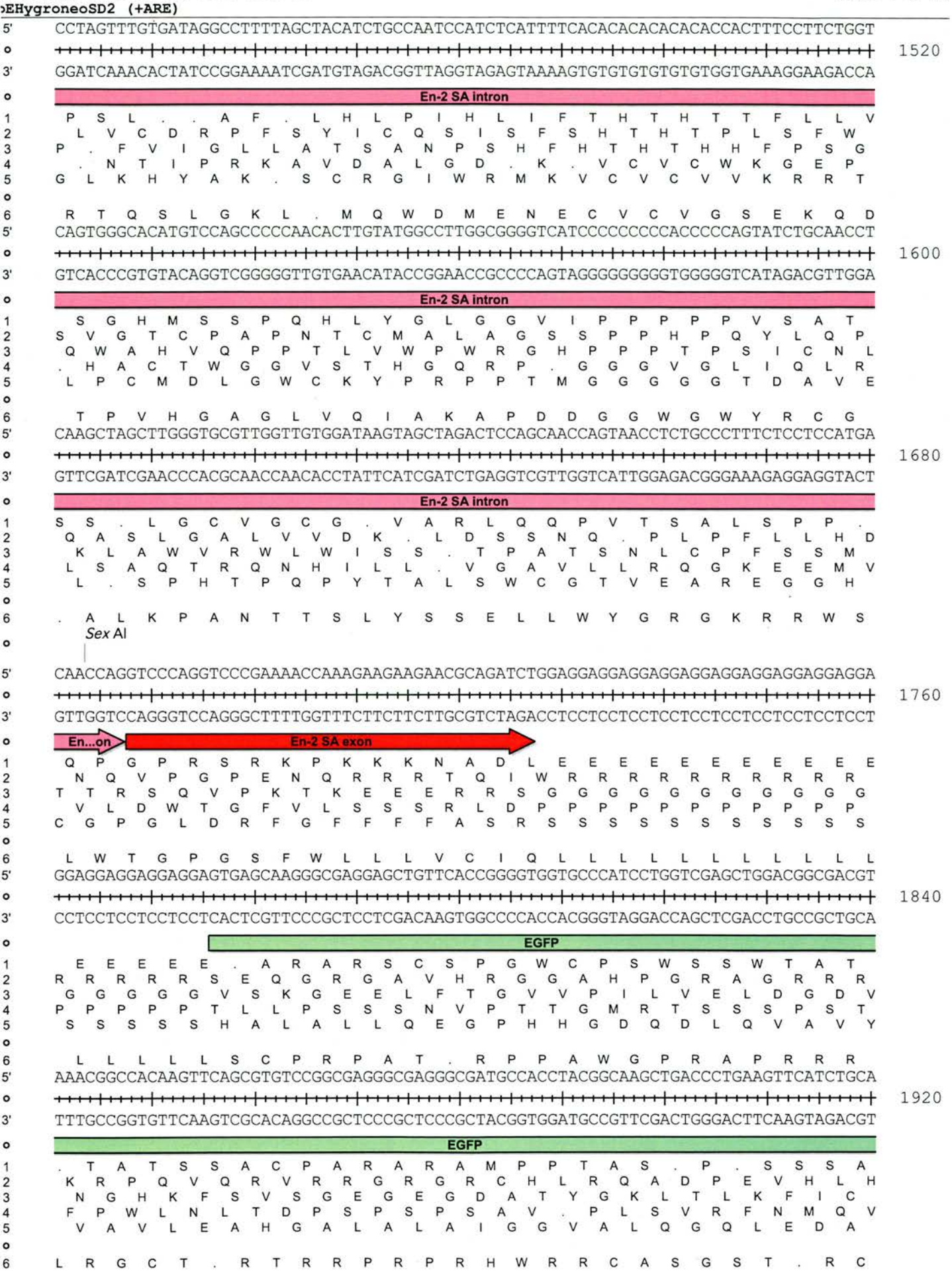
## Map and Sequence of pEHygro2neoSD2 (+ARE) poly(A) trap vector

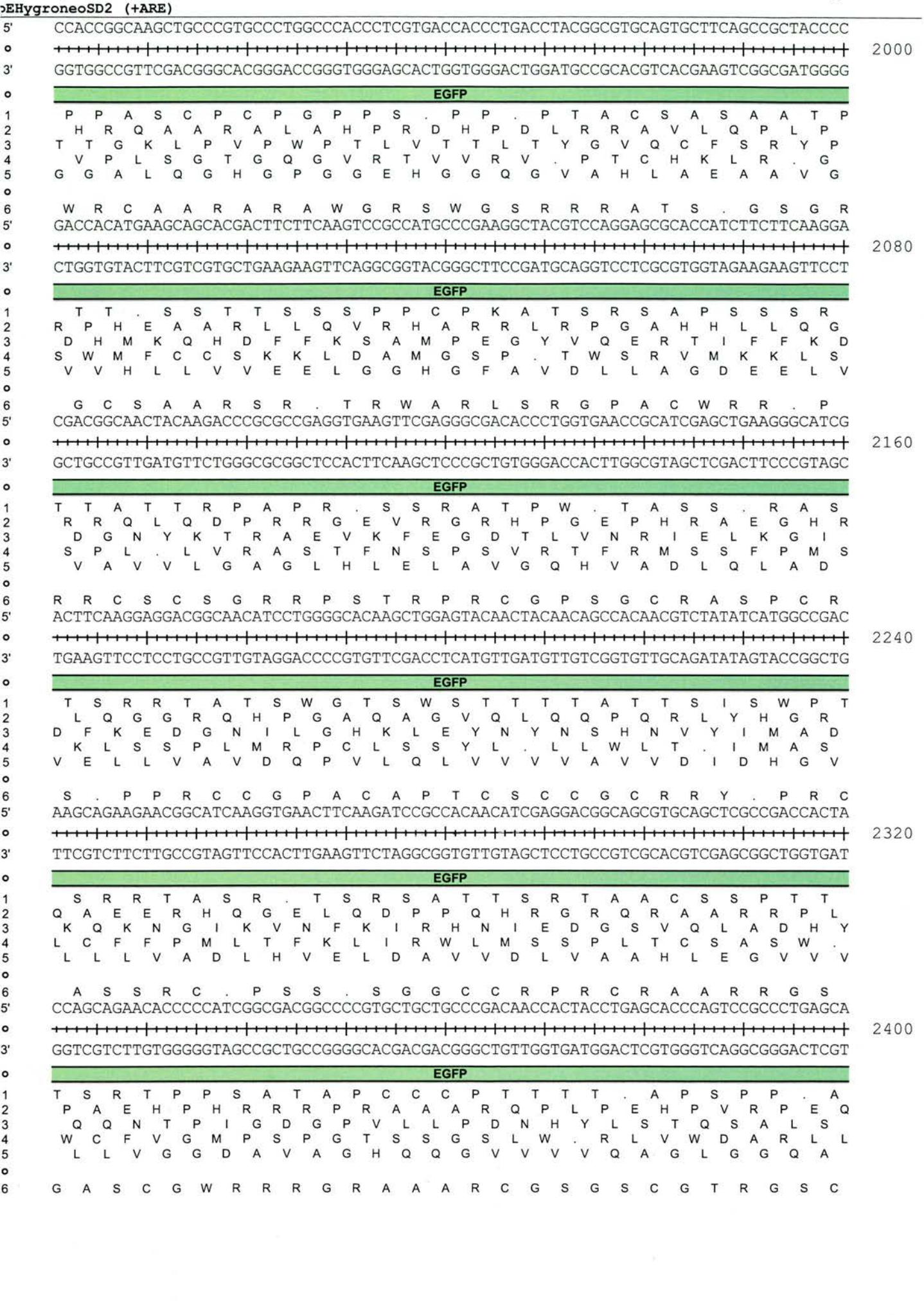


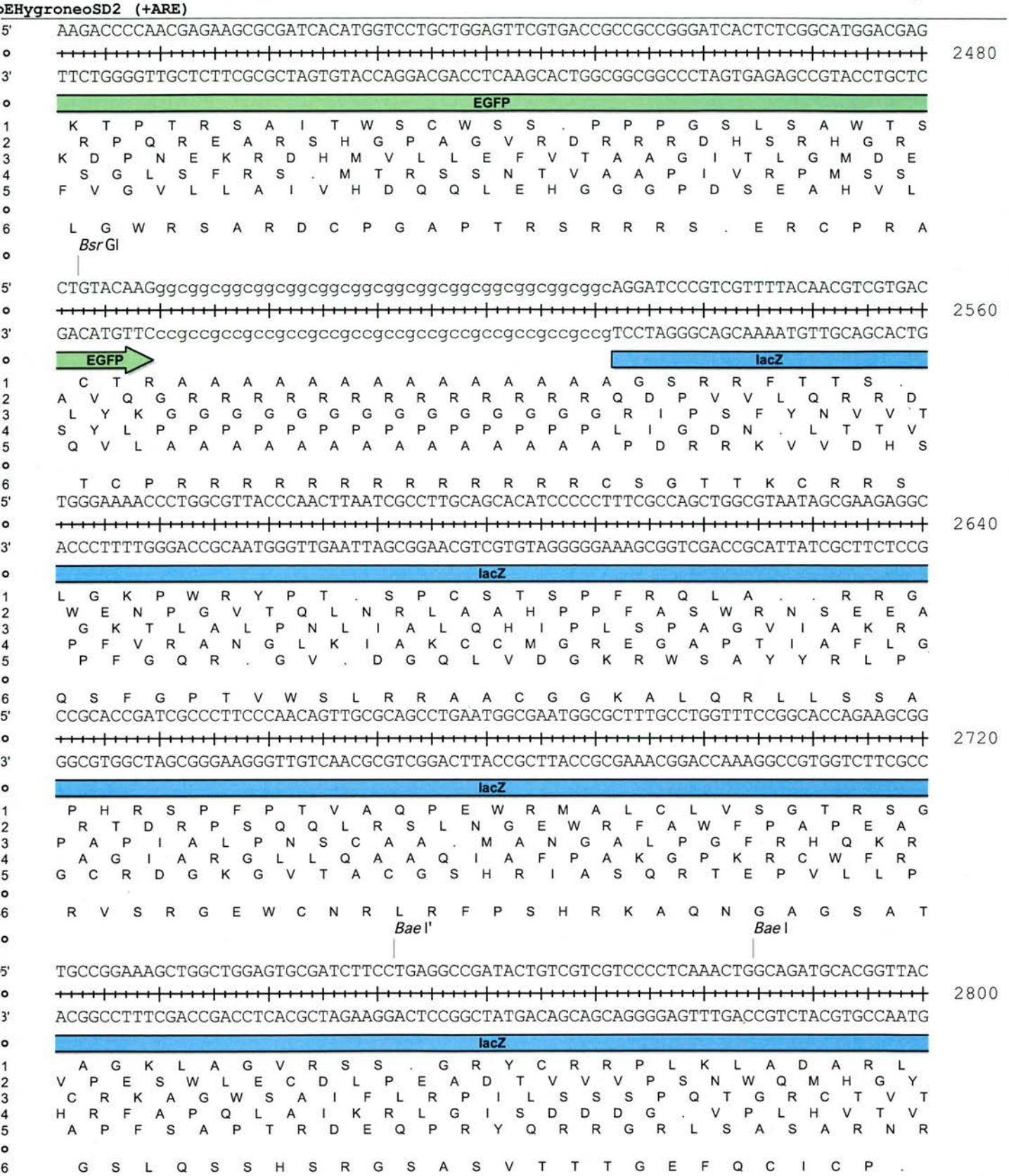

Hin dIII
En-2 SA intron
En-2 SA intron
En-2 SA intron
En-2 SA intron
En-2 SA intron
En-2 SA intron

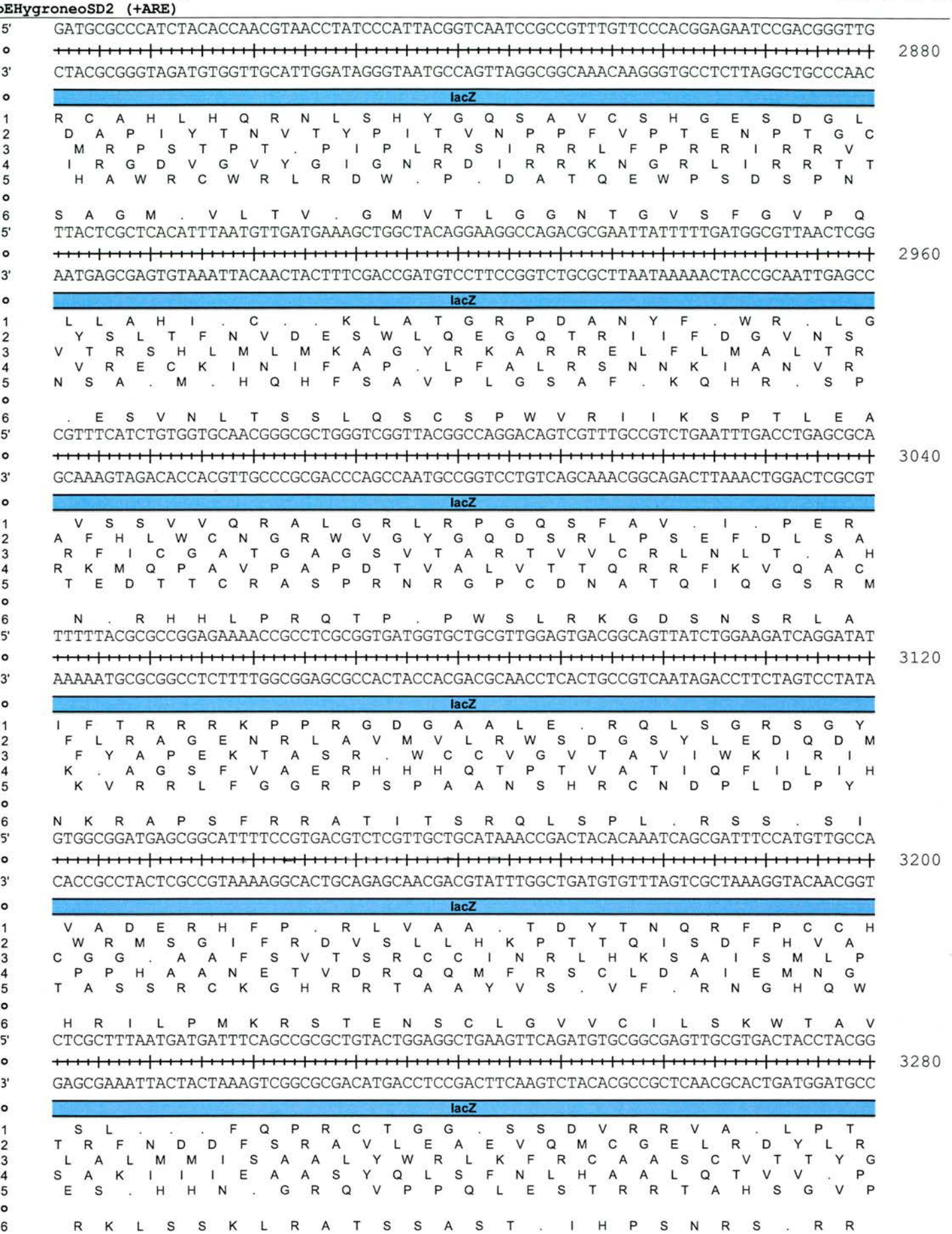

En-2 SA intron
En-2 SA intron
En-2 SA intron
En-2 SA intron
En-2 SA intron
560
640
720
800
880
960

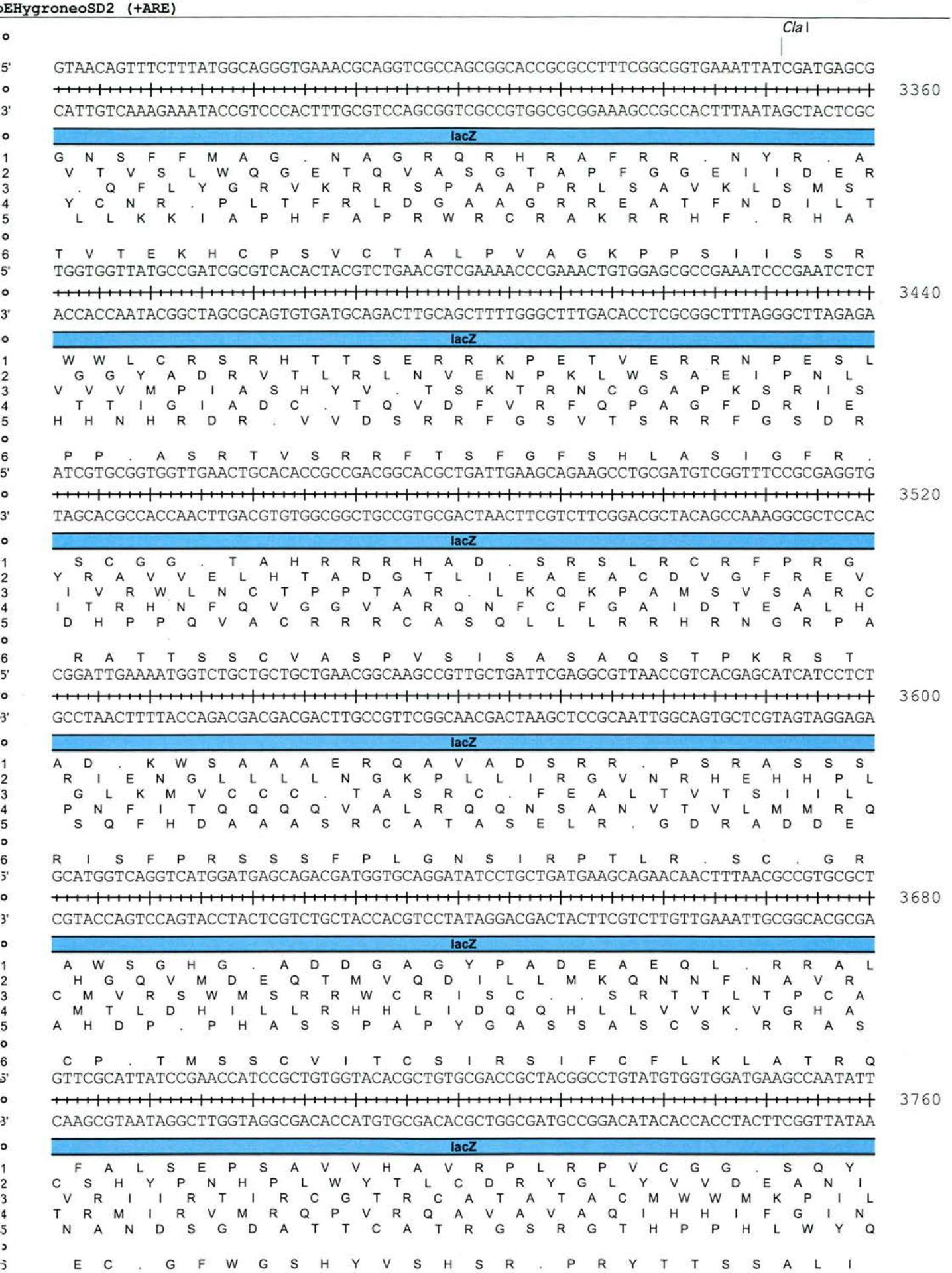

En-2 SA intron
1040
1120
1200
1280
1360
1440

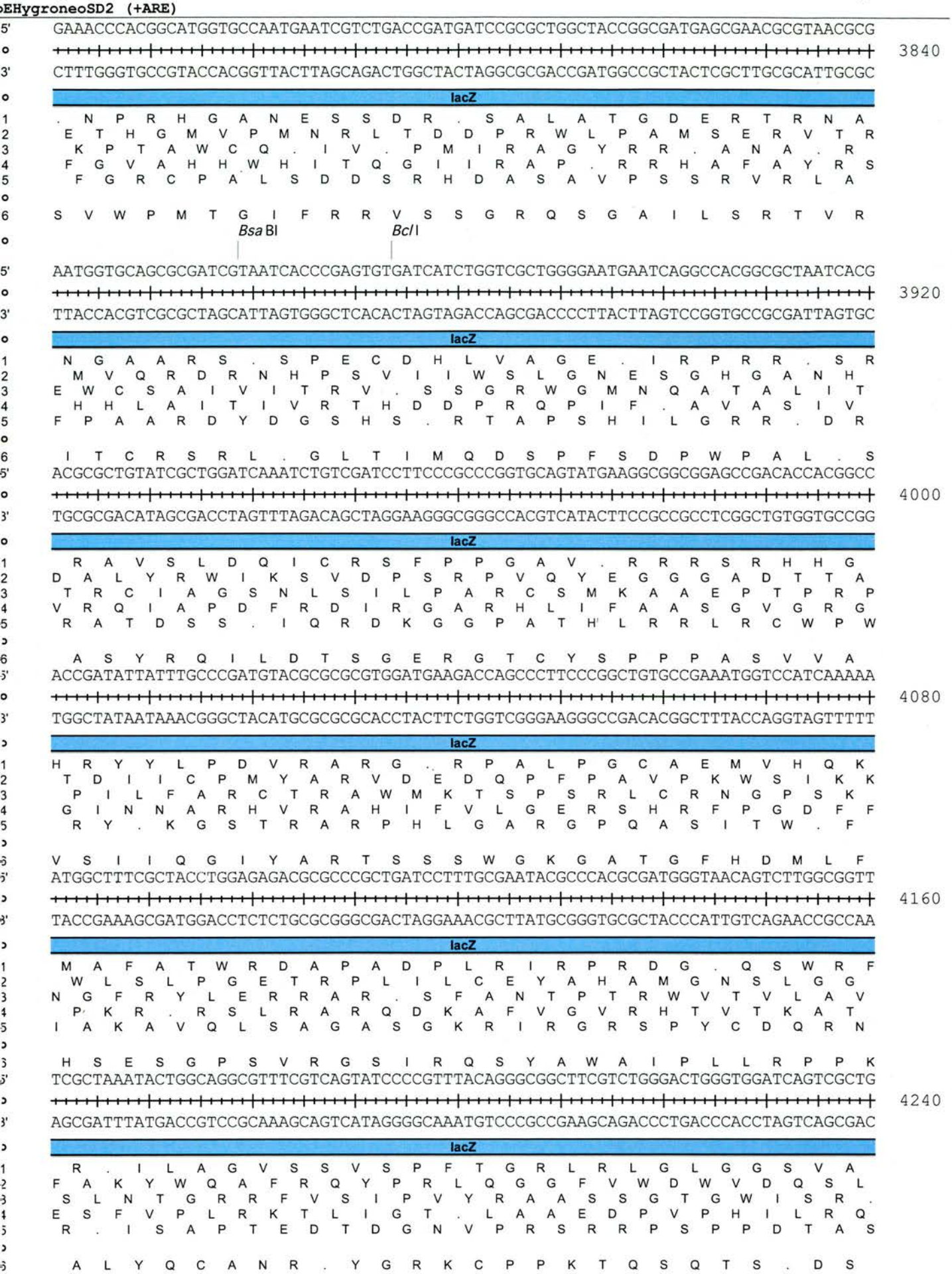

En-2 SA intron
En-2 SA intron
En-2 SA intron
Sex AI
En...on
En-2 SA exon
EGFP
EGFP

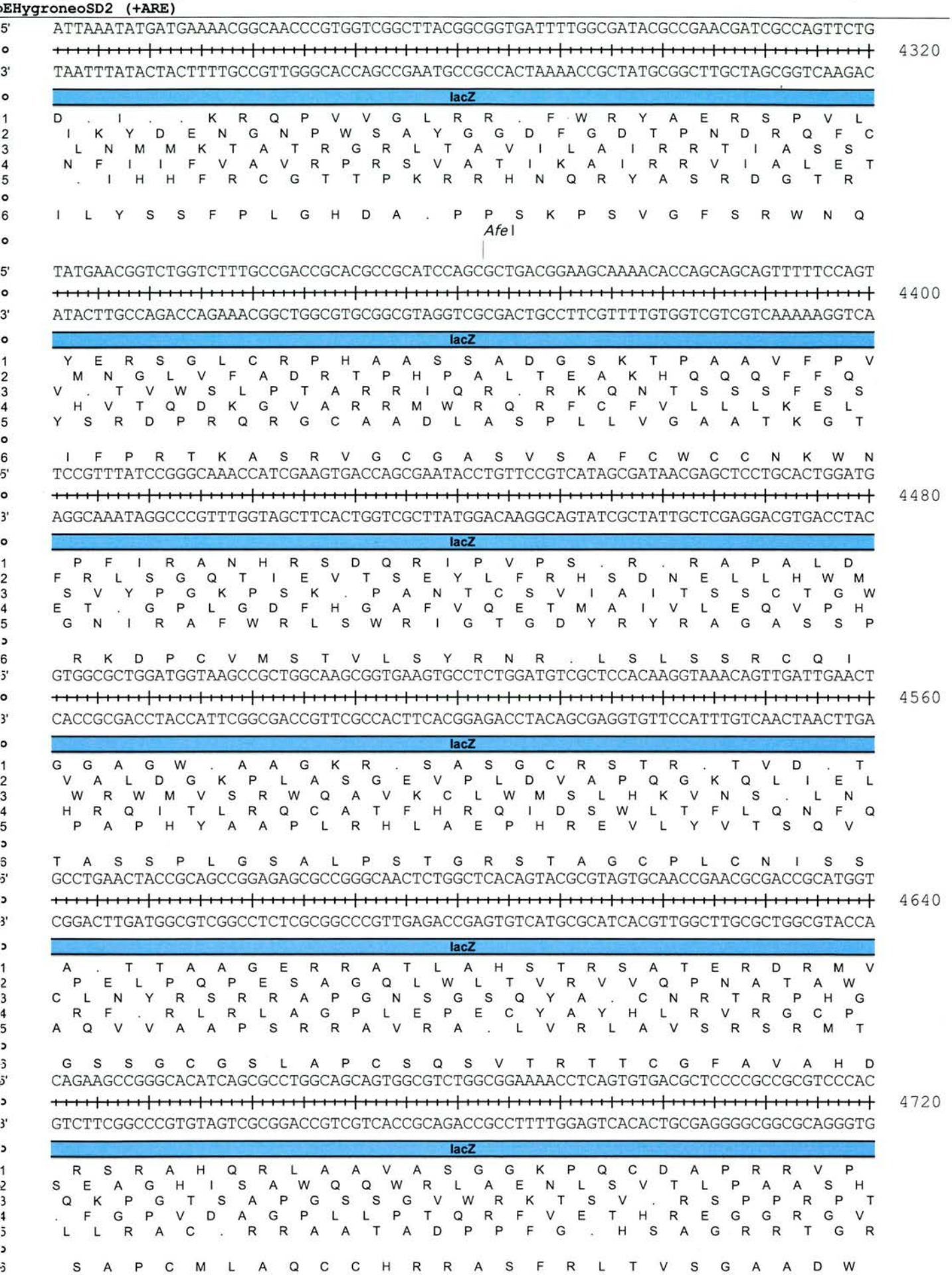

5' CCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCC
2000
3' GGTGGCCGTTCGACGGGCACGGGACCGGGTGGGAGCACTGGTGGGACTGGATGCCGCACGTCACGAAGTCGGCGATGGGG
EGFP
P P A S C P C P G P P S . P P . P T A C S A S A A T P
H R Q A A R A L A H P R D H P D L R R A V L Q P L P
T T G K L P V P W P T L V T T L T Y G V Q C F S R Y P
V P L S G T G Q G V R T V V R V . P T C H K L R . G
G G A L Q G H G P G G E H G G Q G V A H L A E A A V G
W R C A A R A R A W G R S W G S R R R A T S . G S G R
5' GACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGA
2080
3' CTGGTGTACTTCGTCGTGCTGAAGAAGTTCAGGCGGTACGGGCTTCCGATGCAGGTCCTCGCGTGGTAGAAGAAGTTCCT
EGFP
T T . S S T T S S S P P C P K A T S R S A P S S S R
R P H E A A R L L Q V R H A R R L R P G A H H L L Q G
D H M K Q H D F F K S A M P E G Y V Q E R T I F F K D
S W M F C C S K K L D A M G S P . T W S R V M K K L S
V V H L L V V E E L G G H G F A V D L L A G D E E L V
G C S A A R S R . T R W A R L S R G P A C W R R . P
5' CGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCG
2160
3' GCTGCCGTTGATGTTCTGGGCGCGGCTCCACTTCAAGCTCCCGCTGTGGGACCACTTGGCGTAGCTCGACTTCCCGTAGC
EGFP
T T A T T R P A P R . S S R A T P W . T A S S . R A S
R R Q L Q D P R R G E V R G R H P G E P H R A E G H R
D G N Y K T R A E V K F E G D T L V N R I E L K G I
S P L . L V R A S T F N S P S V R T F R M S S F P M S
V A V V V L G A G L H L E L A V G Q H V A D L Q L A D
R R C S C S G R R P S T R P R C G P S G C R A S P C R
5' ACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGAC
2240
3' TGAAGTTCCTCCTGCCGTTGTAGGACCCCGTGTTCGACCTCATGTTGATGTTGTCGGTGTTGCAGATATAGTACCGGCTG
EGFP
T S R R T A T S W G T S W S T T T T A T T S I S W P T
L Q G G R Q H P G A Q A G V Q L Q Q P Q R L Y H G R
D F K E D G N I L G H K L E Y N Y N S H N V Y I M A D
K L S S P L M R P C L S S Y L . L L W L T . I M A S
V E L L V A V D Q P V L Q L V V V V A V V D I D H G V
S . P P R C C G P A C A P T C S C C G C R R Y . P R C
5' AAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTA
2320
3' TTCGTCTTCTTGCCGTAGTTCCACTTGAAGTTCTAGGCGGTGTTGTAGCTCCTGCCGTCGCACGTCGAGCGGCTGGTGAT
EGFP
S R R T A S R . T S R S A T T S R T A A C S S P T T
Q A E E R H Q G E L Q D P P Q H R G R Q R A A R R P L
K Q K N G I K V N F K I R H N I E D G S V Q L A D H Y
L C F F P M L T F K L I R W L M S S P L T C S A S W .
L L L V A D L H V E L D A V V D L V A A H L E G V V V
A S S R C . P S S . S G G C C R P R C R A A R R G S
5' CCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCA
2400
3' GGTCGTCTTGTGGGGGTAGCCGCTGCCGGGGCACGACGACGGGCTGTTGGTGATGGACTCGTGGGTCAGGCGGGACTCGT
EGFP
T S R T P P S A T A P C C C P T T T T . A P S P P . A
P A E H P H R R R P R A A A R Q P L P E H P V R P E Q
Q Q N T P I G D G P V L L P D N H Y L S T Q S A L S
W C F V G M P S P G T S S G S L W . R L V W D A R L L
L L V G G D A V A G H Q Q G V V V V Q A G L G G Q A
G A S C G W R R R G R A A A R C G S G S C G T R G S C

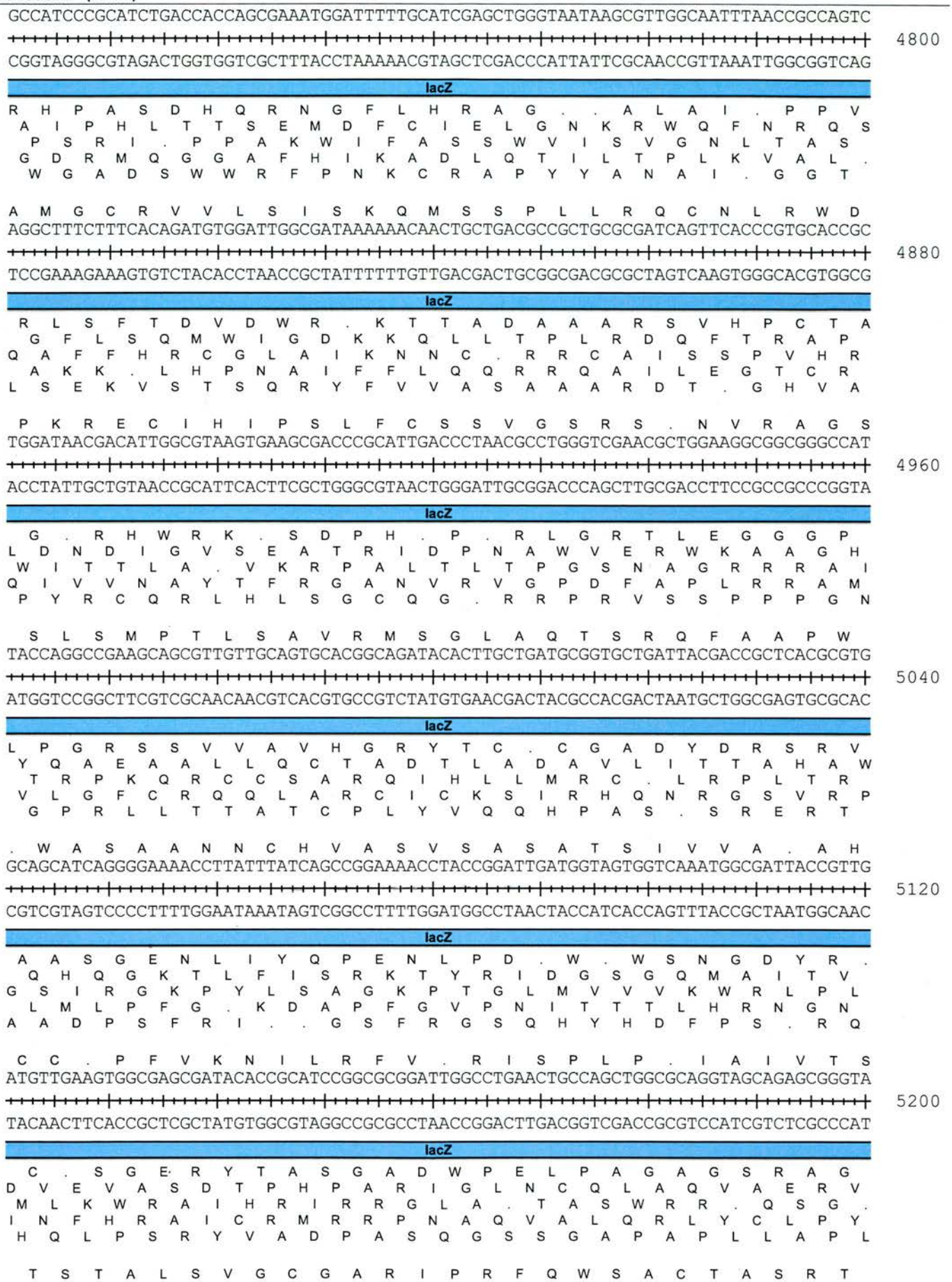

EGFP
lacZ
BsrGI
BaeI'
BaeI

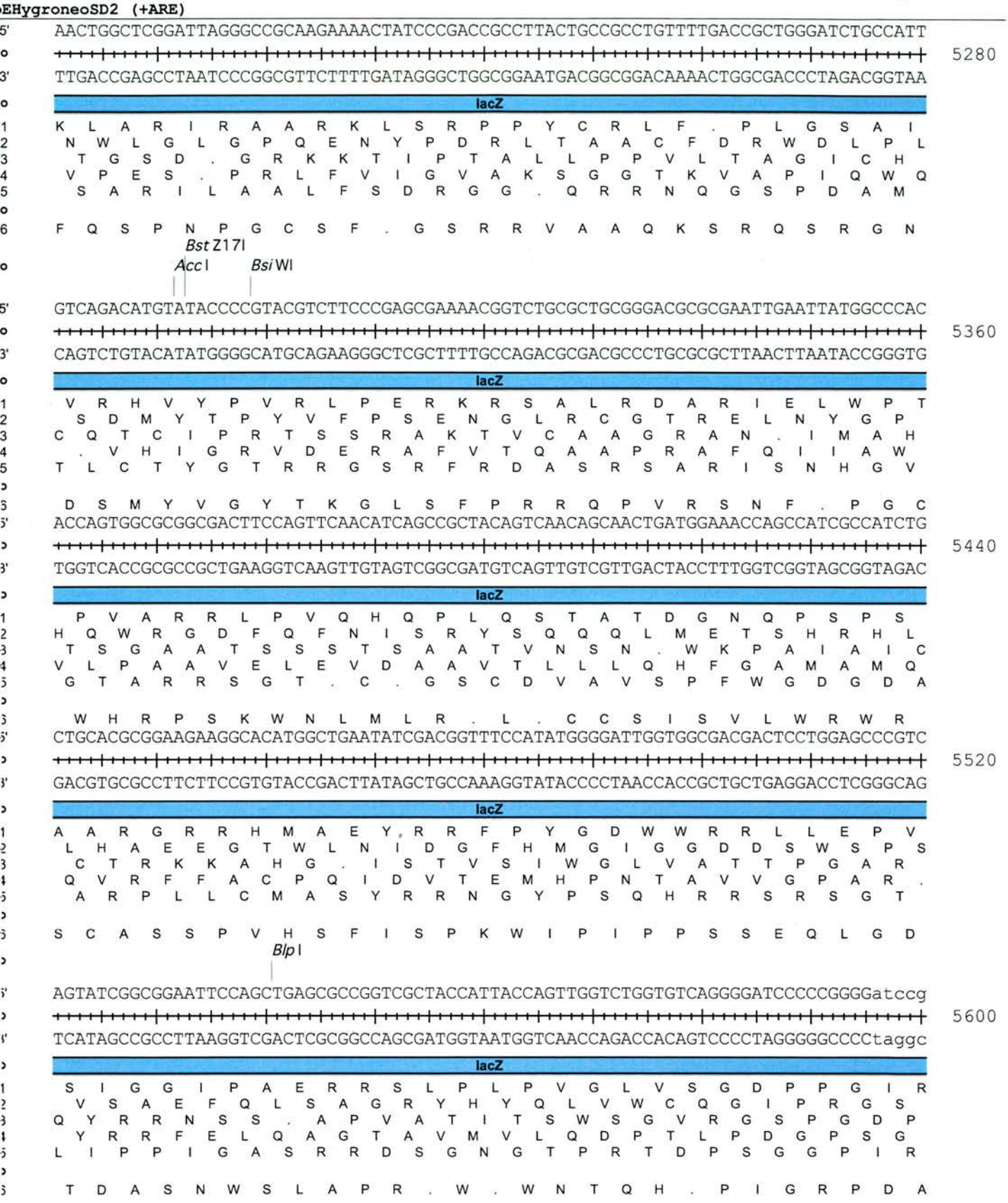

lacZ
lacZ
lacZ
lacZ
lacZ
2880
2960
3040
3120
3200
3280

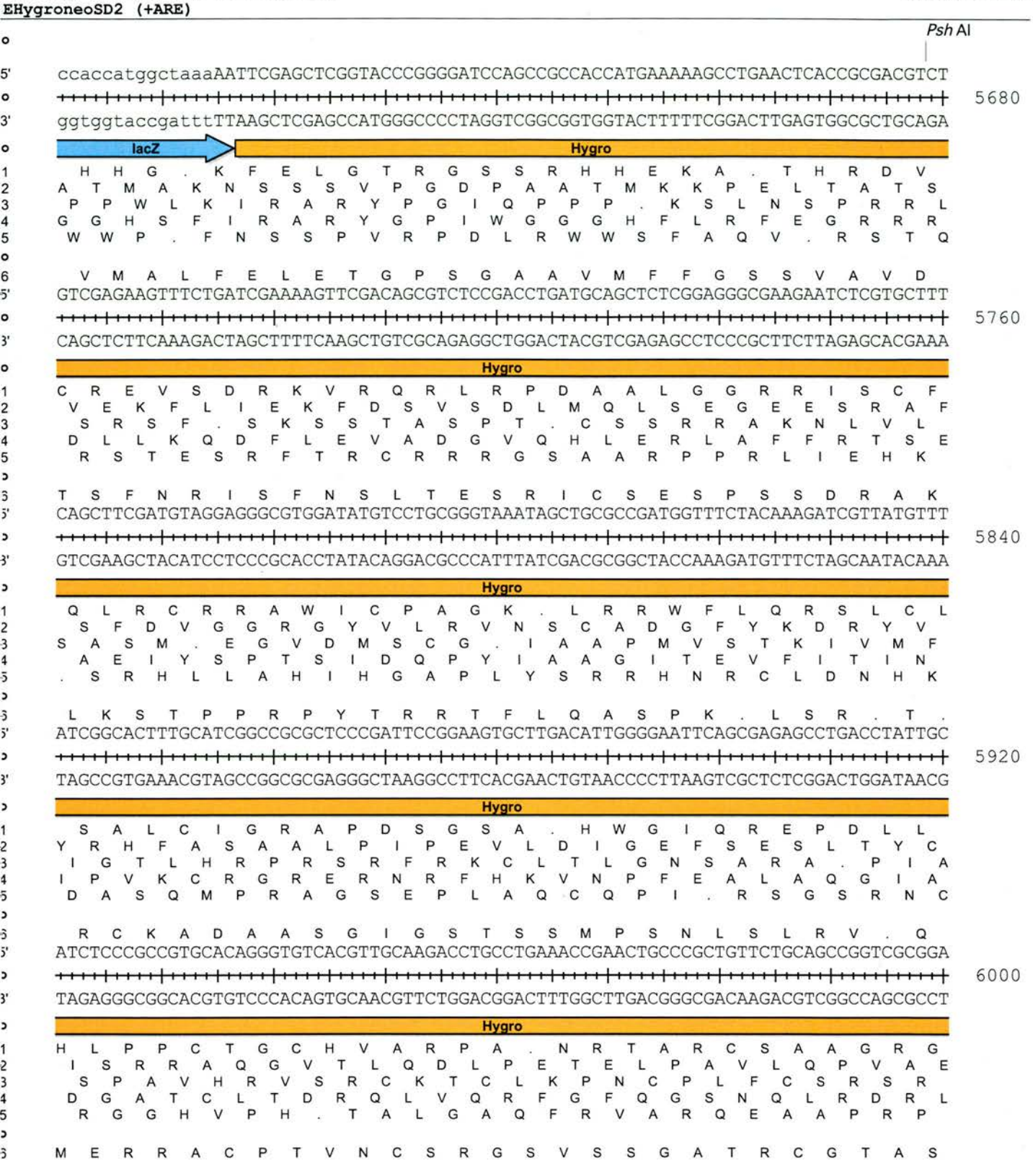

ClaI
lacZ
3360
3440
3520
3600
3680
3760

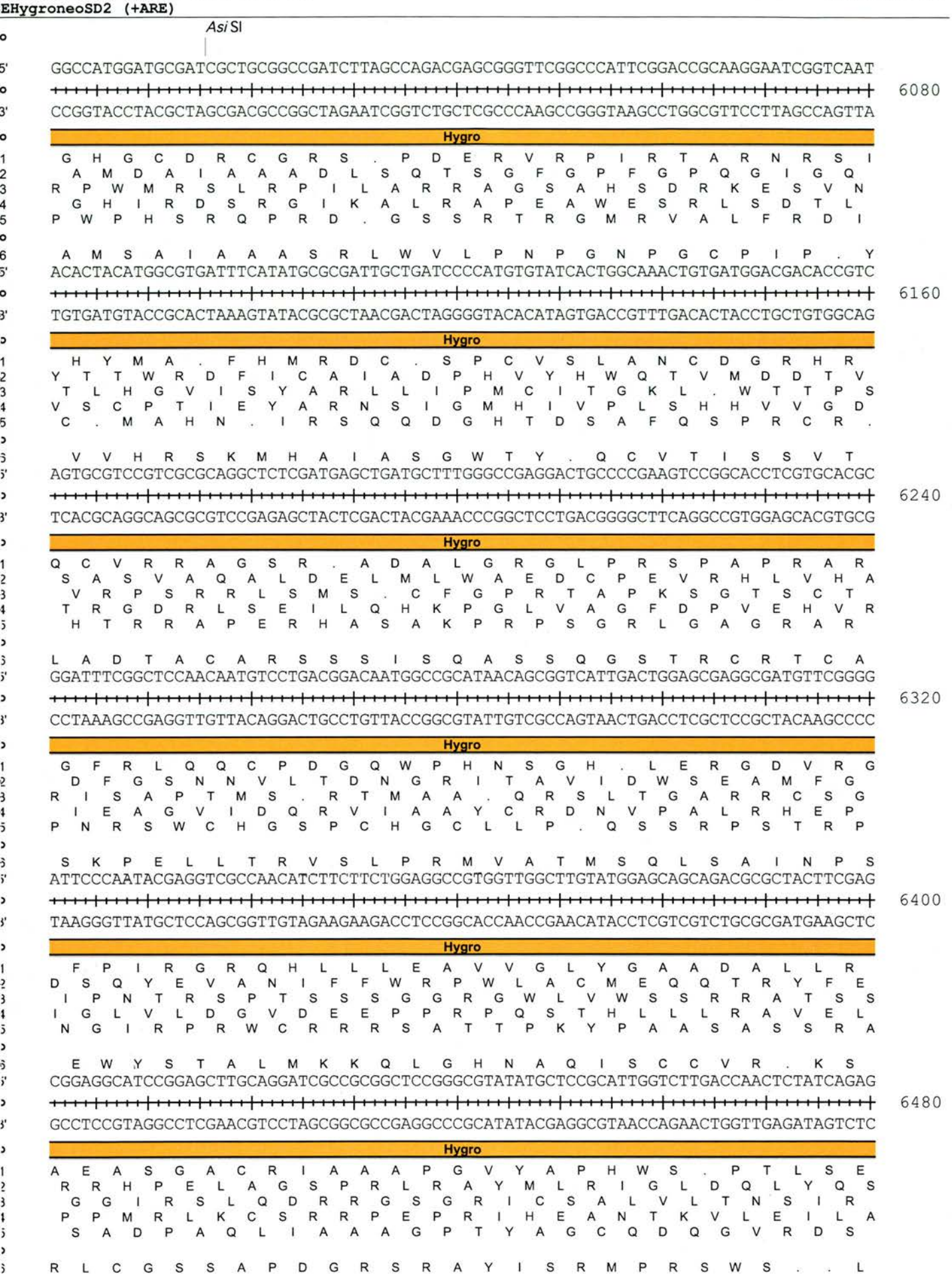

lacZ
BsaBI
BclI
lacZ
lacZ
lacZ
lacZ

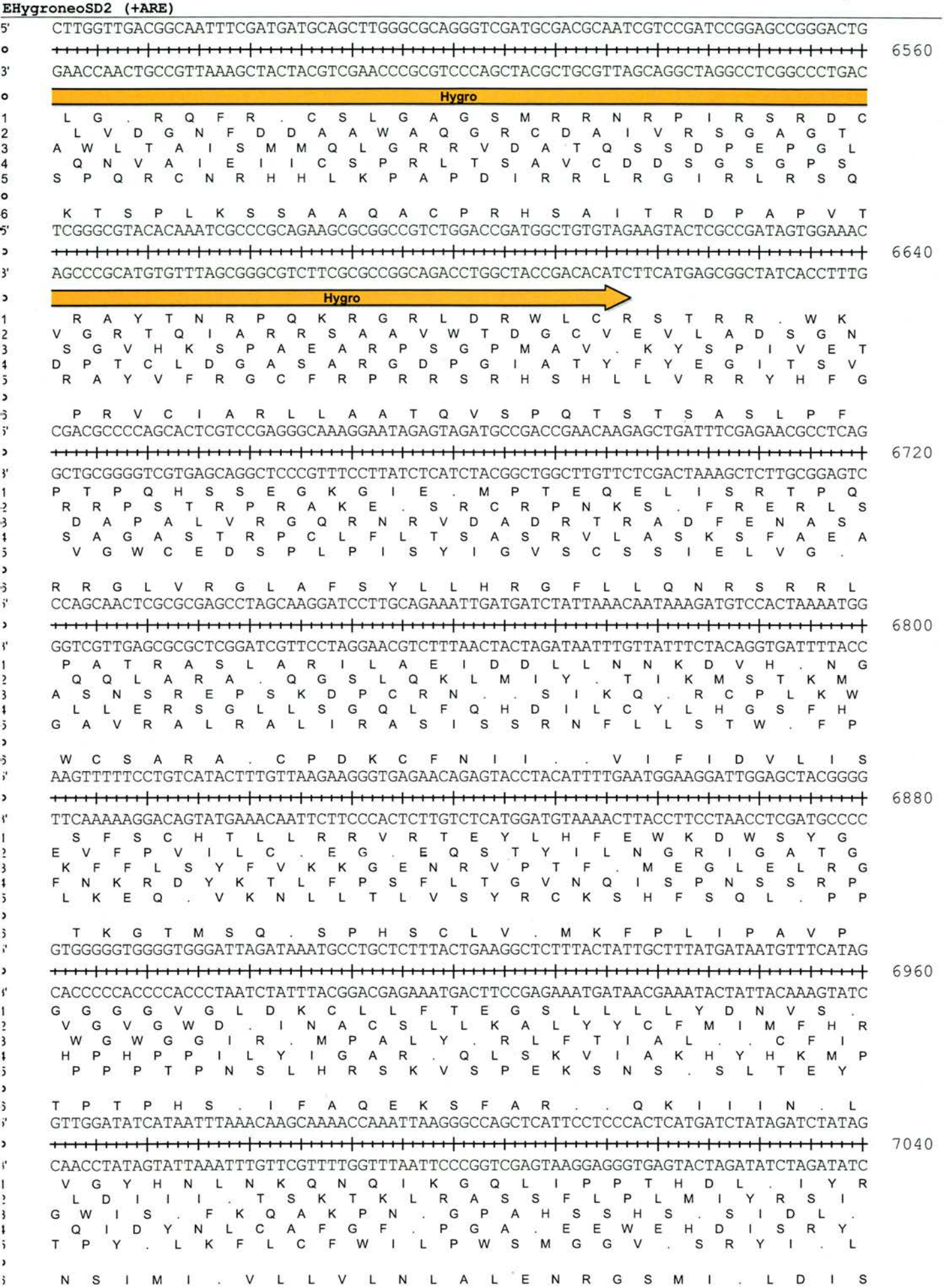

lacZ
AfeI
4320
4400
4480
4560
4640
4720

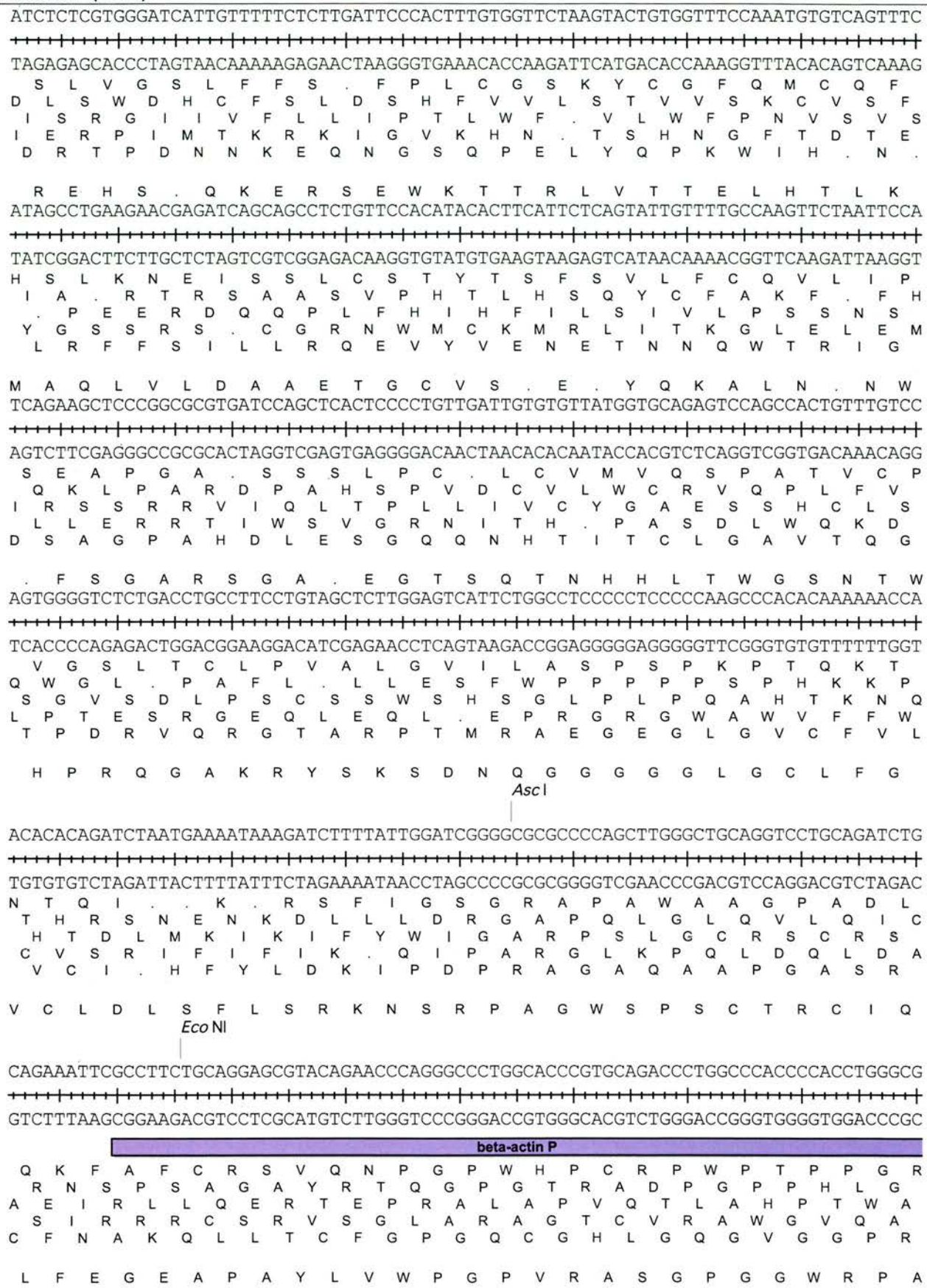

lacZ
4800
4880
4960
5040
5120
5200

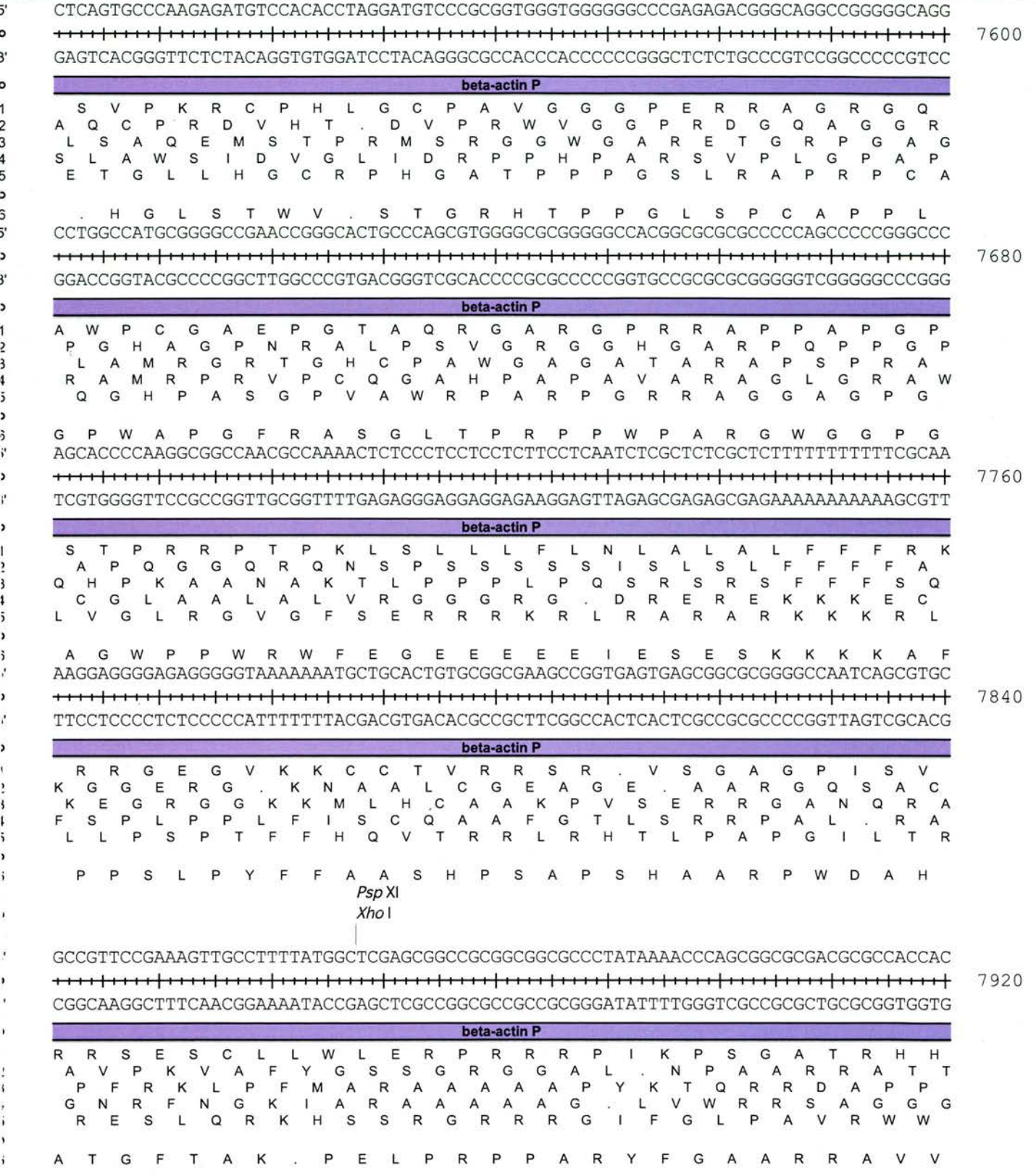

5' AACTGGCTCGGATTAGGGCCGCAAGAAAACTATCCCGACCGCCTTACTGCCGCCTGTTTTGACCGCTGGGATCTGCCATT
5280
3' TTGACCGAGCCTAATCCCGGCGTTCTTTTGATAGGGCTGGCGGAATGACGGCGGACAAAACTGGCGACCCTAGACGGTAA
lacZ
K L A R I R A A R K L S R P P Y C R L F . P L G S A I
N W L G L G P Q E N Y P D R L T A A C F D R W D L P L
T G S D . G R K K T I P T A L L P P V L T A G I C H
V P E S . P R L F V I G V A K S G G T K V A P I Q W Q
S A R I L A A L F S D R G G . Q R R N Q G S P D A M
F Q S P N P G C S F . G S R R V A A Q K S R Q S R G N
BstZ17I
AccI
BsiWI
5' GTCAGACATGTATACCCCGTACGTCTTCCCGAGCGAAAACGGTCTGCGCTGCGGGACGCGCGAATTGAATTATGGCCCAC
5360
3' CAGTCTGTACATATGGGGCATGCAGAAGGGCTCGCTTTTGCCAGACGCGACGCCCTGCGCGCTTAACTTAATACCGGGTG
lacZ
V R H V Y P V R L P E R K R S A L R D A R I E L W P T
S D M Y T P Y V F P S E N G L R C G T R E L N Y G P
C Q T C I P R T S S R A K T V C A A G R A N . I M A H
. V H I G R V D E R A F V T Q A A P R A F Q I I A W
T L C T Y G T R R G S R F R D A S R S A R I S N H G V
D S M Y V G Y T K G L S F P R R Q P V R S N F . P G C
5' ACCAGTGGCGCGGCGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAACTGATGGAAACCAGCCATCGCCATCTG
5440
3' TGGTCACCGCGCCGCTGAAGGTCAAGTTGTAGTCGGCGATGTCAGTTGTCGTTGACTACCTTTGGTCGGTAGCGGTAGAC
lacZ
P V A R R L P V Q H Q P L Q S T A T D G N Q P S P S
H Q W R G D F Q F N I S R Y S Q Q Q L M E T S H R H L
T S G A A T S S S T S A A T V N S N . W K P A I A I C
V L P A A V E L E V D A A V T L L L Q H F G A M A M Q
G T A R R S G T . C . G S C D V A V S P F W G D G D A
W H R P S K W N L M L R . L . C C S I S V L W R W R
5' CTGCACGCGGAAGAAGGCACATGGCTGAATATCGACGGTTTCCATATGGGGATTGGTGGCGACGACTCCTGGAGCCCGTC
5520
3' GACGTGCGCCTTCTTCCGTGTACCGACTTATAGCTGCCAAAGGTATACCCCTAACCACCGCTGCTGAGGACCTCGGGCAG
lacZ
A A R G R R H M A E Y R R F P Y G D W W R R L L E P V
L H A E E G T W L N I D G F H M G I G G D D S W S P S
C T R K K A H G . I S T V S I W G L V A T T P G A R
Q V R F F A C P Q I D V T E M H P N T A V V G P A R .
A R P L L C M A S Y R R N G Y P S Q H R R S R S G T
S C A S S P V H S F I S P K W I P I P P S S E Q L G D
BlpI
5' AGTATCGGCGGAATTCCAGCTGAGCGCCGGTCGCTACCATTACCAGTTGGTCTGGTGTCAGGGGATCCCCCGGGGatccg
5600
3' TCATAGCCGCCTTAAGGTCGACTCGCGGCCAGCGATGGTAATGGTCAACCAGACCACAGTCCCCTAGGGGGCCCCtaggc
lacZ
S I G G I P A E R R S L P L P V G L V S G D P P G I R
V S A E F Q L S A G R Y H Y Q L V W C Q G I P R G S
Q Y R R N S S . A P V A T I T S W S G V R G S P G D P
Y R R F E L Q A G T A V M V L Q D P T L P D G P S G
L I P P I G A S R R D S G N G T P R T D P S G G P I R
T D A S N W S L A P R . W . W N T Q H . P I G R P D A

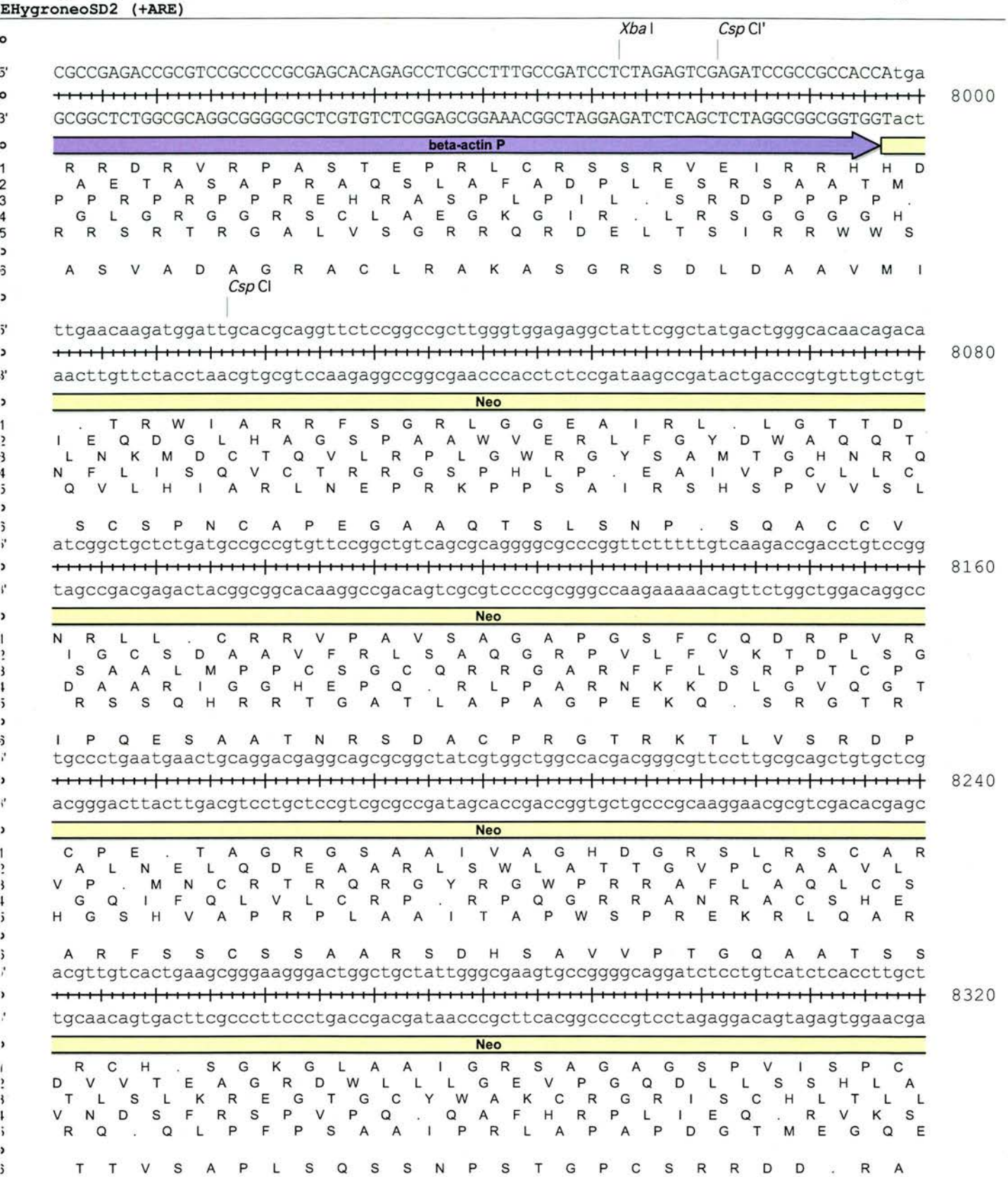

Psh AI
lacZ
Hygro
5680
5760
5840
5920
6000

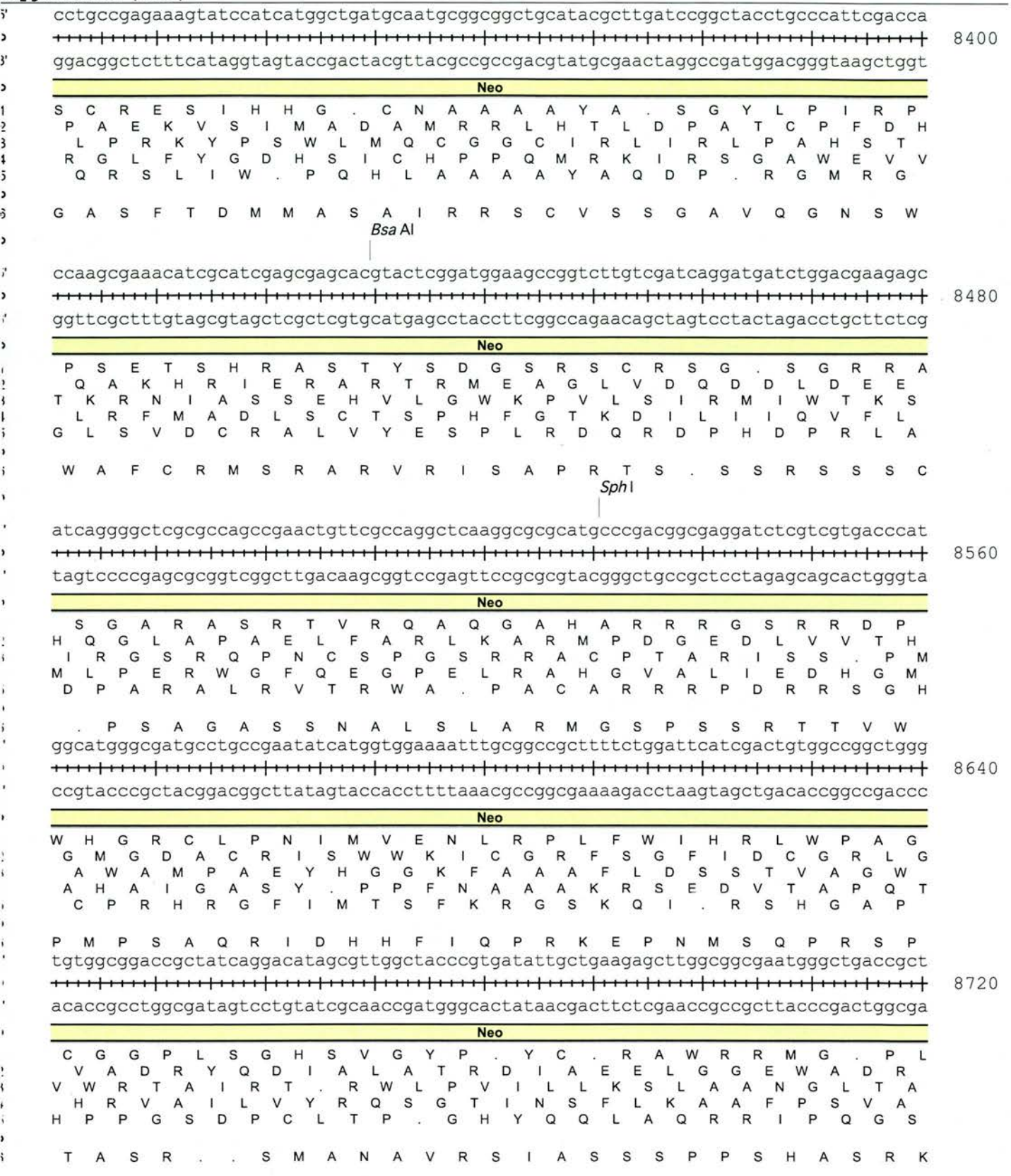

AsiSI
Hygro
6080
6160
6240
6320
6400
6480

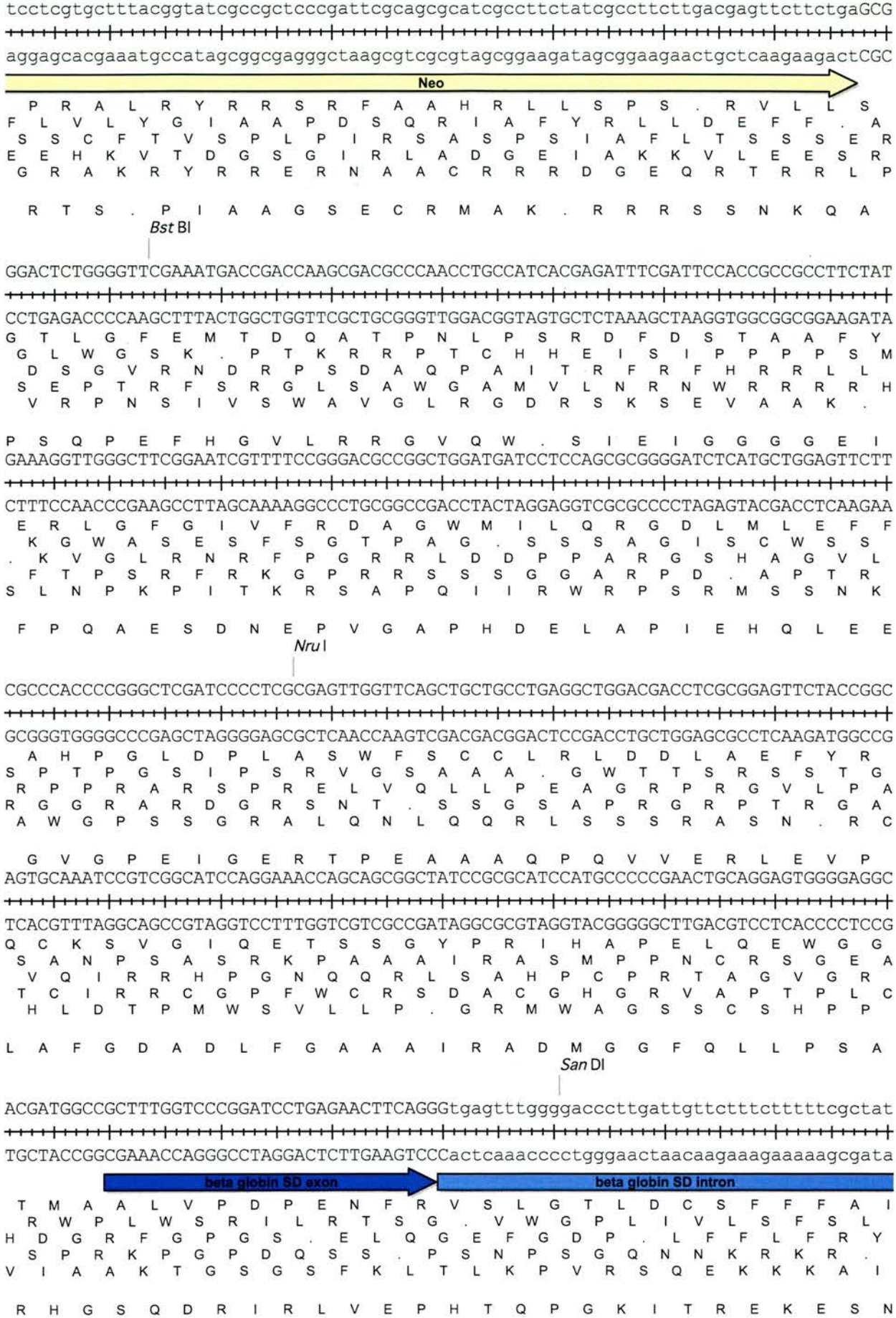

Hygro
Hygro

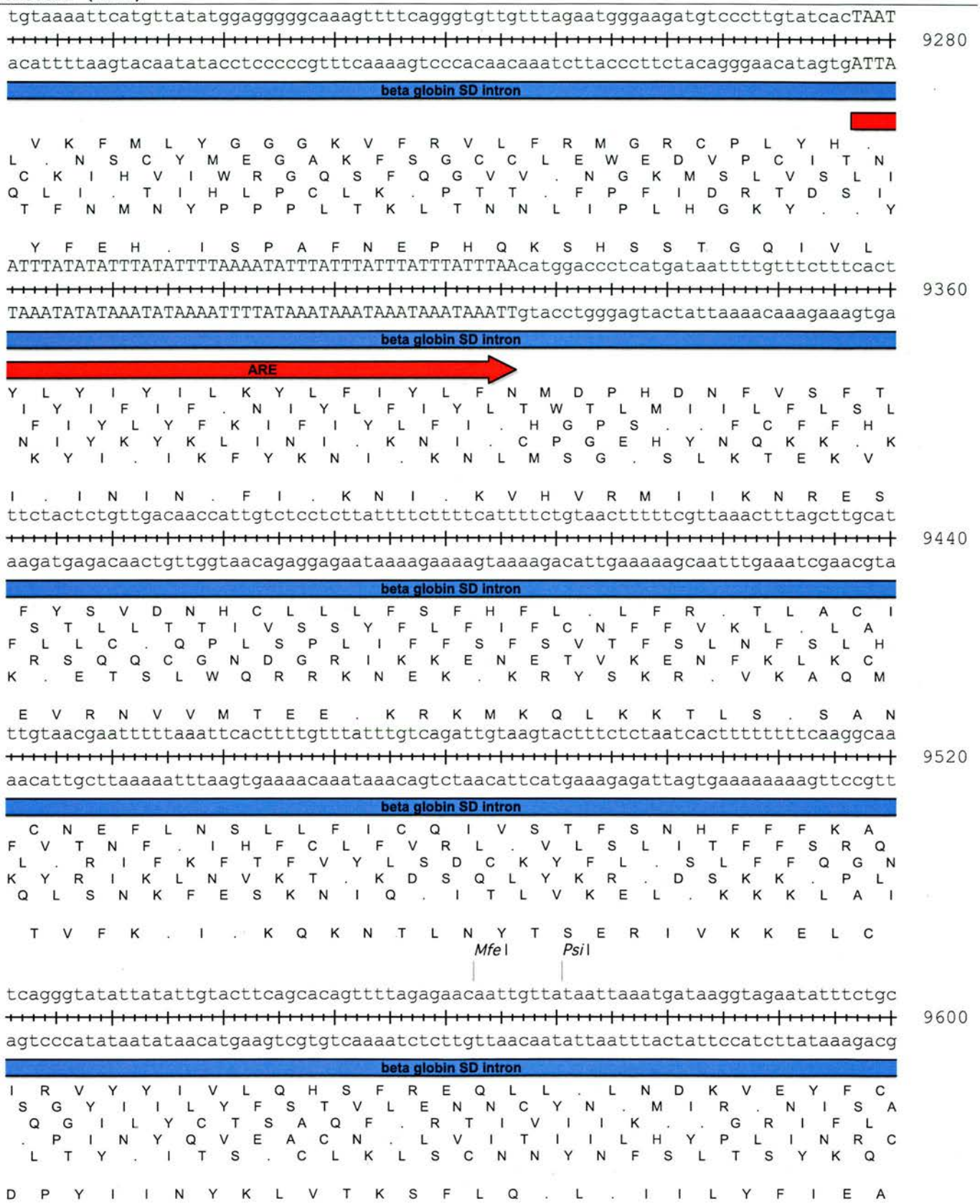

AscI
EcoNI
beta-actin P

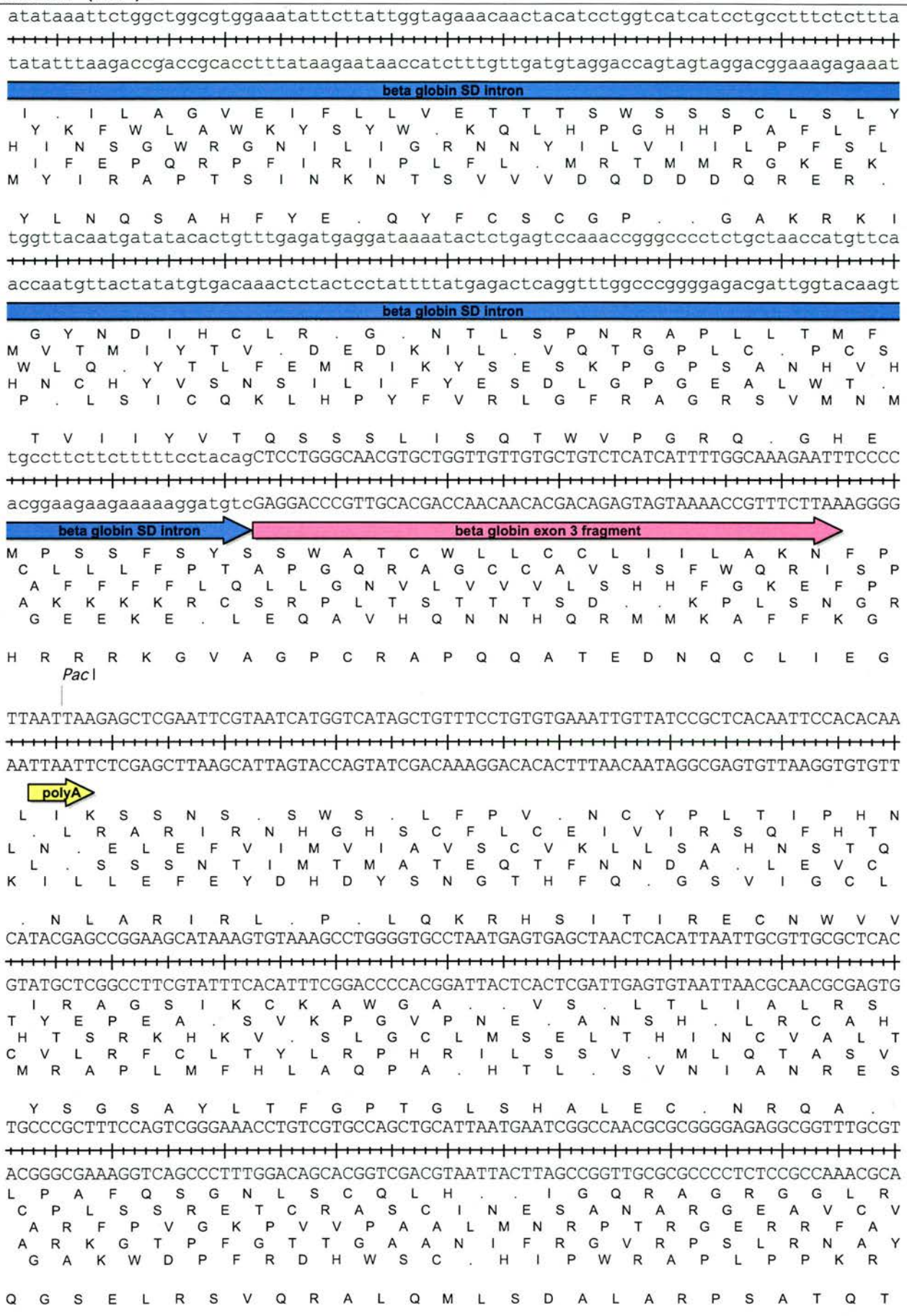

beta-actin P
beta-actin P
beta-actin P
beta-actin P
beta-actin P
Psp XI
Xho I
7600
7680
7760
7840
7920


Xba I
Csp CI'
beta-actin P
8000
Csp CI
Neo
8080
Neo
8160
Neo
8240
Neo
8320
Neo

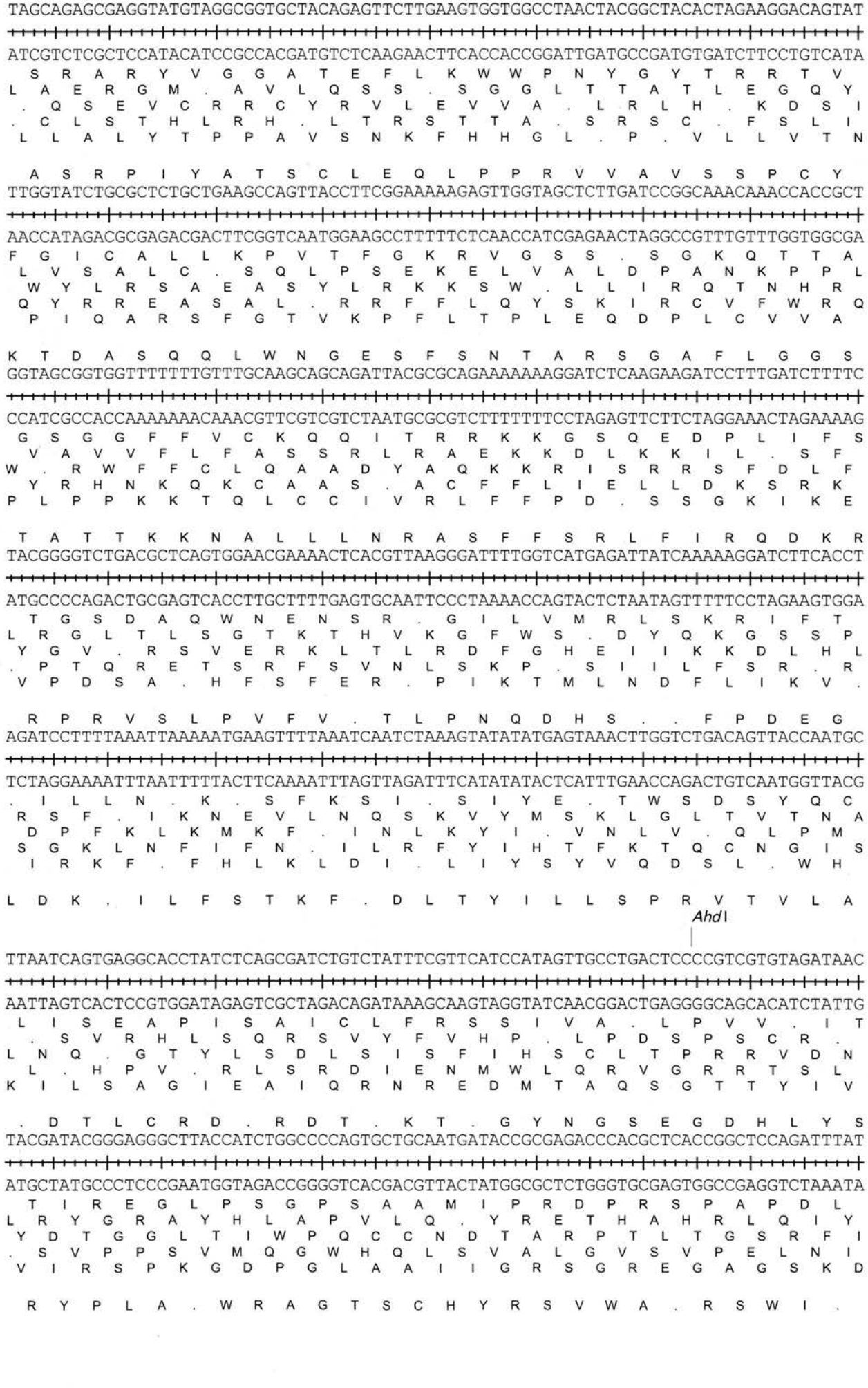

cctgccgagaaagtatccatcatggctgatgcaatgcggcggctgcatacgcttgatccggctacctgcccattcgacca
8400
ggacggctctttcataggtagtaccgactacgttacgccgccgacgtatgcgaactaggccgatggacgggtaagctggt
Neo
S C R E S I H H G . C N A A A A Y A . S G Y L P I R P
P A E K V S I M A D A M R R L H T L D P A T C P F D H
L P R K Y P S W L M Q C G G C I R L I R L P A H S T
R G L F Y G D H S I C H P P Q M R K I R S G A W E V V
Q R S L I W . P Q H L A A A A Y A Q D P . R G M R G
G A S F T D M M A S A I R R S C V S S G A V Q G N S W
Bsa AI
ccaagcgaaacatcgcatcgagcgagcacgtactcggatggaagccggtcttgtcgatcaggatgatctggacgaagagc
8480
ggttcgctttgtagcgtagctcgctcgtgcatgagcctaccttcggccagaacagctagtcctactagacctgcttctcg
Neo
P S E T S H R A S T Y S D G S R S C R S G . S G R R A
Q A K H R I E R A R T R M E A G L V D Q D D L D E E
T K R N I A S S E H V L G W K P V L S I R M I W T K S
L R F M A D L S C T S P H F G T K D I L I I Q V F L
G L S V D C R A L V Y E S P L R D Q R D P H D P R L A
W A F C R M S R A R V R I S A P R T S . S S R S S S C
Sph I
atcaggggctcgcgccagccgaactgttcgccaggctcaaggcgcgcatgcccgacggcgaggatctcgtcgtgacccat
8560
tagtccccgagcgcggtcggcttgacaagcggtccgagttccgcgcgtacgggctgccgctcctagagcagcactgggta
Neo
S G A R A S R T V R Q A Q G A H A R R R G S R R D P
H Q G L A P A E L F A R L K A R M P D G E D L V V T H
I R G S R Q P N C S P G S R R A C P T A R I S S . P M
M L P E R W G F Q E G P E L R A H G V A L I E D H G M
D P A R A L R V T R W A . P A C A R R R P D R R S G H
. P S A G A S S N A L S L A R M G S P S S R T T V W
ggcatgggcgatgcctgccgaatatcatggtggaaaatttgcggccgcttttctggattcatcgactgtggccggctggg
8640
ccgtacccgctacggacggcttatagtaccacctttaaacgccggcgaaaagacctaagtagctgacaccggccgaccc
Neo
W H G R C L P N I M V E N L R P L F W I H R L W P A G
G M G D A C R I S W W K I C G R F S G F I D C G R L G
A W A M P A E Y H G G K F A A A F L D S S T V A G W
A H A I G A S Y . P P F N A A A K R S E D V T A P Q T
C P R H R G F I M T S F K R G S K Q I . R S H G A P
P M P S A Q R I D H H F I Q P R K E P N M S Q P R S P
tgtggcggaccgctatcaggacatagcgttggctacccgtgatattgctgaagagcttggcggcgaatgggctgaccgct
8720
acaccgcctggcgatagtcctgtatcgcaaccgatgggcactataacgacttctcgaaccgccgcttacccgactggcga
Neo
C G G P L S G H S V G Y P . Y C . R A W R R M G . P L
V A D R Y Q D I A L A T R D I A E E L G G E W A D R
V W R T A I R T . R W L P V I L L K S L A A N G L T A
H R V A I L V Y R Q S G T I N S F L K A A F P S V A
H P P G S D P C L T P . G H Y Q Q L A Q R R I P Q G S
T A S R . . S M A N A V R S I A S S S P P S H A S R K

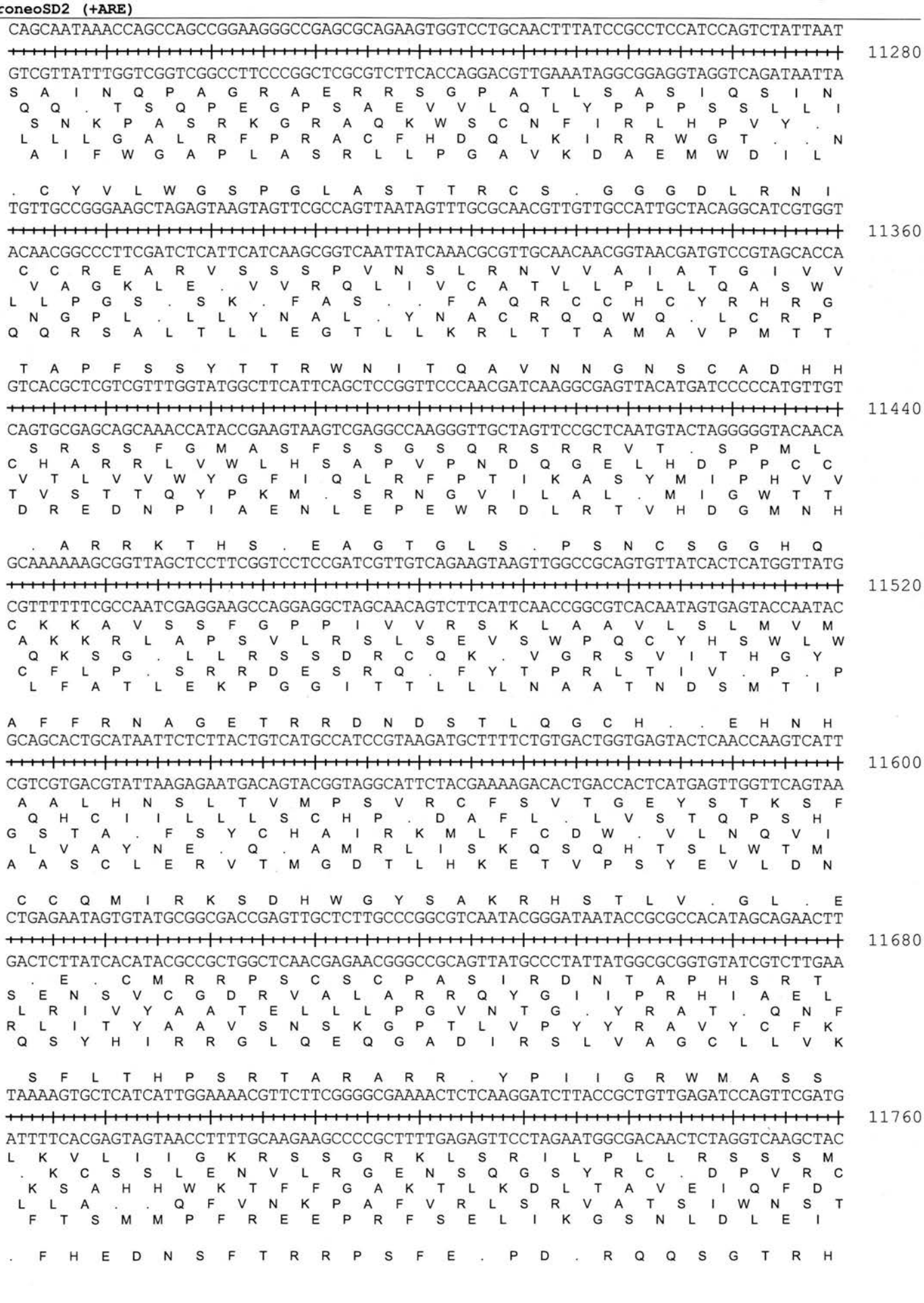

Neo
BstBI
NruI
SanDI
beta globin SD exon
beta globin SD intron

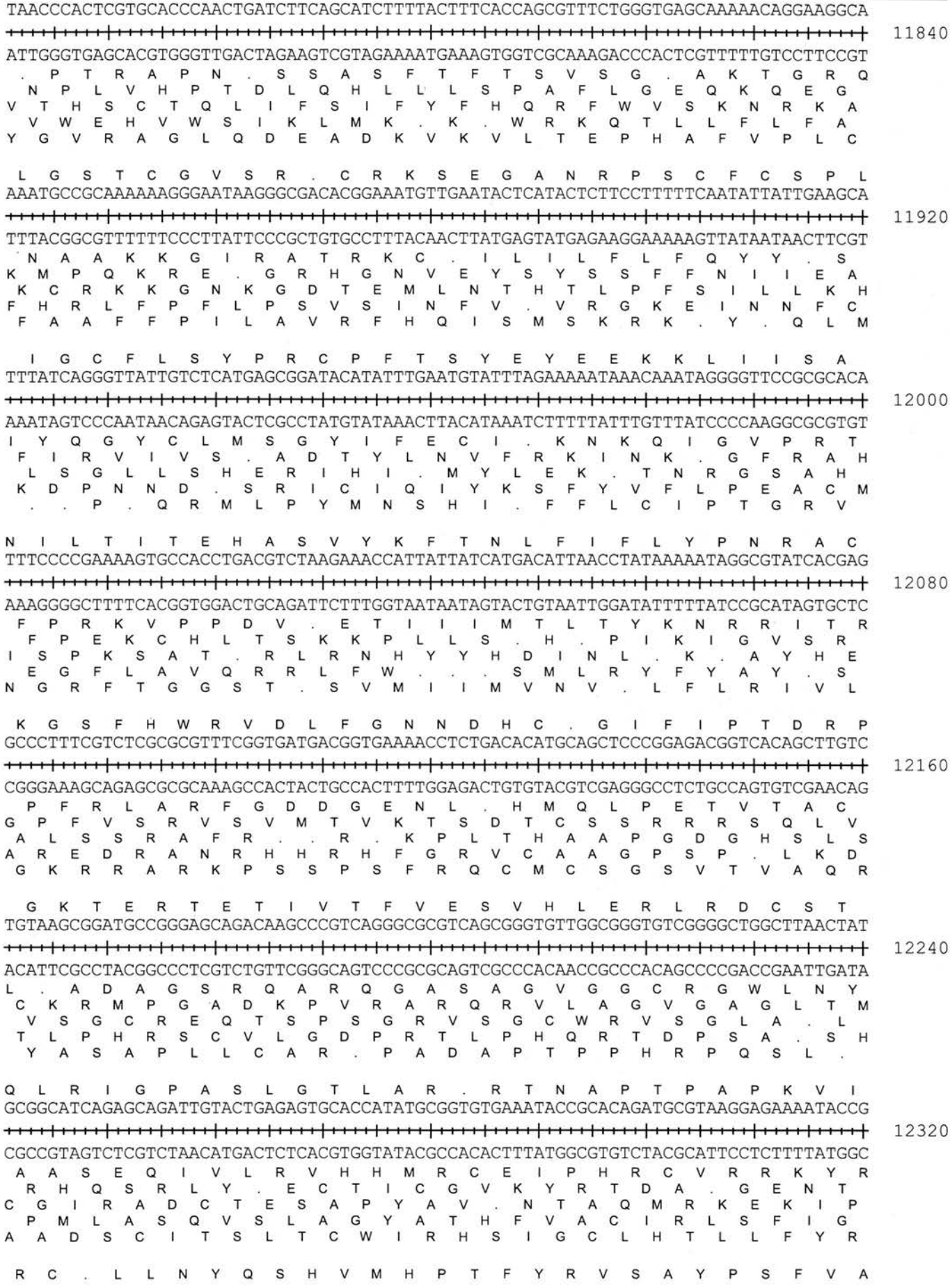

beta globin SD intron
ARE
beta globin SD intron
beta globin SD intron
beta globin SD intron
beta globin SD intron
MfeI
PsiI
beta globin SD intron


beta globin SD intron
beta globin SD intron
beta globin SD intron
beta globin exon 3 fragment
PacI
polyA
9680
9760
9840
9920
10000
10080

```
    ATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTC
    ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++ 10160
    TAACCCGCGAGAAGGCGAAGGAGCGAGTGACTGAGCGACGCGAGCCAGCAAGCCGACGCCGCTCGCCATAGTCGAGTGAG
      I  G  R  S  S  A  S  S  L  T  D  S  L  R  S  V  V  R  L  R  R  A  V  S  A  H  S
    L  G  A  L  P  L  P  R  S  L  T  R  C  A  R  S  F  G  C  G  E  R  Y  Q  L  T
    Y  W  A  L  F  R  F  L  A  H  .  L  A  A  L  G  R  S  A  A  A  S  G  I  S  S  L
      Q  A  S  K  R  K  R  A  .  Q  S  A  A  S  P  R  E  A  A  A  L  P  I  L  E  S
    I  P  R  E  E  A  E  E  S  V  S  E  S  R  E  T  T  R  S  R  R  A  T  D  A  .  E

     N  P  A  R  G  S  G  R  E  S  V  R  Q  A  R  D  N  P  Q  P  S  R  Y  .  S  V  .
    AAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCA
    ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++ 10240
    TTTCCGCCATTATGCCAATAGGTGTCTTAGTCCCCTATTGCGTCCTTTCTTGTACACTCGTTTTCCGGTCGTTTTCCGGT
      K  A  V  I  R  L  S  T  E  S  G  D  N  A  G  K  N  M  .  A  K  G  Q  Q  K  A
    Q  R  R  .  Y  G  Y  P  Q  N  Q  G  I  T  Q  E  R  T  C  E  Q  K  A  S  K  R  P
     K  G  G  N  T  V  I  H  R  I  R  G  .  R  R  K  E  H  V  S  K  R  P  A  K  G  Q
    L  P  P  L  V  T  I  W  L  I  L  P  Y  R  L  F  S  C  T  L  L  L  G  A  F  P  W
     F  A  T  I  R  N  D  V  S  D  P  S  L  A  P  F  F  M  H  A  F  P  W  C  F  A  L

      L  R  Y  Y  P  .  G  C  F  .  P  I  V  C  S  L  V  H  S  C  F  A  L  L  L  G
    GGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCA
    ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++ 10320
    CCTTGGCATTTTTCCGGCGCAACGACCGCAAAAAGGTATCCGAGGCGGGGGGACTGCTCGTAGTGTTTTTAGCTGCGAGT
    R  N  R  K  K  A  A  L  L  A  F  F  H  R  L  R  P  P  D  E  H  H  K  N  R  R  S
     G  T  V  K  R  P  R  C  W  R  F  S  I  G  S  A  P  L  T  S  I  T  K  I  D  A  Q
      E  P  .  K  G  R  V  A  G  V  F  P  .  A  P  P  P  .  R  A  S  Q  K  S  T  L
     S  G  Y  F  P  R  T  A  P  T  K  G  Y  A  G  G  G  Q  R  A  D  C  F  D  V  S  L
       F  R  L  F  A  A  N  S  A  N  K  W  L  S  R  G  G  S  S  C  .  L  F  R  R  E

     P  V  T  F  L  G  R  Q  Q  R  K  E  M  P  E  A  G  R  V  L  M  V  F  I  S  A  .
    AGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGT
    ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++ 10400
    TCAGTCTCCACCGCTTTGGGCTGTCCTGATATTTCTATGGTCCGCAAAGGGGGACCTTCGAGGGAGCACGCGAGAGGACA
     S  Q  R  W  R  N  P  T  G  L  .  R  Y  Q  A  F  P  P  G  S  S  L  V  R  S  P  V
      V  R  G  G  E  T  R  Q  D  Y  K  D  T  R  R  F  P  L  E  A  P  S  C  A  L  L
    K  S  E  V  A  K  P  D  R  T  I  K  I  P  G  V  S  P  W  K  L  P  R  A  L  S  C
      D  S  T  A  F  G  S  L  V  I  F  I  G  P  T  E  G  Q  F  S  G  R  A  S  E  Q
    L  .  L  H  R  F  G  V  P  S  Y  L  Y  W  A  N  G  G  P  L  E  R  T  R  E  G  T

     T  L  P  P  S  V  R  C  S  .  L  S  V  L  R  K  G  R  S  A  G  E  H  A  R  R  N
    TCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTA
    ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++ 10480
    AGGCTGGGACGGCGAATGGCCTATGGACAGGCGGAAAGAGGGAAGCCCTTCGCACCGCGAAAGAGTATCGAGTGCGACAT
      P  T  L  P  L  T  G  Y  L  S  A  F  L  P  S  G  S  V  A  L  S  H  S  S  R  C
    F  R  P  C  R  L  P  D  T  C  P  P  F  S  L  R  E  A  W  R  F  L  I  A  H  A  V
     S  D  P  A  A  Y  R  I  P  V  R  L  S  P  F  G  K  R  G  A  F  S  .  L  T  L  .
    E  S  G  A  A  .  R  I  G  T  R  R  E  G  K  P  F  R  P  A  K  E  Y  S  V  S  Y
     G  V  R  G  S  V  P  Y  R  D  A  K  R  G  E  P  L  T  A  S  E  .  L  E  R  Q  L

      R  G  Q  R  K  G  S  V  Q  G  G  K  E  R  R  S  A  H  R  K  R  M  A  .  A  T
    GGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCC
    ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++ 10560
    CCATAGAGTCAAGCCACATCCAGCAAGCGAGGTTCGACCCGACACACGTGCTTGGGGGGCAAGTCGGGCTGGCGACGCGG
    R  Y  L  S  S  V  .  V  V  R  S  K  L  G  C  V  H  E  P  P  V  Q  P  D  R  C  A
     G  I  S  V  R  C  R  S  F  A  P  S  W  A  V  C  T  N  P  P  F  S  P  T  A  A  P
      V  S  Q  F  G  V  G  R  S  L  Q  A  G  L  C  A  R  T  P  R  S  A  R  P  L  R
     T  D  .  N  P  T  P  R  E  S  W  A  P  S  H  A  R  V  G  R  E  A  R  G  S  R  R
       Y  R  L  E  T  Y  T  T  R  E  L  S  P  Q  T  C  S  G  G  T  .  G  S  R  Q  A

     P  I  E  T  R  H  L  D  N  A  G  L  Q  A  T  H  V  F  G  G  N  L  G  V  A  A  G
    TTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGAT
    ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++ 10640
    AATAGGCCATTGATAGCAGAACTCAGGTTGGGCCATTCTGTGCTGAATAGCGGTGACCGTCGTCGGTGACCATTGTCCTA
     L  S  G  N  Y  R  L  E  S  N  P  V  R  H  D  L  S  P  L  A  A  A  T  G  N  R  I
      Y  P  V  T  I  V  L  S  P  T  R  .  D  T  T  Y  R  H  W  Q  Q  P  L  V  T  G
    L  I  R  .  L  S  S  .  V  Q  P  G  K  T  R  L  I  A  T  G  S  S  H  W  .  Q  D
      I  R  Y  S  D  D  Q  T  W  G  P  L  V  R  S  I  A  V  P  L  L  W  Q  Y  C  S
    K  D  P  L  .  R  R  S  D  L  G  T  L  C  S  K  D  G  S  A  A  A  V  P  L  L  I

      .  G  T  V  I  T  K  L  G  V  R  Y  S  V  V  .  R  W  Q  C  C  G  S  T  V  P  N
```


TAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTAT
10720
ATCGTCTCGCTCCATACATCCGCCACGATGTCTCAAGAACTTCACCACCGGATTGATGCCGATGTGATCTTCCTGTCATA
S R A R Y V G G A T E F L K W W P N Y G Y T R R T V
L A E R G M . A V L Q S S . S G G L T T A T L E G Q Y
. Q S E V C R R C Y R V L E V V A . L R L H . K D S I
. C L S T H L R H . L T R S T T A . S R S C . F S L I
L L A L Y T P P A V S N K F H H G L . P . V L L V T N
A S R P I Y A T S C L E Q L P P R V V A V S S P C Y
TTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCT
10800
AACCATAGACGCGAGACGACTTCGGTCAATGGAAGCCTTTTTCTCAACCATCGAGAACTAGGCCGTTTGTTTGGTGGCGA
F G I C A L L K P V T F G K R V G S S . S G K Q T T A
L V S A L C . S Q L P S E K E L V A L D P A N K P P L
W Y L R S A E A S Y L R K K S W . L L I R Q T N H R
Q Y R R E A S A L . R R F F L Q Y S K I R C V F W R Q
P I Q A R S F G T V K P F L T P L E Q D P L C V V A
K T D A S Q Q L W N G E S F S N T A R S G A F L G G S
GGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTC
10880
CCATCGCCACCAAAAAAACAAACGTTCGTCGTCTAATGCGCGTCTTTTTTTCCTAGAGTTCTTCTAGGAAACTAGAAAAG
G S G G F F V C K Q Q I T R R K K G S Q E D P L I F S
V A V V F L F A S S R L R A E K K D L K K I L . S F
W . R W F F C L Q A A D Y A Q K K R I S R R S F D L F
Y R H N K Q K C A A S . A C F F L I E L L D K S R K
P L P P K K T Q L C C I V R L F F P D . S S G K I K E
T A T T K K N A L L L N R A S F F S R L F I R Q D K R
TACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCT
10960
ATGCCCCAGACTGCGAGTCACCTTGCTTTTGAGTGCAATTCCCTAAAACCAGTACTCTAATAGTTTTTCCTAGAAGTGGA
T G S D A Q W N E N S R . G I L V M R L S K R I F T
L R G L T L S G T K T H V K G F W S . D Y Q K G S S P
Y G V . R S V E R K L T L R D F G H E I I K K D L H L
. P T Q R E T S R F S V N L S K P . S I I L F S R . R
V P D S A . H F S F E R . P I K T M L N D F L I K V .
R P R V S L P V F V . T L P N Q D H S . . F P D E G
AGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGC
11040
TCTAGGAAAATTTAATTTTTACTTCAAAATTTAGTTAGATTTCATATATACTCATTTGAACCAGACTGTCAATGGTTACG
. I L L N . K . S F K S I . S I Y E . T W S D S Y Q C
R S F . I K N E V L N Q S K V Y M S K L G L T V T N A
D P F K L K M K F . I N L K Y I . V N L V . Q L P M
S G K L N F I F N . I L R F Y I H T F K T Q C N G I S
I R K F . F H L K L D I . L I Y S Y V Q D S L . W H
L D K . I L F S T K F . D L T Y I L L S P R V T V L A
AhdI
TTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAAC
11120
AATTAGTCACTCCGTGGATAGAGTCGCTAGACAGATAAAGCAAGTAGGTATCAACGGACTGAGGGGCAGCACATCTATTG
L I S E A P I S A I C L F R S S I V A . L P V V . I T
. S V R H L S Q R S V Y F V H P . L P D S P S C R .
L N Q . G T Y L S D L S I S F I H S C L T P R R V D N
L . H P V . R L S R D I E N M W L Q R V G R R T S L
K I L S A G I E A I Q R N R E D M T A Q S G T T Y I V
. D T L C R D . R D T . K T . G Y N G S E G D H L Y S
TACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTAT
11200
ATGCTATGCCCTCCCGAATGGTAGACCGGGGTCACGACGTTACTATGGCGCTCTGGGTGCGAGTGGCCGAGGTCTAAATA
T I R E G L P S G P S A A M I P R D P R S P A P D L
L R Y G R A Y H L A P V L Q . Y R E T H A H R L Q I Y
Y D T G G L T I W P Q C C N D T A R P T L T G S R F I
. S V P P S V M Q G W H Q L S V A L G V S V P E L N I
V I R S P K G D P G L A A I I G R S G R E G A G S K D
R Y P L A . W R A G T S C H Y R S V W A . R S W I .


CAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAAT
11280
GTCGTTATTTGGTCGGTCGGCCTTCCCGGCTCGCGTCTTCACCAGGACGTTGAAATAGGCGGAGGTAGGTCAGATAATTA
S A I N Q P A G R A E R R S G P A T L S A S I Q S I N
Q Q . T S Q P E G P S A E V V L Q L Y P P P S S L L I
S N K P A S R K G R A Q K W S C N F I R L H P V Y .
L L L G A L R F P R A C F H D Q L K I R R W G T . . N
A I F W G A P L A S R L L P G A V K D A E M W D I L
. C Y V L W G S P G L A S T T R C S . G G G D L R N I
TGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGT
11360
ACAACGGCCCTTCGATCTCATTCATCAAGCGGTCAATTATCAAACGCGTTGCAACAACGGTAACGATGTCCGTAGCACCA
C C R E A R V S S S P V N S L R N V V A I A T G I V V
V A G K L E . V V R Q L I V C A T L L P L L Q A S W
L L P G S . S K . F A S . . F A Q R C C H C Y R H R G
N G P L . L L Y N A L . Y N A C R Q Q W Q . L C R P
Q Q R S A L T L L E G T L L K R L T T A M A V P M T T
T A P F S S Y T T R W N I T Q A V N N G N S C A D H H
GTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGT
11440
CAGTGCGAGCAGCAAACCATACCGAAGTAAGTCGAGGCCAAGGGTTGCTAGTTCCGCTCAATGTACTAGGGGGTACAACA
S R S S F G M A S F S S G S Q R S R R V T . S P M L
C H A R R L V W L H S A P V P N D Q G E L H D P P C C
V T L V V W Y G F I Q L R F P T I K A S Y M I P H V V
T V S T T Q Y P K M . S R N G V I L A L . M I G W T T
D R E D N P I A E N L E P E W R D L R T V H D G M N H
. A R R K T H S . E A G T G L S . P S N C S G G H Q
GCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATG
11520
CGTTTTTTCGCCAATCGAGGAAGCCAGGAGGCTAGCAACAGTCTTCATTCAACCGGCGTCACAATAGTGAGTACCAATAC
C K K A V S S F G P P I V V R S K L A A V L S L M V M
A K K R L A P S V L R S L S E V S W P Q C Y H S W L W
Q K S G . L L R S S D R C Q K . V G R S V I T H G Y
C F L P . S R R D E S R Q . F Y T P R L T I V . P . P
L F A T L E K P G G I T T L L L N A A T N D S M T I
A F F R N A G E T R R D N D S T L Q G C H . . E H N H
GCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATT
11600
CGTCGTGACGTATTAAGAGAATGACAGTACGGTAGGCATTCTACGAAAAGACACTGACCACTCATGAGTTGGTTCAGTAA
A A L H N S L T V M P S V R C F S V T G E Y S T K S F
Q H C I I L L L S C H P . D A F L . L V S T Q P S H
G S T A . F S Y C H A I R K M L F C D W . V L N Q V I
L V A Y N E . Q . A M R L I S K Q S Q H T S L W T M
A A S C L E R V T M G D T L H K E T V P S Y E V L D N
C C Q M I R K S D H W G Y S A K R H S T L V . G L . E
CTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTT
11680
GACTCTTATCACATACGCCGCTGGCTCAACGAGAACGGGCCGCAGTTATGCCCTATTATGGCGCGGTGTATCGTCTTGAA
. E . C M R R P S C S C P A S I R D N T A P H S R T
S E N S V C G D R V A L A R R Q Y G I I P R H I A E L
L R I V Y A A T E L L L P G V N T G . Y R A T . Q N F
R L I T Y A A V S N S K G P T L V P Y Y R A V Y C F K
Q S Y H I R R G L Q E Q G A D I R S L V A G C L L V K
S F L T H P S R T A R A R R . Y P I I G R W M A S S
TAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATG
11760
ATTTTCACGAGTAGTAACCTTTTGCAAGAAGCCCCGCTTTTGAGAGTTCCTAGAATGGCGACAACTCTAGGTCAAGCTAC
L K V L I I G K R S S G R K L S R I L P L L R S S S M
. K C S S L E N V L R G E N S Q G S Y R C . D P V R C
K S A H H W K T F F G A K T L K D L T A V E I Q F D
L L A . . Q F V N K P A F V R L S R V A T S I W N S T
F T S M M P F R E E P R F S E L I K G S N L D L E I
. F H E D N S F T R R P S F E . P D . R Q Q S G T R H


TAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCA
11840
ATTGGGTGAGCACGTGGGTTGACTAGAAGTCGTAGAAAATGAAAGTGGTCGCAAAGACCCACTCGTTTTTGTCCTTCCGT
. P T R A P N . S S A S F T F T S V S G . A K T G R Q
N P L V H P T D L Q H L L L S P A F L G E Q K Q E G
V T H S C T Q L I F S I F Y F H Q R F W V S K N R K A
V W E H V W S I K L M K . K . W R K Q T L L F L F A
Y G V R A G L Q D E A D K V K V L T E P H A F V P L C
L G S T C G V S R . C R K S E G A N R P S C F C S P L
AAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCA
11920
TTTACGGCGTTTTTTCCCTTATTCCCGCTGTGCCTTTACAACTTATGAGTATGAGAAGGAAAAAGTTATAATAACTTCGT
N A A K K G I R A T R K C . I L I L F L F Q Y Y . S
K M P Q K R E . G R H G N V E Y S Y S S F F N I I E A
K C R K K G N K G D T E M L N T H T L P F S I L L K H
F H R L F P F L P S V S I N F V . V R G K E I N N F C
F A A A F F P I L A V R F H Q I S M S K R K . Y . Q L M
I G C F L S Y P R C P F T S Y E Y E E K K L I I S A
TTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACA
12000
AAATAGTCCCAATAACAGAGTACTCGCCTATGTATAAACTTACATAAATCTTTTTATTTGTTTATCCCCAAGGCGCGTGT
I Y Q G Y C L M S G Y I F E C I . K N K Q I G V P R T
F I R V I V S . A D T Y L N V F R K I N K . G F R A H
L S G L L S H E R I H I . M Y L E K . T N R G S A H
K D P N N D . S R I C I Q I Y K S F Y V F L P E A C M
. . P . Q R M L P Y M N S H I . F F L C I P T G R V
N I L T I T E H A S V Y K F T N L F I F L Y P N R A C
TTTCCCCGAAAAGTGCCACCTGACGTCTAAGAAACCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAG
12080
AAAGGGGCTTTTCACGGTGGACTGCAGATTCTTTGGTAATAATAGTACTGTAATTGGATATTTTTATCCGCATAGTGCTC
F P R K V P P D V . E T I I I M T L T Y K N R R I T R
F P E K C H L T S K K P L L S . H . P I K I G V S R
I S P K S A T . R L R N H Y Y H D I N L . K . A Y H E
E G F L A V Q R R L F W . . . S M L R Y F Y A Y . S
N G R F T G G S T . S V M I I M V N V . L F L R I V L
K G S F H W R V D L F G N N D H C . G I F I P T D R P
GCCCTTTCGTCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTC
12160
CGGGAAAGCAGAGCGCGCAAAGCCACTACTGCCACTTTTGGAGACTGTGTACGTCGAGGGCCTCTGCCAGTGTCGAACAG
P F R L A R F G D D G E N L . H M Q L P E T V T A C
G P F V S R V S V M T V K T S D T C S S R R R S Q L V
A L S S R A F R . . R . K P L T H A A P G D G H S L S
A R E D R A N R H H R H F G R V C A A G P S P . L K D
G K R R A R K P S S P S F R Q C M C S G S V T V A Q R
G K T E R T E T I V T F V E S V H L E R L R D C S T
TGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCTGGCTTAACTAT
12240
ACATTCGCCTACGGCCCTCGTCTGTTCGGGCAGTCCCGCGCAGTCGCCCACAACCGCCCACAGCCCCGACCGAATTGATA
L . A D A G S R Q A R Q G A S A G V G G C R G W L N Y
C K R M P G A D K P V R A R Q R V L A G V G A G L T M
V S G C R E Q T S P S G R V S G C W R V S G L A . L
T L P H R S C V L G D P R T L P H Q R T D P S A . S H
Y A S A P L L C A R . P A D A P T P P H R P Q S L .
Q L R I G P A S L G T L A R . R T N A P T P A P K V I
GCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCG
12320
CGCCGTAGTCTCGTCTAACATGACTCTCACGTGGTATACGCCACACTTTATGGCGTGTCTACGCATTCCTCTTTTATGGC
A A S E Q I V L R V H H M R C E I P H R C V R R K Y R
R H Q S R L Y . E C T I C G V K Y R T D A . G E N T
C G I R A D C T E S A P Y A V . N T A Q M R K E K I P
P M L A S Q V S L A G Y A T H F V A C I R L S F I G
A A D S C I T S L T C W I R H S I G C L H T L L F Y R
R C . L L N Y Q S H V M H P T F Y R V S A Y P S F V A

```
CATCAGGCGCCATTCGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGC
+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++| 12400
GTAGTCCGCGGTAAGCGGTAAGTCCGACGCGTTGACAACCCTTCCCGCTAGCCACGCCCGGAGAAGCGATAATGCGGTCG
  I  R  R  H  S  P  F  R  L  R  N  C  W  E  G  R  S  V  R  A  S  S  L  L  R  Q
A  S  G  A  I  R  H  S  G  C  A  T  V  G  K  G  D  R  C  G  P  L  R  Y  Y  A  S
 H  Q  A  P  F  A  I  Q  A  A  Q  L  L  G  R  A  I  G  A  G  L  F  A  I  T  P  A
C  .  A  G  N  A  M  .  A  A  C  S  N  P  L  A  I  P  A  P  R  K  A  I  V  G  A
 M  L  R  W  E  G  N  L  S  R  L  Q  Q  S  P  R  D  T  R  A  E  E  S  N  R  W  S

  D  P  A  M  R  W  E  P  Q  A  V  T  P  F  P  S  R  H  P  G  R  R  .  .  A  L
TGGCGAAAGGGGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACG
+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++| 12480
ACCGCTTTCCCCCTACACGACGTTCCGCTAATTCAACCCATTGCGGTCCCAAAAGGGTCAGTGCTGCAACATTTTGCTGC
L  A  K  G  G  C  A  A  R  R  L  S  W  V  T  P  G  F  S  Q  S  R  R  C  K  T  T
 W  R  K  G  D  V  L  Q  G  D  .  V  G  .  R  Q  G  F  P  S  H  D  V  V  K  R  R
  G  E  R  G  M  C  C  K  A  I  K  L  G  N  A  R  V  F  P  V  T  T  L  .  N  D
 P  S  L  P  I  H  Q  L  A  I  L  N  P  L  A  L  T  K  G  T  V  V  N  Y  F  S  P
  A  F  P  P  H  A  A  L  R  N  L  Q  T  V  G  P  N  E  W  D  R  R  Q  L  V  V

Q  R  F  P  S  T  S  C  P  S  .  T  P  Y  R  W  P  K  G  L  .  S  T  T  F  R  R
GCCAGTGCC
+++++++++                                                                        12489
CGGTCACGG
 A  S  A
  P  V  P
G  Q  C
  W  H  W
A  L  A

 G  T  G
```

# APPENDIX 2

## Sequences of primers used in RT PCR experiments (5' to 3')

1) **Fbox-ex1:**
CTTCGCCTGCTTCCACTTACTTGC

2) **Fbox-ex2:**
GGGACTGAGCACAACTACTAGAT

3) **EST1F:**
ATAAAGGTGACTCCCTAGTGCTGG

4) **EST-H3-1F:**
GACCAGTTTAGCAGATGTGTC

5) **H3-10:**
ACTGCCTCAGGGGAGTTTGAG

6) **5B1:**
CGTCGCTGTCTTGGTGTGCT

7) **5C1:**
GAGGACTGCCGCCACCACCGC

8) **1B3:**
ATCAGAAACAGTTCCAGTCAGCC

9) **BAT2D-1:**
CAGCCAGTCTAGCAAAATGAACAG

10) **Phlda2-ex1F:**
GGTATGGAAGAAGAAGCGCTGCGT

11) **Phlda2-ex2R1:**
GCTCAACTGGTCCCGTGCGTTT

12) **Ylpm1R:**
TCCTTTTAGTACCAGTCCTCAAGC

13) **Ylpm1F:**
GCCACCAGCCACCTCCAGTTCCTC

14) **Rfx4-1R:**
GGATCACCGTTCGAGATGCCT

15) **Rfx4-1F:**

TACTGGACACTGTAATAAGAGCCA

**16) Tmem57-1R:**
TGCAGTGAGCAGCAAACGGTCGAC

**17) Tmem57-1F:**
GAGGGTATCTACGGCAGTACA

## Sequences of primers used for sequencing of vector pEHygro2neoSD2 (5' to 3')

**1) Hygro1F:**
TGTCGAGAAGTTTCTGATCG

**2) Hygro1R:**
CGATCAGAAACTTCTCGACA

**3) Hygro2R:**
GTACTCGCCGATAGTGGAAA

**4) Globin1F:**
CATCATTTTGGCAAAGAATT

**5) Globin1R:**
AATTCTTTGCCAAAATGATG

**6) Globin2F:**
GTGCTGTCTCATCATTTTGGCA

**7) End1F:**
ATGGGAAGATGTCCCTTGTA

**8) End2F:**
ACTCTGTTGACAACCATTGT

**9) End3F:**
AAGGCAATGAGGGTATATTA

**10) Vec1:**
CACAGGAAACAGCTATGACC

**11) Vec2:**
GACCATGATTACGCCAAGCTT

**12) Vec3:**
TTATCCGCTCACAATTCCAC

**13) Vec4:**

TACGAGCCGGAAGCATAAAG

**14) Vec5:**
TATAGTCCTGTCGGGTTTCG

**15) Vec8:**
AATCGACGCTCAAGTCAGAG

**16) SA1:**
TCCTCTTCCCATGAATTCCA

**17) SA2:**
TACTTTCGGTTCCTCTTCCC

**18) EGFP1R:**
TTCTCGTTGGGGTCTTTGCT

**19) lacZ5R:**
ATTCAGGCTGCGCAACTGTT

**20) lacZ4R:**
GTGTAGTCGGTTTATGCAGC

**21) Amp1:**
CATCCATAGTTGCCTGACTC

**22) Amp2:**
ATACGGGAGGGCTTACCATCT

**23) Amp3:**
TCGTCGTTTGGTATGGCTTC

**24) Amp4:**
GCGCCACATAGCAGAACTTT

**25) Amp5:**
CAGTTCGATGTAACCCACTC

**26) Hygro3F:**
GGACCGATGGCTGTGTAGAA

**27) AHygro1:**
CAGGATAGAGTAGATGCCGA

**28) AHygro2:**
AGAACGCCTCAGCCAGCAACT

**29) AHygro3:**
CTAAGAAGGGTGAGAACAG

**30) Ahygro4:**
GAATGGAAGGATTGGAGCTAC

**31) ActinP1F:**
TGGACATCTCTTGGGCACTGA

**32) ActinP1R:**
TCAGTGCCCAAGAGATGTCCA

**33) ActinP2F:**
TTCTGCAGATCTGCAGGACC

# APPENDIX 3

## Detailed Analysis of 5'Race Sequence Tags

**1) RACE tag ID:** 5B1
**Vector:** pEGeo2
**Sequence:**
73 bp
CTCTNTAGGCGTCGCTCTCTTGGTGTGCTTGTTCTTANCCCGGTCAATG
ATTTCAGGTACTTTGTTGATGGAG

**Chromosome Position:** 4 (-): 131624739-131625709 (Ensembl v39)

**Top BLAST hits (NCBI BLASTN):**
***Sequence in red represents the region of homology to the BLAST hits**
**Accession:** AK076188
**Description:** Mus musculus 18 days pregnant adult female placenta and extra embryonic tissue cDNA, RIKEN full-length enriched library, clone:3830421G02 product: Mus musculus RNA transcript from U17 small nucleolar RNA host gene, full insert sequence
**% Match:** 95 (70/73 bp)

**Accession:** BC100513
**Description:** Mus musculus RNA, U17d small nucleolar, mRNA (cDNA clone IMAGE:30918832)
**% Match:** 95 (70/73 bp)

**Accession:** AJ006837
**Description:** Mus musculus RNA transcript from U17 small nucleolar RNA host gene
**% Match:** 95 (70/73 bp)

**Gene Information**
**Official Symbol:** Rnu17d **and Name:** RNA, U17d small nucleolar [*Mus musculus*]
**Other Aliases:** U17HG
**Chromosome:** 4; **Location:** 4 D2.3
**GeneID:** 399101

**Function:** No protein-coding potential. Transcript serves as a vehicle for the biogenesis of small nucleolar RNAs (snoRNAs) of the H/ACA-box class (Pelczar and Filipowicz, 1998).

**Predicted translation:**

```
CT CTC TAG GCG TCG CTC TCT TGG TGT GCT TGT TCT TGA CCC
    L   *   A   S   L   S   W   C   A   C   S   *   P
GGT CAA TGA TTT CAG GTA CTT TGT TGA TGG AGG TCC CAG GTC
 G   Q   *   F   Q   V   L   C   *   W   R   S   Q   V
CCG AAA ACC AAA GAA GAA GAA CGC AGA TCT GGA GGA GGA GGA
  P   K   T   K   E   E   E   R   R   S   G   G   G   G
GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GTG AGC
 G   G   G   G   G   G   G   G   G   G   G   G   V   S
AAG GGC GAG GAG CTG
 K   G   E   E   L
```

**Sequence in black:** RACE product sequence and predicted translation
**Sequence in pink:** Vector's *En*-2 splice acceptor exon
**Sequence in blue:** stretch of glycines
**Sequence in green:** beginning of egfp coding region

i

**2) RACE tag ID:** 5C1
**Vector:** pEGeo2
**Sequence:**
32 bp
CTGCTGTGTGAGGACTNCNGCCACCACCGCTG

**Chromosome Position:** 5 (+): 125679658-125679689 (Ensembl v39)

**Top BLAST hits (NCBI BLASTN):**
**Accession:** BC008661
**Description:** Mus musculus ubiquitin C, mRNA (cDNA clone IMAGE: 3598724), complete cds
**% Match:** 93 (30/32)

**Accession:** AF285161
**Description:** Mus musculus cell-line C3H/He polyubiquitin C (Ubc) gene, complete cds
**% Match:** 93 (30/32)

**Gene Information**
**Official Symbol:** Ubc **and Name:** ubiquitin C [*Mus musculus*]
**Other Aliases:** 2700054O04Rik, AI194771, TI-225
**Other Designations:** polyubiquitin C
**Chromosome:** 5; **Location:** 5 64.0 cM
**GeneID:** 22190
**Function:** Ubiquitin protein serves as a tag in the selective proteolysis of abnormal/foreign proteins by the 26S proteasome. Gene contains multiple tandem ubiquitin coding regions and the resulting protein is processed to ubiquitin monomers.

**Predicted translation:**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CTG | CTG | TGT | GAG | GAC | TGC | CGC | CAC | CAC | CGC | TGG | TCC | CAG | GTC | |
| L | L | C | E | D | C | R | H | H | R | W | S | Q | V | |
| CCG | AAA | ACC | AAA | GAA | GAA | GAA | CGC | AGA | TCT | GGA | GGA | GGA | GGA | |
| P | K | T | K | E | E | E | R | R | S | G | G | G | G | |
| GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GTG | AGC | |
| G | G | G | G | G | G | G | G | G | G | G | G | V | S | |
| AAG | GGC | GAG | GAG | CTG | | | | | | | | | | |
| K | G | E | E | L | | | | | | | | | | |

**3) RACE tag ID:** 1B3
**Vector:** pEGeo2
**Sequence:**
76 bp
GTATCAGAAACAGTTCCAGTCAGCCCCCGCCACAGTGCGCATGGCAC
AGCCGTTTCCTGCACAGTTTGCACCCCAG
**Chromosome Position:** 1 (+):164513366-164513464 bp (Ensembl v39)

**Top BLAST hits (NCBI BLASTN):**
**Accession:** XM_917352
**Description:** PREDICTED: Mus musculus BAT2 domain containing 1, transcript variant 9 (Bat2d), mRNA
**% Match:** 100 (76/76)

**Accession:** XM_001002638
**Description:** PREDICTED: Mus musculus BAT2 domain containing 1, transcript variant 12 (Bat2d), mRNA
**% Match:** 100 (76/76)

**Accession:** BC099612
**Description:** Mus musculus BAT2 domain containing 1, mRNA (cDNA clone IMAGE: 3417702), partial cds
**% Match:** 100 (76/76)

**Gene Information**
**Official Symbol:** Bat2d **and Name:** BAT2 domain containing 1 [*Mus musculus*]
**Other Aliases:** 1810043M20Rik, 9630039I18Rik, A630006J20, Bat2d1, E130112L15Rik, mKIAA1096
**Chromosome:** 1; **Location:** 1 H1
**GeneID:** 226562
**Function:** Unknown
**Predicted translation:**

```
TGA ACA GTG TTG TGT ATC AGA AAC AGT TCC AGT CAG CCC CCG
 *   T   V   L   C   I   R   N   S   S   S   Q   P   P
CCA CAG TGC GCA TGG CAC AGC CGT TTC CTG CAC AGT TTG CAC
 P   Q   C   A   W   H   S   R   F   L   H   S   L   H
CCC AGG TCC CAG GTC CCG AAA ACC AAA GAA GAA GAA CGC AGA
 P   R   S   Q   V   P   K   T   K   E   E   E   R   R
TCT GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA
 S   G   G   G   G   G   G   G   G   G   G   G   G   G
GGA GGA GGA GTG AGC AAG GGC GAG GAG CTG
 G   G   G   V   S   K   G   E   E   L
```

**Sequence in orange:** Upstream endogenous sequence from *Bat2d* exon 21

**4) RACE tag ID:** 1A1
**Vector:** pEGeo2
**Sequence:**
82 bp
CTTTCGGAGCTGTGCGNCATTCTGAGCAGGAATGGCAATGTGGACCT
CCGTGATGGGACATCTTGTGGGATCTCACAGCCAG

**Chromosome Position:** 1 (+): 162871843-162872481 (Ensembl v39)

**Top BLAST hits (NCBI BLASTN):**
**Accession:** BC059726
**Description:** Mus musculus growth arrest specific 5, mRNA (cDNA clone IMAGE: 6582586)
**% Match:** 96 (79/82)

**Gene Information**
**Official Symbol:** Gas5 **and Name:** growth arrest specific 5 [*Mus musculus*]
**Other Aliases:** Gas-5, MGC6251
**Chromosome:** 1; **Location:** 1 H2.1
**GeneID:** 14455
**Function:** No protein-coding potential. Locus transcribes snoRNAs (class box C/D)-similar to Rnu17d (see RACE tag 5B1 above).

**Predicted translation:**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TTC | GGA | GCT | GTG | CGN | CAT | TCT | GAG | CAG | GAA | TGG | CAA | TGT | GGA |
| F | G | A | V | R | H | S | E | Q | E | W | Q | C | G |
| CCT | CCG | TGA | TGG | GAC | ATC | TTG | TGG | GAT | CTC | ACA | GCC | AGG | TCC |
| P | P | * | W | D | I | L | W | D | L | T | A | R | S |
| CAG | GTC | CCG | AAA | ACC | AAA | GAA | GAA | GAA | CGC | AGA | TCT | GGA | GGA |
| Q | V | P | K | T | K | E | E | E | R | R | S | G | G |
| GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA |
| G | G | G | G | G | G | G | G | G | G | G | G | G | G |
| GTG | AGC | AAG | GGC | GAG | GAG | CTG | | | | | | | |

**5) RACE tag ID:** 1A6
**Vector:** pEGeo2
**Sequence:**
121 bp
GGTCCCTGGAGGAGAGAGCAAGGAAGGAGAGGAAGAGTTTTTTTAC
CAGCTTTGGAGGGAGAAGAACCCGGCAGCCATGTGAGGTCTCTGGA
GGAGCTGGAGCCTGTGGCCACTATTACAG

**Chromosome Position:** 3 (-): 108083536-108083657 (Ensembl v39)

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AC079042
**Description:** Mus musculus strain C57BL/6J chromosome 3 clone rp23-313f7, complete sequence
**% Match:** 100 (121/121)

**Accession:** BY727029
**Description:** BY727029 RIKEN full-length enriched, adult male corpora quadrigemina Mus musculus cDNA clone B230112L17 5', mRNA sequence.
**% Match:** matched 89% identity over 60% length (last 73 bp) of the 1A6 sequence tag.

**Predicted translation:**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GGT | CCC | TGG | AGG | AGA | GAG | CAA | GGA | AGG | AGA | GGA | AGA | GTT | TTT |
| G | P | W | R | R | E | Q | G | R | R | G | R | V | F |
| TTA | CCA | TGC | TTT | GGA | GGG | AGA | AGA | ACC | CGG | CAG | CCA | TGT | GAG |
| L | P | C | F | G | G | R | R | T | R | Q | P | C | E |
| GTC | TCT | GGA | GGA | GCT | GGA | GCC | TGT | GGC | CAC | TAT | TAC | AGG | TCC |
| V | S | G | G | A | G | A | C | G | H | Y | Y | R | S |
| CAG | GTC | CCG | AAA | ACC | AAA | GAA | GAA | GAA | CGC | AGA | TCT | GGA | GGA |
| Q | V | P | K | T | K | E | E | E | R | R | S | G | G |
| GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA |
| G | G | G | G | G | G | G | G | G | G | G | G | G | G |
| GTG | AGC | AAG | GGC | GAG | GAG | CTG | | | | | | | |
| V | S | K | G | E | E | L | | | | | | | |

**6) RACE tag ID:** 1B2

**Vector:** pEGeo2
**Sequence:**
58 bp
TCTTCCGCCGCGNCCCGTTCCTGATGAGTGCTGTGATCCAGGATGATC
NGGCTCCCAG
**Chromosome Position:** 5 (-): 23222939-23224488 (Ensembl v39)

**Top BLAST hits (NCBI BLASTN):**
**Accession:** XM_001000526
**Description:** PREDICTED: Mus musculus expressed sequence AI506816 (AI506816), mRNA
**% Match:** matched 91% identity over 83% length (first 48 bp) of the 1B2 sequence tag.

**Accession:** AK157274
**Description:** Mus musculus activated spleen cDNA, RIKEN full-length enriched library, clone: F830209B14 product: unclassifiable, full insert
**% Match:** matched 91% identity over 83% length (first 48 bp) of the 1B2 sequence tag.

**Accession:** AK151523
**Description:** Mus musculus bone marrow macrophage cDNA, RIKEN full-length enriched library, clone:I830031F13 product:unclassifiable, full
**% Match:** matched 91% identity over 83% length (first 48 bp) of the 1B2 sequence tag.

**Gene Information**
**Official Symbol:** AI506816 **and Name:** expressed sequence AI506816 [*Mus musculus*]
**Chromosome:** 5; **Location:** 5 A3
**GeneID:** 433855
**Function:** Unknown-EST gene

**Predicted translation:**

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TTC | CGC | CGC | GGC | CCG | TTC | CTG | ATG | AGT | GCT | GTG | ATC | CAG | GAT | GAT | CTG | GCT | |
| F | R | R | G | P | F | L | **Met** | S | A | V | I | Q | D | D | L | A | |
| CCC | AGG | TCC | CAG | GTC | CCG | AAA | ACC | AAA | GAA | GAA | GAA | CGC | AGA | TCT | GGA | GGA | |
| P | R | S | Q | V | P | K | T | K | E | E | E | R | R | S | G | G | |
| GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GTG | AGC | AAG | |
| G | G | G | G | G | G | G | G | G | G | G | G | G | G | V | S | K | |
| GGC | GAG | GAG | CTG | | | | | | | | | | | | | | |
| G | E | E | L | | | | | | | | | | | | | | |

**7) RACE tag ID:** H3-10-1
**Vector:** pEHygro2neoSD2

**Sequence:**
137 bp
GTTGGAATCTGCTTCNNCAGAAGACCAGCTGAAACAAATAGCTTCGT
GGGACTGAGCACAACTACTAGATTCTTGGACTTCCGTTCACAGCTGCC
AATTGTTGGGAGTACAATAATGGAGGAGTCGGAATTGGAGAT

**Chromosome Position:** 18 (+): 85091512-85115585 (Ensembl v39)

**Top BLAST hits (NCBI BLASTN):**
**Accession:** NM_015798
**Description:** Mus musculus F-box protein 15 (Fbxo15), mRNA
**% Match:** 98 (135/137)

**Accession:** AK166599
**Description:** Mus musculus morula whole body morula cDNA, RIKEN full-length enriched library, clone: I0C0032P13 product: F-box only protein
**% Match:** 98 (135/137)

**Accession:** AK163224
**Description:** Mus musculus 2 cells egg cDNA, RIKEN full-length enriched library, clone:B020003D24 product:F-box only protein 15, full insert
**% Match:** 98 (135/137)

**Gene Information**
**Official Symbol:** Fbxo15 **and Name:** F-box protein 15 [*Mus musculus*]
**Other Aliases:** AU019763, Fbx15, ecat3
**Other Designations:** F-box only protein 15
**Chromosome:** 18; **Location:** 18 E4
**GeneID:** 50764
**Function:** Involved in phospohorylation-dependent ubiquitination through binding to E3 ubiquitin protein ligases called SCFs. Shown to be a target for Oct3/4.

**Predicted translation:**

```
GTT GGA ATC TGC TTC TAC AGA AGA CCA GCT GAA ACA AAT AGC
 V   G   I   C   F   Y   R   R   P   A   E   T   N   S
TTC GTG GGA CTG AGC ACA ACT ACT AGA TTC TTG GAC TTC CGT
```

```
  F    V    G    L    S    T    T    T    R    F    L    D    F    R
TCA  CAG  CTG  CCA  ATT  GTT  GGG  AGT  ACA  ATA  ATG  GAG  GAG  TCG
  S    Q    L    P    I    V    G    S    T    I  Met    E    E    S
GAA  TTG  GAG  ATG  TCC  CAG  GTC  CCG  AAA  ACC  AAA  GAA  GAA  GAA
  E    L    E  Met    S    Q    V    P    K    T    K    E    E    E
GAA  CGC  AGA  TCT  GGA  GGA  GGA  GGA  GGA  GGA  GGA  GGA  GGA  GGA
  E    R    R    S    G    G    G    G    G    G    G    G    G    G
GGA  GGA  GGA  GGA  GGA  GGA  GTG  AGC  AAG  GGC  GAG  GAG  CTG
  G    G    G    G    G    G    V    S    K    G    E    E    L
```

**8) RACE tag ID:** H3-10-2
**Vector:** pEHygro2neoSD2

**Sequence:**
31 bp
ACTGCCTCACCCCACTTTGAGGACTTGTAAG

**Chromosome Position:** 18 (+): 85068799-85069929 (Ensembl v39)

**Top BLAST hits(NCBI BLASTN):**
**Accession:** AC134545
**Description:** Mus musculus BAC clone RP24-470H2 from chromosome 18, complete sequence
**% Match:** 96 (30/31)

**Predicted translation:**

```
AC  TGC  CTC  ACC  CCA  CTT  TGA  GGA  CTT  GTA  AGG  TCC  CAG  GTC
    C    L    T    P    L    *    G    L    V    R    S    Q    V
CCG  AAA  ACC  AAA  GAA  GAA  GAA  CGC  AGA  TCT  GGA  GGA  GGA  GGA
 P    K    T    K    E    E    E    R    R    S    G    G    G    G
GGA  GGA  GGA  GGA  GGA  GGA  GGA  GGA  GGA  GGA  GGA  GGA  GTG  AGC
 G    G    G    G    G    G    G    G    G    G    G    G    V    S
AAG  GGC  GAG  GAG  CTG
 K    G    E    E    L
```

**9) RACE tag ID:** H3-1
**Vector:** pEHygro2neoSD2

**Sequence:**
84 bp
TAAGANNAGTTTAGCAGATCTGTCAGACACAATGACCCCNNTGGCA
GNTCTCTTAACTCCATAGCTCTACCCGAAGCTATCCAG

**Chromosome Position:** 2 (-): 180638399-180638481 (Ensembl v39)

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AK036185
**Description:** Mus musculus 16 days neonate cerebellum cDNA, RIKEN full-length enriched library, clone: 9630044K06 product: unclassifiable, full insert sequence
**% Match:** 92 (78/84)

**Accession:** AK036185
**Description:** Mus musculus 12 days embryo spinal ganglion cDNA, RIKEN full-length enriched library, clone: D130048F08 product: unclassifiable, full insert sequence
**% Match:** 92 (78/84)

**Gene Information**
**Official Symbol:** Dido1 **and Name:** death inducer-obliterator 1 [*Mus musculus*]
**Other Aliases:** 6720461J16Rik, C530043I07, D130048F08Rik, DIO-1, Datf1, Dido2, Dido3, mKIAA0333
**Other Designations:** death inducer-obliterator-2; death inducer-obliterator-3
**Chromosome:** 2; **Location:** 2 H4
**GeneID:** 23856
**Function:** Three isoforms described (Futterer et al., 2005). One of them was shown to be involved in apoptosis (Garcia-Domingo et al., 1999). Exact role of the other two is unknown. Targeting of a segment common to all three isoforms was associated with a myelodysplastic/myeloproliferative (MDS/MPD)-type disease (Futterer et al., 2005).

**Predicted translation:**
```
TAA GAC CAG TTT AGC AGA TCT GTC AGA CAC AAT GAC CAC AGT
*   D   Q   F   S   R   S   V   R   H   N   D   H   S
GGC AGT CTC TTA ACT CCA TAG CTC TAC CCG AAG CTA TCC AGG
```

```
 G   S   L   L   T   P   *   L   Y   P   K   L   S   R
TCC CAG GTC CCG AAA ACC AAA GAA GAA GAA CGC AGA TCT GGA
 S   Q   V   P   K   T   K   E   E   E   R   R   S   G
GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA
 G   G   G   G   G   G   G   G   G   G   G   G   G   G
GGA GTG AGC AAG GGC GAG GAG CTG
 G   V   S   K   G   E   E   L
```

**10) RACE tag ID:** H4H-1
**Vector:** pEHygro2neoSD2

**Sequence:**
71 bp
GAGACAGCCGCATCTTCTTGTGCAGTGCCAGCCTCGTCCCGTAGACA
AAATGGTGAAGGTCGGTGTGAACG

**Chromosome Position:** Multiple (Ensembl v39)

**Top BLAST hits (NCBI BLASTN):**
**Accession:** BC085274
**Description:** Mus musculus similar to glyceraldehyde-3-phosphate dehydrogenase, mRNA (cDNA clone MGC:103192 IMAGE:6530979), complete cds
**% Match:** 100 (71/71)

**Accession:** BC083065
**Description:** Mus musculus glyceraldehyde-3-phosphate dehydrogenase, mRNA (cDNA clone MGC:103190 IMAGE:5371133), complete cds
**% Match:** 100 (71/71)

**Accession:** NG_005470
**Description:** Mus musculus glyceraldehyde-3-phosphate dehydrogenase pseudogene (LOC654475) on chromosome 4
**% Match:** 100 (71/71)

**Accession:** NG_005467
**Description:** Mus musculus glyceraldehyde-3-phosphate dehydrogenase pseudogene (LOC654475) on chromosome 5
**% Match:** 100 (71/71)

**Accession:** NG_005469
**Description:** Mus musculus glyceraldehyde-3-phosphate dehydrogenase pseudogene (LOC654475) on chromosome X
**% Match:** 100 (71/71)

**Accession:** NG_005233
**Description:** Mus musculus glyceraldehyde-3-phosphate dehydrogenase pseudogene (LOC433921) on chromosome 5
**% Match:** 100 (71/71)

**Predicted translation:**

```
TCT CTG CTC CTC CCT GTT CCA GAG ACA GCC GCA TCT TCT TGT
 S   L   L   L   P   V   P   E   T   A   A   S   S   C
GCA GTG CCA GCC TCG TCC CGT AGA CAA AAT GGT GAA GGT CGG
 A   V   P   A   S   S   R   R   Q   N   G   E   G   R
TGT GAA CGG TCC CAG GTC CCG AAA ACC AAA GAA GAA GAA CGC
 C   E   R   S   Q   V   P   K   T   K   E   E   E   R
AGA TCT GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA GGA
 R   S   G   G   G   G   G   G   G   G   G   G   G   G
GGA GGA GGA GGA GTG AGC AAG GGC GAG GAG CTG
 G   G   G   G   V   S   K   G   E   E   L
```

**11) RACE tag ID:** H3-17
**Vector:** pEHygro2neoSD2
**Sequence:**
245 bp
CGTCCGCCAGTCACGGCCGCCGCCCCCAGCGACGTCACCCACGCGCG
CAGAAGCGGACGCCGCGGATCAAGATGTCTCTGCCATGCCCACGGG
ACGCACGGACGCANGGANGGACTCCACNAGCATCNACCGNTCACTG
CCGCCGCCNACAGTGACGTCNCCNANNAAAGCACACACNNAGNTG
NGGACGCCGTGGNCAAGATGTNTCTGCCATCCCCACAGGANGGACG
GANGGACTCCACAAG
**Chromosome Position:** Unlocalised

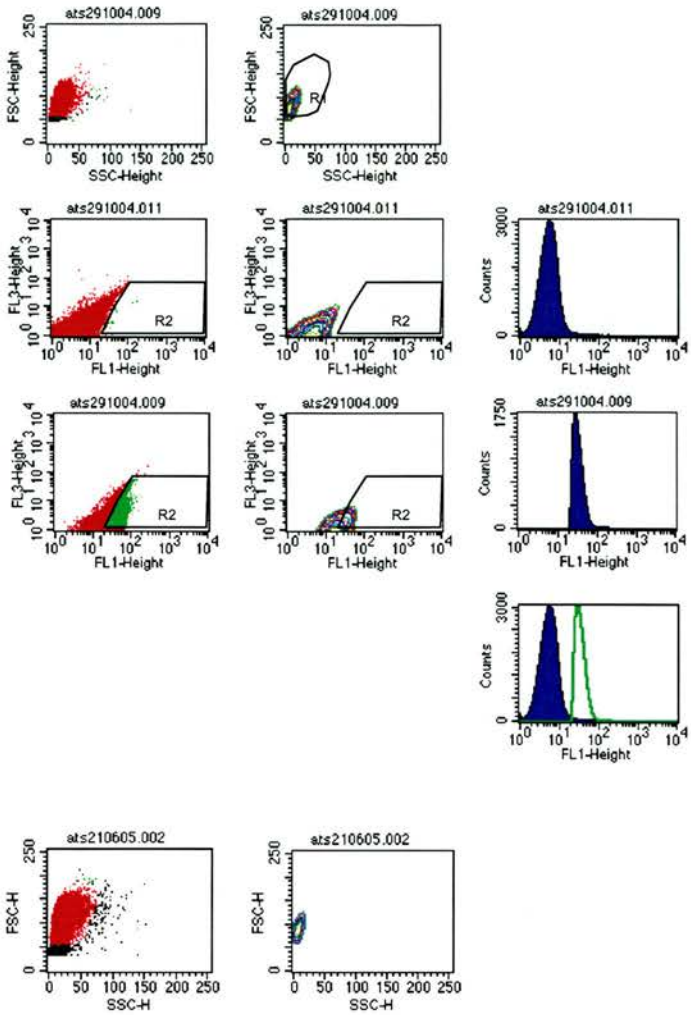
**Top BLAST hits (NCBI BLASTN):**
**Accession:** BC058113
**Description:** Mus musculus erythroid differentiation regulator 1, mRNA (cDNA clone MGC:69587 IMAGE:6820436), complete cds
**% Match:** two regions of homology-one showed 97% identity over the first 120 bp of the sequence tag (117/120) and the other 86% identity over the last 117 bp of the sequence tag (101/117).

**Gene Information**
**Interim Symbol:** Erdr1 **and Name:** erythroid differentiation regulator 1 [*Mus musculus*]
**Other Aliases:** MGC5764, edr
**Chromosome:** Un
**GeneID:** 170942
**Function:** Induction of haemoglobin synthesis, cell survival and growth control (Dormer et al., 2004).

**Predicted translation:**

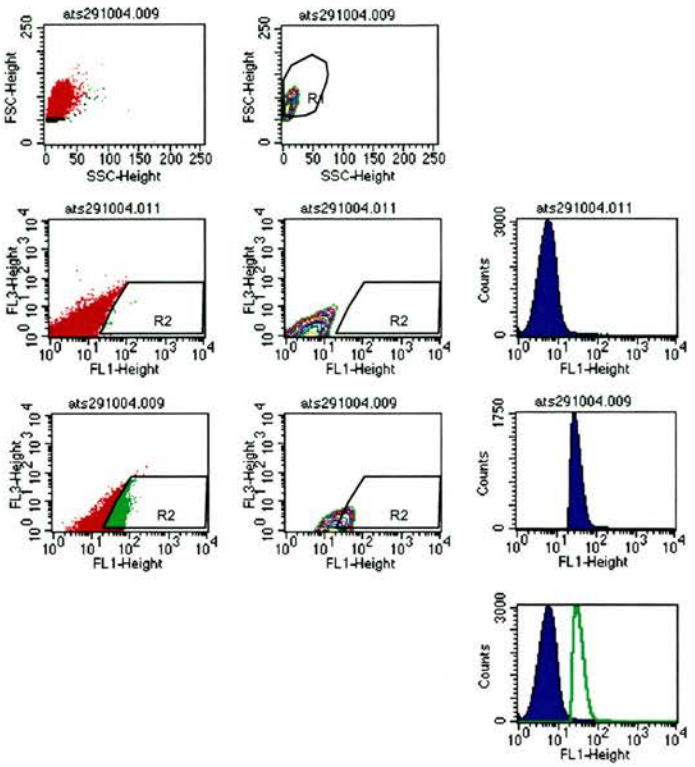
| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CGT | CCG | CCA | GTC | ACG | GCC | GCC | GCC | CCC | AGC | GAC | GTC | ACC | CAC |
| R | P | P | V | T | A | A | A | P | S | D | V | T | H |
| GCG | CGC | AGA | AGC | GGA | CGC | CGC | GGT | CAA | GAT | GTC | TCT | GCC | ATG |
| A | R | R | S | G | R | R | G | Q | D | V | S | A | **Met** |
| CCC | ACG | GGA | CGC | ACG | GAC | GCA | CGG | ACG | GAC | TCC | ACA | CCG | GTC |
| P | T | G | R | T | D | A | R | T | D | S | T | P | V |
| ACT | GCC | GCC | GCC | CAC | AGT | GAT | GTC | ACC | CAC | GAA | AGC | ACA | CAC |
| T | A | A | A | H | S | D | V | T | H | E | S | T | H |
| GTA | GAA | GCG | GAC | GCC | GTG | GTC | AAG | ATG | TCT | CTG | CCA | TCC | CCA |
| V | E | A | D | A | V | V | K | Met | S | L | P | S | P |
| CAG | GAC | GGA | CGG | ACG | GAC | TCC | ACA | AGG | TCC | CAG | GTC | CCG | AAA |
| Q | D | G | R | T | D | S | T | R | S | Q | V | P | K |
| ACC | AAA | GAA | GAA | GAA | CGC | AGA | TCT | GGA | GGA | GGA | GGA | GGA | GGA |
| T | K | E | E | E | R | R | S | G | G | G | G | G | G |
| GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GGA | GTG | AGC | AAG | GGC |
| G | G | G | G | G | G | G | G | G | G | V | S | K | G |

**Flow cytometry profiles of some of the gene trap clones analysed by 5'RACE PCR and whose expression profile is not shown in Chapter 3**

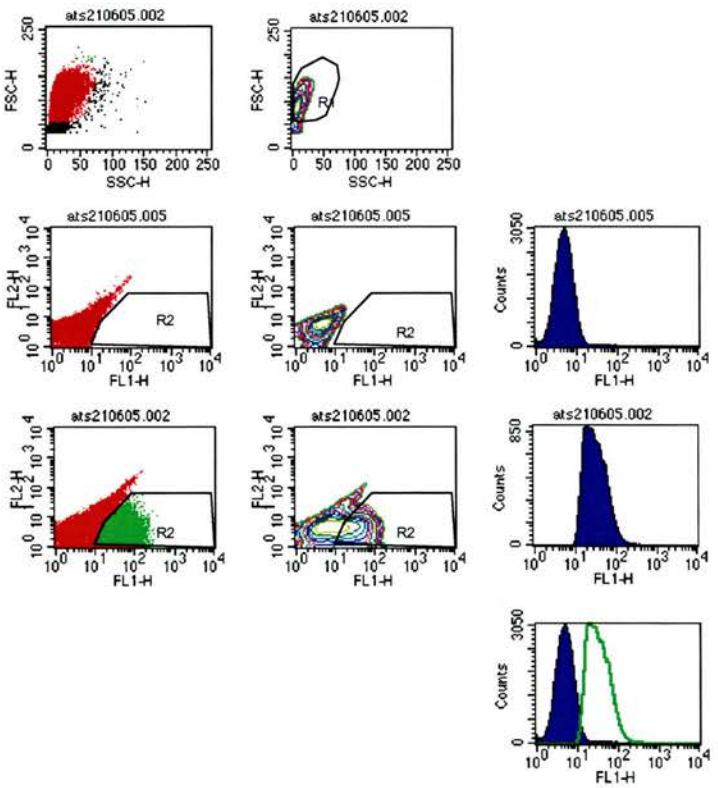


**Flow cytometry profiles of some of the gene trap clones analysed by 5'RACE PCR and whose expression profile is not shown in Chapter 3**

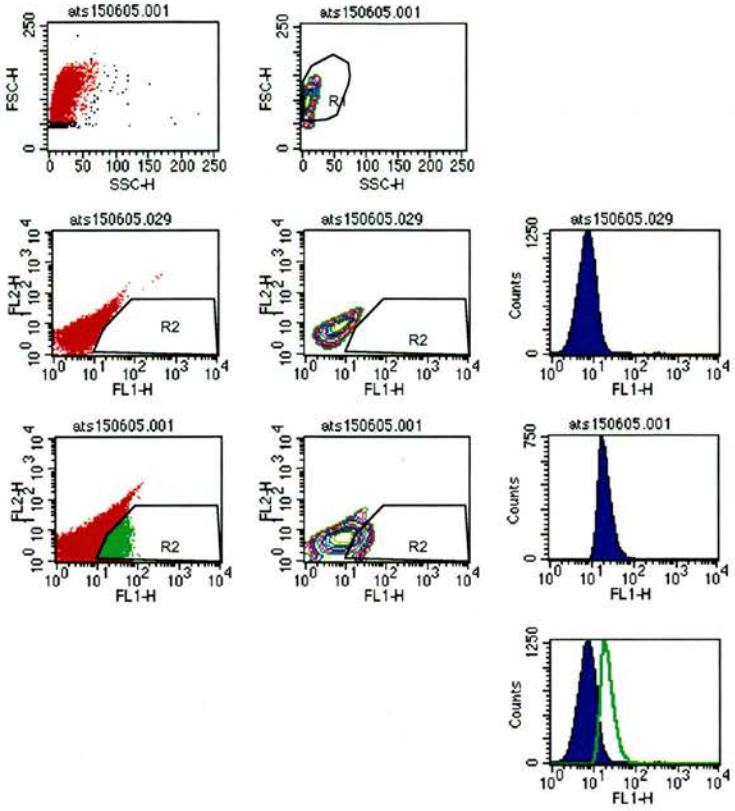
**Clone H3-1**



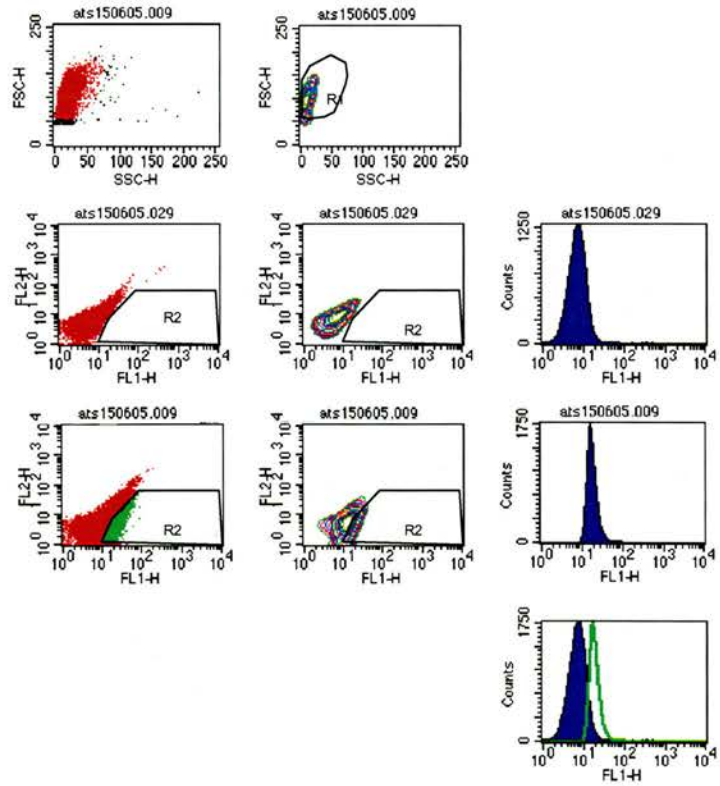

**Clone 5B1**



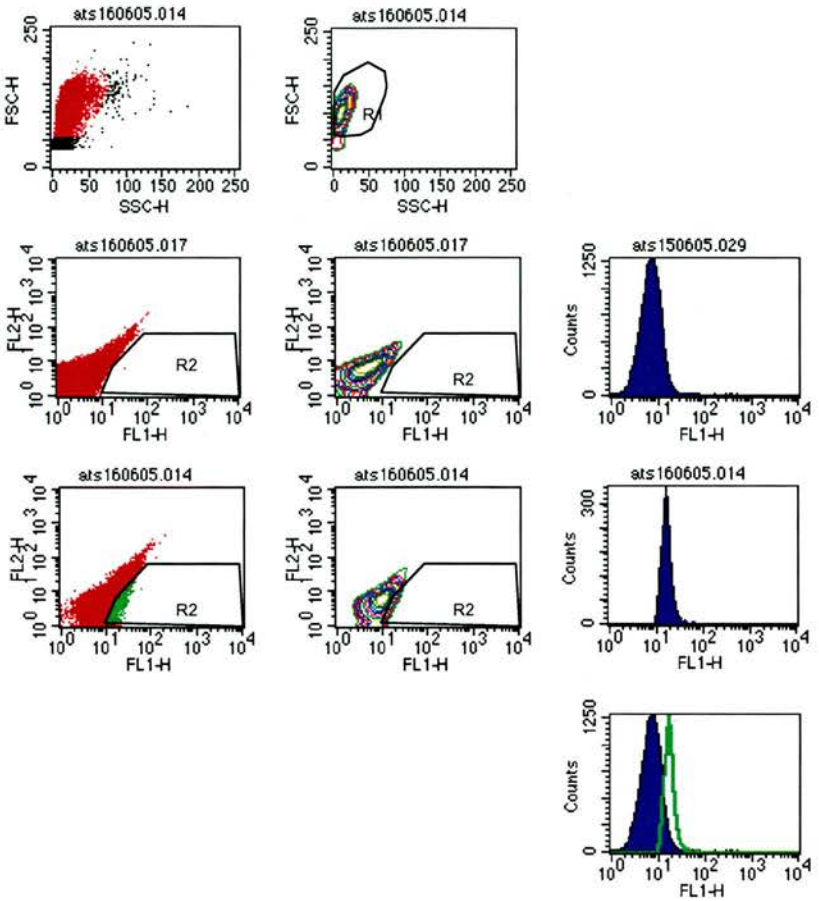
**Clone 1A1**




**Clone 1B3**



**Clone 1A6**


ats160605.014
FSC-H
SSC-H
R1
ats160605.017
FL2-H
FL1-H
R2
ats150605.029
Counts
ats160605.014
Counts

# APPENDIX 4

## 3'Race Sequence tag Analysis and Integration details for clones characterised by proper SD Function

### 1) CLONES N2 AND N29

**Vector:** pEHygro2neoSD2 (HindIII/MfeI)
**Sequence:**
**211 bp**
GAGGAGATGTAGTCTCCCCTCCCCCAGCCTGAAACCTGCTTGCTCGGGGT
GGAGCTTCCTGCTCATTCGTTCTGCCACGCCCACTGCTGGAACCTGAGGA
GCCACACACGTGCACCTTTCTACTGGACCAGAGATTATTCGGCGGGAATC
GGGTCCCCTCCCCCTTCCTTCATAACTGGTGTCACAACAATAAAATTTGA
GCCTTGATCAGAAAAAAAAAAAAAAAAAAAAA
**Chromosome Position:** Multiple (Ensembl v39)

**Top BLAST hits (NCBI BLASTN):**
Multiple high homology hits on different chromosomes
Hits are related to a sequence from 3'LTR of mouse ETn transposon
**Accession:** Y17106
**Description:** Mus musculus transposon ETn, SELH/L3A strain
**% Match:** 99% (210/211)

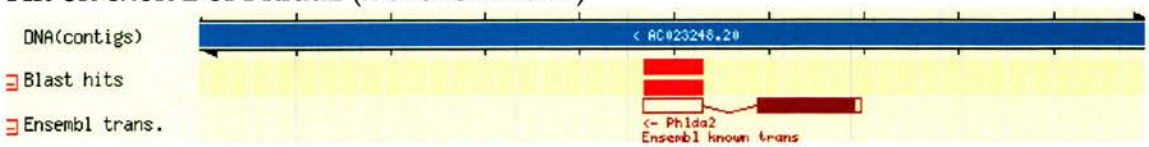
### 2) CLONE N19

**Vector:** pEHygro2neoSD2 (HindIII/MfeI)
**Sequence:**
290 bp
AAACGCACGGGACCAGTTGAGCCNGGAACCAGCCCACGCTGCCGCATCA
GACAGGAGCCCAGGAGCCNGGGGGATGCCCCGGGTGACGCCGCCGCTGC
TCTTCACCGAAGATATTCCCGTGCTTGCTGAGCCTTGCCCCGCTGTGCGG
GTGCGCTAACTTATTGGACCGTATTTATATTGGTTACCTGCTTCCAACCCA
CCATTCCGTGTTAATATTTTTTATACCATATTTTCATTCCAAATAAACAAT
GTCACTTTTCTTTTTTTAAAAAAAANNAAAAAAAAAAAAA
**Chromosome Position:** 7 (-): 131206322 -131206403

**Top BLAST hits (NCBI BLASTN):**
**Accession:** BC019141
**Description:** Mus musculus pleckstrin homology-like domain, family A, member 2, mRNA (cDNA clone MGC: 29089 IMAGE: 5042041), complete cds
**% Match:** 98 (257/260)
**Accession:** NM_009434

**Description:** Mus musculus pleckstrin homology-like domain, family A, member 2 (Phlda2), mRNA
**% Match:** 98 (257/260)

Hit on exon 2 of Phlda2 (2 exons in total)

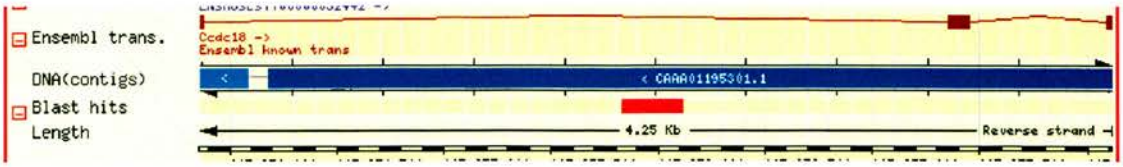



**Gene trap orientation relative to gene's transcription:** +

**Gene Information**
**Official Symbol:** Phlda2 **and Name:** pleckstrin homology-like domain, family A, member 2 [Mus musculus]
**Other Aliases:** Ipl, Tssc3
**Other Designations:** tumor-suppressing subchromosomal transferable fragment 3
**Chromosome:** 7; **Location:** 7 69.5 cM
**GeneID:** 22113

**3) CLONE N34**
**Vector:** pEHygro2neoSD2 (HindIII/MfeI)
**Sequence:**
255 bp
AATGAAGATTTTACTGGGGNTGCTTTAGGAAGCCATTAAAAATCAATATA
ATCTACGTTTATAATCAGAAACATAAATTCTATGGAAAATGGATTAGTAA
TAAAATAAGAACATATAAGGAGTCCATTAGGAAGGCCTAGCTAAACACC
ATGAAAGAAACAATAATGGTCTGTGCATAGTGGTAATATAAAGTCAATG
AATTTTATAATATATTTTAGAAATCAGGTTGACAAGATCTGCTAATAAAN
AAAAANA
**Chromosome Position:** 5 (-): 105647255-105647507

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AC159996
**Description:** Mus musculus chromosome 5, clone RP23-46J11, complete sequence
**% Match:** 99 (243/244)
BLAST hit (opposite orientation) within a region located on intron 21-22 of Ccdc18 gene (29 exons in total)




**Conservation (USCS genome server):** Region homologous to the RACE tag is conserved between mouse and rat

**4) CLONE N23**

**Vector:** pEHygro2neoSD2 (HindIII/MfeI)

**Sequence:**

186 bp

NGNTNGNTTTTANCAACTCAGAAACTNGCCNAAACCACNGATAAGTAAA
AGANATGNANAGNGNNGNACNGANAACTTGAATTAANNNTGNACTTTCT
GGGCANACGGCAGCTTGGATACAGCTAGGTTTCTTNACACAGNGATTGN
TTTTAACAACTCAGAANCNNGCCAAAACTNCNNNCANNN

**Chromosome Position (UCSC genome server):** 6 (+): 10067072-10067099 (region in red) and 6(+): 10081909-10081943 (region in blue)

**Top BLAST hits (NCBI BLASTN):**

**Accession:** AC144519

**Description:** Mus musculus BAC clone RP24-325A1 from chromosome 6, complete sequence

**% Match:** Two regions (shown in red and blue on the RACE tag) of homology-each is homologous to two different regions on chromosome 6-red: 96% (27/28)-blue: 88% (30/34)

No transcript present in the region

**Conservation (USCS genome server):** Mouse chromosome regions 6: 10067072 – 10067152 and 10081856 - 10081943 are highly conserved between mouse and rat

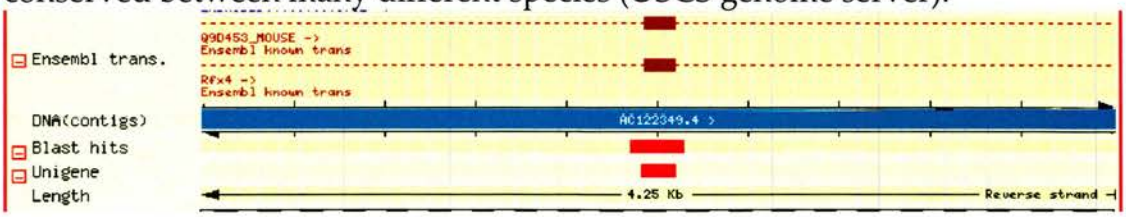
**5) CLONES: 1, 6, 7, 30, 32, 33, 35**
**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
GAGGAGATGTAGTCTCCCCTCCCCCAGCCTGAAACCTGCTTGCTCGGGGT
GGAGCTTCCTGCTCATTCGTTCTGCCACGCCCACTGCTGGAACCTGAGGA
GCCACACACGTGCACCTTTCTACTGGACCAGAGATTATTCGGCGGGAATC
GGGTCCCCTCCCCCTTCCTTCATAACTGGTGTCACAACAATAAAATTTGA
GCCTTGATCAGAAAAAAAAAAAAAAAAAAAA
Same as clones N2 and N29 (see above)

**6) CLONE 2**
**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
270 bp
AAACAGAAGGAGCACAGCCCCTGGCGGGACACCTTACACTGGATGATGA
GTCTGCGGTGCTCATCCCGAGAGTCCTCCATGGTATACAGAGTCTGCTTG
GTGATGCTACTCAGGTCCACATTCCTCCAGTCCTCCAGCATCTGGAACGT
GATGTCTGCACTGTGGATCACCGTTCGAGATGCCTGAAATCCAATCCAAC
CAACTTGTCAGCACATATGCTGTCAGGTTAAATAAAAGCCCTCCTCTCTG
TCTNAAAAAAAANAANAAAAA
**Chromosome position (USCS genome server):** 10 (-): 84292973-84293223

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AY102010
**Description:** Mus musculus winged-helix transcription factor RFX4 variant 3 (Rfx4) mRNA, complete cds, alternatively spliced
**% Match:** 100 (148/148)

**Accession:** AY342003
**Description:** Mus musculus regulatory factor X 4 variant (Rfx4) mRNA
**% Match:** 100 (148/148)

RACE tag is homologous to exon 11 of Rfx4 gene (18 exons in total)-note homology also spans flanking intronic sequence which was found to be conserved between many different species (USCS genome server).



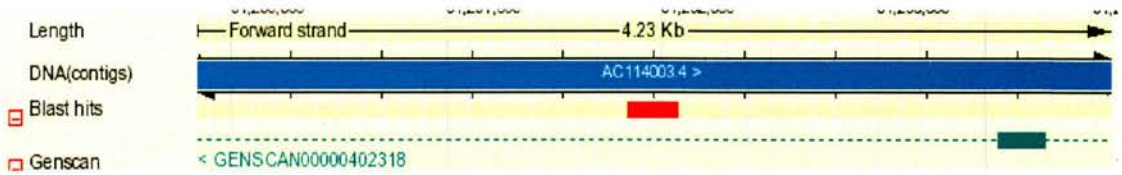

**Gene trap orientation relative to gene's transcription:** -

**Gene Information**
**Official Symbol:** Rfx4 **and Name:** regulatory factor X, 4 (influences HLA class II expression) [Mus musculus]
**Other Aliases:** 4933412G19Rik, NYD-sp10
**Other Designations:** regulatory factor X, 4; winged-helix transcription factor RFX4
**Chromosome:** 10; **Location:** 10 C1
**GeneID:** 71137

**7) CLONE 4**
**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
250 bp
GGATGCTAGACCNCAGGAAGGATGTCNCCAGCCTCACANTGAATTGCTG
CCTGTCACAGCCTTTGAAAATGTTANNTGCCANAGAAGGNNCCANCAGG
GCTTCTGCTAACCTTGTCTTATTGTTCTACCAAATAACCAAGGCGATGGC
ATGACATTCCCTACACGGAGTTTAACATATCATTTGCATTTACAGTGTAA
AGTTGTTATTAAAGGCAGGATATTTTTACGAGCAAAAAAAAAAAAAAAA
AAA
**Chromosome position (USCS genome server):** 18 (-): 34232342-34232573

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AC114003
**Description:** Mus musculus chromosome 18 clone RP23-174C24, complete sequence
**% Match:** 95 (221/231)

Hit within intron of a predicted transcript (Genscan ID: GENSCAN000000402318)




**Conservation (USCS genome server):** Mouse chromosome region 18: 34232342–34232573 is highly conserved between mouse, rat, human and dog

**8) CLONE 5**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

692 bp

ACAAGTTGGATGGCTTGAGGACTGGTACTAAAAGGAAGCGAGACTGGGA
GGCGATTGCCAGCAGAATGGAGGATTACCTTCAGCTCCCTGACGACTACG
AGACTCGCGCTTCTGAGCCTGGGAAGAAGAGGACNTGNTTTCTCTNCCTA
GNCCNGGNTGTCCTGAAACTCANGGACACCTACAACCCCTGNGATACCA
CACCCACNGANTGAAAAAAAANNANAAATGAGAGTGGTCNCCTGGCTGA
AAGAGCCCTCAATCGGACCAAATATATATGAGACTTAGTTTTGAATGGAG
TCATGTTCCTCTAAGGTGGTTCGATGTGAGGCGACCTGAAGCTCAGCCCT
CGGGAAGCTTCTTGTGTCTGGAAGCTTCTTGTGTCTCCACGTTTCAATTTT
GTTTTGTTTTACCACTTTCATTTTTAAACGTTTTCCAGTATCCCCCCCAAA
ATTTTANAATCTTGCTATATNCCAAGCAAGACATTATTAATATTTTTTTTN
AANTAATTTGGGGGAGGGAGGGTTAGCTTAANTGCTAAATCAGGNGTGG
ATCCNCNNGANGGTTCCAACNAAANTGTTTCCTCCCNTGTATNCTGNCAC
ANANTATCNTCNTTGCTTGGGGNTGGTTTTNTTTTTTANNGNTNNAAAAN
NNNNCTNATTNNNCTTGGANAAAAAACAANAAATTTTNCCTNT

**Chromosome position (USCS genome server):** 12 (+): 85954074-85959476

**Top BLAST hits (NCBI BLASTN):**

**Accession:** BC055465

**Description:** Mus musculus YLP motif containing 1, mRNA (cDNA clone IMAGE:3992501), complete cds

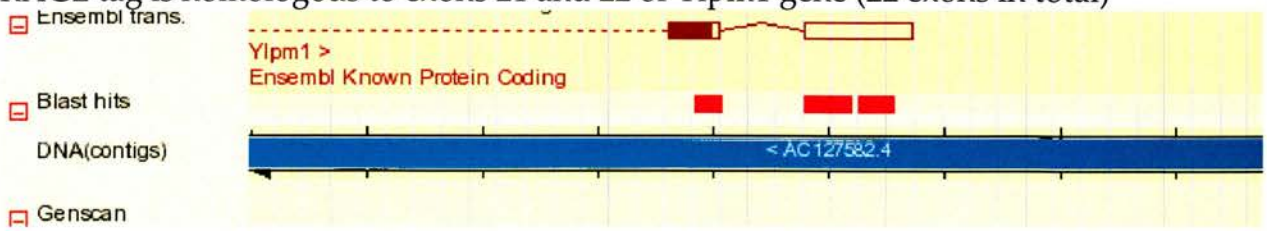
**% Match:** Three regions of homology (shown in blue, red and green)-blue: 100% (110/110), red: 99% (160/161), green: 85% (164/191)

**Accession:** NM_178363

**Description:** Mus musculus YLP motif containing 1 (Ylpm1), mRNA

**% Match:** Three regions of homology (shown in blue, red and green)-blue: 100% (110/110), red: 99% (160/161), green: 85% (164/191)

RACE tag is homologous to exons 21 and 22 of Ylpm1 gene (22 exons in total)




**Gene trap orientation relative to gene's transcription:** +

**Gene Information**

**Official Symbol:** Ylpm1 **and Name:** YLP motif containing 1 [Mus musculus]

**Other Aliases:** A930013E17Rik, AI851834, ZAP, Zap3

**Other Designations:** ZAP3 protein; nuclear protein ZAP
**Chromosome:** 12
**GeneID:** 56531

**9) CLONE 8**
**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
414 bp
GTTCCCAGACTGCCTTTTGAGAATAAAATGGGAGGCCAGAACCAAAGTC
TTTTGAATAAAGCACCACAACTCTAACCTGTTTGGCTGCCTTCCTTCCCAA
GGCACAGATCTTTCCCAGCATGGAAAAGCATGTAGCAGTTGTAGGACAC
ACTAGACGAGAGCACCAGATCTCATTGTGGGTGGTTGTGAACCACCCACC
ATGTGGTTGCTGGGATTTGAACTCANGATCTTCAGAAGAGCAGTCAGGGC
TCTAAACCGATGAGCCATCTCTCCAGCCTCCTACATTCCTTCTTAAGGCAT
GAATGATCCCAGCATGGGAAGACAGTCTGCCCTCCCTCCTTTTTGAGCCA
TTTTCCCTCTTTCACCATATACTCAATAAAATAAGTAAATGAAAAAAAAA
AAAAAAAAAAAAAN
**Chromosome Position (UCSC genome server):** 11(-): 53983451-53983810

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AF441733
**Description:** Mus musculus clone RP23-225M6 45S pre-ribosomal RNA gene, partial sequence; and intergenic spacer, partial sequence
**% Match:** 98 (388/391)

**Accession:** BK000964
**Description:** TPA_exp: Mus musculus ribosomal DNA, complete repeating unit
**% Match:** 98 (388/391)

Trapped sequence corresponds to a repeat

**10) CLONE 11**
**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
189 bp
GGAGGACCAGCAGCCCAGATCTCCTGACTCCAGCCACACATCTTTCACAG
CTCACTGGAGAGAGGTCAAATACCACCAATGCCAATGTGACTGTGGCAC
TGCCTGGCCCCAGGACAGGGCACTGAACAATCAACATTCACATGAATTA
AAAACCAGCTTTTAAACAGAAAAAAAAAAAAAAAAAACCN
**Chromosome Position (UCSC genome server):** 19 (+): 7303181-7303348

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AC140307
**Description:** Mus musculus BAC clone RP23-389D15 from chromosome 19, complete sequence
**% Match:** 100 (167/167)

**Top BLAST hits (NCBI BLASTN-mouse EST):**
**Accession:** AI506879
**Description:** vl56e06.x1 Stratagene mouse skin (#937313) Mus musculus cDNA clone IMAGE: 976258 3', mRNA sequence.
**% Match:** 100 (167/167)

**Accession:** AA562649
**Description:** vl56e06.r1 Stratagene mouse skin (#937313) Mus musculus cDNA clone IMAGE: 976258 5', mRNA sequence.
**% Match:** 99 (166/167)

Homology (Opposite orientation) with intronic area (1st intron) of known transcript Nat11 (Ensembl transcript ID: ENSMUST00000025675; 9 exons in total)


**Conservation (USCS genome server):** Mouse chromosome region 19: 7303181-7303348 is conserved between mouse, rat and human

**Gene trap orientation relative to gene's transcription:** -

**11) CLONE 12**
**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
222 bp
NTNCNCGNNCTNTCTTANCTGTCNTATGCCNGTGCTGCTTGNCTGCCTGN
NTGCTCTGCCTGTCGNNATGTTNNCTTTCCTGCTNNTNTGTGNANNTGCC
NNNNCCACACNGATAGNCAACCTNCATTTACGGGNNTTNNTCGNTAGCA
TGCAAAAGGGAATAATGGCCGTGGAGCCAACAACTTCAATAGGAAATGA
AAAAAAAAAAAAAAAANCCNNNNN
**Chromosome Position (UCSC genome server):** 11 (+): 97461499-97461582

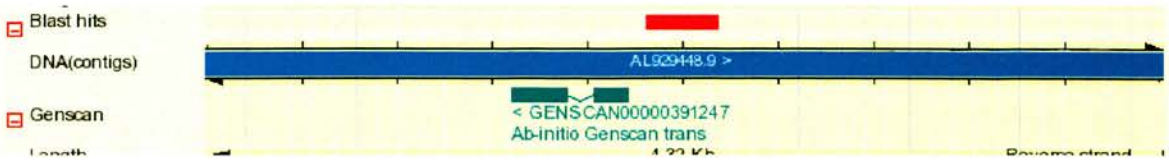
**Top BLAST hits (NCBI BLASTN):**
**Accession:** AK155239
**Description:** Mus musculus NOD-derived CD11c +ve dendritic cells cDNA, RIKEN full-length enriched library, clone: F630210K12 product: unclassifiable, full insert sequence
**% Match:** 98 (51/52) (Region in red)

**Accession:** AK052674
**Description:** Mus musculus 0 day neonate kidney cDNA, RIKEN full-length enriched library, clone: D630020N09 product: unclassifiable, full insert sequence
**% Match:** 98 (51/52) (Region in red)

**Conservation (USCS genome server):** Mouse chromosome region 11: 97461499-97461582 is conserved between mouse, rat and human

**12) CLONE 13**
**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
380 bp
ANACATTCCTNGGACATAAACAACTTCCATATGCTGCTGTTCCTGGAGGC
ACANAATGGAATGGCCTCAGCCCACAGTTGAGAGCAGCATTGACTCCAG
CCGCATATGTAGCAGAGGATGGCCTTGTTGGACATCAGTGGGAGGAGAG
GCCCTTGGTCCTGAGAAGATAAAGGCCTGGGGAATGCTTTCTCTTTAGAC
TCTACTGCTCCAGGGATGTGACAATGCAGAACAAGAAAGACCCTGGAAT
TGCAATTATGAGCTTATGAGGACGATGTTGTTTAATAGCATGCTTCAGGA
ATGCACTGCCAAAGAACTTTCAAAGACATTTGGCTGTAAAACAGTATGG
ATCCACTGGAAAACGTNNAAAAAAAAAAAAAAAA
**Chromosome Position (UCSC genome server):** 4 (-): 85492307- 85492647

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AL929448
**Description:** Mouse DNA sequence from clone RP23-313G2 on chromosome 4, complete sequence
**% Match:** 98 (319/324)
No known transcripts present in the region
Hit adjacent to ab-initio Genscan transcript GENSCAN00000391247




**Conservation (USCS genome server):** Mouse chromosome region 4: 85492307–85492493 is conserved between mouse, rat, human and dog.

**13) CLONE 16**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

701 bp

CTAGGACTTCTCTCNAACTTGTGTGCTGAGGAGACTCGATGTGGACCTCA
TCTCCTAGGCTGAANTCGGCGGACTGGCTCATGNNCACTTGTGTACTGAG
NCCNAGGAAAGGATCTAGCAGAAAGCNNCGTCTCTTATCCTGGGCTTGG
CANCTNGGAANAGGACNNGTAGTGGAAATCCTGCAATCTGAAAAGCTTA
CTGAAAGGTGACAAANAAGCTGAAGATGGGTGGCGGAGAGAGGTATAA
CATTCCAGACCCTCAATCTAGAAATGCTAGTAAGAACCNAGAACAGCAA
AATAGACAGAAGAGCAAGGATCAGAATTCTTCCCAGACGAAGATTGCTC
TTAAGAAAAAGGAACGAGGACATGGGTACAATCCAGCAGCAGCAGCATG
GCAGGCCATGCAAAATGGGGGAAAGACCAANAGCCTTTCTAACAACTCC
AACTGGAATGCTGGTTTATCAAGTCCTAGCTTGCTTTTTAAGTCTCAAGCT
AGTCAGAACTATGCTGGAGCCAAATTTAGTGAANCACCATCACCAAGTG
TTCTCCCCAAGCCACCAAGCCACTGGGTTCATGTTTCCTTGAACCCTTCCN
ATAAGGGAACGATGACNTTTCAACTTAAANCCTTACTTANGGTACAGGTA
TAAGTTAGTGTTAACTTTNAATTGNGGNCCTTACNCATGNACATCTGATT
TATGTGCN

**Chromosome Position (UCSC genome server):** 4 (-): 135144032-135145175

**Top BLAST hits (NCBI BLASTN):**

**Accession:** BC006598

**Description:** Mus musculus proline-rich nuclear receptor coactivator 2, mRNA (cDNA clone MGC:11707 IMAGE:3965195), complete cds

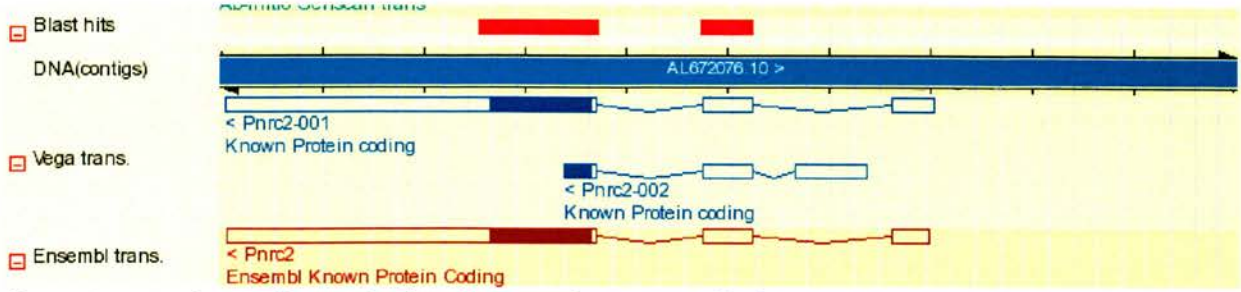
**% Match:** 93 (660/705)

**Accession:** AK077403

**Description:** Mus musculus 6 days neonate head cDNA, RIKEN full-length enriched library, clone: 5430438M09 product:similar to PROLINE-RICH NUCLEAR RECEPTOR COACTIVATOR 2 [Homo sapiens], full insert sequence

**% Match:** 93 (660/705)

RACE tag homologous to exons 2 and 3 of Pnrc2 gene (3 exons in total)




**Gene trap orientation relative to gene's transcription:** +

**Gene Information**

**Official Symbol:** Pnrc2 **and Name:** proline-rich nuclear receptor coactivator 2 [Mus musculus]

**Other Aliases:** 0610011E17Rik, D4Bwg0593e, MGC11707

**Chromosome:** 4; **Location:** 4 66.7 cM

**GeneID:** 52830

**14) CLONE 19**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

375 bp

ATGTGTGACTTTGAAATGATTGGGATCTTCTAGAGCATTCATTTTTCTACC
CCAAAAGGAGAATTTGAACCTTCTTTCAGTATTTTTGAGGACCCACCTCT
GGGACTCTTGAGCTGACNGCTTTGGNACTTGNANGNGATTTGNAAAAGA
AAACTAGAAGCCATGTCTGACTCACCTCAGTCAATAAGGTATTTGTGCTG
GCTGCAGATGTCTCTGAGCACCCAAGTAAAGAAACAATGTGGCAAACTG
CTGCAAGCTCTTGCTTCTCCATCATGAAGAACTCCATCATGATGAACATC
ATGGCTCCCTGGAACTGGGAGCCAAGAATGGACTGTGGACCTTTGAACT
GCAAGCCAAAATAAACCCTACTCCCTT

**Chromosome Position (UCSC genome server):** 8 (-): 25504433-25507388

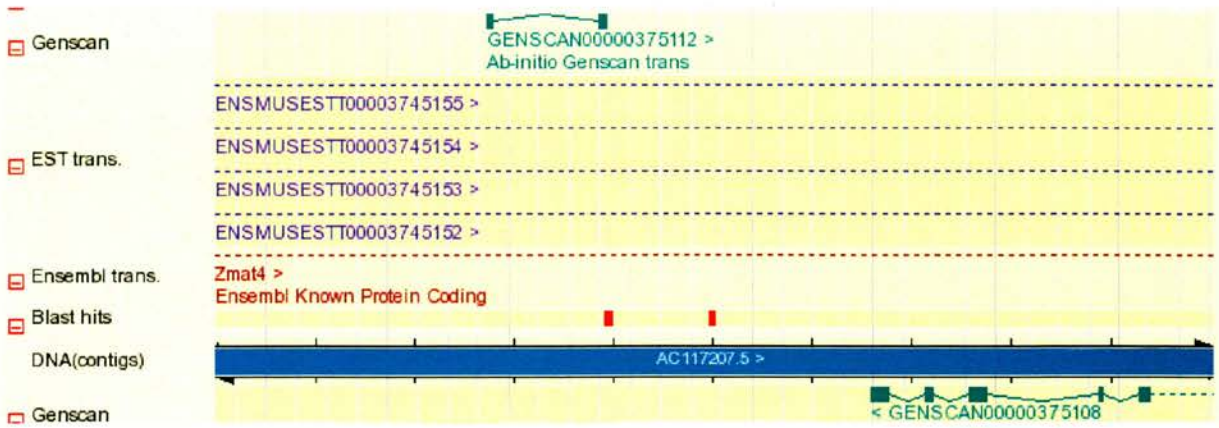
**Top BLAST hits (NCBI BLASTN):**

**Accession:** AC117207

**Description:** Mus musculus BAC clone RP23-278A7 from 8, complete sequence

**% Match:** three regions of homology (shown in blue, green and red)-blue: 100% (96/96), red: 95% (129/135), green: 100% (146/146)

BLAST hits homologous (opposite orientation) to areas within an intronic region of Zmat4 gene-also homologous to exon of ab-initio transcript GENSCAN00000375112




**Conservation (USCS genome server):** Mouse chromosome region 8: 25504433–25504653 is conserved between mouse and rat. Mouse chromosome regions 8: 25505335-25505774 and 25505843–25506091 are conserved between mouse, rat, human and dog. Mouse chromosome region 25506446-25507388 is conserved between mouse, rat and dog.

**15) CLONE 22**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

394 bp

NTTACNGACATTAACNTCTGAACATCNCTAACAAGTNTTGNAATTCACAC
CTTCAGGAAGAGAAAGCGCCAANNAAGGAGGCCTTCCCATCTCAGCAAG
ANCAGTCTCCATTGAAGAAAAAAACTTGACAAAGACCNNCAGTAACNNC
NGCCCTGTCANTCGNTAGAAATGCCANAGGAGGAAGATGAGAAACACAC
CAAGCCAATATGCATAGAAAATCTGANCCAGGTCTTTCAAAGTTCTTAAN
AGTCTATTCTAAATNCATCAGCTACAACAGCATTTAAAAGACTAAGCATC
GAATAAATGCCATTCTCAGTGAGTAGCATTGTATTAATTATCATAGAAAT
AAAATTGTTGAATGAAAAANAAAAAAAAAAAAAAAAAAAAANCNTT

**Chromosome Position (UCSC genome server):** 3 (+): 127531848-127532209

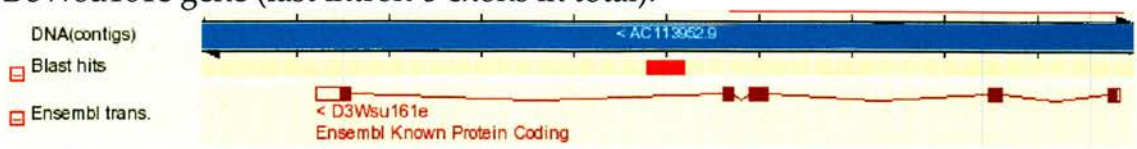
**Top BLAST hits (NCBI BLASTN):**

**Accession:** AC113952

**Description:** Mus musculus chromosome 3, clone RP24-166J20, complete sequence

**% Match:** 93 (336/360)

BLAST hit homologous (opposite orientation) to intronic region within D3Wsu161e gene (last intron-5 exons in total).




**Conservation (USCS genome server):** Mouse chromosome region 3: 127531848-127532083 is conserved between mouse, human and dog.

**16) CLONE 23**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

278 bp

GNNTGACNNACNTCCNGCTTGGTATGNNANATAGTCAGTACTGGNNCTN
TGGNTGCTGCNCTAGGAGTGNTNACNNANANAGNGTTGNTNTNTGGNGN
CCCCTNNTNNGGNAANCCACGACTAANGCCNTCTCTCTGATTNGNTTTGC
AATGTACCATTAGAAATTATTTCNNNCCTNCCCNCNCNNGNAAAANNAN
AACTAATGCCTTCTCTCTGATTTATTTTCGAATGTACCATTAAAAATTAAT
TCCCCCCTTAAAAAAAAAAAAAAAAAAAA

**Chromosome Position (UCSC genome server):** 12 (-): 87314170-87314228

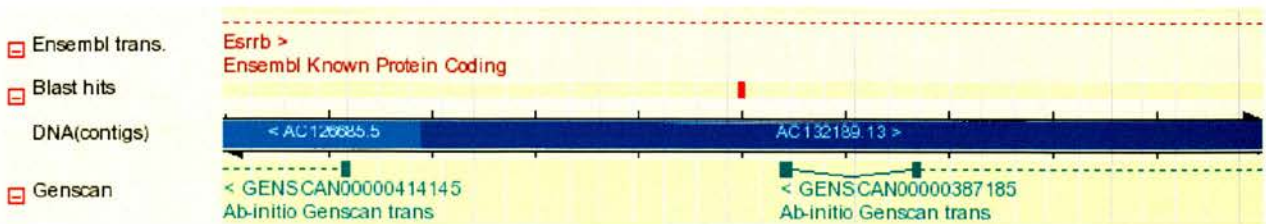
**Top BLAST hits (NCBI BLASTN):**

**Accession:** AC145163

**Description:** Mus musculus BAC clone RP24-238C5 from chromosome 12, complete sequence

**% Match:** Two regions of homology (shown in blue and red)-blue, 84% (45/53); red, 100% (58/58)

BLAST hit is homologous (opposite orientation) to intronic area (1st intron) within Essrb gene (7 exons in total)-also adjacent to ab-initio transcript GENSCAN00000387185.




**Conservation (USCS genome server):** Mouse chromosome region 12: 87314170–87314228 is conserved between mouse, rat, human, dog and opossum.

**17) CLONE 28**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

582 bp

TACATTTTTATACCTGAAATTCCTGGTAGTGTGGGCACTTGTCCTCCTTGC
CGACTTTGTCCTGGAGTTCCGATTTGAATACCTGTGGCCGTTCTGGCTTTT
CATCAGAAGCGTCTATGATTCCTTCAGATACCAAGGACTGGTATGTTTCT
ATTGCATGTTCTCCAACTGTGAGGGATGTAGTTGCCCTTCTAATTAAGAG
AATATCTAAAGTGGCTGATACTGACCTGTTTTAAGTGGAGTTTATTAGGA
ATAAACTAGCTGTTAGCAATGGAAGTGCCTGTTTTGGGATGAGTTAATTG
GGTCTGTAGATTGAGCCCACTGATGAATGGTTCTGTTTTGTTCTTTCTCAC
TTGAGTAGAGAGAAATGATTGGGCAGAGGATTACAAGAAGGAAGTAGCA
ATGGAGGACTCTGTGTTGCTTAGTCTTACCTGGAGTGCTCTGGAACTAAG
TTGTCATTGAATCAAAGCAGTGATTCTCATTGGGTCTGAAGACTCCCTTTT
AAGTAAATTAAAAAAATTCTTGGAGAACTCCAAAAAAGCCTTCTNGTTA
AAATGTAAGGTTNCCAAAATAAACCAAGGG

**Chromosome Position (UCSC genome server):** 4 (-): 134109536-134110066

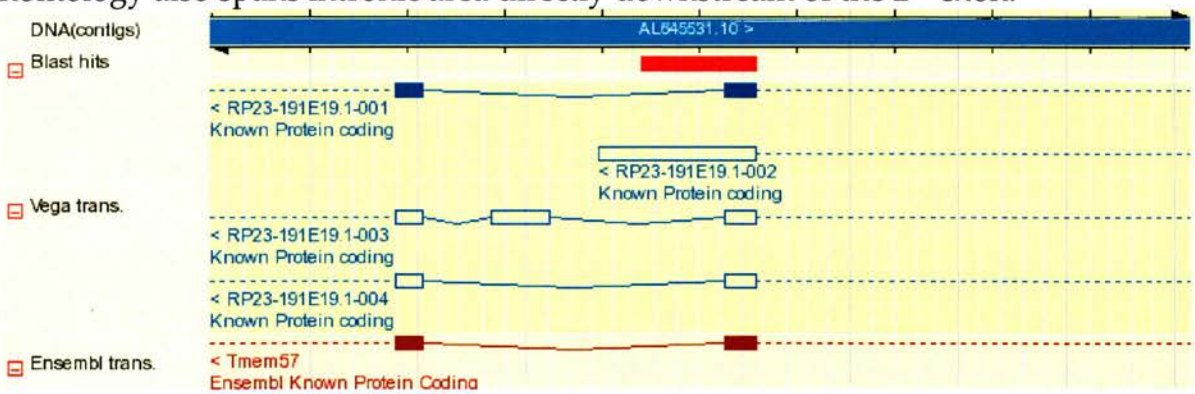
**Top BLAST hits (NCBI BLASTN):**

**Accession:** BC037192

**Description:** Mus musculus transmembrane protein 57, mRNA (cDNA clone IMAGE: 4480906), with apparent retained intron

**% Match:** 99 (507/509)

RACE tag is homologous to second exon of Tmem57 gene (11 exons in total)- homology also spans intronic area directly downstream of the 2nd exon.




**Gene trap orientation relative to gene's transcription:** +

**Gene Information**

**Official Symbol:** Tmem57 **and Name:** transmembrane protein 57 [Mus musculus]

**Other Aliases:** 1110007C24Rik, 9230118A01Rik, AI317300, AI606104, C61
**Other Designations:** macoilin
**Chromosome:** 4; **Location:** 4 65.69 cM
**GeneID:** 66146

**18) CLONE 29**
**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
183 bp
TTCCAAGATGCCTTCCAGGCTCNAACCCAGGACATGTGAGCGGCTGGNC
GGCCTCAACAACTTGNCNAACCAACTCAGGACCTGANAATGGCANAGTA
CATTTGGAANAAGCACTGCTGGTTTGNCNCTCTGTGGTAAAAAAAAANN
AAAAATTTNAACCTTGANCATACNNNNAAAAAAAAA
**Chromosome Position (UCSC genome server):** 3 (-): 32897106-32897254

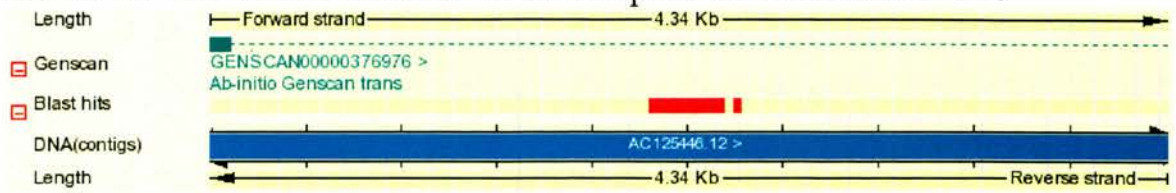
**Top BLAST hits (NCBI BLASTN):**
**Accession:** AC111140
**Description:** Mus musculus chromosome 3, clone RP23-369L16, complete sequence
**% Match:** 92 (125/135)

RACE tag homologous to a repeat sequence (RLTR9E-int; information from USCS RepeatMasker software).

**19) CLONE 34**
**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
446 bp
GCTGGTTTTAAAATTGGATCCCCCCATCTCAGTCTCCGGAGTGTGGATAC
CACAAACATATACCACCACGCCTGGCTGCAACTGGGGAACTTTTGAGTAA
GGAAGACACCGGGAATATTGTTTCAATGAATTGTCTGCATGAAATTCCTA
GAGATGGTTCAGAACTCTGGCTGAGGCAGTCTCAAGGCATTGAGTCTCTT
CAAGCGGAAGAGAAGATGAACTCAAATGTAAGGGTACAAAGAGACAAA
GTAGAGATATATGGTGTTGGTGTTTCTCCACATTAGTAAAGCATTAGCTG
ATAAAAATTTAGACAGAGCATGAGTTTGTTGATAACTTCCTCTACAGANT
GNNTTNCNGGTNNACCANGGNTATCCGAAANANCCCTGGCNCGAATATN
CGTGCATAAAACCATACTGTCAGTCNCTAAAAAAAAAAAAAAAAAAAAA
**Chromosome Position (UCSC genome server):** 14 (+): 99191588-99192276

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AC125446
**Description:** Mus musculus chromosome 14, clone RP24-501J23, complete sequence
**% Match:** 100 (345/345)

BLAST hit within intron of ab-initio transcript GENSCAN00000376976




**Conservation (USCS genome server):** Mouse chromosome region 14: 99191588– 99192276 is conserved between mouse and rat. Also partially conserved between mouse, rat and dog.

**20) CLONE 44**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

191 bp

GCAAAAGAGNCACCTTCCCNCGTTNCTTCANANNANGNTGTTGAATATTT
GCTGAACTTTCATGTCATTGAACACCTGTGAAGACNGNCGGAGCAGACG
CCATGGTTCCGGAAATGAAAGGGAGACAGAAAGAGCCACTGANGACCTG
AGCCCCTGCAGCTGCTGGTNTCTGGNTGAACCTGATGNCANAA

**Chromosome Position (UCSC genome server):** 6 (-): 122271468-122271613

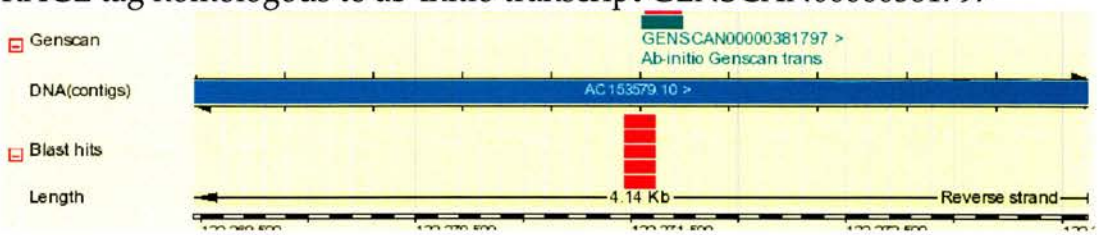
**Top BLAST hits (NCBI BLASTN):**

**Accession:** AC153579

**Description:** Mus musculus 6 BAC RP23-21F2 (Roswell Park Cancer Institute (C57BL/6J Female) Mouse BAC Library) complete sequence

**% Match:** 92 (135/146)

RACE tag homologous to ab-initio transcript GENSCAN00000381797




**Conservation (USCS genome server):** Mouse chromosome region 6: 122271468–122271529 is partially conserved between mouse, rat, human and dog. Chromosome region 6: 122271530- 122271578 is well conserved between mouse, rat, human, dog, opossum, chicken, Xenopus tropicalis and Tetraodon.

**21) CLONE 46**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

564 bp

GTAACTTAGAAGATGATAATTAATGTGGTTGCTGATAATTCTGAATAAAT
ACAGCTTTTATCCCAGGTGTGCCATTTTGAAGACTGAGACCATAGAGTTC
TAAGAATAAAGGA*AAGAGCCCTTGGG*AAATTATTATATATAGCAAAAATG
TGAATCCTCAGATGGAATGAAAGGCCTGCACCATAGACATCGAAGCATT
TACACCCCGCTTGAAGAGTTTGAAATGGACTTTACCACTGAGAAATCAAG
ATGGCAGCCCATTATGGGGAATTGAGGAAAATGGATTAATGCAAGAATG
CTGTAATATTATACAACCAACACAGGATTCTTTTAATGTGGATTCCATGA
AATGAATGATTCTTACCCAACACAAATGGACAGTGGAATTTACTTCCTAA
AGACTTGTTACATGTCATGTACATTTTTGACATCTGGAGAAGACTCTACA
ATTCTACAAATGGTAGTTTGTATTCCTGGAATTTCTTGCAGTTTGATCTGA
AGTGACCTTATGGAATGTTAACTTTAATAAAATCTCTAAAACTTAAAAAA
AAAAAAAAAAAAAAA

**Chromosome Position (UCSC genome server):** 7 (-): 73417549-73429433

**Top BLAST hits (NCBI BLASTN):**

**Accession:** AK140938

**Description:** Mus musculus 16 days embryo head cDNA, RIKEN full-length enriched library, clone: C130076G01 product: unclassifiable, full insert sequence

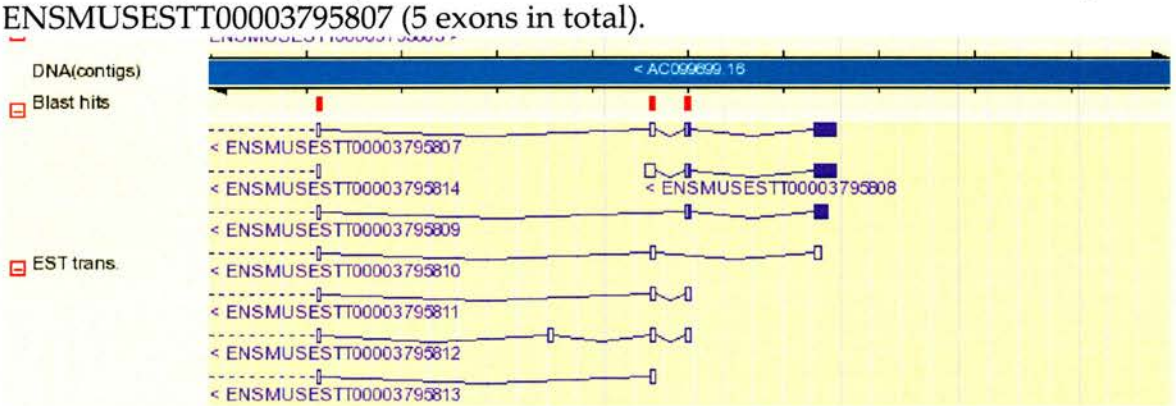
**% Match:** 100 (543/543)

**Accession:** AK007604

**Description:** Mus musculus 10 day old male pancreas cDNA, RIKEN full-length enriched library, clone: 1810026B05 product: unclassifiable, full insert sequence

**% Match:** 100 (430/430)

RACE tag homologous to exons 2-4 of EST transcript ENSMUSESTT00003795807 (5 exons in total).




**Gene trap orientation relative to gene's transcription:** +

**22) CLONE 2D4**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

538 bp

NNNNNNNNNNNNNNNNNNNNGNNNNNGNNCCTGAGCTNANNAAGCC
TGGAGAGACAGTCNNNNNNNNNGNNNGGNTTCTGGGTATACCTTCNCAA
ACTATGGAANNNNNTGGGTGAAGCAGGCTCCAGGAAAGGGTTTAAAGTG
GATGGGCTGGATAAACACCTACACTGGAGAGCCAACATATGCTGATGAC
TTCAAGGGACGGTTTGCCTTCTCTTTGGAAACCTCTGCCAGCACTGCCTAT
TTGNAGATCANCAACCTCAAAAATGAGGACACGGCTACATATTTCTGTGC
AAGACACAGTGTGAAAACCACANCCTGAGGGTGTCAAAAACCATGAGGA
GAANGTGGTTCAGCTGTGTCCNNNANCAACCAGAGGAAACANTCTCTCC
TTGATGTTCGGCTTNNTTNTGAGATTACTGACAACACATATAATAGTCAT
NNANGGTCAGCCACANAATGTTCTCAGTGGNANTGTGNCAGANCANATT
NANGAAAGGAACTNGGNCCATTTNANANCATCTTTAATATGGTAN

**Chromosome Position (UCSC genome server):** 12 (-): 114556604-114588772

**Top BLAST hits (NCBI BLASTN):**

**Accession:** AJ851868

**Description:** Mus musculus immunoglobulin heavy chain locus constant region and partial variable region, strain 129/Sv

**% Match:** 91 (424/461)

**Accession:** AY169677

**Description:** Mus musculus clone VG2J2.8 immunoglobulin heavy chain variable region precursor, gene, partial cds

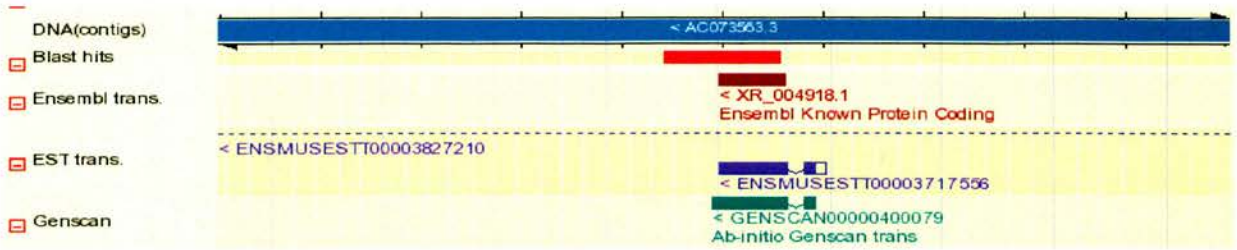
**% Match:** 96 (269/279)

**Accession:** XM_903470

**Description:** PREDICTED: Mus musculus similar to Ig heavy chain V-I region HG3 precursor (LOC630342), mRNA

**% Match:** 96 (238/247)

RACE tag is homologous to exon of EST transcript ENSMUSESTT00003717556-also homologous to similar to Ig heavy chain V-I region V35 precursor (LOC676419), mRNA.




**Gene trap orientation relative to gene's transcription:** +

**23) CLONE 2A5**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

170 bp

GNNNNNNNNNNNNNNNNNNNNANNTNNNNNNNGNCAAGTCTTGCTAC
AATAGAGAGATGGAGAGCANNNANANNATGNNNGNNTNTNNTNTNNNN
NCCNCNGGNNNGAANCNNNCNTGCTCNTGNTGNNNCTTTCNGNNTNTNC
NNNNTTCCTGNACCCCTGNTGGNNN

**Chromosome Position (UCSC genome server):** 6 (-): 122763428-122763463

**Top BLAST hits (NCBI BLASTN):**

**Accession:** AC163108

**Description:** Mus musculus BAC clone RP23-46L17 from chromosome 6, complete sequence

**% Match:** 100 (31/31)-sequence in red

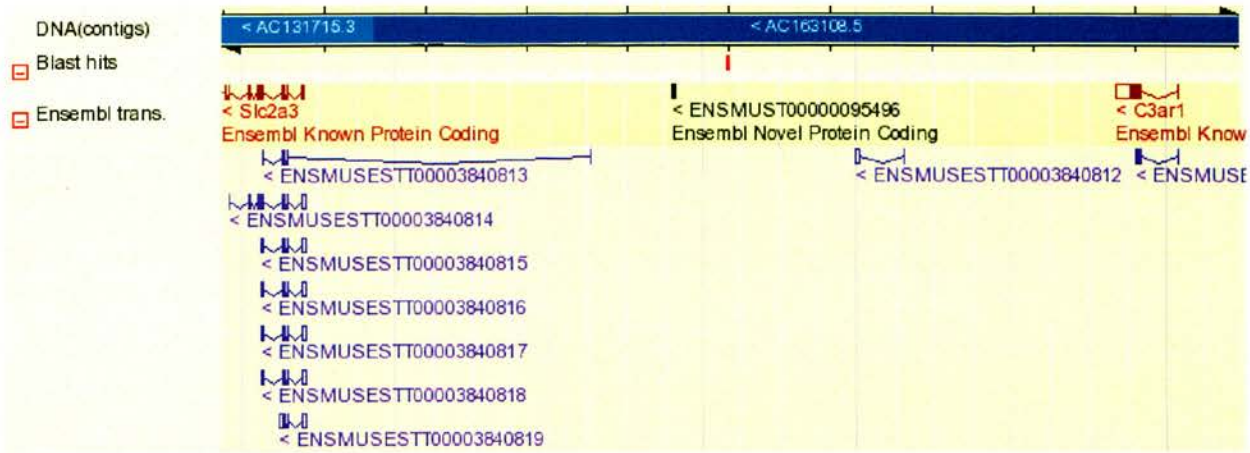
**Top BLAST hits (NCBI BLASTN-mouse EST):**

**Accession:** BF019192

**Description:** uy02c04.y1 McCarrey Eddy spermatocytes Mus musculus cDNA clone IMAGE: 3656838 5' similar to SW:GTR3_MOUSE P32037 GLUCOSE TRANSPORTER TYPE 3, BRAIN. ;mRNA sequence.

**% Match:** 100 (31/31)-sequence in red

BLAST hit is linked to a Slc2a3 gene-homologous to a region upstream of the gene (10 exons in total).




**Gene trap orientation relative to gene's transcription:** +

**Conservation (USCS genome server):** Mouse chromosome region 6: 122763428 – 122763463 is conserved between mouse, rat, human and dog.

**24) CLONE 2C1**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

477 bp

GAACCNGTCTTATGATCCCCACTTGCCNCANATNAANTTGATTGTNGANG
AGANNAGANNNCTCATGCAGCANTAATTNNTATCCACAACTNCTGATTN
NGANTGNGAGGCCATCTCTNGNNNANNTNAGANNCAANTTTCGTTTGNN
ATTCCNNCNNTGNCATGCNANCTTCNCNATTCTTTTATTGAAAAANANAN
AANAANANNATTTNCGNGGGGGTCTNTAAAANANGAAAAAAAAAANAA
ATAAAAACTTGNNTGTGNTGTCCATATCATNGATNNANNGAANCNCTTCT
ATTTCAANATATTTACCTCTGGATAGTTAANAAAANTCTTANACTTCTNA
GTTTTCCNTTCACAAAANTTTATAAAATCNTTATTCATGGGNCTGATAAA
ACTTGACTGANTATTTCACTAANATATNATAATTNTGGNTTNNGGACTTC
NCNNANAANATATTCCAACATGNCTATNTGN

**Chromosome Position (UCSC genome server):** 13 (+): 29240648-29240856

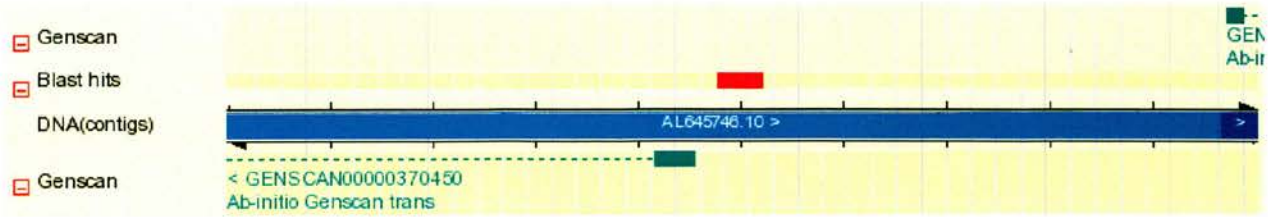
**Top BLAST hits (NCBI BLASTN):**

**Accession:** AL645746

**Description:** Mouse DNA sequence from clone RP23-153B6 on chromosome 13, complete sequence

**% Match:** 88 (121/136)-sequence in red

BLAST hit adjacent to ab-initio Genscan transcript GENSCAN00000370450




**Conservation (USCS genome server):** Mouse chromosome region 13: 29240648 – 29240757 is conserved between mouse, rat, human and dog. Mouse chromosome region 13: 29240758 – 29240856 is conserved between mouse, rat, human, dog and Tetraodon.

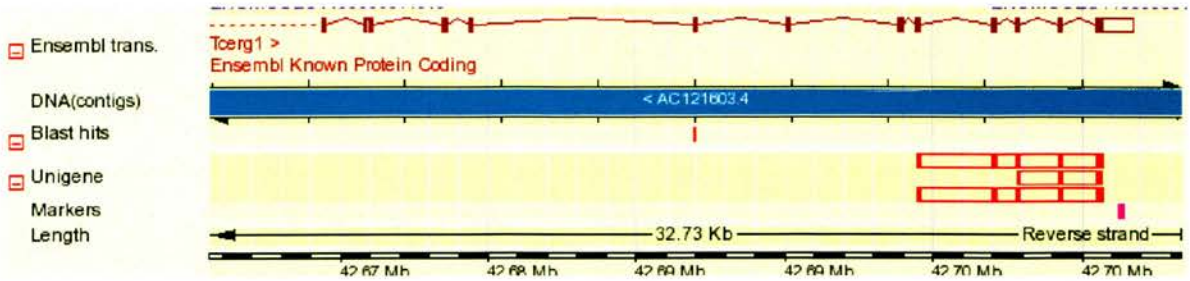
**25) CLONE 1C6**
**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
106 bp
GTGTTTGATCAGGTATGTCAAGACCAGAGCAGAGGAAGAACGCAGGGAA
AAGAAAAACAAAATAATGCAAGCCAAGGAAGATTTCAAAAAAAAAAAA
AAAAAAAAA
**Chromosome Position (UCSC genome server):** 18 (+): 42686907 - 42686997

**Top BLAST hits (NCBI BLASTN):**
**Accession:** BC040284
**Description:** Mus musculus transcription elongation regulator 1 (CA150), mRNA (cDNA clone MGC: 36862 IMAGE: 4460736), complete cds
**% Match:** 98 (84/85)

**Accession:** BC039185
**Description:** Mus musculus transcription elongation regulator 1 (CA150), mRNA (cDNA clone MGC: 28934 IMAGE: 3982736), complete cds
**% Match:** 98 (84/85)

**Accession:** AK150312
**Description:** Mus musculus bone marrow macrophage cDNA, RIKEN full-length enriched library, clone: I830002K14 product: transcription elongation regulator 1 (CA150), full insert sequence
**% Match:** 98 (84/85)

RACE tag is homologous to exon 15 of Tcerg1 gene (22 exons in total).




**Gene trap orientation relative to gene's transcription:** +

**Gene Information**

**Official Symbol:** Tcerg1 **and Name:** transcription elongation regulator 1 (CA150) [Mus musculus]

**Other Aliases:** 2410022J09Rik, 2900090C16Rik, AI428505, CA150b, FBP 23, Taf2s, ca150, p144

**Other Designations:** TATA box binding protein (TBP)-associated factor, RNA polymerase II, S, 150kD; coactivator of 150 kD; transcription factor CA150b

**Chromosome:** 18; **Location:** 18 B3

**GeneID:** 56070

**26) CLONE 3C4**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

869 bp

GAATCAGAGTCCGAAGCAAAGCTAGATGGAGAGACAGCGTCTGACAGCG
AGAGCCGAGCGGAAACAGCTCCCCTACCAACCTCCGTAGATGACACCCC
AGAGGTCCTTAACAGGGCCCTTTCCAACCTGTCCTCAAGGTGGAAGAACT
GGTGGGTGAGAGGCATCCTGACTTTGGCCATGATCGCGTTTTTCTTCATT
ATCATTTACCTGGGACCAATGGTTTTGATGATGATTGTTATGTGTGTCCAG
ATTAAGTGTTTCCATGAAATAATCACTATTGGCTACAATGTATACCACTC
CTACGACCTGCCCTGGTTCAGGACCCTCAGCTGGTAAGCTCTCTGCTCCA
CTGGGGCAGGCAGGCACCTTGCCTGATTTGAGTGATGTGATTTGTGGCAG
GTACTTACAGTGCTGGTGGAGACATTGCCACAGGCTTGCAGGAGATTGA
GCCTGGATTGTATGGCACGTTTGAGGAGACTTGAGTAGCTGAGGACTTCG
TGTCGTCTGTGACAGATGTGTCTATGGTTGGGCTTAGCTCANCAGAAGGG
TCCCTGGGCTCCTGCAGCCAGGGACTGTATTGTGCATCGCACCCGAGGTC
AGCGTGCTTTTTATGCTGCAGCCCAGCGTGTCACTGGAGTCCTCTGGGAG
CTGGCAGGTCCCGGTGCCGGTGTGGCTGTCTCTGGAACTTGATTNGCTCA
TCACTACAGCAGACTGTTGAGGTTACTCTCTTTAANGCACTCGGGGCCGG
CGTGGTTCANATCTGANCTCCTTCANAAACTGCTTTTTTTTCCTTCTCNNA
AAAGGATTGTCTANCACAANCTGCTAATAAANACTGTNNCCAAAAAAAA
AAAAAAAAAANNNCTTTTCCT

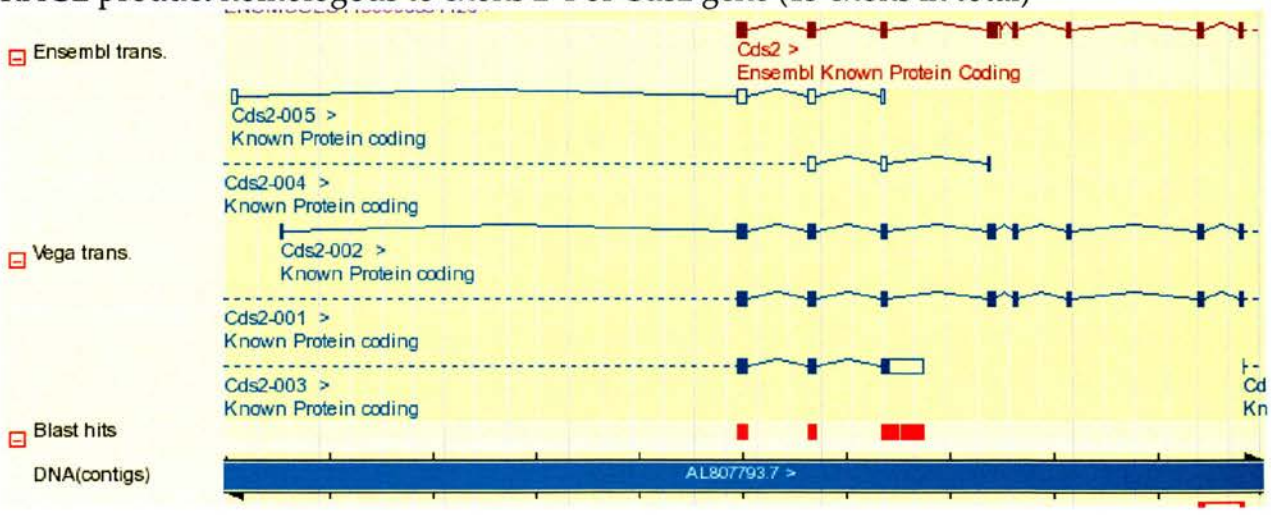
**Chromosome Position (UCSC genome server):** 2 (+): 131984686-131987634

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AK170888
**Description:** Mus musculus NOD-derived CD11c +ve dendritic cells cDNA, RIKEN full-length enriched library, clone: F630207J03 product: CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 2, full insert sequence
**% Match:** 99 (774/781)

**Accession:** AK050589
**Description:** Mus musculus 2 days neonate thymus thymic cells cDNA, RIKEN full-length enriched library, clone: C920007D24 product: SIMILAR TO CDP-DIACYLGLYCEROL SYNTHASE (PHOSPHATIDATE CYTIDYLYLTRANSFERASE) 2 homolog [Mus musculus], full insert sequence
**% Match:** 99 (519/520)

**Accession:** BC059776
**Description:** Mus musculus CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 2, mRNA (cDNA clone IMAGE: 6484748), complete cds
**% Match:** 99 (519/520)

RACE product homologous to exons 2-4 of Cds2 gene (13 exons in total)




**Gene trap orientation relative to gene's transcription:** +

**Gene Information**
**Official Symbol:** Cds2 **and Name:** CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 2 [Mus musculus]
**Other Aliases:** 5730450N06Rik, 5730460C18Rik, AI854580, D2Wsu127e

**Other Designations:** phosphatidate cytidylyltransferase 2
**Chromosome:** 2; **Location:** 2 73.0 cM
**GeneID:** 110911

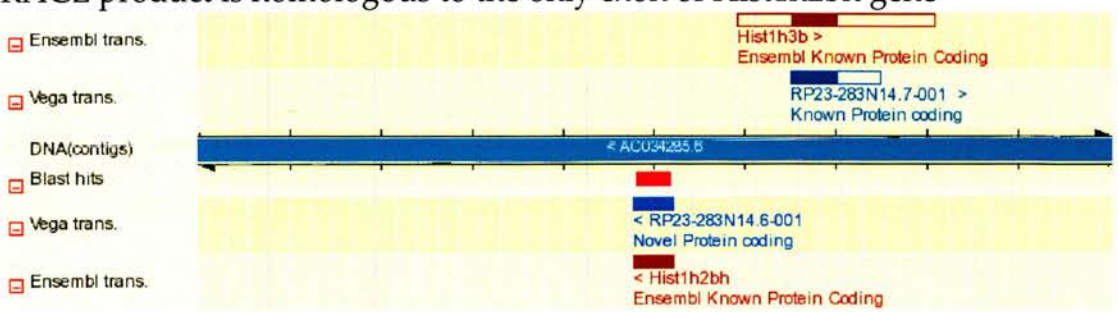
### 27) CLONE 2C4

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
639 bp
GGAANAATTTCACCGTTGCNNTGCCAGTAAAAGGCANACACNNGCTATT
GGTGGTCGTNCGGGAACANTCCCGGAATTTCCGNTAGATGTTGNNGCAG
ACNNGTCNGGAGACACNGNAAACNGCACAACCACNGCNCGGTTGTCTGA
NGTGACATCAGGCCCGTCNCCACGCGTCTAATTCCCNAACTGTNCANCNN
CCTNGCATCTCGTGGCTGTTGTCCTTGCAACCTGTNNCCATATNNACAAA
AAANTTTATACTTCCTACTTCAATAAATGCNCGCTGCCCGAATATAAAAA
AAAAAAAAAAAAAAAAAANCCNNNTNNNGNNTTTNNNNGGGGANGGNT
NCAAGAANGNCCTGACCAAGGCCCAAAAGAANGANGGNAAGAAGGGCA
AGCGCNGCNGCANGGATAGNTACTCTGCGTACNTGTNCAAGGTGTGNAA
GNAAGTGCNCCCCGACACCGTTNATCTCCTCCAAAGCCATGNGNATCNTG
AACNNGNTNGTGAANNACATCTTNNAGCGCNTCNNGNGCGAGGTGTCNC
GCCTGGNNNATTACAACTAGCGCTCGANCATNANGTNCCGGGANATNCA
AANNGNCNTNCGCCTGCTGATGNCGNGGAGCTGGCAACNCGCCNNGTCG
**Chromosome Position (UCSC genome server):** 13 (-): 23550495-23550780

**Top BLAST hits (NCBI BLASTN):**
**Accession:** BC092138
**Description:** Mus musculus histone 1, H2bh, mRNA (cDNA clone MGC: 106612 IMAGE:30613720), complete cds
**% Match:** 77 (147/189)

**Accession:** NM_178197
**Description:** Mus musculus histone 1, H2bh (Hist1h2bh), mRNA
**% Match:** 77 (147/189)

RACE product is homologous to the only exon of Hist1h2bh gene




**Gene trap orientation relative to gene's transcription:** +

**Gene Information**
**Official Symbol:** Hist1h2bh **and Name:** histone 1, H2bh [Mus musculus]
**Other Aliases:** RP23-283N14.6, H2b-221, MGC106612
**Other Designations:** OTTMUSP00000000538
**Chromosome:** 13
**GeneID:** 319182

**28) CLONE 2C6**
**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
368 bp
CTGCCAAAAGAAGAGAGTTTACCCTGGCTGCCTTCAGGATACAGTCCCCA
TCTGCTGCCTGTGGATCAAGATAGAAGACTTCTGACTCCTCCAGCACCAT
GCCTGCTTGTATGCTGCCGTGCCTCCCACCATGATGATAATGCACTGAAC
CTCTGAAGCTGTAAGCCAGCCCCAATTAAATGTTGTCTTTTATAAGANNT
NNCTTGNANCATGGTGCCTCTTCACAGCAATGGAACCTAAAGCATGTACA
CATCCTCCTGTATCCATGCGAGCCAATGAGATGCTGTGTCTCCCTCACATT
CGCCTATGGATCTTGTCACTTTAGTAAATGTTAATGGCCTGAAAAAAAAA
AAAAAAAAAAAAAANNN
**Chromosome Position (UCSC genome server):** 7 (+): 43147752-43148093

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AC165418
**Description:** Mus musculus chromosome 7, clone RP23-148M16, complete sequence
**% Match:** 97 (335/342)

**Top BLAST hits (NCBI BLASTN-mouse EST):**
**Accession:** BY432161
**Description:** BY432161 RIKEN full-length enriched, pooled tissues, 16 days embryo, etc. Mus musculus cDNA clone I920118K02 3', mRNA sequence.
**% Match:** 97 (303/311)

BLAST hit is not related to any known transcript

**Conservation (USCS genome server):** Mouse chromosome region 7: 43147997–43148093 is conserved between mouse and human.

**29) CLONE 2A4**

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)

**Sequence:**

429 bp

AGAAAGCAAGCACCCCGCCTTGCTATGTAGTTCTGTATTGATCAGACTAC
ATCTCCAGCCTTGAGCTCCATTTTAGACACGCAGATTCCTGACGGACATA
TATGAGTGGAAGCGCACACTGGAGAATGACCCTTTTGAGGAGGGGAACA
GTTAGAGGAAGTGCCTGTGACTGGCCTGAACTGGAGACAAAGATGAAAA
CTTTGAGATACTTTCTAACACCTGGAAGAAAGATGACTCANCATGCAAAG
GTGCCTGCTACTGAACCTGATAACCTGAGAGTTTGATCTCTCAGGTTCTA
CAAGGCACAAGGACAGAACTAACTCACATGTAGTCCTCAGACCGCTGTA
TAGTACCTACCGTCTGTCACGCCCCAATTCTAAATAAATAACGAAACCTA
ANNAAAAAAAAAAAAAAAAAAACNCNNTTTTT

**Chromosome Position (UCSC genome server):** 10 (-): 69302407-69302804

**Top BLAST hits (NCBI BLASTN):**

**Accession:** AB057357

**Description:** Mus musculus gene for Ankyrin 3, partial cds

**% Match:** 99 (395/396)

**Accession:** AK013487

**Description:** Mus musculus adult male hippocampus cDNA, RIKEN full-length enriched library, clone: 2900006A17 product:unclassifiable, full

**% Match:** 99 (261/263)

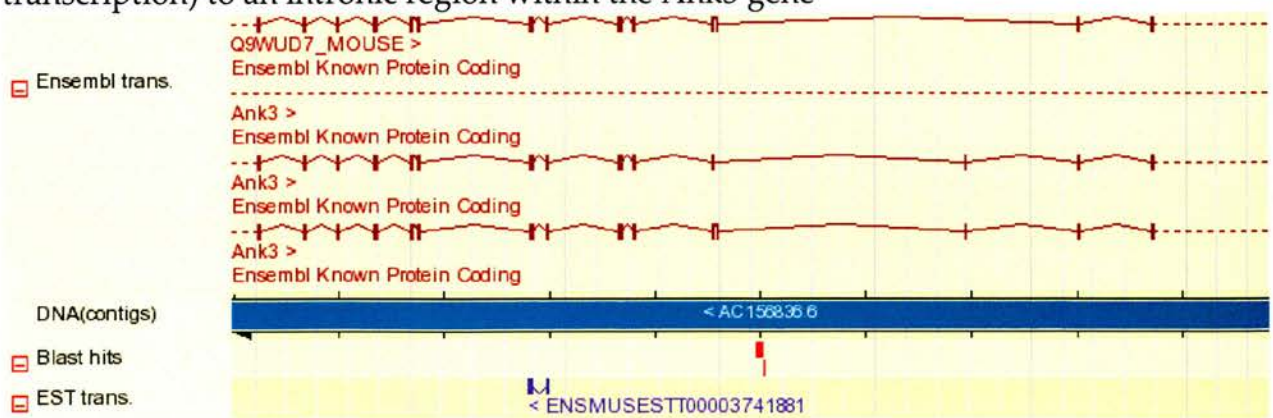
**Top BLAST hits (NCBI BLASTN-mouse EST):**

**Accession:** DV656129

**Description:** DS138B_H02 Eppig/Hampl oocyte Mus musculus cDNA clone DS138B_H02 5', mRNA sequence.

**% Match:** 98 (324/328)

RACE product is homologous (in the opposite orientation to the gene's transcription) to an intronic region within the Ank3 gene



**Conservation (USCS genome server):** Mouse chromosome region 10: 69302577– 69302659 is conserved between mouse, rat and human. Chromosome region 10: 69302660 – 69302804 is conserved between mouse, rat, human and dog.

### 30) CLONE 1A4

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
309 bp
GNTGNCNTGNTCANGGTNATCTCTTCACNGCNGATAACAACATTGGGGT
GAGGACAGATACTGCTNGCAATGGAGCCAGCCAACTGTAGACTGAAGCC
TTTGGGTCTAGGAGCCATGTCAAGTCTTTCCTTATCTCACGTATTTGGTCA
CAGCAGCNAAATGCTAAGGAGACATCCCCTCTTGATCCTTACTAATATTT
TAGTTTTCACCACAATAAAACTATAGACAAATCAAAANNAAAAAAAAAA
ANNNNCCTCTTGACCCTNCCTANNANNGNAGTTNTCACCACAATAAAAC
TATTGACAAATC
**Chromosome Position (UCSC genome server):** 6 (-): 133985138-133985316

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AC145116
**Description:** Mus musculus BAC clone RP23-434H14 from 6, complete sequence
**% Match:** 98 (163/166)

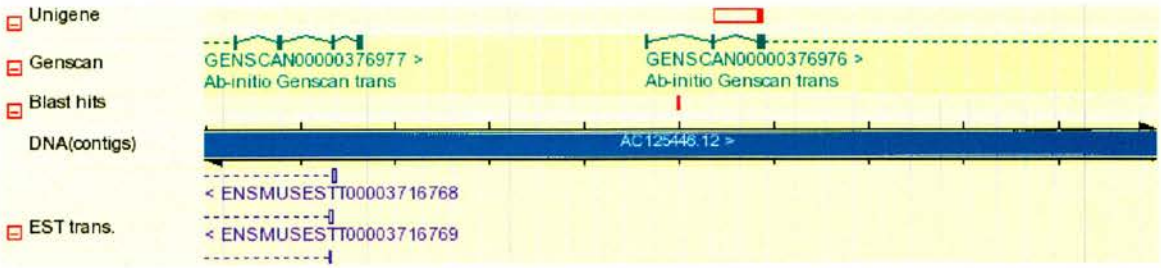
Trapped sequence corresponds to a repeat sequence (LINE class; information from USCS RepeatMasker software).

### 31) CLONE 1D6

**Vector:** pEHygro2neoSD2 (+ARE) (HindIII/MfeI)
**Sequence:**
298 bp
NCATCTCNNGACCCCGCGAAGNGTGGATACCAGNANGANGGTNNACCAC
CCCCCTTNNGNGTGCNACTGGGGAACTGTGTTGATTANGGAAGTACACCC
TTAAATAGGTGGCTNNAATGNNTTGTCTGCANGAAATTCCATAGAGATG
GNTNATAACTNTGGCTGACGNAGTNTCGGGGCATTGAGTNNCTNCNAGN
GGAAGAGAAGATGAACTCGGCTGNAANGGTACANAAANACNAAGNAGA
GNTATATGGTGTNGGTGTTTCTGNNCCTCAGCAAAGCNTNNGCTGATNAA
ANT
**Chromosome Position (UCSC genome server):** 14 (+): 99191627-99191862

**Top BLAST hits (NCBI BLASTN):**
**Accession:** AC125446
**Description:** Mus musculus chromosome 14, clone RP24-501J23, complete sequence
**% Match:** 80 (118/146)

RACE tag homologous to intronic region of ab-initio GENSCAN transcript GENSCAN00000376976




**Conservation (USCS genome server):** Mouse chromosome region 14: 99191627 -99191676 is conserved between mouse and rat. Chromosome region 14: 99191677– 99191862 is conserved between mouse, rat, and dog.

# LIST OF FIGURES

## CHAPTER 1



## CHAPTER 2

## CHAPTER 3

### CHAPTER 4



## LIST OF TABLES

### CHAPTER 1



### CHAPTER 3

# REFERENCES


Allmang, C., Petfalski, E., Podtelejnikov, A., Mann, M., Tollervey, D., Mitchell, P. (1999). The yeast exosome and human PM-Scl are related complexes of 3′→5′ exonucleases. Genes and Development *13*, 2148–2158.

Amsterdam, A., Nissen, R.M., Sun, Z., Swindell, E.C., Farrington, S., and Hopkins, N. (2004). Identification of 315 genes essential for early zebrafish development. Proceedings of the National Academy of Sciences of the United States of America *101*, 12792-12797.

Araki, K., Imaizumi, T., Sekimoto, T., Yoshinobu, K., Yoshimuta, J., Akizuki, M., Miura, K., Araki, M., and Yamamura, K. (1999). Exchangeable gene trap using the Cre/mutated lox system. Cellular and Molecular Biology (Noisy-le-Grand) *45*, 737-750.

Arney, K.L., and Fisher, A.G. (2004). Epigenetic aspects of differentiation. Journal of Cell Science *117*, 4355-4363.

Arning, S., Gruter, P., Bilbe, G., and Kramer, A. (1996). Mammalian splicing factor SF1 is encoded by variant cDNAs and binds to RNA. RNA *2*, 794–810.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Smith, J.A., Struhl, K., Albright, L.M., Coen, D.M., Varki, A., and Janssen, K. (1994). Current protocols in molecular biology Vol 2 (8th Edn). New York, John Wiley and Sons Inc.

Avilion, A.A., Nicolis, S.K., Pevny, L.H., Perez, L., Vivian, N., and Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on SOX2 function. Genes and Development *17*, 126–140.

Bachiller, D., Klingensmith, J., Kemp, C., Belo, J.A., Anderson, R.M., May, S.R., McMahon, J.A., McMahon, A.P., Harland, R.M., Rossant, J., and De Robertis, E.M. (2000). The organizer factors Chordin and Noggin are required for mouse forebrain development. Nature *403*, 658-661.

Baghdoyan, S., Dubreuil, P., Eberle, F., and Gomez, S. (2000). Capture of cytokine-responsive genes (NACA and RBM3) using a gene trap approach. Blood *95*, 3750-3757.

Baker, R. K., Haendel, M.A., Swanson, B.J., Shambaugh, J.C., Micales, B.K., and Lyons, G.E. (1997). In vitro preselection of gene-trapped embryonic stem cell clones for characterising novel developmentally regulated genes in the mouse. Developmental Biology *185*, 201-214.

Barreau, C., Paillard, L., and Osborne, H.B. (2005). AU-rich elements and associated factors: are they unifying principles? Nucleic Acids Research *33*, 7138-7150.

Baust, C., Baillie, G.J., and Mager, D.L. (2002). Insertional polymorphisms of ETn retrotransposons include a disruption of the wiz gene in C57BL/6 mice. Mammalian Genome *13*, 423–428.

Baust, C., Gagnier, L., Baillie, G.J., Harris, M.J., Juriloff, D.M., and Mager, D.L. (2003). Structure and expression of mobile ETnII retroelements and their coding-competent MusD relatives in the mouse. Journal of Virology *77*, 11448–11458.

Beddington, R.S.P. and Robertson, E.J. (1998). Anterior patterning in mouse. Trends in Genetics *14*, 277-284.

Bedford, M.T., Reed, R., and Leder, P. (1998). WW domain-mediated interactions reveal a spliceosome-associated protein that binds a third class of proline-rich motif: the proline glycine and methionine-rich motif. Proceedings of the National Academy of Sciences of the United States of America *95*, 10602–10607.

Berglund, J.A., Chua, K., Abovich, N., Reed, R. and Rosbash, M. (1997). The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell* 89, 781–787.

Berglund, J.A., Fleming, M.L., Rosbash, M. (1998). The KH domain of the branchpoint sequence binding protein determines specificity for the pre-mRNA branchpoint sequence. RNA *4*, 998–1006.

Blackshear, P.J., Graves, J.P., Stumpo, D.J., Cobos, I., Rubenstein, J.L., and Zeldin, D.C. (2003). Graded phenotypic response to partial and complete deficiency of a brain-specific transcript variant of the winged helix transcription factor RFX4. Development *130*, 4539-4552.

Blencowe, B.J. (2000). Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends in Biochemical Sciences *25*, 106–110.

Bonaldo, P., Chowdhury, K., Stoykova, A., Torres, M., and Gruss, P. (1998). Efficient gene trap screening for novel developmental genes using IRES*βgeo* vector and *in vitro* preselection. Experimental Cell Research *244*, 125-136.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., Gifford, D.K., Melton, D.A., Jaenisch, R., and Young, R.A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947–956.

Bradley, A., Evans, M., Kaufman, M.H., and Robertson, E. (1984). Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. Nature *309*, 255-256.

Brewer, G., and Ross, J. (1988). Poly(A) shortening and degradation of the 30 A+U-rich sequences of human c-myc mRNA in a cell-free system. Molecular and Cellular Biology *8*, 1697–1708.

Brewer, G. (1998). Characterization of c-myc 3′ to 5′ mRNA decay of mRNA decay activities in an in vitro system. Journal of Biological Chemistry *273*, 34770–34774.

Bridge, A.J., Pebernard, S., Ducraux, A., Nicoulaz, A.L., and Iggo, R. (2003). Induction of an interferon response by RNAi vectors in mammalian cells. Nature Genetics *34*, 263–264.

Bronchain,O. J., Hartley, K. O., and Amaya, E. (1999). A gene trap approach in Xenopus. Current Biology *9*, 1195-1198.

Brulet, P., Kaghad, M., Xu, Y.S., Croissant, O., and Jacob, F. (1983). Early differential tissue expression of transposon-like repetitive DNA sequences of the mouse. Proceedings of the National Academy of Sciences of the United States of America *80*, 5641–5645.

Brulet, P., Condamine, H., and Jacob, F. (1985). Spatial distribution of transcripts of the long repeated ETn sequence during early mouse embryogenesis. Proceedings of the National Academy of Sciences of the United States of America *82*, 2054–2058.

Bultman, S., Gebuhr, T., Yee, D., La Mantia, C., Nicholson, J., Gilliam, A., Randazzo, F., Metzger, D., Chambon, P., Crabtree, G., and Magnuson, T. (2000). A *Brg1* Null Mutation in the Mouse Reveals Functional Differences among Mammalian SWI/SNF Complexes. Molecular Cell *6*, 1287-1295.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology *268*, 78-94

Camus, A., Kress, C., Babinet, C., and Barra, J. (1996). Unexpected behaviour of a gene trap vector comprising a fusion between the Sh *ble* and the *lacZ* genes. Molecular Reproduction and Development *45*, 255-263.

Cannon, J.P., Colicos, S.M., and Belmont, J.W. (1999). Gene trap screening using negative selection: identification of two tandem, differentially expressed loci with potential hematopoietic function. Developmental Genetics *25*, 49-63.

Cao, S., Bendall, H., Hicks, G.G., Nashabi, A., Sakano, H., Shinkai, Y., Gariglio, M., Oltz, E.M., and Ruley, H.E. (2003). The high-mobility-group box protein SSRP1/T160 is essential for cell viability in day 3.5 mouse embryos. Molecular and Cellular Biology *23*, 5301-5307.

Caplen, N.J., Parrish, S., Imani, F., Fire, A., and Morgan, R.A. (2001). Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. Proceedings of the National Academy of Sciences of the United States of America *98*, 9742–9747.

Carlson, C.M., Dupuy, A.J., Fritz, S., Roberg-Perez, K.J., Fletcher, CF., and Largaespada, D.A. (2003). Transposon mutagenesis of the mouse germline. Genetics *165*, 243-256.

Carlson, C.M. and Largaespada, D.A. (2005). Insertional mutagenesis in mice: new perspectives and tools. Nature Reviews Genetics *6*, 568-580.

Carmeliet, P., Ferreira, V., Breier, G., Pollefeyt, S., Kieckens, L., Gertsenstein, M., Fahrig, M., Vandenhoeck, A., Harpal, K., Eberhardt, C., Declercq, C., Pawling, J., Moons, L., Collen, D., Risau, W., and Nagy, A. (1996). Abnormal blood vessel development and lethality in embryos lacking a single VEGF allele. Nature *380*, 435-439.

Carroll, P., Renoncourt, Y., Gayet, O., De Bovis, B., and Alonso, S. (2001). Sorting nexin-14, a gene expressed in motoneurons trapped by an in vitro preselection method. Developmental Dynamics *221*, 431-442.

Cecconi, F., Alvarez-Bolado, G., Meyer, B.I., Roth, K.A., and Gruss, P. (1998). Apaf1 (CED-4 homolog) regulates programmed cell death in mammalian development. Cell *94*, 727-737.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell *113*, 643–655.

Chambeyron, S., Da Silva, N.R, Lawson, K.A., and Bickmore, W.A. (2005). Nuclear re-organisation of the *Hoxb* complex during mouse embryonic development. Development *132*, 2215-2223.

Chen, C.Y., and Shyu, A.B. (1994). Selective degradation of early-response-gene mRNAs: functional analyses of sequence features of the AU-rich elements. Molecular and Cellular Biology *14*, 8471–8482.

Chen, C.-Y.A., Chen, T.-M. and Shyu, A.B. (1994). Interplay of two functionally and structurally distinct domains of the c-fos AU-rich element specifies its mRNA-destabilizing fuction. Molecular and Cellular Biology *14*, 416–426.

Chen, C.-Y.A., Xu, N., and Shyu, A.B. (1995). mRNA decay mediated by two distinct AU-rich elements from c-fos and granulocyte-macrophage colony-stimulating factor transcripts : different deadenylation kinetics and uncoupling from translation. Molecular and Cellular Biology *15*, 5777–5788.

Chen, C.Y., Gherzi, R., Ong, S.E., Chan, E.L., Raijmakers, R., Pruijn, G.J., Stoecklin, G., Moroni, C., Mann, M., and Karin, M. (2001). AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. Cell *107*, 451– 464.

Chen, H., Bagri, A., Zupicich, J.A., Zou, Y., Stoeckli, E., Pleasure, S.J., Lowenstein, D.H., Skarnes, W.C., Chedotal, A., and Tessier-Lavigne, M. (2000). Neuropilin-2 regulates the development of selective cranial and sensory nerves and hippocampal mossy fiber projections. Neuron *25*, 43-56.

Chen, W.V. and Soriano, P. (2003). Gene trap mutagenesis in embryonic stem cells. Methods in Enzymology *365*, 367-386.

Chen, W.V., and Chen, Z. (2004). Differentiation trapping screen in live culture for genes expressed in cardiovascular lineages. Developmental Dynamics *229*, 319-327.

Chen, W.V., Delrow, J., Corrin, P.D., Frazier, J.P., and Soriano, P. (2004a). Identification and validation of PDGF transcriptional targets by microarray-coupled gene-trap mutagenesis. Nature Genetics *36*, 304-312.

Chen, Y-T., Liu, P., and Bradley, A. (2004b). Inducible gene trapping with drug-selectable markers and Cre/*loxP* to identify developmentally regulated genes. Molecular and Cellular Biology *24*, 9930-9941.

Chowdhury, K, Bonaldo, P., Torres, M., Stoykova, A., and Gruss, P. (1997). Evidence for the stochastic integration of gene trap vectors into the mouse germline. Nucleic Acids Research *25*, 1531-1536.

Ciruna, B. and Rossant, J. (2001). FGF signaling regulates mesoderm cell fate specification and morphogenetic movement at the primitive streak. Developmental Cell *1*, 37-49.

Cobellis, G. Nicolaus, G., Iovino, M., Romito, A., Marra, E., Barbarisi, M., Sardiello, M., Di Giorgio, F.P., Iovino, N., Zollo, M., Ballabio, A., and Cortese, R. (2005). Tagging genes with cassette-exchange sites. Nucleic Acids Research *33*, e44.

Constantini, F. and Lacy, E. (1981). Introduction of a rabbit β-globin gene into the mouse germ line. Nature *294*, 92-94.

Danckwardt, S., Neu-Yilik, G., Thermann, R., Frede, U., Hentze, M.W., and Kulozik, A.E. (2002). Abnormally spliced β-globin mRNAs: a single point mutation generates transcripts sensitive and insensitive to nonsense-mediated mRNA decay. Blood *99*, 1811-1816.

DeGregori, J., Russ, A., von Melchner, H., Rayburn, H., Priyaranjan, P., Jenkins, N.A., Copeland, N.G., and Ruley, H.E. (1994). A murine homolog of the yeast RNA1 gene is required for postimplantation development. Genes and Development *8*, 265-276.

Deng, J.M. and Behringer, R.R. (1995). An insertional mutation in the BTF3 transcription factor gene leads to an early postimplantation lethality in mice. Transgenic Research *4*, 264-269.

De Palma, M., Montini, E., Santoni de Sio, F.R., Benedicenti, F., Gentile, A., Medico, E., and Naldini, L. (2005). Promoter trapping reveals significant differences in integration site selection between MLV and HIV vectors in primary hematopoietic cells. Blood *105*, 2307-2315.

De-Zolt, S., Schnutgen, F., Seisenberger, C., Hansen, J., Hollatz, M., Floss, T., Ruiz, P., Wurst, W., and von Melchner, H. (2006). High-throughput trapping of secretory pathway genes in mouse embryonic stem cells. Nucleic Acids Research *34*, e25.

Dhara, S.K., and Benvenisty, N. (2004). Gene trap as a tool for genome annotation and analysis of X chromosome inactivation in human embryonic stem cells. Nucleic Acids Research *32*, 3995-4002.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005). Efficient transposition of the *piggyBac* (PB) transposon in mammalian cells and mice. Cell *122*, 473-483.

Doetschman, T.C., Eistetter, H., Katz, M., Schmidt, W., and Kemler, R. (1985). The in vitro development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium. Journal of Embryology and Experimental Morphology *87*, 27-45.

Donnelly, M.L., Hughes, L.E., Luke, G., Mendoza, H., ten Dam, E., Gani, D., and Ryan, M.D. (2001). The 'cleavage' activities of foot-and-mouth disease virus 2A site-directed mutants and naturally occurring '2A-like' sequences. The Journal of General Virology *82*, 1027-1041.

Dormer, P., Spitzer, E., Frankerberger, M., and Kremmer, E. (2004). Erythroid differentiation regulator (EDR), a novel, highly conserved factor I. Induction of haemoglobin synthesis in erythroleukaemic cells. Cytokine *26*, 231-242.

Drabkin, H.J., and Rajbhandary, U.L. (1998). Initiation of protein synthesis in mammalian cells with codons other than AUG and amino acids other than methionine. Molecular and Cellular Biology *18*, 5140-5147.

Dunwoodie, S.L., and Beddington, R.S. (2002). The expression of the imprinted gene Ipl is restricted to extra-embryonic tissues and embryonic lateral mesoderm during early mouse development. The International Journal of Developmental Biology *46*, 459-466.

Dupuy, A.J., Fritz, S., and Largaespada, D.A. (2001). Transposition and gene disruption in the male germline of the mouse. Genesis *30*, 82-88.

Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494-498.

Elefanty, A.G., Robb, L., Birner, R., and Begley, C.G. (1997). Hematopoietic-specific genes are not induced during in vitro differentiation of scl-null embryonic stem cells. Blood *90*, 1435-1447.

Endoh, M., Ogawa, M., Orkin, S., and Nishikawa, S. (2002). SCL/tal-1-dependent process determines a competence to select the definitive hematopoietic lineage prior to endothelial differentiation. The EMBO Journal *21*, 6700-6708.

Episkopou, V., Arkell, R., Timmons, P.M., Walsh, J.J., Andrew, R.L., and Swan, D. (2001). Induction of the mammalian node requires Arkadia function in the extraembryonic lineages. Nature *410*, 825-830.

Evans, M. J. and Kaufman, M. H. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature *292*, 154-156.

Eystathioy, T., Jakymiw, A., Chan, E.K., Seraphin, B., Cougot, N., and Fritzler, M.J. (2003). The GW182 protein colocalizes with mRNAdegradation associated proteins hDcp1 and hLSm4 in cytoplasmic GW bodies. RNA *9*, 1171–1173.

Farley, A.H., Luning Prak, E.T., and Kazazian, H.H. Jr. (2004). More active human L1 retrotransposons produce longer insertions. Nucleic Acids Research *32*, 502-510.

Faisst, A.M., and Gruss, P. (1998). Bodenin: a novel murine gene expressed in restricted areas of the brain. Developmental Dynamics *212*, 293-303.

Fan, X.C., and Steitz, J.A. (1998). Overexpression of HuR, a nuclear-cytoplasmic shuttling protein, increases the in vivo stability of ARE-containing mRNAs. The EMBO Journal 17, 3448–3460.

Fischer, S.E., Wienholds, E., and Plasterk, R.H. (2001). Regulated transposition of a fish transposon in the mouse germ line. Proceedings of the National Academy of Sciences of the United States of America *98*, 6759-6764.

Fong, G.H., Zhang, L., Bryce, D.M., and Peng, J. (1999). Increased hemangioblast commitment, not vascular disorganization, is the primary defect in flt-1 knock-out mice. Development *126*, 3015-3025.
Ford, L.P., Watson, J., Keene, J.D., and Wilusz, J. (1999). ELAV proteins stabilize deadenylated intermediates in a novel in vitro mRNA deadenylation/degradation system. Genes and Development *13*, 188–201.

Forrester, L.M., Nagy, A., Sam, M., Watt, A., Stevenson, L., Bernstein, A., Joyner, A.L., and Wurst, W. (1996). An induction gene trap screen in embryonic stem cells: identification of genes that respond to retinoic acid *in vitro*. Proceedings of the National Academy of Sciences of the United States of America *93*, 1677-1682.

Francastel, C., Schubeler, D., Martin, D.I., and Groudine, M. (2000). Nuclear compartmentalization and gene activity. Nature Reviews Molecular Cell Biology *1*, 137-143.

Frank, D., Fortino, W., Clark, L., Musalo, R., Wang, W., Saxena, A., Li, C.M., Reik, W., Ludwig, T., and Tycko, B. (2002). Placental overgrowth in mice lacking the imprinted gene Ipl. Proceedings of the National Academy of Sciences of the United States of America *99*, 7490-7495.

Friedel, R.H., Plump, A., Lu, X., Spilker, K., Jolicoeur, C., Wong, K., Venkatesh, T.R., Yaron, A., Hynes, M., Chen, B., Okada, A., McConnell, S.K., Rayburn, H., and Tessier-Lavigne, M. (2005). Gene targeting using a promoterless gene trap vector ("targeted trapping") is an efficient method to mutate a large fraction of genes. Proceedings of the National Academy of Sciences of the United States of America *102*, 13188-13193.

Friendrich, G., and Soriano, P. (1991). Promoter traps in embryonic stem cells: A genetic screen to identify and mutate developmental genes in mice. Genes and Development *5*, 1513-1523.

Frohman, M.A., Dush, M.K., and Martin, G.R. (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proceedings of the National Academy of Sciences of the United States of America *85*, 8998-9002.

Fujikura, J., Yamato, E., Yonemura, S., Hosoda, K., Masui, S., Nakao, K., Miyazaki, Ji, J., and Niwa, H. (2002). Differentiation of embryonic stem cells is induced by GATA factors. Genes and Development *16*, 784-789.

Futterer, A., Campanero, M.R., Leonardo, E., Criado, L.M., Flores, J.M., Hernandez, J.M., San Miguel, J.F., and Martinez-A, C. (2005). Dido gene expression alterations are implicated in the induction of hematological myeloid neoplasms. Journal of Clinical Investigation *115*, 2351-2362.

Gajovic, S., Chowdhury, K., and Gruss, P. (1998). Genes expressed after retinoic acid-mediated differentiation of embryoid bodies are likely to be expressed during embryo development. Experimental Cell Research *242*, 138-143.

Gao, M., Wilusz, C.J., Peltz, S.W., and Wilusz, J. (2001). A novel mRNA decapping activity in HeLa cytoplasmic extracts is regulated by AU-rich elements. The EMBO Journal *20*, 1134–1143.

Garcia-Domingo, D., Leonardo, E., Grandien, A., Martinez, P., Albar, J.P., Izpisua-Belmonte, J.C., and Martinez-A, C. (1999). DIO-1 is a gene involved in onset of apoptosis in vitro, whose misexpression disrupts limb development. Proceedings of the National Academy of Sciences of the United States of America *96*, 7992-7997.

Garcia-Meunier, P., Etienne-Julan, M., Fort, P., Piechaaczyk, M., and Bonhomme, F. (1993). Concerted evolution in the GAPDH family of retrotransposed pseudogenes. Mammalian Genome 4, 695-703.

Gilbert, N., Lutz-Prigge, S., and Moran, J.V. (2002). Genomic deletions created upon LINE-1 retrotransposition. Cell *110*, 315-325.

Goldstrohm, A.C., Greenleaf, A.L., and Garcia-Blanco, M.A. (2001). Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. Gene *277*, 31-47.

Gonzalez-Billault, C., Demandt, E., Wandosell, F., Torres, M., Bonaldo, P., Stoykova, A., Chowdhury, K., Gruss, P., Avila, J., and Sanchez, M.P. (2000). Perinatal lethality of microtubule-associated protein 1B-deficient mice expressing alternative isoforms of the protein at low levels. Molecular and Cellular Neuroscience *16*, 408-421.

Gordon, J. W. and Ruddle, F. H. (1981). Integration and stable germ line transmission of genes injected into mouse pronuclei. Science *214*, 1244-1246.

Gossler, A., Joyner, A.L., Rossant, J., and Skarnes, W.C. (1989). Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. Science *244*, 463-465.

Grandvaux, N., tenOever, B.R., Servant, M.J., and Hiscott, J. (2002). The interferon antiviral response: from viral invasion to evasion. Current Opinion in Infectious Diseases *15*, 259–267.

Gridley, T.P., Soriano, P., and Jaenisch, R. (1987). Insertional mutagenesis in mice. Trends in Genetics *3*, 162-166.

Guth, S., Martinez, C., Gaur, R.K., and Valcarcel, J. (1999). Evidence for substrate-specific requirement of the splicing factor U2AF(35) and for its function after polypyrimidine tract recognition by U2AF(65). Molecular and Cellular Biology *19*, 8263–8271.

Haerry, T.E., and Gehring, W.J. (1997). A conserved cluster of homeodomain binding sites in the mouse Hoxa-4 intron functions in Drosophila embryos as an enhancer that is directly regulated by Ultrabithorax. Developmental Biology *186*, 1-15.

Hall, G.W., and Thein, S. (1994). Nonsense codon mutations in the terminal exon of the β-globin gene are not associated with a reduction in β-mRNA accumulation: a mechanism for the phenotype of dominant β-thalassemia. Blood *83*, 2031-2037.

Hamatani, T., Carter, M.G., Sharov, A.A., and Ko, M.S. (2004). Dynamics of global gene expression changes during mouse preimplanation development. Developmental Cell *6*, 117-131.

Hansen, J., Floss, T., Van Sloun, P., Fuchtbauer, E.M., Vauti, F., Arnold, H.H., Schnutgen, F., Wurst, W., von Melchner, H., and Ruiz, P. (2003). A large-

scale, gene-driven mutagenesis approach for the functional analysis of the mouse genome. Proceedings of the National Academy of Sciences of the United States of America *100*, 9918-9922.

Harbers, K., Jahner, D., and Jaenisch, R. (1981). Microinjection of cloned retroviral genomes into mouse zygotes: integration and expression in the animal. Nature *293*, 540-542.

Harbers, K., Kuehn, M., Delius, H., and Jaenisch, R. (1984). Insertion of retrovirus into the first intron of α1(I) collagen gene leads to embryonic lethal mutation in mice. Proceedings of the National Academy of Sciences of the United States of America *81*, 1504-1508.

Hardouin, N. and Nagy, A. (2000). Gene-trap-based target site for Cre-mediated transgenic insertion. Genesis *26*, 245-252.

Hart, A.H., Hartley, L., Sourris, K., Stadler, E.S., Li, R., Stanley, E.G., Tam, P.P., Elefanty, A.G., and Robb, L. (2002). Mixl1 is required for axial mesendoderm morphogenesis and patterning in the murine embryo. Development *129*, 3597-3608.

Hicks, G.G., Shi, E.G., Li, X.M., Li, C.H., Pawlak, M., and Ruley, H.E. (1997). Functional genomics in mice by tagged sequence mutagenesis. Nature Genetics *16*, 338-344.

Hidaka, M., Caruana, G., Stanford, W.L., Sam, M., Correll, P.H., and Bernstein, A. (2000). Gene trapping of two novel genes, *Hzf* and *Hhl*, expressed in hematopoietic cells. Mechanisms of Development *90*, 3-15.

Hildebrand, J.D. and Soriano, P. (2002). Overlapping and unique roles for C-terminal binding protein 1 (CtBP1) and CtBP2 during mouse development. Molecular and Cellular Biology *22*, 5296-5307.

Hill, D.P., and Wurst, W. (1993). Screening for novel pattern formation genes using gene trap approaches. Methods in Enzymology *225*, 664-681.

Hirashima, M., Bernstein, A., Stanford, W.L., and Rossant, J. (2004). Gene trap screening for endothelial-specific genes. Blood *104*, 711-718.

Hitotsumachi, S., Carpenter, D.A., and Russell, W.L. (1985). Dose-repetition increases the mutagenic effectiveness of N-ethyl-N-nitrosourea in mouse

spermatogonia. Proceedings of the National Academy of Sciences of the United States of America *82*, 6619-6621.

Hooper, M., Hardy, K., Handyside, A., Hunter, S., and Monk, M. (1987). Hprt-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured-cells. Nature *326*, 292-295.

Horie, K., Kuroiwa, A., Ikawa, M., Okabe, M., Kondoh, G., Matsuda, Y., and Takeda, J. (2001). Efficient chromosomal transposition of a *Tc1/mariner*-like transposon *Sleeping Beauty* in mice. Proceedings of the National Academy of Sciences of the United States of America *98*, 9191-9196.

Horie, K., Yusa, K., Yae, K., Odajima, J., Fischer, S.E., Keng, V.W., Hayakawa, T., Mizuno, S., Kondoh, G., Ijiri, T., Matsuda, Y., Plasterk, R.H., and Takeda, J. (2003). Characterization of *Sleeping Beauty* transposition and its application to genetic screening in mice. Molecular and Cellular Biology *23*, 9189-9207.

Hrabe de Angelis, M.H., Flaswinkel, H., Fuchs, H., Rathkolb, B., Soewarto, D., Marschall, S., Heffner, S., Pargent, W., Wuensch, K., Jung, M., Reis, A., Richter, T., Alessandrini, F., Jakob, T., Fuchs, E., Kolb, H., Kremmer, E., Schaeble, K., Rollinski, B., Roscher, A., Peters, C., Meitinger, T., Strom, T., Steckler, T., Holsboer, F., Klopstock, T., Gekeler, F., Schindewolf, C., Jung, T., Avraham, K., Behrendt, H., Ring, J., Zimmer, A., Schughart, K., Pfeffer, K., Wolf, E., and Balling, R. (2000). Genome-wide, large-scale production of mutant mice by ENU mutagenesis. Nature Genetics *25*, 444-447.

Imataka, H., Olsen,H.S., and Sonenberg, N. (1997). A new translational regulator with homology to eukaryotic translation initiation factor 4G. The EMBO Journal *16*, 817-825.

Ingelfinger, D., Arndt-Jovin, D.J., Luhrmann, R. and Achsel, T. (2002). The human LSm1-7 proteins colocalize with the mRNA-degrading enzymes Dcp1/2 and Xrnl in distinct cytoplasmic foci. RNA *8*, 1489–1501.

Ishida, Y., and Leder, P. (1999). RET: a polyA-trap retrovirus vector for reversible disruption and expression monitoring of genes in living cells. Nucleic Acids Research *27*, e35

Ivics, Z., Hackett, P.B., Plasterk, R.H., and Izsvak, Z. (1997). Molecular reconstruction of *Sleeping Beauty*, a *Tc-1*-like transposon from fish, and its transposition in human cells. Cell *91*, 501-510.

Jackson, A.L., Bartz, S.R., Schelter, J., Kobayashi, S.V., Burchard, J., Mao, M., Li, B., Cavet, G., and Linsley, P.S. (2003). Expression profiling reveals off-target gene regulation by RNAi. Nature Biotechnology *21*, 635-637.

Jaenisch, R. (1976). Germ line integration and Mendelian transmission of the exogenous Moloney leukaemia virus. Proceedings of the National Academy of Sciences of the United States of America *73*, 1260-1264.

Jaenisch, R. (1988). Transgenic animals. Science *240*, 1468-1474.

Kanai-Azuma, M., Kanai, Y., Gad, J.M., Tajima, Y., Taya, C., Kurohmaru, M., Sanai, Y., Yonekawa, H., Yazaki, K., Tam, P.P., and Hayashi, Y. (2002). Depletion of definitive gut endoderm in Sox17-null mutant mice. Development *129*, 2367-2379.

Kedersha, N. and Anderson, P. (2002). Stress granules: sites of mRNA triage that regulate mRNA stability and translatability. Biochemical Society Transactions *30*, 963–969.

Kedersha, N., Stoecklin, G., Ayodele, M., Yacono, P., Lykke-Andersen, J., Fitzler, M.J., Scheuner, D., Kaufman, R.J., Golan, D.E., and Anderson, P. (2005). Stress granules and processing bodies are dynamically linked sites of mRNP remodeling. Journal of Cell Biology *169*, 871–884.

Keller, G. (1995). In vitro differentiation of embryonic stem cells. Current Opinion in Cell Biology *7*, 862-869.

Keller. G. (2005). Embryonic stem cell differentiation: emergence of a new era in biology and medicine. Genes and Development *19*, 1129-1155.

Kile, B.T., and Hilton, D.J. (2005). The art and design of genetic screens: mouse. Nature Reviews Genetics *6*, 557-567.

Kim, J.H., Auerbach, J.M., Rodriguez-Gomez, J.A., Velasco, I., Gavin, D., Lumelsky, N., Lee, S.H., Nguyen, J., Sanchez-Pernaute, R., Bankiewicz, K., and McKay, R. (2002). Dopamine neurons derived from embryonic stem cells function in an animal model of Parkinson's disease. Nature *418*, 50-56.

Klochendler-Yeivin, A., Fiette, L., Barra, J., Muchardt, C., Babinet, C., and Yaniv, M. (2000). The murine SNF5/INI1 chromatin remodeling factor is

essential for embryonic development and tumor suppression. EMBO Reports *1*, 500-506.

Kluppel, M., Vallis, K., and Wrana, J. L. (2002). A high-throughput induction gene trap approach defines C4ST as a target of BMP signalling. Mechanisms of Development 118, 77-89.

Kohtz, J.D., Jamison, S.F., Will, C.L., Zuo, P., Luhrmann, R., Garcia-Blanco, M.A., and Manley, J.L. (1994). Protein-protein interactions and 50-splice-site recognition in mammalian mRNA precursors. Nature *368*, 119–124.

Komada, M., McLean, D.J., Griswold, M.D., Russell, L., and Soriano, P. (2000). E-MAP-115, encoding a microtubule-associated protein, is a retinoic acid-inducible gene required for spermatogenesis. Genes and Development *14*, 1332-1342.

Korn, R., Schoor, M., Neuhaus, H., Henseling, U., Soininen, R., Zachgo, J., and Gossler, A. (1992). Enhancer trap integrations in mouse embryonic stem cells give rise to staining patterns in chimaeric embryos with a high frequency and detect endogenous genes. Mechanisms of Development *39*, 95-109.

Kuvbachieva, A., Bestel, A.M., Tissir, F., Maloum, I., Guimiot, F., Ramoz, N., Bourgeois, F., Moalic, J.M., Goffinet, A.M., and Simonneau, M. (2005). Identification of a novel brain-specific and Reelin-regulated gene that encodes a protein colocalized with synapsin. The European Journal of Neuriscience *20*, 603-610.

Lai, W.S., Carballo, E., Strum, J.R., Kennington, E.A., Phillips, R.S., and Blackshear, P.J. (1999). Evidence that tristetraprolin binds to AU-rich elements and promotes the deadenylation and destabilization of tumor necrosis factor alpha mRNA. Molecular and Cellular Biology *19*, 4311–4323.

Lai, Z., Han,I., Park, M., and Brady, R.O. (2002). Design of an HIV-1 lentiviral-based gene-trap vector to detect developmentally regulated genes in mammalian cells. Proceedings of the National Academy of Sciences of the United States of America *99*, 3651-3656.

Lapoumeroulie, C., Pagnier, J., Bank, A., Labie, D., and Krishnamoorthy, R. (1986). β-Thalassemia due to a novel mutation in IVS 1 sequence donor site

consensus sequence creating a restriction site. Biochemical and Biophysical Research Communications *139*, 709-713.

Leighton, P.A., Mitchell, K.J., Goodrich, L.V., Lu, X., Pinson, K., Scherz, P. Skarnes, W.C., and Tessier-Lavigne, M. (2001). Defining brain wiring patterns and mechanisms through gene trapping in mice. Nature *410*, 174-179.

Li, X., Zhang, G., Ngo, N., Zhao, X., Kain, S.R., and Huang, C.C. (1997). Deletions of the Aequorea victoria green fluorescent protein define the minimal domain required for fluorescence. Journal of Biological Chemistry *272*, 28545-28549.

Li, M., Pevny, L., Lovell-Badge, R., and Smith, A. (1998). Generation of purified neural precursors from embryonic stem cells by lineage selection. Current Biology *8*, 971-974.

Lickert, H., Kutsch, S., Kanzler, B., Tamai, Y., Taketo, M. M., and Kemler, R. (2002). Formation of multiple hearts in mice following deletion of β-catenin in the embryonic endoderm. Developmental Cell *3*, 171-181.

Lien, C.L., McAnally, J., Richardson, J.A., and Olson, E.N. (2002). Cardiac-specific activity of an Nkx2-5 enhancer requires an evolutionarily conserved Smad binding site. Developmental Biology *244*, 257-266.

Loebel, D.A., Watson, C.M., De Young, R.A., and Tam, P.P. (2003). Lineage choice and differentiation in mouse embryos and embryonic stem cells. Developmental Biology *264*, 1-14.

Loflin, P., Chen, C.Y., and Shyu, A.B. (1999). Unraveling a cytoplasmic role for hnRNP D in the in vivo mRNA destabilization directed by the AU-rich element. Genes and Development *13*, 1884–1897.

Lothian, C., and Lendahl, U. (1997). An evolutionary conserved region in the second intron of the human nestin gene directs gene expression to CNS progenitor cells and to early neural crest cells. European Journal of Neuroscience *9*, 452-462.

Luo, G., Ivics, Z., Izsvak, Z., and Bradley, A. (1998). Chromosomal transposition of a *Tc1/mariner*-like element in mouse embryonic stem cells.

Proceedings of the National Academy of Sciences of the United States of America *95*, 10769-10773.

Lyon, M.F., and Morris, T. (1966). Mutation rates at a new set of specific loci in the mouse. Genetical Research *7*, 12-17.

Mager, D.L., and Freeman, J.D. (2000). Novel mouse type D endogenous proviruses and ETn elements share long terminal repeat and internal sequences. Journal of Virology *74*, 7221–7229.

Mainguy, G., Montesinos, M.L., Lesaffre, B., Zevnik, B., Karasawa, M., Kothary, R., Wurst, W., Prochiantz, A., and Volovitch, M. (2000). An induction gene trap for identifying a homeoprotein-regulated locus. Nature Biotechnology *18*, 746-749.

Maquat, L.E. (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nature Reviews Molecular Cell Biology *5*, 89-99.

Marikawa, Y., Fujita, T.C., and Alarcon, V.B. (2004). An enhancer-trap LacZ transgene reveals a distinct expression pattern of *Kinesin family 26B* in mouse embryos. Development Genes and Evolution *214*, 64-71.

Martin, G. R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proceedings of the National Academy of Sciences of the United States of America *78*, 7634-7638.

Marzluff, W.F., Gongidi, P., Woods, K.R., Jin, J., and Maltais, L.J. (2002). The human and mouse replication-dependent histone genes. Genomics *80*, 487-498.

Matsuda, E., Shigeoka, T., Iida, R., Yamanaka, S., Kawaichi, M., and Ishida, Y. (2004). Expression profiling with arrays of randomly disrupted genes in mouse embryonic stem cells leads to in vivo function analysis. Proceedings of the National Academy of Sciences of the United States of America *101*, 4170-4174.

McClive, P., Pall, G., Newton, K., Lee, M., Mullins, J., and Forrester, L. (1998). Gene trap integrations expressed in the developing heart: insertion site

affects splicing of the PT1-ATG vector. Developmental Dynamics *212*, 267-276.

Medico, E., Gambarotta, G., Gentile, A., Comoglio, P.M., and Soriano, P. (2001). A gene trap vector system for identifying transcriptionally responsive genes. Nature Biotechnology *19*, 579-582.

Meshorer, E., and Mistelli, T. (2006). Chromatin in pluripotent embryonic stem cells and differentiation. Nature Reviews Molecular Cell Biology *7* 540-546.

Mitchell, K.J., Pinson, K.I., Kelly, O.G., Brennan, J., Zupicich, J., Scherz, P., Leighton, P.A., Goodrich, L.V., Lu, X., Avery, B.J., Tate, P., Dill, K., Pangilinan, E., Wakenight, P., Tessier-Lavigne, M., and Skarnes, W.C. (2001). Functional analysis of secreted and transmembrane proteins critical to mouse development. Nature Genetics *28*, 241-249.

Mitchell, P., and Tollervey, D. (2000a). mRNA stability in eukaryotes. Current Opinion in Genetics and Development *10*, 193–198.

Mitchell, P., and Tollervey, D. (2000b). Musing on the structural organization of the exosome complex. Nature Struct Biol. 7, 843–846.

Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell *113*, 631–642.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, H.H. Jr. (1996). High frequency retrotransposition in cultured mammalian cells. Cell *87*, 917-927.

Mukherjee, D., Gao, M., O'Connor, J.P., Raijmakers, R., Pruijn, G., Lutz, C.S., and Wilusz, J. (2002). The mammalian exosome mediates the efficient degradation of mRNAs that contain AU-rich elements. The EMBO Journal *21*, 165– 174.

Muth, K., Bruyns, R., Thorey, I.S., and von Melchner, H. (1998). Disruption of genes regulated during hematopoietic differentiation of mouse embryonic stem cells. Developmental Dynamics *212*, 277-283.

Nadeau, J.H., Balling, R., Barsh, G., Beier, D., Brown, S.D., Bucan, M., Camper, S., Carlson, G., Copeland, N., Eppig, J., Fletcher, C., Frankel, W.N., Ganten, D., Goldowitz, D., Goodnow, C., Guenet, J.L., Hicks, G., Hrabe de Angelis, M., Jackson, I., Jacob, H.G., Jenkins, N., Johnson, D., Justice, M., Kay, S., Kingsley, D., Lehrach, H., Magnuson, T., Meisler, M., Poustka, A., Rinchik, E.M., Rossant, J., Russell, L.B., Schimenti, J., Shiroishi, T., Skarnes, W.C., Soriano, P., Stanford, W., Takahashi, J.S., Wurst, W., Zimmer, A; International Mouse Mutagenesis Consortium. (2001). Sequence interpretation. Functional annotation of mouse genome sequences. Science *291*, 1251-1255.

Nagai, T., Ibata, K., Park, E.S., Kubota, M., Mikoshiba, K., and Miyawaki, A. (2002). A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. Nature Biotechnology *20*, 87-90.

Nakano, T., Kodama, H. and Honjo, T. (1994). Generation of lymphohematopoietic cells from embryonic stem cells in culture. Science *265*, 1098-1101.

Natarajan, D., and Boulter, C.A. (1995). A lacZ-hygromycin fusion gene and its use in gene trap vector for marking embryonic stem cells. Nucleic Acids Research *23*, 4003-4004.

Neilan, E.G., and Barsh, G.S. (1999). Gene trap insertional mutagenesis in mice: new vectors and germ line mutations in two novel genes. Transgenic Research *8*, 451-458.

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Scholer, H., and Smith, A. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. Cell *95*, 379–391.

Nishikawa, S., Nishikawa, S., Hirashima, M., Matsuyoshi, N., and Kodama, H. (1998). Progressive lineage analysis by cell sorting and culture identifies FLK+VE-cadherin cells at a diverging point of endothelial and hemopoietic lineages. Development *125*, 1747-1757.

Niwa, M., Rose, S.D., and Berget, S.M. (1990). In vitro polyadenylation is stimulated by the presence of an upstream intron. Genes and Development *4*, 1552–1559.

Niwa, M., and Berget, S.M. (1991). Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns. Genes and Development *5*, 2086–2095.

Niwa, H., Araki, K., Kimura, S., Taniguchi, S., Wakasugi, S., and Yamamura, K. (1993). An efficient gene-trap method using polyA trap vectors and characterization of gene-trap events. Journal of Biochemistry (Tokyo) *113*, 343-349.

Niwa, H., Miyazaki, J. & Smith, A. G. (2000). Quantitative expression of *Oct-3/4* defines differentiation, dedifferentiation or self-renewal of ES cells. Nature Genetics *24*, 372–376.

Nord, A.S., Chang, P.J., Conklin, B.R., Cox, A.V., Harper, C.A., Hicks, G.G., Huang, C.C., Johns, S.J., Kawamoto, M., Liu, S., Meng, E.C., Morris, J.H., Rossant, J., Ruiz, P., Skarnes, W.C., Soriano, P., Stanford, W.L., Stryke, D., von Melchner, H., Wurst, W., Yamamura, K., Young, S.G., Babbitt, P.C., and Ferrin, T.E. (2006). The International Gene Trap Consortium Website: a portal to all publicly available gene trap cell lines in mouse. Nucleic Acids Research *34*, D642-648.

O'Kane, C.J., and Gehring, W.J. (1987). Detection *in situ* of genomic regulatory elements in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America *84*, 9123-9127.

Osipovich, A.B., White-Grindley, E.K., Hicks, G.G., Roshon, M.J., Shaffer, C., Moore, J.H., and Ruley, H.E. (2004). Activation of cryptic 3' splice sites within introns of cellular genes following gene entrapment. Nucleic Acids Research *32*, 2912-2924.

Osipovich, A.B., Singh, A., and Ruley, E.H. (2005). Post-entrapment genome engineering: first exon size does not affect the expression of fusion transcripts generated by gene entrapment. Genome Research *15*, 428-435.

Ostertag, E.M., DeBerardinis, R.J., Goodier, J.L., Zhang, Y., Yang, N., Gerton, G.L., and Kazazian, H.H. Jr. (2002). A mouse model of human L1 retrotransposition. Nature Genetics *32*, 655-660.

Pall, G.S., Wallis, J., Axton, R., Brownstein, D.G., Gautier, P., Buerger, K., Mulford, C., Mullins, J.J., and Forrester, L.M. (2004). A novel transmembrane

MSP-containing protein that plays a role in right ventricle development. Genomics *84*, 1051-1059.

Pawlak, M.R., Scherer, C.A., Chen, J., Roshon, M.J., and Ruley, H.E. (2000). Arginine N-methyltransferase 1 is required for early postimplantation mouse development, but cells deficient in the enzyme are viable. Molecular and Cellular Biology *20*, 4859-4869.

Pedelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C., and Waldo, G.S. (2006). Engineering and characterization of a superfolder green fluorescent protein. Nature Biotechnology *24*, 79-88.

Pelczar, P., and Filipowicz, W. (1998). The host gene for intronic U17 small nucleolar RNAs in mammals has no protein-coding potential and is a member of the 5′-Terminal Oligopyrimidine gene family. Molecular and Cellular Biology 18, 4509-4518.

Peng, S.S., Chen, C.Y., Xu, N., and Shyu, A.B. (1998). RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein. The EMBO Journal *17*, 3461–3470.

Perelygin, A.A., Kondrashov, F.A., Rogozin, I.B., and Brinton, M.A. (2002). Evolution of the mouse polyubiquitin-C gene. Journal of Molecular Evolution *55*, 202-210.

Pevny, L.H., Sockanathan, S., Placzek, M., and Lovell-Badge, R. (1998). A role for SOX1 in neural determination. Development *125*, 1967-1978.

Reddy, S., DeGregory, J.V., von Melchner, H., and Ruley, H.E. (1991). Retrovirus promoter-trap vector to induce *lacZ* gene fusions in mammalian cells. Journal of Virology *65*, 1507-1515.

Ren, J., Shi, M., Liu, R., Yang, Q.H., Johnson, T., Skarnes, W.C., and Du, C. (2005). The Birc6 (Bruce) gene regulates p53 and the mitochondrial pathway of apoptosis and is essential for mouse embryonic development. Proceedings of the National Academy of Sciences of the United States of America *102*, 565-570.

Ridgeway, A.G., Petropoulos, H., Wilton, S., and Skerjanc, I.S. (2000). Wnt signaling regulates the function of MyoD and myogenin. Journal of Biological Chemistry *275*, 32398-32405.

Rinchik, E.M., and Carpenter, D.A. (1999). N-ethyl-N-nitrosourea mutagenesis of a 6- to 11-cM subregion of the Fah-Hbb interval of mouse chromosome 7: Completed testing of 4557 gametes and deletion mapping and complementation analysis of 31 mutations. Genetics *152*, 373-383.

Robberson, B.L., Cote, G.J., and Berget, S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Molecular and Cellular Biology 10, 84–94.

Roberg-Perez, K., Carlson, C.M., and Largaespada, D.A. (2003). MTID: a database of Sleeping Beauty transposon insertions in mice. Nucleic Acids Research *31*, 78-81.

Romao, L., Inacio, A., Santos, S., Avila, M., Faustino, P., Pacheco, P., and Lavinha, J. (2000). Nonsense mutations in the human β-globin gene lead to unexpected levels of cytoplasmic mRNA accumulation. Blood *96*, 2895-2901.

Rossant, J. (2004). Lineage development and polar asymmetries in the peri-implantation mouse blastocyst. Seminars in Cell and Developmental Biology *15*, 573-581.

Ruiz-Echevarria, M.J., and Peltz, S.W. (2000). The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. Cell *101*, 741–751.

Russ, A.P., Friedel, C., Ballas, K., Kalina, U., Zahn, D., Strebhardt, K., and von Melchner, H. (1996). Identification of genes induced by factor deprivation in hematopoietic cells undergoing apoptosis using gene-trap mutagenesis and site-specific recombination. Proceedings of the National Academy of Sciences of the United States of America *93*, 15279-15284.

Russ, A.P., Wattler, S., Colledge, W.H., Aparicio, S.A., Carlton, M.B., Pearce, J.J., Barton, S.C., Surani, M.A., Ryan, K., Nehls, M.C., Wilson, V., and Evans, M.J. (2000). Eomesodermin is required for mouse trophoblast development and mesoderm formation. Nature *404*, 95–99.

Russell, W. L., Kelly, E.M., Hunsicker, P.R., Bangham, J.W., Maddux, S.C., and Phipps, E.L. (1979). Specific-locus test shows ethylnitrosourea to be the most potent mutagen in the mouse. Proceedings of the National Academy of Sciences of the United States of America *76*, 5818-5819.

Salminen, M., Meyer, B.I., and Gruss, P. (1998). Efficient PolyA trap approach allows the capture of genes specifically active in differentiated embryonic stem cells and mouse embryos. Developmental Dynamics *212*, 326-333.

Sam, M., Wurst, W., Kluppel, M., Jin,O., Heng, H., and Bernstein, A. (1998). Aquarius, a novel gene isolated by gene trapping with an RNA-dependent polymerase motif. Developmental Dynamics *212*, 304-317.

Scheel, J.R., Ray, J., Gage, F.H., and Barlow, C. (2005). Quantitative analysis of gene expression in living adult neural stem cells by gene trapping. Nature Methods *2*, 363-369.

Scherer, C.A., Chen, J., Nachabeh, A., Hopkins, N., and Ruley, H.E. (1996). Transcriptional specificity of the pluripotent embryonic stem cell. Cell Growth and Differentiation *7*, 1393-1401.

Shalaby, F., Rossant, J., Yamaguchi, T.P., Gertsenstein, M., Wu, X.F., Breitman, M.L., and Schuh, A.C. (1995). Failure of blood-island formation and vasculogenesis in Flk-1-deficient mice. Nature *376*, 62-66.

Shalaby, F., Ho, J., Stanford, W.L., Fischer, K.D., Schuh, A.C., Schwartz, L., Bernstein, A., and Rossant, J. (1997). A requirement for Flk1 in primitive and definitive hematopoiesis and vasculogenesis. Cell *89*, 981-990.

Shaw, G. and Kamen, R. (1986). A conserved AU sequence from the 3' untranslated region of GM-CSF messenger-RNA mediates selective messenger-RNA degradation. Cell *46*, 659-667.

Sheth, U. and Parker, R. (2003) Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. Science *300*, 805–808.

Shigeoka, T., Kawaichi, M., and Ishida, Y. (2005). Suppression of nonsense-mediated mRNA decay permits unbiased gene trapping in mouse embryonic stem cells. Nucleic Acids Research *33*, e20.

Shim, J., Lim, H., R Yates, J., and Karin, M. (2002). Nuclear export of NF90 is required for interleukin-2 mRNA stabilization. Molecular Cell *10*, 1331–1344.

Schnutgen, F., De-Zolt, S., Van Sloun, P., Hollatz, M., Floss, T., Hansen, J., Altschmied, J., Seisenberger, C., Ghyselinck, N.B., Ruiz, P., Chambon, P.,

Wurst, W., and von Melchner, H. (2005). Genomewide production of multipurpose alleles for the functional analysis of the mouse genome. Proceedings of the National Academy of Sciences of the United States of America *102*, 7221-7226.

Shyu, A.B., Greenberg, M.E., and Belasco, J.G. (1989). The c-fos transcript is targeted for rapid decay by two distinctmRNAdegradation pathways. Genes and Development *3*, 60–72.

Shyu, A.B., Belasco, J.G., and Greenberg, M.E. (1991). Two distinct destabilizing elements in the c-fos message trigger deadenylation as a first step in rapid mRNA decay. Genes and Development *5*, 221–231.

Serafini, T., Colamarino, S.A., Leonardo, E.D., Wang, H., Beddington, R., Skarnes, W.C., and Tessier-Lavigne, M. (1996). Netrin-1 is required for commissural axon guidance in the developing vertebrate nervous system. Cell *87*, 1001-1014.

Shatkin, A.J., Manley, J.L. (2000). The ends of the affair: capping and polyadenylation. Nature Structural Biology *7*, 838–842.

Shirai, M., Miyashita, A., Ishii, N., Itoh, Y., Satokata, I., Watanabe, Y.G., and Kuwano, R. (1996). A gene trap strategy for identifying the genes expressed in the embryonic nervous system. Zoological Science *13*, 277-283.

Silva, J.M., Li, M.Z., Chang, K., Ge, W., Golding, M.C., Rickles, R.J., Siolas, D., Hu, G., Paddison, P.J., Schlabach, M.R., Sheth, N., Bradshaw, J., Burchard, J., Kulkarni, A., Cavet, G., Sachidanandam, R., McCombie, W.R., Cleary, M.A., Elledge, S.J., and Hannon, G.J. (2005). Second-generation shRNA libraries covering the human and mouse genomes. Nature Genetics 37, 1281-1288.

Sirard, C., de la Pompa, J.L., Elia, A., Itie, A., Mirtsos, C., Cheung, A., Hahn, S., Wakeham, A., Schwartz, L., Kern, S.E., Rossant, J., and Mak, T.W. (1998). The tumor suppressor gene Smad4/Dpc4 is required for gastrulation and later for anterior development of the mouse embryo. Genes and Development *12*, 107-119.

Skarnes, W.C., Auerbach, B.A. and Joyner, A.L. (1992). A gene trap approach in mouse embryonic stem cells: The *lacZ* reporter gene is activated by splicing, reflects endogenous gene expression, and is mutagenic in mice. Genes and Development *6*, 903-918.

Skarnes, W.C., Moss, J.E., Hurtley, S.M. and Beddington, R.S. (1995). Capturing genes encoding membrane and secreted proteins important for mouse development. Proceedings of the National Academy of Sciences of the United States of America *92*, 6592-6596.

Skarnes, W.C., von Melchner, H., Wurst, W., Hicks, G., Nord, A.S., Cox, T., Young, S.G., Ruiz, P., Soriano, P., Tessier-Lavigne, M., Conklin, B.R., Stanford, W.L., Rossant, J; International Gene Trap Consortium. (2004). A public gene trap resource for mouse functional genomics. Nature Genetics *36*, 543-544.

Skarnes, W.C. (2005). Two ways to trap a gene in mice. Proceedings of the National Academy of Sciences of the United States of America *102*, 13001-13002.

Sledz, C.A., Holko, M., de Veer, M.J., Silverman, R.H., and Williams, B.R. (2003). Activation of the interferon system by short-interfering RNAs. Nature Cell Biology *5*, 834–839.

Smith, A.G. (1991). Culture and differentiation of embryonic stem cells. Journal of Tissue Culture Methods *13*, 89-94.

Smith, C.M., and Steitz, J.A. (1998). Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. Molecular and Cellular Biology *12*, 6897-6909.

Soriano, P., Gridley, T., and Jaenisch, R. (1987). Retroviruses and insertional mutagenesis in mice: proviral integration at the Mov 34 locus leads to early embryonic death. Genes and Development *1*, 366-375.

Spence, S.E., Gilbert, D.J., Swing, D.A., Copeland, N.G., and Jenkins, N.A. (1989). Spontaneous germ line virus infection and retroviral insertional mutagenesis in eighteen transgenic Srev lines of mice. Molecular and Cellular Biology *9*, 177-184.

Stanford, W.L., Caruana, G., Vallis, K.A., Inamdar, M., Hidaka, M., Bautch, V.L., and Bernstein, A. (1998). Expression trapping: Identification of novel genes expressed in hematopoietic and endothelial lineages by gene trapping in ES cells. Blood *92*, 4622-4631.

Stanford, W. L., Cohn, J. B. and Cordes, S. P. (2001). Gene-trap mutagenesis: past, present and beyond. Nature Reviews Genetics *2*, 756-768.

Stark G.R., Kerr, I.M., Williams, B.R., Silverman, R.H., and Schreiber, R.D. (1998). How cells respond to interferons. Annual Review of Biochemistry *67*, 227–264.

Stoecklin, G., Colombi, M., Raineri, I., Leuenberger, S., Mallaun, M., Schmidlin, M., Gross, B., Lu, M., Kitamura, T., and Moroni, C. (2002). Functional cloning of BRF1, a regulator of ARE-dependent mRNA. The EMBO Journal *21*, 4709–4718.

Strumpf, D., Mao, C.A., Yamanaka, Y., Ralston, A., Chawengsaksophak, K., Beck, F., and Rossant, J. (2005). Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. Development *132*, 2093–2102.

Stryke, D., Kawamoto, M., Huang, C.C., Johns, S.J., King, L.A., Harper, C.A., Meng, E.C., Lee, R.E., Yee, A., L'Italien, L., Chuang, P.T., Young, S.G., Skarnes, W.C., Babbitt, P.C., and Ferrin, T.E. (2003). BayGenomics: a resource of insertional mutations in mouse embryonic stem cells. Nucleic Acids Research *31*, 278-281.

Stuhlmann, H. (2003). Gene trap vector screen for developmental genes in differentiating ES cells. Methods in Enzymology *365*, 386-406.

Sutherland, H.G.E., Mumford, G.K., Newton, K., Ford, L.V., Farrall, R., Dellaire, G., Caceres, J.F., and Bickmore, W.A. (2001). Large-scale identification of mammalian proteins localized to nuclear sub-compartments. Human Molecular Genetics *10*, 1995-2001.

Szymczak, A.L., Workman, C.J., Wang, Y., Vignali, K.M., Dilioglou, S., Vanin, E.F., and Vignali, D.A. (2004). Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide-based retroviral vector. Nature Biotechnology *22*, 589-594.

Takahashi, H., Maeda, M., Sawa, H., Hasegawa, H., Moryiama, M., Sata, T., Hall, W.W., Kurata, T. (2006). Dicer and positive charge of proteins decrease the stability of RNA containing the AU-rich element of GM-CSF. Biochemical and Biophysical Research Communications *340*, 807-814.

Tanaka, I., and Ishihara, H. (2001). Enhanced expression of the early retrotransposon in C3H mouse-derived myeloid leukemia cells. Virology *280*, 107–114.

Taniwaki, T., Haruna, K., Nakamura, H., Sekimoto, T., Oike, Y., Imaizumi, T., Saito, F., Muta, M., Soejima, Y., Utoh, A., Nakagata, N., Araki, M., Yamamura, K., and Araki, K. (2005). Characterization of an exchangeable gene trap using pU-17 carrying a stop codon-*βgeo* cassette. Development Growth and Differentiation *47*, 163-172.

Tarrant, J.M., Groom, J., Metcalf, D., Li, R., Borobokas, B., Wright, M.D., Tarlinton, D., and Robb, L. (2002). The absence of Tssc6, a member of the tetraspanin superfamily, does not affect lymphoid development but enhances in vitro T-cell proliferative responses. Molecular and Cellular Biology *22*, 5006-5018.

Tate, P., Lee, M., Tweedie, S., Skarnes, W.C., and Bickmore, W.A. (1998). Capturing novel mouse genes encoding chromosomal and other nuclear proteins. Journal of Cell Science *111*, 2575-2585.

Tateossian, H., Powles, N., Dickinson, R., Ficker, M., and Maconochie, M. (2004). Determination of downstream targets of FGF signalling using gene trap and cDNA subtractive approaches. Experimental Cell Research *292*, 101-114.

Tazi, J., Durand, S., and Jeanteur, P. (2005). The spliceosome: a novel multi-faceted target for therapy. Trends in Biochemcal Sciences *30*, 469-478.

Thorey, I.S., Muth, K., Russ, A.P., Otte, J., Reffelmann, A., and von Melchner, H. (1998). Selective disruption of genes transiently induced in differentiating mouse embryonic stem cells by using gene trap mutagenesis and site-specific recombination. Molecular and Cellular Biology *18*, 3081-3088.

To, C., Epp, T., Reid, T., Lan, Q., Yu, M., Li, C.Y., Ohishi, M., Hant, P., Tsao, N., Casallo, G., Rossant, J., Osborne, L.R., and Stanford, W.L. (2004). The Centre for Modelling Human Disease Gene Trap resource. Nucleic Acids Research *32*, D557-D559.

Tokuzawa, Y., Kaiho, E., Maruyama, M., Takahashi, K., Mitsui, K., Maeda, M., Niwa, H., and Yamanaka, S. (2003). Fbx15 is a novel target of Oct3/4 but

is dispensable for embryonic stem cell self-renewal and mouse development. Molecular and Cellular Biology 23, 2699-2708.

Torres, M., Stoykova, A., Huber, O., Chowdhury, K., Bonaldo, P., Mansouri, A., Butz, S., Kemler, R., and Gruss, P. (1997). An alpha-E-catenin gene trap mutation defines its function in preimplantation development. Proceedings of the National Academy of Sciences of the United States of America *94*, 901-906.

Townley, D.J., Avery, B.J., Rosen, B., and Skarnes, W.C. (1997). Rapid sequence analysis of gene trap integrations to generate a resource of insertional mutations in mice. Genome Research *7*, 293-298.

Tran, H., Schilling, M., Wirbelauer, C., Hess, D., and Nagamine, Y. (2004). Facilitation of mRNA Deadenylation and Decay by the Exosome-Bound, DExH Protein RHAU. Molecular Cell *13*, 101-111.

Tsuiji, M., Fujimori, M., Ohashi, Y., Higashi, N., Onami, T.M., Hedrick, S.M., and Irimura, T. (2002). Molecular cloning and characterization of a novel mouse macrophage C-type lectin, mMGL2, which has a distinct carbohydrate specificity from mMGL1. Journal of Biological Chemistry *277*, 28892-28901.

Tsukahara, M., Suemori, H., Noguchi, S., Ji, Z.S., and Tsunoo, H. (2000). Novel nucleolar protein, midnolin, is expressed in the mesencephalon during mouse development. Gene *254*, 45-55.

Tsukahara, M., Ji, Z.S., Noguchi, S., and Tsunoo, H. (2001). A novel putative transmembrane protein, IZP6, is expressed in neural cells during embryogenesis. Development, Growth and Differentiation *43*, 285-293.

Tzouanacou, E., Tweedie, S., and Wilson, V. (2005). Identification of Jade1, a gene encoding a PHD zinc finger protein, in a gene trap mutagenesis screen for genes involved in anteroposterior axis development. Molecular and Cellular Biology *23*, 8553-8562.

Vagner, S., Vagner, C., and Mattaj, I.W. (2000). The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF65 to couple 30-end processing and splicing. Genes and Development *14*, 403–413.

Vidal, F., Lopez, P., Lopez-Fernandez, L.A., Ranc, F., Scimeca, J.C., Cuzin, F., Rassoulzadegan, M. (2001). Gene trap analysis of germ cell signalling to

Sertoli cells: NGF-TrkA mediated induction of Fra1 and Fos by post-meiotic germ cells. Journal of Cell Science *114*, 435-443.

Vijaya, S., Steffen, D.L., and Robinson, H.L. (1986). Acceptor sites for retroviral integrations map near DNaseI-hypersensitive sites in chromatin. Journal of Virology *60*, 683-692.

Visa, N., Izaurralde, E., Ferreira, J., Daneholt, B., and Mattaj, I.W. (1996). A nuclear cap-binding complex binds Balbiani ring pre-mRNA cotranscriptionally and accompanies the ribonucleoprotein particle during nuclear export. Journal of Cell Biology *133*, 5–14.

von Melchner, H., and Ruley, H. E. (1989). Identification of cellular promoters by using a retrovirus promoter trap. Journal of Virology *63*, 3227-3233.

von Melchner, H., Reddy, S., and Ruley, H.E. (1990). Isolation of cellular promoters by using a retrovirus promoter trap. Proceedings of the National Academy of Sciences of the United States of America *87*, 3733-3737.

von Melchner, H. DeGregori, J.V, Rayburn, H., Reddy, S., Friedel, C., and Ruley, H.E. (1992). Selective disruption of genes expressed in totipotent embryonal stem cells. Genes and Development *6*, 919-927.

Voss, A.K., Thomas, T., and Gruss, P. (1998). Efficiency assessment of the gene trap approach. Developmental Dynamics *212*, 171-180.

Voss, A.K., Thomas, T., Petrou, P., Anastassiadis, K., Scholer, H., and Gruss, P. (2000). Taube nuss is a novel gene essential for the survival of pluripotent cells of early mouse embryos. Development *127*, 5449-5461.

Wagner, E.F., Stewart, T.A., and Mintz, B. (1981). The human β-globin gene and a functional viral thymidine kinase gene in developing mice. Proceedings of the National Academy of Sciences of the United States of America *78*, 5016-5020.

Wang, H. and Dey, S.K. (2006). Roadmap to embryo implantation: clues from mouse models. Nature Reviews Genetics *7*, 185-199.

Waterston, R.H. et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Watt, A.J., Jones, E.A., Ure, J.M., Peddie, D., Wilson, D.I., and Forrester, L.M. (2001). A gene trap integration provides an early in situ marker for hepatic specification of the foregut endoderm. Mechanisms of Development *100*, 205-215.

Weiher, H., Noda, T., Gray, D.A., Sharpe, A.H., and Jaenisch, R. (1990). Transgenic mouse model of kidney disease: insertional inactivation of ubiquitously expressed gene leads to nephrotic syndrome. Cell *62*, 425-434.

Wempe, F., Yang, J-Y, Hammann, J., and von Melchner, H. (2001). Gene trapping identifies transiently induced survival genes during programmed cell death. Genome Biology 2, research0023.1-0023.10.

Whitney, M., Rockenstein, E., Cantin, G., Knapp, T., Zlokarnik, G., Sanders, P., Durick, K., Craig, F.F., and Negulescu, P.A. (1998). A genome-wide functional assay of signal transduction in living mammalian cells. Nature Biotechnology *16*, 1329-1333.

Williamson, D.J., Banik-Maiti, S., DeGregori, J., and Ruley, H.E. (2000). hnRNP C is required for postimplantation mouse development but is dispensable for cell viability. Molecular and Cellular Biology *20*, 4094-4105.

.

Wilson, V., Manson, L., Skarnes, W.C., and Beddington, R.S.P. (1995). The *T* gene is necessary for normal mesoderm morphogenetic cell movements during gastrulation. Development *121*, 877-886.

Wilson, T., and Treisman, R. (1988). Removal of poly(A) and consequent degradation of c-fos mRNA facilitated by 3′ AU-rich sequences. Nature *336* 396– 399.

Winnier, G., Blessing, M., Labosky, P.A., and Hogan, B.L. (1995). Bone morphogenetic protein-4 is required for mesoderm formation and patterning in the mouse. Genes and Development *9*, 2105-2116.

Wolpert, L., Beddington, R., Jessell, T., Lawrence, P., Meyerowitz, E., and Smith, J. (2002). Principles of Development. Oxford, Oxford University Press.

Wu, J.Y., and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. Cell *75*, 1061–1070.

Wu, S., Romfo, C.M., Nilsen, T.W., and Green, M.R. (1999). Functional recognition of the 30 splice site AG by the splicing factor U2AF35. Nature *402*, 832–835.

Wurst, W., Rossant, J., Prideaux, V., Kownacka, M., Joyner, A., Hill, D.P., Guillemot, F., Gasca, S., Cado, D., Auerbach, A., et al. (1995). A large-scale gene-trap screen for insertional mutations in developmentally regulated genes in mice. Genetics *139*, 889-899.

Xin, H-B., Deng, K.Y., Shui, B., Qu, S., Sun, Q., Lee, J., Greene, K.S., Wilson, J., Yu, Y., Feldman, M., and Kotlikoff, M.I. (2005). Gene trap and gene inversion methods for conditional gene inactivation in the mouse. Nucleic Acids Research *33*, e14.

Xiong, J-W., Battaglino, R., Leahy, A., and Stuhlmann, H. (1998). Large-scale screening for developmental genes in embryonic stem cells and embryoid bodies using retroviral entrapment vectors. Developmental Dynamics *212*, 181-197.

Xu, N., Chen, C.Y., and Shyu, A.B. (1997). Modulation of the fate of cytoplasmic mRNA by AU-Rich elements: key sequence features controlling mRNA deadenylation and decay. Molecular and Cellular Biology *17*, 4611–4621.

Yamaguchi, Y., Ogura, S., Ishida, M., Karasawa, M., and Takada, S. (2005). Gene trap screening as an effective approach for identification of Wnt-responsive genes in the mouse embryo. Developmental Dynamics *233*, 484-495.

Yant, S.R., Wu, X., Huang, Y., Garrison, B., Burgess, S.M., and Kay, M.A. (2005). High-resolution genome-wide mapping of transposon integration in mammals. Molecular and Cellular Biology *25*, 2085-2094.

Ye, W., Shimamura, K., Rubenstein, J.L., Hynes, M.A., and Rosenthal, A. (1998). FGF and Shh signals control dopaminergic and serotonergic cell fate in the anterior neural plate. Cell *93*, 755-766.

Yoshida, M., Yagi, T., Furuta, Y., Takayanagi, K., Kominami, R., Takeda, N., Tokunaga, T., Chiba, J., Ikawa, Y., and Aizawa, S. (1995). A new strategy of gene trapping in ES cells using 3'RACE. Transgenic Research *4*, 277-287.

Yoshikawa, Y., Fujimori, T., McMahon, A.P., and Takada, S. (1997). Evidence that absence of Wnt-3a signaling promotes neuralization instead of paraxial mesoderm development in the mouse. Developmental Biology *183*, 234-242.

Yuan, L., Moyon, D., Pardanaud, L., Breant, C., Karkkainen, M.J., Alitalo, K., and Eichmann, A. (2002). Abnormal lymphatic vessel development in neuropilin 2 mutant mice. Development *129*, 4797-4806.

Zambrowicz, B.P., Friedrich, G.A., Buxton, E.C., Lilleberg, S.L., Person, C., Sands, A.T. (1998). Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. Nature *392*, 608-611.

Zambrowicz, B.P., Abuin, A., Ramirez-Solis, R., Richter, L.J., Piggott, J., BeltrandelRio, H., Buxton, E.C., Edwards, J., Finch, R.A., Friddle, C.J., Gupta, A., Hansen, G., Hu, Y., Huang, W., Jaing, C., Key, B.W. Jr., Kipp, P., Kohlhauff, B., Ma, Z.Q., Markesich, D., Payne, R., Potter, D.G., Qian, N., Shaw, J., Schrick, J., Shi, Z.Z., Sparks, M.J., Van Sligtenhorst, I., Vogel, P., Walke, W., Xu, N., Zhu, Q., Person, C., and Sands, A.T. (2003). Wnk1 kinase deficiency lowers blood pressure in mice: A gene-trap screen to identify potential targets for therapeutic intervention. Proceedings of the National Academy of Sciences of the United States of America *100*, 14109-14114.

Zhang, T., Kruys, V., Huez, G., and Gueydan, C. (2002). AU-rich element mediated translational control: complexity and multiple activities of trans-activating factors. Biochemical Society Transactions *30*, 952-958.

Zheng, X.H., and Hughes, S.H. (1999). An avian sarcoma/leucosis virus-based gene trap vector for mammalian cells. Journal of Virology *73*, 6946-6952.

Zhou, H.M., Weskamp, G., Chesneau, V., Sahin, U., Vortkamp, A., Horiuchi, K., Chiusaroli, R., Hahn, R., Wilkes, D., Fisher, P., Baron, R., Manova, K., Basson, C.T., Hempstead, B., Blobel, C.P. (2004). Essential role for ADAM19 in cardiovascular morphogenesis. Molecular and Cellular Biology *24*, 96-104.

Zlokarnik, G., Negulescu, P.A., Knapp, T.E., Mere, L., Burres, N., Feng, L.X., Whitney, M., Roemer, K., and Tsien, R.Y. (1998). Quantitation of transcription and clonal selection of single living cells with beta-lactamase as reporter. Science *279*, 84-88.