BMC Genomics

BioMed Central

Research article

Open Access

Gene trap mutagenesis of hnRNP A2/B1: a cryptic 3' splice site in the neomycin resistance gene allows continued expression of the disrupted cellular gene

Michael Roshon*[1,2], James V DeGregori[1,3] and H Earl Ruley[1]

Address: [1]Department of Microbiology and Immunology, Room AA4210 MCN, Vanderbilt University School of Medicine, 1161 21st Ave South, Nashville, TN. 37232-2363, USA, [2]Present Address: Department of Emergency Medicine, Carolinas Medical Center, PO Box 32861, Charolette, NC 28232-2861, USA and [3]Univ. of Colorado Health Sci. Center, 4200 E. 9th Ave., Box C229 (or room BRB802), Denver, CO 80262, USA

Email: Michael Roshon* - mroshon@carolinas.org; James V DeGregori - james.degregori@uchsc.edu; H Earl Ruley - ruleye@email.mc.vanderbilt.edu

* Corresponding author

Published: 20 January 2003

BMC Genomics 2003, **4**:2

Received: 7 October 2002
Accepted: 20 January 2003

This article is available from: http://www.biomedcentral.com/1471-2164/4/2

Abstract

Background: Tagged sequence mutagenesis is a process for constructing libraries of sequenced insertion mutations in embryonic stem cells that can be transmitted into the mouse germline. To better predict the functional consequences of gene entrapment on cellular gene expression, the present study characterized the effects of a U3Neo gene trap retrovirus inserted into an intron of the hnRNP A2/B1 gene. The mutation was selected for analysis because it occurred in a highly expressed gene and yet did not produce obvious phenotypes following germline transmission.

Results: Sequences flanking the integrated gene trap vector in 1B4 cells were used to isolate a full-length cDNA whose predicted amino acid sequence is identical to the human A2 protein at all but one of 341 amino acid residues. hnRNP A2/B1 transcripts extending into the provirus utilize a cryptic 3' splice site located 28 nucleotides downstream of the neomycin phosphotransferase start codon. The inserted Neo sequence and proviral poly(A) site function as an 3' terminal exon that is utilized to produce hnRNP A2/B1-Neo fusion transcripts, or skipped to produce wild-type hnRNP A2/B1 transcripts. This results in only a modest disruption of hnRNPA2/B1 gene expression.

Conclusions: Expression of the occupied hnRNP A2/B1 gene and utilization of the viral poly(A) site are consistent with an exon definition model of pre-mRNA splicing. These results reveal a mechanism by which U3 gene trap vectors can be expressed without disrupting cellular gene expression, thus suggesting ways to improve these vectors for gene trap mutagenesis.

Background

Gene entrapment has provided effective strategies for insertional mutagenesis of mammalian cells in culture. The mutagens permit direct selection of clones in which cellular genes have been disrupted and simplify the characterisation of genes associated with recessive mutations [1]. Mutagenesis of embryo-derived stem (ES) cells, coupled with in vitro genetic screens, has been widely used to analyse gene functions in mice [1]. These have included screens for mutations in developmentally regulated genes [2–5], in genes regulated by extracellular agonists [6,7], and in genes encoding secreted and transmembrane proteins [8,9]. Characterized mutations include genes involved in intracellular trafficking [10], transcriptional regulation [11,12], signal transduction [7,8,11,13–15], neural development [16] and neural wiring [17], and axial

patterning [18,19]. The rapid expansion of the nucleic acid databases has had a tremendous impact on the identification of genes disrupted by gene entrapment. This has led to the development of tagged sequence mutagenesis, a process by which genes disrupted in ES cells are characterized at the nucleotide level prior to germline transmission [20–24].

Gene trap retroviruses developed in our laboratory contain a selectable marker in the U3 region of the long terminal repeat (LTR) of a replication-defective Moloney murine leukemia virus. Selection for U3 gene expression generates clones in which the provirus is positioned in or near exons of actively transcribed genes and is expressed on transcripts originating in the flanking cellular DNA [25]. The vectors appear to be effective mutagens. Single-gene mutation frequencies are 100–1000 fold higher in cells isolated after gene trap selection than in cells containing randomly integrated retroviruses [26]. These targeting frequencies also support the idea that retrovirus can integrate throughout the genome and that most, if not all, expressed genes can be disrupted. Finally, approximately 40% of inserts selected in ES cells result in obvious phenotypes following transmission into the mouse germline [27–29]. In the four cases examined, the virus appeared to induce null mutations [10,30–32].

In order to best utilise gene traps in genetic studies, it is necessary to understand the factors that allow expression of the entrapment vectors and that determine whether expression of the occupied gene will be disrupted. This is particularly true for tagged sequence mutagenesis, where one would like to predict by sequence alone the effects of the targeting vector on cellular gene expression. For this, a representative number inserts must be characterised including those not associated with any discernible phenotype. Most previously analysed mutations were selected because of phenotypes observed after germline transmission, and thus are unlikely to reveal mechanisms that could allow expression of the entrapment vector without disrupting expression of the occupied gene.

The present study characterized a mutation in the 1B4 cell line induced by the U3Neo gene trap retrovirus [29]. This insert was selected for study because the provirus inserted into a widely expressed gene and yet no phenotype was observed in mice homozygous for the provirus. Further analysis revealed that the provirus integrated into an intron of murine homologue of the human hnRNP A2/B1 gene. The gene encodes two related nuclear ribonucleoproteins, hnRNP A2 and hnRNP B1, members of a large family of RNA binding proteins found associated with mammalian heterogeneous nuclear RNA [33].

Results

The 1B4 Provirus integrates into the hnRNPA2/B1 gene

The 1B4 cell line was isolated by infecting D3 ES cells with the U3Neo gene trap retrovirus and selecting for G418 resistant clones [29]. 1B4 cells contain a single, intact provirus as assessed by Southern blot hybridization (data not shown). Sequences flanking the virus, isolated by inverse PCR, hybridized to a transcript of approximately 1.8 Kb, and were used to screen a PCC3 embryonal carcinoma cell cDNA library. A total of 55 positive plaques were identified among 1×10^6 plaques screened. Further analysis of ten cDNAs revealed two overlapping clones covering the entire 1.8 Kb transcript. The composite cDNA contained an open reading frame encoding a polypeptide of 341 amino acids (Figure 1). Comparison of the translated sequence to the GenBank database using the BLASTP program [34] revealed a significant match with human hnRNP A2/B1 [35]. The human and mouse proteins are identical except asparagine 287 in the human sequence was replaced by a threonine (Figure 1). The mouse cDNA sequence has been deposited in GenBank (accession number AF073993).

In order to determine where the provirus inserted within the hnRNP A2/B1 gene, genomic DNA flanking the 1B4 provirus was sequenced. The flanking DNA isolated by inverse PCR extended to a HinfI site 226 nucleotides upstream and downstream of the provirus (Figure 2). Sequences matching the cloned cDNA and the human hnRNP A2/B1 cDNA extended 64 nucleotides downstream of the HinfI site, while the remaining 159 nucleotides did not match the cDNA sequence. A consensus 5' splice site was located at the point where the genomic and cDNA sequences diverged. Therefore the 1B4 provirus appeared to integrate 159 nucleotides into an intron of the hnRNP A2/B1 gene. The flanking sequence has been submitted to GenBank (accession number AF073990).

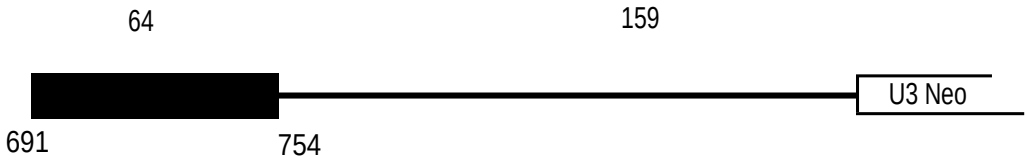
The fact that the flanking genomic DNA hybridized to a single genomic DNA fragment suggested that the provirus inserted into the hnRNP A2/B1 gene and not into a related but uncharacterized gene. However, since the match was based on a relatively short stretch of exon, several experiments were performed to confirm linkage between provirus and the hnRNP A2/B1 gene. First, two primers complementary to hnRNP A2/B1 sequences located upstream of the provirus were used together with a neo specific primer in separate PCR reactions. In each case, the size of the amplified product was consistent with insertion of the provirus into the hnRNPA2/B1 gene (data not shown). Second, cDNA sequences predicted to lie downstream of the integration site were used to probe Southern blots. The 3' hnRNP A2/B1 cDNA probe hybridized to a 18 kB EcoR1 fragment corresponding to the wild type gene and to a 22 kB fragment in DNA from mice

```
CACAGCGCCAGGACGAGTCCCGTGCGCGTCCGTCCGCGGAGGTCTTTCTCATCTCGCTCGGCTGCGGGAAATCGGGCTGA   80
AGCGACTGAGTCCGCG ATG GAG AGA GAA AAG GAA CAG TTC CGA AAG CTC TTT ATT GGT GGC TTA  144
                  M   E   R   E   K   E   Q   F   R   K   L   F   I   G   G   L    16

AGC TTT GAA ACC ACA GAA GAA AGT TTG AGA AAC TAC TAT GAG CAA TGG GGA AAG CTC ACA   204
 S   F   E   T   T   E   E   S   L   R   N   Y   Y   E   Q   W   G   K   L   T    36

GAC TGT GTG GTT ATG CGG GAT CCT GCA AGC AAA AGA TCA AGA GGA TTT GGC TTT GTA ACT   264
 D   C   V   V   M   R   D   P   A   S   K   R   S   R   G   F   G   F   V   T    56

TTC TCA TCC ATG GCC GAG GTT GAC GCT GCC ATG GCT GCA AGG CCT CAT TCC ATT GAT GGC   324
 F   S   S   M   A   E   V   D   A   A   M   A   A   R   P   H   S   I   D   G    76

AGG GTA GTT GAG CCA AAA CGT GCT GTA GCA AGA GAG GAG TCT GGA AAA CCA GGA GCC CAT   384
 R   V   V   E   P   K   R   A   V   A   R   E   E   S   G   K   P   G   A   H    96

GTG ACT GTG AAG AAG CTG TTT GTT GGT GGA ATT AAG GAA GAT ACT GAG GAA CAC CAC CTT   444
 V   T   V   K   K   L   F   V   G   G   I   K   E   D   T   E   E   H   H   L   116

AGA GAT TAC TTT GAA GAG TAT GGA AAA ATT GAT ACT ATT GAA ATA ATT ACC GAT AGG CAG   504
 R   D   Y   F   E   E   Y   G   K   I   D   T   I   E   I   I   T   D   R   Q   136

TCT GGA AAG AAA AGA GGC TTT GGC TTT GTT ACT TTT GAT GAC CAT GAT CCT GTG GAT AAA   564
 S   G   K   K   R   G   F   G   F   V   T   F   D   D   H   D   P   V   D   K   156

ATT GTC TTG CAA AAA TAT CAC ACC ATA AAT GGT CAC AAT GCA GAA GTT AGA AAG GCA TTG   624
 I   V   L   Q   K   Y   H   T   I   N   G   H   N   A   E   V   R   K   A   L   176

TCT AGA CAA GAA ATG CAG GAA GTC CAA AGT TCT AGG AGT GGA AGA GGA GGA AAC TTT GGT   684
 S   R   Q   E   M   Q   E   V   Q   S   S   R   S   G   R   G   G   N   F   G   196

TTT GGG GAT TCT CGT GGT GGC GGT GGC AAT TTT GGA CCA GGA CCA GGA AGC AAC TTT AGG   744
 F   G   D   S   R   G   G   G   G   N   F   G   P   G   P   G   S   N   F   R   216

GGG GGA TCT GAT GGA TAC GGA AGT GGA CGT GGA TTT GGG GAT GGC TAT AAT GGG TAT GGA   804
 G   G   S   D   G   Y   G   S   G   R   G   F   G   D   G   Y   N   G   Y   G   236

GGA GGA CCT GGA GGT GGC AAT TTT GGA GGT AGC CCT GGT TAT GGA GGA GGA AGA GGA GGA   864
 G   G   P   G   G   G   N   F   G   G   S   P   G   Y   G   G   G   R   G   G   256

TAT GGT GGT GGA GGA CCT GGA TAT GGC AAC CAG GGT GGG GGC TAC GGA GGT GGT TAT GAC   924
 Y   G   G   G   G   P   G   Y   G   N   Q   G   G   G   Y   G   G   G   Y   D   276

AAC TAT GGA GGA GGA AAT TAT GGA AGT GGA AGT TAC AAT GAT TTT GGA AAT TAT AAC CAG   984
 N   Y   G   G   G   N   Y   G   S   G   T   Y   N   D   F   G   N   Y   N   Q   296

CAG CCT TCT AAC TAT GGT CCA ATG AAG AGT GGA AAC TTT GGG GGT AGC AGG AAC ATG GGA  1044
 Q   P   S   N   Y   G   P   M   K   S   G   N   F   G   G   S   R   N   M   G   316

GGA CCA TAT GGT GGA GGG AAC TAT GGT CCT GGA GGA AGT GGA GGA AGT GGT GGC TAT GGT  1104
 G   P   Y   G   G   G   N   Y   G   P   G   G   S   G   G   S   G   G   Y   G   336

GGA AGG AGC AGA TAT TAA GCTTCTTCCTACTTACCATGGGCTTCACTGTATAAATAGGAGAGGATGAGAGCCC  1177
 G   R   S   R   Y  ***                                                          341

AGAGGTAACAGAACAGCTTCAGGTTATGGAAATAACAATGTTAAGGAAACTCTTACTCTCAGTCATGCATAAATATCAGT 1257
GGTATGGCAGAAGACACCAGAGCAGATGCAGACAGCCATTTTGTGAATGGATTGGATTATTTAATAACATTACCTTACTG 1337
TGGAGGAAGGATTGTAAAAAAAAATGCCTTTGAGACAGTTCTTAGGCTTTTTAATTGTTGTTTCTTTCTAGTGGTCTTTG 1417
TAAGAGTGTAGAAGCATTCCTTCTTTGATAATGTTAAATTTGTAAGTTTCAGGTGACATGTGAAACCTTTTTTAAGATTT 1497
TTCTCAAAAGTTTTGAAAAGCTATTAGCCAGGATCATGGTGTAATAAGACATAACGTTTTTCCTTCCCAATTTTAAGTGC 1577
GTGTGTAGAGTTAAGAAGCTGTTGTACATTTATGATTTAATAAAATATTCTAAAGGAAAAAAAAAAAAAAAAAAAAAAA 1657
AAAAAAAAAAAAAAAAAAAA                                                             1677
```

Figure 1

Sequence of the murine hnRNP A2/B1 gene. The complete nucleotide sequence and predicted amino acid sequence of the murine hnRNP A2/B1 cDNA was determined from two overlapping clones. The predicted amino acid sequence is identical to the predicted human protein with the exception of the substitution of threonine for asparagine at amino acid 287 (shaded). The site of provirus integration in 1B4 cells is indicated. Flanking sequences isolated by inverse PCR colinerar with the cDNA and a polyadenylation signal are underlined.

A



B

GATTCTCGTGGTGGCGGTGGCAATTTTGGACCAGGACCAGGAAGCAACTTTAGGGGGGGA -164

TCTGGTAAGTTTGTTTCTTAAGGGTGTGATTGCATTTTGAATCAGAATGTCTTCAGTGAG -104

ATAAATTTTGCAGTTAGCCTGTAGTGTGTAGTGCGAGAAAGAAAGGAGACCCTGTGCTCA -44

GACAAGGTTAAGATGTTGCCCTCTGACCACCATCATGTAAACT U3 NEO

Figure 2
Integration of the 1B4 provirus occurs in an intron of the hnRNP A2/B1 gene. (A) Schematic representation of upstream flanking sequence isolated by inverse PCR (1B4 Flanking) shows that the 1B4 provirus integrated 159 nucleotides into an intron of the hnRNP A2/B1 gene. Exon and intron sequences are indicated by the black box and thin line, respectively. Nucleotide numbers correspond to the cDNA sequence in Figure 1. (B) Genomic sequences are colinear with the cloned cDNA sequences for 64 nt. (shaded region) until a consensus 5' splice site at which point they diverge.

containing the 1B4 provirus (data not shown). The difference in the size of the wild type and mutant alleles resulted from the inserted U3Neo provirus Finally, as described below, U3Neo transcripts expressed in 1B4 cells are fused to upstream hnRNP A2/B1 sequences.

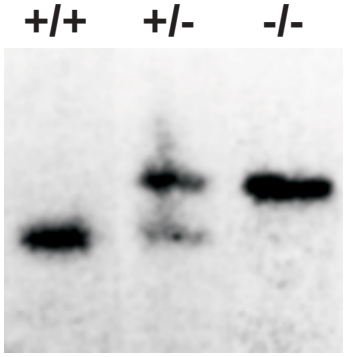
Provirus integration does not disrupt expression of the hnRNP A2/B1 gene

Of the sixteen U3 gene trap proviruses selected in ES cells that we have introduced into the germline, six resulted in obvious phenotypes (typically embryonic death) when bred to homozygosity [10,21,28–32]. In cases where no obvious phenotypes are observed, it is important to determine if the insert did not disrupt gene expression or if the gene is dispensable. Inheritance of the 1B4 provirus followed a Mendelian distribution and no phenotypic changes were observed. Among the 58 offspring analyzed after crossing mice heterozygous for the 1B4 provirus, 13 failed to inherit the provirus, while 32 and 13 were heterozygous and homozygous for the provirus, respectively (Figure 3). A representative Southern blot used to genotype offspring is shown in Figure 3.

To test whether the 1B4 provirus disrupts expression of the hnRNP A2/B1 gene, RNA from wild-type mice and mice homozygous for the 1B4 provirus were analyzed by Northern blot hybridization, using hnRNP A2/B1 cDNA probes derived from sequences upstream and downstream of the integration site. All tissues from wild type mice expressed a single, 1.8 kb transcript, consistent with previous studies [36,37]. Mice homozygous for the 1B4 provirus expressed the 1.8 kb transcript as well as an additional, larger transcript (Figure 4). The size of this larger transcript, approximately 2.3 kb as compared to the migration of 18S and 28S RNAs, is consistent with fusion of upstream hnRNP A2/B1 exons to the Neo gene. Sequences derived from hnRNPA2/B1 sequences downstream of the site of integration also hybridized to the 1.8 kb transcript in both wild type and 1B4 homozygous mice indicating that transcription of full-length hnRNP A2/B1 transcripts is not completely disrupted by the 1B4 provirus.

Mouse embryonic fibroblasts (MEF) were isolated from both wild type and homozygous mutant embryos. MEF isolated from embryos homozygous for the 1B4 provirus showed no obvious differences from those isolated from

A



B

Inheritance			
Total Offspring	**+/+ (%)**	**+/- (%)**	**-/- (%)**
58	**13 (22.4)**	**32 (55.2)**	**13 (22.4)**

Figure 3
Inheritance of the 1B4 provirus. Mice heterozygous for the 1B4 provirus were crossed and their progeny were analyzed by Southern blot analysis following *Eco*R1 digestion and hybridization to the 1B4 flanking sequence. (A) Representative Southern blot hybridization patterns for wild-type (+/+), heterozygous (+/-), and homozygous mutant (-/-) mice. (B) Inheritance of the 1B4 provirus follows a Mendelian distribution as assessed by the distribution of genotypes of progeny from intercrossing 1B4 heterozygotes.

heterozygotes or wild type embryos. RNA isolated from these MEFs was used to quantify the extent of message reduction in 1B4 homozygous cells by Northern blot hybridization. When using probes derived from cDNA sequences downstream of the integration site northern blot analysis revealed a 50% reduction of transcripts in mutants compared to wild type using the glyceraldehyde 3-phosphate dehydrogenase (GAPDH) message as an internal standard (Figure 4).

Splicing and polyadenylation of hnRNP A2/B1-Neo fusion transcripts

Each LTR of the U3Neo provirus contains sequences for 3' processing and polyadenylation. Continued expression of hnRNP A2/B1 transcripts suggests that use of the viral poly(A) sites is less efficient than removal of the intron in which the provirus resides. To determine whether mutation of viral 3' processing signals was responsible for continued hnRNP A2/B1 expression, a 500 base pair region spanning the polyadenylation signal in the 5' LTR was amplified from integrated provirus DNA and sequenced. However, the sequence of the PCR product was identical to the wild-type Moloney murine leukemia virus LTR (data not shown).

The question remained as to how hnRNP A2/B1-neo fusion transcripts are expressed. Previous Northern blot analysis found high levels of a 2.3 Kb fusion transcript in 1B4 cells [29], approximately the size expected for hnRNPA2/neo fusion transcripts terminating in the 5' LTR. One possibility is that fusion transcripts may combine the proximal upstream hnRNP A2/B1 exon, 5' splice site, flanking intron and 5' LTR into a single, terminal exon. However, this possibility contradicts current models of exon definition in which exons in pre-mRNA are first defined by proteins interacting across exons and then processed as relatively autonomous units. Alternatively, the proximal hnRNP A2/B1 exon may maintain its autonomy and splice to a cryptic 3' splice site, located either in Neo or in the adjacent intron.

To distinguish between these alternatives, hnRNP A2/B1/Neo fusion transcripts expressed in MEF and ES cells were analyzed by reverse transcriptase PCR (RT-PCR). A primer complementary to the Neo gene (NeoA) was used to prime first strand cDNA synthesis. Primers complementary to adjacent hnRNP A2/B1 exon sequences (PR1 or PR2) and a neo-specific primer (NeoB) were used to amplify transcripts extending from the hnRNP A2/B1 gene into the provirus (Figure 5A). Transcripts extending through the 5' splice site, proximal intron and into the provirus would produce RT-PCR products of 917 and 598 nucleotides with PR1 and PR2, respectively. As shown in Figure 5B, the size of the major PCR product from each reaction was significantly smaller than expected for transcripts colinear with the flanking DNA. Moreover, the major PCR products did not hybridize to a U3-specific oligo probe (Figure 5C). Three independent RT-PCR products were cloned from separate amplification reactions and sequenced. As shown in Figure 5D, all of these transcripts spliced from the proximal 5' splice site in the hnRNP A2/B1 gene to a cryptic 3' splice site located in the Neo gene (Figure 5D). Characteristic of 3' splice sites, the Neo splice site contained PyAG and a potential branch point

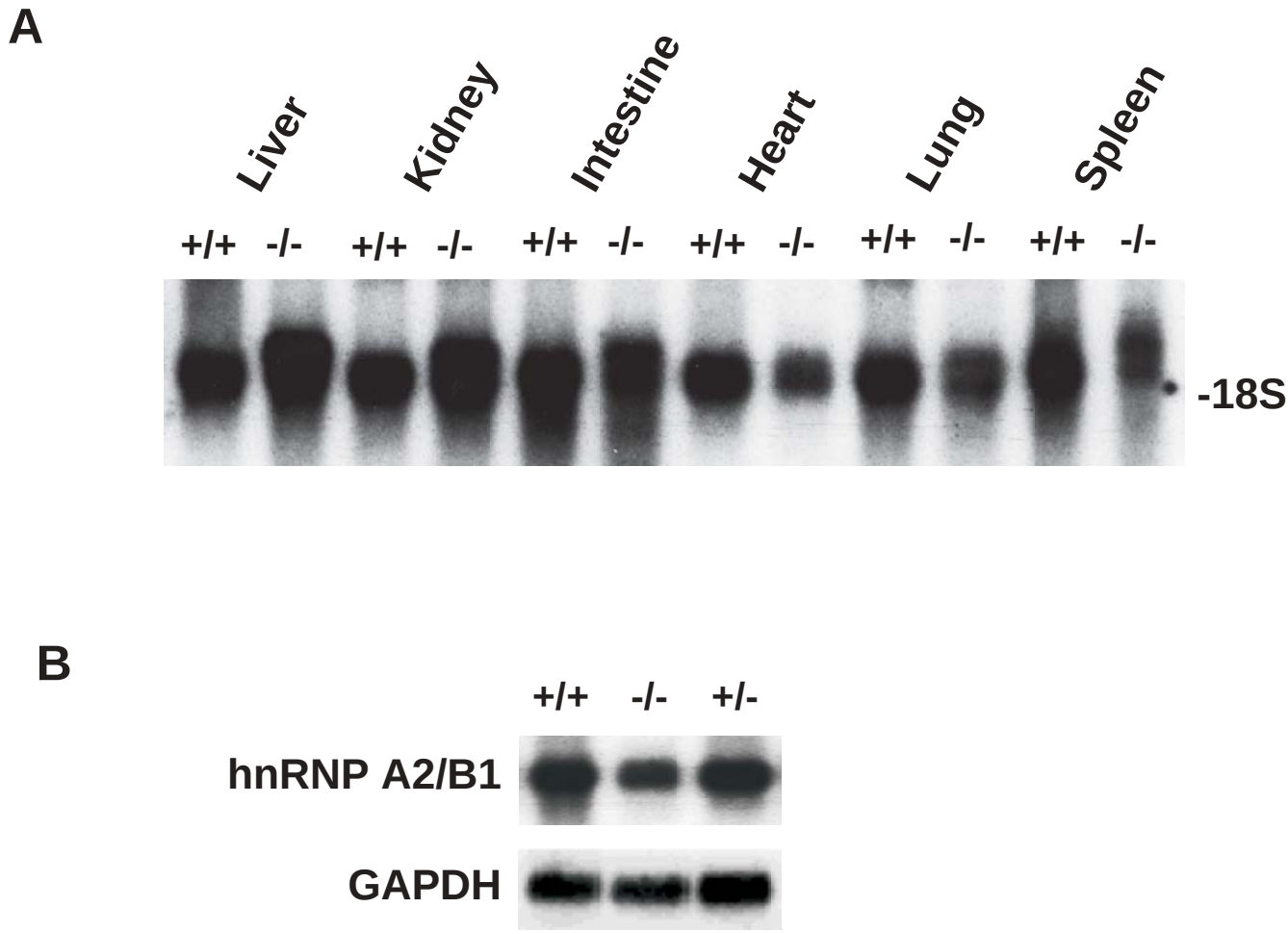


Figure 4

HnRNP A2/B1 expression in tissues and cells from 1B4 homozygous and wild type mice. (A) Analysis of hnRNP A2/B1 expression in wild type (+/+) and homozygous mutant (-/-) mice. Approximately 20 μg of total RNA was fractionated on a 1% formaldehyde agarose gel and hybridized to a ^{32}P labelled probe corresponding to the flanking sequence isolated by inverse PCR. The migration of the 18S ribosomal RNA is indicated. (B) Analysis of hnRNP A2/B1 expression in embryonic fibroblast cell lines derived from wild type (+/+), heterozygous (+/-), and homozygous mutant (-/-) mice. Approximately 20 μg of total RNA was fractionated on a 1% formaldehyde agarose gel and hybridized to a ^{32}P labelled probe corresponding to hnRNP A2/B1 exons located 3' of the site of provirus integration. Levels of GAPDH transcripts were used as an internal control for RNA loading. Homozygous mutant cells displayed a two-fold reduction in hnRNPA2/B1 transcript levels, as determined by phosphoimager analysis.

sequence but lacked a poly-pyrimidine stretch. The 3' splice site is downstream of the initiation codon for neomycin phosphotransferase. Therefore, hnRNPA2/B1-Neo transcripts are expected to encode a fusion protein consisting of the amino terminal 219 amino acids of hnRNP A2/B1 fused to amino acid 10 of the neomycin phosphotransferase (NPT) protein.

The U3 probe also detected several minor RT-PCR products upon prolonged exposure (Figure 5C). We were unable to clone these products due to their low abundance. However, they were smaller than expected for transcripts co-linear with the flanking DNA and may arise from cryptic splice sites in the flanking intron.

Discussion

Several large scale screens of insertion mutations in mouse embryo-derived stem (ES) cells rely on DNA sequence analysis to select mutations for germline transmission [21–24]. The process, designated "tagged sequence

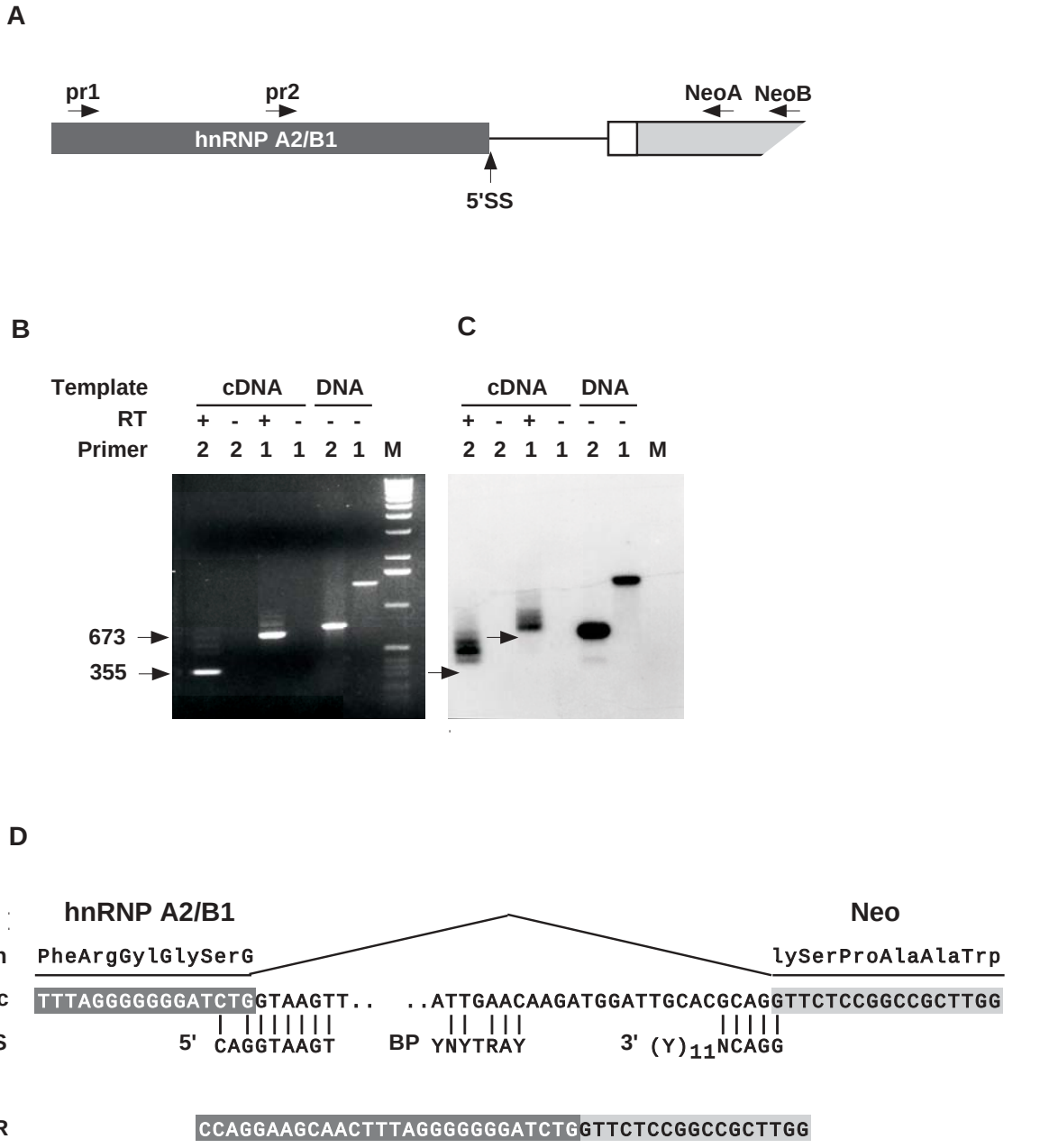


Figure 5

Analysis of hnRNP A2/B1-neo fusion transcripts. (A) Strategy used to amplify fusion transcripts. The provirus is shown downstream of a 5' splice site (5'SS) in the hnRNP A2/B1 gene. The positions of the primers used for reverse transcriptase PCR (RT-PCR) are indicated. (B) Gel analysis of RT-PCR products of. PCR reactions were performed before (-RT) or after (+RT) first strand cDNAsynthesis with the neoA primer. PCR reactions using DNA templates were performed to provide controls for products amplified from unspliced transcripts. PCR products were fractionated on a 1% agarose gel and stained with ethidium bromide. Sizes of the RT-PCR products are indicated. (C) PCR products from B were analyzed by Southern blot using a U3 oligo as a probe. Arrows indicate the mobility of RT-PCR products seen in C. (D) Sequence alignment of RT-PCR products with genomic sequences. 5' splice site (5'), 3' splice site (3'), and branch point (BP) consensus sequences are shown for comparison.

mutagenesis", involves sequencing short segments of DNA isolated from each mutation to identify genes disrupted by the targeting vector. Sequence-based screens are faster and less expensive than phenotype-based screens, and provide centralised collections of characterised mutations available for germline transmission. However, to maximise the utility of tagged sequence mutagenesis, one would like to predict, from the sequence alone, the functional consequences of the inserted targeting vector on cellular gene expression. Toward this end, the present study characterised a mutation generated by insertion of the U3Neo gene trap retrovirus into an intron of the hnRNP A2/B1 gene. Expression of the Neo gene involved splicing of some hnRNP A2/B1 transcripts to a cryptic splice acceptor site located 28 nucleotides downstream of the neomycin phosphotransferase (NTP) initiation codon. Other hnRNP A2/B1 transcripts splice normally, removing the provirus along with other intron sequences. Therefore, expression of the hnRNPA2/B1 gene was only reduced to about half of wild type levels, and the mutation caused no obvious phenotype in mice.

U3 gene trap vectors were designed to disrupt cellular gene function by usurping the promoter of the occupied genes and by ablating transcription downstream of the two poly(A) sites (one in each LTR) carried by the provirus. Since poly(A) sites are not usually recognised when located in introns [38–40], U3 gene traps were expected to select for clones in which the provirus had inserted into exons of transcriptionally active genes. However, in approximately half of the targeted genes we have analysed [21], the provirus has inserted into introns.

The present study identified a mechanism that allows expression of a U3Neo gene from a provirus positioned within an intron. The majority of hnRNP A2/B1-Neo fusion transcripts utilised a cryptic 3' splice site within the NPT coding sequence. This places the 5' proviral poly(A) site at the end of an alternative exon that can be utilised to produce fusion transcripts, or excluded to produce wild type transcripts. Since the initiation codon for NPT lies upstream of the Neo 3' splice site, NTP is expressed as a fusion protein in which the first 219 amino acids of hnRNP A2/B1 are appended to codon 10 of NTP.

Because of these results, we have analyzed the expression of 6 other U3Neo proviruses located in the introns of different genes. Transcripts in all but one of clone splice to the cryptic Neo splice site; while the one exception utilizes a cryptic splice site located in the proximal intron (E. White, G. Hicks, M. Roshon and H. E. Ruley, in preparation). Thus, utilization of the Neo cryptic site appears to provide the predominant mechanism by which the U3Neo gene is expressed following insertion into introns. These results are consistent with an exon definition model in which splicing and polyadenylation require interactions between factors acting across exons [40,41]. This model predicts that polyadenylation signals are not recognised unless they can be defined as part of a 3' terminal exon. Accordingly, poly(A) sites are not efficiently recognised when positioned between 5' and 3' splice sites [38,39], and insertion of a 5' splice site into a 3' terminal exon suppresses polyadenylation [40,42]. Conversely, upstream 3' splice sites can enhance polyadenylation [43–45]. We find that the proximal hnRNP A2/B1 exon upstream of the provirus does not lose its identity; rather, the exon splices either to the Neo splice site or to the next hnRNP A2/B1 exon. Moreover, the poly(A) site in the 5' LTR appears to be used exclusively in conjunction with a cryptic 3' splice site.

Utilization of the Neo 3' splice site is likely to have two important consequences with regard to the use of U3Neo vectors for insertional mutagenesis. First, insertion into an intron may not disrupt cellular gene expression. In the present study, levels of hnRNP A2/B1 transcripts were reduced only about two fold in homozygous mutant cells, and the relative amount of the A2/B1 protein in hnRNP complexes was unaffected (G. Dreyfuss, personal communication). This may explain the absence of an obvious phenotype in mice. Alternatively, other hnRNP proteins compensate for reduced levels of the A2/B1 transcripts, just as cells can tolerate severe reductions in the levels of the related hnRNP A1 protein [46,47].

Second, since the Neo 3' splice site is downstream of the NPT initiation codon, its use may skew the targeting to favor of those genes capable of splicing upstream exons in-frame, to produce enzymatically active fusion proteins. The magnitude of the potential bias is difficult to assess. A variety of amino-terminal fusions maintain enzymatic activity (or produce enzymatically active breakdown products) including those fused to codon 12 of NPT [48–52]. Moreover, selection of resistant cell clones requires only minimal levels of Neo gene expression [53]. Still, genes providing the appropriate introns are expected to provide larger targets for gene trap mutagenesis than genes lacking such introns. This could contribute to the fact that 3 of 400 inserts characterised in an earlier study occurred in the same intron of the L29 gene [21].

The Neo 3' splice site contains a potential branch point sequence, and the sequence (CAGG) across the intron-exon boundary is optimal according to the scanning model of 3' splice site selection [54,55]. The Neo site differs from the typical 3' splice site in that it lacks a polypyrimidine tract; however, this feature is often missing from alternative splice sites [56]. Both the A of the branch point and the intron-terminal AG dinucleotide are considered

invariant; therefore, one may be able to enhance the mutagenic efficiency of U3Neo vectors by altering these nucleotides. Alternatively, the problem may be avoided by using other selectable markers, assuming their sequences lack cryptic splice sites or by using gene trap vectors that rely on splicing to activate the expression of genes carried by the targeting vector. The latter vectors contain strong splice sites, either in front of [2,4] or behind [23] the entrapment cassette, allowing efficient expression from within introns.

Conclusions

hnRNP A2 and hnRNP B1 are members of a large family of RNA binding proteins found associated with mammalian heterogeneous nuclear RNA. The proteins are thought to participate in the processing mRNA precursors [33], and they can influence splice site selection and promote exon skipping in vitro [57–59]. Consistent with a fundamental role in RNA metabolism, the human and mouse hnRNP A2 sequences are highly conserved with only one amino acid difference out of 341 residues. However, since the 1B4 mutation did not ablate hnRNP A2/B1 gene expression, it is unlikely to be useful for studies of hnRNPA2/B1 gene function. While further analysis might uncover phenotypes associated with this hypomorphic mutation, detailed examination of either cells or mice seemed unjustified in the absence of any greater effect on gene expression. Our results illustrate the interplay between polyadenylation and splicing as predicted by an exon definition model. Moreover, the 1B4 mutation reveals a mechanism by which U3 gene trap vectors can be expressed without disrupting cellular gene expression and suggests ways to improve the vectors for gene trap mutagenesis.

Methods

Isolation of cDNA clones encoding the hnRNP A2/B1 protein

DNA sequences (260 nt.) adjacent to the 1B4 provirus were isolated by inverse polymerase chain reaction (PCR) as reported elsewhere [29]. This flanking sequence was used as a probe to genotype mutant mice and cells by Southern blot hybridization and was also used to isolate cDNA clones encoding the murine hnRNP A2/B1 protein from a PCC3 embryonal carcinoma cell cDNA library. 55 hybridizing plaques were identified from a total of 1×10^6 plaques screened. Initial characterization of 10 strongly hybridizing plaques identified two overlapping clones corresponding to the full-length transcript.

Sequencing

cDNA templates were subcloned into the pBluescript KS- plasmid and completely sequenced from both strands. Plasmid DNA was isolated by the boiling lysis method [60], followed by precipitation with polyethylene glycol 8000. 5 μg of plasmid DNA was used in each sequencing reaction [10,61]. Initial sequences were determined by using T3 and T7 primers, and extended by using custom 17–18 nt primers (Gibco-BRL).

PCR amplification of 5' polyadenylation site

The region of the provirus LTR containing the 5' poly(A) site was amplified by PCR from genomic DNA isolated from 1B4 homozygous mice. PCR reactions (10 mM Tris.HCl, pH 8.3, 5 mM KCl, 1.5 mM $MgCl_2$, 200 μM of each deoxyribonucleoside triphosphate, each primer at 2 μM, and 2.5 units of Amplitaq (Perkin-Elmer/Cetus) involved 35 cycles of denaturation (95°C for 1.0 min), primer annealing (55°C for 1.0 min), and primer extension (72°C for 2 min). The only product generated by upstream (5'-CTTCTATCGCCTTCTTGACG) and downstream (5'-ACACAGATAAGTTGCTGGCC) primers was of the predicted size (529 bp). This product was subcloned into the Invitrogen TA cloning vector and sequenced.

Reverse transcriptase PCR

RT-PCR was performed as described [62]. 20 μg RNA was treated with 1 unit of RNAse free DNase (Gibco BRL) (20 mM Tris-HCl pH 8.4, 50 mM KCL 2.5 mM MgCl2) for 15 minutes at room temperature. First strand cDNA synthesis was performed at 42°C for 30 min in a 20 μl reaction containing: 5 μg RNA, 500 nM NEO A primer (5'-ATTGTCTGTTGTGCCCAGTCATA), 20 mM Tris-HCl (pH 8.4), 50 mM KCl 2.5 mM $MgCl_2$, 10 mM DTT, 400 μM each dNTP, and 8 units Super Script II reverse transcriptase (Gibco-BRL). 2 units of RNAse H was added and incubated for 10 min at 55°C. 2 μl of the single strand cDNA was amplified through 35 cycles (95°C for 1.0 min; 55°C for 1.0 min and 72°C for 2 min.) in a 50 μl reaction containing: 2 μM Neo and hnRNP A2/B1 specific primers, 10 mM Tris.HCl, pH (8.3), 5 mM KCl, 1.5 mM $MgCl_2$, 200 μM of each dNTP, and 2.5 units of Amplitaq (Perkin-Elmer/Cetus). Neo B (5'-CGAATAGCCTCTCCACCCAA) was used as the Neo specific primer, and either PR1 (5'-GGAACAGTTCCGAAAGCTC) or PR2 (5'-GAGGAACACCACCTTAG) were used as the hnRNP A2/B1 specific primers.

Authors' contributions

JD introduced the 1B4 mutation into the germline, MR analyzed the 1B4 mutation in mice and cells and characterized the hnRNPB1/A2 cDNA and ER supervised the project.

List of abbreviations

dNTP, deoxyribonucleoside triphosphate; hnRNP heterogeneous nuclear ribonuclear protein; nt, nucleotide; LTR, long terminal repeat; MEF, mouse embryonic fibroblast; NPT, neomycin phosphotransferase; RT PCR, reverse transcriptase polymerase chain reaction.

Acknowledgements

We thank Gideon Dreyfuss for communicating unpublished observations and Geoff Hicks for a critical reading of the manuscript. This work was supported by Public Health Service Grants (R01HG00684, R01GM51201, and R01RR13166 to HER) and by a grant from the Kleberg Foundation. Additional support was provided by Cancer Center (Core) grant P30CA42014. M.R. was supported by a Medical Scientist Training Grant 5T32GM07347.

References

1. Stanford WL, Cohn JB and Cordes SP **Gene-trap mutagenesis: past, present and beyond.** *Nat Rev Genet* 2001, **2:**756-768
2. Friedrich G and Soriano P **Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice.** *Genes Dev* 1991, **5:**1513-1523
3. Stanford WL, Caruana G, Vallis KA, Inamdar M, Hidaka M, Bautch VL and Bernstein A **Expression trapping: identification of novel genes expressed in hematopoietic and endothelial lineages by gene trapping in ES cells.** *Blood* 1998, **92:**4622-4631
4. Skarnes WC, Auerbach BA and Joyner A **A gene trap approach in mouse embryonic stem cells: the lacZ reporter is activated by splicing reflects endogenous gene expression, and is mutagenic in mice.** *Genes Dev* 1992, **6:**903-918
5. Wurst W, Rossant J, Prideaux V, Kownacka M, Joyner A, Hill DP, Guillemot F, Gasca S, Cado D and Auerbach A **A large-scale gene-trap screen for insertional mutations in developmentally regulated genes in mice.** *Genetics* 1995, **139:**889-899
6. Forrester LM, Nagy A, Sam M, Watt A, Stevenson L, Bernstein A, Joyner AL and Wurst W **An induction gene trap screen in embryonic stem cells: Identification of genes that respond to retinoic acid in vitro.** *Proc Natl Acad Sci (USA)* 1996, **93:**1677-1682
7. Russ AP, Friedel C, Ballas K, Kalina U, Zahn D, Strebhardt K and von Melchner H **Identification of genes induced by factor deprivation in hematopoietic cells undergoing apoptosis using gene-trap mutagenesis and site-specific recombination.** *Proc Natl Acad Sci (USA)* 1996, **93:**15279-15284
8. Skarnes WC, Moss JE, Hurtley SM and Beddington RSP **Capturing genes encoding membrane and secreted proteins important for mouse development.** *Proc Natl Acad Sci (USA)* 1995, **92:**6592-6596
9. Mitchell KJ, Pinson KI, Kelly OG, Brennan J, Zupicich J, Scherz P, Leighton PA, Goodrich LV, Lu X and Avery BJ **Functional analysis of secreted and transmembrane proteins critical to mouse development.** *Nat Genet* 2001, **28:**241-249
10. DeGregori JV, Russ A, von Melchner H, Rayburn H, Priyaranjan P, Jenkins N, Copeland N and Ruley HE **A murine homolog of the yeast RNA1 gene is required for post-implantation development.** *Genes Dev* 1994, **8:**265-276
11. Chen Z, Friedrich GA and Soriano P **Transcriptional Enhancer Factor 1 Disruption by a Retroviral Gene Trap Leads to Heart Defects and Embryonic Lethality in Mice.** *Genes Dev* 1994, **8:**2293-2301
12. Deng JM and Behringer RR **An insertional mutation in the BTF3 transcription factor gene leads to an early postimplantation lethality in mice.** *Transgenic Res* 1995, **4:**264-269
13. Gogos JA, Thompson R, Lowry W, Sloane B, Weintraub H and Horwitz M **Gene trapping in differentiating cell lines: regulation of the lysosomal protease cathepsin B in skeletal myoblast growth and fusion.** *J Cell Biol* 1996, **134:**837-847
14. Kerr WG, Heller M and Herzenberg LA **Analysis of lipopolysaccharide-response genes in B-lineage cells demonstrates that they can have differentiation stage-restricted expression and contain SH2 domains.** *Proc Natl Acad Sci (USA)* 1996, **93:**3947-3952
15. Sterner-Kock A, Thorey IS, Koli K, Wempe F, Otte J, Bangsow T, Kuhlmeier K, Kirchner T, Jin S and Keski-Oja J **Disruption of the gene encoding the latent transforming growth factor-beta binding protein 4 (LTBP-4) causes abnormal lung development, cardiomyopathy, and colorectal cancer.** *Genes Dev* 2002, **16:**2264-2273
16. Takeuchi T, Yamazaki Y, Katoh-Fukui Y, Tsuchiya R, Kondo S, Motoyama J and Higashinakagawa T **Gene trap capture of a novel mouse gene, *jumonji*, required for neural tube formation.** *Genes Dev* 1995, **9:**1211-1222
17. Leighton PA, Mitchell KJ, Goodrich LV, Lu X, Pinson K, Scherz P, Skarnes WC and Tessier-Lavigne M **Defining brain wiring patterns and mechanisms through gene trapping in mice.** *Nature* 2001, **410:**174-179
18. Gasca S, Hill DP, Klingensmith J and Rossant J **Characterization of a gene trap insertion into a novel gene, cordon-bleu, expressed in axial structures of the gastrulating mouse embryo.** *Developmental Genetics* 1995, **17:**141-154
19. Episkopou V, Arkell R, Timmons PM, Walsh JJ, Andrew RL and Swan D **Induction of the mammalian node requires Arkadia function in the extraembryonic lineages.** *Nature* 2001, **410:**825-830
20. Chowdhury K, Bonaldo P, Torres M, Stoykova A and Gruss P **Evidence for the stochastic integration of gene trap vectors into the mouse germline.** *Nucleic Acids Res* 1997, **25:**1531-1536
21. Hicks GG, Shi EG, Li XM, Li CH, Pawlak M and Ruley HE **Functional genomics in mice by tagged sequence mutagenesis.** *Nat Genet* 1997, **16:**338-344
22. Holzschu D, Lapierre L, Neubaum D and Mark WH **A molecular strategy designed for the rapid screening of gene traps based on sequence identity and gene expression pattern in adult mice.** *Transgenic Res* 1997, **6:**97-106
23. Zambrowicz B, Friedrich GA, Buxton EC, Lilleberg SL, Person C and Sands AT **Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells.** *Nature* 1998, **392:**608-611
24. Wiles MV, Vauti F, Otte J, Fuchtbauer EM, Ruiz P, Fuchtbauer A, Arnold HH, Lehrach H, Metz T and von Melchner H **Establishment of a gene-trap sequence tag library to generate mutant mice from embryonic stem cells.** *Nat Genet* 2000, **24:**13-14
25. Hicks GG, Shi E-G, Chen J, Roshon M, Williamson D, Scherer C and Ruley HE **Retrovirus Gene Traps.** *Methods Enzymol* 1995, **254:**263-275
26. Chang W, Hubbard C, Friedel C and Ruley HE **Enrichment of insertional mutants following retrovirus gene trap selection.** *Virology* 1993, **193:**737-747
27. Reddy S, Rayburn H, von Melchner H and Ruley HE **Fluorescence-activated sorting of totipotent embryonic stem cells expressing developmentally regulated *lac*Z fusion genes.** *Proc Natl Acad Sci USA* 1992, **89:**6721-6725
28. Scherer CA, Chen J, Nachabeh A, Hopkins N and Ruley HE **Transcriptional specificity of the pluripotent embryonic stem cell.** *Cell Growth Diff* 1996, **7:**1393-1401
29. von Melchner H, DeGregori JV, Rayburn H, Reddy S, Friedel C and Ruley HE **Selective disruption of genes expressed in totipotent embryonal stem cells.** *Genes Dev* 1992, **6:**919-927
30. Chen J, Nachabeh A, Scherer C, Ganju P, Reith A, Bronson R and Ruley HE **Germline inactivation of the murine eck receptor tyrosine kinase by gene trap retroviral insertion.** *Oncogene* 1996, **12:**979-988
31. Pawlak MR, Scherer CA, Chen J, Roshon MJ and Ruley HE **Arginine N-methyltransferase 1 is required for early postimplantation mouse development, but cells deficient in the enzyme are viable.** *Mol Cell Biol* 2000, **20:**4859-4869
32. Williamson DJ, Banik-Maiti S, DeGregori J and Ruley HE **hnRNP C is required for postimplantation mouse development but Is dispensable for cell viability.** *Mol Cell Biol* 2000, **20:**4094-4105
33. Dreyfuss G, Matunis MJ, Pinol-Roma S and Burd CG **hnRNP proteins and the biogenesis of mRNA.** *Ann Rev Biochem* 1993, **62:**289-321
34. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402
35. Burd CG, SM S, Gorlach M and Dreyfuss G **Primary stuctures of the heterogeneous nuclear ribonucleoprotein A2, B1, and C2 proteins: A diversity of RNA binding proteins generated by small peptide inserts.** *Proc Natl Acad Sci (USA)* 1989, **86:**9788-9792
36. Faura M, Renau-Piqueras J, Bachs O and Bosser R **Differential distribution of heterogeneous nuclear ribonucleoproteins in rat tissues.** *Biochem Biophys Res Comm* 1995, **217:**554-560
37. Kozu T, Henrich B and Schafer KP **Structure and expression of the gene (*HNRNPA2B1*) encoding the human hnRNP protein A2/B1.** *Genomics* 1995, **25:**365-371
38. Adami G and Nevins JR **Splice site selection dominates over poly (A) site choice in RNA production from complex adenovirus transcription units.** *EMBO J* 1988, **7:**2017-2116

39. Levitt N, Briggs D, Gil A and Proudfoot NJ **Definition of an efficient synthetic poly (A) site.** *Genes Dev* 1989, **3:**1019-1025
40. Niwa M, MacDonald C and Berget S **Are vertebrate exons scanned during splice-site selection?** *Nature* 1992, **360:**277-280
41. Robberson BL, Cote GJ and Berget SM **Exon definition may facilitate splice site selection in RNAs with multiple exons.** *Mol Cell Biol* 1990, **10:**84-94
42. Furth PA, Choe W-T, Rex JH, Byrne JC and Baker CC **Sequences homologous to 5' splice sites are required for the inhibitory activity of papillomavirus late 3' untranslated regions.** *Mol Cell Biol* 1994, **14:**5278-5289
43. Miller JT and Stoltzfus CM **Two distant upstream regions containing cis-acting signals regulating splicing facilitate 3' end processing of avian sarcoma virus RNA.** *J Virol* 1992, **66:**4242-4251
44. Niwa M, Rose SD and Berget SM **In vitro polyadenylation is stimulated by the presence of an upstream intron.** *Genes Dev* 1990, **4:**1552-1559
45. Wasserman KM and Steitz JA **Association with terminal exons in pre-mRNAs: a new role for the U1 snRNP.** *Genes Dev* 1993, **7:**647-659
46. Ben-David Y, Bani MR, Chabot B, De Koven A and Bernstein A **Retroviral insertions downstream of the heterogeneous nuclear ribonucleoprotein A1 gene in erythroleukemia cells: evidence that A1 is not essential for cell growth.** *Mol Cell Biol* 1992, **12:**4449-4455
47. Yang X, Bani MR, Lu SJ, Rowan S, Ben-David Y and Chabot B **The A1 and A1B proteins of heterogeneous nuclear ribonucleoparticles modulate 5' splice site selection in vivo.** *Proc Natl Acad Sci (USA)* 1994, **91:**6924-6928
48. Chui YL, Lozano F, Jarvis JM, Pannell R and Milstein C **A reporter gene to analyse the hypermutation of immunoglobulin genes.** *J Mol Biol* 1995, **249:**555-563
49. Reiss B, Sprengel R and Schaller H **Protein fusions with the kanamycin resistance gene from transposon Tn5.** *EMBO J* 1984, **3:**3317-3322
50. Schwartz F, Maeda N, Smithies O, Hickey R, Edelmann W, Skoultchi A and Kucherlapati R **A dominant positive and negative selectable gene for use in mammalian cells.** *Proc Natl Acad Sci (USA)* 1991, **88:**10416-10420
51. Sedivy JM and Sharp PA **Positive genetic selection for gene disruption in mammalian cells by homologous recombination.** *Proc Natl Acad Sci (USA)* 1989, **86:**227-231
52. Van den Broeck G, Timko MP, Kausch AP, Cashmore AR, Van Montagu M and Herrera-Estrella L **Targeting of a foreign protein to chloroplasts by fusion to the transit peptide from the small subunit of ribulose 1,5-bisphosphate carboxylase.** *Nature* 1985, **313:**358-363
53. Jeannotte L, Ruiz JC and Robertson EJ **Low level of Hox1.3 expression does not preclude the use of promoterless vectors to generate a targeted gene disruption.** *Mol Cell Biol* 1991, **11:**5578-5585
54. Smith CW, Chu TT and Nadal-Ginard B **Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns.** *Mol Cell Biol* 1993, **13:**4939-4952
55. Smith CW, Porro EB, Patton JG and Nadal-Ginard B **Scanning from an independently specified branch point defines the 3' splice site of mammalian introns.** *Nature* 1989, **342:**243-247
56. Jackson IJ **A reappraisal of non-consensus mRNA splice sites.** *Nucleic Acids Res.* 1991, **19:**3795-3798
57. Mayeda A, Helfman DM and Krainer AR **Modulation of exon skipping and inclusion by heterogeneous nuclear ribonucleoprotein A1 and pre-mRNA splicing factor SF2/ASF.** *Mol Cell Biol* 1993, **13:**2993-3001
58. Mayeda A and Krainer AR **Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2.** *Cell* 1992, **68:**365-375
59. Mayeda A, Munroe SH, Caceres JF and Krainer AR **Function of conserved domains of hnRNP A1 and other hnRNP A/B proteins.** *EMBO Journal* 1994, **13:**5483-5495
60. Maniatis TE, Fritsch F and Sambrook J **Molecular Cloning: A Laboratory Manual** *Cold Spring Harbor: Cold Spring Harbor Laboratory* 1989,
61. Hsiao K **A fast and simple procedure for sequencing double stranded DNA with Sequenase.** *Nucleic Acids Res* 1991, **19:**2787
62. Kawasaki ES **Amplification of RNA.** *In: PCR Protocols: A Guide to Methods and Applications (Edited by: Innis MA, Gelfand DH, Sninsky JJ, White TJ) San Diego: Academic Press, Inc* 1990, 21-27