SUB-CELLULAR LOCALISATION AND *TRANS*- SPLICING OF FUSION TRANSCRIPTS IN MOUSE EMBRYONIC STEM CELLS CONTAINING GENE TRAP INSERTIONS.

JUDITH E. SLEEMAN

Thesis presented for the degree of Doctor of Philosophy

University of Edinburgh

1995

I declare that this work is my own, except where stated otherwise.

Judith E. Sleeman

ACKNOWLEDGEMENTS

The work contained in this thesis was carried out in the laboratories of W. C. Skarnes, A. G. Smith and R. S. P. Beddington at the Centre for Genome Research, University of Edinburgh. I would like to thank Bill Skarnes for his support over the last four years.

In my early days at the CGR, Julie Moss ran the lab., Val Wilson gastrulated regularly in spectacular fashion, Barry Rosen was an encyclopaedia of the useful and the ridiculous and Linda Manson was terrifyingly organised but still completely bonkers. Ian Chambers pulled very strange faces in lab. meetings. Jenny Nichols was Jenny Nichols. Gary Robertson fixed things (usually) and Morag Robertson was the best person to borrow things from. Dougie Colby took lots of unwarranted abuse in tissue culture without throwing heavy objects at me. Kate Cripps provided a stable home environment. Later on, Julie still ran the lab., Andrew Jeske kept Bills' mice from taking over the building, Joanne Broadbent and associates provided evening entertainment, Dave Townley was large and Jane Brennan was indescribable. Alison Cozens fed me coffee and doughnuts when things got too much for me. Dani Klewe-Nebenius was absolutely one hundred percent crazy. Sheila Laird was indispensable for making those overlooked solutions. Morag was still the best person to borrow things from. Ken Lee borrowed from everyone but was useful when computers misbehaved. Right at the very end, Anne Corlett was very tolerant of my rapidly building hysteria, Bill was tireless in his reading and re-reading of chapters and Karl Jamieson dragged me out of the pub to get this thing printed.

I owe a great debt of thanks to the people who have supported me over the last four years, especially my family. without whose constant encouragement I would never have seen this through.

ABSTRACT

The work presented in this thesis combines the gene trap approach and whole mount *in situ* hybridisation to study the sub-cellular localisation of RNAs. All known RNA localisation signals lie within the 3' untranslated regions (3'UTRs) of genes. A modified gene trap vector designed to trap the 3' regions of endogenous genes was used in mouse embryonic stem (ES) cells to assess the range of RNA localisation patterns that can be directed by these regions. Conventional gene trap vectors that form fusion transcripts containing the 5' regions of endogenous genes were also used both in ES cells and in fibroblasts to investigate the possibility that RNA localisation signals also exist in these regions.

Using vectors which contain a splice acceptor site to trap the 5' regions of genes, a number of cell lines showed nuclear localisation of *βgeo* fusion transcript. These nuclear transcripts contained intron sequences from the vector, indicating that they were inefficiently or incorrectly spliced. In two of these cell lines, although most of the *βgeo* transcript remained unspliced, a proportion was processed by an accurate inter-molecular spicing reaction, joining sequences from a variety of endogenous mouse genes to the vector splice acceptor site. Each of these cell lines was shown to contain a single site of integration of the vector, within the 5' external transcribed spacer (5'ETS) of a ribosomal transcription unit.

These integrations are predicted to lead to the synthesis of uncapped pol I transcripts containing the ribosomal 5'ETS followed by the splice acceptor, *βgeo* and polyadenylation signal from the gene trap vector. These pol I transcripts, containing protein coding sequences linked to a splice acceptor site with no upstream splice donor are analogous to the VSG and PARP transcription units that are *trans-* spliced in trypanosomes. This analogy suggests that a mechanism similar to the *trans-* splicing of the trypanosome VSG and PARP genes is being used in these ES cell lines to produce translated products from gene trap insertions into RNA pol I transcription units.

TABLE OF CONTENTS

FIGURES AND TABLES

LIST OF ABBREVIATIONS

ATP	Adenosine Triphosphate
BCIP	5-bromo-4-chloro-3-indolyl phosphate-p-toluidine salt
βgeo	β-galactosidase/ neomycin fusion
bp	Base Pair(s)
BSA	Bovine Serum Albumen
CBP20/80	Cap Binding Protein 20/80
CTP	Cytosine Triphosphate
DAPI	4',6-Diamidine-2-Phenylindole Dihydrochloride
dATP	Deoxyadenosine Triphosphate
dCTP	Deoxycytosine Triphosphate
DEPC	Diethyl Pyrocarbonate
DFC	Dense Fibrillar Component
dGTP	Deoxyguanosine Triphosphate
DIA/LIF	Differentiation Inhibiting Activity/ Leukaemia Inhibitory Factor
DIG	Digoxigenin
DMSO	Dimethly Sulphoxide
dNTP	Deoxynucleotide Triphosphate
DTT	Dithiothrietol
dTTP	Deoxythymidine Triphosphate
EDTA	Ethylenediaminetetra-acetic acid
EGTA	1,2-Di(2-aminoethoxy) ethane-N,N,N',N'-tetra-acetic acid
En-2	Engrailed 2
ES Cell	Embryonic Stem Cell
5'ETS	5' External Transcribed Spacer
FC	Fibrillar Centre

FITC	Flourescine isothiocyanate
GC	Granular Component
GMEM	Glasgow Modified Eagle's Medium
GTP	Guanosine Triphosphate
HEPES	N-2-Hydroxyethylpeperazine-N'-2-ethane-sulphonic acid
hnRNA	Heterogeneous Nuclear RNA
hnRNP	Heterogeneous Nuclear Ribonucleoprotein
IMS	Industrial Methylated Spirit
ITS	Internal Transcribed Spacer
LacZ	β-galactosidase
MAP-2	Microtubule Associated Protein 2
MBP	Myelin Basic Protein
MMG	Monomethylguanosine
MOPS	3-(N-Morpholino)propanesulphonic Acid
mRNA	Messenger RNA
NBT	Nitroblue Tetrazolium Chloride
neo	Neomycin
NLS	Nuclear Localisation Signal
NOR	Nucleolar Organising Region
NTP	Nucleotide Triphosphate
NPC	Nuclear Pore Complex
NuMA	Nuclear Mitotic Apparatus Associated Protein
PARP	Procyclic Acid Repetitive Protein
PCR	Polymerase Chain Reaction
PEG	Polyethylene Glycol
PGK	Phosphoglycerate Kinase
PIPES	Piperazine-N,N'-bis[2-etane sulphonic acid]
pol I/II/III	RNA Polymerase I/II/III

PVP	Polyvinyl Pyrrolidone
5'RACE	5' Rapid Amplification of cDNA Ends
RIPA	Radio-immunoprecipitation Assay
RNP	Ribonucleoprotein
rRNA	Ribosomal RNA
SDS	Sodium Dodecyl Sulphate
SL RNA	Spliced Leader RNA
snRNP	Small Nuclear Ribonucleoprotein
SSC	Standard Sodium Citrate
TBE	Tris Borate EDTA buffer
TEMED	N,N,N',N'-tetramethyl-ethylene diamine
TMG	Trimethylguanosine
tRNA	Transfer RNA
U2AF	U2 Accessory Factor
UTP	Uridine Triphosphate
UTR	Untranslated Region
VSG	Variant Surface Glycoprotein
X-gal	5-bromo-4-chloro-indolyl-β-galactopyroanoside

SUMMARY OF IMPORTANT CELL LINES

LINES	PARENTAL LINE	VECTOR	CHARACTERISTICS
TbP20,8/ TbP20,29	10T1/2 fibroblasts	pGT1.8βgeo	Inefficient splicing and nuclear localisation of fusion transcript, probably due to insertions into exons of pol II genes.
ST416	CGR8 ES Cells	pGT1.8tm	Inefficient splicing and nuclear localisation of fusion transcript, probably due to an insertion into the exon of a pol II gene.
ST576/ ST478	CGR8 ES Cells	pGT1.8tm	Nuclear localisation of fusion transcript associated with *trans*- splicing of the introduced splice acceptor site to a number of endogenous splice donors.
ST567	CGR8 ES Cells	pGT1.8tm	Insertion into the *embigin* gene.
ST402	CGR8 ES Cells	pGT1.8tm	Inefficient splicing of fusion transcript. Probable second insertion into the *embigin* gene within an exon.

Chapter 1

INTRODUCTION

Following its initial transcription, eukaryotic RNA undergoes a number of processing events before the execution of its final function. Depending on the type of RNA, these include cleavage, splicing, association with accessory molecules and export from the nucleus. In order to carry out their final function, many mature RNA species must also reach their correct final destination within the cell. In the case of snRNPs this involves import back into the nucleus. For a number of protein coding mRNAs, the localisation to particular regions of the cytoplasm is important.

There is a growing body of evidence that spatial aspects of these processing events are more tightly controlled than was previously believed, particularly within the nucleus. Cellular RNAs can be divided into three major classes according to the enzyme by which they are transcribed. RNA pol I is responsible for the transcription of ribosomal RNA, pol II for the transcription of mRNA and snRNAs and pol III for the transcription of tRNA, 5S ribosomal RNA and U6 snRNA. In this introduction, I will concentrate on aspects of pol I and pol II transcription and processing.

Transcription and Processing of RNA Species

Pol I Transcription and Processing

RNA polymerase I is used exclusively for the formation of ribosomes by the transcription of rDNA. Eukaryotic cells contain many copies of the rDNA transcription unit (150-200 in mammals). In most species, these transcription units are arranged in several tandem arrays, with regions

coding for the primary transcript separated by non-transcribed spacer regions. rDNA transcription is regulated in accordance with cellular growth rate, being down-regulated when cells approach stationary phase or are starved of an essential nutrient. The synthesis of mature ribosomes from the primary rRNA transcript involves the removal of the transcribed spacer regions of the transcript to release the 18S, 5.8S and 28S rRNAs and their association with 5S RNA (transcribed by RNA polymerase III) and approximately 85 ribosomal proteins (transcribed by pol II) (reviewed in Sollner-Webb and Mougy, 1991). In contrast to the pre-messenger RNA transcripts synthesised by RNA polymerase II, nascent rRNA transcripts are not modified by the addition of a methylated cap structure to their 5' ends

Mature rRNAs are liberated from the 45S primary transcript by a series of cleavages. There is a preferential order to these cleavages (reviewed in Sollner-Webb and Tower, 1986). In most organisms, the first cleavage event occurs within the 5' external transcribed spacer (5'ETS). In mouse, this primary processing event occurs at residue 650, at the 5' border of a 200bp sequence conserved among mammals. The 5'ETS is liberated as an unstable 24S species, concomitant with the appearance of the 41S pre-rRNA (Perry 1976). Experiments using gel retardation and UV cross linking have revealed that this first event occurs in a large complex (Kass and Sollner-Webb, 1990) containing at least six polypeptides, ranging from 52 to 250 kilodaltons in size. The small nucleolar ribonucleoprotein snRNP U3 has also been demonstrated to bind to this 5'ETS processing complex in vitro (Kass et al, 1990). It has been proposed that these processing complexes are the structures previously visualised at the electron microscopic level (Mougy et al, 1993) (Miller and Beatty, 1969)

and described as 'terminal balls'. These structures are seen at the ends of elongating rRNA molecules furthest from the rDNA. The terminal balls have been demonstrated to be associated with U3 snRNP (Maser and Calvert, 1989) and to contain the antigen fibrillarin (Scheer and Benavente, 1990). Fibrillarin is one of the major protein components of the interphase nucleolus, and is found associated with the nucleolar snRNPs U3, U8, U13 and U14 (reviewed in Fournier and Maxwell, 1993).

The snRNP U3 is the most widely studied of the nucleolar snRNPs. In vitro experiments using mouse cell extracts (Kass et al, 1990) demonstrated that U3 is required for the first cleavage event within the 5'ETS. Injection of anti-sense oligonucleotides into Xenopus oocytes in order to form a DNA/RNA hybrid with U3, causing its destruction by endogenous RN'ase, led to the disruption of a later cleavage event at the boundary between the first internal transcribed spacer (ITS1) and the 5.8S region (Savino and Gerbi, 1990). U3 snRNP is thus predicted to be of fundamental importance in both early and late cleavage steps.

Physical Location of rRNA Transcription and Processing.

The activity of ribosomal genes leads to the formation of the nucleolus, a highly dynamic nuclear structure which breaks down during cell division, reforming immediately after mitosis around the nucleolar organising regions (NORs) which consist of the tandemly repeated rRNA genes. P-element mediated insertion of single rRNA genes into various chromosomal sites in Drosophila led to the generation of extra nucleoli (Karpen et al 1988) confirming that the NORs are the ribosomal genes themselves. Active transcription of the genes has been shown to be necessary for nucleolus assembly (Benavente et al, 1987) leading to the

suggestion that nascent transcripts act as the nucleation sites for nucleolus assembly. This is also supported by the observation that the 5' regions of the growing pre-rRNA transcripts can interact with fibrillarin, one of the major protein components of the nucleolus (Scheer and Benavente, 1990).

The nucleolus is the most prominent feature of the interphase nucleus, and is easily seen using phase contrast microscopy. At the EM level, the nucleolus can be resolved as a membrane free organelle comprising one or more fibrillar centres (FCs) usually 0.2-0.4μ in diameter, each bounded by a dense fibrillar component (DFC), 0.05-0.1μ thick, surrounded by the granular component (GC) that makes up the bulk of the nucleolus (Jordan, 1984). Ribosomal DNA and RNA pol I have been demonstrated to be present in the fibrillar centre, leading to the conclusion that the FC is the site of rDNA transcription (Scheer and Benavente, 1990). The dense fibrillar component contains the abundant nucleolar proteins fibrillarin and nucleolin, in addition to many of the nucleolar snRNPs and is predicted to be the site of early processing events. The granular component contains pre ribosomal particles, presumably undergoing the final stages of processing (Warner, 1990).

A detailed study of the physical location of transcription and the early stages of processing of rRNA within the nucleoli of HeLa cells has been carried out by Puvion-Dutilleul et al (1991). Their results suggest that rRNA genes which are actively transcribed are located at the borders of the FCs, with elongating transcripts extending into the adjacent DFC. The absence of signal in more peripheral areas of the nucleolus using the 5'ETS probe suggests that the first cleavage event occurs in the DFC, immediately adjacent to the sites of transcription. The rDNA detected

within the FC in this and previous studies (Scheer and Benavente, 1990) is not associated with rRNA transcripts, and is predicted to represent a transient storage site for inactive genes.

In summary, the nucleolus can be envisaged as a site of both transcription and processing of the pol I transcribed rRNA genes, comprising the genes in close spatial association with the molecules required for the synthesis of their product, the ribosome. At the core of the nucleolus (the fibrillar centre) lie the transcription units. The nascent transcripts are generated at the boundary between the fibrillar centre and the surrounding dense fibrillar component, where the initial steps of processing are thought to occur. Further processing events, including the association of the pol I transcribed rRNAs with the pol II transcribed ribosomal proteins and the pol III transcribed 5S RNA, are thought to occur within the DFC and the surrounding granular component. Thus, pol I transcripts demonstrate a strong association between temporal and spatial aspects of their synthesis and processing.

Pol II Transcription and Processing

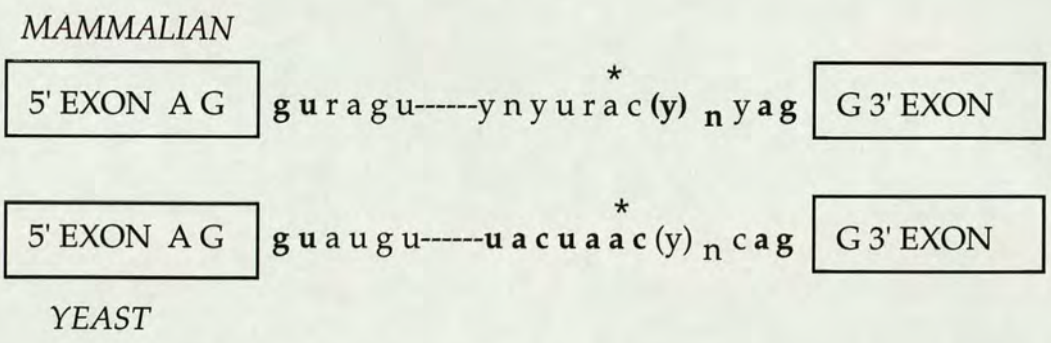
RNA polymerase II is responsible for the transcription of heterogeneous nuclear RNA (hnRNA), the majority of which is processed into messenger RNA (mRNA). These RNA molecules are covalently modified at their 5' ends by the addition of a methylated G residue (the cap) and at their 3' ends by cleavage and the addition of a polyA tail. The cleavage and polyadenylation is signaled by an AAUAAA sequence, 10 to 30 base pairs upstream of the cleavage site. hnRNA molecules contain long stretches of RNA which are not present in the mature mRNA. These sequences are referred to as introns (Catterall et al, 1978) and are removed

from the precursor molecule by splicing, to leave the functional exons together with 5' and 3' untranslated regions in the mature transcript. In higher eukaryotes, *cis*-splicing links exons from a single pre-mRNA. In a number of lower eukaryotes, inter-molecular splicing, or *trans*- splicing, has also been documented.

Cis-Splicing

The joining of exons from the same molecule through a 3'-5' phosphodiester bond is achieved by two sequential trans-esterification reactions. In the first reaction, the phosphodiester bond at the 5' splice site (splice donor) is attacked by the 2'-OH of an adenosine residue at the branch point in the intron. This produces a free upstream exon, leaving the intron region, with its 5' end covalently linked to the branch point, attached to the downstream exon. In the second reaction, the splice donor site is covalently linked to the 3' splice site (splice acceptor) of the downstream exon. The intronic region is released as a circular 'lariat' structure, and subsequently degraded. The splicing reaction takes place in a spliceosome, a large (60S) complex (Brody and Abelson, 1985; Frendewey and Keller, 1985) comprising the pre-mRNA itself, together with the snRNP's U1, U2, U4, U5 and U6, and a large number of accessory proteins.

Figure 1.1: Consensus Sequences for Splicing in Mammalian and Yeast pre-mRNAs.



Upper case denotes exon sequence
Lower case denotes intron sequence
r is any purine, y is any pyrimidine
The adenosine residue at the branch point is marked with an asterisk
The most highly conserved regions are in bold type

Assembly of the *Cis* -Spliceosome

In order to position the splice sites in the catalytic centre of the spliceosome in the correct orientation and order, the assembly of the catalytic snRNPs and accessory proteins into the spliceosome must be tightly controlled. The assembly of the spliceosome has been studied in yeast and in mammalian cell extracts. Spliceosome formation in lower and higher eukaryotes shows a general similarity. In each case, the assembly of the spliceosome can be broken down into several stages.

The earliest functional intermediate in yeast spliceosome assembly is referred to as the "commitment complex", in which U1 snRNP is bound to the 5' splice site by sequence homology and also contacts the branch point sequence, either directly or via another factor (Seraphin and Rosbash, 1989). This complex is ATP-independent and is thought to be the functional homologue of the E complex (early complex) of mammalian spliceosome assembly (Michaud and Reed, 1991), which is also ATP

independent and contains the non-snRNP splicing factor U2AF and several spliceosome associated proteins (SAPs) in addition to U1 snRNP (Michaud and Reed, 1993).

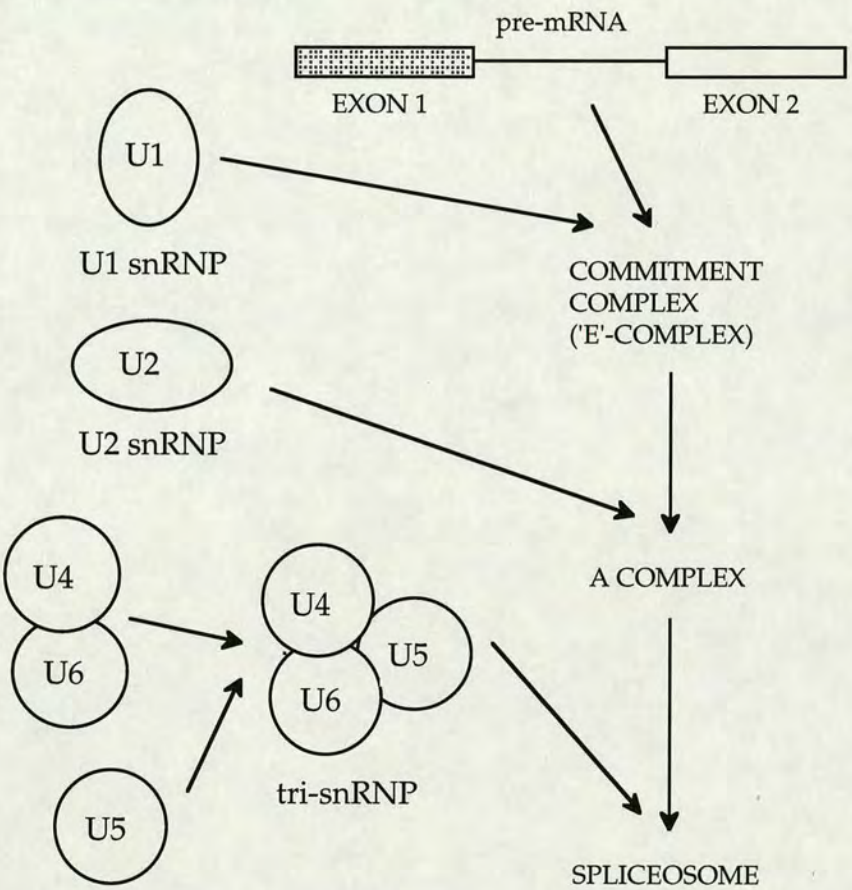
The next snRNP to bind to the forming spliceosome is U2 which interacts simultaneously with the branch point and 3' splice site. There is limited sequence homology between U2 and the branch point. Accessory proteins such as U2AF in mammalian cells (Green, 1991) bind near the 3' splice site and direct the U2 snRNP towards the branch site.

The complete spliceosome is then formed by the interaction of the pre-spliceosome with a pre-formed U4-U5-U6 tri-snRNP particle (Newman, 1994). Once the spliceosome has been assembled, conformational changes occur, particularly within U6 (Wolff and Bindereif, 1993), resulting in interactions between U6 and U2 at two different domains. U6 also forms a sequence specific interaction with the 5' splice site, juxtaposing the guanosine residue of the 5' splice site and the adenosine residue of the branch point within the catalytic site of the spliceosome in order for the first trans-esterification reaction to proceed. A uridine rich loop within U5 snRNP interacts first with the upstream exon, then with the downstream exon by non Watson- Crick base pairing, ensuring that first the 5' and then the 3' splice site are correctly aligned (Newman and Norman, 1992; Teigelkamp et al, 1995).

The later stages of spliceosome assembly are ATP-dependant, largely because of the requirement for conformational changes within the snRNPs and the stabilisation of weak base pair interactions. These

functions have been associated with a large number of accessory proteins in both yeast and mammalian cells.

Figure 1.2: Assembly of the Major Components of the Spliceosome



***Trans* -Splicing**

Trans-splicing is a less usual form of splicing, first characterised in the Trypanosomatidae, a family of parasitic protozoa. *Trans* -splicing involves the joining of sequences from different mRNA precursor molecules through an intermolecular reaction that is biochemically and mechanistically very similar to *cis*-splicing. The consensus splice sites used for *trans*-splicing are equivalent to those used for *cis*-splicing in mammalian cells. Since its initial characterisation in trypanosomes, *trans*-

splicing has been documented in a growing number of eukaryotic organisms including parasitic and free living nematodes, the trematode *Schistosoma mansoni* and *Euglena gracilis*. Computer searches using a canonical *trans*-splicing structure, designed by reference to known *trans*-splicing structures, have predicted that *trans*- splicing may occur in a wide range of organisms, including vertebrates (Dandekar and Sibbald,1990).

The majority of work on *trans*- splicing has concentrated on Trypanosomes, using the permeabilised cell system developed by Ullu et al (1990) and the nematodes *C. elegans* (in vivo studies) and *A. lumbricoides* (cell free in vitro studies). These organisms employ "spliced-leader" type *trans*- splicing, whereby transcripts receive an identical 5' mini-exon. There are, however, significant differences between the *trans*-splicing reactions of trypanosomes and those of nematodes.

No *cis*- splicing has been characterised in trypanosomes. mRNAs are processed exclusively by *trans*- splicing. A single spliced leader molecule is involved, which donates a 39 nucleotide sequence to each mRNA. In contrast to this, nematodes such as *C. elegans* exhibit both *cis*- and *trans*-splicing with an estimated 10-15% of genes being *trans*- spliced, sometimes with a single gene containing both *cis*- and *trans*- spliced exons (reviewed by Blumenthal, 1995). Thus the trypanosome system provides a model in which the minimum requirements for *trans*- splicing can be studied, while nematodes provide systems for the investigation of the functional similarities and co-operation between *cis*- and *trans*- splicing.

Involvement of snRNPs in *Trans*- Splicing

Trypanosomes contain homologues of several of the snRNPs involved in *cis*- splicing. U2 snRNP, which binds to the branch point, and the heterodimer of U4 and U6 snRNPs have been characterised. U2 has a domain structure that is largely conserved between *cis*- and *trans*- splicing, but some aspects of its assembly differ in trypanosomes, and some *trans*- splicing specific RNA-protein interactions have also been identified (Gunzl et al, 1992 and 1993). Of the trypanosome snRNPs, U6 shows the most conservation to the *cis*- spliceosomal snRNP. A base-pairing interaction has been identified between a sequence near the 3' end of U6 and the spliced leader RNA (Hannon et al, 1992). This interaction may be responsible for bringing the two splicing substrates into proximity in the *trans*- spliceosome.

A 72 nucleotide RNA, termed spliced leader associated RNA (SLA RNA) has recently been identified in *Trypanasoma brucei* (Watkins et al, 1994). This RNA is unlike any known small RNA except for the presence of a small domain, CUUUUA, which resembles the domain GCCUUUUAC from U5 which is involved in interactions with exon sequences near the splice sites in the *cis*- spliceosome. The interaction between SLA RNA and the 5' splice site of the spliced leader RNA has been demonstrated by psoralen cross linking and suggests that SLA RNA is the trypanosome homologue of U5 snRNP.

No homologue of U1 has been identified in trypanosomes. However, the Spliced leader RNA itself participates in *trans*- splicing as part of a snRNP. The structure of trypanosome SL RNA is analogous to an exon fused to a snRNP. Intramolecular base pairing within the SL RNA has been

demonstrated to involve the 5' splice site in a manner similar to the intermolecular binding of U1 to the 5' splice site in *cis-* splicing. Furthermore, the snRNP-like domain of the SL RNA can be used to deliver heterologous exons to the 3' splice site through *trans-* splicing (Maroney et al, 1991). It seems, therefore, that the 'intron' region of the trypanosome SL RNA functions as a snRNP, with some homology to U1, delivering 5' exons to the *trans-* spliceosome. Mutagenesis studies (Yu et al, 1993) have demonstrated that U6 snRNP can be induced to behave in a similar manner as SL RNA, donating a fragment that can then proceed through the next step of splicing to form the 5' exon, suggesting an evolutionary relationship between the spliced leader RNA and the snRNPs.

The Significance of *Trans-* Splicing

Unlike trypanosomes, nematodes posses all of the components required for conventional *cis-* splicing and indeed appear to process the majority of their transcripts by *cis-* splicing. There has been a great deal of debate about the significance of *trans-* splicing, particularly in nematodes where it occurs in parallel with *cis-* splicing. The basic requirement for an exon to undergo *trans-* splicing in *C. elegans* is that its upstream 'intron' does not contain a splice donor site (Conrad et al, 1993). This intron-like structure containing a splice acceptor site and branch point, but no splice donor has been termed an 'outron'. The introduction of a splice donor site between 50 and 250 nucleotides upstream of the *trans-* splice acceptor of the gene *rol-6* is sufficient to convert it into a *cis-* spliced gene (Conrad et al, 1993). Until recently, there has been no clear rationale for the processing of certain transcripts by *trans-* splicing in these organisms.

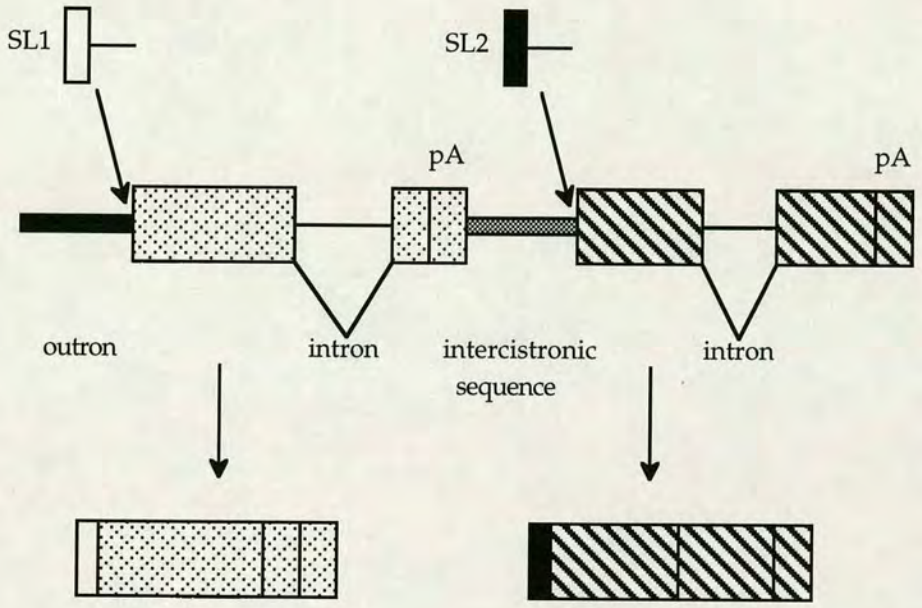
Trans- splicing in *C.elegans* produces a sub-set of mRNAs which contain an unusual tri-methyl guanosine (TMG) cap, along with a unique 22 base pair non-translated sequence. It had been suggested that this subset of RNAs may have a specialised sub-cellular localisation, an unusual stability, or a unique mode of translation (Liou and Blumenthal, 1990). However, there is no firm evidence for any of these suggestions, although the extreme inter-specific sequence conservation between spliced leader mini-exons suggest that some important function may be attributable the this region of the mRNA. Mature RNAs arising from *trans*-splicing do not seem to be processed differently from *cis-* spliced messages. In fact, the first report of *trans-* splicing in *C-elegans* involved the actin genes (Krause and Hirsh, 1987), three of which are tightly clustered on chromosome V and processed by *trans-* splicing, with the fourth being located on the X chromosome and *cis-* spliced. There is no evidence that the mature transcripts from the different genes have distinct functions.

More recently, poly-cistronic pre-mRNAs have been described in *C-elegans* (Speith et al, 1993). During the splicing of these poly-cistronic messages, some genes receive a 22 base pair mini exon from an alternative spliced leader molecule, SL2. The choice of spliced leader sequence appears to be dependent solely on the chromosomal arrangement of the genes. Just as genes whose mRNAs are *trans-* spliced do not share any obvious functional attributes, the sub-set that receive SL2 do not have anything in common except for their position within the polycistronic transcription unit. Coding regions which are between 96 and 294 base pairs downstream of another gene in the same orientation receive the SL2 sequence, while all other genes receive SL1. It is, therefore,

possible that the major and perhaps only similarity between genes which are *trans-* spliced is their genomic structure.

An RNA transcribed as part of a polycistronic precursor is not expected to receive a 7-methyl guanosine cap structure in the normal manner. All cellular mRNAs acquire a cap structure immediately after transcription by mono-methylation of the N-7 nitrogen of the terminal guanosine and, in some cases, the next one or two adjacent riboses. The cap structure contributes to the stability of the message and enhances its translation. The acquisition of a first exon by *trans-* splicing to a spliced leader molecule possessing a cap (albeit a TMG cap, rather than the conventional MMG) would provide a mechanism for polycistronic messages to be capped and subsequently efficiently exported from the nucleus and translated. Mature *trans-* spliced mRNAs in *C.elegans* have been demonstrated to retain this TMG cap (Liou and Blumenthal, 1990; Van Doren and Hirsch, 1990). The hypothesis of *trans-* splicing as a way of capping otherwise incapable transcripts is in agreement with the presence of polycistronic transcription units in both *C.elegans* and trypanosomes. In trypanosomes, the genes coding for the variable surface glycoprotein (VSG) and pro-cyclic acid repetitive protein (PARP) genes are transcribed by an α- amanitin resistant polymerase, most probably pol I. These transcripts are proposed to receive a cap structure by *trans-* splicing to the pol II transcribed spliced leader RNA (Zomerdijk et al, 1991; Sherman et al, 1991).

Figure 1.3: Polycistronic Transcription and *Trans*- Splicing in *C. elegans*.



Splice Site Selection in *Cis*- and *Trans* -Splicing

The accurate juxtaposition of the splice donor with its adjacent splice acceptor is of vital importance for the accuracy of pre-mRNA splicing. The mechanisms employed for splice site selection are not fully understood. Many vertebrate genes are made up of numerous small exons, separated by much larger introns. Splice sites in vertebrate cells are, however, considerably less well conserved than their counterparts in yeast, whose genes usually contain a single intron, or no intron at all (Guthrie. 1991; Ruby and Abelson, 1991). Thus, in vertebrate cells, a highly complex series of splicing events is carried out using very little sequence information. Models for the accurate recognition of splice sites rely on the observation that splice sites are recognised in pairs, rather than as individual sequence motifs. Interactions between the 5' and 3' splice sites have been detected

in the earliest pre-spliceosome complexes in both yeast and mammalian cells (Legrain et al, 1988; Seraphin & Roshbash, 1989; Michaud & Reed, 1993). In yeast, interactions at the 5' splice site are essential for the formation of interactions at the 3' splice site, whereas in mammalian cells, the presence of the 5' splice site is important for the efficiency and stability of these interactions but is not an absolute requirement. Analyses of complexes formed on pre-RNA substrates containing only a 3' splice site, or only a 5' splice site have demonstrated that forms of the E complex (early complex) , designated E3' and E5' can form independently on mutant substrates (Michaud and Reed, 1993). In E3', the splicing factor U2AF is enriched, by comparison to the normal E complex. In E5', U1 snRNP is enriched. The interaction, either direct or indirect, of U2AF with U1 snRNP is thought to be the mediator of the functional association of the splice sites in the normal E complex.

Early models for the recognition of splice site pairs invoked a scanning mechanism for the correct alignment of exon ends (Lang and Spritz, 1983; Lewin, 1980). Initial splice site recognition occurs at the 5' splice site, which is recognised in conjunction with the branch point. A scanning mechanism then selects the nearest downstream AG conforming to the 3' splice site consensus. Recent modifications to this model include an element of competition between AGs by way of explanation of the presence of cryptic 3' splice sites within mammalian introns which can be activated by the mutation of the genuine downstream 3' splice acceptor (Smith et al, 1993). The data used as a basis for this model are from in vitro studies using precursors with short, or artificially shortened intron sequences (214bp in the study by Smith et al). Although this model, involving the pairing of splice sites at either end of an intron, works well

for these artificial substrates, or for yeast RNAs containing single short introns, scanning through natural vertebrate introns which may be up to 100 Kb in length and contain many cryptic splice sites seems less plausible. If the assumption is made that splice site recognition in *cis*- and *trans*-splicing occurs by a similar mechanism, then this model also poses a problem when considering *trans*-splice events, where the two splice site to be joined occur on separate molecules. Scanning from the 5' splice site in these genes would not locate the 3' splice site.

A more recent model for splice site selection involves the recognition of splice site pairs in the orientation in which they are found in an exon, i.e.. a 3' splice donor site upstream of a 5' splice acceptor (Robberson et al, 1990). In this model, splice site recognition is seen as two separate stages. The initial interactions of U1 and U2 snRNPs with the splice sites is referred to as "exon definition". Following exon definition, neighbouring exons are aligned with each other by exon juxtaposition. This presumably involves interactions between the factors bound to individual exons. The addition of U4, U5 and U6 to the splicing complex is predicted to occur during exon juxtaposition (Berget, 1995). The first and last exons of a gene would require special mechanisms for their recognition. The first exon may be recognised as a 5' splice site with an upstream cap structure. The 7-methyl guanosine cap structure has been demonstrated to be essential for the in vitro splicing of pre-mRNAs containing a single intron (Izaurralde et al, 1994). Factors recognising 3' splice sites may interact with factors recognising poly (A) sites in order to define the last exon of a gene.

In general, it seems likely that, in pre-mRNAs with small introns, the initial pairing of splice sites occurs across the intron (Guthrie 1991; Ruby

and Abelson 1991; Rosbash and Seraphin, 1991), whereas in genes with small exons and large introns, the initial pairing occurs across the exon. This would explain why similar splice site mutations have different effects in genes with different exon structures (Berget, 1995). The model becomes more complex in transcripts containing a mixture of exons and introns of different lengths, possibly with some combination of the two proposed methods of splice site selection being employed. Genes which are *trans*-spliced would be expected also to use the exon definition model of splice site selection.

Spatial Aspects of Pol II Transcript Processing.

Although considerably more is known about the molecular events of processing of pol II transcripts than those of pol I transcripts, the spatial aspects are not as well understood. Recent data (Huang et al, 1994) have begun to correlate sub-nuclear structures visible using electron microscopy and the distribution of factors involved in pre-mRNA splicing. The use of conventional uranyl acetate/lead citrate staining of nuclei for electron microscopy reveals little of the sub-structure, largely due to the masking of such structures by a large quantity of soluble protein (Hendzel and Bazett-Jones, 1995). Nucleoplasmic depletion can be carried out using a non-ionic detergent prior to electron microscopy. This allows the resolution of three major extra-nucleolar structures: perichromatin fibrils, heterochromatin and inter chromatin granules.

A number of studies have documented intra-nuclear speckles associated with the transcription and processing of mRNA. The labeling of nascent transcripts with 5-bromouridine 5'-triphosphate and its subsequent detection using antibodies has revealed the accumulation of nascent

transcript in numerous speckles within the nucleus. Jackson et al (1993) reported 300 to 400 speckles per nucleus in HeLa cells, while Wansink et al reported around 100 speckles per nucleus in T24 human bladder carcinoma cells. Poly A+ RNA has been demonstrated to concentrate primarily within 20 - 50 discrete transcript domains in primary fibroblasts and myoblasts (Carter et al, 1991). Autoimmune antibodies have been used to investigate the sub-nuclear localisation of snRNPs of the splicing complex (Spector, 1990). The snRNPs studied formed 20-50 domains, with small amounts also detectable in the surrounding nucleoplasm. The non-snRNP splicing factor SC-35 is also concentrated in these nuclear domains (Fu and Maniatis, 1990; Carmo-Fonseca et al, 1991; Spector et al, 1991; Huang and Spector, 1992 and Carter et al, 1993). The patterns seen tend to be complex, with domains of varying sizes being seen, along with some specific staining outside the intensely staining areas.

There is some debate as to whether the domains containing high concentration of transcript and those containing high concentration of splicing factors represent the same nuclear structure. Wansink et al detected no correlation between the domains of RNA transcription seen in T24 cells and accumulations of the splicing factor SC-35 seen in the same cells. Jackson et al, however, reported that most of the transcript domains seen in HeLa cells also contain a component of the splicing apparatus detected by an anti-Sm antibody. Poly A+ RNA-rich domains in fibroblasts have also been demonstrated to be co-incident with snRNP and SC-35 rich regions (Carter et al 1991; 1993). The majority of evidence available suggests that nuclear RNA and splicing factors do show at least some degree of co-localisation within the nucleus. It has been proposed

that both transcription and RNA processing take place in or near these transcript domains (Carter et al, 1993).

A comparison of *in situ* hybridisation data to observations made by electron microscopy (Fakan, 1994) suggest that the speckles of splicing factors observed using immunoflourescence are analogous to clusters of perichromatin granules seen using the electron microscope. Perichromatin fibrils, first seen by W. Bernhard (1969), are found in close proximity to these perichromatin granules and are thought to be a direct visualisation of nascent transcript. It seems likely, therefore, that the sites of transcription and processing of transcripts are closely linked, rather than strictly co-incident. A recent study combining electron microscopy with *in situ* hybridisation has detected poly A+ RNA in association with both the peri-nuclear fibrils and the inter chromatin granules (Huang et al, 1994). The poly A+ RNA associated with the inter chromatin granules remains in the nucleus on treatment of the cell with inhibitors of transcription, leading to the suggestion that it may be a stable population of RNA with an important role to play in nuclear function, rather than nascent transcript.

These studies of the spatial organisation of pol II transcription and processing can be summarised by viewing each transcription domain as a cluster of several active gene loci (Hendzel and Bazett-Jones, 1995) around which the protein and RNA factors necessary for processing of the nascent transcripts accumulate. This model bears a striking resemblance to the organisation of RNA pol I transcription and processing within the nucleolus, albeit on a much smaller scale. As yet, the structural basis for the spatial organisation of pol II transcription and processing is not

understood. However, a protein found in the nuclear matrix and mitotic apparatus (nuclear-mitotic apparatus protein, NuMA) has recently been demonstrated to co-localise with splicing factors in interphase nuclei and to be associated with snRNPs in HeLa cell extracts (Zeng et al, 1994). This protein may have some role in the tethering of pol II transcripts during processing.

Nuclear Export of RNA Species

The traffic of macromolecules between the nucleus and cytoplasm occurs via the nuclear pore complex (NPC). The NPC has been extensively studied by electron microscopy and current models describe it as an hourglass-like structure inserted into the nuclear membrane (Maquat, 1991). The function of the NPC is in the selective transport of protein and RNA molecules into and out of the nucleus. The number of NPCs present on the nuclear membrane of a cell ranges from 100 to 5 X 10^7 (Maquat, 1991). Accumulations of particles around the cytoplasmic face of nuclear pores have been demonstrated (Unwin and Milligan, 1982). These are detectable by electron microscopy and appear, by their size and shape, to be ribosomes. One of the major functions of the NPC is in the export of mature transcripts from the nucleus. This process is energy dependent and selective, with different classes of RNA being exported using different molecular pathways (Reviewed in Zapp, 1992).

The 18S, 5.8S and 28S ribosomal RNAs transcribed by pol I are thought to become functional ribosomes during their transportation out of the nucleus (Zapp,1992). The transport of these species has not been widely studied, although it is thought that interactions between the small ribosomal sub-unit containing the 18S rRNA and the large sub-unit,

containing the 5.8S and 28S rRNAs interact in eukaryotic cells to aid export from the nucleus (Khanna-Gupta and Ware, 1989).

Pol II transcripts can be divided into three functional groups: mRNA, histone mRNA and snRNAs. All of these RNA species receive a mono methyl guanosine cap structure immediately after transcription. mRNAs (excluding histone mRNAs) are also polyadenylated. Histone mRNAs terminate in a highly conserved stem loop structure, formed by endonucleocytic cleavage between two conserved sequence motifs.

The presence of a monomethylguanosine (m7GpppN) cap structure is believed to play a role in the export of pol II transcribed RNAs, as micro-injection of large amounts of free competitor m7GpppG leads to the retention of pol II transcripts in the nucleus (Hamm and Mattaj, 1990). The cap is essential for the export of snRNAs, and increases the efficiency of mRNA export. mRNA species with altered cap structures can be exported from the nucleus, but with a greatly decreased efficiency. It is proposed that other, as yet undiscovered, factors are involved in mRNA export. Histone mRNA export shows different characteristics again, with the presence of a mature 3' end being involved in the transport of these species into the cytoplasm (Eckner et al 1991).

Studies of splicing mutants in yeast (Legrain and Rosbash, 1989) have implicated the presence of intron sequences in pre-mRNAs as a strong signal for their retention within the nucleus. The 5' splice site and the branch point are thought to be of particular importance in this nuclear retention. It is proposed that the interaction of these sequences with elements of the splicing machinery is responsible for their function as

nuclear retention signals. Thus, an mRNA needs to contain a 5' cap and lack intron sequences in order to be efficiently transported across the nuclear membrane.

Different mechanisms are, therefore, used for the export of different classes of RNA. Even within the group of RNAs transcribed by RNA pol II, three distinct mechanisms operate in the export of the three different classes of transcript. Post-transcriptional processing events at the 3' and 5' ends of the RNAs and within the body of the RNAs are necessary for their efficient export, with correctly transcribed but partially or incorrectly processed transcripts being retained within the nucleus.

Sub-Cellular Localisation of Mature pol II Transcripts.

The localisation of certain mature pol II transcripts to distinct regions of the cell has been demonstrated to be necessary for their final function. The localisation of the snRNP splicing factors and its significance for their function has already been discussed. Three systems have been used for the study of sub-cellular localisation of mature mRNA species: the early *Xenopus* embryo, the early *Drosophila* embryo and differentiated chicken somatic cells.

One of the earliest demonstrations of localised mRNAs was made in *Xenopus* oocytes (Rebagliati et al, 1985). Several maternal mRNAs were demonstrated by *in situ* hybridisation to show preferential localisation to either the animal or the vegetal pole. A great deal of subsequent work has been carried out on one of these mRNAs, *vg-1*, which encodes a transforming growth factor homologue (Weeks and Melton, 1987). This mRNA shows a homogeneous distribution in early oocytes, but becomes

restricted to the vegetal cortex in middle and late stage oocytes. The micro-injection of truncated *vg-1* transcripts lacking the 5' regions of the message have demonstrated that the localisation is independent of protein synthesis and that the sequences responsible for the localisation lie within the 3'UTR of the transcript (Yisreali and Melton, 1988). The localisation signal has been mapped to a 340 nucleotide sequence within the 3'UTR (Mowry and Melton). The mechanism of *vg-1* transport and anchorage is thought to involve cytoskeletal or cytoskeletal-associated proteins (Yisreali et al, 1990).

The localisation of maternal transcripts is also important in the early *Drosophila* embryo. A gradient of the Bicoid protein is necessary in the oocyte for the correct development of the head and thorax. The establishment of this gradient is achieved by the localisation of the *bicoid (bcd)* mRNA. *Cis* acting RNA sequences have been demonstrated to be responsible for the anterior localisation of *bcd* RNA (MacDonald and Struhl, 1988). The localisation pathway of *bicoid* mRNA involves a number of tightly controlled steps, involving a number of genes including *expurantia, swallow* and *staufen,* whose transcripts are themselves localised to the anterior pole of the embryo (St Johnston et al, 1989; Stephenson et al, 1988).

Localised maternal mRNAs are also important in the posterior development of the *Drosophila* embryo. Posterior localisation of *nanos* mRNA determines abdominal segmentation (Wang and Lehmann, 1991) while *oskar* mRNA is involved in the specification of the pole plasm (germ line) (Ephrussi et al, 1991). Again, a number of different genes have been shown to affect the localisaton of these mRNAs. Indeed, *oskar* is

required for the correct localisation of *nanos* mRNA (Ephrussi et al, 1991), indicating that the pathways of localisation of messages involved in the patterning of the *Drosophila* embryo are closely interlinked.

Later in *Drosophila* development, at the early blastoderm stage, the pair-rule genes are expressed. The mRNAs for these genes are localised to the apical periplasm (i.e. below the nucleus). The mechanism of this localization has been studied by the expression of hybrid transcripts between the pair-rule genes *fushi-tarazu, hairy* and *even-skipped* and the *β–galactosidase* reporter (Davis and Ish-Horowicz, 1991). These experiments located the signals governing the polarised distribution of the messages to short regions of sequence within the 3'UTRs.

The early gradients of morphogens established in *Xenopus* and *Drosophila* embryos appear over a time span of hours or days, reflecting the stability of these messages, and the complexity of the pathways by which these gradients are achieved (Kislauskis and Singer, 1992). By contrast, the messages of the later pair rule genes are highly unstable (Edgar et al, 1991) so their localisation is likely to be achieved by a different mechanism, probably by leaving the nucleus in a polarised fashion (Kislauskis and Singer, 1992). In both cases, however, the developmentally regulated localisation of specific transcripts is mediated by signals within the 3'UTRs of the messages themselves.

The localisation of specific mRNAs in the cytoplasm of somatic cells has also been studied by *in situ* hybridisation. The earliest work on somatic cells was a study of the mRNAs for cytoskeletal components in chick embryonic fibroblasts and myoblasts (Lawrence and Singer, 1986). Actin

mRNA was shown to be localised to the leading edge of motile cells. Vimentin and tubulin were more peri-nuclear. Further studies on actin mRNA have shown that the message localises to regions of the cell where the corresponding protein is required. In cells undergoing a response to injury of a confluent monolayer, actin mRNA is present in the lamellipodia of spreading cells (Hoock et al, 1991).

Differential localisation of specific mRNAs has also been documented in a number of other differentiated cell types. In intestinal epithelium, actin mRNA localises to the apical end of the cell, where actin filaments polymerize to form the microvilli (Cheng and Bjerknes, 1989). In neurons, the mRNA for microtubule associated protein 2 (MAP-2) is found in the dendrites, while the transcripts for Gap-43 and α-tubulin are restricted to the cell bodies (Garner et al, 1988; Bruckenstein et al, 1990). Myosin mRNA has been shown to accumulate near sarcomeres in muscle (Pomeroy et al, 1991).

In each of the documented cases of sub-cellular mRNA localisation, the message appears to localise in regions of the cell where the product of the message is required. The use of protein synthesis inhibitors has demonstrated that, at least for actin mRNA, the transcript localisation occurs independently of protein synthesis (Sundell and Singer, 1990). The sequences responsible for actin mRNA localisation have been mapped to the 3'UTR of the transcript (Kislauskis et al, 1993). The use of drugs to disrupt the cytoskeleton suggest that microfilaments of the cytoskeleton are required for the correct localisation of actin mRNA (Sundell and Singer, 1991). Actin filaments are thought to be involved in the anchoring of transcripts once their localisation has been achieved (Singer, 1992). A

study of the sub-cellular localisation and transport of myelin basic protein (MBP) mRNA in living oligodendrocytes (Ainger et al, 1993) demonstrated the accumulation of the mRNAs for MBP, actin and globin as cytoplasmic granules. These granules were relatively uniform in size (about 0.3m) and, in the case of MBP mRNA, were demonstrated to be motile and associated with the cytoskeletal matrix. Similar granules were also detected in neuroblastoma cells, suggesting that they are not specific to oligodendrocytes. Granular sub-cellular localisation has also been reported for Vimentin mRNA in myoblast, myotubes and fibroblasts (Cripe et al, 1993). The observation of these granules has been explained as a visualisation of the structures involved in the localisation of transcripts (Ainger et al, 1993). These results have lead to the proposal of a model involving the formation of ribonucleoprotein particles (RNPs) to account for the sub-cellular localisation of certain transcripts (Wilhelm and Vale, 1993). It appears, therefore, that, in somatic cells, certain mRNAs show localisation to the specific regions of the cytoplasm where their protein products are required. This localisation appears to be achieved, at least for some of the transcripts, by the transport of the RNA as RNP particles in association with the cytoskeleton.

Although the majority of steady state messenger RNA species are difficult to detect within the nucleus due to their rapid export into the cytoplasm, the transcript of the *Xist* gene is restricted to the nucleus, as determined by sub-cellular fractionation in mouse cells (Brockdorff et al, 1992) and *in-situ* hybridisation in human cells (Brown et al 1992). The *Xist*- transcript is, however, highly unusual. The transcripts are very large (17Kb in human and 15Kb in mouse) and contain no substantial open reading frames. It has been suggested that the *Xist* transcripts function as RNAs,

rather than by being translated (Rastan, 1994). In all female cells, one copy of the X-chromosome is found in an inactive state. The transcript from the *Xist* gene in humans has been demonstrated by *in situ* hybridisation to be associated with the barr body, which represents the inactive X chromosome, so it is thought to play a role in the inactivation of the X chromosome. The *Xist* gene may be unique among pol II transcribed genes in having a nuclear localised transcript, but it may represent the first identified member of a new class of pol II transcripts.

The Development of The Gene Trap Technique To Create and Analyze Developmentally Important Mutations in the Mouse.

The gene trap technique was developed to facilitate large scale screens for genes involved in the control of murine embryonic development. Traditionally, such studies have proved more difficult in the mouse than in invertebrate organisms such as *Caenorhabditis elegans* and *Drosophila melanogaster*. This results from the much larger size of the mammalian genome, together with the inaccessibility of the embryo within the uterus of the mouse and the expense of large breeding programs.

Experiments on mice containing retrovirus or transgene insertions have demonstrated that the introduction of exogenous DNA into the mouse germ line has the potential to generate mutations (reviewed in Jaenisch, 1988). In these cases, the insertion of the exogenous DNA disrupts the structure of an endogenous gene, producing the mutant phenotype. For example, insertion of a retrovirus into the germ line of mice, leading to an insertion into the a1(I) collagen gene, gave rise to an embryonic lethal mutation (Schnieke et al, 1983). In mice homozygous for this mutation,

developmental arrest co-incided with high expression of the a1(I) collagen gene in normal embryos. In another study, exposure of mouse embryos to a recombinant retrovirus resulted in the insertional mutagenesis of a ubiquitously expressed gene whose mutant phenotype suggested an important role in renal function (Weiher et al, 1990). The gene *limb deformity* was also identified following its mutation by the insertion of a transgene (Woychick et al, 1990). However, in general, the endogenous transcription units associated with such insertional mutations have proved difficult to clone.

In addition to the ability of exogenous DNA to affect endogenous chromosomal DNA by mutation, it has been known for some time that chromosomal DNA can affect the activity of inserted DNA. Casadaban and Cohen (1979) used the introduction of exogenous DNA to identify transcriptionally active regions of the bacterial genome. Constructs were introduced which required the acquisition of certain *cis*- acting DNA sequences to activate the expression of a reporter gene. The genes associated with these elements could then be cloned from DNA sequences flanking the site of insertion. The effect of chromosomal position on the expression of exogenous DNA has also been observed in the mouse. For example, sub-strains of mice containing insertions of the Moloney leukaemia virus at different chromosomal positions activated the virus at different times, depending on the position of the viral DNA within the genome (Jaenisch et al, 1981).

Modifications of the technique used in bacteria to identify regions of transcriptional activity have been used in eukaryotic systems. In *Drosophila*, the transposable P-element system has been widely used to

introduce exogenous DNA into the genome. An 'enhancer trap' vector driven by the weak P-element promoter was used to screen for chromosomal elements acting at a distance to stimulate expression (O'Kane and Gehring, 1987). This vector contained an in-frame translational fusion of *lacZ* to the second exon of the P-element; both ends of the P-element necessary in cis for transposition and the *rosy* gene as a marker with which to select for transformants. The vector was injected into *Drosophila* embryos together with a helper P-element that produces transposase but is itself incapable of transposing. Multiple insertions occurred in many of the injected embryos. Forty nine different strains were bred from the transformants. Thirty seven of these lines showed spatially regulated expression of *lacZ*, demonstrating the potential for this type of approach in identifying elements within the *Drosophila* genome involved in the regulated expression of endogenous genes and creating cell type specific markers to facilitate further genetic analyses. A larger scale screen was carried out using a similar vector (Bellen et al, 1989) in which >500 strains were analyzed. Again, a large number of these strains (65%) showed restricted patterns of *lacZ* activity. Additionally, cytological mapping of 68 of these insertions revealed that 6 of the insertions mapped to the positions of genes that have a similar pattern of expression to that seen with the *lacZ* marker. This confirmed that the genetic elements responsible for the restricted expression from the P-element promoter are also involved in the restricted expression of nearby endogenous genes. A screen of 3768 transformants by Bier et al (1989) used a vector containing the mini-*white* gene as a marker for transformants, forming a smaller, more efficiently transposable vector. These experiments also noted a correlation between the patterns of *lacZ* activity seen and the expression patterns of nearby genes. In addition, 480

embryonic lethal mutations were documented. Classifiable developmental defects were observed in 68 of these lines (2% of the total number of strains analyzed) demonstrating that enhancer trap vectors can be mutagenic. The vector used in this study also contained bacterial plasmid sequences (ampicillin resistance and the origin of replication) to allow cloning of DNA flanking the transposon insertions in both the 5' and 3' directions.

The potential for the use of enhancer trap insertions to identify genes important for development in the mouse was demonstrated using a transgene with the weak herpes simplex virus thymidine kinase promoter linked to the *lacZ* reporter gene (Allen et al, 1988). Transgenic foetuses displayed unique *lacZ* staining patterns which were faithfully transmitted to the next generation, indicating that the pattern of transgene activation was dependent on its chromosomal position. In studies using transgene integration, the transgene serves as a genetic marker to facilitate the analysis of genetic loci associated with different patterns of reporter activation. The use of similar *lacZ* containing vectors in embryonic stem (ES) cells allows pre-screening for insertions of particular interest to be carried out in vitro and in chimaeric animals, without the need to make transgenic mouse lines (Gossler et al, 1989). These enhancer trap vectors also contain the bacterial neomycin resistance gene for selection of ES cells containing vector integrations. Staining of neomycin resistant ES cell clones for lacZ protein activity enables the selection of potentially interesting insertions, for example those giving expression in a sub-set of differentiated cells. Selected ES cell clones can then be used to generate chimaeric and transgenic mice. Enhancer trap vectors need only to insert near to an endogenous

transcription unit for the reporter gene to be activated, so they are not necessarily mutagenic, and cloning of the endogenous sequence elements associated with their activation is not straightforward.

Modifications of the enhancer trap vectors were designed to increase their mutagenic capacity and to allow for easier cloning of endogenous transcription units. The gene trap approach (Gossler et al, 1989; Friedrich and Soriano, 1991) uses a vector with no promoter and places a splice acceptor site 5' to the start of the *lacZ* gene. Integration of the gene trap construct into the introns of genes in the correct orientation should create a spliced *lacZ* fusion transcript and a functional fusion protein if the reading frame is maintained. Insertion of the promoter trap or gene trap vectors in this manner should lead to mutation of endogenous genes. In the early promoter trap and gene trap vectors, the neomycin selectable marker was present under the control of a separate promoter. Modifications of these types of vector allowed for direct selection of ES cell clones expressing the *lacZ* reporter construct by using a novel reporter gene, β-geo, which is a *lacZ/ neomycin* fusion gene (Friedrich and Soriano, 1991). Promoter trap vectors (von Melchner et al, 1990; Macleod et al, 1991, and Reddy et al, 1991) contain *lacZ* genes which lack their own promoter. Thus, the vectors need to insert within an endogenous transcription unit to obtain reporter expression.

The development of enhancer trap, promoter trap and gene trap vectors in mice was based on the assumption that the expression pattern of the reporter gene reflects the expression of the endogenous gene into which the vector has integrated. Experiments by Skarnes et al (1992) used the gene trap vector, pGT4.5 (Gossler et al, 1989) to confirm that activation of

the *lacZ* reporter gene reflected endogenous gene expression. 5' RACE (rapid amplification of cDNA ends) cloning was used to demonstrate the correct use of the introduced splice acceptor site and to generate probes for the further analysis of the gene trap integrations. The expression patterns of the endogenous genes associated with two gene trap insertions were demonstrated to be similar to the expression patterns of the *lacZ* reporter gene. In each case, negligible amounts of normal endogenous transcript were detected, confirming that gene trap insertions are mutagenic. Transmission of ES cells carrying these mutations into the germline produced developmental abnormalities in progeny homozygous for the gene trap mutation.

Recent modification of the gene trap approach (Skarnes et al, 1995) has also demonstrated the potential of this technique preferentially to mutate and clone genes of particular classes. βgeo protein experimentally fused to the signal sequence from the rat *CD4* gene was demonstrated by immunoflourescence to accumulate within the lumen of the endoplasmic reticulum (ER). This protein, however, did not have any β-gal enzyme function. β-gal activity was restored by including the trans-membrane domain of *CD4* downstream of the signal sequence. This activity showed a distinctive localisation pattern, giving X-gal staining in the ER and multiple cytoplasmic inclusions. On the basis of these data, the *CD4* trans-membrane domain was included in the vector pGT1.8tm upstream of βgeo. Using this 'secretory trap' vector, β-gal enzyme activity should be obtained only as a result of insertions downstream of a signal sequence. Thus, the secretory trap vector pGT1.8tm is expected to select for integrations into genes coding for secreted and type I membrane proteins.

Experimental Design

The gene trap vectors used in this study were developed from those of Gossler et at, 1989. The basic design of the vectors relies on a reporter gene which lacks a promoter and translational initiation codon, but is placed immediately downstream of intron sequences including a splice acceptor site (3' splice site) and branch point consensus. Insertion of the vector into the intron of an endogenous gene produces a primary fusion transcript in which a splice donor site from the endogenous gene at the site of insertion is positioned upstream of the splice acceptor site in the vector. Splicing of this transcript leads to a fusion mRNA containing 5' regions of the endogenous gene and the reporter construct. Provided that the reading frame of the endogenous gene is the same as that of the reporter gene, this transcript is translated into a functional fusion protein, using the initiation ATG from the endogenous gene. The formation of the fusion transcript disrupts the function of the endogenous gene, leading in some cases to a mutant phenotype. The interrupted gene responsible for the phenotype is then accessible to cloning using a 5' RACE (rapid amplification of cDNA ends) strategy.

A variety of different gene trap vectors were used. The vector pGT1.8K (Skarnes, unpublished) uses *lacZ* as the reporter gene, and has the selectable marker, *neomycin*, present on a separate promoter. pGT1.8βgeo (Skarnes et al, 1995) contains the *βgeo* gene, which is a fusion of *lacZ* and *neomycin*. This produces a single fusion protein that both confers neomycin resistance on the cells for the selection procedure and has β-galactosidase activity to allow screening for positive clones. The secretory trap vector pGT1.8tm is a modification of the pGT1.8βgeo vector, containing a trans membrane domain from the rat *CD4* gene. As

described above, this vector was designed to trap genes coding for secreted molecules (Skarnes et al, 1995).

The ability of gene trap vectors to produce fusion transcripts containing known sequences together with sequences from endogenous genes has been exploited to study the sub-cellular localisation of RNA species. The recent improvements in whole mount *in situ* hybridisation techniques allow the distribution of fusion transcripts to be monitored at sub-cellular resolution using non-radioactive (digoxigenin) anti-sense riboprobes. Large scale screens were undertaken using gene trap vectors and a vector modified to trap the 3' regions of genes, in order to assess the contribution of different regions of the endogenous transcript to the control of RNA localisation

Results obtained from these initial screens led to the analysis of the processing of fusion transcripts in a number of these lines. These analyses revealed insertions of the vectors into non-pol II transcription units and also, probably, into the exons of pol II transcripts. These unexpected integration events have shed light on the range of gene trap insertions that can produce viable neomycin resistant colonies, which will be useful in the analysis of data generated using gene trap technology in the future. The processing of fusion transcripts in two lines electroporated with the secretory trap vector (pGT1.8tm) has also demonstrated that mouse embryonic stem cells are able to accurately *trans* -splice consensus mammalian 5' and 3' splice sites. This observation has implications both for the study of the normal processing of endogenous genes, and for the artificial manipulation of gene expression in mammalian cells.

Chapter 2

MATERIALS AND METHODS

2.1 Tissue Culture

All tissue culture manipulations were undertaken inside a laminar flow sterile hood (ICN, Flow). All objects were washed with 70% industrial methylated spirits (IMS, BDH) before being placed in the hood. Cells were maintained in tissue culture grade plastic flasks and culture dishes (Corning). The plastic was first gelatinised by the application of a 0.1% (w/v) solution of gelatin (Sigma) in UHP water (purified by passage through an Elga-Prima reverse osmosis unit, Elgastat) for 10 minutes. The cells were grown in a humidified incubator at 37°C, in a 6% CO_2 atmosphere. For fibroblast culture, GMEM was supplemented with 10% foetal calf serum. For embryonic stem cell culture, 100u/ml of DIA/LIF was also added. Medium in the flasks was replaced when it began to turn from orange to yellow, usually every 2 days. Confluent cultures were passaged into fresh flasks at 1/10 their original density. The medium was removed from the flasks. the cells were then rinsed twice with phosphate buffered saline (PBS Dulbecco 'A', Unipath Ltd) pre-warmed to 37°C, then overlaid with trypsin and incubated at 37°C until the cells lifted from the surface of the flask. The flasks were tapped to remove the cells from the base. The trypsin was neutralised by the addition of 5ml of fresh medium and a single cell suspension formed by drawing the medium several times though a glass pipette. The cells were pelleted by centrifugation at 1200rpm for 5 minutes, the medium removed, and the pellet resuspended in 5ml of fresh culture

medium. 0.5ml of cell suspension was then added to pre-warmed fresh culture medium, and placed in a new gelatinised flask.

Freezing Cells

Cells were removed from the growth surface by trypsin digestion and resuspended in culture medium to which 1/10 volume of DMSO (AnalaR, BDH) was then added. The cell suspension was then centrifuged at 1200rpm for 5 minutes, the medium removed, and the pellet resuspended in freezing medium (culture medium suplemented with 10% DMSO). 0.5ml of freezing medium was used for each 5 X 10^6 cells. The suspension was then placed into cryotubes (Nunc) in 0.5ml aliquots and frozen at -80°C. After 24hrs, the vials were transferred to a liquid nitrogen cell bank (XLC110, Minnesota Valley Engineering Cryogenics).

Thawing cells

Vials of frozen cells were removed from the liquid nitrogen storage bank and thawed rapidly. The cell suspension was then transferred to a 15ml centrifuge tube (Corning) containing 4.5ml of culture medium using a sterile pasteur pipette. The cells were pelleted by centrifugation at 1200rpm for 5 minutes, resuspended in 10ml of fresh culture medium and transferred to a gelatinised 25cm^2 flask. The medium was changed when the cells were firmly attached to the growth surface (usually after about 7 hours). The cells were subsequently maintained as above.

Electroporation of Cells

150μg of the plasmid for electroporation was digested with the appropriate restriction enzyme overnight at 37°C, precipitated with 100% ethanol (Hayman Limited) and resuspended in 100μl of sterile PBS.

10^8 cells were removed from culture dishes by trypsin digestion and resuspended in 0.8ml of sterile PBS. The cells were mixed with the DNA and placed in a 0.4cm electroporation cuvette (Gene Pulser). Fibroblasts were electroporated using conditions of 250V, 250μF. Embryonic stem cells were electroporated using 800V, 3μF. The cells were then plated in 10cm diameter culture dishes. One plate was also seeded with an equal number of non-electroporated cells, to serve as a control for the selection procedure. After 24hrs, the medium was changed and G418 was added (400μg/ml for fibroblasts, 200μg/ml for ES cells). The medium was then changed every 2 days until selection was complete ie. until all of the cells on the control plate had been killed. For large scale screening of transcript localisation in ES cells, the plates of cells were fixed at this stage with 4% (w/v) paraformaldehyde in PBS. For the isolation of clonal cell lines, individual colonies were picked from the plates using a mouth pipette, and these were cultured in 24 well plastic dishes until ready for further processing.

Once confluent, each clone was split into wells on a number of separate 24 well dishes. One of these was cultured further, while the others were used to screen for fusion RNA and/or fusion protein localisation. Clones required for further study were then expanded into 6 well dishes, then 25cm^2 flasks, and frozen down for later use.

Replica Plating of ES Cells (Hill and Wurst, 1993)

Replica filters produced by the following method are suitable for fixation with 4% (w/v) paraformaldehye and hybridisation using the whole mount *in situ* technique (2.12).

1) Electroporate the cells and grow under selection until distinct colonies have formed.

2) Aspirate the medium from the plates and carefully place autoclaved, marked polyester filters (1mm pore size) on top of the colonies.

3) Cover 80% of the filter with sterile 3mm diameter glass beads and add 10ml of ES culture medium containing G418.

4) After 2 days of incubation, aspirate the medium, pour off the glass beads, mark the orientation of the filter on the back of the plate and peel off the filter with sterile forceps.

5) Rinse the filter with PBS prior to fixation with 4% (w/v) paraformaldehyde. Add fresh medium to the 'master' plate, and return to the incubator for further culture.

Once the whole mount *in situ* procedure has been carried out on the filter using a *lacZ* riboprobe, colonies which show high expression can be picked and cultured further.

Cell Culture Medium

UHP Water	170ml
10 x GMEM (Gibco)	20ml
Sodium Bicarbonate (7.5% stock, Gibco)	6.6ml
Non-essential Amino Acids (100 x, Gibco)	2ml
Glutamine, 200mM + Pyruvate, 100mM Gibco)	4ml

β-mercaptoethanol (Sigma)	200µl
Foetal Calf Serum (Gibco)	20ml
DIA/ G418	as required

Trypsin

0.25% (w/v) trypsin (from 2.5% solution in normal saline, Gibco)

1mM EDTA (AnalaR, BDH)

1% (v/v) Chick Serum (Gibco)

Store at -20°C.

2.2 General Cloning Techniques

Restriction Digestion of DNA

Restriction enzymes and buffers were supplied by Boehringer Mannheim. The conditions recommended by the manufacturers were generally followed. For most applications, and incubation time of 1-2 hours was used. For digestion of genomic DNA, an overnight incubation at 37°C was employed.

The products of DNA restriction digests were analysed by electrophoresis using agarose gels (Seakem LE agarose, FMC Bioproducts) at appropriate concentrations in TBE buffer.

Isolation of DNA Fragments for Cloning Procedures

Restriction digests were run on a low melting point agarose gel (Gibco ultrapure) of appropriate concentration (usually 0.7% w/v) in TBE buffer containing 1µg/ml ethidium bromide (Sigma). The gel was flouresced under

UV light and the required bands excised from the gel using a scalpel blade. The LMP gel containing the DNA was melted at 65°C for 10 minutes, then extracted twice with phenol (Fisons, Tris washed, liquified), twice with phenol/chloroform and once with chloroform (BDH AnalaR). The DNA was then precipitated with 2.5 volumes of ethanol and 1/10 volume of 3M sodium acetate (BDH) and resuspended at an appropriate concentration for further manipulation.

Blunting of 5' Overhangs Resulting from Restriction Digestion

In order to ligate DNA fragments which did not have complementary overhangs following restriction digestion, these ends were first blunted. In the case of a 5' overhang, this was done by filling in the overhang with Klenow.

The following components were assembled at room temperature:

DNA to be blunted
2.5µl 10mM dNTPs (Pharmacia)
1µl Klenow (Boehringer)
ddH2O to 25µl final volume

The reaction was incubated at 37°C for 30 minutes, then heated at 70°C to inactivate the enzyme. The DNA was then precipitated using 2.5 volumes of ethanol and 1/10 volume of 3M sodium acetate, spun down, rinsed with 70% (v/v) ethanol and resuspended in an appropriate volume of TE for further manipulation.

Dephosphorylation of Linearised Plasmid DNA

Removal of the 5' terminal phosphate groups of linearised plasmid DNA with complementary termini was required to prevent recircularisation of the plasmid DNA in subsequent ligation steps.

The following components were assembled at room temperature:

Linearised plasmid DNA (usually about 10µg)
1µl (1 unit) Calf Intestinal Alkaline Phosphatase (Boehringer)
10µl reaction buffer (supplied with the enzyme)
ddH2O to 100µl total volume

The reaction was incubated at 37°C for 30 minutes, then the following components added:

5µl of 10% SDS (0.5% final concentration, Sigma)
1µl of 500mM EDTA (5mM final, BDH)
1µl of 100mg/ml proteinase K (100µg/ml final, Sigma)

and the reaction incubated at 56°C for 30 minutes. The reaction was then extracted once with phenol and once with phenol/chloroform and precipitated with 2.5 volumes of ethanol and 1/10 volume of 3M sodium acetate, and resuspended at approximately 100ng/ml in TE.

Ligation of DNA Fragments

A ratio of 3:1 vector:insert DNA was normally used, with a 10ng input of vector. A control ligation containing no insert DNA was always included to assess the degree of re-circularisation of the plasmid. 1µl (1 unit) of T4 DNA

ligase (Boehringer) was added to the DNA, together with 1.5µl of 10 x ligase buffer (supplied with the enzyme) and the volume adjusted to 15µl with ddH2O. The reaction was left to proceed at 16°C overnight.

Preparation of Electrocompetent E.coli (Hanahan, 1985).

E.coli strains DH5α and JM101 were used to make electroporation competent cells according to the following protocol. The competence of the resultant cells was usually around 10^8/µg.

1) Inoculate 1 litre of LB growth medium with a single colony of *E.coli* and grow in an incubator shaker (New Brunswick Scientific) at 37°C until their OD600 reaches 0.6-0.9.

2) Chill the flasks on ice for 5 minutes and do not allow to warm up hereafter.

3) Spin the cells in 250ml bottles (Sorvall) at 4000rpm, 4°C for 20 minutes. Remove the supernatant immediately after centrifugation to prevent lifting of the pellet.

4) Gently resuspend each pellet in 250ml of ice cold water. Transfer each 250ml suspension to two pre-chilled 250ml bottles. Spin at 5000rpm, 4°C for 10 minutes. Remove the supernatant, resuspend in 200ml of ice cold water and spin as before.

5) Gently resuspend each pellet in 40ml of ice cold 10% glycerol (BDH) in water. Transfer to pre-chilled 50ml conical tubes (Corning) and spin at 4000rpm, 4°C for 10 minutes.

6) Resuspend in 2ml of ice cold 10% (v/v) glycerol per litre of original culture volume. Transfer to pre-chilled microfuge tubes and snap freeze in liquid nitrogen. Store at -70°C.

Electroporation of E.coli (Dower et al, 1988)

Electroporation of plasmids into *E.coli* was carried out according to the following protocol, using a Gene Pulser electroporator and 0.2cm Gene Pulser electroporation cuvettes.

1) Chill the electroporation cuvettes and the sliding cuvette holder on ice.

2) Set the gene pulser apparatus to 25μF capacitance, 1.8KV and 200Ω.

3) Thaw the electrocompetent cells at room temperature and place on ice.

4) To a cold 1.5ml microfuge tube, add 40μl of competent cell suspension and the DNA for electroporation (4μl of a ligation or 1μl of a stock plasmid).

5) Transfer the mixture of cells and DNA to a cold 0.2cm electroporation cuvette.

6) Apply one pulse at the above settings. This should result in a pulse of 12.5KV/cm with a time constant of 4.5 to 5.0msec.

7) Immediately add 1ml of LB medium (at room temperature) to the cuvette and gently resuspend the cells with a pasteur pipette.

8) Transfer the cells to a 15ml polypropylene pop-top tube (Falcon) and place in an incubator shaker at 37°C for one hour to recover.

9) Plate 100μl of the cell suspension on a 9cm diameter LB agar plate (Gibco), containing appropriate selection agent (usually 10μg/ml of ampicillin). Pellet the remainder of the cells by centrifugation, resuspend in 100μl of LB medium and plate onto a second LB agar plate. Grow both plates overnight at 37°C.

2.3 Screening of λ DASH Genomic Library

The murine 129 genomic library was a gift from Dr J. Rossant. RACE cloned DNA from line ST576 was used to probe the library to obtain genomic clones of a number of the genes which undergo *trans*- splicing in this line. The

screening was carried out according to the plaque hybridisation method of Benton and Davies (1977).

Titration of Library

A 10 ml culture of *E.coli* NM675 or LE392 was grown to stationary phase overnight, harvested by centrifugation at 3000rpm for 10 minutes and resuspended in 1/3 volume of 20mM magnesium chloride. This plating stock was stored at 4°C for up to one week. Serial dilutions of the library phage stock were made, and 10µl of each dilution added to 100µl of plating stock. These were incubated at 37°C for 30 minutes. Molten λ-top agarose (4ml) was added, mixed with the cells, and poured onto 9cm diameter LB agar plates. Once set, these were incubated overnight at 37°C. The number of plaques obtained with each dilution of the library stock was used to calculate the titre of the original library stock.

Plating of the λ-DASH Library

The library was plated onto ten 14cm diameter plates at a density of 50 000 plaques per plate using the method described above. Phage were transferred onto 13.2cm nylon filters (Biodyne A Transfer Membrane, Pall Europe Ltd) by laying the dry filters directly onto the agarose. Two filters were taken per plate. The first was left in contact with the agarose for 1 minute, the second for 2 minutes. Orientation marks were made by stabbing through both the filter and the agar with a needle. Each filter was peeled catrefully from the surface of the agar using forceps and floated, DNA side up, on a shallow tray of denaturation buffer for 3 minutes. The filters were then transferred to neutralising buffer for 3 minutes and finally rinsed briefly in 2 x SSC, air dried and baked at 80°C for 90 minutes. Duplicate pairs of filters were

hybridised with the appropriate random primed probes, exposed to autoradiographic film (Kodak X-OMAT, IBI) and positively hybridising plaques identified by alignment of the plates with the autoradiographs. The positive plaques were picked using the blunt end of a pasteur pipette and each was placed into 1ml of SM buffer containing 10mM magnesium sulphate. The plaques were stored in this buffer at 4°C for a minimum of 16 hours to allow the phage to adsorb from the agarose before secondary screening was carried out.

Secondary Screening

10µl of positively hybridising phage suspension was added to 100µl of plating stock and plated on 9cm diameter LB agar plates as described above. Duplicate filters were taken from each plate as described. Following hybridisation, positive plaques were picked into SM buffer using the thin end of a pasteur pipette, and the screening procedure repeated until the phage stocks were plaque pure ie. every plaque produced from the stock gave a positive signal on hybridisation.

λ Top Agarose

Seakem LE Agarose	7g
Sodium Chloride	2.5g
Trypone (Difco)	10g
ddH2O	1 Litre

SM Buffer

100mM Sodium Chloride (BDH)

10mM Magnesium Sulphate (BDH)

50mM Tris-HCl pH7.5 (Boehringer)

0.01% (v/v) gelatin (BDH)

Denaturation Buffer

1.5M Sodium Chloride

0.5M Sodium Hydroxide (BDH)

Neutralising Buffer

1.5M Sodium Chloride

0.5M Tris-HCl pH 8.0

2.4 Isolation of Nucleic Acids

Small Scale Preparation of Plasmid DNA

A single colony was picked with a sterile gilson tip and used to inoculate 2ml of L broth, containing 5µg/ml ampicillin (Sigma). This culture was shaken at 37°C overnight and DNA prepared according to the following protocol:

1) Fill an eppendorf with the overnight culture and spin 15s to pellet the bacteria.

2) Remove and discard supenatant.

3) Add 100µl solution I and resuspend by vortexing

4) Add 200µl solution II and mix gently.

5) Add 150µl solution III and mix by inversion.

6) Spin 5 mins in microcentrifuge.

7) Transfer supernatant to new tube, Add 250µl phenol and 200µl chloroform. Vortex. Spin 3 mins and remove aqueous phase.

8) Precipitate with 2 volumes of ethanol and 1/10 volume 3M Sodium Acetate (BDH) on ice for 10 minutes.

9) Spin in microcentrifuge for 10 minutes, rinse in 70% (v/v) EtOH, dry and resuspend in 50µl TE plus 100µg/ml RNA'se A (Sigma). Incubate at 37°C for 30 mins.

Solution I (TGE)

50mM Glucose (BDH)

25mM Tris-HCl pH8.0

10mM EDTA pH8.0

Solution II (Lysis Buffer)

0.2M NaOH

1% (w/v) SDS

Solution III

5M Potassium Acetate (BDH)	60ml
Glacial Acetic Acid (Fisons)	11.5ml
ddH_2O	28.5ml

Medium Scale Preparation of Plasmid DNA

Plasmids required for probe templates or cloning procedures were prepared from 50ml overnight cultures of E.coli using a Qiagen Plasmid Midi kit (Qiagen Inc.) according to the manufacturers instructions.

Large Scale Preparation of Plasmid DNA

Plasmids required for electroporation into cells were prepared according to the following protocol.

1) Pick a single colony of *E. coli* containing the vector and drop into 10ml of LB medium supplemented with 10µg/ml Ampicillin.

2) Grow overnight in an automatic shaker at 37°C.

3) Add 2ml of this overnight culture to 500ml of LB medium containing ampicillin and shake overnight at 37°C.

4) Divide the culture into two 250ml centrifuge bottles and spin at 6000rpm, 4°C for 10mins.

5) Discard the supernatant and resuspend each pellet in 25ml of TGE (see small scale plasmid preparation, above).

6) Add 5ml of 10µg/ml lysozyme (Sigma) in TGE and allow the bottles to stand for 10mins on ice.

7) Add 60ml of a fresh 9:1 dilution of 0.22M NaOH: 10% SDS and leave the bottles on ice for 5mins.

8) Add 30ml of solution III (see small scale plasmid preparation, above), mix gently and leave on ice for a further 15 minutes.

9) Centrifuge the samples at 6000rpm for 10 minutes.

10) Transfer the supernatant into a 500ml bottle, add 0.6 volumes of isopropanol (BDH) and pellet the DNA by spinning at 8000rpm and 4°C for 15 minutes.

11) Discard the supernatant and resuspend the pellet in 8ml of TE.

12) Titrate to neutrality with 3M Tris-HCl and transfer to 30ml corex tubes.

13) Add an equal volume of tris saturated phenol (Fisons) and shake well.

14) Spin at 8000rpm for 5 miutes and remove the aqueous phase to a fresh tube.

15) Add an equal volume of chloroform, shake well, spin at 8000rpm for 5 minutes and remove the aqueous phase to a fresh tube.

16) Add 1ml of 3M Sodium acetate pH5.0 and 18ml of ethanol to each tube and precipitate the DNA at -20°C for at least one hour.

17) Pellet the DNA at 8000rpm, 4°C for 15 minutes and discard the supernatant.

18) Wash the pellet twice with 70% (v/v) ethanol and dissolve in 5ml of TE.

19) For each tube, weigh 10g of caesium chloride (Boehringer) and dissolve in 5ml of TE.

20) Add the CsCl solution to the DNA and add 1ml of 10mg/ml ethidium bromide.

21) Transfer the solution to polyallomer quick seal tubes (Beckman), balance using 1g + 1ml CsCl and spin overnight at 50 000rpm, 16°C (L7 Ultracentrifuge, Beckman).

22) Flouresce the tubes using long wave UV light and remove the lower DNA band to a 15ml centrifuge tube (Corning) by side puncture using a needle and syringe.

23) Add TE to give a volume of 6ml and extract the mixture with equal volumes of n-butanol (BDH) until all traces of ethidium bromide are removed.

24) Add an equal volume of sterile water and two volumes of ethanol, mix and spin at 10 000rpm, 4°C for 10 minutes.

25) Rinse the pellet in 70% (v/v) ethanol, air dry and resuspend in 200µl of TE.

Preparation of Genomic DNA From Tissue Culture Cells

Cells were grown in a $25cm^2$ flask until confluent. The cells were harvested by trypsinisation, pelleted by centrifugation and resuspended in 3ml of TEN (10mM Tris; 50mM EDTA; 100mM NaCl) to lyse the cells. SDS was added to a final concentration of 0.5%. Proteinase K (Sigma) was added to a final concentration of 0.5µg/ml and the lysate incubated at 37°C for 2hrs to overnight. The lysate was then phenol/chloroform extracted twice, with gentle vortexing and an equal volume of isopropanol added to the aqueous phase to precipitate the DNA. The genomic DNA was then spooled with a sealed pasteur pipette and rinsed in 70% (v/v) ethanol. The DNA was air dried and resuspended in TE at a concentration of 1mg/ml.

Preparation of Phage DNA From Liquid Culture

Preparation of Liquid Phage Lysate (Ausubel et al (ed.) 1987).

A single plaque was picked and placed in 0.5ml of SM. This was stored at 4°C for a minimum of 2 hours. 100µl of this phage stock was then added to 100µl of SM and 100µl of a fresh overnight culture of plating cells and this mixture incubated for 20 minutes at 37°C to allow the phage to adsorb to the cells. The cells and phage were then added to 50ml of L broth, supplemented with 20mM $MgSO_4$ at neutral pH and shaken at 37°C for at least 5 hours, until the cells had grown up to form an opaque culture, and then lysed. 250µl of chloroform was added to 50ml of lysate, which was shaken and then centrifuged at 4000rpm for 10 minutes. The supernatant was carefully removed, and the lower phase discarded.

Preparation of DNA From Liquid Lysate (Davis et al (eds), 1986).

For each 10ml of lysate, the following protocol was then used:

1) To 10ml of lysate, add 10ml of TM buffer (10mM MgSO4; 10mM Tris-HCl pH7.5); 32µl of 10mg/ml DNase I (Boehringer) and 32µl of 10mg/ml RNase A. Mix gently.

2) Incubate for 15 minutes at room temperature.

3) Add 2ml of 5M NaCl and 2.2g of solid Polyethylene Glycol (PEG, Sigma). Leave at room temperature until PEG is fully dissolved.

4) Place on ice for 15 minutes.

5) Spin for 10 minutes at 12 000 x G and 4°C.

6) Discard supernatant and resuspend phage pellet in 300µl of TM buffer. Transfer to a microfuge tube.

7) Extract twice with 300µl chloroform, making sure not to carry over the PEG interface between the two phases.

8) To the aqueous phase, add 15µl of 0.5M EDTA pH8.0 and 30µl of 5M NaCl.

9) Add 350µl of phenol, mix by vortexing, spin for 5 minutes in microcentrifuge and remove aqueous phase to a fresh tube.

10) Extract with 350µl chloroform.

11) Add 875µl ethanol and precipitate on ice for 10 minutes. Spin for 10 minutes in microcentrifuge, rinse pellet with 70% (v/v) ethanol, vacuum dry and resuspend in 50µl of TE plus RNase A.

Expected yield is 2-5 mg of DNA per 1ml of original lysate.

Preparation of total RNA

Cells were grown to confluency in $25cm^2$ tissue culture flasks. 2.5ml of guanidinium lysis buffer was added to each flask, and the flasks laid flat to ensure that the lysis buffer covered the entire monolayer. The lysate was removed and drawn through a 23 gauge needle several times to shear the

genomic DNA. A cushion of 3.6ml of 5M caesium chloride was placed in an SW50.1 polyallomer centrifuge tube (Beckman) and 1.4 ml of lysate gently layered on top. The tubes were spun at 35, 000rpm, 16°C overnight. The supernatant was removed, and the walls of the tube rinsed with guanidinium lysis buffer. The bottom of the tube was removed with a scalpel blade, and the RNA pellet washed with 70% (v/v) ethanol. The pellet was then resuspended in 300µl of urea buffer and 100µl of DEPC treated water. The RNA was then precipitated with 40µl of 2M sodium acetate and 800µl of ethanol and stored at -20°C as a precipitate.

Guanidinium Lysis Buffer

4M Guanidinium

0.1M Tris-HCl pH7.5

Add 7.2µl of β-mercaptoethanol per 1ml immediately before use.

Urea Buffer

8M Urea (BDH)

10mM HEPES (Gibco)

Preparation of Cytoplasmic RNA (Ausubel et al, eds, 1989).

Cells were grown in 25cm^2 tissue culture flasks. Each flask was rinsed twice with ice cold PBS and the cells scraped from the growth surface. The cells were pelleted by centrifugation at 2000rpm in an F28/50 rotor (Sorvall) at 4°C for 5 minutes. The supernatant was removed and the cell pellet loosened by vortexing at half maximum speed for a few seconds. 4ml of NP-40 lysis buffer was then added slowly, while continuing to vortex. This mixture was incubated on ice for 5 minutes, then homogenised gently in a dounce homogeniser to ensure that the cells had lysed, leaving the nuclei

intact. The condition of the nuclei was checked using a microscope. The nuclei were pelleted by centrifugation for 5 minutes at 2000rpm in an F-28/50 rotor and the supernatant containing the cytoplasmic RNA removed. The supernatent was centrifuged for 10 minutes at 9000rpm in an F-28/50 rotor and the supernatant removed to a 50ml plastic centrifuge tube (Corning). 4ml of 2 x proteinase K buffer was added, together with 80µl of 20mg/ml proteinase K (Sigma). The mixture was incubated at 37°C for 30 minutes. The proteinase K digestion was quenched by the addition of 4ml of phenol and 4ml of chloroform. This was mixed well, then spun at 3000rpm for 5 minutes. The aqueous phase was removed and the the RNA precipitated from it using 2.5 volumes of ethanol. The RNA was pelleted by centrifugation, rinsed with 70% (v/v) ethanol, air dried and resuspended in 200µl of DEPC treated water.

NP-40 Lysis Buffer

10mM Tris-HCl pH7.4

10mM NaCl

3mM $MgCl_2$

0.5% (v/v) Nonidet P-40 (Sigma)

Autoclave the first three components and allow to cool prior to addition of NP-40.

2 x Proteinase K Buffer

0.2M Tris-Cl pH7.5

0.44M NaCl

2% (w/v) SDS

25mM EDTA

2.5 Nucleic Acid Filter Preparation

Southern Blotting.

A) Small DNA Targets (Plasmid, PCR Products etc.)

Samples were loaded onto a 0.8% (w/v) agarose gel made up in 1 x TBE buffer and run at 60V. The gel was then soaked in 0.4M sodium hydroxide to denature the samples, then lain DNA side up on a clean glass plate. A piece of hybond N+ nylon membrane (Amersham) was cut to the size of the gel and pre-wet in 0.4M NaOH. The membrane was lain on the gel, and air bubbles gently expelled. Two sheets of 3MM filter paper were wet in 0.4M NaOH and placed on top of the membrane, followed by 2 sheets of dry 3MM paper and a one inch thickness of paper towels. A second glass plate and a weight were placed on top of the towels and the gel left to blot overnight. The blotting apparatus was then dismantled, and the nylon membrane rinsed in 0.5M sodium phosphate buffer prior to hybridisation.

B) Large DNA Targets (Genomic DNA)

Samples were loaded onto a 0.8% (w/v) agarose gel and run at 30V overnight. The gel was then soaked in 0.25M hydrochloric acid to nick the DNA, then in 0.4M sodium hydroxide for 2 x 45 minutes to denature the DNA. The gel was transferred by capilliary blot using a 20 x SSC transfer buffer (see figure 2.1, below). The filter was rinsed with 0.5M sodium phosphate buffer, then UV crosslinked using a stratagene stratalinker on a 1200μJ setting prior to hybridisation.

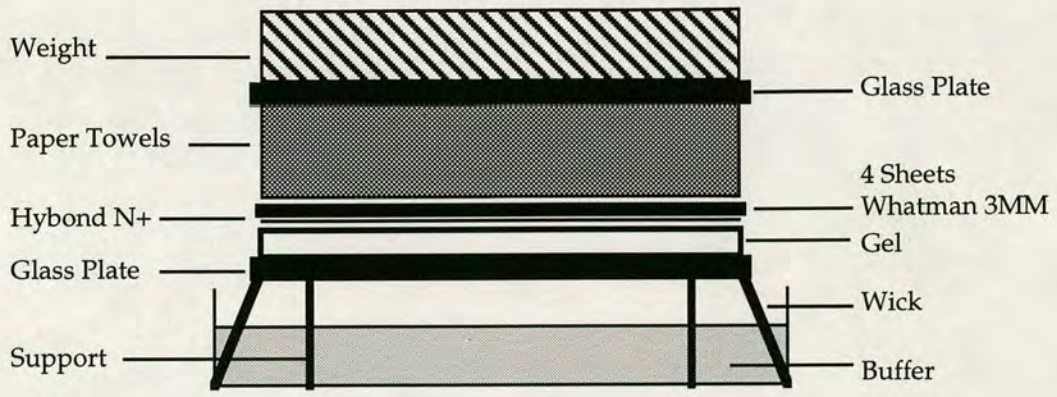


Figure 2.1: Construction of a Capilliary Southern Blot.

Northern Blotting.

10µg samples of each RNA were pelleted and resuspended in 4.5µl of DEPC treated water, 2µl of 5 x formaldehyde gel running buffer; 3.5µl of formaldehyde and 10µl of formamide. The samples were then denatured at 55°C for 15 minutes. 2µl of sample dye and 0.2µl of 10mg/ml ethidium bromide were added to each sample.

The samples were run on a 0.9% (w/v) agarose gel containing 2.2M formaldehyde, (made up in formadehyde gel running buffer) at 30V overnight. The gel was rinsed in sterile water for 1 hour to remove the formaldehyde prior to photography and further processing.

The gel was soaked in 50mM sodium hydroxide for 45 minutes to nick the RNA and then neutralised for 30 minutes in 0.5M sodium phosphate buffer. A nylon filter (Hybond N+) was pre-wet in 0.5M sodium phosphate buffer and placed in contact with the gel. Three sheets of 3MM paper were placed on top of the filter, followed by about 3cm thickness of paper towels. A weight was placed on the top and the samples allowed to transfer overnight.

The filter was then UV cross linked (Stratagene Stratalinker) before hybridisation.

Formaldehyde Gel Running Buffer (10X)

1M MOPS pH7.0	400ml
3M Sodium Acetate pH4-5	33ml
0.5M EDTA pH8.0	20ml
DEPC H_2O	547ml

1M MOPS

41.8g MOPS (3-[N-Morpholino]propane-sulphonic acid, Sigma)

4.1g Sodium Acetate

10ml 500mM EDTA pH8.0

pH to 7.0, autoclave and store in the dark.

1M Sodium Phosphate Buffer

70.5g Na_2HPO_4 (anhydrous, BDH)

4ml Phosphoric Acid (BDH)

ddH_2O to 1 litre final volume.

Filter before use, using a bottle top filter (Nalgene).

2.6 Radiolabelling of DNA

Random Primed DNA Probes

The probe fragment was released by digestion of plasmid DNA with the appropriate restriction enzymes. The digest was electrophoresed through a 0.8% (w.v) low melting point agarose gel, made up in TBE buffer and the probe fragment excised. The concentration of the fragment was adjusted to 5mg/ml by the addition of sterile water. The gel containing the probe

template was melted at 60°C and 5µl (25µg) mixed with 5µl of sterile water, 5µl of DNA labelling mix (Containing 2µl of hexanucleotide mix and 3µl of 0.6mM adenosine, guanidine and thymidine triphosphates from the Boehringer Random Prime DNA Labelling Kit), 5µl of ^{32}P dCTP (10mCi/µl, Amersham) and 1µl (2 units) of Klenow (Boehringer). The probe mixture was incubated at 37°C for 30 minutes, then unincorporated nucleotides were removed by spinning the probe through a G-50 sephadex column. Immediately prior to adding the probe to the hybridisation mixture, it was denatured at 100°C for 2 minutes.

5' End-Labelling of Oligonucleotides

The following components were mixed together in a microfuge tube:

10.5pmol oligonucleotide

7µl [γ-32P] rATP (3000 Ci/mmol, Amersham)

2.5µl 10 x T4 Polynucleotide Kinase buffer

1µl (10 units) T4 polynucleotide kinase (Boehringer)

ddH2O to 25µl final volume

The reaction was allowed to proceed at 37°C for 30 minutes, then placed at 95°C for 5 minutes to terminate the reaction. The concentration of end-labelled primer produced by this protocol is 0.4pmol/µl.

2.7 Filter Hybridisations

Southern and Northern blots were hybridised using a technique based on that of Church and Gilbert (1984).

Southern Blots

Filters were pre-hybridised in glass bottles (Techne) for 1 hour at 65°C in 0.5M sodium phosphate buffer, 7% (w/v) SDS with 5mg/ml skimmed milk powder (Marvel). This buffer was then removed and replaced with fresh buffer containing approximately half of the yield from a random primed

DNA probe reaction. The probe was boiled for 3 minutes before its addition to the hybridisation mixture. Hybridisation was carried out overnight at 65°C and the filters washed for 1 x 15 minutes and 2 x 30 minutes with 150mM sodium phosphate, 0.1% (w/v) SDS proir to exposure to autoradiographic film (Kodak X-OMAT).

Northern Blots

Filters were pre-hybridised in glass bottles for 1 hour at 60°C in 175mM sodium phosphate buffer, 3.5% SDS, 30% formamide (BDH) with 5mg/ml BSA fraction V (Sigma). This buffer was removed and replaced with fresh buffer containing half of the yield from a random prime DNA probe reaction. The probe was boiled for 3 minutes before its addition to the hybridisation mixture. Hybridisation was carried out overnight at 60°C and the filters washed for 1 x 15 minutes and 2 x 30 minutes with 30mM sodium phosphate, 0.1% SDS prior to exposure to autoradiographic film.

Library Filters

Filters were pre-hybridised in deep glass petri dishes for 1 hour at 65°C in 0.5 x Denhardts; 6 x SSC; 0.5% (w/v) Sarkosyl (N-lauroylsarcosine, Sigma) with 100µg/ml herring sperm DNA. This buffer was then replaced with 0.5 x Denhardts; 6 X SSC; 0.5% (w/v) Sarkosyl; 10% (w/v) Dextran Sulphate (sodium salt, Pharmacia) containing 100µg/ml herring sperm DNA and the yield from a random prime DNA probe reaction. The probe was boiled for 3 minutes before being added to the hybridisation mixture. The filters were hybridised at 65°C overnight, then washed for 2 x 30 minutes with 2 x SSC, 0.1% (w/v) SDS prior to exposure to autoradiographic film.

10 X Denhardt's Solution

2g Polyvinylpyrrolidone (PVP, Sima)

2g Ficoll (Pharmacia)

2g BSA (Sigma)

ddH_2O to 1 litre final volume. Filter before use.

Stripping of Filters

Filters to be re-probed were stripped using the following protocols. In each case, the stripped filters were exposed to autoradiographic film for 48 hours before re-probing to ensure that all previous signals had been removed.

Southern Blots and Library Filters

These were stripped by placing them in boiling 0.1% (w/v) SDS and leaving the solution to cool to room temperature. This was repeated until no signal could be detected using a geiger counter (usually 3 times).

Northern Blots

These were stripped by incubating with 50% (v/v) formamide, 15mM sodium phophate buffer for 30 minutes or longer, until no signal could be detected using a geiger counter.

2.8 Sequencing of DNA

Plasmid DNA

plasmid DNA was prepared according to the following protocol before sequencing using the Sequenase 2.0 Double Stranded DNA Sequencing Kit (USB) according to the manufacturers instructions.

1) Resuspend approximately 2.5µg of DNA prepared using the small scale protocol (see 2.4) in 50µl of TE containing 100µg/ml RNA'se A.

2) Digest at 37°C for 30 minutes.

3) Add 30µl of 20% (w/v) Polyethylene Glycol (PEG), 2.5M sodium chloride.

4) Precipitate on ice for 1 hour.

5) Pellet the DNA by centrifugation and rinse the pellet with 70% (v/v) ethanol.

6) Vacuum dry the DNA and resuspend in 18µl of ddH_2O.

7) Add 2µl of 2M sodium hydroxide and incubate at room temperature for 5 minutes.

8) Add 4µl of 10M ammonium acetate and 500µl of ethanol simultaneously and precipitate on ice for 15 minutes.

9) Pellet the DNA by centrifugation, wash with 70% (v/v) ethanol, vacuum dry and store at -20°C prior to sequencing.

λ Genomic Clones

λ DNA prepared as described (see 2.4) was sequenced using the CircumVent Thermal Cycle Dideoxy DNA Sequencing Kit (NEB). 50ng of template DNA was used for each reaction, and the manufacturers protocol for sequencing using 5' end-labelled primers was followed.

Primers Used (Synthsised by Oswell DNA Ltd)

Oligo 107 (4B-11): TTGCTGGCTATGATGAC

Oligo 122 (4B-2): TGAAGGTTAGCGCGACA

Oligo 123 (4B-15): GTTGCTCTAAGGACTGT

Oligo 124 (4B-8): GGTCAGTTTGCGCTCTC

Oligo 142 (4B-7): GTCCTTGCTGGTGAAGC

2.9: 5' Rapid Amplification of cDNA Ends (RACE) Cloning

Protocol A (W.C Skarnes,1990)

1) First strand cDNA synthesis

Resuspend 5µg of RNA in 8.5µl of DEPC treated water. Add 5ng of oligo 1 (*lacZ* primer). Heat at 68°C for 5 minutes and cool on ice.

Add: 1µl of 20mM DTT (Dithiothrietol, Boehringer)

1µl of 10mM dNTPs (Pharmacia)

0.25µl of RNase inhibitor (Boehringer)

4µl of 5 x reverse transcriptase buffer (Supplied with enzyme)

1µl (10 units) reverse transcriptase (Life Sciences)

DEPC water to 20µl total.

Incubate at 42°C for 2 hours.

Add: 80µl of DEPC treated water

20µl of 10M ammonium acetate

320µl ethanol

Precipitate on ice for 30 minutes, spin down, wash pellet with 70% ethanol and vacuum dry.

2) Alkaline Hydrolysis and Purification of First Strand cDNA

Resuspend cDNA reaction in 17µl of ddH_2O. Add 18µl of 0.2M sodium hydroxide and incubate at 65°C for 1 hour. Neutralise with 25µl of 1M Tris pH8.0 and 25µl of 0.1M hydrochloric acid (BDH).

Add 5µl of glycogen (Boehringer); 20µl of 10M ammonium acetate and 250µl ethanol and precipitate for 30 minutes on ice. Spin down the pellet, wash with 70% (v/v) ethanol, dry and resuspend in 100µl of TE.

Extract with phenol/chloroform, then add 2µl of 5M sodium chloride and 250µl of ethanol. Precipitate for 30 minutes on ice, spin down, wash with 70% (v/v) ethanol and vacuum dry.

3) A-tailing of First Strand cDNA.

Resuspend cDNA in: 26µl ddH_2O

8µl 5 x TdT buffer (Supplied with Enzyme)

4µl 2mM dATP

2µl (50 units) terminal transferase (Gibco)

ddH_2O to 40µl total

Incubate at 37°C for 7.5 minutes. Add 110µl of ddH_2O and extract the sample with phenol, phenol/chloroform them chloroform. Add 50µl of 10M ammonium acetate, then precipitate with 0.5ml of ethanol on ice for 10 minutes. Centrifuge to pellet the DNA, rinse with 70% (v/v) ethanol and vacuum dry.

4) Second Strand Synthesis

Resuspend A-tailed cDNA in: 12µl ddH_2O

1.5µl 10 x Klenow buffer

0.5µl 10mM dNTPs

10ng T-tailed primer (oligo 56)

1µl (5 units) Klenow

ddH_2O to 15µl total

Incubate at room temperature for 30 minutes, then 37°C for a further 30 minutes. Add 65µl ddH_2O; 20µl of 10M ammonium acetate and 250µl ethanol. Precipitate on ice for 10 minutes, centrifuge, wash pellet with 70% (v/v) ethanol and vacuum dry.

5) Polymerase Chain Reaction

Resuspend sample in: 40µl ddH_2O

5µl of 5 x PCR buffer (250mM KCl/ 50mM Tris, pH8.3 @ 37°C)

1µl 10mM dNTPs

1µg oligo 59 (oligo 56 without the T tail)

1µg oligo 55 (*En-2* primer)

1µl Taq polymerase (Promega)

ddH_2O to 50µl final volume

Overlay the samples with a drop of mineral oil (Sigma), then put through 40 cycles of: 1.5 minutes @ 94°C

2 minutes @ 65°C

15 minutes @ 72°C

At the end of the final cycle, add 1µl of 10mM dNTPs and 0.5µl of Taq. Incubate at 72°C for 30 minutes, then cool slowly to room temperature.

6) Size Selection of PCR Products

Run the PCR products on a 0.8% (w/v) agarose gel containing ethidium bromide. Make a small cut in the lane containing the products at a distance down the gel representing about 400bp. Prepare DEAE cellulose paper (Schleicher and Schuell) by soaking for 5 minutes in 10mM EDTA (pH8.0), 5 minutes in 0.5M NaOH, then rinsing thouroughly in ddH_2O. Insert a small piece of DEAE paper into the cut and continue to run the gel until products ranging from 400bp up to 700bp have been collected on the paper. Elute the cDNA from the paper by soaking in 1.5M sodium chloride in TE buffer at 37°C for 1 hour. Add 5µg of carrier tRNA (Boehringer) and 2.5 volumes of

ethanol and precipitate on ice for 10 minutes. Spin for 15 minutes to pellet the cDNA, wash with 70% (v/v) ethanol, vacuum dry and resuspend in 160ml of ddH_2O. Precipitate again using 0.2M final ammonium acetate and 2.5 volumes of ethanol.

7) Cloning of Amplified cDNA

Digest the cDNA with XbaI and Asp718. Phenol extract, precipitate with 0.2M final ammoninum acetate and 2.5 volumes of ethanol. Clone the cDNA fragments into KpnI/XbaI digested Bluescript II SK+ (Promega) using an insert:vector ratio of 3:1 for the ligation, electroporate into DH5α cells and sequence the inserts using Sequenase 2.0 double stranded DNA sequencing kit (USB).

Protocol B

1) 1st strand cDNA Synthesis

Resuspend 5µg total RNA or 10µg cytoplasmic RNA in 5µl of DEPC treated water, add 10ng of oligo 1 (*lacZ* primer) and make the volume up to 11µl with DEPC water. Heat at 70°C for 5 minutes and cool on ice.

Add on ice: 2µl 0.1M DTT
1µl 10mM dNTPs (Pharmacia)
0.5µl Rnase Inhibitor (Boehringer)
4µl 5 x 1st strand buffer (Supplied with Superscript II)

Heat to 37°C for 2 mins, then add 1µl Superscript II (Gibco) and incubate at 37°C for 1 hour.

2) Alkaline Hydrolysis and Purification of First Strand DNA.

Add 2.2µl of 1M NaOH and incubate at 65°C for 20 minutes. Neutralise with 2.2µl of 1M hydrochloric acid (BDH).

Load total volume of 1st strand reaction onto a 0.1mm microdialysis filter (Millipore) floating in a petri dish of TE. and leave for 4 hours. Take up in 20µl of ddH_2O.

3) A-tailing of First Strand cDNA

To 20µl of 1st strand reaction, add: 6µl 5 x TdT buffer
2µl 2mM dATP

Incubate for 2mins at 37°C, then add 2µl terminal transferase and incubate for a further 5 minutes at 37°C, then for 2 minutes at 70°C.

4) Second Strand Synthesis.

Mix together: 15µl tailed cDNA
2.0µl 10 x restriction buffer M (Boehringer)
1µl 10mM dNTP's
1ng T-tailed primer (Oligo 56)
1µl (2 units) Klenow

Incubate at room temperature for 30 mins, then 37°C for a further 30 minutes, followed by 5 mins at 70°C

5) Size Selection of Second Strand Products

Load total volume into microdialysis 0.1 mm filter floating on a dish of T.E. and leave for 4 hrs. Take up in 36.5 µl of ddH_2O.

6) First Round Polymerase Chain Reaction.

Mix together: 36.5 µl DNA in H_2O

5µl 10 x PCR buffer (500mM KCl/ 100mM Tris, pH8.3 @ 37°C)

4µl $MgCl_2$ (25mM)

1µl 10mM dNTP's

500ng Oligo 59 (oligo 56 without the T tail)

500ng Oligo 67 (CD4 primer)

1.0µl Amplitaq (Perkin Elmer)

ddH_2O to 50µl total volume

Overlay the samples with a drop of mineral oil, then put through 40 cycles of: 1.5 minutes @ 94°C

1.5 minutes @ 60°C

3 minutes @ 72°C

7) Size Selection of First Round PCR Products.

Run 25µl of the first round PCR on a 1% (w/v) low melting point agarose gel. Cut out a region of the gel of volume about 100µl at aroud 600bp. Melt the gel @ 70°C for 5 minutes.

8) Second Round Nested PCR

Mix together:

5µl low melt gel

5µl 10 x PCR buffer

4µl 25mM $MgCl_2$

1µl 10mM dNTP's

500ng Oligo 55 (*En-2* primer)

500ng Oligo 59

1µl Amplitaq

ddH_2O to 50µl total volume

Overlay the samples with a drop of mineral oil and put through 30 cycles of:

1.5 minutes @ 94°C

1.5 minutes @ 60°C

3 minutes @ 72°C

After the final round, add:

250ng of each primer

0.5µl Taq

0.5µl 10 mM dNTP's

and put through a single cycle of:

1.5 minutes @ 94°C

1.5 minutes @ 60°C

20 minutes @ 72°C

9) Size Selection of Second Round PCR Products

Microdialyse on 0.1mm filter against TE for 4 hours. Recover in 80µl H_20.

10) Cloning of Amplified cDNA

Digest the cDNA with XbaI and Asp718. Phenol extract, precipitate with 0.2M final ammoninum acetate and 2.5 volumes of ethanol. Clone the cDNA fragments into KpnI/XbaI digested Bluescript II SK+ (Promega) using an insert:vector ratio of 3:1 for the ligation, electroporate into DH5α cells and sequence the inserts using Sequenase 2.0 double stranded DNA sequencing kit (USB).

Primers Used for 5' RACE Cloning

Oligo 1: AGGGTTTTCCCAGTCACGACG

Oligo 55: TGCTCTGTCAGGTACCTGTTGG

Oligo 56: GGTTGTGTCGACTATCGATGGGTTTTTTTTTTTTTTTT

Oligo 59: GGTTGTGTCGACTATCGATGGG

Oligo 67: AGTAGACTTCTGCACAGACACC

All primers were synthesised by Oswell DNA.

2.10 Extended Range Polymerase Chain Reaction Using Genomic DNA Template

PCR of genomic sequences was carried out using a mixture of Taq and Pfu polymerases. A ratio of 1 unit Pfu (Stratagene) to 100 units AmpliTaq (Perkin Elmer) was used. The reaction buffer consisted of 20mM Tris-Cl pH 8.55; BSA at 150μg/ml; 16mM $(NH_4)_2SO_4$; 3.5mM $MgCl_2$ (Barnes, 1994). 1μg of genomic DNA was used as template, and 30 cycles carried out using 30sec at 99°C; 30 sec at 67°C and 10 minutes at 68°C.

Primers Used

Oligos 55 and 67 to vector sequences (see above)

Oligo 83 to 5'ETS Sequence: TCCTTCCATCTCTCGCGCAATG

2.11 RNAse Protection Assays

These assays were based on the protocol of King and Melton, (1987).

Probe Templates

Probe templates were made by restriction digest of bluescript plasmid containing RACE cloned products. In each case, the T7 polymerase site 3' of the sequence of interest was retained. The digests were then run on a 1.2% (w/v) agarose gel, and the fragments representing the probe templates excised. The DNA was extracted from the agarose by spinning the gel slice through synthetic wool (Filter Floss, Crystal Clear Ltd. Bolton, Lancs) at 4°C and then ethanol precipitated and resuspended in DEPC treated water at approximately 250ng/µl.

Probe synthesis

The following components were mixed together at room temperature:

template	500ng (approximately 2µl)
200mM DTT	0.75µl
2mg/ml BSA	0.75µl
3.3mM ATP/UTP/GTP	2.25µl
RN'ase inhibitor (Boehringer)	0.5µl
[α32P] CTP (800Ci/mmol, DuPont)	6.25µl
10 x transcription buffer	1.5µl
Polymerase (usually T7, Boehringer)	1µl

For GAPDH loading control probe, 0.25µl of [α32P] CTP was used, together with 6µl of 1mM unlabelled CTP.

The reaction mixtures were then incubated at 37°C (40°C for GAPDH) for 1hr.

1µl of RNase free DNase (Boehringer) was then added to each reaction to digest the probe template, and incubation at 37°C continued for 10 minutes. A further 1µl of DNase was added, followed by a further 10 minute incubation to ensure maximum degradation of the probe template.

The volume of each probe was made up to 100µl with sterile water, and each probe extracted once with phenol/chloroform and spun through a sephadex G-50 column to remove unincorporated nucleotides.

Gel Purification of Probes

12µl of each probe was mixed with 8µl of loading dye (stop solution from Sequenase 2.0 sequencing kit, USB), denatured briefly, and run on a 6% (w/v) polyacrylamide sequencing gel for 1hr at 60W.

The wet gel was wrapped in saran wrap and orientation marks made. A piece of X-ray film was then placed on the gel, and the same orientation marks made on the film. After developing the film, the two sets of marks were aligned, and the area of the gel containing the probe band carefully excised and removed to a microfuge tube. A further piece of X-ray film was then placed on the gel and developed to ensure that the correct region of the gel had been removed. 100µl of probe elution buffer was then added to each gel slice, and the gel crushed with a disposable spatula. The probes were then shaken at 37°C for 2hrs, at the end of which time the gel was spun to the bottom of the tube and the supernatant, containing the probe, removed. To

ensure maximum yield of probe from the gel, a further 50µl of elution buffer was added to the gel, vortexed briefly, and the supernatant pooled with the previous 100µl. 1µl of each probe was then added to scintillation fluid and the incorporation of each probe measured.

Hybridisation

10µg samples of target RNAs were precipitated, spun down and resuspended in 2µl DEPC water. A volume of eluted probe containing 350 000cpm was added to each sample (50 000cpm for GAPDH loading control) and the volume made up to 100µl with DEPC water. These mixtures were then precipitated at -70°C using 250µl of ethanol, with 1µl of glycogen (Boehringer) as carrier. The precipitated RNA and probe mixtures were then spun down, air dried and resuspended in 3µl of DEPC water, 3µl of 10 x hybridisation buffer and 24µl of deionized formamide (IBI). The samples were then denatured at 85°C for 15 minutes and hybridised overnight at 55°C.

RNase Digestion

Digestion buffer was made up as follows:

2M tris pH7.5	30µl
500mM EDTA	60µl
5M NaCl	360µl
dH_2O	5.55ml
25mg/ml RNase A	9.6µl
RNase T1 (Gibco)	2.1µl

and 350µl added to each sample. The samples were then put at 30°C for 30 minutes to digest away unhybridised RNA.

20µl of 10% (w/v) SDS and 5µl of 10mg/ml proteinase K (Boehringer) were then added to stop the digestion, and the samples incubated at 30°C for a further 10 minutes.

The samples were then extracted once with phenol/chloroform and precipitated at -70°C with 1ml of ethanol using 5µg of tRNA as carrier.

The precipitated RNAs were then spun down, air dried and resuspended in 1µl of DEPC water and 4µl of loading dye, denatured briefly at 95°C and run on a 6% (w/v) polyacrylamide gel. The gel was dried down in a Bio-Rad gel drier before exposure to autoradiographic film.

10 x Transcription Buffer

400mM Tris-HCl pH7.5

60mM $MgCl_2$

20mM Spermidine (Sigma)

Probe Elution Buffer

0.5M ammonium acetate

1mM EDTA

0.2% (w/v) SDS

5 x Hybridisation Buffer

2M NaCl

200mM PIPES pH6.4 (piperazine-N,N'-bis[2-etanesulphonic acid], Sigma)

10mM EDTA

2.12 X-gal Staining of Gene Trap Cell Lines to Detect lacZ Fusion Protein

Cells were rinsed in PBS, then fixed for 5 minutes in X-gal fix. The fix was removed, and the cells rinsed 3 times for 5 minutes in X-gal wash. Enough X-gal stain to cover the cell monolayer was added, and staining left to proseed at 37°C overnight. The staining reaction was stopped by further washing, and the samples stored in X-gal fix.

X-gal Fix

Gluteraldehyde (Sigma)	0.2% (w/v)
Phosphate buffer	0.1M
$MgCl_2$	2mM
EGTA (pH8.0)	5mM

X-gal Wash

Phosphate buffer	0.1M
$MgCl_2$	2mM
Sodium Deoxycholate (Sigma)	0.1% (w/v)
Nonidet P-40	0.02% (v/v)
BSA	0.05% (w/v)

X-gal Stain

X-gal (Sigma)	25mg
(5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside)	
dissolve in dimethy formamide (Sigma)	0.5ml
X-gal wash	25ml
$K_3Fe(CN)_6$ (Sigma)	41mg
$K_4Fe(CN)_6$ (Sigma)	52.5mg
0.085% (w/v) NaCl	0.4ml

1M Phosphate Buffer

Add 33 parts of 1M Na_2HPO_4 to 77 parts of 1M NaH_2PO_4 to give pH7.3.

2.13 Whole Mount *in situ* Hybridisation of Gene Trap Cell Lines

Digoxigenin Riboprobe Synthesis

The following components were mixed together at room temperature:

1-3µg linearised template DNA

10µl 5 x NTP mix

5µl 10 x Transcription buffer

50 units RNase inhibitor (Boehringer)

50 units of appropriate phage RNA polymerase

DEPC water to 50µl total volume

The labelling reaction was allowed to proceed for 2 hours at 37°C. The probe template was then digested by adding 20 units of DNase (RNase free, Boehringer) and incubating at 37°C for a further 15 minutes. 4µl of 250mM EDTA was added to stop the reaction, and the probe RNA precipitated with 5.5µl of 4M Lithium Chloride (BDH) and 165µl of ethanol at -70°C for 30 minutes. The precipitated probe was then pelleted by centrifugation at 4°C, washed in 70% (v/v) ethanol, dried briefly under vacuum and resuspended in 100µl of DEPC treated water. Probes prepared in this way were then run on a 0.8% (w/v) agarose gel to check for size and degradation.

Preparation of cells.

Embryonic stem cells and fibroblasts were seeded on gelatin coated glass coverslips in 6 well plastic dishes at a density of 2 x 10^5 cells per well (previously unscreened clones were grown in 24 well dishes, as described

elsewhere). They were grown in Glasgow Modified Eagles Medium (GMEM) suplemented with 10% (v/v) foetal calf serum, DIA (ES cells only) and G418 at 200μg/ml for ES cells and 400μg/ml for fibroblasts (electroporated lines only).

The cells were then rinsed at room temperature with DEPC PBS and fixed in 4% (w/v) paraformaldehyde/DEPC PBS for one hour at room temperature, or overnight at 4°C. The fixative was removed by rinsing twice on ice with PBT (PBS with 0.1% (v/v) Tween-20, Sigma) and the cells dehydrated through a methanol/PBS series (25%, 50%, 75%, 100% v/v). The dishes were then sealed and stored at -70°C until required.

The whole mount *in-situ* hybridisation was then carried out according to the following protocol:

1) Post-Fixation and Hybridisation.

1) Rehydrate cells on ice through methanol series to PBS (5 minutes each step)

2) Rinse cells 3 times with PBT at room temperature.

3) Wash for 30mins in 1% (v/v) triton X100 (Sigma) in PBS.

4) Fix cells for 15 minutes in 4% (w/v) paraformaldehyde in PBT with 0.2% gluteraldehyde added just before use.

5) Wash 3 x 5 mins in PBT

6) Wash cells with 1:1 hybridisation buffer: PBT

7) Wash cells with hybridisation buffer.

8) Add hybridisation buffer containing 10μg/ml tRNA plus 10μg/ml herring sperm DNA (Boehringer) to the wells for pre-hybridisation.

9) Place the dishes of cells in a perspex box humidified with 50:50 formamide (BDH): water. Seal the box with tape.

10) Place at 70°C for 1-5 hours.

2) Hybridisation

1) Split hybridisation buffer into the amounts needed for each seperate probe.

2) Aliquot amounts of the probes to give concentrations of 1:100 to 1:200 for each when added to the hybridisation buffer.

3) Denature the probes for 10 mins at 80°C, cool on ice then add to the aliquots of hybridisation buffer.

4) Add the probes to the cells, return them to the perspex boxes, seal with tape and incubate at 70°C overnight.

3) Antibody Conjugate Binding

1) Wash cells for 5 mins at 65°C with wash buffer

2) Wash cells 3 x 30 minutes with wash buffer at 65°C.

3) During these washes, heat inactivate sheep serum (Gibco) at 65°C for 30 minutes. Dilute with 1 x TBST to make a 10% (v/v) solution for blocking and a 1% (v/v) solution for antibody binding.

4) Allow cells to cool to room temperature.

5) Rinse cells 3 times with TBST.

6) Add 10% (v/v) sheep serum in TBST to the cells, and leave at room temperature for about 1hr (blocking step)

7) Dilute anti-DIG alkaline phosphatase conjugated Fab fragments (Boehringer) in 1% sheep serum in TBST to give a concentration of 1:2000.

8) Remove the 10% serum from the cells and add the antibody solution.

9) Return the plates to the perspex boxes, this time humidified with water. Incubate at 4°C overnight.

4) Colorimetric Detection

1) Wash cells 3 x 5 minutes with TBST at room temperature.

2) Wash for at least 2hrs using at least 3 changes of TBST.

4) Wash cells 3 x 10 minutes in alkaline phosphatase buffer.

5) Make up the stain in a foil-wrapped tube by adding 4.5µl of NBT and 3.5µl of BCIP (X-phosphate) per ml of alkaline phosphatase buffer.

6)Remove the final wash and replace with stain.

7) Cover the dishes with foil and leave at room temperature to allow stain to develop.

8) Stop the reaction by rinsing 3 x in PBT with 1mM EDTA. Cells can be stored in this at 4°C short term.

10 x Transcription buffer

400mM Tris-HCl pH 8,0

60mM $MgCl_2$

100mM DTT

20mM Spermidine

100mM NaCl

5 x DIG NTP Mix

5mM ATP

5mM CTP

5mM GTP

3.25mM UTP

1.75mM DIG-11-UTP (Boehringer)

Hybridisation Buffer

50% (v/v) Ultrapure Formamide (Gibco)

5 x SSC pH4.5 (DEPC treated)

5µg/ml Heparin (Sigma)

0.1% (v/v) Tween-20

Post- Hybridisation Wash Buffer

2 x SSC

50% (v/v) Ultrapure Formamide

0.1% (v/v) Tween-20

10 x TBST

NaCl	8g
KCl	0.2g
1M Tris-HCl pH7.5	25ml
Tween-20	1ml

ddH2O to 100ml final volume.

Alkaline Phosphatase Buffer

NaCl	100mM
$MgCl_2$	50mM
Tween-20	0.1% (v/v)
Tris-HCl pH9.5	100mM

For stain, add 4.5ml of NBT (Nitroblue tetrazolium chloride, 75mg/ml stock, Gibco) and 3.5ml of BCIP (5-bromo-4-chloro-3-indolyl phosphate-p-toluidine salt, 50mg/ml stock, Gibco) per 1ml of buffer.

Glass coverslips, where used, were mounted on glass microscope slides (Chance Propper Ltd) using Mowiol mounting medium (Harlow Chemical Co. Ltd.). For cells with indistinct morphologies, Hoescht 33258 was added at a final concentration of 1μM to counterstain nuclear DNA.

2.14 Flourescent *in situ* Hybridisation (FISH)

Sample Preparation

4×10^6 cells were seeded into a $25cm^2$ flask and grown overnight, to give a 50% confluent culture. The medium was removed and replace with fresh GMEM containing 0.05mg/ml of colcemid (Gibco). The cells were cultured for a further 1.5 hours, then harvested and chromosomes prepared from them according to the following protocol:

1) Remove the growth medium and replace with 10ml of PBS.

2) Gently wash the cells from the growth surface into the PBS by pipetting.

3) Pellet the cells by centrifugation at 1000rpm for 5 minutes. Discard the supernatant.

4) Resuspend the cell pellet in 5ml of hypotonic lysis buffer and leave at room temperature for 15 minutes.

5) Pellet the cells, and resuspend in 5ml of 3:1 methonol:acetic acid fix. Add the fix slowly, with gentle vortex mixing to prevent the cells from forming clumps.

6) Pellet the cells, and resuspend in a further 5ml of fix. Store at -20°C overnight.

7) Pellet the cells, and resuspend in sufficient fix to give an opaque suspension.

8) Using a pasteur pipette, drop the cell suspension onto clean, acid/alcohol treated glass microscope slides. Breathe on the slides to moisten them before dropping the cells.

9) Allow the slides to air dry, then leave overnight at room temperature to 'age' them.

10) Store the slides in a dessicator at room temperature for 2-7 days before using them for in-situ hybridisation.

Probe Labelling

This labelling protocol is designed for use with cosmid templates, but also works well for the plasmid and λ templates used in this study.

Mix together:

1µg DNA template
4µl 10 x NTS
4µl 2mM dATP
4µl 2mM dGTP
4µl 2mM dCTP
2µl 0.5mM dTTP
4µl 1mM biotin-16-dUTP or
1mM digoxigenin-11-dUTP (Boehringer)
1µl DNA pol I (Gibco)
2µl fresh 1:500 dilution DNase I (Gibco)

Incubate for 90 minutes at 16°C.

Add TE to give a final volume of 100µl, and spin through a G-50 sephadex column to remove unincorporated nucleotides.

Assessment of Probe Quality

1) Wash a gridded circular nitrocellulose filter in ddH_2O, soak in 20 x SSC for 10 minutes and air dry.

2) Make 10^{-2} and 10^{-3} dilutions of the probe reactions and spot 1µl and 2µl samples of each onto the filter. Use 1, 2, 10, and 20pg standards of appropriately labelled lambda DNA as a control.

3) UV cross link the filter in a Stratalinker (Stratagene) using the 250mJ programme.

4) Wash the filter for 5 minutes in buffer 1.

5) Incubate for 30 minutes at 37°C in buffer 1 with 3% BSA fraction V (Sigma).

6) Incubate in 10ml of buffer 1 containing 10µl of streptavidin-alkaline phosphatase (for biotin labelled probes) or antidigoxigenin-alkaline phosphatase (for digoxigenin labelled probes) for 30 minutes at room temperature.

7) Wash 2 x 15 minutes in 200ml of buffer 1.

8) Wash 5 minutes in buffer 3.

9) Place filter in a hybridisation bag. Add 5ml of buffer 3 containing 2 drops each from bottles 1, 2 and 3 of the Vector NBT/BCIP alkaline phosphatase detection kit (VectorLabs). Incubate in the dark for 2-4 hours to allow the stain to develop.

The concentration of the probes can then be estimated by comparison to the standards. Probes of 1ng/ml or greater are acceptable.

Hybridisation

1) Dispense 50ng of each probe per slide. Add 3µg of Cot I DNA (Gibco BRL) and 5µg of salmon sperm DNA per slide. Add 2 volumes of ethanol and spin vac to pellet and dry the probe.

2) Resuspend the probe mixture in 10µl of hybridisation mix per slide and leave to dissolve at room temperature.

3) Incubate the aged slides in 2 x SSC containing 100µg/ml RNase at 37°C for 1 hour.

4) Wash briefly in 2 x SSC.

5) Dehydrate through fresh 70%, 90% and 100% (v/v) ethanols for 2 minutes each, then dessicate under a vacuum for 10 minutes.

6) Warm the slides in a 70°C oven for 5 minutes, then denature in pre-warmed 70% (v/v) formamide, 2 x SSC for 3 minutes at 70°C.

7) Transfer quickly to ice-cold 70% (v/v) ethanol for 2 minutes, then dehydrate through 90% and 100% (v/v) ethanol and dry under vacuum as before.

8) While the slides are drying, denature the probe and competitor DNAs in hyb. mix at 70°C for 5 minutes.

9) Transfer the probes to a 37°C water bath to pre-anneal for 15 minutes.

11) While the probes are annealing, warm the slides and some 22 x 22mm coverslips on a 37°C hotplate.

12) Pipette the 10µl of pre-annealed probe onto a coverslip and pick up the coverslip with the prepared slide. Gently expel any bubbles and seal the coverslip onto the slide with rubber cement (TipTop). Incubate in a covered tray in a 37°C waterbath overnight. From this stage onwards, the slides must not be allowed to dry out.

Slide Washing and Detection

1) Make up 10ml of blocking buffer and use this to make dilutions of antibodies and conjugates. Spin at 4°C for 15 minutes to pellet any clumps of antibody. For simultaneous detection of digoxigenin labelled probes with FITC and biotin labelled probes with texas red, the following reagents are required:

a) Anti-dig FITC (sheep)	1:20 dilution
b) Avidin-texas red	1:500
c) Anti-sheep FITC (rabbit)	1:100
d) Anti-avidin biotin (goat)	1:100

All of the above reagents are supplied by Vector.

2) Peel the rubber solution off the coverslips but do not remove the coverslips. Put the slides into a glass rack.

3) Wash 4 x 3 minutes in 50% (v/v) formamide, 2 x SSC at 45°C, agitating occasionally to cause the coverslips to fall off.

4) Wash 4 x 3 minutes with 2 x SSC at 45°C.

5) Wash 4 x 3 minutes with 0.1 x SSC at 60°C.

6) Transfer the slides to 4 x SSC, 0.1% (v/v) Tween-20.

7) Incubate each slide with 40µl blocking buffer under a 22 x 40mm coverslip for 5 minutes at room temperature.

8) Working with one slide at a time, remove the coverslips and drain excess blocking buffer from the slide. Place 40µl of the first antibody mixture on a coverslip and pick up with the slide. Incubate in a moistened chamber at 37°C for 30-60 minutes.

9) Remove the coverslips and wash the slides 3 x 2 minutes in 4 x SSC, 0.1% (v/v) tween-20 at 37°C.

10) Add the next antibody, incubate and wash as above.

Potential cross-reactivity between antibodies must be taken into consideration when deciding on the order in which to apply the antibodies. In this study, the following order of reagents was employed:

a) FITC anti-digoxigenin and avidin-texas red.

b) FITC anti-sheep.

c) Biotin anti-avidin.

d) Avidin-texas red.

The chromosomes were then counterstained by the addition of 50µl of AFT10 mounting medium (Citiflour) containing 2µg/ml DAPI. A 22mm x 22mm coverslip was applied, and excess mounting medium squeezed out. The coverslip was sealed in place using rubber cement (Pang).

Image Capture and Analysis

Slides were viewed using a Zeiss Axioplan flourescence microscope with a triple band pass filter set (Chroma) to allow sequential visualisation of FITC, Texas red and DAPI images using a computer driven excitation filter wheel. Image registration was perfect, as the polychroic filter and emission filter remained in place while aquiring all three images. Metaphase spreads and interphase nuclei were imaged used a cooled CCD camera fitted with a KAF 1400 chip (Photometrics). Seperate images of the two probe signals and counterstain were pseudocoloured and merged using an Apple Mackintosh Quadra 900 computer with software developed by Digital Scientific. Background hybridisation was removed by normalisation and removal of the lowest intensity pixels. Care was taken not to interfere with chromosome hybridisation signals.

Hypotonic Lysis Buffer

0.25% (w/v) Potassium Chloride

0.5% (w/v) Tri-sodium Citrate

10 X NTS

0.5M Tris-HCl pH7.5

0.1M $MgSO_4$

1mM DTT

0.5mg/ml BSA fraction V

Buffer 1

0.1M Tris-HCl pH7.5

0.15M NaCl

Buffer 3

0.1M Tris pH9.5

Hybridisation Mix

50% (v/v) deionised Formamide

2 x SSC

1% (v/v) Tween-20

10% (w/v) Dextran Sulphate (diluted from frozen 50% (w/v) stock)

Blocking Buffer

4 x SSC

5% (w/v) Skimmed Milk Powder (Marvel)

2.15 Miscellaneous Methods

LB Agar Plate Preparation

A 1.5% (w/v) solution of bactoagar (Difco) in LB growth medium was made. Amplicillin was added at a concentration of 10µg/ml if required. Approximately 30ml of this solution was poured into for each 9cm dish, and the plates left to set at room temperature.

G-50 Sephadex Columns

A small plug of synthetic wool was placed in the tip of a 1ml syringe (Plastipak, Benton Dickinson Ltd). 10% (w/v) G-50 sephadex (Pharmacia) in TE was pipetted into the top of the syringe. The TE drained through to leave a solid column of G-50 sephadex. The column was centrifuged in a plastic centrifuge tube for 2.5 minutes at 100rpm. A microfuge tube was placed in the centrifuge tube, underneath the column and 100µl of sample was added to the top of the column. The column was re-centrifuged and the sample collected in the microfuge tube.

DEPC Treatment of Solutions for RNA Work

Diethyl Pyrocarbonate (DEPC, Sigma) was added to the solution to a final concentration of 0.1% (v/v). The solution was left in a fume cupboard overnight, then autoclaved to inactivate the DEPC before use.

2.16 General Solutions

6% Acrylamide Gel Mix

For 500ml

Urea	210.21g (7M final concentration)
40% (w/v) acrylamide (BDH)	71ml
2% (w/v) bis-acrylamide (BDH)	75ml
10 x TBE	25ml

For each 100ml of gel mix, 400µl of 10% (w/v) ammonium persulphate (Sigma) and 100µl of TEMED (N,N,N',N'-tetramethyl-ethylenediamine, Sigma) were added immediately proir to pouring the gel. Gels were run at 60W constant power using 0.5 x TBE buffer in a BRL model S2 sequencing gel apparatus.

LB Bacterial Growth Medium

Tryptone (Difco)	10g
Yeast Extract (Difco)	5g
NaCl	5g
$MgCl_2$	2g

ddH_2O to 1 litre final volume.

Autoclave before use.

Loading Dye For Electrophoresis (10 x Stock)

50% (v/v) Glycerol

0.4% (w/v) Bromophenol Blue (BDH)

0.4% (w/v) Xylene Cyanol (Sigma)

125mM EDTA

20 x SSC (pH7.0)

NaCl	876.5g
Tri-Sodium Citrate (BDH)	441g

ddH_2O to 5 litres final volume.

10 x TBE

Tris	108g
Boric Acid (BDH)	55g
EDTA	9.3g

ddH_2O to 1 litre final volume.

TE

10mM Tris-HCl pH7.5

1mM EDTA pH8.0

Unless otherwise stated, all molecular biology techniques were based on those described in Sambrook et al, 1989.

Chapter 3

The Use of Gene Trap Vectors to Assess the Involvement of 3' and 5' RNA Regions in the Sub-Cellular Localisation of Specific Transcripts.

INTRODUCTION

A number of mRNAs have been demonstrated to localise to distinct sub-cellular regions. Differential distribution of mRNAs within the cell has been studied most widely in the oocytes and early embryos of *Drosophila* and *Xenopus*. The *Xenopus* maternal mRNA, *vg-1*, which encodes a transforming growth factor homolog (Weeks and Melton, 1987) shows a homogeneous distribution in early oocytes, but becomes restricted to the vegetal cortex in middle and late stage oocytes. Microinjection of radiolabelled *vg-1* transcripts lacking the 5'UTR and initiation codon leads to their co-localisation with the endogenous transcript (Yisreali and Melton, 1988). This confirms that the localisation seen is a function of the transcript itself, rather than a result of translation of the *vg-1* protein. Further dissection of the *vg-1* 3'UTR has demonstrated that a 340 nucleotide region of the 3'UTR is required for the correct localisation of *vg-1* RNA and is sufficient to confer a *vg-1* like localisation on a reporter mRNA (Mowry and Melton, 1992). The mechanism of *vg-1* transport and anchorage is thought to involve cytoskeletal or cytoskeletal-associated proteins (Yisreali et al, 1990).

The localisation of maternal transcripts is also important in the early *Drosophila* embryo. A gradient of the Bicoid protein is necessary in the oocyte for the correct development of the head and thorax. The establishment of this gradient is preceeded by localisation of the *bicoid (bcd)*

mRNA. *Cis* acting RNA sequences have been demonstrated to be responsible for the anterior localisation of *bcd* RNA (MacDonald and Struhl, 1988). The genes *expurantia, swallow* and *staufen* have been identified as determinants of *bcd* RNA localisation by mutational analyses (St Johnston et al, 1989; Stephenson et al, 1988.), indicating that the localisation pathway of the anterior fate determinant, *bicoid*, involves a number of tightly controlled steps.

Localised maternal mRNAs are also important in the posterior development of the *Drosophila* embryo. Posterior localisation of *nanos* mRNA determines abdominal segmentation (Wang and Lehmann, 1991) while *oskar* mRNA is involved in the specification of the pole plasm (germ line) (Ephrussi et al, 1991). Again, a number of different genes have been shown to affect the localisaton of these mRNAs. Indeed, *oskar* is required for the correct localisation of *nanos* mRNA (Ephrussi et al, 1991), indicating that the pathways of localisation of messages involved in the patterning of the *Drosophila* embryo are closely interlinked.

Later in *Drosophila* development, at the early blastoderm stage, the pair-rule genes are expressed. The mRNAs for these genes are localised to the apical periplasm (ie below the nucleus). The mechanism of this localisation has been studied by the expression of hybrid transcripts between the pair-rule genes *fushi-tarazu, hairy* and *even-skipped* and the *β-galactosidase* reporter (Davis and Ish-Horowicz, 1991). These experiments mapped the signals governing the polarised distribution of the messages to short regions of sequence within the 3'UTRs.

The early gradients of morphogens established in *Xenopus* and *Drosophila* embryos appear over a time span of hours or days, reflecting the stability of these messages, and the complexity of the pathways by which these gradients are achieved (Kislauskis and Singer, 1992). By contrast, the messages of the later pair rule genes are highly unstable (Edgar et al, 1991) so their localisation is likely to be acheived by a different mechanism, probably by leaving the nucleus in a polarised fashion (Kislauskis and Singer, 1992). In both cases, however, the developmentally regulated localisation of specific transcripts is mediated by signals within the 3'UTRs of the messages themselves.

The localisation of specific mRNAs in the cytoplasm of somatic cells has also been studied by *in situ* hybridisation. The earliest work on somatic cells was a study of the mRNAs for cytoskeletal components in chick embryonic fibroblasts and myoblasts (Lawrence and Singer, 1986). Actin mRNA was shown to be localised to the leading edge of motile cells. Vimentin and tubulin were more peri-nuclear. In cells undergoing a response to injury of a confluent monolayer, actin mRNA is present in the lamellipodia of spreading cells (Hoock et al, 1991). Differential localisation of specific mRNAs has also been documented in a number of other differentiated cell types. In intestinal epithelium, actin mRNA localises to the apical end of the cell, where actin filaments polymerise to form the microvilli (Cheng and Bjerknes, 1989). In neurons, the mRNA for microtubule associated protein 2 (MAP-2) is found in the dendrites, while the transcripts for Gap-43 and α-tubulin are restricted to the cell bodies (Garner et al, 1988; Bruckenstein et al, 1990). Myosin mRNA has been shown to accumulate near sarcomeres in muscle (Pomeroy et al, 1991).

In each of the documented cases of sub-cellular mRNA localisation in somatic cells, the message appears to localise in regions of the cell where the product of the message is required. The use of protein synthesis inhibitors has demonstrated that, at least for actin mRNA, the transcript localisation occurs independantly of protein synthesis (Sundell and Singer, 1990). The sequences responsible for actin mRNA localisation have been mapped to the 3'UTR of the transcript (Kislauskis et al, 1993).

The use of drugs to disrupt the cytoskeleton suggest that microfilaments of the cytoskeleton are required for the correct localisation of actin mRNA (Sundell and Singer, 1991). Actin filaments are thought to be involved in the anchoring of transcripts once their localisation has been acheived (Singer, 1992). A study of the sub-cellular localisation and transport of myelin basic protein (MBP) mRNA in living oligodendrocytes (Ainger et al, 1993) demonstrated the accumulation of the mRNAs for MBP, actin and globin as cytoplasmic granules. These granules were relatively uniform in size (about 0.3m) and, in the case of MBP mRNA, were demonstrated to be motile and associated with the cytoskeletal matrix. Similar granules were also detected in neuroblastoma cells, suggesting that they are not specific to oligodendrocytes. Granular sub-cellular localisation has also been reported for Vimentin mRNA in myoblast, myotubes and fibroblasts (Cripe et al, 1993). The observation of these granules has been explained as a visualisation of the stuctures involved in the localisation of transcripts (Ainger et al, 1993). These results have lead to the proposal of a model involving the formation of ribonucleoprotein particles (RNPs) to account for the sub-cellular localisation of certain transcripts (Wilhelm and Vale, 1993). It appears, therefore, that the localisation of specific transcripts is

acheived, at least for some of the transcripts, by the transport of the RNA as RNP particles in association with the cytoskeleton.

The observation of sub-cellular RNA localisation in somatic cells, combined with the results from developmental systems implicate RNA localisation as a general mechanism for creating asymmetric distributions of proteins in the cytoplasm. In all cases where the signals responsible for sub-cellular RNA localisation have been mapped, they lie within the 3'UTR of the transcript. No signals determining the sub-cellular distribution of endogenous messages have been mapped to the 5'UTRs of genes, but there is no published exploration of the possibility that localisation signals may lie in these regions.

The work presented in this chapter uses a gene trap approach, combined with whole mount *in situ* hybridisation to assess the involvement of sequences in the 5' and 3' regions of endogenous genes in the formation of sub-cellular patterns of transcript localisation. The gene trap vector pGT1.8K (figure 3.1A) was used to trap the 5' regions of endogenous genes in mouse ES cells. Mouse embryonic stem cells were used for the initial large scale screen of the variety of RNA localisation patterns that can be obtained using 5' gene trap vectors because of their rapid growth characteristics. The sub-cellular localisations of the resultant fusion transcripts were determined in a large number of colonies of cells containing gene trap integrations in order to investigate the involvement of 5' sequences in the control of transcript distribution. A strategy using the gene trap vector pGT1.8βgeo (figure 3.1B) in 10T1/2 fibroblast cells was adopted to investigate any correlation between the distribution of fusion transcripts containing the 5' regions of

endogenous genes and that of the fusion proteins translated from them. Fibroblasts were chosen for this study as their larger size and more flattened morphology allow a more detailed analysis of sub-cellular localisation patterns.

To investigate the involvement of 3' sequences in transcript localisation, a gene trap vector was modified to form fusion transcripts containing the 3'UTR regions of endogenous genes and the neomycin coding region. This vector was used for electroporation into ES cells in order to assess the range of sub-cellular RNA localisations that can be dictated by elements within the 3'UTR.

The results presented document a range of patterns of sub-cellular fusion transcript localisation. In most of the clones the fusion transcripts were distributed widely within the cytoplasm, suggesting that sub-cellular localization of transcripts is not of importance for the majority of mammalian genes. Two particularly striking RNA localisation patterns were observed, however. A small number of clones were obtained using the 5' gene trap vectors which showed an accumulation of fusion transcript within the nucleus of the cell. No nuclear localised clones were seen with the 3' trap vector, but a large number of clones electroporated with this vector showed a characteristic single dot of transcript accumulation in each cell. The significance of these localisations is discussed in this and following chapters.

Electroporation of Embryonic stem Cells

A) 5' Gene Trap Vectors

150µg of the plasmid pGT1.8K (figure 3.1A) was digested overnight with HindIII and resuspended in 100µl of sterile PBS. CGR8 cells were cultured as described in chapter 2 and 10^8 cells were electroporated with the vector using settings of 800V, 3µF. The electroporated cells were plated onto ten 10cm diameter tissue culture dishes, and maintained in 200µg/ml G418 until colonies had formed. 10^7 non-electroporated cells were also plated and maintained in G418 to ensure that the selection procedure was complete. The plates of colonies were then fixed for one hour in 4% paraformaldehyde in PBS and taken through the whole mount *in situ* procedure.

B) 3' Trap Vector

Digestion of 150µg of the plasmid pKnSD with EcoRI and HindIII was carried out overnight at 37°C (Figure 3.2). The parental cell line, CGR8 was cultured as described and 10^8 cells were electroporated with the vector using settings of 3µF, 800V. The electroporated cells were plated onto ten 10cm diameter tissue culture dishes, and maintained in 200µg/ml G418 until colonies had formed and all non-electroporated cells on the control plate had been killed. The G418 resistant colonies were then picked using a mouth pipette, and cultured in 24 well plates until confluent. The confluent clones were trypsinised and split into two sets of 24 well plates. When approximately 50% confluent, one set of plates was frozen at -80°C, and the other fixed in 4% (w/v) paraformaldehyde and taken through the whole mount *in situ* procedure as described in chapter two

Establishment of Electroporated Fibroblast Cell Lines.

150μg of the plasmid pGT1.8βgeo was linearised with HindIII and resuspended in 100ml of sterile PBS (Figure 3.1B). 10T1/2 fibroblast cells were cultured as described and 10^8 cells were electroporated with pGT1.8βgeo using conditions of 250V, 250μF. The electroporated cells were plated onto ten 10cm diameter tissue culture dishes and selected in 400μg/ml G418. A control dish was seeded with 10^7 non-elctroporated cells and maintained in 400μg/ml G418 to ensure that selection was complete.

The electroporated cells formed semi-confluent mono-layers rather than individual colonies, so the cells were harvested by trypsinisation and frozen down as ten pools of cells, each representing one dish from the electroporation. Pools 4, 9, 18, 19 and 20 were then thawed and replated at a lower density. Individual colonies were picked from each pool using a mouth pipette and grown to confluence in 24 well dishes. Once confluent, these clones were trypsinised and split into three separate sets of 24 well plates. One set of plates was fixed in 4% (w/v) paraformaldehyde in PBS and put through the whole mount *in situ* procedure using a *lacZ* riboprobe, one set was stained with X-gal to determine the localisation of the fusion protein. Both procedures are detailed in chapter two. The third set of plates was maintained in culture until the RNA and protein localisation studies were complete. Once this information was available, any clones showing potentially interesting patterns of either protein or RNA localisation were expanded into $25cm^2$ dishes and frozen down for further study at a later date. Further details of these culture procedures are given in chapter two.

Analysis of Sub-Cellular RNA Patterns

Probes

Digoxigenin labelled probes were synthesised as described in chapter two. The anti-sense neomycin probe was transcribed using T7 bacteriophage RNA polymerase, from the plasmid pNeoBlue, which comprises a 1Kb fragment of the neomycin gene cloned into Bluescript IISK (promega, figure 3.3A). The plasmid was first linearised by digestion with SacI. The α–Actin probe was transcribed using SP6 polymerase using the plasmid pActin, linearised at the EcoRI site as a template (figure 3.3B). This probe contains the coding region of the gene, and so will detect cytoplasmic βactin in addition to α (cardiac) actin. Each probe was used at a 1:200 dilution. The anti-sense riboprobe to *lacZ* was transcribed using T7 bacteriophage polymerase using the plasmid pΔEK, linearised at the SacI site as a template (figure 3.3C). The probe was used at 1:500 dilution.

Whole Mount *in situ* Hybridisation

The whole mount *in situ* protocol used was based on a protocol optimised for mouse embryos (Rosen and Beddington, 1993). Alterations were made to the fixation and permeabilisation conditions in this protocol to ensure that the cellular architecture remained intact to allow sub-cellular RNA distribution to be monitored. At the same time, it was important to maintain the high signal to background ratio obtained with the original protocol. The length of fixation with 4% (w/v) paraformaldehyde was cut to one hour from the overnight fixation employed in the original protocol. The permeabilisation of 3X 20 minutes with RIPA (a mixture of ionic and non-ionic detergents) was also replaced with a 30 minute permeabilisation with 1% (v/v) triton X 100 in PBS.

Figure 3.1: 5' Gene Trap Vectors.

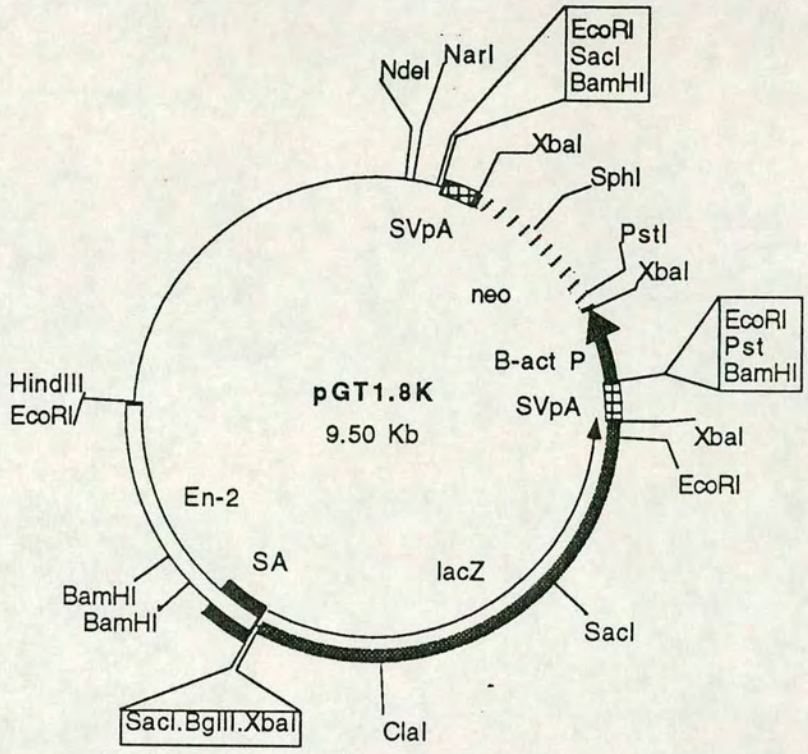
A) pGT1.8K (Constructed by W. C. Skarnes)

Mouse *En-2* genomic DNA (white box) consisting of a 5' 1.8Kb EcoRI/BglII fragment including the homeobox-containing exon (Joyner and Martin, 1987) is joined in frame to the *lacZ* gene (pMC1871; Pharmacia, black box). The *En-2* DNA includes a splice acceptor site (SA). The human β-actin promotor (black arrow) is present as a SstI/Sau3A fragment (Frederickson et al, 1989) driving a 1Kb BglII/SmaI fragment of the *neomycin* gene (stiped box), containing the entire coding region (Southern and Berg, 1982). The *neomycin* gene has been modified to contain a Kozak consensus sequence at the start of translation (W.C.Skarnes, unpublished). Polyadenylation signals (checked boxes) are obtained from the vector pECE (Ellis et al, 1986) and contain the BclI(2770)/BamHI(2533) fragment of the SV40 genome.

B) pGT1.8βgeo (Constructed by W. C. Skarnes)

This plasmid was made by replacement of the ClaI/SphI fragment of pGT1.8K with a ClaI/SphI fragment from the vector pSAβgeo (Friedrich and Soriano, 1991). This fragment replaces the independent *lacZ* and *neomycin* genes with the *βgeo* fusion of *lacZ* and *neomycin*.

Both vectors are linearised with HindII prior to electroporation.


NdeI
NarI
EcoRI
SacI
BamHI
XbaI
SphI
SVpA
PstI
XbaI
neo
EcoRI
Pst
BamHI
B-act P
SVpA
XbaI
EcoRI
HindIII
EcoRI
pGT1.8K
9.50 Kb
En-2
SA
lacZ
SacI
BamHI
BamHI
SacI.BglII.XbaI
ClaI

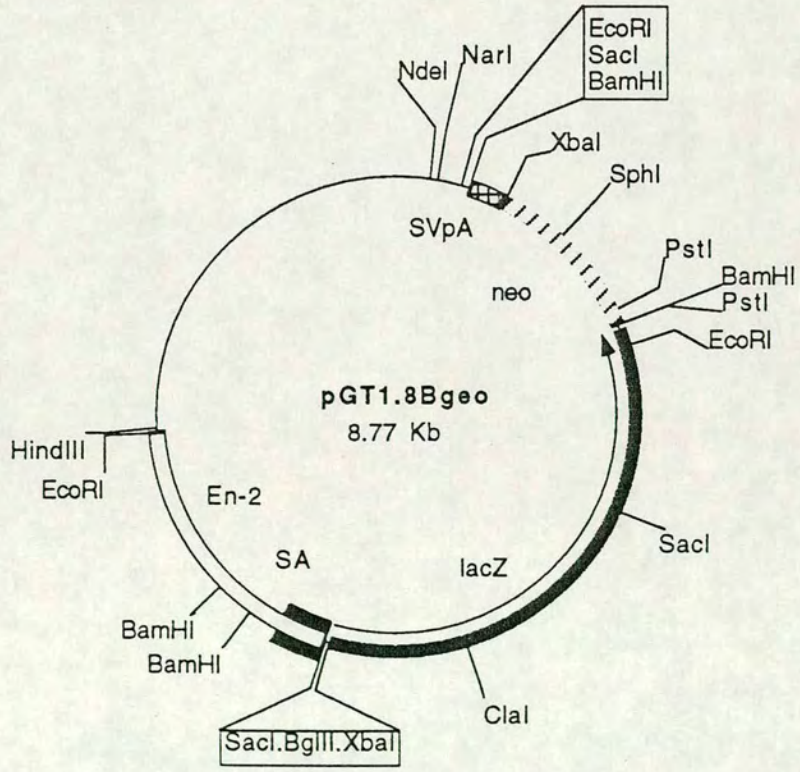

NdeI
NarI
EcoRI
SacI
BamHI
XbaI
SphI
SVpA
PstI
BamHI
PstI
neo
EcoRI
pGT1.8Bgeo
8.77 Kb
HindIII
EcoRI
En-2
SacI
SA
lacZ
BamHI
BamHI
ClaI
SacI.BglII.XbaI

Figure 3.2: 3' Trap Vector, pβKnSD (Constructed by W. C. Skarnes)

The human β-actin promoter (black arrow) is present as a SstI/Sau3A fragment (Frederickson et al, 1989) driving the *neomycin* gene (striped box), as before. The last 15 base pairs of the first exon and the entire first intron of the *β-globin* gene (white box) are fused in frame to the *neomycin* gene. This vector was made by replacement of an SphI/EcoRI fragment from the vector pβKnpA (W. C. Skarnes, unpublished), which comprises the human *β-actin* promoter driving the entire *neomycin* gene, by an SphI/EcoRI fragment from the vector pPGKneoβ, which contains the *neomycin/βglobin* fusion.

This vector is cut with HindIII and EcoRI to remove plasmid sequences prior to electroporation.

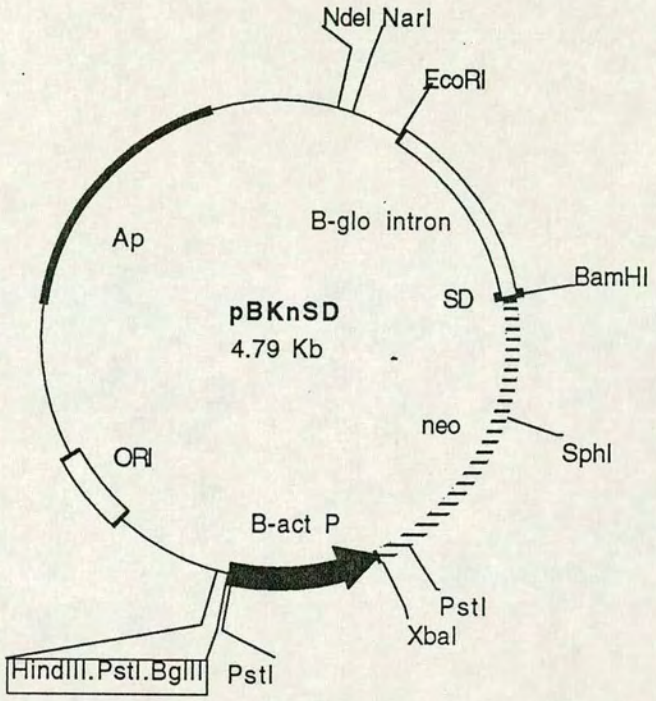

NdeI NarI
EcoRI
B-glo intron
Ap
BamHI
SD
pBKnSD
4.79 Kb
neo
SphI
ORI
B-act P
PstI
XbaI
HindIII.PstI.BglII
PstI

Figure 3.3: Templates for Digoxigenin Labelled Riboprobes.

A) pSAβgeoΔEK.

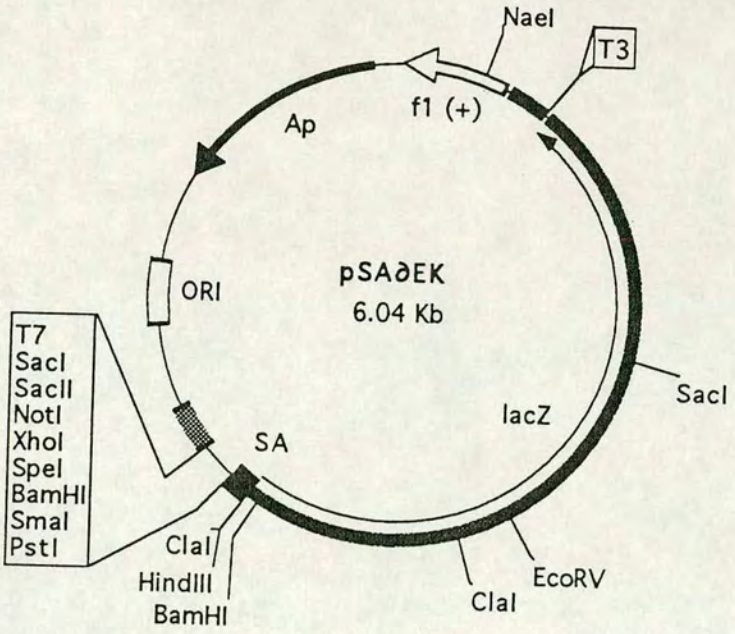
This plasmid is a KpnI/EcoRI deletion from the vector pSAβgeo (Friedrich and Soriano, 1989) which removes the *neomycin* region from the *βgeo* fusion. Linearisation of this plasmid with SacI provides a template for an anti-sense *lacZ* riboprobe, transcribed by T3 RNA polymerase.

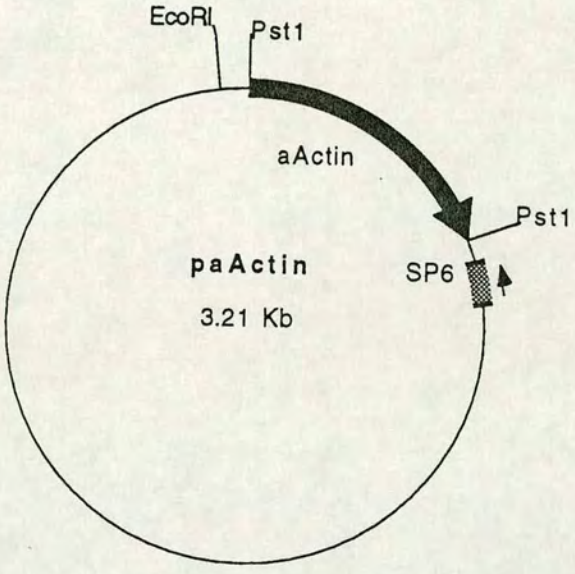
B) pα-actin.

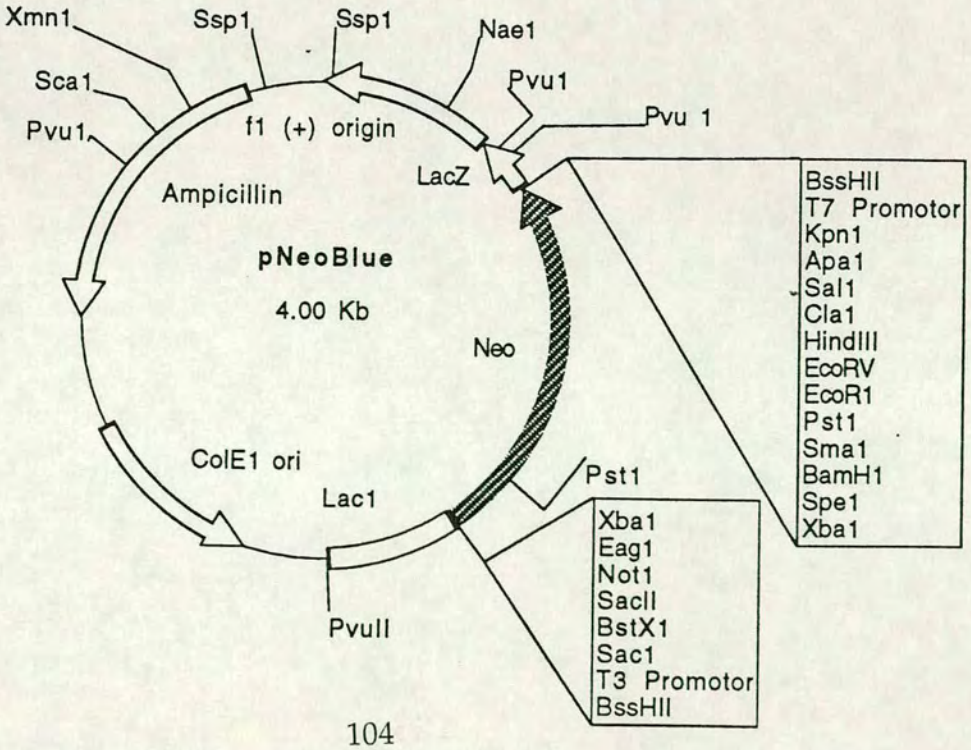
This plasmid contains a 601bp PstI fragment of *α–actin*, cloned in pGEM. Linearisation with EcoRI provides a template for transcription of an anti-sense riboprobe.

C) pNeoBlue.

This plasmid contains a 1Kb XbaI fragment of *neomycin* (Southern and Berg, 1982), cloned into Bluescript II SK+. Linearisation with SacI provides a template for transcription of an anti-sense riboprobe.


NaeI
T3
f1 (+)
Ap
ORI
pSA∂EK
6.04 Kb
SacI
lacZ
SA
T7
SacI
SacII
NotI
XhoI
SpeI
BamHI
SmaI
PstI
ClaI
HindIII
BamHI
ClaI
EcoRV


EcoRI
Pst1
aActin
Pst1
paActin
3.21 Kb
SP6


Xmn1
Ssp1
Ssp1
Nae1
Sca1
Pvu1
Pvu1
f1 (+) origin
Pvu 1
LacZ
Ampicillin
pNeoBlue
4.00 Kb
Neo
ColE1 ori
Lac1
Pst1
PvuII
BssHII
T7 Promotor
Kpn1
Apa1
Sal1
Cla1
HindIII
EcoRV
EcoR1
Pst1
Sma1
BamH1
Spe1
Xba1
Xba1
Eag1
Not1
SacII
BstX1
Sac1
T3 Promotor
BssHII

RESULTS

Gene trap vectors are designed to produce fusion transcripts containing sequences from the endogenous gene at the site of insertion of the vector spliced to reporter sequences, leading to mutation of the endogenous gene and production of a tagged transcript under the control of the endogenous promotor. Translation of this fusion transcript into a fusion protein occurs only if the structure of the vector integration site satisfies certain criteria detailed below. Standard gene trap screens rely on the production of an active reporter protein (lacZ) to identify gene trap lines carrying a mutation. The idea to use gene trap vectors to study sub-cellular fusion transcript distribution arose from experiments carried out to investigate the frequency of gene trap events producing detectable fusion transcript, but no detectable fusion protein. Mouse ES cells were electroporated with the gene trap vector pGT1.8K and the resultant colonies were used for whole mount *in situ* hybridisation using a *lacZ* probe. In this experiment, a number of different sub-cellular patterns of fusion transcript distribution were observed (W. C. Skarnes, unpublished) including accumulation of fusion transcripts in the nucleus and in single large dots within the cytoplasm. These distributions were unexpected, so the work presented in this chapter was initiated both to quantitate the frequency of occurence of different patterns of fusion transcript localisation and to investigate further the nature of these patterns.

Gene Trap Vectors GT1.8K and GT1.8βgeo

Both of the vectors used contain a promotorless *lacZ* gene, downstream of a splice acceptor region from the mouse *engrailed-2 (En-2)* gene. This region contains the last 1762 base pairs of an *En-2* intron including the branch

point and the first 120 base pairs of the following exon. All cells containing an insertion of the vector pGT1.8K will be neomycin resistant, due to the presence of the neomycin gene driven by the β-actin promotor. Cells in which the vector is inserted within the intron of a pol II transcription unit in the correct orientation are also predicted to form a fusion transcript between endogenous and *lacZ* sequences. Provided that the reading frame of the endogenous gene is the same as that of the *lacZ* gene, a fusion protein is produced containing 5' regions of the endogenous gene and an active *lacZ* (figure 3.4). Using this vector, 1% of neomycin resistant colonies produce an active fusion protein. In pGT1.8βgeo, in contrast to pGT1.8K, the *lacZ* gene is present as a fusion with neomycin (*βgeo*). Thus, only cells containing insertions in the correct orientation and reading frame will be neomycin resistant, leading to a considerably reduced background of neomycin resistant but *lacZ* negative colonies (figure 3.5).

In cells containing an integration of either vector, the localisation of the fusion RNA can be visualised by whole mount *in situ* hybridisation using probes to detect *lacZ* or *neomycin* sequences. The distribution of active fusion proteins can be monitored using X-gal staining.

Modified 3' Trap Vector pβKnSD

Since all of the mRNA localisation signals identified so far lie within the 3'UTRs of the transcripts, a modified gene trap vector was designed to form fusions containing the 3' regions of endogenous genes spliced to a neomycin reporter gene. The vector used contains the β–actin promotor driving a neomycin selectable marker lacking a polyadenylation signal, linked to the last 15 base pairs of the first exon and the entire first intron of the β–globin

gene. This intron/exon boundary contains a splice donor site. To produce a stable fusion mRNA, the vector must integrate within an endogenous gene. Splicing must occur between the β–globin splice donor and a splice acceptor from a downstream exon of the gene into which the vector has inserted. The fusion transcript should contain the entire neomycin gene, fused to 3' exons and the 3'UTR of the endogenous gene (figure 3.6). Any mRNA localisation signals contained in the 3'UTR should now determine the sub-cellular localisation of the neomycin fusion transcript which can be detected by whole mount *in situ* hybridisation. A similar approach has recently been used by Yoshida et al (1995) using a PGK promotor to drive a neomycin gene terminating in a splice donor site. 3' RACE cloning from lines electroporated with this vector suggested that the vector was functioning as predicted and led to the isolation of a novel gene.

Validity of the Whole Mount *in situ* Protocol for the Study of Sub-cellular RNA Localisation.

The use of an actin riboprobe on the control cell line, CGR8, shows the already well documented sub-cellular distribution of cytoplasmic actin, with concentrations of signal in the leading processes of well spread, motile cells. This confirms the suitability of this whole mount in-situ protocol for the investigation of sub-cellular RNA localisation. At the same time, the use of a neomycin probe on line CGR8 gave no background staining after an overnight staining reaction (figure 3.7), indicating that the changes made to the protocol had not affected the low level of background staining seen with the original protocol.

Sub-cellular Localisation of *LacZ* Fusion Transcripts in Mouse Embryonic Stem Cells Electroporated with pGT1.8K.

To assess the frequency with which each of the previously observed patterns of fusion transcript localisation occurs, a large scale screen of mouse ES cells electroporated with the gene trap vector pGT1.8K was undertaken. The electroporation of 10^8 ES cells with 100μg of the vector pGT1.8K gave rise to approximately 3000 colonies. *LacZ* fusion transcripts were present at detectable levels in 19% of the colonies analysed. This is in marked contrast to the 1% of colonies observed to produce detectable fusion protein in previous screens (Gossler et al, 1989 and W. C. Skarnes, unpublished). The large number of colonies synthesising detectable levels of fusion transcript, but no detectable levels of fusion protein presumably represent integrations producing improperly spliced or non-translated RNAs, perhaps as a result of insertion of the vector immediately downstream of a 5' untranslated exon, or RNAs encoding non-functional β-gal proteins. It is known, for example, that fusions containing the signal sequence of an endogenous gene in the absence of a transmembrane domain are enzymatically inactive (Skarnes et al, 1995). From this result, it is clear that the major proportion of gene trap insertions producing a fusion transcript do not produce detectable fusion protein due to a number of RNA processing or translational constraints. Gene trap events that do not produce detectable fusion protein are missed by conventional gene trap screens which use X-gal staining for protein activity to detect cell lines containing gene trap insertions. A gene trap integration producing a non-translated fusion transcript is more likely to represent a genuine null mutation in an endogenous gene than is one where fusion protein containing regions of the endogenous gene product is generated. Therefore, current gene trap screens may be missing a large number of

mutations of potential interest. The adoption of a transcript based screening procedure, although initially more time consuming, may be a more thorough way to exploit gene trap technology.

The colonies that stained with the *lacZ* RNA probe showed a variety of sub-cellular transcript localisations. Table 3.8 shows the frequency of each localisation pattern seen. Figure 3.9 gives typical examples of each of the patterns. The majority of colonies either gave no staining (81%), or gave uniform (13%) or grainy (5%) cytoplasmic signal. A small fraction of colonies (1-2%) showed nuclear localisation of the *lacZ* fusion transcript. Approximately half of these also contained a discrete dot of transcript in or very close to the nucleus. Colonies which gave staining in a very few, rounded up cells were counted as not staining, since these rounded up cells appear to be dead or dying so the staining seen is likely to be background, rather than genuine hybridisation of the probe to its target sequence.

Comparison of Fusion RNA and Fusion Protein Patterns in Gene Trap Electroporated Fibroblasts

Distinctive dots of fusion transcript were seen in the cytoplasm of about 1% of ES cell lines electroporated with pGT1.8K as described above. A small proportion of colonies also showed nuclear localisation of fusion RNA. It had previously been noted that the fusion protein detected in gene trap cell lines also shows a variety of sub-cellular localisation patterns (W. C. Skarnes, unpublished) with nuclear protein and cytoplasmic dots of protein being the most common patterns seen. As described previously, in most cases of restricted cytoplasmic transcript localisation the reason for the localisation is to control the distribution of the protein translated from the

mRNA. To investigate whether the presence of dots of transcript localisation correlated with the presence of dots of protein localisation a screen was undertaken using the gene trap vector pGT1.8βgeo in mouse (10T1/2) fibroblast cells. This vector was chosen to minimise the number of neomycin resistant lacZ negative colonies obtained. Fibroblast cells were chosen as their flat morphology and extensive cytoplasm aids visualisation of sub-cellular patterns.

10^8 fibroblast cells were electroporated with 100μg of the vector pGT1.8βgeo. The cells were plated onto twenty 10 cm diameter culture dishes and selected with 400mg/ml G418. After 2 weeks of culture, the cells had formed a confluent monolayer of neomycin resistant cells, rather than individual colonies. The cells were removed from the growth surface by trypsinisation and frozen as twenty pools of cells for further study. Three of these pools were then thawed and plated at a lower density so that individual colonies formed. As a result of these additional manipulations, this experiment cannot be used to quantify the relative numbers of colonies obtained showing each localisation pattern.

The *lacZ* probe detected fusion transcript in 70-80% of the neomycin resistant clones recovered in this experiment. Since activation of the selectable marker, neomycin, requires the synthesis of a *βgeo* fusion transcript, a larger proportion of clones obtained with this vector were expected to have detectable levels of *lacZ* containing fusion transcript. The 20-30% of colonies which did not stain must either have contained amounts of fusion transcript too low to be detected by whole mount *in situ* hybridisation or have lost *lacZ* sequences whilst retaining *neomycin*

sequences. A similar range of cytoplasmic transcript distribution patterns were detected with this vector as with pGT1.8K. Table 3.10A shows the numbers of clones obtained showing each RNA pattern. Percentages are not given, as the freezing down and replating of the cells from this electroporation may have altered the representation of certain patterns within the population. The majority of clones showed widespread cytoplasmic staining. A number of clones showed cytoplasmic staining with an increase in intensity close to the nucleus (scored as cytoplasmic/peri-nuclear). This pattern was not seen with embryonic stem cells, but it is likely that this difference is a result of the very different morphology of fibroblast cells, rather than a difference in the genes trapped by the two vectors. A smaller number of clones showed grainy cytoplasmic or peri-nuclear staining. A few clones were also isolated that showed nuclear accumulation of fusion transcripts or single dots of stain within the cytoplasm.

LacZ fusion protein was detected in 30-40% of the neomycin resistant clones isolated. Table 3.10B shows the number of clones showing each protein distribution. Again, fewer clones produce fusion protein with detectable levels of β-gal activity than produce detectable fusion transcript. In order to obtain neomycin resistant clones using the pGT1.8βgeo vector, translation of the fusion transcript must occur, as the neomycin gene is present as a fusion with *lacZ*. Clones producing no detectable β-gal must either produce a protein in which the β-gal region is inacitive or produce levels of protein activity undetectable by the X-gal staining procedure used. The large difference between the numbers of colonies showing detectable transcript and those showing detectable protein suggests that the whole mount *in situ* technique is more efficient in detecting gene trap events than X-gal staining.

Figure 3.11 shows representative examples of the fusion RNA patterns seen in fibroblasts electroporated with pGT1.8βgeo, together with the β-gal protein expression pattern seen for each of the selected clones. RNA and protein do not co-localise in these cell lines, even in cases where the RNA (H: TβP9,7) or protein (A: TβP9,16) show highly restricted sub-cellular localisations. In particular, clone TβP9,55 (I) has a lacZ protein activity distribution which suggests that the gene product of the trapped gene is associated with the cytoskeleton. The fusion RNA in this line, however, shows an even distribution throughout the cytoplasm.

Sub-cellular Localisation of Neomycin Fusion Transcripts Using the Modified 3' Trap Vector.

Mouse embryonic stem cells were electroporated with the modified 3' trap vector in a parallel electroporation with that using pGT1.8K described previously. The number of G418 colonies obtained with the vector pKnSD was significantly lower than the numbers obtained using the 5' gene trap vector pGT1.8K. Only 134 clones were recovered from the electroporation of 10^8 cells. This represents about 1/20 of the number of colonies recovered from electroporations using pGT1.8K. The number of colonies recovered was, however, comparable to the number usually obtained using the vector pGT1.8βgeo (this study and W.C Skarnes, unpublished). This result suggests that the vector pKnSD enriches for insertions in transcription units in the correct orientation and in appropriate positions within genes. Both pGT1.8βgeo and pKnSD are expected to require insertions into genes in the correct orientation to produce neomycin resistance. pGT1.8K in contrast requires only that the vector inserts stably at some point within the genome.

The *in situ* procedure revealed cytoplasmic neomycin transcript in >70% of the cell lines isolated from this electroporation. No cloning has yet been undertaken from these lines, so it has not been conclusively demonstrated that these are fusion transcripts generated by splicing to endogenous splice acceptors. However, a vector of similar design has recently been demonstrated to function as a gene trap vector in this manner (Yoshida et al, 1995). Table 3.12 shows the frequency of each localisation pattern seen. Figure 3.13 shows typical examples of each pattern recorded. In all cases, the transcripts detected were cytoplasmic and had a highly granular appearance. Approximately 40% of the clones also showed intense peri-nuclear accumulation of transcript, usually as a single dot of signal, but in some cases two or more dots were seen per cell. One cytoplasmic structure known to occur as a single entity in each cell is the centrosome. During early mitosis, the centrosome divides and one centrosome migrates to each pole of the cell where it is involved in spindle formation. In mitotic cells from the lines showing dots of transcript single dots of neomycin RNA can be seen, occupying the space between the two sets of condensed chromosomes (figure 3.12). This confirms that the accumulations of neomycin transcript seen are not associated with the centrosome.

Comparison of the Range of Patterns detected Using the two Types of Vector.

Each of the three screens gave a number of different transcript localisation patterns. The most striking patterns observed are the nuclear localisation seen exclusively in clones electroporated with 5' gene trap vectors, and the peri-nuclear dot seen predominantly in clones electroporated with the modified 3' trap vector. Additionally, it was noted that the 5′ gene trap

vectors gave rise to some colonies showing a grainy transcript distribution and some showing very uniform staining. In contrast to this, all of the clones obtained by electroporation with the 3′ trap vector showed grainy fusion transcript. All of the RNA localisation signals described to date have been mapped to the 3' UTRs of genes. The comparison of fusion transcript and fusion protein localisation carried out in fibroblast cells suggests that signals involved in the control of protein distribution by means of transcript localisation will not be found in the 5' regions of transcripts. However, the detection of both uniform and grainy cytoplasmic distributions of fusion RNA suggests that some signals involved in the cytoplasmic trafficking of transcripts may lie in the 5' regions of genes.

Figure 3.4: Generalised Structure of Fusion Transcripts Generated by the Insertion of the 5' Gene Trap Vector pGT1.8K into the Intron of an Endogenous Gene.

A) Structure of Vector.

The *En-2* intron (represented by the thick black line), splice acceptor site (SA) and 120bp of *En-2* exon sequence (dark grey box) are upstream of the *lacZ* gene (shaded box). The *neomycin* gene (pale grey box) is driven by a seperate β-actin promotor (white box). The polyadenylation signal (pA) is from SV40.

B) The predicted structure of the integration of pGT1.8K into the intron of an endogenous gene. White boxes represent endogenous exons, thin lines represent endogenous introns.

C) The predicted transcript produced from this integration. Dotted lines represent splicing between the endogenous splice donor site (SD) and the *En-2* splice acceptor site.

D) The predicted fusion proteins. If the reading frame of the endogenous gene is the same as that of the *lacZ* gene a fusion protein will be produced. Fusion transcripts from pGT1.8K will be translated to give a fusion protein containing 5' regions of the endogenous gene and the *lacZ* protein. A neomycin protein will be produced regardless of the reading frame of the endogenous gene.

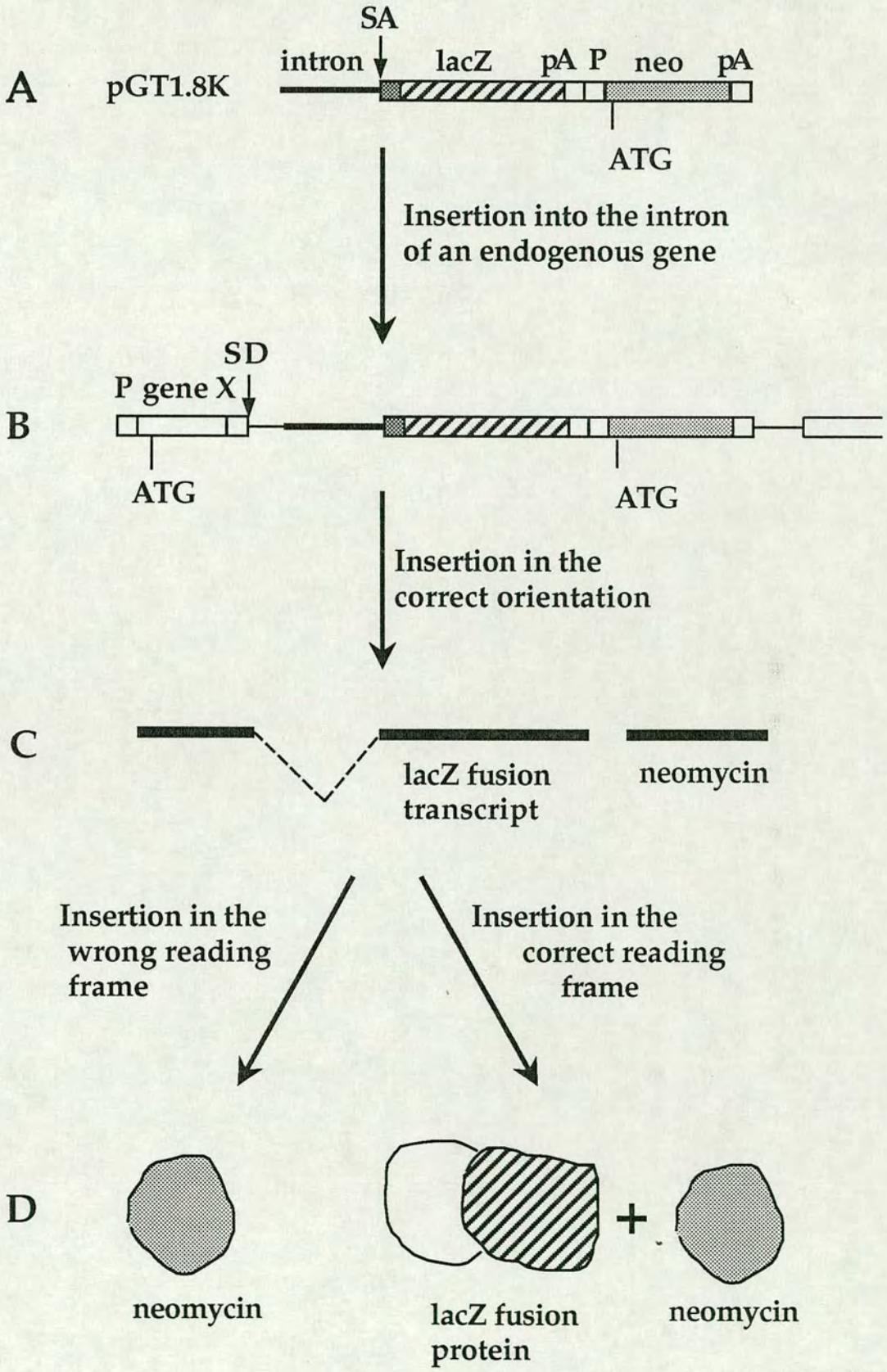

A
pGT1.8K
SA
intron
lacZ
pA
P
neo
pA
ATG
Insertion into the intron of an endogenous gene
B
SD
P gene X
ATG
ATG
Insertion in the correct orientation
C
lacZ fusion transcript
neomycin
Insertion in the wrong reading frame
Insertion in the correct reading frame
D
neomycin
lacZ fusion protein
+
neomycin

Figure 3.5: Generalised Structure of Fusion Transcript Generated by the Insertion of the 5' Gene Trap Vector pGT1.8βgeo into the Intron of an Endogenous Gene.

A) Structure of Vector.

The *En-2* intron (represented by the thick black line), splice acceptor site (SA) and 120bp of *En-2* exon sequence (dark grey box) are upstream of the *βgeo* fusion gene (shaded box). The polyadenylation signals (pA) is from SV40.

B) The predicted structure of the integration of pGT1.8βgeo into the intron of an endogenous gene. White boxes represent endogenous exons, thin lines represent endogenous introns.

C) The predicted transcripts produced from these integrations. Dotted lines represent splicing between the endogenous splice donor site (SD) and the *En-2* splice acceptor site.

D) The predicted fusion proteins. If the reading frame of the endogenous gene is the same as that of the *βgeo* fusion gene, a fusion protein will be produced. Transcripts from pGT1.8βgeo will be translated to give a fusion protein containing 5' regions of the endogenous gene, *lacZ* and *neomycin*.

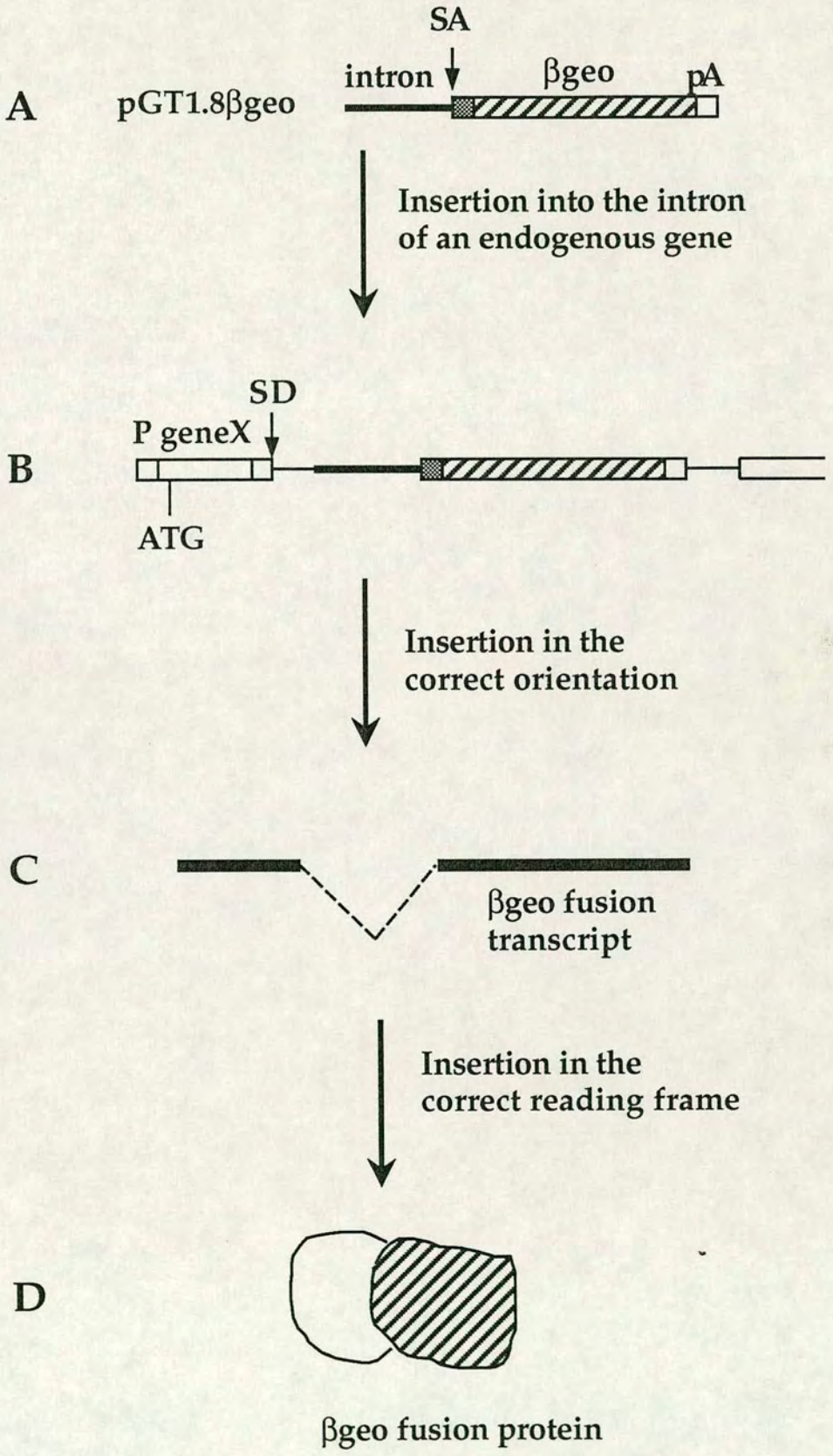

SA
intron
βgeo
pA
A
pGT1.8βgeo
Insertion into the intron
of an endogenous gene
SD
P geneX
B
ATG
Insertion in the
correct orientation
C
βgeo fusion
transcript
Insertion in the
correct reading frame
D
βgeo fusion protein

Figure 3.6: Generalised Structure of Fusion Transcripts Generated by the Insertion of the 3' Trap Vector pβKnSD into an Endogenous Gene.

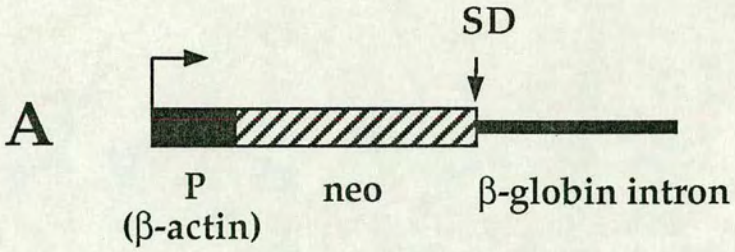
A) Structure of pβKnSD.

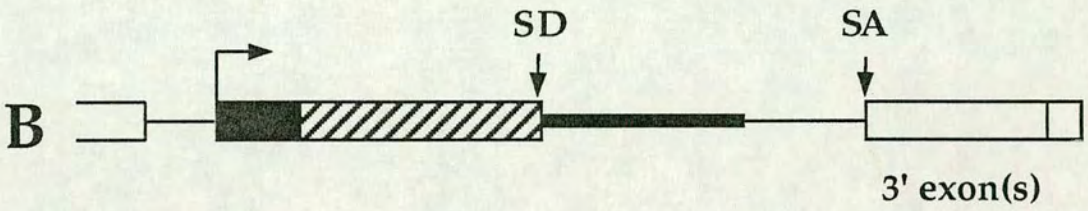
The constitutive *β–actin* promotor (represented by the black box) is placed upstream of the *neomycin* gene (the shaded box) and a region of *β–globin* genomic DNA including the splice donor (represented by the thick line). This region includes the last 15 base pairs of the first exon of the gene and the whole of the first intron.

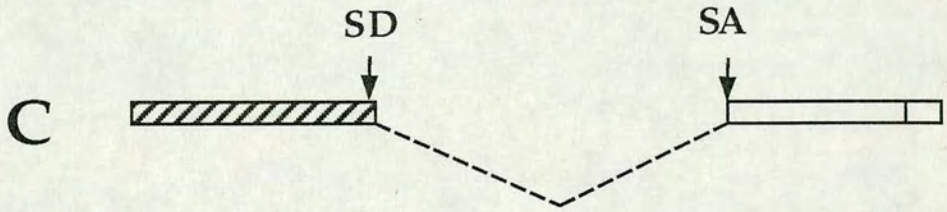
B) The predicted structure of the integration of pβKnSD into an endogenous gene. White boxes represent endogenous exons, thin lines represent endogenous introns.

C) The predicted splicing of the primary transcript driven by the β–actin promotor.

D) The predicted final transcript, comprising the *neomycin* gene and 3' regions of the endogenous gene at the site of insertion, including the 3'UTR. The transcript is expected to be polyadenylated as dictated by the endogenous polyadenylation signal.


A
SD
P
(β-actin)
neo
β-globin intron


B
SD
SA
3' exon(s)


C
SD
SA

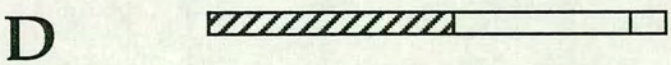

D

Figure 3.7: Whole Mount *In Situ* Hybridisations Using *Actin* and *LacZ* Probes on the ES Cell Line CGR-8.

A) An anti sense *actin* probe shows the characteristic staining pattern in well spread, differentiated cells. Concentration of signal can be seen in the leading edges of the cell. Cells were stained for 2 hours.

B) An anti-sense *LacZ* probe demonstrates the absence of background staining using this whole mount *in situ* protocol. Cells were stained overnight.

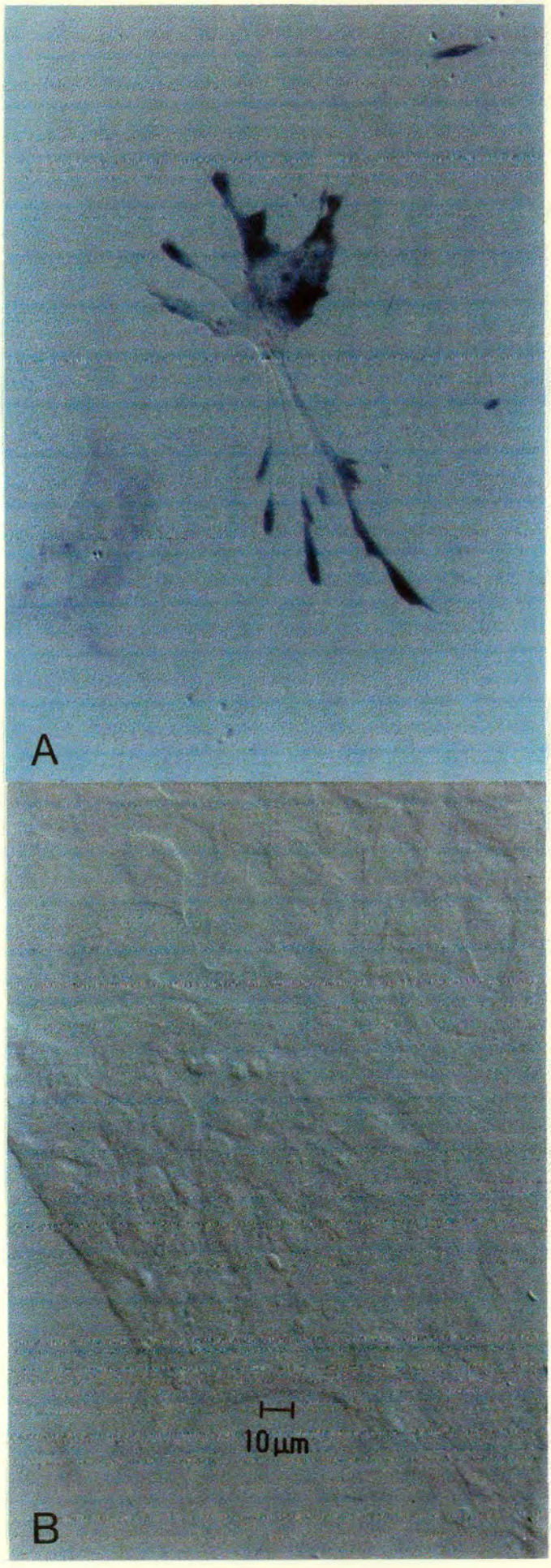

A
10 µm
B

Table 3.8: Summary of the Patterns of Localisation of *lacZ* Fusion Transcript Following the Electroporation of Embryonic Stem Cells With the Gene Trap Vector pGT1.8K.

LOCALISATION PATTERN	NUMBER OF CLONES	% OF TOTAL
No staining	1321	81
Uniform cytoplasmic	208	13
Grainy cytoplasmic	78	5
Nuclear	20	1
Nuclear with intense dots	3	<1
TOTAL	**1630**	**100**

Figure 3.9: Typical Examples of Transcript Localisations Seen in Embryonic Stem cells electroporated With the 5' Gene Trap Vector pGT1.8K.

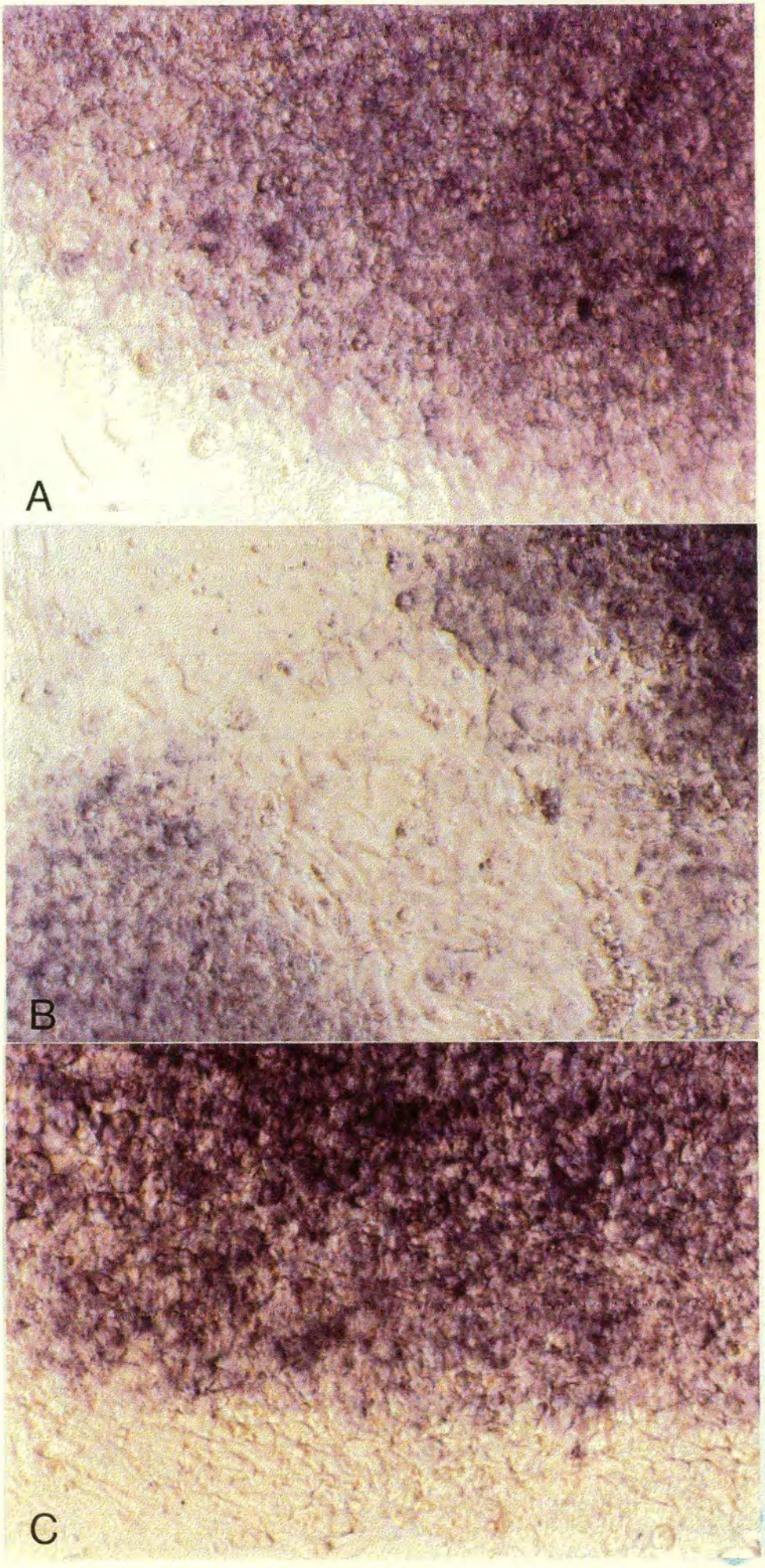
Each photograph represents cells of a single colony. All cells were viewed using the X40 objective of an olympus inverted microscope after overnight staining. Photographs were taken using Kodak GPF160 Ektacolor gold II colour negative film.

Panels A, B and C all show colonies scored as having uniform cytoplasmic staining.

A) Uniform Cytoplasmic Staining. Signal can be seen in all cells. The darker appearance of more central regions of the colony (the top right of the photograph) can be attributed to optical effects caused by piling of the cells.

B) Patchy Cytoplasmic Staining. The staining within each cell appears to be uniform, but is restricted to patches of cells within the colony.

C) Patchy Cytoplasmic Staining. A small number of colonies showed definite borders between staining and non staining cells. The non-staining cells were the more differentiated cells on the edges of the colonies.

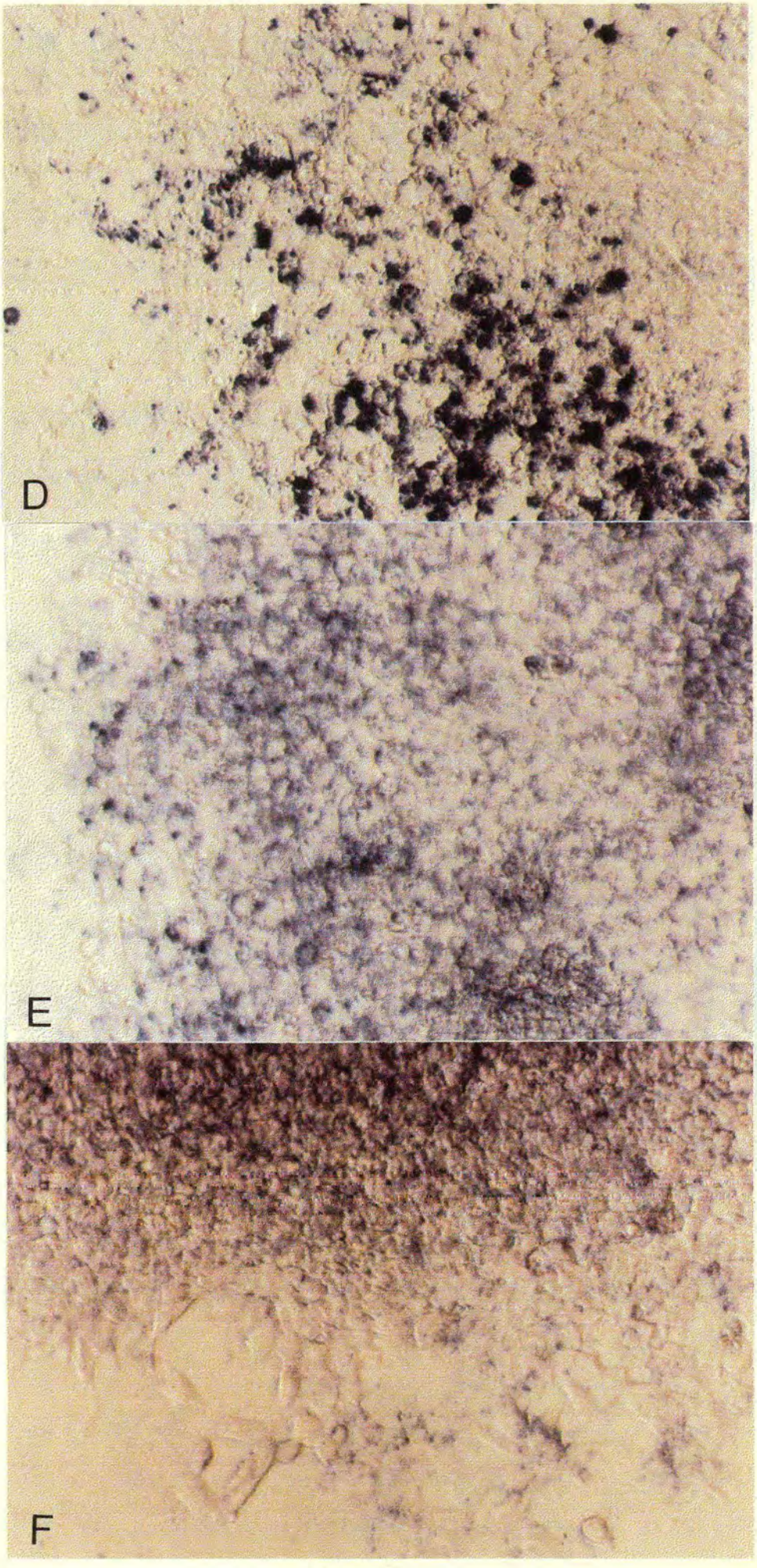

A
B
C

Panels D, E and F all show colonies scored as having grainy cytoplasmic staining.

D) Patchy, Grainy Cytoplasmic Staining. Intense grains of stain can be seen in some parts of the colony. The use of a UV counter stain for DNA reveals the staining to be cytoplasmic, at least in well spread cells, where this can be readily determined.

E) Grainy Cytoplasmic Staining. All of the cells of the colony stain. The staining has a grainy appearance, used to distinguish these colonies from those described as having uniform cytoplasmic staining.

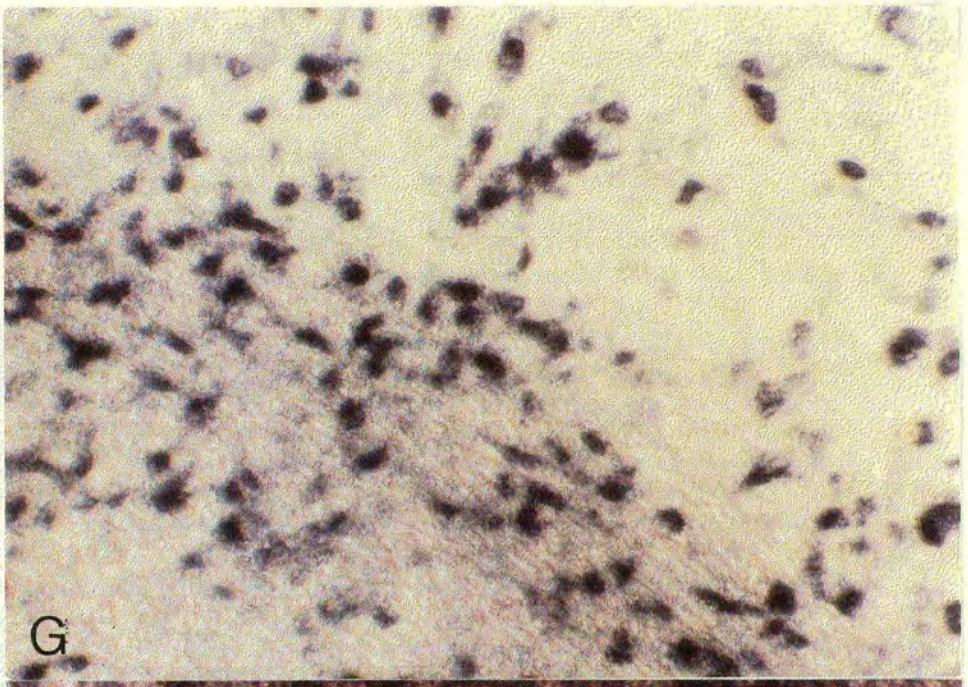
F) Grainy Cytoplasmic Staining. This clone also shows grainy staining. The grains appear to be generally smaller than those in E above.

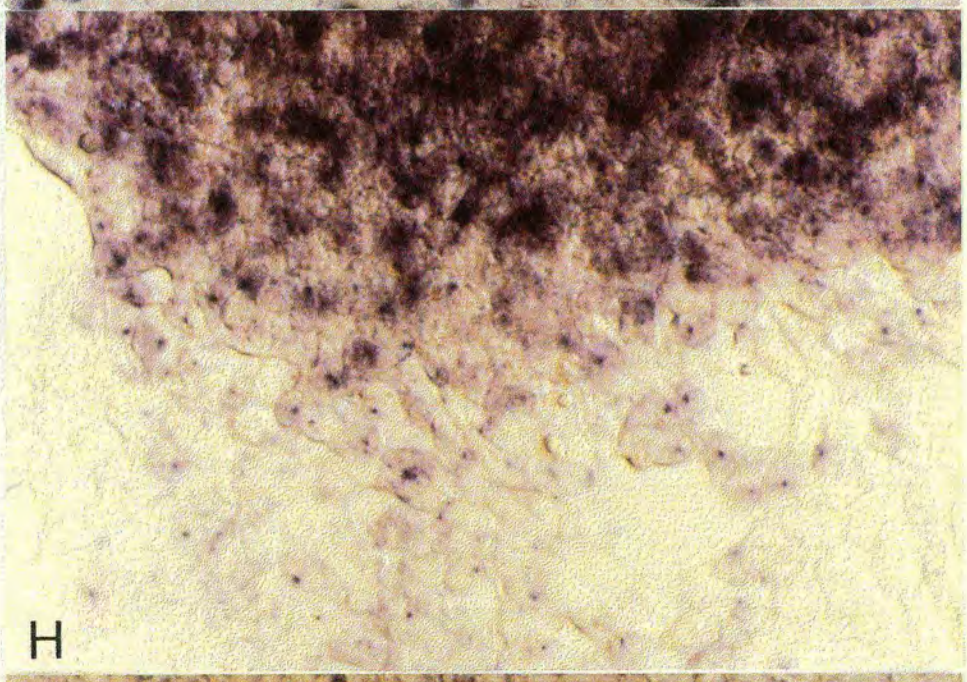

D
E
F

G) Nuclear Staining. Signal is strongest within the nucleus, with some staining also visible in the cytoplasm. The stain has a grainy appearance throughout the cell, best demonstrated in well spread cells.

H) Nuclear Staining With intense Dots. The staining in these cells is also mostly nuclear. A number of cells also show intense dots of staining within the nucleus.

I) No Staining. A few rounded up cells appear to give some signal, but this is likely to be non-specific background staining.


G


H


I

Table 3.10: Summary of the Localisation Patterns of βgeo Fusion Transcript and Fusion Protein Following Electroporation of Mouse Fibroblasts with the Gene Trap Vector pGT1.8βgeo.

A) *βgeo* Fusion Transcript Distribution

LOCALISATION PATTERN	NUMBER OF CLONES
Cytoplasmic	143
No staining	61
Cytoplasmic/peri-nuclear	23
Grainy cytoplasmic	17
Grainy cytoplasmic/peri-nuclear	4
Nuclear	4
Intense dots of stain	2
TOTAL	**254**

b) LacZ Fusion Protein Distribution

LOCALISATION PATTERN	NUMBER OF CLONES
No staining	166
Whole cell	27
Nuclear	19
Peri-nuclear dots	22
Nuclear plus peri-nuclear dots	5
Cytoplasmic/peri-nuclear	7
Cytoplasmic	6
Cytoskeletal	2
TOTAL	**254**

Figure 3.11: Typical Examples of Sub-Cellular Fusion Transcript and Protein Patterns Observed Following the Electroporation of 10T1/2 Fibroblast Cells With the Gene Trap Vector pGT1.8βgeo.

Each photograph represents cells from a single clonal cell line. Cells were viewed using an inverted Olympus microscope and images captured using the Colorvision programme on a Macintosh Quadra 900 computer. Composites were made using Adobe Photoshop and printed using a dye sublimation colour printer.

A) TβP9,16

i) *LacZ* RNA whole mount *in situ* hybridisation shows uniform cytoplasmic staining.

ii) X-gal staining reveals peri-nuclear fusion protein, with chains of dots of intense staining surrounding the nucleus. This pattern has since been demonstrated to be indicative of an insertion of the vector into a gene coding for a secreted protein (Skarnes et al, 1995).

B) TβP9,50

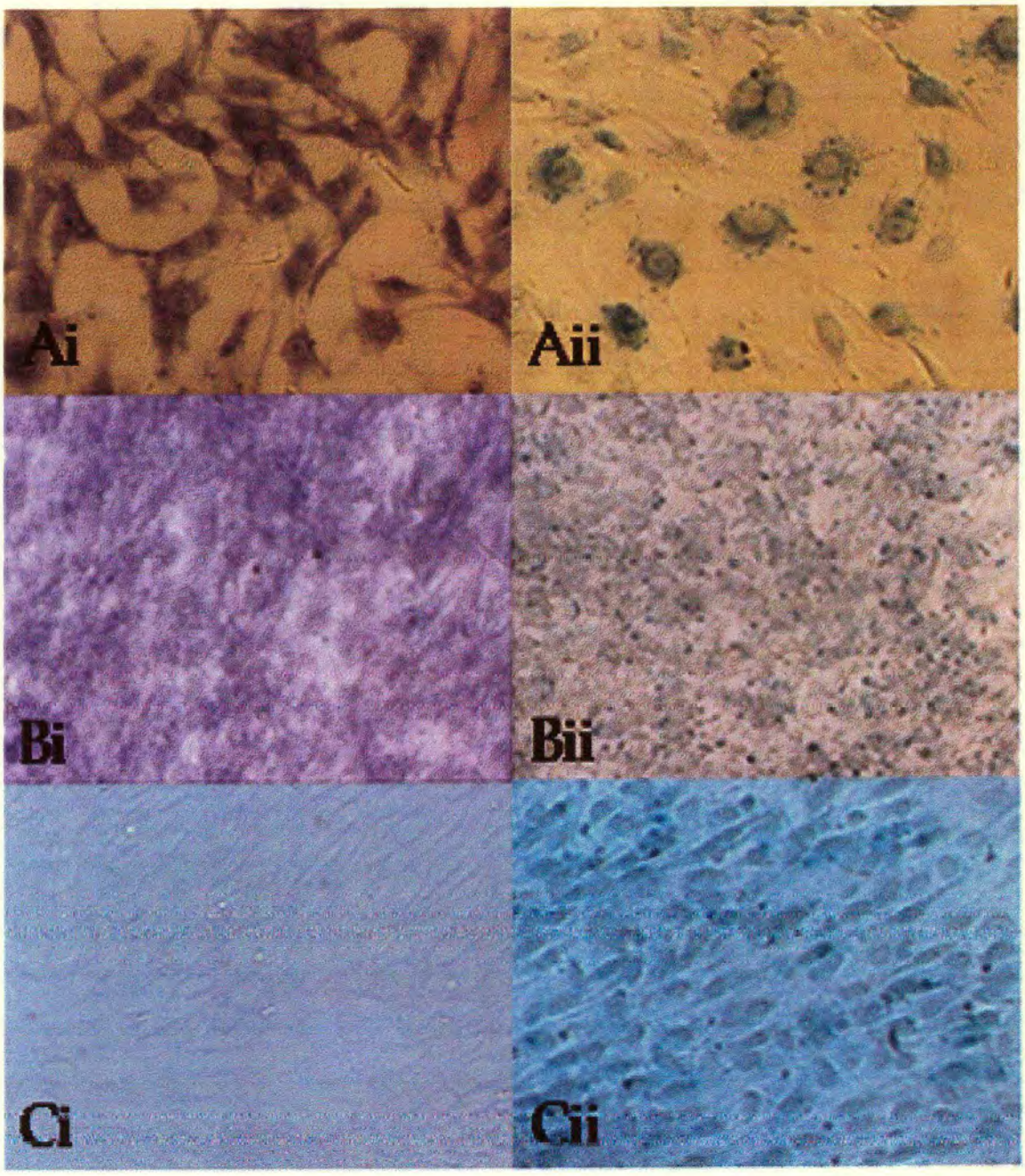
i) Fusion transcript shows a granular cytoplasmic distribution.

ii) Fusion protein shows a diffuse cytoplasmic localisation, with large accumulations of stain in many cells.

C) TβP9,10

i) Fusion transcript is not detectable.

ii) Fusion protein is easily detected, showing a predominantly nuclear localisation, with large accumulations of stain in many cells. This cell line confirms that high concentrations of fusion transcript are not required to generate significant levels of fusion protein.


Ai
Aii
Bi
Bii
Ci
Cii

D) TβP20,14

i) Fusion transcript in this line shows a peri-nuclear distribution.

ii) Fusion protein is not detectable.

E) TβP4,26

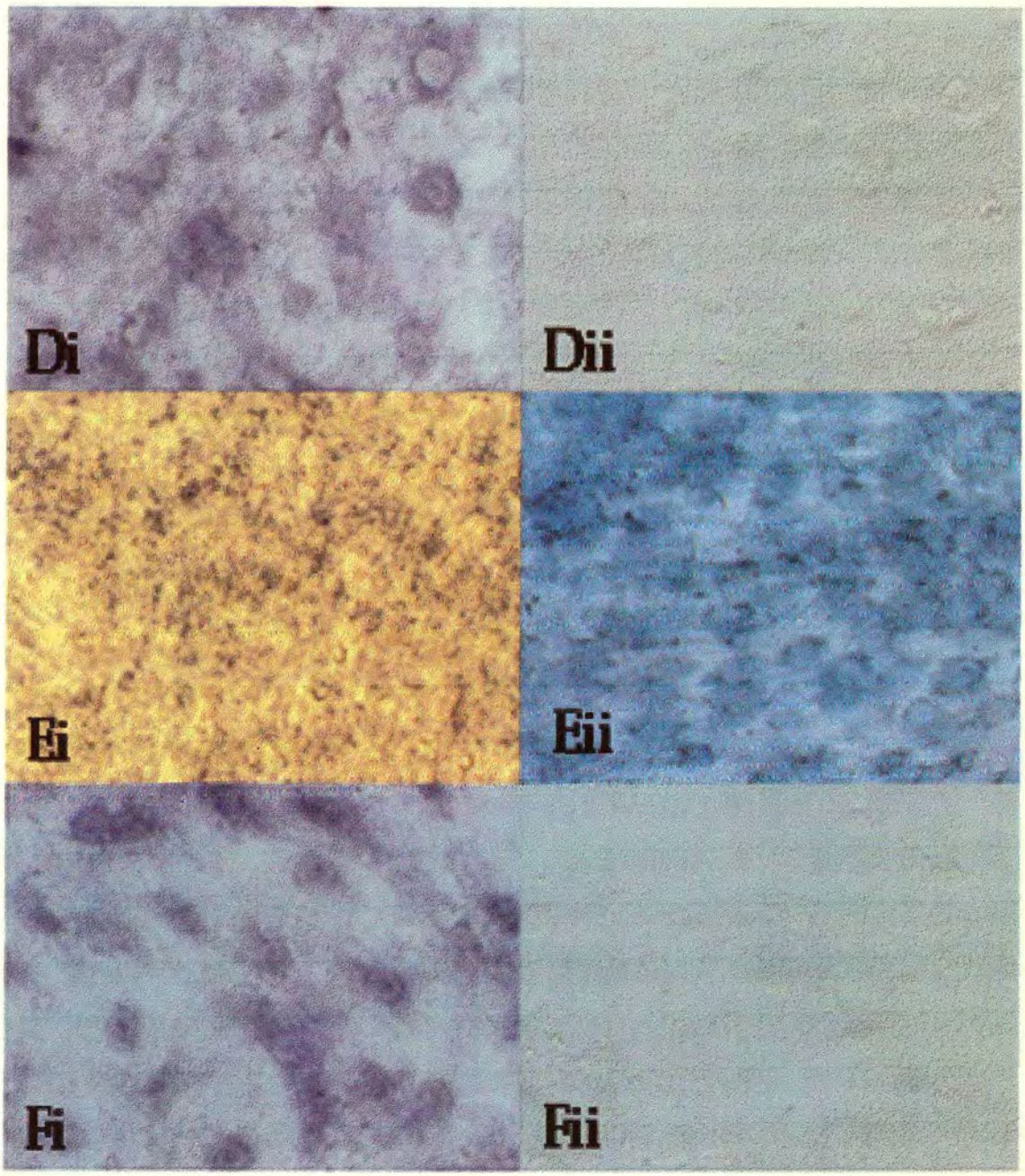
i) Fusion transcript shows a highly granular appearance throughout the cytoplasm.

ii) Fusion protein shows a peri-nuclear distribution.

F) TβP4,35

i) Fusion transcript shows a grainy cytoplasmic distribution, with a tendancy to accumulate around the nucleus.

ii) Fusion protein is not detectable.


Di
Dii
Ei
Eii
Fi
Fii

G) TβP20,8

i) Fusion transcript is largely restricted to the nucleus, with a small amount also present in the cytoplasm.

ii) Fusion protein is not detectable.

H) TβP9,7

i) In this line, the fusion transcript is present in large dots of stain within the cytoplasm.

ii) Fusion protein shows a uniform distribution throughout the cell, confirming that the dot of transcript seen does not give rise to a dot of protein.

I) TβP9,55

i) Fusion transcript shows a uniform cytoplasmic distribution.

ii) Fusion protein shows a cytoskeletal distribution. Transcripts for some cytoskeletal proteins have been shown to have characteristic sub-cellular distributions. This result suggests either that this is not the case for all cytoskeletal proteins, or that the signals required for transcript localisation have been lost from the fusion transcript in this line.

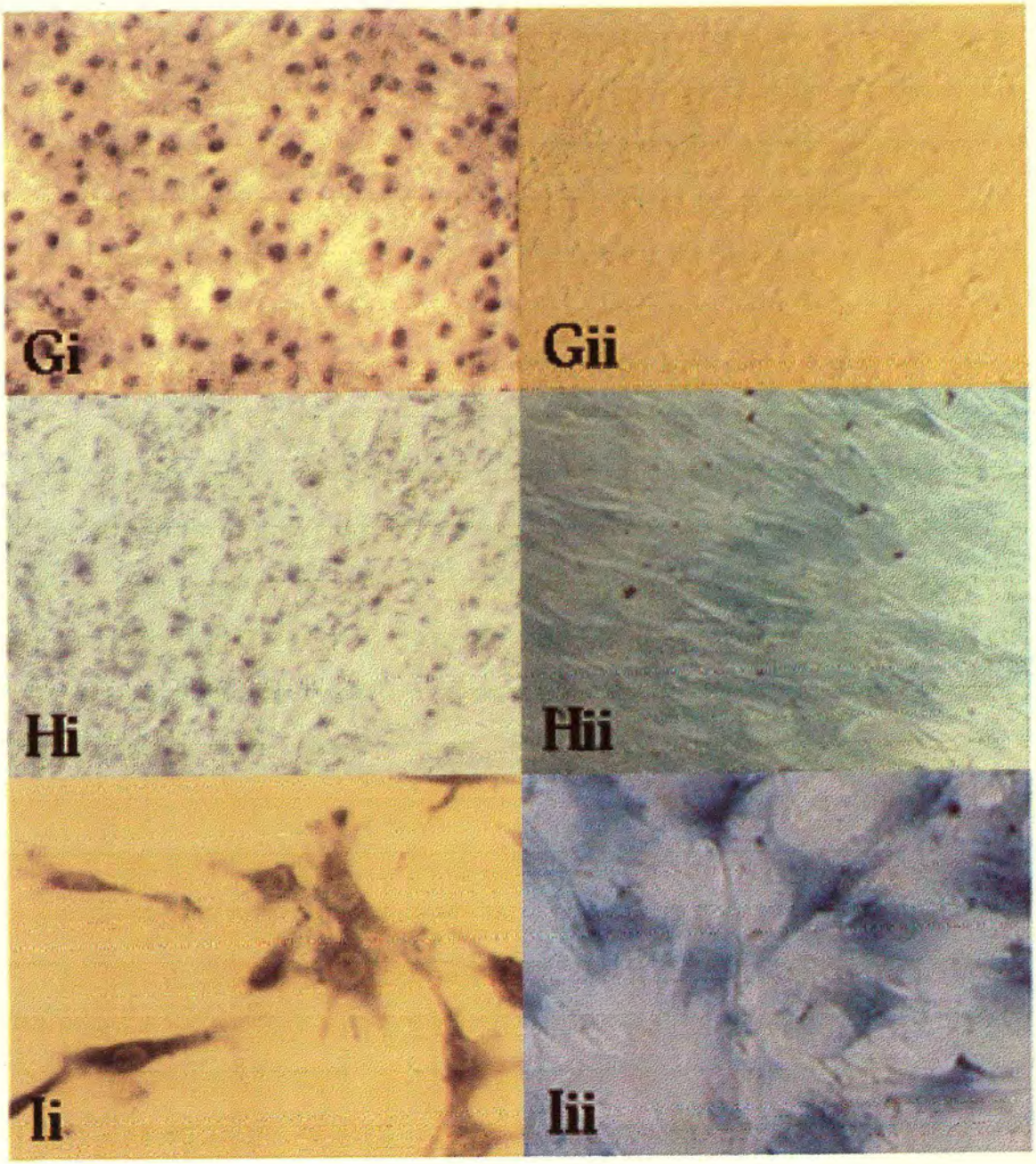

Gi
Gii
Hi
Hii
Ii
Iii

Figure 3.12: Summary of the RNA Localisation Patterns Observed Following Electroporation of Embryonic Stem cells With the 3' Trap Vector pβKnSD.

RNA DISTRIBUTION	NUMBER OF CLONES	% OF TOTAL
No staining	37	27
Grainy Cytoplasmic	48	35
Grainy Cytoplasmic With Focal Accumulations	35	26
Focal Accumulations Alone	16	12
TOTAL	**136**	**100**

A total of 38% of the clones recovered showed large dots of neomycin fusion transcript.

Figure 3.13: Typical Examples of Transcript Localisations Seen in Embryonic Stem Cells Electroporated With the 3' Gene Trap Vector pβKnSD.

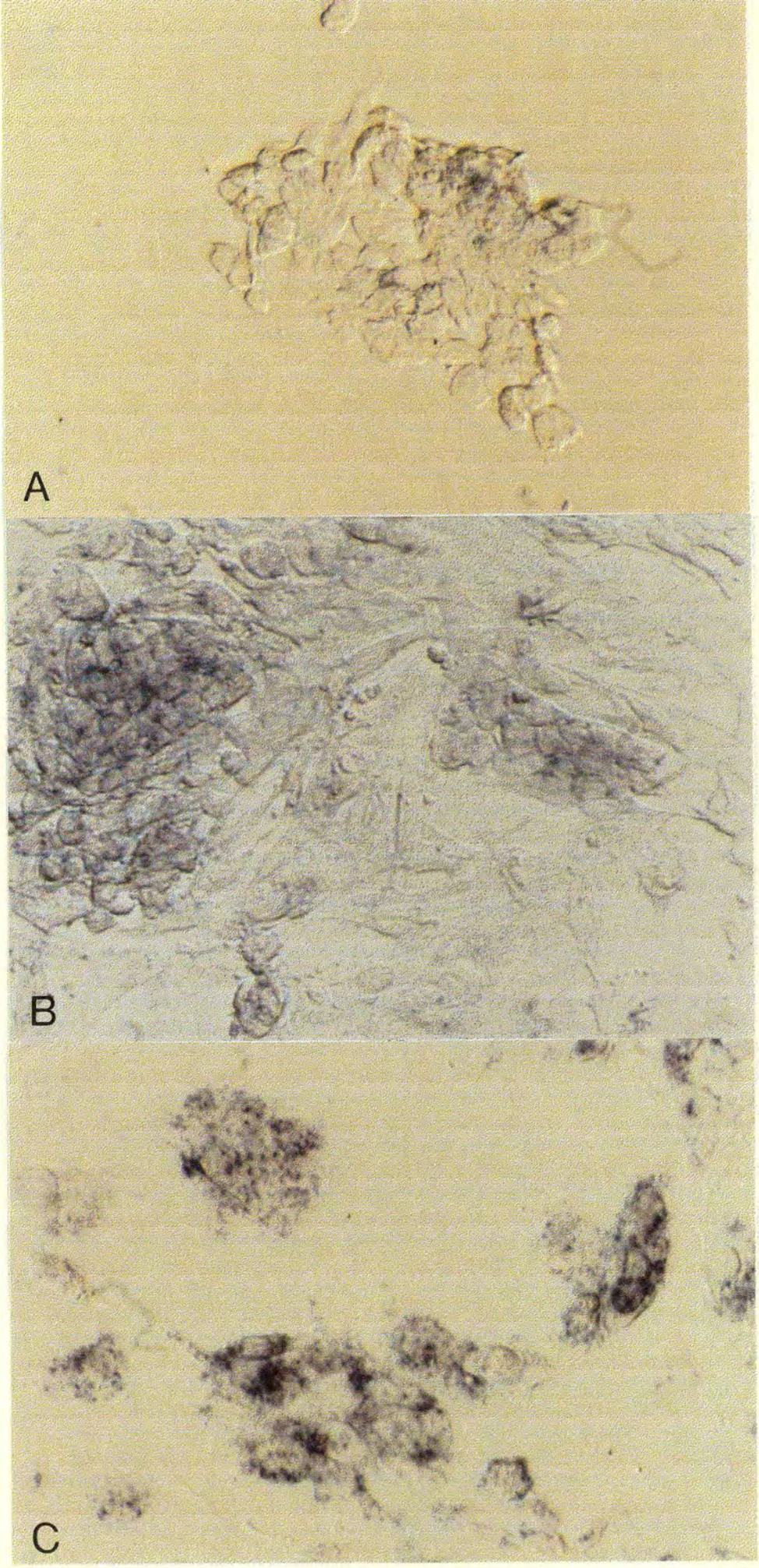
Each photograph represents cells of a single isolated cell line. All cells were viewed using the X40 objective of an olympus inverted microscope after overnight staining. Photographs were taken using Kodak GPF160 ektacolor gold II colour negative film.

Panels A, B and C all show cell lines scored as having grainy cytoplasmic staining.

A) Only a few cells in the colony stain. Staining does not appear to be dependent on the state of differentiation of the cells.

B) Again, patches of cells show staining. In this line, staining appears to be more intense in less differentiated cells.

C) All cells of this line give an intense granular signal.


A
B
C

Panels D, E and F all show cell lines containing focal accumulations or dots of stain.

D) Each cell of this line contains a single accumulation of signal. There is very little staining in other regions of the cell.

E) Again, each cell contains a single dot of signal, with a significant amount of grainy cytoplasmic signal also detected.

F) This line shows strong general cytoplasmic staining in addition to dots of stain. The cytoplasmic staining is largely restricted to nests of undifferentiated cells, while the dots are more readily visible in more differentiated cells, perhaps due to the absence of other cytoplasmic staining in these cells.

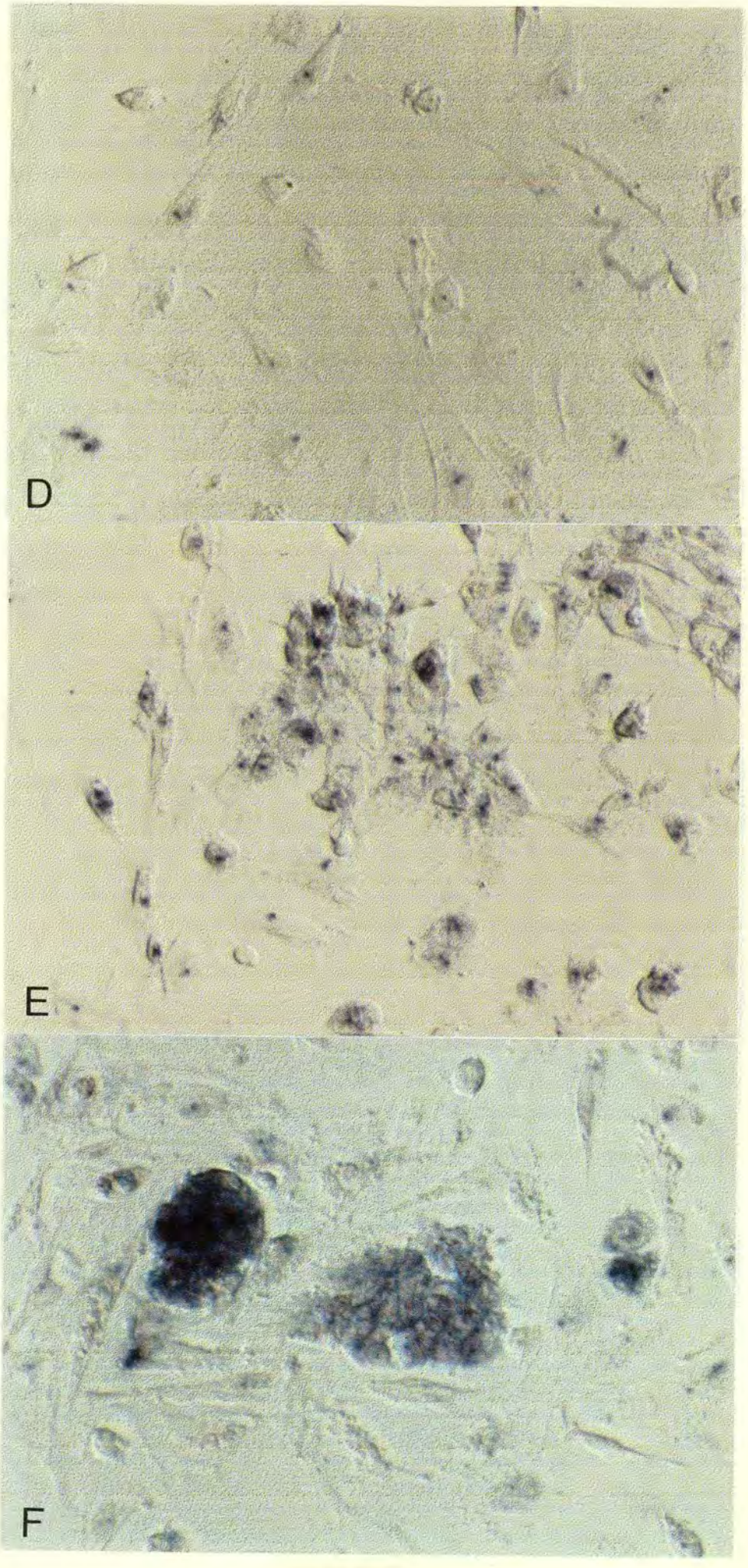
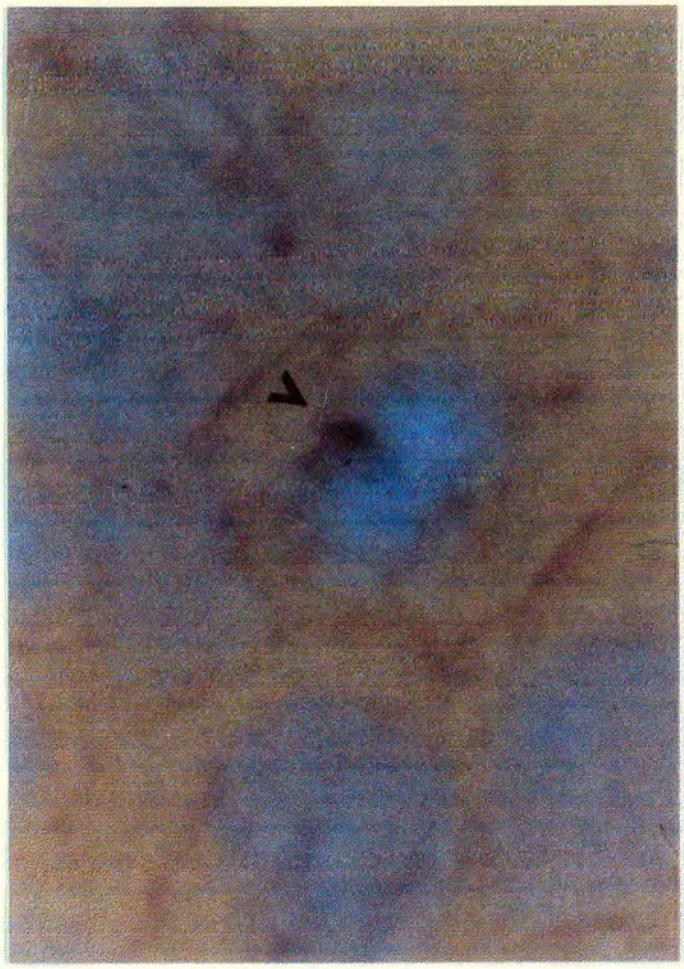

D
E
F

Figure 3.14: The Focal Accumulation of Signal Seen in 40% of Embryonic Stem Cell Lines Electroporated with the 3' Trap Vector pβKnSD Occupies a Position Between the Two Sets of Chromosomes During Mitosis.

The mitotic cell pictured here is from line SD3. Whole mount *in situ* hybridisation using a *neomycin* anti-sense riboprobe showed a dot of stain in cells of this line (arrowed). Counterstaining of genomic DNA with 1μM Hoechst 33258 (light blue staining) demonstrates that the accumulation of signal occupies a position between the sets of condensed chromosomes during mitosis rather than forming two dots at either end of the mitotic spindle. The cell pictured is characteristic both of line SD3 and of other lines showing a dot of signal.



DISCUSSION

The electroporation of gene trap vectors into cells leads to the synthesis of *lacZ* fusion transcripts and fusion proteins. The vector pGT1.8K has previously been observed to yield detectable levels of fusion protein in only 1% of neomycin resistant colonies (W. C. Skarnes, unpublished). The data presented here indicate that 22% of neomycin resistant colonies obtained using pGT1.8K produce levels of *lacZ* fusion transcript detectable by whole mount *in situ* hybridisation. 30-40% of clones obtained by electroporating mouse fibroblasts with pGT1.8βgeo showed detectable levels of fusion protein, while 70-80% of the same clones contained detectable levels of fusion transcript. Colonies producing detectable fusion transcript but no detectable fusion protein are probably due to insertion events generating fusion transcripts that are inefficiently processed, inefficiently translated, or translated into fusion proteins lacking β-gal activity. Transcripts in which the endogenous sequence and *lacZ* sequence are not in the same frame would not be translated. Insertions into certain classes of genes may produce inactive β-gal protein. The active form of β-gal is a tetramer. Any fusion protein that cannot tetramerise either by virtue of the structure of each molecule or their localisation within the cell, will lack β-gal activity, while retaining neomycin activity. For example, certain integrations into genes coding for secreted molecules may lead to the production of fusion proteins that become sequestered within the lumen of the golgi apparatus, where they are non-functional (Skarnes et al, 1995). A difference in sensitivity between the two techniques may also contribute to the discrepancy seen.

A variety of patterns of sub-cellular fusion transcripts have been documented using conventional 5' gene trap vectors and a vector designed to trap the 3' regions of endogenous genes. The number of different patterns seen was similar for the two types of vector. The limited range of patterns of sub-cellular transcript localisation seen in the experiments suggests that RNA localisation is of minimal importance to the correct function of the majority of mammalian genes. This is not unexpected, since the genes for which RNA localisation has been previously shown to be important are involved in the formation of the basic structure either of the developing embryo, or of the cell itself. The observation of a small number of different localisation patterns, however, suggests that the cell may sort different broad classes of transcript along slightly different localised pathways within the cytoplasm. The cloning of genes associated with the localisation patterns seen would perhaps give some indication as to the criteria by which cytpolasmic RNAs are sorted into these classes.

The RNA localisation signals reported to date have resided exclusively in the 3'UTRs of genes. The data presented in this chapter suggest that there may be elements in the 5' regions of genes which play a role either in the final localisation of transcripts in the cell, or in the pathways by which these localisations are achieved. Fusion transcripts including the 5' regions of endogenous genes do not show any co-localisation with the fusion proteins translated from them, suggesting that any localisation signals present in 5' regions have a role other than in the control of protein localisation. Using 5′ gene trap vectors, a distinction can be made between cells which show a uniform distribution of fusion transcript within the cytoplasm, and those in which the RNA has a highly granular appearance. This distinction is not

apparant in clones electroporated with the 3′ trap vector, in which all fusion transcripts show a granular distribution. The nature of the colorimetric detection used could give rise to a certain graininess. The colour reaction is initiated by the cleavage of the phosphate group from the BCIP by alkaline phosphatase. This reaction yields a blue colour and produces a proton which reduces NBT to yield a purple insoluble precipitate. However, this cannot account for the large grains of signal detected in many of the clones.

A study of the sub-cellular localisation and transport of myelin basic protein (MBP) mRNA in living oligodendrocytes and neuroblastoma cells (Ainger et al, 1993) demonstrated the accumulation of the mRNAs for MBP, actin and globin as cytoplasmic granules following microinjection. These granules are about 0.3m diameter, motile and associated with the cytoskeletal matrix. The granular localisation seen with these RNAs and with vimentin mRNA in a variety of cell types (Cripe et al, 1993) is proposed to be a visualisation of the transport of transcripts within the cytoplasm as ribonucleoproteins. The detection of granules of globin mRNA suggests that the formation of RNPs is not unique to those RNA that have a tightly controlled distribution within the cell, since globin transcripts are present throughout the cytoplasm. Thus, it is proposed that all mRNAs are transported through the cytoplasm in the form of RNA transport particles containing specific proteins, with localised and non-localised transcripts being segregated into different particles (Wilhelm and Vale, 1993). Although all of the signals involved in transcript localisation mapped so far have resided in the 3′UTR, no study has been made of potential sites of RNA/ protein interaction in non-localised transcripts. The size of the granules of fusion transcript detected in the electroporated cell lines is similar to that of

the particles seen in the study of MBP and globin transcripts, suggesting that they may be the proposed RNA transport particles.

If the assumption is made that transcripts are normally present in the cytoplasm as RNP particles, then the uniform distribution seen in some clones containing 5′ gene trap integrations represents an abnormal situation. This may result from the loss of 3′ sequences from these transcripts. If this is the case, then it is likely that some transcripts contain sequences in their 5′ regions that are sufficient for particle formation, while others require either 3′ signals alone or a combination of 3′ and 5′ signals. Alternatively, it may be that the large granules observed using the 3′ trap vector are artefactual and result from sequences within the pKnSD vector. The neomycin sequence is unlikely to be responsible for granule formation, as a significant number of the fibroblast lines electroporated with the vector pGT1.8βgeo showed uniform distribution of the fusion transcript, despite the presence of the neomycin sequence within it. The β-actin promotor fragment used in the construction of the 3′ trap vector does contain 39bp of the β-actin 5′ UTR. However, since the signals required for actin mRNA localisation have been mapped to the 3′UTR, it is highly unlikely that any additional localisation signals are present in this small region. Although many genes have now been studied using colorimetric whole mount *in situ* hybridisation, particularly in developing embryos, published examples are rarely shown at high enough magnification to examine the texture of the staining. Thus it is difficult to judge whether most transcripts are normally present as cytoplasmic granules.

There are other significant differences in the patterns seen with the two types of vector. Intense dots of transcript accumulation were seen in the cytoplasm of cells from 40% of the lines derived using the 3' trap vector. This type of localisation was seen in only two of the large number of 5' gene trap clones studied. Nuclear localisation of transcript, in contrast, was seen exclusively in cell lines containing insertions of 5' gene trap vectors. Both of these patterns of transcript distribution were unexpected and warrant further discussion.

The peri-nuclear dots seen in cells producing fusions of neomycin with the 3' regions of endogenous genes may be associated with structural elements within the cell. Sub-cellular structures associated with the interface between the nucleus and the cytoplasm have been described previously. These are the centrosome and the nuclear pore complex (NPC).

The centrosome lies to one side of the nucleus, and acts as microtuble organising centre, from which the microtubules of the cytoskeleton radiate (Karsenti and Maro, 1986). One of the earliest events in mitosis is the replication of the centrosome, and the migration of the two daughter centrosomes to opposite poles of the cells, where they direct the formation of the mitotic spindle. If the accumulation of fusion transcript seen in the 3' gene trap cell lines corresponded to the centrosome, the expected localisation of neomycin signal in mitotic cells would be two dots, on at each pole of the mitotic spindle. Figure 3.14 demonstrates that this is not the case, since some mitotic cells from lines show a single dot of fusion transcript, visible in the area of the cell beween the two sets of chromosomes. It can

therefore be concluded that the accumulations of fusion transcript in this line are not associated with the centrosome.

The NPC has been extensively studied by electron microscopy and current models describe it as an hourglass-like structure inserted into the nuclear membrane (Izzuralde and Mattaj, 1995). The function of the NPC is in the selective transport of protein and RNA molecules into and out of the nucleus. The number of NPCs present on the nuclear membrane of a cell ranges from 100 to 5 X 10^7 (Maquat,1991). Accumulations of particles around the cytoplasmic face of nuclear pores have been demonstrated (Unwin and Milligan, 1982). These are detectable by electron microscopy and appear, by their size and shape, to be ribosomes. One of the major functions of the NPC is in the export of mature transcripts from the nucleus. This process is energy dependant and selective, with incompletely spliced or uncapped RNAs being prevented from leaving the nucleus (Zasloff, 1983; Legrain and Rosbash, 1989; Hamm and Mattaj, 1990). It is possible that the accumulations of neomycin transcript seen are a visualisation of the export of mRNA from the nucleus at restricted sites. No direct demonstration of transport of RNA from the nucleus at a discrete site has been made. However, localised tracks of transcripts leading from the sites of transcription of certain genes towards the periphery of the nucleus have been demonstrated (Lawrence et al, 1989). This suggests that at least some RNAs may be exported from the nucleus at specific sites. The selective transport of different mRNAs from different areas of the nucleus has also been suggested as the mechanism by which assymetric transcript localisation is acheived in cells of the *Drosophila* embryo (Kislauskis and Singer, 1992). RNA coated gold particles have been visualised within the NPC and in the surrounding cytoplasm by electron

microscopy (Dworetzsky and Feldherr, 1988). In these studies the particles were not seen to accumulate in the peri-nuclear cytoplasm, but appeared to move away from the nuclear membrane following nucleocytoplasmic transport. These particles, however, were injected into the cell rather than being synthesised in the nucleus, so their transport may not be restricted in the normal manner.

It has been proposed that the localisation of RNA is controlled at every stage along the processing pathway, during transcription and processing in the nucleus, export into the cytoplasm and translation within cytoplasm and, perhaps, also during its final degradation (Maquat, 1991). The 3'UTRs of genes have recently been implicated in the maintenance of steady state mRNA levels by their involvement in the control of mRNA degradation (Harford and Klausner, 1991 and reviewed in Jackson, 1993). Introns excised from the primary transcript of the Delta gene have a long half life in the Drosophila embryo, and can be detected by *in situ* hybridization. These introns have been shown to form discrete dots within the nucleus prior to their degradation (Kopczynski and Muskavitch, 1992). This suggests that the degradation of excised introns is spatially restricted within the nucleus, so it is possible that the degradation of mature transcripts that are no longer required may also be spatially restricted within the cytoplasm, perhaps in a discrete domain like the dot seen in this study.

It cannot be ruled out that the accumulation of neomycin transcript seen is a function of the nascent fusion polypeptide, rather than of the transcript itself. A rapid association of nascent polypeptides with each other could bring the transcripts with which they are associated into a focus. Sundell and

Singer used puromycin, which dissociates ribosomes from mRNA to distinguish between protein and mRNA mediated effects in their study of actin mRNA localisation (Sundell and Singer, 1990). Puromycin treatment of the gene trapped cell lines prior to in situ hybridisation could be used to confirm that the localisations seen are a function of the transcript itself.

The nuclear accumulation of fusion transcript seen in lines electroporated with 5' gene trap vectors was unexpected, as the only published example of nuclear localisation of a non- snRNA pol II transcript is that of the *Xist* gene (Brockdorff et al, 1992; Brown et al, 1992). In contrast to the nuclear localisation seen in this study, the *Xist* RNA is restricted to a small area of the nucleus, corresponding to the Barr body. Nuclear localisation of *βgeo* fusion transcript occurs as a result of inefficient splicing and is discussed at length in chapters 4 and 5.

It is possible that the patterns of mRNA localisation seen occur as a result of expressing a fusion transcript, and would not be seen if transcripts from the endogenous genes at the sites of vector insertion were studied. The cloning of endogenous genes associated with potentially interesting patterns could be undertaken using 5' RACE cloning for the conventional gene trap vectors and 3' RACE cloning for the modified 3' trap vector. This would not only provide molecular probes with which to study sub-cellular localisation of endogenous genes, but also may uncover sequence motifs involved in the formation of different transcript distributions.

A greater variety of patterns was seen in fibroblasts than in ES cells, with a number of clones showing peri-nuclear transcript localisation being seen.

This is likely to be as a result of the flatter morphology and more extensive cytoskeleton of fibroblasts. Sub-cellular RNA patterns were, generally, more difficult to distinguish in ES cells than in fibroblasts. This is partly a result of their rounded morphology and small cytoplasmic volume and partly a result of the limitations placed on the magnification at which they could be sucessfully viewed when grown on tissue culture plastic. For a screen of this size, the preparation of a seperate glass slide from each colony is impractical, so ES cells are not the ideal cell type to use for this type of analysis. The use of these vectors in a range of differentiated cell types, or the use of a range of differentiation strategies on the ES lines already obtained may reveal more distinct, cell type specific RNA distributions. It would be interesting, for example, to use this approach in cultures neurons to identify novel mRNAs that localise to synapses.

In general, these data suggest that the occurance of highly restricted patterns of mRNA localisation, such as those exhibited by transcripts coding for cytoskeletal proteins, is not common. The genes for which such patterns have previously been described are genes of importance in the formation of structures within developing embryos and those important in the formation and maintenance of the structure of the cell itself. It is not suprising that the majority of transcripts studied do not show such tightly controlled localisation. The detection of a small number of transcript localization patterns does, however, suggest that mRNAs are transported within the cytoplasm along a number of different pathways. Table 3.15 summarises the transcript localisations obtained with each of the three vectors, giving a possible explanation for each pattern. Sequence data from

cloning genes associated with the different localisation patterns may give insight into how the cell sorts transcripts along these different pathways.

Figure 3.15: Summary of the RNA Localisation Patterns Observed In the Three Screens Carried Out, with Possible Explanations for Each.

LOCALISATION	VECTOR	CELL LINE(S)	EXPLANATION
Grainy cytoplasmic	All	Fibroblast and ES	A visualisation of RNA transport as RNPs.
Uniform cytoplasmic	5' vectors	Fibroblast and ES	Abberant RNA transport due to loss of protein binding signals.
Peri-Nuclear	pGT1.8K	Fibroblast only	Seen only in fibroblasts due to more extensive cytoplasm.
Cytoplasmic Dots	pbKnSD pGT1.8bgeo	ES Fibroblast	Visualisation of RNA export or degradation.
Nuclear (all cells)	5' vectors	Fibroblast and ES	Inefficient splicing due to insertion into an exon.
Nuclear (variable)	pGT1.8tm	ES only	*Trans*-splicing due to insertion into rRNA genes.

Chapter 4

Nuclear Localisation of *LacZ* Sequences in Gene Trap Cell Lines is Associated With Inefficient Splicing of the Fusion Transcript.

INTRODUCTION

Individual pre-mRNAs are usually difficult to detect within the nucleus, as they are subject to rapid processing and export into the cytoplasm. The first state in which RNAs destined to become mature mRNAs can be identified is pre-mRNA contained within the heterogenous nuclear RNA (hnRNA) population in the nucleoplasm. hnRNA is larger than mRNA and much less stable. hnRNA is found associated with a number of nuclear proteins as heterogeneous nuclear ribonucleoproteins (hnRNPs) Pulse-chase experiments labelling either the 5′ or the 3′ ends of hnRNA have been used to demonstrate that approximately 25% of hnRNA is rapidly processed into mRNA and exported into the cytoplasm. The remaining hnRNA is turned over entirely within the nucleus (Lewin, 1990).

The separation of total cellular RNA into nuclear and cytoplasmic RNA followed by Northern blotting can be used to detect nuclear pre-mRNA molecules corresponding to individual mRNA species. The direct visualisation of protein coding RNA species within the nucleus by the less sensitive technique of *in situ* hybridisation has only recently been achieved. The localisation of viral RNA in cells latently infected with Epstein-Barr Virus was the first study of an individual protein coding transcript within the nucleus (Lawrence et al, 1989). Since then, other RNA sequences, including those of some endogenous genes, have been detected within the nucleus (Lawrence et al, 1990; Huang and Spector, 1991;

Kopczynski and Muskavitch, 1992; Xing et al, 1993). In each case, only very small amounts of transcript have been detected, restricted to discrete areas of the nucleus. The transcripts of the fibronectin and neurotensin genes have been studied in detail using flourescence *in situ* hybridisation (Xing et al, 1993). Simultaneous RNA and DNA *in situ* hybridisation was used to detect the site of transcription of these genes as well as the nascent transcripts. Fibronectin RNA frequently accumulated in elongated tracks that overlapped the site of transcription, but also extended well beyond it. Exon and intron probes were used to demonstrate the absence of intron sequences from the transcripts in areas of the tracks spatially removed from the site of transcription. Although such tracks were not observed for the neurotensin gene whose transcript formed focal points of accumulation, the authors concluded that these tracks were a visualisation of the processing of the fibronectin pre-mRNA, with a progressive loss of intron sequences by splicing as the nascent transcript moves away from the site of transcription (Xing and Lawrence, 1993).

The *Xist* gene has been demonstrated to produce a nuclear localised RNA (Brockdorff et al, 1992; Brown et al, 1992). The *Xist* transcript is unusual in a number of ways. It is transcribed by RNA Pol II from the inactive X chromosome. and the mature transcript is unusually long (17Kb in human and 15Kb in mouse). Despite being spliced, *Xist* RNA contains no open reading frame and has not been demonstrated to associate with the translational machinery of the cell (Brockdorff et al, 1991, 1992). In human cells, the *Xist* transcript has been localised to a single dot within the nucleus, corresponding to the Barr body or inactive X chromosome. These and other data have lead to the proposal that the *Xist* transcript functions as an RNA rather than being translated into a protein, and that it has a role

in the inactivation of one copy of the X chromosome in female cells (Brockdorff et al, 1991, 1992; Kay et al, 1993; Rastan, 1994). It is not known whether the *Xist* gene is unique in producing a transcript which functions as a nuclear RNA in this way, or whether it is the first of a new class of genes to be discovered.

High concentrations of nuclear pre-mRNA can be achieved experimentally by perturbing the processing or export of the molecules. The presence of a monomethylguanosine (m7GpppN) cap structure is believed to play a role in the export of pol II transcribed RNAs, as microinjection of large amounts of free competitor m7GpppG leads to the retention of pol II transcripts in the nucleus (Hamm and Mattaj, 1990). The cap is essential for the export of snRNAs, and increases the efficiency of mRNA export. Two cap binding proteins have recently been shown to be involved in the export of capped transcripts from the nucleus. Cap binding protein 80 (CBP80) and cap binding protein 20 (CBP20) bind to the cap as a heterodimer and are involved both in the nuclear export of RNA and in the splicing of pre-mRNA. mRNA species with altered cap structures can be exported from the nucleus but with a greatly decreased efficiency, suggesting that additional factors are involved in the export process.

The importance for polyadenylation of transcripts in nuclear export is unclear. Histone pre-mRNAs with mature 3' ends formed by ribozyme action rather than the cellular 3' processing mechanism are transport deficient (Eckner et al, 1991). The influenza virus protein NS1 recognises and binds poly (A) sequences in vitro and in vivo, and inhibits the nuclear export of all poly (A) containing mRNAs in infected cells (Krug, 1993). This

suggests that the presence of poly (A) sequences may also be involved in the regulation of nuclear export of transcripts in non-infected cells.

In addition to the requirement for certain structural elements to be included in pol II transcripts for their export, it is also necessary for the intron sequences to be removed. Experiments on yeast splicing mutants have shown that the presence of intron sequences prevents the export of pre-mRNAs to the cytoplasm (Legrain and Rosbash, 1989). A synthetic intron with an open reading frame was inserted into the open reading frame of a *β-galactosidase* gene. This synthetic RNA was poorly spliced and gave rise to only a small amount of β-gal protein, which was translated from the unspliced pre-mRNA. The inefficient translation was concluded to be due to the retention of the pre-mRNA in the nucleus. The deletion of the 5′ splice junction or the branchpoint sequence, which are important for spliceosome assembly in yeast, lead to the efficient export and translation of the unspliced pre-mRNA. The introduction of the construct with both splicing sequences intact into a cell line containing a mutant U1 snRNA also led to efficient export and translation of the pre-mRNA. These results support a model whereby the formation of splicing complexes delays or prevents the export of RNA from the nucleus, with efficient transport from the nucleus occurring after the release of the spliced transcript from the spliceosome. RNA species which do not form splicing complexes due to mutations within the RNA itself or within splicing factors essential for spliceosome assembly by-pass the splicing pathway and so are exported from the nucleus by default (Zapp, 1992). Interestingly, in the study by Legrain and Roshbash, approximately 3% of the pre-mRNA containing intact splice sites was exported from the nucleus and translated, giving rise to the small amount of β-gal protein observed in the initial experiment.

This suggests that the retention of unspliced transcripts within the nucleus is not absolute, with a small amount of pre-mRNA entering the cytoplasm.

The electroporation of ES cells and fibroblasts with 5' gene trap vectors led to the isolation of several cell lines which showed a striking accumulation of *lacZ* fusion transcript within the nucleus, with only a small amount detectable in the cytoplasm. This observation is unexpected, as splicing to the *En-2* splice acceptor in the vector sequence is required to produce a fusion protein and confer neomycin resistance on the cells. This processing should lead to rapid export of the fusion transcript from the nucleus. The nuclear localisation seen in gene trap electroporated lines was thought to represent either a perturbation in the processing of the fusion transcripts in these lines or integrations into a class of endogenous genes that have nuclear localised transcripts. This chapter presents a more detailed analysis of the nuclear accumulations of *lacZ* fusion transcripts seen in gene trap electroporated cell lines, presenting possible explanations for the nuclear accumulation of RNA seen, together with a preliminary analysis of the processing of the fusion transcripts in a number of these cell lines. In each cell line studied, the major fusion transcript has been demonstrated to contain intron sequence from the vector. The use of a probe to this intron region indicates that the presence of the intron sequences in these transcripts is responsible for their retention within the nucleus. The intron region is proposed to remain in the fusion transcript by inefficient *cis*-splicing in some cell lines. In other lines, the processing of the fusion transcript appears to more complex. This will be discussed further in chapter 5.

Origins of Cell Lines Showing Nuclear Localisation of βgeo Fusion Transcript.

Two fibroblast cell lines, TβP20,8 and TβP20,29 were isolated as described in chapter 3. The ES cell lines ST416, ST478 and ST576 were derived by Dr W. C. Skarnes by electroporation of the parental cell line CGR-8 with the Secretory trap vector pGT1.8tm (figure 4.1). This vector contains a trans-membrane domain from the rat CD4 gene and has been designed to preferentially trap genes coding for secretory molecules.

Further Studies of Nuclear Localisation Using Whole Mount *in situ* Hybridisation.

Probes

Digoxigenin labelled probes were synthesised as described in chapter one. The anti-sense *lacZ* probe was transcribed using T7 bacteriophage RNA polymerase, from the plasmid pΔEK (see chapter three). The *En-2* intron probe was transcribed using T7 polymerase using the plasmid pSA1b, linearised at the BamHI site as a template (figure 4.2). These probes are of approximately the same length and both were used at a 1:200 dilution.

Hybridisations.

The *in situs* were carried out using cells plated on glass coverslips in order to increase the resolution available for studying the sub-nuclear localisation of the fusion transcript. The *En-2* intron probe was used to assess the efficiency of splicing to the *En-2* splice acceptor. Cells of each line were cultured on gelatin coated glass coverslips (see chapter two), fixed

with 4% (w/v) paraformaldehyde in PBS and taken through the whole mount procedure as described previously.

Northern Blot Analysis

Probes

A 0.8Kb *LacZ* probe template was made by digestion of pSAβgeoΔEK with ClaI (see figure 3.3). An *En-2* intron template was made by digestion of pSA1b with HindIII and BglII (figure 4.2). Both template fragments were isolated from a low melting point agarose gel and labelled with ^{32}PdCTP by random priming (see chapter 2).

Northern Blotting

Total RNA was isolated from guanidinium lysate of cell lines TβP20,8; TβP20,29; ST416 and ST576 by the caesium chloride method (see chapter 2). 10μg samples of total RNA from each cell line were loaded onto denaturing gels, together with 10μg samples of total RNA from cell lines known to use the *En-2* splice acceptor site efficiently. The gels were run, blotted and hybridised as described in chapter two. The blot of RNA from lines TβP20,8; TβP20,29 and ST416 was probed with the *LacZ* probe, stripped and re-probed with the intron probe. ST576 RNA was run on two duplicate blots and one probed with each of the two probes. The Northern blot analysis of clone ST576 was carried out by Julie Moss.

5' RACE Cloning

5μg samples of total RNA from lines TβP20,8 and ST416 were used for RACE cloning using RACE protocol A, detailed in chapter 2. Cytoplasmic RNA was prepared from lines TβP20,8; TβP20,29 and ST416 using the fractionation protocol described in chapter 2. 10μg samples of this RNA were taken through RACE protocol B. RACE products were cloned into

Bluescript II SK+ vector (Promega) as described in chapter 2. Clones were digested with Asp718 and XbaI to release the insert, run on an agarose gel and Southern blotted (see chapter 2) using an *En*-2 splice acceptor probe prepared by digestion of the plasmid p1.8HX with BamHI and BglII (figure 4.3). RACE clones that hybridised to the *En*-2 splice acceptor sequence were then sequenced using the Sequenase II DNA sequencing kit from (see chapter 2).

Figure 4.1: 5' Secretory Trap Vector pGT1.8TM (Constructed by W. C. Skarnes)

pGT1.8TM was derived from pGT1.8βgeo (figure 3.1B) by the insertion of a 0.7Kb PstI/NdeI fragment of rat *CD4* (grey box) containing the transmembrane domain (TM) in frame with *βgeo* (Skarnes et al, 1995).

The vector is linearised at the HindIII site prior to electroporation.

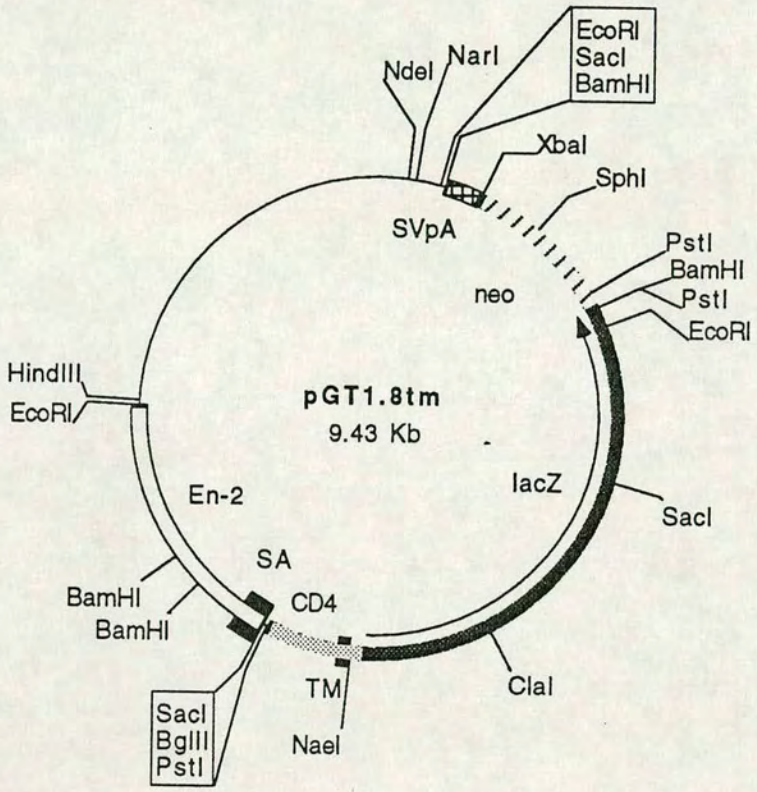

EcoRI
SacI
BamHI
NdeI
NarI
XbaI
SphI
SVpA
PstI
BamHI
PstI
EcoRI
neo
HindIII
EcoRI
pGT1.8tm
9.43 Kb
En-2
lacZ
SacI
SA
BamHI
BamHI
CD4
TM
ClaI
SacI
BglII
PstI
NaeI

Figure 4.2: Vector pSA1b. (Constructed by W. C. Skarnes)

The *En-2* splice acceptor region (white box) was sub-cloned into pBluescript II SK+ (Promega) as a HindIII/XbaI fragment (p1.8HX, see figure 4.3). ExoIII deletions were then made to remove *En-2* exon sequences, to leave 13bp of exon sequence in addition to the intron region. Linearisation of the plasmid with BamHI provides a template for transcription of an anti-sense riboprobe using T7 polymerase.

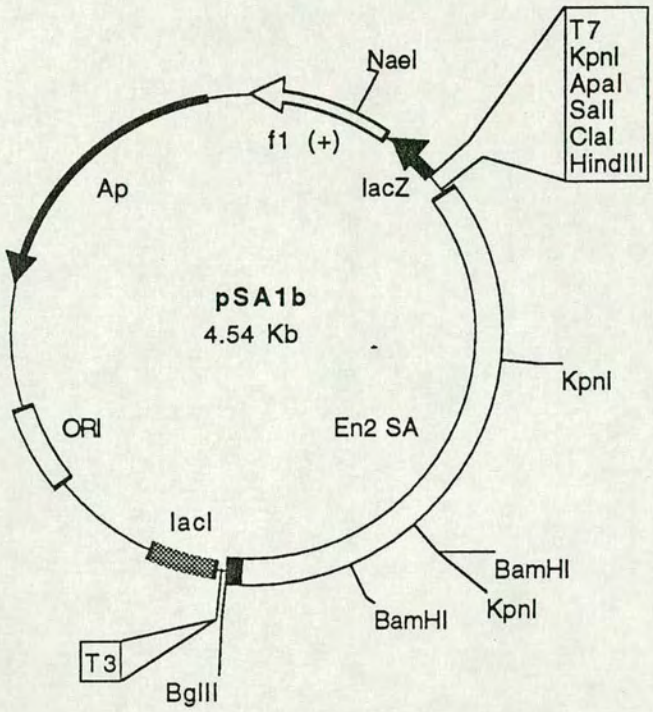
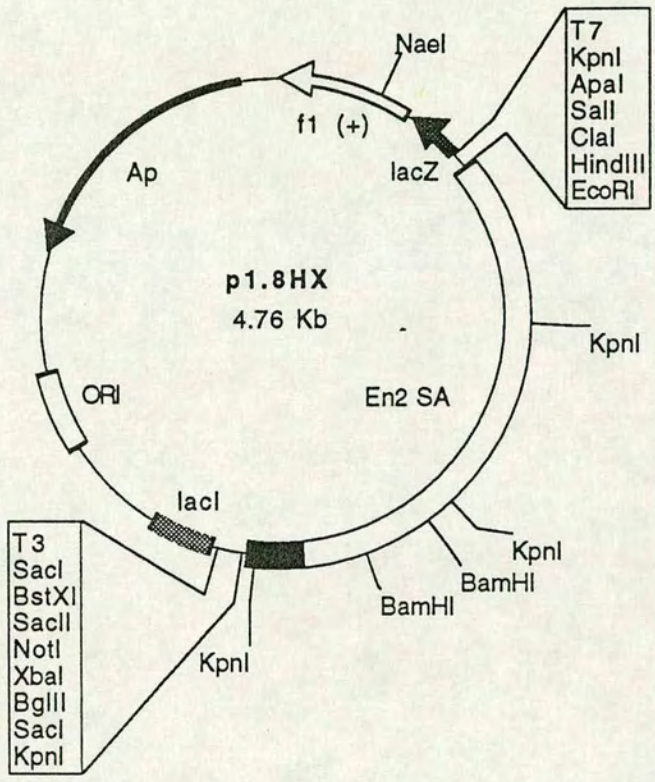

T7
KpnI
ApaI
SalI
ClaI
HindIII
NaeI
f1 (+)
Ap
lacZ
pSA1b
4.54 Kb
KpnI
ORI
En2 SA
lacI
BamHI
KpnI
BamHI
T3
BglII

Figure 4.3: Plasmid p1.8HX. (Constructed by W. C. Skarnes)

A 1.8Kb HindIII/XbaI *En-2* genomic fragment sub-cloned into pBluescript SK+. Digestion with BamHI/BglII yields a 0.5Kb template for a random primed *En-2* splice acceptor probe. This probe contains 120bp of exon sequence (denoted by a black box).


NaeI
T7
KpnI
ApaI
SalI
ClaI
HindIII
EcoRI
f1 (+)
Ap
lacZ
p1.8HX
4.76 Kb
KpnI
ORI
En2 SA
lacI
T3
SacI
BstXI
SacII
NotI
XbaI
BglII
SacI
KpnI
KpnI
BamHI
BamHI
KpnI

A Small Fraction of Gene Trap Cell Lines Show Accumulation of Fusion Transcripts in the Nucleus.

Approximately 1% of gene trap cell lines studied in the screens for sub-cellular RNA localisation patterns described in chapter three showed fusion transcript restricted predominantly to the nucleus. Lines TβP20,8 and TβP20,29 were derived from 10T1/2 fibroblasts electroporated with the vector pGT1.8K. Lines ST416, ST478 and ST576 were derived from CGR8 ES cells electroporated with the vector pGT1.8tm. In each cell line, a small amount of fusion RNA could also be detected in the cytoplasm of the cell (figure 4.4). The transcript seen in the nucleus did not show a homogeneous distribution, but was present as a number of intense speckles. Clone TβP20,8 showed speckles of varying size, with the RNA present in the cytoplasm also showing a very grainy, speckled appearance. Clone TβP20,29 shows a number of small speckles within the nucleus, together with a single, large crescent shaped dot of stain present in each nucleus. The cytoplasmic RNA in this line is barely detectable. The nuclear stain in ES cell line ST416 is made up of a very large number of very small speckles, giving a grainy appearance to the nucleus. Cytoplasmic stain is visible in cells of this line, but lacks the grainy appearance seen in cells of line TβP20,8. Cells of lines ST478 and ST576 show very similar staining patterns. The level of nuclear staining varies greatly from cell to cell, ranging from one intense dot of RNA to a large number of speckles of varying sizes. Again, cytoplasmic *lacZ* transcript is clearly visible in cells of both these lines. The average level of staining is slightly lower in line ST478 than in line ST576. The nuclear retention of the transcript suggested that these lines contained integrations into endogenous genes that have

nuclear localised transcripts, or that the processing of the fusion transcripts in these lines had been perturbed. In order to investigate this, 5′ RACE cloning was carried out using total RNA from lines TβP20,8 and ST416.

Nuclear Accumulation of Fusion Transcripts in Associated With Inefficient Splicing At the Introduced *En*-2 Splice Acceptor Site.

5'RACE cloning was carried out using total RNA from lines TβP20,8 and ST416. First strand cDNA was synthesised from the template RNA using a primer to *lacZ* sequence from the gene trap vector. These products were tailed with A residues and second strand cDNA synthesis primed using a T tailed primer containing an XbaI site to facilitate cloning of the final products (primer 56). Polymerase chain reaction (PCR) was then carried out using primer 59, which has the same sequence as primer 56, but without the T tail, and a primer to the *En*-2 region of the vector, containing a KpnI site. This reaction synthesises a double stranded DNA fragment containing 120bp of *En*-2 exon sequence and the sequence found upstream of the *En*-2 splice acceptor site in the template RNA. The regions of interest are flanked by restriction enzyme sites (XbaI 5' and KpnI 3') to allow directional cloning into pBluescript SK+ (Promega). The PCR products were size selected and the fraction from 400bp to 700bp was used for cloning. A sample of each clone was digested with XbaI and KpnI and used for Southern blotting with an 120bp*En*-2 splice acceptor probe to identify genuine clones. No *En*-2 positive clones were obtained for line ST416. Line TβP20,8 gave ten clones in total of which six hybridised to *En*-2 sequence. Two of these were sequenced and found to contain the expected 120bp of *En*-2 exon sequence joined to *En*-2 intron sequences demonstrating that unspliced fusion transcripts are present in this line (figure 4.5). This result was unlikely to be a result of DNA contamination of the starting population of RNA as the

En-2 intron sequence contains no poly A tracks that may otherwise lead to mis-priming of second strand cDNA synthesis.

5′ RACE cloning results demonstrated that in line TβP20,8 which showed nuclear accumulation of *lacZ* transcript, some of the fusion transcript was unspliced. The presence of intron sequences in pre-mRNA has been implicated in its retention within the nucleus. To investigate whether the fusion transcripts present in the nuclei of these lines were unspliced, whole mount *in situ* hybridistion using a probe to detect *En-2* intron sequence was used to examine the localisation of unspliced fusion transcript relative to that of the total spliced and unspliced *βgeo* transcript (figure 4.4). Staining with this probe was less intense than the lacZ staining in each cell line. Intron-containing transcripts were confined to the nucleus with no cytoplasmic transcript containing intron sequence being detected in the lines tested. The intron probe gave a very similar localisation pattern to that seen with the *lacZ* probe, but the intron containing transcript appeared to be more tightly restricted to the intra-nuclear speckles. In the two fibroblast lines TβP20,8 and TβP20,29, the unspliced transcript tended to accumulate as a single large dot.

Evidence For Two Classes of Insertions Leading To Inefficient Splicing of the Gene Trap Vector.

Total RNA from a number of the nuclear localised cell lines was used for Northern blot analysis using a *lacZ* probe and an *En-2* intron probe. The results of this analysis can be divided into two classes:

1) Discrete Bands Seen With Both Probes (figure 4.6).

Total RNA from lines ST416; TβP20,8 and TβP20,29 each contained a large unspliced primary transcript that hybridised to both probes and one or two

smaller transcripts that gave signal only with the *lacZ* probe and did not contain *En-2* intron sequence. Control cell lines, known to use the *En-2* splice acceptor efficiently, do not give any signal with the intron probe.

2) Heterogeneous Smears Seen With Both Probes (figure 4.7).

The signal given by the *lacZ* probe in line ST576 was a large smear, with an intense band, representing a major *lacZ* containing species at the top. The signal seen with the intron probe was very similar, but of slightly lower intensity

5' RACE Cloning From Cytoplasmic RNA From Lines TβP20,8; TβP20,29 and ST416.

Northern blot analysis demonstrated that these three cell lines contain a small amount of fusion transcripts that are spliced at the *En-2* splice acceptor site. Cytoplasmic RNA was isolated from lines TβP20,8; TβP20,29 and ST416 to enrich for properly spliced mRNAs. This cytoplasmic RNA was used for further 5'RACE cloning to identify the insertion site of the vector in these lines. The input of RNA was increased to 10μg for the second cloning experiment. Cloning from lines TβP20,8 and TβP20,29 was unsucessful. Four RACE clones containing *En-2* exon sequence were obtained from line ST416. Two of these clones also hybridised to the *En-2* intron probe, so represented unspliced fusion transcript. The remaining two were sequenced and found to be independent clones of the same gene, both joined correctly to the *En-2* splice acceptor (figure 4.8). The sequence showed no homology to previously cloned genes in the data base, suggesting that the integration in line ST416 is into a novel gene. Translation of the novel sequence confirmed that it is in the same reading frame as the vector sequence. The recovery of unspliced clones from

cytoplasmic RNA is indicative either of contamination of the cytoplasmic RNA preparation with nuclear RNA, or of the presence of some unspliced fusion transcript in the cytoplasm. Although no intron containing fusion transcript was detected in the cytoplasm by whole mount *in situ*, a small amount of pre-mRNA is known to be exported into the cytoplasm (Legrain and Roshbash, 1989).

The experiments presented in this chapter implicate inefficient splicing of the *βgeo* fusion transcripts as the reason for their accumulation within the nuclei of the cell lines studied. The data are consistent with the view that immature pol II transcripts are retained in the nucleus of the cell, with only fully processed transcripts being exported from the nucleus. The integrations leading to this inefficient splicing appear to fall into two catagories, each resulting from unexpected sites of vector insertion. The observation of intra-nuclear speckles of fusion transcript suggests that partially processed RNAs are not free to diffuse within the nuceoplasm, but are in some way spatially restricted. These observations have implications for models of the spatial organisation of mRNA splicing within the mammalian nucleus.

Figure 4.4: Nuclear Localisation of Unspliced *lacZ* Fusion Transcripts in Fibroblast and Embryonic Stem Cells.

Whole mount *in situ* hybridisations were carried out using *lacZ* and *En-2* intron anti sense riboprobes on cells grown on glass coverslips. Cells were stained overnight and viewed using an Olympus Vanox microscope. Photographs were taken using Kodak GPF160 Ektacolor gold II color negative film.

A) Fibroblast Line TβP20,8. Nuclei of this line were counterstained using 1μM Hoechst 33285, as the cells had a very flat morphology.

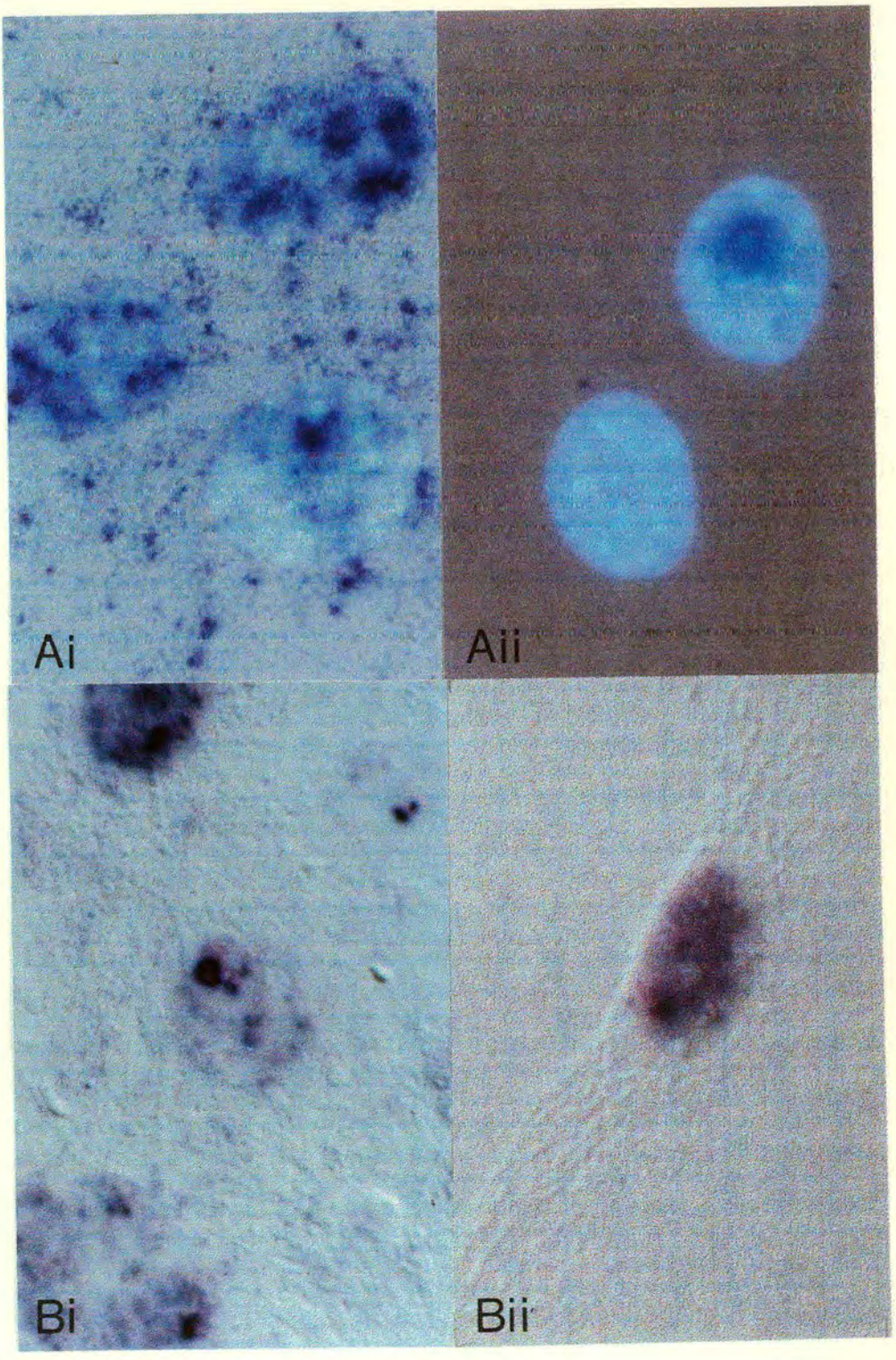
i) *LacZ* Probe. A highly granular stain was seen, restricted mostly to the nucleus, with only a small amount of signal seen in the cytoplasm.

ii) Intron Probe. Signal with this probe was entirely restricted to the nucleus, showing a single large accumulation in most nuclei.

B) Fibroblast Line TβP20,29

i) *LacZ* Probe. This probe gave a predominantly nuclear stain, with a large crescent shaped accumulation visible in each nucleus. A small amount of granular cytoplasmic signal was also seen.

ii) Intron Probe. The signal seen with this probe was entirely nuclear, showing a similar crescent shaped accumulation as seen with the *lacZ* probe.


Ai
Aii
Bi
Bii

C) ES Cell Line ST576

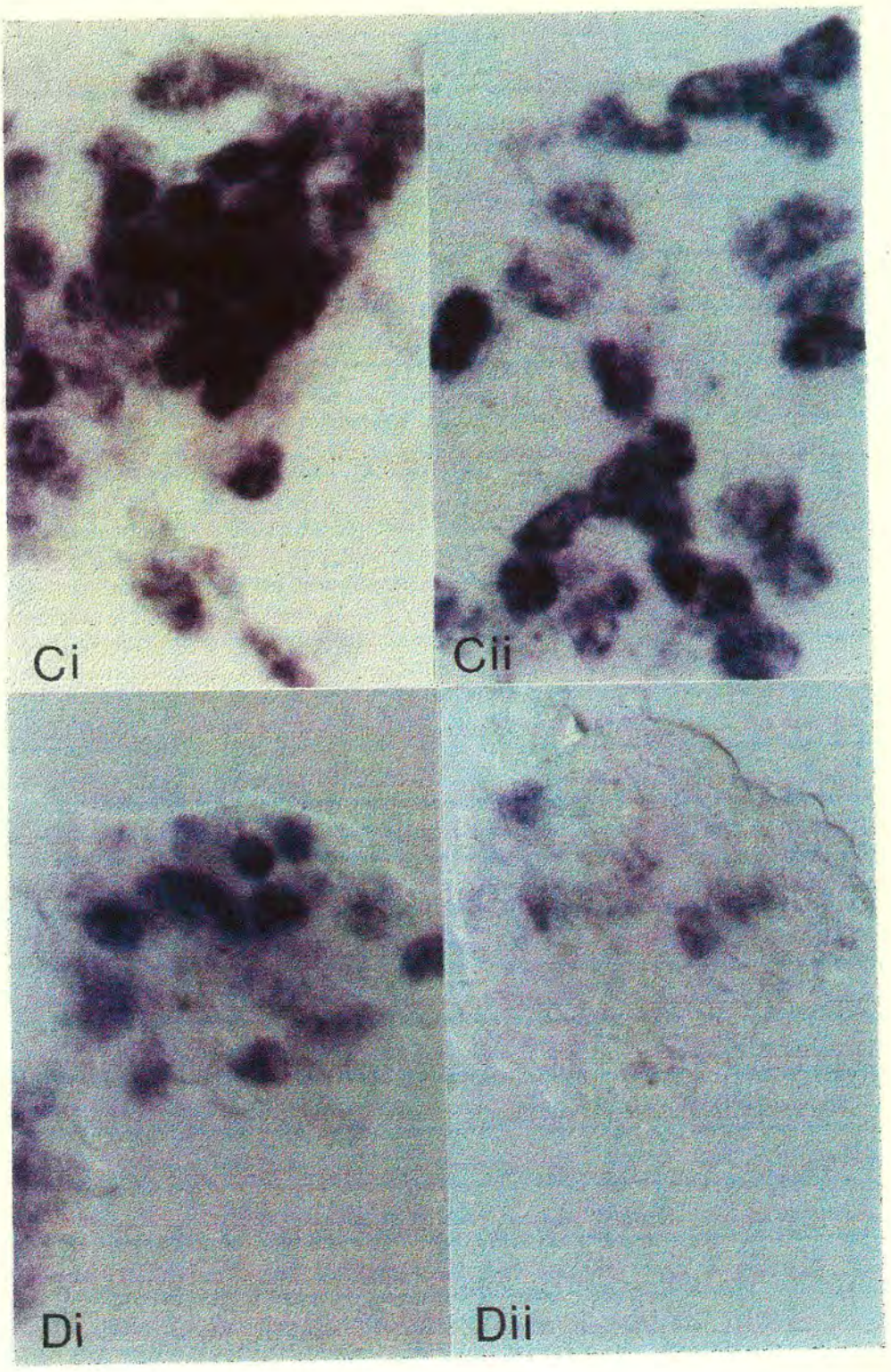
i) *LacZ* Probe. The staining was predominantly nuclear, showing speckles of stain, but with significant amounts of cytoplasmic signal also detected. In contrast to the two fibroblast lines described above, the staining in line ST576 was highly variable with some nuclei containing undetectable amounts of fusion transcript, and some staining very darkly.

ii) Intron Probe. The nuclear staining seen with this probe was very similar to that seen with the *lacZ* probe, but at a slightly lower intensity. No cytoplasmic signal was detected with this probe.

D) ES Cell Line ST478

i) *LacZ* Probe. This cell line showed a staining pattern virtually identical to ST576 above, but at a slighly lower intensity.

ii) Intron Probe. Again, the staining in this line was almost indistinguishable from that of line ST576, but at a lower level.


Ci
Cii
Di
Dii

E) ES Cell Line ST416

i) *LacZ* Probe. This probe gave very intense staining in all nuclei of this cell line, with a small amount also detectable in the cytoplasm.

ii) *LacZ* Anti- Sense Probe. This probe serves as a negative control and gave no signal after overnight staining

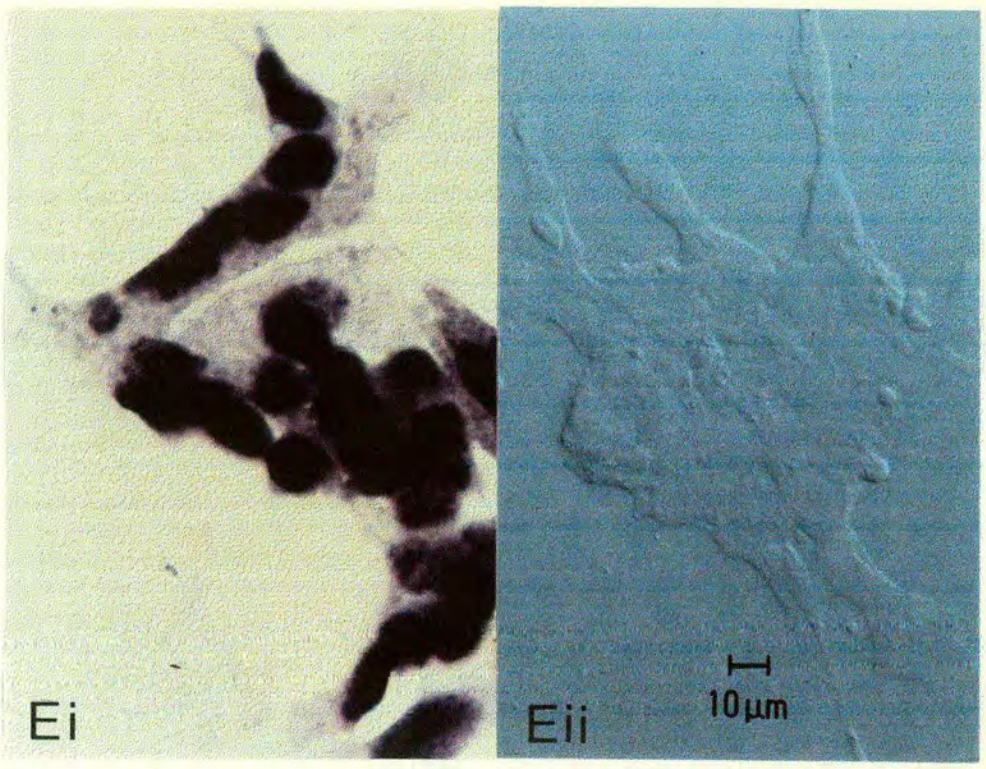

Ei
Eii
10μm

Figure 4.5: 5' RACE Clones Obtained From Total RNA From Line TβP20,8.

CLONE	SEQUENCE	IDENTITY
TbP20,8.6	ctccagcaaccagtaacctctgccctttctcct ccatgaxaaccag	**Unspliced *En*-2 Splice Acceptor**
TbP20,8.7	gcgttggttgtggataagtagctagactccagc aaccagtaacctctgccctttctcctccatgax aaccag	**Unspliced *En*-2 Splice Acceptor**

Figure 4.6: Northern Blots of Total RNA From Lines TβP20,8; TβP20,29 and ST416.

A) *LacZ* Probe

A Northern blot of total RNA from lines TβP20,8; TβP20,29 and ST416 was probed with a *lacZ* random primed probe. RNA from two control cell lines was also used. These lines had each been demonstrated by 5' RACE cloning to contain a single, correctly spliced fusion transcript. Autoradiography film was exposed to the blot for 48 hours. Each of the two control lines shows a single band, indicating the presence of a single *lacZ* containing transcript. Lines TβP20,8; TβP20,29 and ST416 each show a number of bands, indicating the presence of more than one *lacZ* containing transcript. In each line, the major transcript band is around 9.8Kb in length.

B) Intron Probe

The blot was stripped and re-probed with a random primed probe to *En-2* intron sequence. The two control lines gave no signal with this probe, as expected when the *En-2* splice acceptor is used efficiently. In lines TβP20,8; TβP20,29 and ST416, the major *lacZ* containing transcript band hybridised to the *En-2* intron probe, indicating that the major transcript bands in these lines are incompletely spliced and contain intron sequence from the vector.

TbP20,8
TbP20,29
ST416
Control
Control
9.8
6.4
5.2
10.2
6.4
5.2
A
TbP20,8
TbP20,29
ST416
Control
Control
9.8
B

Figure 4.7: Northern Blots of Total RNA From Line ST576

Duplicate Northern blots were made, each containing two samples of total RNA from line ST576 and one sample from a control cell line which was known to produce a single correctly spliced fusion transcript. One blot was probed with a random primed *lacZ* probe, the other with a random primed *En-2* intron probe. Autoradiographic film was exposed to the blots for 72 hours.

A) *LacZ* Probe

The control cell line contained a single discrete band that hybridised to the *lacZ* probe. ST576 gave a smear of *lacZ* signal, beginning with a band of transcript at about 10Kb.

B) *En-2* Intron Probe

The control cell line gave no signal with this probe. Line ST576 gave a smear of signal similar to that seen with the *lacZ* probe, but less intense and extending a shorter distance down the gel.

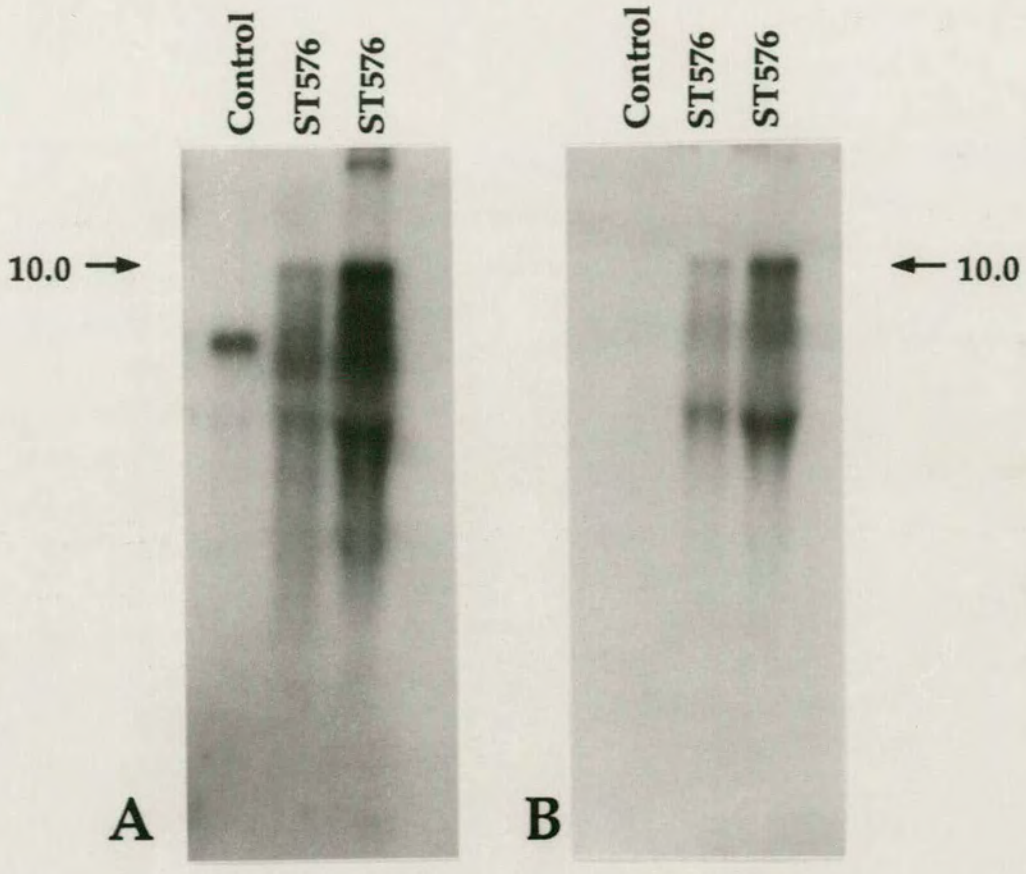
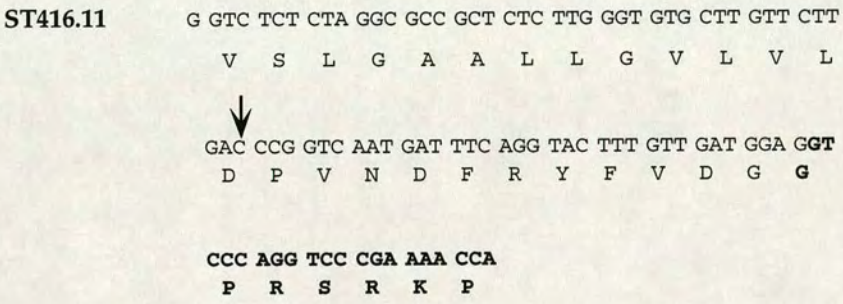

Control
ST576
ST576
10.0
A
Control
ST576
ST576
10.0
B

Figure 4.8: 5' RACE Clones From Cytoplasmic RNA From Line ST416.



The nucleotide sequence is given with its translation underneath. Bold type denotes *En*-2 sequence, confirming that the sequence cloned from line ST416 cytoplasmic RNA has the same reading frame as the vector coding sequence. The end point of clone ST416.8 is marked with an arrow. The two sequences obtained represent independent clones from the same novel gene.

The Retention of Fusion Transcripts Within the Nucleus is Associated with the Presence of Intron Sequences.

Whole mount *in situ* hybridisation has demonstrated that gene trap lines showing nuclear localisation of the *lacZ* fusion transcript contain intron sequences from the gene trap vector in a large proportion of the *lacZ* containing transcripts. These *in-situs* also indicate that the small amount of *lacZ* transcript found in the cytoplasm does not contain intron sequence. This observation has been confirmed by Northern blotting, which also divided the cell lines into two distinct classes, one producing a limited number of *lacZ* containing transcripts of distinct size, and the other producing a wide range of different *lacZ* containing transcripts leading to a smear on a Northern blot

Previously published experiments have demonstrated that transcripts containing intron sequences or lacking a 5′ cap are prevented from leaving the nucleus (Hamm and Mattaj, 1990; Legrain and Roshbash, 1989). Transcripts containing intron sequences are believed to be actively retained within the nucleus by the formation of interactions between splicing signals within the intron and splicing factors of the spliceosome. Upon release of the spliced transcript from the spliceosome, the mature mRNA becomes available as a substrate for the RNA export pathway. In the case of transcripts lacking a cap, the retention has been demonstrated to be due to the failure of these molecules to form associations with cap binding proteins involved in the regulation of RNA export. Additionally, the cap has been shown to play a role in pre-mRNA splicing, with its absence

leading to inefficient splicing and, hence, nuclear retention of uncapped transcripts

The nuclear retention of intron-containing *βgeo* transcripts in the gene trap cell lines studied is suggestive of some perturbation in the processing of the fusion transcripts in these lines. The absence of detectable quantities of unspliced fusion RNA from the cytoplasm suggests that the transcripts are not suitable substrates for nucleocytoplasmic transport. This may be due solely to their being unspliced, or to a combination of this with another factor, such as the absence of a 5′ cap.

The gene trap technique relies on the integration of the vector into the intron of an endogenous pol II transcription unit to allow efficient splicing between the introduce *En-2* splice acceptor and an endogenous splice donor (as described in chapter 3). It is possible that the vector may integrate either into the exon of a pol II gene, or into a non-pol II transcription unit. The integration of the vector into the exon of a pol II transcribed, protein coding gene would lead to competition between the introduced splice acceptor and the endogenous splice acceptor for splicing to the upstream splice donor (figure 4.9). Use of the introduced splice acceptor would produce a translatable fusion transcript. Use of the endogenous splice acceptor would lead to a fusion transcript containing intron sequences which can be predicted mostly to remain in the nucleus. No protein could be translated from such a transcript, as the *En-2* intron sequence contains stop codons in all three frames. Alternatively, the entire altered exon may be skipped, leading to a transcript containing no vector sequences which would not be detected by our analyses. The presence of a small number of *lacZ* containing transcripts in lines TβP20,8; TβP20,29 and ST416, with the

largest transcript in each line also containing intron sequences, suggests that this may be the explanation for the inefficient splicing of transcripts in these lines. RACE cloning of two independant clones representing the same endogenous gene from ST416 cytoplasmic RNA also lends support to this model. Unpublished work using cell lines electroporated with the secretory trap vector may be useful in proving this as a way of retaining intron sequences in *lacZ* fusion transcripts. Line ST567 has been demonstrated to contain an insertion into the previously cloned gene *embigin* (D.Townley, unpublished). Chimeric embryos containing cells from this line show a *lacZ* protein staining pattern consistent with the published expression pattern for the *embigin* gene (Huang et al, 1990). 5' RACE cloning from a second line, ST402, gave only unspliced clones. Cells from line ST402 have been transmitted through the germline and resultant embryos stained for LacZ protein activity (W.C.Skarnes, unpublished). The staining pattern seen in these embryos at 8.5dpc was identical to the pattern observed in ST567 chimaeras, showing restriction to the brain, visceral yolk sac and presumptive foregut. Further analysis of the insertion in line ST402 is required, but these early data suggest that this may represent a second insertion into the *embigin* gene, possibly within an exon.

Insertion of the vector into a non-pol II gene would lead to the formation of a transcription unit in which the splice acceptor site from *En-2* has no consensus splice donor site to splice to. The *lacZ* and neomycin coding regions would also, presumably, be placed under the control of pol I or pol III, rather than being transcribed by the usual pol II mechanism (figure 4.10). If transcripts can be produced from such an integration, they will be uncapped and lack an ATG translation initiation codon, so are unlikely to be translated to produce a fusion protein. The implications of this type of

integration for splicing of the fusion transcript are, however, complex and are explored more fully in chapter five.

The Intron-Containing Transcript is Present in The Nucleus as a Number of Intra-nuclear Speckles.

A number of studies have documented intra-nuclear speckles associated with the transcription and processing of mRNA. Total nascent transcript has been demonstrated to accumulate in numerous speckles within the nucleus. The number of speckles detected varies, with Jackson et al (1993) reporting 300 to 400 speckles per nucleus in HeLa cells, and Wansink et al (1993) reporting around 100 speckles per nucleus in T24 human blader carcinoma cells. Poly (A)+ RNA has been demonstrated to concentrate primarily within 20 to 50 discrete transcript domains in primary fibroblasts and myoblasts (Carter et al, 1991). The differences in these numbers may be due to the cell types used. The sub-nuclear localisation of snRNPs of the splicing complex has been investigating using autoimmune antibodies (Spector, 1990). The snRNPs studied formed 20 to 50 domains, with small amounts also detectable in the surrounding nucleoplasm. The non-snRNP splicing factor SC-35 is also concentrated in these splicing factor rich nuclear domains (Fu and Maniatis, 1990; Carmo-Fonseca et al, 1991; Spector et al, 1991; Huang and Spector, 1992 and Carter et al, 1993). The identity of these transcript domains and splicing factor rich regions has remained elusive, with early attempts to co-localise splicing factors and transcripts within the nucleus producing conficting results. Wansink et al detected no correlation between the domains of RNA transcription seen in T24 cells and accumulations of the splicing factor SC-35 seen in the same cells. Jackson et al, however, reported that most of the transcript domains seen in HeLa cells also contain a component of the splicing apparatus detected by

an anti-Sm antibody. Poly (A)+ RNA-rich domains in fibroblasts have also been demonstrated to be co-incident with snRNP and SC-35 rich regions (Carter et al 1991; 1993). The current view is that the sites of transcription and processing of transcripts are closely linked, rather than strictly co-incident, with each transcript domain representing a cluster of active gene loci, around which protein and RNA factors necessary for the processing of the transcript accumulate (Hendzell and Bazett-Jones, 1995).

The speckles of *lacZ* fusion transcript seen by whole mount *in situ* in the nuclei of lines ST416, TβP20,8 and TβP20,29 are visually very similar to the accumulations of splicing factors previously reported. Since I have demonstrated that the splicing of the fusion transcripts in these clones is inefficient, it is likely that the accumulations of transcript seen represent pre-mRNA molecules in association with components of the splicing machinery. The use of antibodies to splicing factors in conjunction with RNA *in situ* hybridisation could be used to investigate further the nature of the speckles of fusion transcript seen.

Implication of Data Presented for the Study of Intranuclear RNA Transport.

Two major models have been proposed for the spatial control of splicing within the nucleus. The transcripts from the Epstein-Barr virus genome; the *c-fos* gene and the *fibronectin* gene have been visualised within the nucleus as curvilinear tracks (Lawrence et al, 1989; Huang and Spector, 1991; Xing and Lawrence, 1991 and 1993). A detailed examination of the track obtained for the *fibronectin* gene revealed that intron sequences are further restricted to one end of the track. Exon sequences were also demonstrated to co-localise with the *fibronectin* gene (Xing and Lawrence,

1993). These data have been used to propose a model whereby pre-mRNA is transcribed, then transported along a spatially defined track towards the nuclear membrane. Processing of the transcript is proposed to occur within this track. The same study also looked at the transcript of the *neurotensin* gene, which gave a single site of hybridisation, co-incident with the *neurotensin* gene. A track of transcript molecules was not detected for this gene. The much smaller size of the *neurotensin* gene, together with the limitation of the technique used to the detection of steady state RNA can be used to explain the failure to detect a track in this case (Rosbash and Singer, 1993).

An alternative model, proposed by Zachar et al (1993), used studies of *Drosophila* polytene nuclei, over-expressing the *lac1* gene. In this study, a high concentration of *lac1* RNA was seen, co-incident with its template DNA. Transcript was also detected at lower levels throughout the nucleus, apparently contained within an extrachromosomal channel network (ECN). The use of intron and exon probes revealed that the ECN contains both incompletely and fully spliced *lac1* RNA species. Movement of transcript was also studied, suggesting that transcripts travel through these channels by diffusion. These data were used to propose a model whereby RNA molecules move from their sites of transcription by diffusion, with most or all pre-mRNAs using the same route of intranuclear movement. Splicing is predicted to occur both co-transcriptionally, and at sites spatially and temporally removed from the site of transcription.

The localisation of *lacZ* fusion RNA in line TβP20,29 at one major site within the nucleus is consistent with either one of these models. The transcript localisation seen in lines TβP20,8; ST416; ST576 and ST478 are

more difficult to explain, as the transcript visualised is transcribed from a single vector integration site, yet a large number of focal accumulations of RNA are seen. The retention of such a high concentration of transcript molecules within the nucleus is, however, a highly abnormal situation. In cells from lines ST576 and ST478, which show varied levels of expression of *lacZ*, a number of cells can be seen which contain a single dot of signal. This suggests that the fusion transcript may initially be restricted to an area of the nucleus close to its site of transcription, with overexpression causing saturation of some mechanism involved with this restriction, leading to a more widespread distribution of the transcript. Further experiments would be needed to test this hypothesis. Attempts to co-localise the single dot of transcript in line TβP20,29 and some cells of lines ST478 and ST576 with vector DNA would be useful, as would studies to determine which other nuclear components are present in the speckles of transcript seen in each line. The data presented in this chapter suggest that the fusion transcripts are not free to diffuse within the nucleus, as the signal is restricted to defined speckles of high intensity. The observation of numerous speckles produced by transcripts from a single transcription unit, however, argues against the tight spatial restriction of RNA species proposed by Lawrence et al. A model in which pre-mRNAs are preferentially transported along a defined route through the nucleus, but with some degree of flexibility allowing for the movement of transcripts to other regions of the nucleus would fit with the results presented here.

In lines ST416, ST478 and ST576, fusion transcripts containing the *En-2* intron from the vector showed a very similar distribution to that of the total *lacZ* containing transcript, with the major difference being an absence of intron-containing RNA from the cytoplasm. In lines TβP20,29 and

TβP20,8, however, transcript molecules containing the *En-2* intron tended to form a single accumulation within each nucleus, whereas the total *lacZ* containing transcript formed a number of intranuclear speckles. This suggests that a further delay in splicing or export occurs in these lines following the removal of the introduced *En-2* intron. It is possible that the integrations in these lines have disrupted the splicing of other introns within the genes at the sites of insertion. The cloning of the genes into which the vector has inserted in these lines would be necessary to investigate this observation further.

The data presented describe a number of gene trap electroporated cell lines in which the majority of the fusion transcript is retained in the nucleus. Intron probes have been used for whole mount *in situ* hybridisation and Northern blot analyses, demonstrating that these cell lines contain integrations that lead to inefficient splicing of the fusion transcript, rather than integrations into genes whose transcripts are localised to the nucleus in wild type cells. The integration events in these cells can be divided into two categories, each of which reflects unexpected behaviour of the gene trap vector. Lines ST416, TβP20,8 and TβP20,29 appear to contain integrations into exons, rather than introns, of pol II transcribed genes. Although not conclusively proven, this explanation is in agreement with all of the data available at this time. Lines ST478 and ST576 show highly unusual processing of the fusion transcripts. This is attributable to the presence of gene trap insertions into non pol II transcription units, and will be explained more fully in chapter 5.

Figure 4.9: Insertion Of The Vector pGT1.8TM Into The Exon Of A Pol II Gene.

A) Structure of pGT1.8TM. The intron (thick black line) and splice acceptor region (grey box) are from the mouse *engrailed-2* gene, as for previous vectors. The presence of a transmembrane domain from the mouse *CD4* gene (black box) is a modification to detect insertions into genes for secreted proteins (Skarnes et al, 1995). The *βgeo* (striped box) and SV40 poladenylation (grey box) regions are the same as for vector pGT1.8βgeo.

B) Predicted structure of the insertion of this vector into the exon of an endogenous pol II gene. Endogenous exons are represented by white boxes, endogenous introns by thin black lines. Competition would occur between the endogenous splice acceptor (SAi) and the splice acceptor from the vector (SAii) for splicing with the upstream endogenous splice donor (SD). This is predicted to lead to the production of two types of fusion transcript:

C) Intron containing transcript. The use of the endogenous splice acceptor (SAi) would define the region from SAi to the polyadenylation site of the vector as the last exon of the gene, leading to the inclusion of *En-2* intron regions in the transcript. The presence of the vector splice acceptor site and branch point sequence in this transcript would lead to its retention in the nucleus. No functional fusion protein could be translated from this transcript as the *En-2* intron sequence contains stop codons in all three frames.

D) Non-Intron Containing Transcript. The use of the splice acceptor site from the vector (SAii) would remove the partial endogenous exon from the transcript together with the *En-2* intron regions. This transcript is predicted to be exported from the nucleus and translated, provided the insertion is in the correct reading frame.

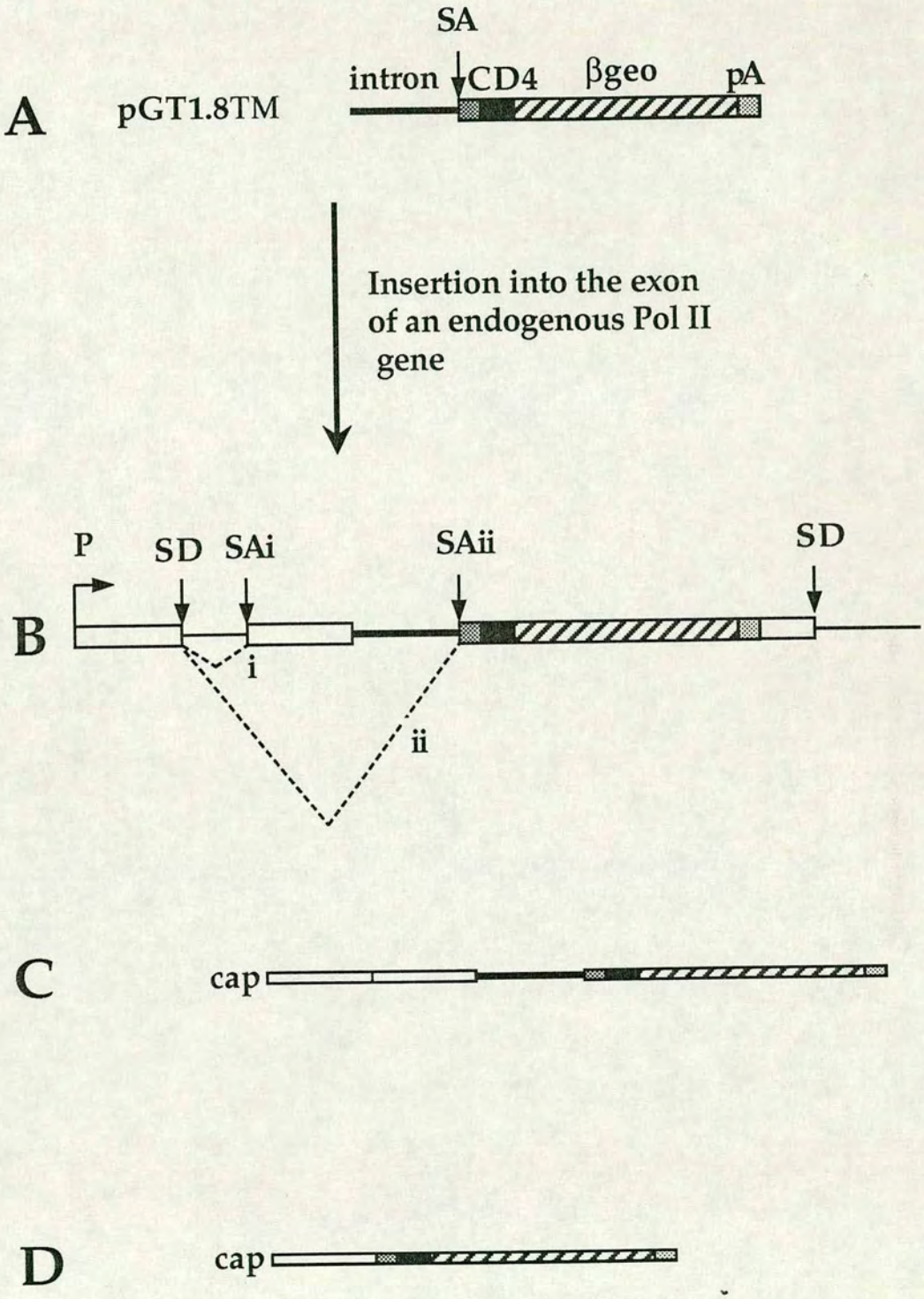
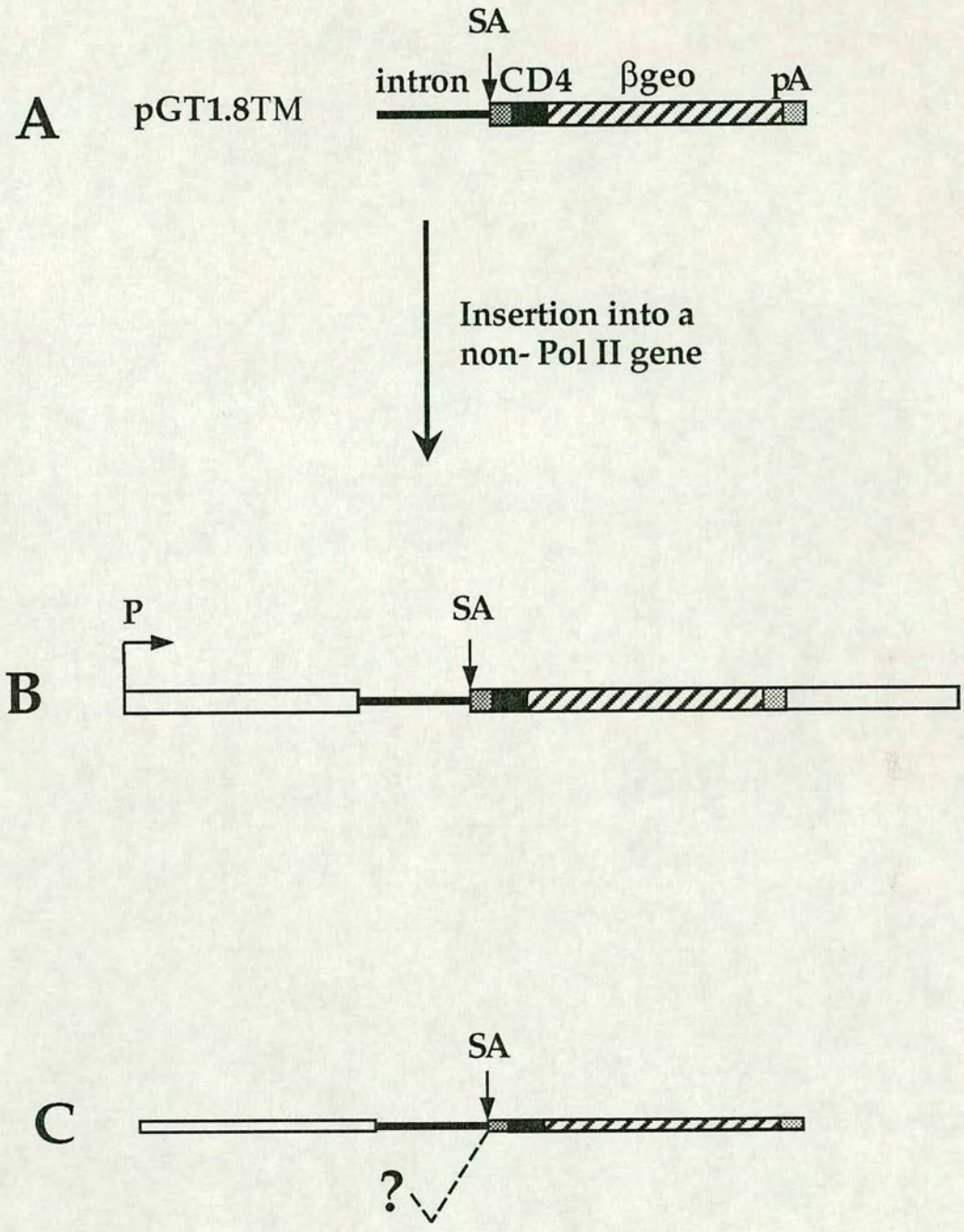

SA
intron
CD4
βgeo
pA
A
pGT1.8TM
Insertion into the exon
of an endogenous Pol II
gene
P
SD
SAi
SAii
SD
B
i
ii
C
cap
D
cap

Figure 4.10: Insertion of the Vector pGT1.8TM Into a Non- pol II Transcription Unit.

A) Structure of pGT1.8TM. See figure 4.9 for full description.

B) Predicted structure of the insertion of this vector into a non- pol II transcription unit. Endogenous sequences are represented by white boxes.

C) Predicted primary transcript produced from this insertion. The transcript produced would not have a 5' cap structure, as it is not a pol II transcript. The En-2 splice acceptor would have no consensus upstream splice donor sequence. If such a transcript is spliced, the choice of splice donor is not obvious. This type of integration is explored in detail in chapter 5.


SA
A
pGT1.8TM
intron
CD4
βgeo
pA
Insertion into a
non- Pol II gene
P
SA
B
SA
C
?

CHAPTER 5

Trans- Splicing in Mouse Embryonic Stem Cells as Revealed by Gene Trap Integrations into Ribosomal RNA Genes.

INTRODUCTION

Trans- splicing of pre-mRNAs has been documented in a number of lower eukaryotes. In trypanosomes, all transcripts are processed by *trans*-splicing (Van der Ploeg, 1986; Borst, 1986), while in nematodes and trematodes, *cis*- and *trans*- splicing reactions are carried out both in the same cell and even within the same pre-mRNA molecules. As described in chapter one, the mechanistic aspects of *cis*- and *trans*- splicing are very similar and most of the same splicing factors are used for both. Additionally, *trans*- and *cis*- splice sites have the same consensus sequences. This poses a potential problem in the processing of transcripts in organisms that carry out both types of processing. The splicing machinery must discriminate between 3' splice sites destined for *cis*-splicing and those destined for *trans*- splicing.

Conrad et al (1991) investigated the genomic structure of the *trans*- spliced *C-elegans* gene *rol-6*. This gene begins with a structure termed an 'outron' which consists of a 3' splice site, with intron-like sequences upstream, but no consensus 5' splice site. The introduction of a 5' splice site 50bp upstream of the *rol-6* 3' *trans*- splice site led to *cis*- splicing. Thus, it appears that there is no specific signal for *trans*- splicing, but that exons that are *trans*- spliced share a common genomic structure.

An additional level of complexity is introduced by the presence of a second spliced leader RNA, SL2 in *C. elegans* (Huang and Hirsh, 1989). A number of *C.elegans* genes are transcribed as polycistronic messages, similar to bacterial operons (Speith et al, 1993; Zorio et al, 1994). Genes which occupy a downstream site in these polycistronic transcripts receive SL2 by *trans*-splicing, whereas the most 5' gene in each polycistronic transcription unit receives SL1. Again, this suggests that the specificity of the splicing machinery in *C. elegans* is governed by genomic structure, rather than sequence based signals.

Trans- splicing of polycistronic messages in trypanosomes and *C.elegans* provides a mechanism by which these transcripts obtain a cap structure at their 5' ends, essential for their subsequent export into the cytoplasm and translation. Experiments using nuclear extracts from *Trypanosoma cruzi* have demonstrated that, at least in vitro, capping of trypanosome RNAs is restricted to the SL RNA and small nuclear RNAs (Zwierzynski and Buck, 1990). The mRNAs for the trypanosome variant-specific surface glycoproteins (VSGs) and procyclic acid repetitive protein (PARP) are believed to be transcribed by RNA pol I and capped by *trans*- splicing (Zomerdijk et al,1991; Rudenko et al, 1991). In this way, *trans*- splicing also provides a mechanism whereby RNAs transcribed by polymerases other than pol II can be used as messenger RNAs. In *C. elegans*, the spliced leader RNAs have trimethylguanosine caps, rather than the monomethylguanosine cap found on most mRNAs. Mature *trans*-spliced messages have been shown to retain this unusual TMG cap structure (Liou and Blumenthal, 1990; Van Doren and Hirsh, 1990) although the significance of this is unclear. *Trans*- splicing can, therefore,

be regarded as a form of *trans-* capping of transcripts which would otherwise not be translated.

The occurrence of *trans-* splicing in higher eukaryotes, particularly mammals has long been a source of debate. Computer searches using a canonical *trans-* splicing structure, designed by reference to known trans-splicing structures, predict that *trans-* splicing may occur in a wide range of vertebrates, including mammals (Dandeker and Sibbald, 1990). *Trans-*splicing of modified adenovirus and *β-globin* transcripts has been documented in cell free splicing systems (Solnick, 1985; Konarska et al, 1985). In these studies, it was demonstrated that the efficiency of *trans-*splicing can be increased by the presence of a small amount of sequence complementarity between the two *trans-* splicing substrates. More recently, an oligonucleotide containing a 5' splice site has been shown to *trans*-splice to a molecule containing a 3' splice site with either a 5' splice site or a splicing enhancer sequence immediately downstream (Chiara and Reed, 1995). Spliced leader RNAs from nematodes can be accurately spliced to nematode or mammalian 3' splice sites in COS cells, confirming that mammalian cells can carry out *trans-* splicing reactions in vivo (Bruzik and Maniatis, 1992).

Trans- splicing of endogenous transcripts by mammalian cells has proved more refractory to investigation. A number of studies have implicated *trans*-splicing in the double isotype production seen in B lymphocytes. Immunoglobulin class switching is a phenomenon whereby a single clone of B-cells synthesises IgM, and subsequently synthesises another immunoglobulin isotype with the same antigen specificity. In certain B-cells, two different isotypes with the same specificity are produced

simultaneously. This is known as double isotype production. Chen et al (1986) reported a neoplastic B-cell line, BCL_1B_1, in which both IgM and IgG1 were produced, apparently from the same unrearranged chromosome. This study proposed a method of complex alternative termination and splicing to account for their results, but also suggested a method of discontinuous transcription, with the polymerase translocating from one DNA template to another during transcription. At the time this paper was written, it was believed that trypanosomes produced their mRNAs by discontinuous transcription. A more recent study of BCL1 cell lines invoked either *trans*-splicing or an RNA ligation based mechanism to account for the double isotype production (Nolan-Willard et al, 1991). The BCL_1 cell lines are, however, neoplastic, and are known to have undergone an unusual translocation between chromosome 12, from which the μ and γ1 heavy chains are expressed, and chromosome 16.

Shimizu et al (1989, 1991) used a transgenic mouse model TG.SA containing a rearranged human gene for IgM to study class switching and double isotype production. In this model, the only immunoglobulin transcripts containing human sequences should be IgM. However, in these mice, 4% of B-cells expressed human IgM and mouse IgG simultaneously. These cells were purified by flourescence activated cell sorting and found to express mRNA containing the human V_HDJ_H region correctly spliced to the mouse Cγ region. In a subsequent study, mRNA containing the human V_HDJ_H region and the mouse E constant region was detected. The presence of the human sequences as a transgene, integrated outside the mouse immunoglobulin locus makes a mechanism of alternative termination and *cis*-splicing for double isotype

expression in this model extremely unlikely in the absence of any DNA rearrangement. The authors therefore proposed *trans-* splicing as the mechanism. No DNA rearrangement could be detected in these mice, but it cannot be conclusively ruled out as a rearrangement in such a small proportion of the B-cells would be difficult to detect. The same mouse model was also found to contain immunoglobulins containing endogenous constant regions with the variable region derived from the transgene (Shimizu and Honjo, 1993). Again, *trans-* splicing was proposed as the mechanism, but DNA rearrangement could not be conclusively ruled out.

The above examples of possible *trans-* splicing in mammalian cells, although provocative, involve a highly specialised sub-set of genes that are known to undergo extensive rearrangement at the DNA level. Reports have also been made of possible *trans-* splicing in mammalian genes not directly involved with the immune system. A cDNA containing sequences from the androgen binding protein (ABP) and the histidine decarboxylase (HDC) genes has been documented in fetal rat liver (Sullivan et al, 1991). Sequence data indicated that regions from the two genes were joined at consensus splice sites. A 4.4Kb transcript on a Northern blot hybridised to ABP and to HDC probes. This suggests that the original cDNA was not a cloning artefact, but represented a genuine endogenous transcript. The presence of two separate transcripts, each of size 4.4Kb could not, however, be formally ruled out. A study on human breast cancer biopsy samples by Murphy et al (1993) has described a transcript apparently representing sequences from the estrogen receptor gene, normally located on chromosome 6, *trans-* spliced to unrelated sequences from chromosome 12. Unusual forms of cytochrome P450

mRNA are also proposed to be generated by *trans*- splicing in rat liver (Zaphiropoulos, 1993). However, in none of these cases have rearrangements at the DNA level been adequately disproven. Recently, Eul et al (1995) have proposed that truncated T-antigens produced in rat cells transformed with the SV40 early region are translated from *trans*-spliced mRNAs. In this case, two identical SV40 transcripts are thought to be spliced together to form the fusion RNA. The primary transcripts each contain a 5' (donor) splice site, with no downstream 3' (acceptor) splice site. The splicing reaction occurs between the splice donor of one molecule, and a cryptic splice acceptor from a second molecule. The cryptic splice acceptor is upstream of the donor in each primary transcript.

In this chapter, I present a more detailed analysis of the gene trap integration events in these lines and the processing of their fusion transcripts. Lines ST576 and ST478 contain integrations into pol I transcribed ribosomal RNA genes. The primary transcripts from these loci are processed by an accurate inter-molecular (*trans*-) splicing reaction, providing a mechanism whereby protein coding sequences transcribed by pol I can obtain a 5' cap and be translated to form an active *βgeo* protein.

Establishment of Cell Lines

Lines ST478 and ST576 were both derived by electroporation of CGR8 cells with the secretory trap vector pGT1.8tm, as described in chapter two. Both lines were isolated by W. C. Skarnes.

Preparation of Genomic DNA from Tissue Culture Cells.

Cells were grown to confluence in a $25cm^2$ flask, harvested by trypsinisation and resuspended in 3ml of TEN (10mM Tris; 50mM EDTA; 100mM NaCl). Genomic DNA was prepared from this lysate (see chapter 2). 10µg samples of this genomic DNA were digested with appropriate enzymes, and used for Southern blotting as described in chapter 2.

RN'ase Protection Assays

These were carried out according to the protocol of King and Melton (1987) with a number of modifications, as outlined in chapter two. The probe templates used for riboprobe synthesis were made by restriction digestion of Qiagen prepared DNA of RACE clones obtained from line ST576.

PCR Amplification of Genomic Sequences.

PCR was carried out using a mixture of 1 unit of cloned pfu (stratagene) to 100 units of amplitaq (perkin elmer). Reaction buffer consisted of 20mM Tris-HCl (pH8.55); 150µg/ml bovine serum albumin; 16mM (NH4)2SO4; 3.5mM $MgCl_2$; and 250mM each dNTP. 30 cycles of PCR were carried out, each cycle being 30sec at 99°C, 30sec at 67°C and 10 minutes at 68°C

(Barnes, 1994). PCR products were gel purified using a 1.2% agarose gel before Kpn1 restiction digestion.

Screening of λ-DASH Genomic Library.

Library screening was carried out according to the plaque hybridisation method of Benton and Davis (1977) as described in chapter two.

Preparation of λ DNA.

50ml liquid lysates were made according to Ausubel et al (ed.) 1987. λ DNA was prepared from these according to Davis et al (eds), 1986 (see chapter 2).

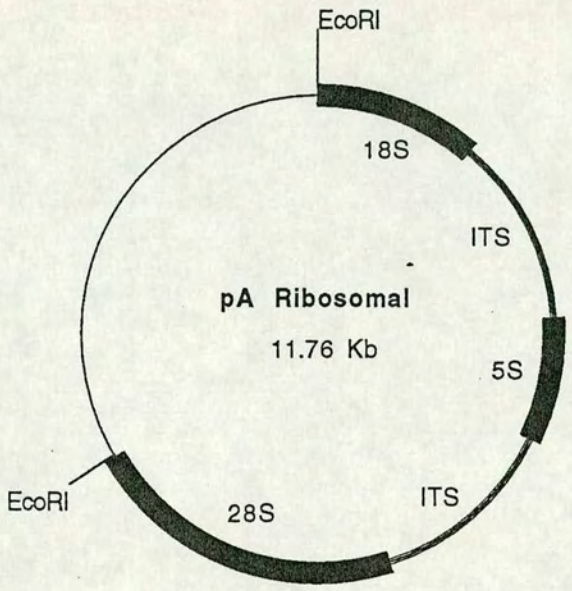
Sequencing of λ Genomic Clones.

Cycle sequencing was carried out using the CircumVent thermal cycle dideoxy DNA sequencing kit (NEB). ^{32}P end labelled primers were used.

Flourescent *in-situ* Hybridisation. Metaphase chromosome spreads and interphase nuclei were prepared from cells of lines ST576 and ST478 using the method described in chapter two. The vector, pGT1.8tm was labelled with digoxigenin, hybridised to the samples and visualised using a FITC detection system. λ genomic clones of cytochrome C oxidase VIIc and the mouse G protein isolated from a λ DASH library were labelled using biotin and visualised using a texas red detection system. A ribosomal clone containing the 5' end of the 18S rRNA, the 5.8S rRNA and the 3' region of the 23S rRNA (figure 5.1) was also labelled with biotin and visualised using texas red.

Figure 5.1: pA Ribosomal Clone Used for Flourescence *In Situ* Hybridisation.

The ribosomal clone pA (a gift from D. Solter) contains a 7.3Kb EcoRI fragment of mouse ribosomal DNA including the 3' region of the 18S rRNA, the whole of the 5S rRNA and the 5' region of the 28S rRNA, together with intervening spacer regions. The fragment is cloned into pBR322. In order to eliminate the risk of a false *in situ* signal due to hybridisation of plasmid sequences from this clone to plasmid sequences from the secretory trap vector, the ribosomal clone was digested with EcoRI and the ribosomal fragment gel purified before labelling.


EcoRI
18S
ITS
pA Ribosomal
11.76 Kb
5S
EcoRI
28S
ITS

Whole mount *in-situ* hybridisation was carried out on cell lines electroporated with the vector pGT1.8tm, using a digoxigenin labelled *lacZ* riboprobe to detect fusion RNAs. In a small number of clones, most of the *lacZ* fusion transcript was located in the nucleus, as described in chapter 4. 5' RACE cloning was carried out using total RNA from a number of these lines containing fusion transcript in the nucleus to investigate the reason for this unusual localisation.

5' RACE Cloning Suggests That Fusion Transcripts in Lines ST576 and ST478 are Processed By An Unusual Mechanism.

5'RACE cloning from lines ST576 and ST478 was carried out by W. C. Skarnes using an updated protocol (protocol B, see chapter two) involving two rounds of PCR using nested primers within the *En-2* sequence. These two lines gave unexpected results (figure 5.2). Three out of twelve clones from line ST576 contained both *En-2* exon sequences from the vector and *En-2* intron sequences indicating, as expected, that a significant proportion of the steady state *lacZ* fusion transcript in this line was unspliced. In addition to these clones, a number of others were obtained. Two clones contained a region of the *En-2* intron from near the linearisation site of the vector, together with sequence from the 5' external transcribed spacer (5'ETS) of an 18S rRNA gene. The remaining clones contained sequences from a number of different endogenous genes. Some of these sequences showed homology to genes previously cloned: *ribosomal protein S12* (Ayane et al, 1989); *cytochrome C oxidase VIIc* (Akamatsu and Grossman, 1990); *Drosophila neuralized* (Boulliane et al,1991); a mouse protein tyrosine phosphatase (PTP) (Matthews et al,

1990) and a mouse G-protein related gene (unpublished: submitted to GenBank database by Raj, 1993) . Two others were not homologous to any previously cloned genes. 5'RACE cloning from line ST478 gave seven clones in total. Three of these were unspliced *En-2* intron sequence. Four different endogenous sequences were also detected. One clone showed homology to the gene for ribosomal protein L22 (unpublished: submitted to GenBank database by Fujita, 1994). None of the other sequences matched previously cloned genes. All of the endogenous sequences cloned were apparently joined correctly to the introduced splice acceptor site in the vector (figure 5.3). These data suggested that intermolecular splicing reactions were taking place in these cell lines between the splice acceptor site from the vector and multiple endogenous transcripts. However, in order to confirm this hypothesis, it was necessary to rule out a number of more trivial explanations of the RACE cloning results.

ST576 is a Clonal Cell Line and Does Not Represent a Mixed Population of Cells.

In order to rule out the possibility that line ST576 was a mixture of several clones containing different vector integrations, sub-clones from the line were analysed. Cells from the original line ST576 were seeded at clonal density and 22 colonies picked. Genomic DNA from the parent line and from 4 of these sub-clones was digested with BglII, which has a single recognition site within the vector sequences, Southern blotted and probed with *lacZ* sequences. The same pattern of bands was seen for each sample, confirming that the original line was not a mixed clone (figure 5.4). The integration of a single copy of the vector would result in a single band using a *lacZ* probe following BglII digestion. The pattern seen for line ST576 was considerably more complex, indicating that there are

multiple copies of the vector present either at a single site of insertion, or at multiple sites within the genome. Data obtained using flourescent *in situ* hybridisation (FISH, see below) indicated that the numerous copies of the vector are present at a single site of insertion.

Trans- Spliced _LacZ_ Fusion Transcripts are Present in Total RNA From Lines ST576 and ST478.

As the RACE cloning protocol relies on the use of PCR, it was necessary to rule out the possibility that the multiple sequences cloned from this cell line were a result of PCR artefacts. Northern blots were carried out using total RNA from line ST576 and a control line, ST534. RACE cloned material from line ST576 was used to make random primed probes. These revealed that the genes cloned from line ST576 were expressed at varying levels, some showing very high expression (ribosomal protein S12) and others being barely detectable *(neu)*. None of the probes detected an extra band representing the *lacZ* fusion transcript in line ST576, indicating that only a small fraction of each endogenous transcript is spliced to the vector sequence. In each case, the level of expression of the endogenous gene remained unaltered in line ST576, compared to the control cell line (figure 5.5A).

A more sensitive technique, RNA'se protection assay, was used to confirm the presence of *trans*- spliced transcripts in total RNA from line ST576. ^{32}P labelled riboprobes were synthesised from a number of the RACE clones obtained from line ST576. Each probe spanned the putative *trans*- splice junction, and was predicted to give three major protected bands: a large band, representing the *trans*- spliced species; an intermediate band, representing normal transcript from the endogenous

gene and a common 120 base pair band, representing *En-2* sequences spliced to sequences other than the one being investigated or remaining unspliced (figure 5.5B).

The splicing of three of the genes cloned from line ST576: cytochrome C oxidase VIIc; the mouse G protein related gene and the novel gene, clone 4B-2, were examined by RNA'se protection assay using total RNA from lines ST576; ST478 and a control line ST534. In each case, the *trans-* spliced product could be detected. Only a small proportion of the transcript from each of the endogenous genes is involved in *trans-* splicing to the *En-2* splice acceptor. (2% for cytochrome C VIIc Oxidase; 4% for clone 4B-2; 6% for G-protein; table 5.6). Surprisingly, *trans-* spliced products from all three of these genes, originally cloned from line ST576, were also detected in total RNA from line ST478, suggesting that these two independent cell lines both use a similar set of endogenous transcripts as substrates for *trans-* splicing to the *En-2* splice acceptor.

The Integration Site of the Vector in Each of the *Trans-* Splicing Lines is Within the 5' External Transcribed Spacer (5'ETS) of an 18S Ribosomal RNA Gene.

Clone 4A-8, obtained by 5'RACE cloning from line ST576, contained sequence from the 5' ETS of a ribosomal RNA transcription unit. The boundary between the intron sequences and the *En-2* exon sequence appeared to have arisen as a result of splicing to a cryptic splice donor within the intron sequence. The junction between the 5'ETS sequence and the intron sequence therefore was concluded to represent the site of integration of the vector in this line (figure 5.7). To confirm this, PCR was carried out on genomic DNA using a primer to the 5'ETS sequence and

nested primers from the vector *En-2* sequence (figure 5.8A). These data showed that the 5'ETS and vector sequences are linked at the DNA level in line ST576. The integration site in line ST478 was also demonstrated to be within a ribosomal transcription unit. The size of the fragment obtained from line ST4978 indicated that the integration in this line occurred further downstream in a ribosomal gene. KpnI digestion of the PCR product followed by Southern hybridisation using probes to *En-2* intron sequence and to the 5'ETS predicted that the integration site is within the 5'ETS region, approximately 400bp downstream of the integration site in line ST576 (figure 5.8B). Northern blot hybridisation using a *lacZ* probe against total RNA from line ST576 gives a smear of signal, containing an intense band (see chapter 4). A similar smear was seen on a Northern blot of total RNA from line ST478 (data not shown). The sizes of these major products (10Kb for ST576 and 10.5Kb for ST478) are consistent with these PCR results.

Efficiency of *Cis*- and *Trans*- Splicing Reactions

The use of the cryptic *cis* splice site predicted from the structure of clone 4A-8 was investigated by RNA'se protection assay using total RNA from line ST576 and from the parental cell line CGR8 as a control for endogenous *En-2* expression. A probe spanning the unspliced intron/exon boundary from the secretory trap vector (probe 405) was used to quantify the total amount of *En-2* containing transcript that was taking part in splicing reactions (figure 5.9A). A second probe (4A-8) spanning the putative cryptic splice was also used to assess the efficiency of *cis*-splicing at the cryptic splice (figure 5.9B).

Probe 405, representing the unspliced intron/exon boundary gave two bands from ST576 RNA. The larger (165bp) band represented unspliced RNA and the smaller (120bp) band represented RNA which has undergone splicing at the introduced *En-2* splice acceptor site. No bands were seen in the CGR8 control lane, showing that endogenous *En-2* transcript is not present in large enough quantities to be detected.

Probe 4A-8, representing the sequence obtained by *cis-* splicing to the cryptic splice donor gave three major bands. The largest of these (309bp) represented transcripts which had undergone *cis-* splicing to the cryptic donor. The smallest (120bp) band represented *En-2* containing transcript which had been spliced to other splice donors or remained unspliced.

Phosphorimage analysis indicated that approximately 18% of the *En-2* containing transcript in line ST576 was undergoing splicing at the *En-2* splice acceptor site, either to the cryptic *cis-* splice donor or to endogenous splice donors. The remaining 82% was unspliced.

Splicing to Endogenous Genes Predominantly Involves Genuine Splice Donor Sites.

Genomic clones of five of the genes apparently *trans-* spliced in lines ST576 and ST478 were isolated from a λ DASH library using RACE clones derived from line ST576 as probes. Primers were designed to sequences immediately upstream of the *En-2* splice junction. Cycle sequencing was carried out using these primers to investigate splice donor site usage in the splicing to *En-2*. For all but one of the genes (Cytochrome C Oxidase VIIc) the splice donor site used appeared to be a genuine splice donor (figure 5.10) both by homology to the consensus mammalian splice donor

sequence, and by comparison of downstream sequences to the published cDNA sequence. Genomic DNA for the mouse homolog of *Drosophila neuralized* was sub-cloned into a plasmid vector (Julie Moss) and sequenced using sequenase 2.0 (Jane Brennan). This gave clearer sequence than did the original cycle sequencing, and confirmed that the splice donor used in this line is a consensus donor. Comparison to the *Drosophila* gene suggests that the intron/exon boundaries are different in the two species.

Flourescent *in-situ* Hybridisation Confirms That Each of the *Trans*-splicing Lines Contains a Single Integration Site of Vector Sequences, Linked to Ribosomal Transcription Units and on a Different Chromosome from the Endogenous Genes Involved in *Trans*- splicing.

FISH was carried out using metaphase chromosome spreads and interphase nuclei from each of the two *trans*- splicing lines (figure 5.11). Vector sequences were labelled with digoxigenin and detected using an FITC antibody system. Genomic clones of cytochrome C Oxidase VIIc, the G protein, and a ribosomal clone were labelled with biotin and detected using a texas red antibody system. Chromosomal DNA was counterstained using DAPI.

The use of the vector probe and a probe to the cloned G-protein (clone 4B-8 from line ST576) showed a single signal with the vector probe (green) and the expected duplicate signals with the G-protein probe (red) for cell lines ST576 (panel A) and ST478 (panel B). The metaphase chromosome spreads showed that the site of integration of the vector in each line is near the centromere on a chromosome other than the one which carries the G-protein gene. This result rules out the possibility that

the splicing seen is an unusual long range *cis-* splicing reaction, and confirms that it must involve two distinct transcript molecules. The signals obtained with the interphase nuclei suggest that the site of transcription of the *βgeo* gene is not in close proximity to the site of transcription of the G-protein gene. The result obtained using a probe for the Cytochrome C Oxidase VIIc gene was similar to that for the G-protein probe both in line ST576 (panel C) and in line ST478 (panel D).

The use of a probe to ribosomal sequences in conjunction with the vector probe in line ST576 (panel E) and ST478 (panel F) confirmed a linkage between the sites of insertion of the vector and ribosomal gene clusters in metaphase chromosome spreads More interestingly, the vector DNA also shows a close association with the ribosomal genes in the interphase nuclei of each cell line, indicating that the rRNA/*βgeo* hybrid transcription unit is sequestered within the nucleolus. FISH experiments were carried out at the MRC Human Genetics Unit, Edinburgh with help from W. Bickmore.

The Endogenous Transcripts Which Undergo *Trans-* splicing in these Lines do not Show Nuclear Localisation.

Northern blot analyses have ruled out the simplest hypothesis that the genes used for *trans-* splicing in these two cell lines are present at unusually high levels within the cell. Another possible explanation for their use in the *trans-* splicing reaction is that they are normally spliced inefficiently, and so are present at relatively high local concentrations within the nucleus. In order to investigate this possibility, whole mount *in situ* hybridisations were carried out using anti-sense riboprobes to a number of the endogenous sequences cloned from line ST576 (figure

5.12). The transcripts show cytoplasmic localisation, with some being ubiquitously expressed, and some showing a more patchy expression pattern. No difference could be detected between the expression of these transcripts in line ST576 and in the control line CGR8. Thus, these genes are not *trans*- spliced in these lines due to high expression, or to a natural accumulation of their transcripts within the nucleus. Additionally, the steady state expression of these genes is not altered by their involvement in *trans*- splicing.

Figure 5.2: 5' RACE Clones Obtained From Lines ST576 and ST478.

CLONE	SEQUENCE	IDENTITY
LINE ST576		
4B-10/4B-6	acacaccact ttccttctgg tcagtgggca catgtccagc cxxxcccaac acttgtatgg ccttggcggg gtcatccccc ccccaccccc agtatctgca acctcaagct agcttgggtg cgttggttgt ggataagtag ctagactcca gcaaccagta acctctgccc tttctcctcc atgacaacca g	**Unspliced *En*-2 Acceptor**
4A-8/4B-16	AAGTCGCTCG TCGACCTCCC CTCCTCCGTC CTTCCATCTC TCGCGCAATG GCGCCGCCCG AGTTCACGGT GGGTTCGTCC TCCGCCTCCG CTTCTCGCCG GGGGCTGGCC GCTGTCCGGT CTCTCCTGCC CGACCCCCGT TGGCGTGGTC TTCTCTCGCC agccctctcc cgtggtctcg ccctcttgtc ctagaagcct cactggccag gtgtaagcca ggtcgtgggt gccgagccct gctccctcat cctcagcatg gatgtgaaga ggactgtatg gcgtgcgggt gtgtgtgacc gtgggtacac ttaaaacacc gggttttgga tctgcactgt cccggatgtc ctctggtgct caaagaccct tctgggtttg ccctttg	**18SrRNA 5'ETS/ *En*-2 intron (M20154)**
4A-9	TCGGCGGCGG CCNANCGGGT GCGTTCAAGA TTCGGCGTCA CCCGTGATTC ACCGCCATGG CCGAGGAAGG CATAGCTGCT GGAGGTGTAA TGGACGTCAA CACTGCTCTA CAAGAGGTGC TGAAGACCGC CCTCATCCAC GATGGCCTAG CACGTGGCAT ACGCGAAGCT GCCAAAGCCT TAGACAAGCG CCAAGCCCAT CTGTGTGTGC TCGCATCCAA CTGTGATGAG CCCATGTATG TCAAGCTGGT GGAGGCACTT TTGGCTGAGC ACCAAATCAA CCTGATAAAG GTTGATGACA ACAAGAAACT AGGGGAATGG GTAGGCCTCT GTAAAATCGA TCGAGAGGGC AAACCACGGA AGGTGGTTGG TTGCAGTTGC GTAGTGGTTA AG	**Mouse Ribosomal Protein S12 (X15962)**
4B-2	CTCTTTGGCT GCGATGCGCC TACGAGTATG TAGTTGTATG GTGGGGATGA TATCTTGGGC CTTATTTGTG AGTGGCCGAG ATACGGGAGA CACAAGGTGG CTATGGTGCT GGTCGAAAGC CATCAACTCT GCAAGAGCGC CGTGAAGGTT AGCGCGACAA CCACCATGTA AAGCCGATAC GGTGTTGGGC GGCGTGACGC AAGTTCGCCA CCGCTGGGTC GGATG	**Novel**
4B-7	TGGCCAGATC ACCAAGAAGC AATGCTGCTG GAGCGGGGCC CTGCGACTTG GCTTCACCAG CAAGGACCCT TCCCGCATCC ACCCCGACTC GCTGCCCAAG TACGCCTACC CTGACCTGGT GTCTCAGAGT GGCTTCTGGG CCAAAGCATT GCCTGAGGAG TTTGCCAACG AGGGCAACAT CATTgccttc tgggtggaca agaagggccg cgtcttctac cggatcaatg agtcagctgc tatgcttttc ttcagtGGGG TCCGGACGGT GGACCCGCTC TGGGCCCTGG TGGACGTCTA CGGCCTCACG CGGGGTGTCC AGCTGCTAG	**Homologous to *Drosophila neu* (X61617)**
4B-8	CCACAGGGTC GTGGTGGCAG CCGCTGTGGT GCTTGGCTCT CTGAGCTATC CGGTGCCATC CTTGTCGCTG CGGCGACCCT CGCATCAACT GCAGCCATGA CCGAGCAGAT GACCCTTCGC GGGACCCTTA AGGGCCACAA TGGATGGGTA ACACAGATCG CAACCACACC GCAGTTCCCG GACATGATCC TGTCTGCGTC TCGAGACAAG ACCATCATCA TGTGGAAGCT GACCAGAGAT GAGACCAACT ATGGCATACC ACAGCGTGCT CTGAGAGGTC ACTCCCACTT CGTTAGTGAT GTTGTTATCT CCTCTGATGG TCAGTTTGCG CTCTCGGGCT CCTGGGACGG AACGCTGCGC CTCTGGGATC TCACAAC	**G-Protein Related Gene (X75313)**

4B-11	AAGGAAGTTA GGTGGTACGG CCATTTCTTC CGCCTTCCGT GTCTGCGGCC CTCGCAGAAC TTCCAGCAGC GACATGTTGG GCCAGAGTAT CCGGAGGTTC ACGACCTCCG TGGTCCGTCG CAGCCACTAT GAGGAGGGTC CGGGGAAGAA TTTGCCATTT TCAGTGGAAA ACAAGTGGCG GTTGCTGGCT ATGATGACCG TGTACTTTGG ATCTGGGTTT GCCGCACCTT TCTTTATAGT AAGACACCAG CTACTTAAAA AATAAGGATA TTTAATTCAT CCCTTTAACA G	**Cytochrome C Oxidase VIIc (X52940)**
4B-15	ACCCGGTGTG GACCAACTAC ATTATCCTCA TTCTGTTGCT CTAAGGACTG TGCTCGCCTA TAGACCTCCT AATGTTGTCT TCTAATACAC GAGGTGCAGC ACGTGCGCCC CCTCCGTTGC G	**Novel**
4B-23	CTCTCTTCAT CCGCTGTCCT GTCGGTTTTC ATTATGGCGC CTTGTACTTC TACCCAACGT TGCCCGCCCC CCATCTTCCC AGCCCACCAG CTCCGTCCGC CTCTCCGGAA GGTTCGTGAG ATACCGAGCT GCCGGGAAGG GACCGGATAT TTCCGTCTGG TCCAAGGAAG CACGGGTTAT GGCCACATGC GCAGTGTATA ACTCAGGGCC CCACGCCTTT TGACCC	**Leukocyte Common Antigen-related Phosphatase (M36033, M33671)**
LINE ST478		
478.17	gattccgcgg cctccaatcg gaggtcctgt ggggaggagc tgcagggcca gtttattttg gcggttcaat ttcaacatgg ctgaggtgag ccgagatagc gaggctgcgg aaaggggcc tgagggctcc tctccggaag ctgtgccagg ggccgcgacc atccccaggg tgaaactcct ggacgccata gtagacactt tcctccagaa gctagtcgcc gacaggag	**Novel**
478.19	ccacaagtgt tctaaaacac agctagctct aaaagcaagc caagctcagt ttctgtcact gtccagtgag tcctatttgt accctcgaaa catcatgtgc cctccacggg tcttgcctca tgtgcacctg tctcagcaag tttgtctctc agctgacatc attctgccaa tcagcccaag gagtccaagg aattgaaaag aaactacagc acaccac	**Novel**
478.66	aggttttgaa gttcaccctg gactgcactc accctgtaga agatggaatc atggacgctg ccaattttga gcagttcctc caggagagaa tcaaggtgga tgggaaagct ggcaacctcg ggggaggagt cgtgaggatc gaaggcagga acacggaaga tcactgtcac ttcagagatg cctttctcca aaag	**Ribosomal Protein L22 (D17653)**
478.68	GATTCCATAG ACTACNCCCT AGCCTGCACG TCATCACCTG CAGCTGGTTC CAAGCTCCAA GGATGAATTG TTTCCCTCTC TTTCAGCTCC GGCCTGCCTC TCAGGTGTAC CCTATCCATG TGTGCCATTG GCGGCTCGCA ACAATCTGGC GCCCTGACCG TGACCTGAGG AAGGTCAGGA AGACACTGAA GGAACGCTGT TTTGTGTGGA GCTCCAGGGG AAGAAAAAGA ATCAGGAAAT TCGTATTGGG GCAAGGTGAA GCAACAG	**Novel**

Numbers given in brackets are the accession numbers for the endogenous gene sequences.

Figure 5.3: Splice Junctions Found in 5' RACE Clones From Lines ST576 and ST478.

RACE Clone	Sequence		Identity
LINE ST576			
4B-6/4B-10	tctgccctttctctccatgacaaccag	**GTCCCAGGT**	En-2 Splice Acceptor
4A-8/4B-16	caaagacccttctgggtttgcccttg	**GTCCCAGGT**	18S rRNA 5' ETS/ En-2
4A-9	GTTGGTTGCAGTTGCGTAGTGGTTAAG	**GTCCCAGGT**	Ribosomal Protein S12
4B-2	GCAAGTCTGCCACCGCTGGGTCGGATG	**GTCCCAGGT**	Novel (no ORF)
4B-7	GCCTCACGCGGGGTGTCCAGCTGCTAG	**GTCCCAGGT**	Drosophila neu Homolog
4B-8	GAACTGGCGCCTCTGGGATCTCACAAC	**GTCCCAGGT**	Mouse G Protein
4B-11	AGGATATTTAATTCATCCCTTTAACAG	**GTCCCAGGT**	Cyt. C Oxidase VIIc
4B-15/4B-18	TGCTGCACGTGCGCCCCCTCCGTTGCG	**GTCCCAGGT**	Novel
4B-23	AACTCAGGGCCCCACGCCTTTTGACCC	**GTCCCAGGT**	Muslrpa PTP Homolog
LINE ST478			
478.17	CCTCCAGAAGCTAGTCGCCGACAGGAG	**GTCCCAGGT**	novel
478.19	AATTGAAAAGAAACTACAGCACACCAC	**GTCCCAGGT**	novel
478.66	CACTTCAGAGATGCCTTTCTCCAAAAG	**GTCCCAGGT**	Ribosomal Protein L22
478.68	TCGTATTGGGGCAAGGTGAAGCAACAG	**GTCCCAGGT**	novel

Bold type denotes *En-2* sequence from the secretory trap vector. Lower case denotes intron sequence.

Figure 5.4: Southern Blot of Genomic DNA From Sub-Clones of Line ST576.

Genomic DNA was extracted from sub-clones of line ST576 and digested overnight with BglII. Following electrophoresis and Southern blotting, the samples were probed with a *LacZ* random primed probe. Each of the sub-clones gave the same pattern of three bands, which was the same as the pattern given by the parental line. This confirms that ST576 is a clonal cell line and not a mixture of a number of cell lines.

BglII has one cut site within the vector, between the *En-2* and *lacZ* sequences. A cell line containing a single copy of the vector would be expected to give a single *lacZ* containing band, representing a junctional fragment containing vector sequences and endogenous sequence 3' to the insertion site. The insertion in line ST576 can be seen to be more complex than this. The band at 8.77Kb represents repeat units of the vector. One of the other two bands probably represents the 3' junctional fragment, with the presence of a third band indicating the probability of at least one truncated copy of the vector also being present. These data suggest the presence of a number of copies of the vector in line ST576, but does not allow us to determine whether they are at a single site within the genome or at a number of different sites.

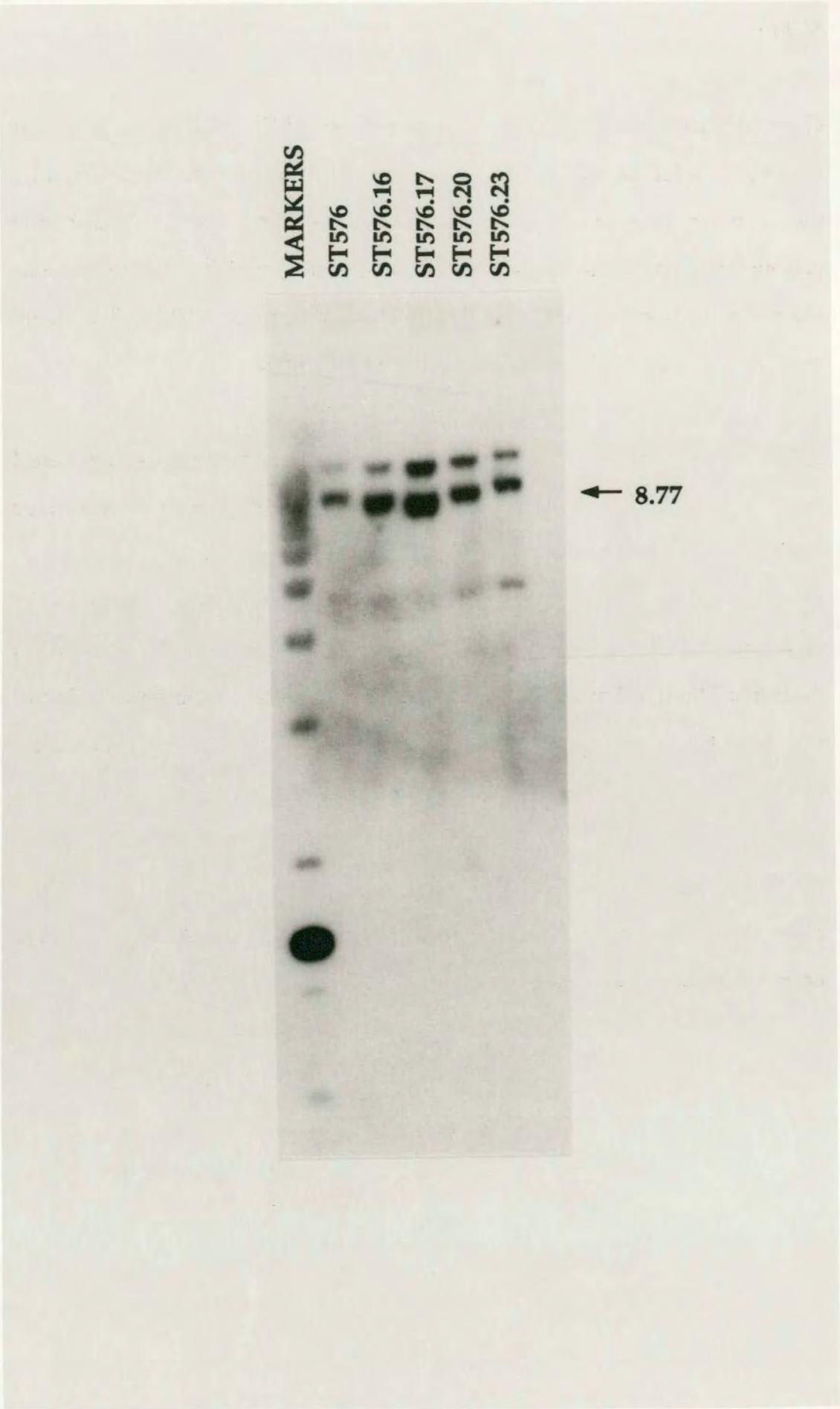

MARKERS
ST576
ST576.16
ST576.17
ST576.20
ST576.23
8.77

Figure 5.5: Confirmation of the Presence of *Trans* -Spliced Fusion Transcripts in lines ST576 and ST478.

A) Northern blots of total RNA from line ST576 and from a control line which is known to produce a single, correctly spliced fusion transcript. Probes used are clones obtained by RACE from line ST576. Each of the genes investigated shows an equivalent level of expression in both cell lines. No additional bands representing *trans*-spliced fusion transcripts can be detected in line ST576. Each autoradiograph was exposed overnight.

B) RNAse protection assays using total RNA from lines ST576, ST478 and a control line. Probe templates were prepared by restriction digestion of sequences from line ST576 cloned into Bluescript SK+ (promega). (Digestion was carried out overnight at 37°C using PvuI and XbaI for clones 4B-2 and 4B-11 and BssHII and XhoI for clone 4B-8. The probe template was then gel purified.) Probes were transcribed using T7 polymerase in the presence of ^{32}P CTP, and span the *trans* -splice junctions between endogenous sequences and engrailed sequence from the vector (see diagram). Each probe is predicted to give three major bands. Product sizes for each probe are given in the table. The largest in each case represents the *trans* -spliced product. Autoradiographs of protections of 4B-11 and 4B-2 were exposed overnight. 4B-8 was exposed for 72 hours. Over- exposure of these gels (4 weeks) confirmed that there was no band representing *trans*- spliced RNA in the control cell line, ST534 (data not shown).

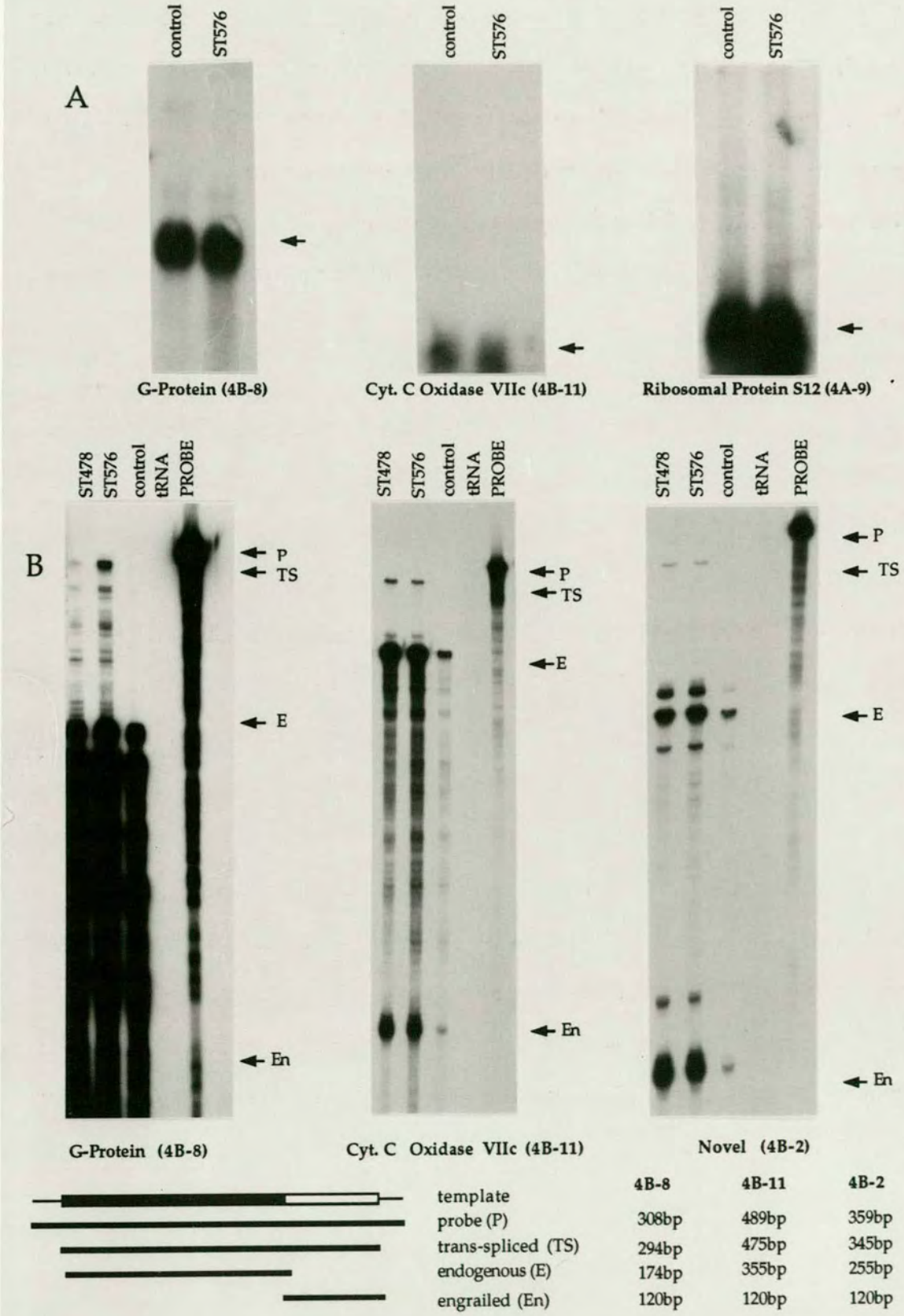

A
control
ST576
G-Protein (4B-8)
Cyt. C Oxidase VIIc (4B-11)
Ribosomal Protein S12 (4A-9)
B
ST478
ST576
control
tRNA
PROBE
P
TS
E
En
G-Protein (4B-8)
Cyt. C Oxidase VIIc (4B-11)
Novel (4B-2)
template
probe (P)
trans-spliced (TS)
endogenous (E)
engrailed (En)
4B-8
4B-11
4B-2
308bp
489bp
359bp
294bp
475bp
345bp
174bp
355bp
255bp
120bp
120bp
120bp

Table 5.6: Densitometric Analysis of RNAse Protections Shown in Figure 5.5 Reveal That Only a Small Percentage of the Transcript From Each Endogenous Gene is *Trans*-spliced.

BAND	MEAN DENSITY	NUMBER OF C RESIDUES	ADJUSTED DENSITY	% OF TOTAL ENDOGENOUS TRANSCRIPT
4B-2 *Trans*-spliced	7.16	86	8.6	4%
4B-2 Endogenous splice	98.51	49	206.9	96%
4B-8 *Trans*-spliced	160.12	86	192.1	6%
4B-8 Endogenous splice	1457.78	49	3061.3	95%
4B-11 *Trans*-spliced	118.1	104	118.1	2%
4B-11 Endogenous splice	4911.3	67	4911.3	98%

An adjustment has been made to take into account the base composition of each probe, as only C residues carry ^{32}P label.

Figure 5.7: Comparison of the Sequence of Clone 4A-8 From Line ST576 With 5'ETS and *En*-2 Intron Sequences Suggests That The Junction Between 5'ETS and Intron Sequences in Clone 4A-8 Represents the Breakpoint at the Site of Integration of the Vector.

The top line, in upper case, represents 18SrRNA 5'ETS sequence. The middle line, in italics, represents clone 4A-8 which begins with 5'ETS sequence (upper case) then moves into *En*-2 intron sequence (lower case). The bottom line, in lower case, represents *En*-2 intron sequence.Potential splice sites at the junction between 5'ETS and *En*-2 intron sequences and between *En*-2 intron and exon sequences are underlined. The consensus splice site sequences are given alongside them for comparison. From this comparison, it seems likely that the junction between *En*-2 intron and exon sequences is a result of a splicing event, involving a cryptic donor from the *En*-2 intron which is a fair match to the consensus. The junction between 5'ETS sequence and *En*-2 sequence does not appear to be a result of a splicing event.

```
18S 5'ETS       CCTCCCTCTCGCGGGGTTCAAGTCGCTCGTCGACCTCCCCT
4A-8            ----------------------------AAGTCGCTCGTCGACCTCCCCT
En-2 Intron     ------------------------------------------------------------

18S 5'ETS       CCTCCGTCCTTCCATCTCTCGCGCAATGGCGCCGCCCGAGTT
4A-8            CCTCCGTCCTTCCATCTCTCGCGCAATGGCGCCGCCCGAGTT
En-2 Intron     ----------------------------------------

18S 5'ETS       CACGGTGGGTTCGTCCTCCGCCTCCGCTTCTCGCCGGGGGCT
4A-8            CACGGTGGGTTCGTCCTCCGCCTCCGCTTCTCGCCGGGGGCT
En-2 Intron     ------------------------------------------------------------

18S 5'ETS       GGCCGCTGTCCGGTCTCTCCTGCCCGACCCCCGTTGGCGTGG
4A-8            GGCCGCTGTCCGGTCTCTCCTGCCCGACCCCCGTTGGCGTGG
En-2 Intron     ------------------------------------------------------------
                          AGgtaagt
18S 5'ETS       TCTTCTCTCGCCGGCTTCGCGGACTCCTGGCTTCGCCCGAG
4A-8            TCTTCTCTCGCCagccctctcccgtggtctcgccctcttgt
En-2 Intron     ----------agccctctcccgtggtctcgccctcttgt

18S 5'ETS       GGTCAGGGGGCTTCCCGGTTCCCCGAC--------------
4A-8            cctagaagcctcactggccaggtgaagccaggtcgtgggtgc
En-2 Intron     cctagaagcctcactggccaggtgaagccaggtcgtgggtgc

18S 5'ETS       ------------------------------------------------------------
4A-8            cgagccctgctccctcatcctcagcatggatgtgaagaggt
En-2 Intron     cgagccctgctccctcatcctcagcatggatgtgaagaggt

18S 5'ETS       ------------------------------------------------------------
4A-8            ctgtatggcgtgcgggtgtgtgtgaccgtgggtacacttaa
En-2 Intron     ctgtatggcgtgcgggtgtgtgtgaccgtgggtacacttaa

18S 5'ETS       ------------------------------------------------------------
4A-8            aacaccgggttttggatctgcactgtcccggatgtcctctg
En-2 Intron     aacaccgggttttggatctgcactgtcccggatgtcctctg

18S 5'ETS
4A-8            gtgctcaaagacccttctgggtttgcccttg----------
En-2 Intron     gtgctcaaagacccttctgggtttgcccttggtaagagcgc
                                              AGgtaagt
18S 5'ETS       ------------------------------------------------------------
4A-8            ------------------------------------------------------------
En-2 Intron     cgggatctacttgtctggaggccagggagtcctcagccgagg
```

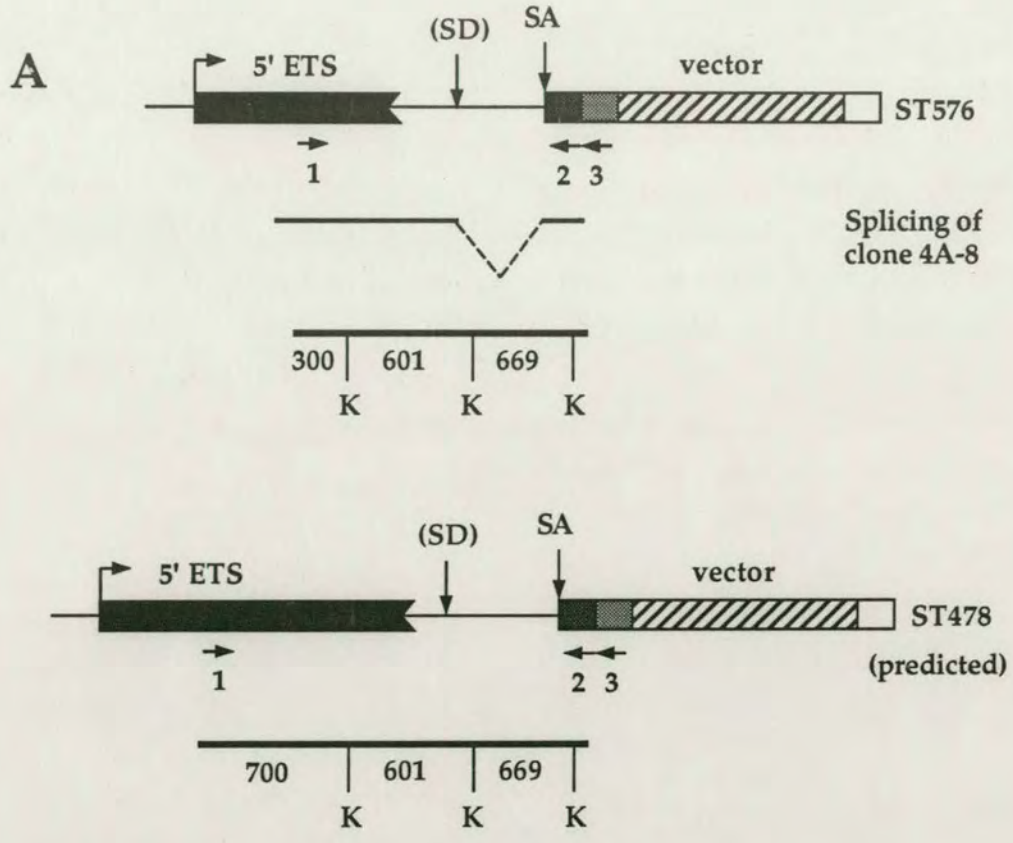
Figure 5.8: Determination of the Sites of Vector Integration in Lines ST576 and ST478

A) The Proposed Structures of the Integration Sites of pGT1.8TM in Lines ST576 and ST478

Endogenous 5' ETS sequences are represented by a black box, intron sequences from the vector by a line. The dark grey and light grey boxes represent *En-2* splice acceptor and *CD4* trans-membrane domain sequences respectively. The hatched box represents *βgeo* sequences and the white box the SV40 polyadenylation signal. PCR was carried out using primers indicated by arrows to confirm that the integration in line ST576 was within the 5'ETS of a ribosomal gene, as predicted, and to investigate the integration site in line ST478. A primer was designed to the 5'ETS sequence (primer 1) and used with nested primers to the vector sequences. First round PCR used primer 3 to *CD4* sequences. The products obtained from this PCR were purified using low melting point agarose, and used for a further round of PCR using primer 2 to *En-2* sequence. The products obtained were, again, gel purified and digested with KpnI, which cuts twice in the vector sequences predicted to be in the PCR product, and once in the 5'ETS.

B) Southern Blot Data Confirms The Identity of the PCR Fragment.

Samples of the undigested PCR products and the KpnI digested products were run on an agarose gel, Southern blotted and probed with a random primed probes to *En-2* intron sequences (1.7Kb), stripped, and reprobed with 5' ETS sequence (170bp). The fragments in the digested samples which hybridise to both probes represent the junction between 5'ETS and *En-2* intron sequences. These fragments contain only a small amount of intron sequence (100bp in the case of ST576), explaining the weak signal obtained with the intron probe.


A
(SD)
SA
5' ETS
vector
ST576
1
2 3
Splicing of
clone 4A-8
300
601
669
K
K
K
(SD)
SA
5' ETS
vector
ST478
(predicted)
1
2 3
700
601
669
K
K
K

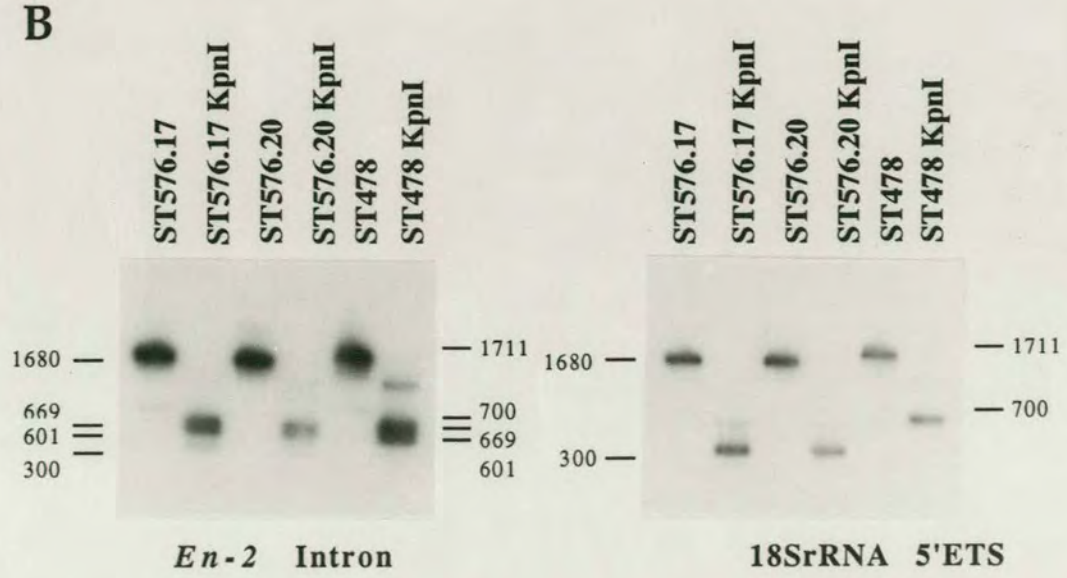

B
ST576.17
ST576.17 KpnI
ST576.20
ST576.20 KpnI
ST478
ST478 KpnI
1680
669
601
300
1711
700
669
601
ST576.17
ST576.17 KpnI
ST576.20
ST576.20 KpnI
ST478
ST478 KpnI
1680
300
1711
700
En-2 Intron
18SrRNA 5'ETS

Figure 5.9: Confirmation of the Use of the Cryptic Splice Donor Site from the *En-2* Intron.

A) RNase Protection Demonstrates the Poor Efficiency of Splicing of the Fusion Transcript in Line ST576.

The probe used was transcribed by T7 bacteriophage RNA polymerase in the presence of α^{32}P CTP. The probe spans the *En-2* intron/exon boundary present in the gene trap vector. Target RNA samples are total RNA from line ST576 and from the parental cell line CGR8. Two clean bands are seen in the ST576 lane. The larger, 165bp band represents unspliced fusion transcript, while the smaller 120bp band represents fusion transcripts in which the *En-2* splice acceptor site has been used correctly. Densitometric analysis reveals that approximately 20% of the fusion transcript in this line is spliced at the *En-2* splice acceptor (data not shown).

B) RNase protection confirms the use of the cryptic splice donor site in the *En-2* intron.

The probe used for this experiment, again transcribed by T7 polymerase, spans the junction between *En-2* intron and exon sequences seen in clone 4A-8, proposed to result from splicing of the *En-2* splice acceptor site to a cryptic splice donor within the *En-2* intron sequence. Target RNAs are total RNAs from lines ST576 and CGR8. The expected bands of 309bp, representing use of the cryptic donor, and 189bp, representing unspliced fusion transcript can be seen in the ST576 lane. A 120bp band representing unspliced fusion transcript as well as that spliced to donors other than the one being investigated is also expected. This band can be seen, but is partly obscured by non-specific background bands also present in the control line and in the tRNA control. These bands presumably occur as a result of incomplete digestion of unhybridised probe RNA. The intensity of the band representing use of the cryptic donor indicates that only a small proportion of the fusion transcript is being processed at this site. The presence of the background bands prevents any numerical assessment of the use of this cryptic donor site.

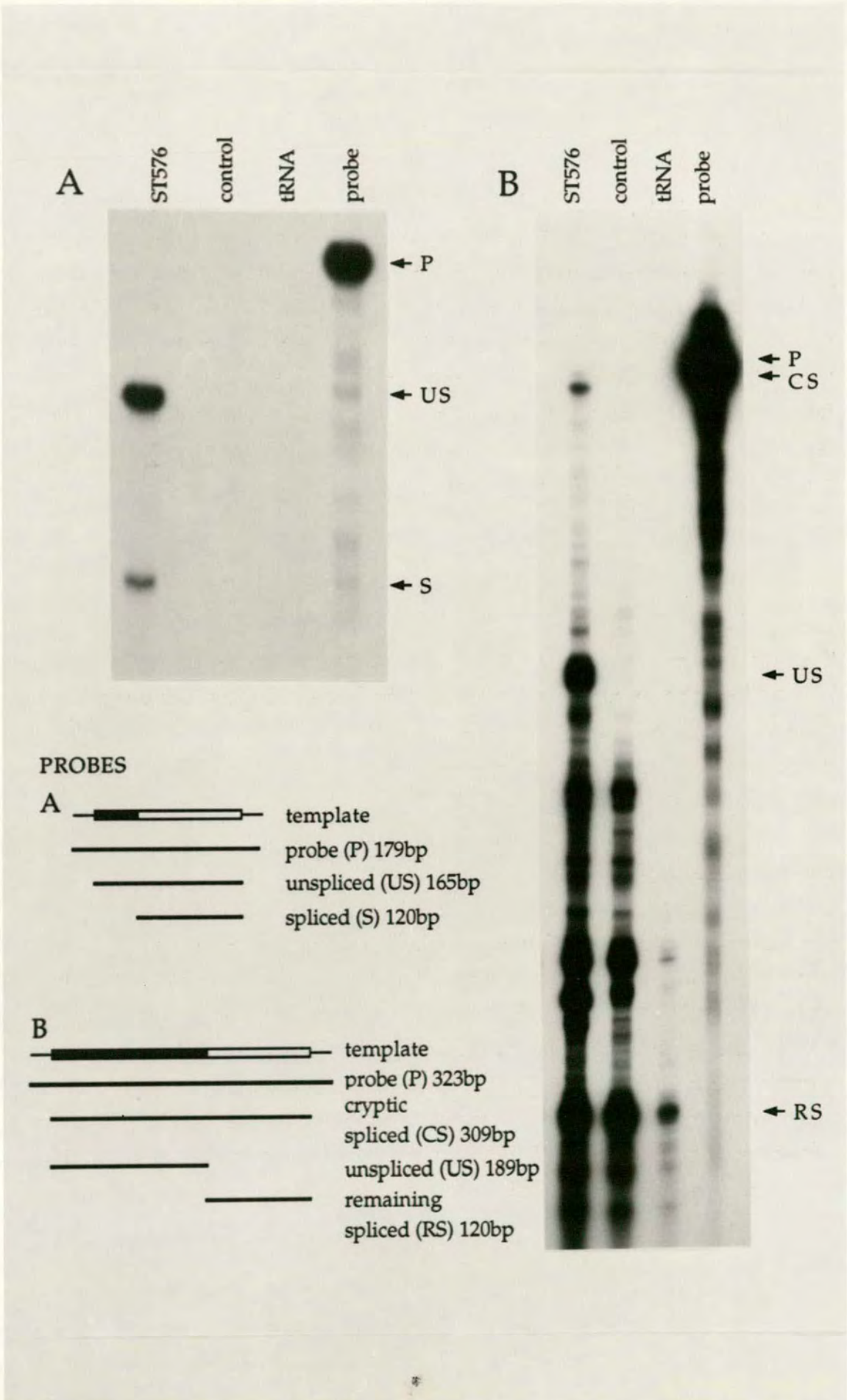

A
ST576
control
tRNA
probe
P
US
S
B
ST576
control
tRNA
probe
P
CS
US
RS
PROBES
A
template
probe (P) 179bp
unspliced (US) 165bp
spliced (S) 120bp
B
template
probe (P) 323bp
cryptic
spliced (CS) 309bp
unspliced (US) 189bp
remaining
spliced (RS) 120bp

Figure 5.10: Sequences Used as Splice Donor Sites in *Trans*- Splicing in Line ST576

	SEQUENCE
CONSENSUS	**A G g u a a g u**
4B-15 (novel)	**C G g u** g **a g u**
4B-2 (novel)	**U G g u a** g g c
4B-8 (G-protein)	**A G g** c **a a g u**
4B-7 (neu homolog)	**A G g u** g **a g u**
4B-11 (Cyt. C. Oxidase)	**A G** a a u g u c

Nucleotides in bold type are those which conform to the consensus mammalian splice donor sequence.

Figure 5.11: Flourescent *In Situ* Hybridisation Confirms The Presence Of A Single Site Of Vector Integration Linked To Ribosomal Gene Clusters In Each Cell Line.

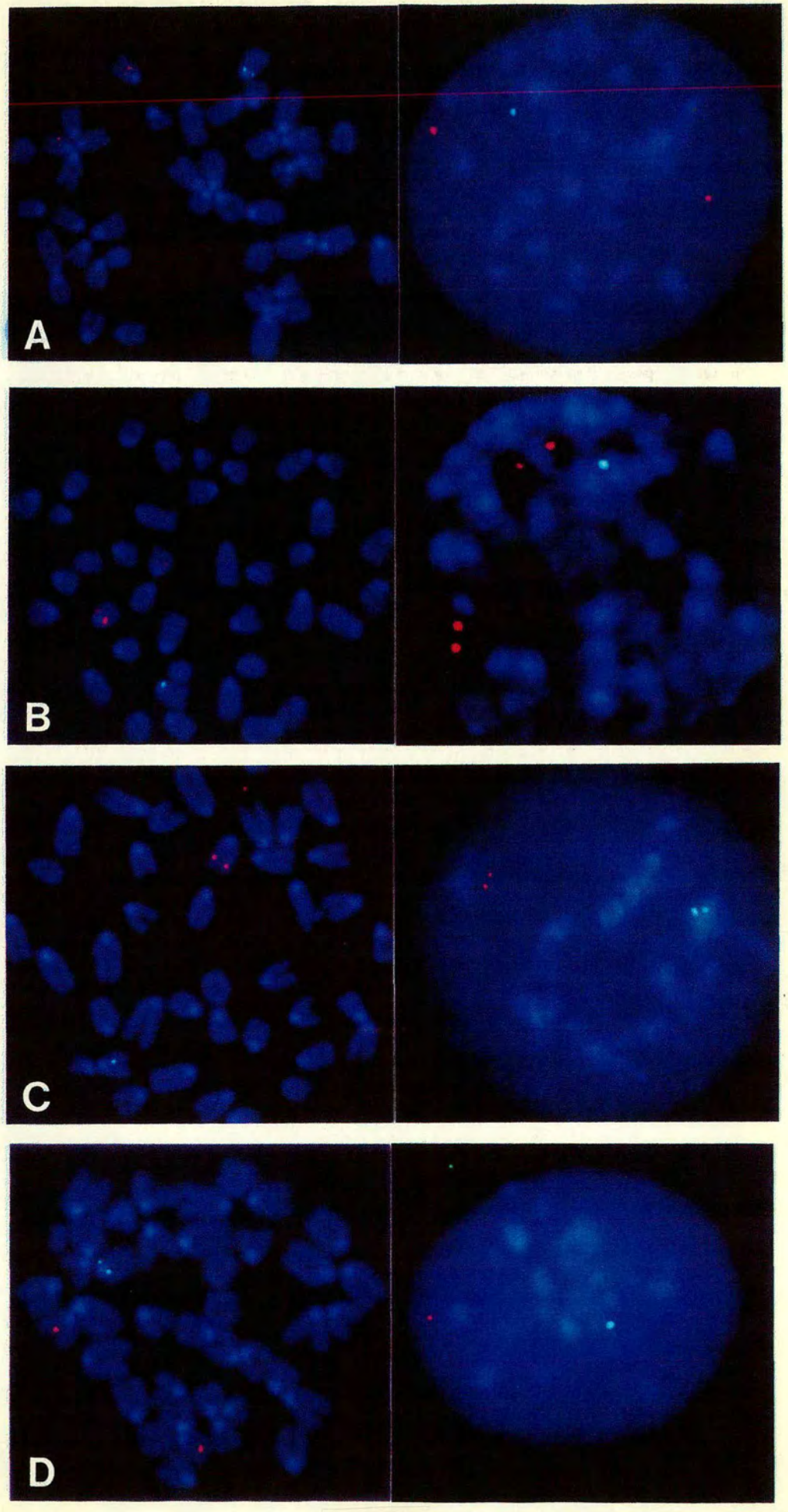
Probes were made by nick translation of plasmid and λ DNA. The gene trap vector was labelled with digoxigenin and visualised using an FITC antibody detection system (green). A ribosomal clone and λDASH genomic clones of the G-protein and cytochrome C Oxidase VIIc originally cloned from line ST576 were labelled with biotin and visualised using a texas red antibody detection system. Each panel shows a representative chromosome spread (left hand side) and a representative interphase nucleus (right hand side).

A) G-protein and vector probes in line ST576.
The integration site of the vector (green) is on a different chromosome from the one which carries the gene for the G-protein cloned from line ST576 in cells of this line. Additionally, the two signals do not co-localise within the interphase nucleus.

B) G-protein and vector probes in line ST478
As for line ST576, the integration of the vector is on a chromosome other than that which carries the G-protein gene, and the two do not co-localise within the interphase nucleus. The double signal seen with G-protein probe (red) demonstrates that chromosome replication had begun in this nucleus prior to fixation

C) Cytochrome C oxidase VIIc and vector probes in line ST576.
The vector integration in lines ST576 is also on a different chromosome from the gene for cytochrome C oxidase VIIc, whose transcript is *trans*-spliced to the vector transcript. Again, co-localisation within the interphase nucleus is not seen.

D) Cytochrome C oxidase VIIc and vector probes in line ST478.
The vector integration is not on the same chromosome as cytochrome C oxidase VIIc in line ST478. No co-localisation is seen in the interphase nucleus.


A
B
C
D

E) Ribosomal DNA and vector probes in line ST576.

In line ST576, the vector integration (green) co-localises with ribosomal DNA (red) towards the centromere of a single chromosome from line ST576. Co-localisation of the two sequences is also seen in interphase nuclei from this line, presumably within the nucleous.

F) Ribosomal DNA and vector probes in line ST478.

Co-localisation of vector and ribosomal DNA sequences can also be seen in chromosome spreads and interphase nuclei from line ST478.

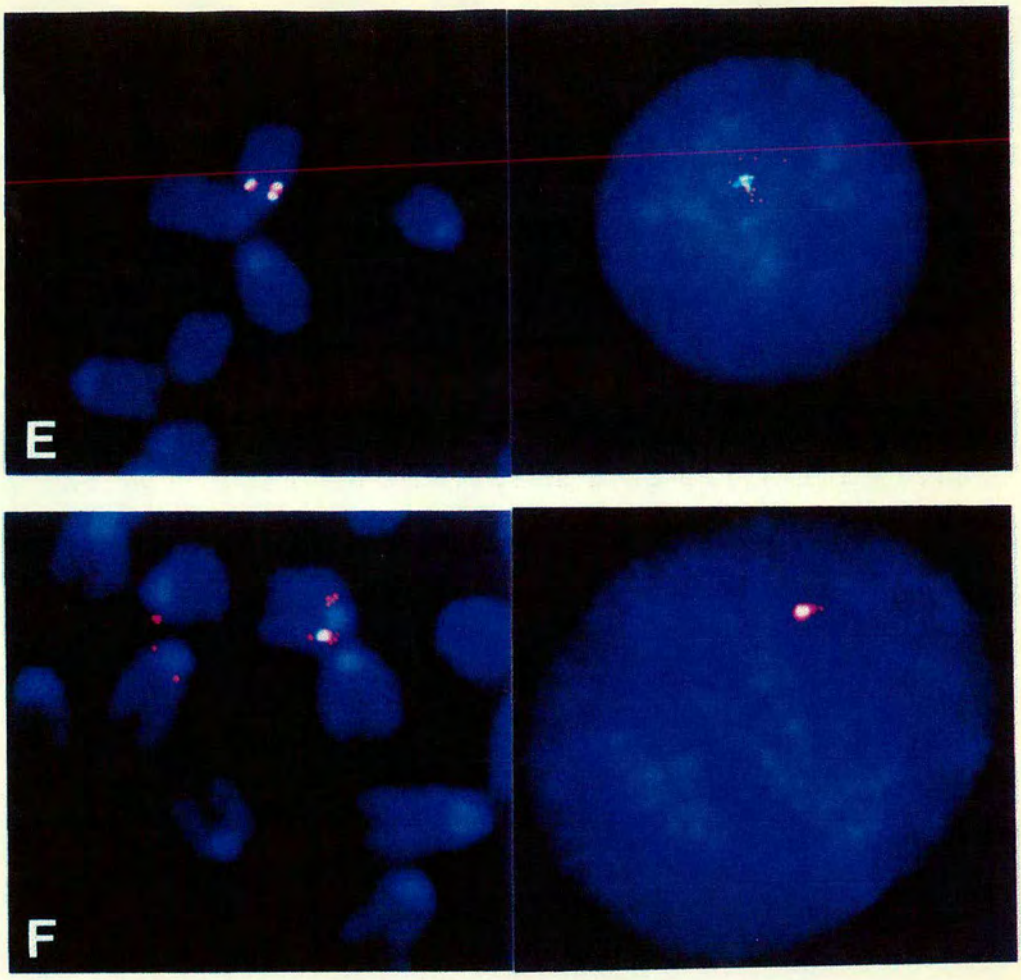

E
F

Figure 5.12: Whole Mount *In-Situ* Hybridisation Demonstrates That the Endogenous Transcripts Used for *Trans*-Splicing in Line ST576 do not Show Nuclear Localisation.

Whole mount *in situ* hybridisation was carried out on cells from lines ST576 and CGR8, grown on glass coverslips. The probes used were transcribed from 5' RACE clones obtained from Line ST576 Using the T7 RNA polymerase transcription initiation signal from the Bluescript II SK+ cloning vector (Promega). Cells were stained overnight and viewed using an Olympus Vanox microscope. Photographs were taken using Kodak GPT160 Ektacolor gold II colour negative film. No difference was seen between the hybridisation of each probe to ST576 cells and its hybridisation to wild type CGR8 cells.

A) 4B-8 (G-Protein), line ST576.

A grainy cytoplasmic signal is seen in all cells.

B) 4B-7 (neu homolog), line ST576.

Grainy cytoplasmic staining is seen in most cells. The staining is more intense in small nests of cells of undifferentiated morphology.

C) 4A-9 (Ribosomal Protein S12), line ST576

An intense grainy cytoplasmic signal is seen in all cells.

D) 4B-2 (novel), line ST576

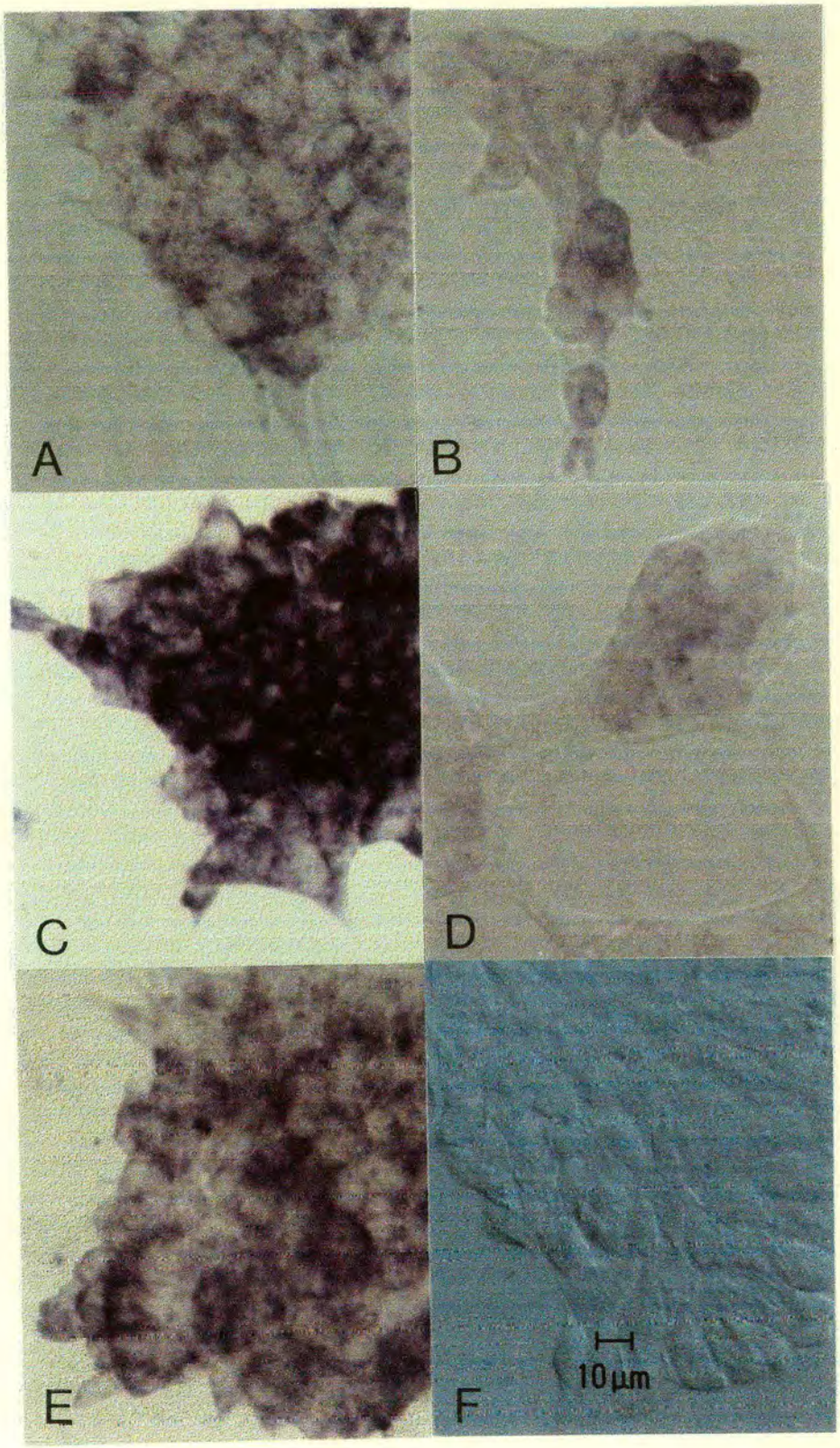
Weak, grainy cytoplasmic staining is seen, largely restricted to cells of undifferentiated morphology.

E) 4B-11 (Cytochrome C Oxidase VIIc), line ST576

Intense grainy cytoplasmic staining is seen in all cells.

F) *LacZ*, line CGR8

This serves as a negative control, confirming that no background staining is seen after overnight staining using this protocol.


A
B
C
D
E
F
10 µm

Two independant mouse embryonic stem cell lines have been described, each of which contains a single site of insertion of the secretory trap vector pGT1.8tm into the 5' external transcribed spacer region of an 18SrRNA gene. In each of these lines, the primary *βgeo* fusion transcript is involved in accurate inter-molecular *trans-* splicing reactions, whereby a number of different fusion transcripts are formed, each containing the 5' region of an endogenous gene and the 3' protein coding regions from the secretory trap vector.

The gene trap integrations characterised in lines ST576 and ST478 are in rRNA genes and result in the formation of 'chimaeric' pol I transcripts that contain the 5'ETS of the ribosomal gene upstream of the *βgeo* protein coding regions. The recovery of gene trap insertions into pol I transcription units was unexpected, since the *βgeo* reporter does not contain a translation initiation signal and so requires splicing to upstream coding exons of pol II genes for its activation. Furthermore, pol I transcripts lack a monomethylguanosine cap structure required for efficient splicing, nuclear export and translation. *Trans-* splicing provides an alternative and unexpected mechanism to explain how gene trap insertions into non-pol II transcription units can potentially produce translated products.

Similarity Between Transcription Units Which are *Trans-* spliced in Lower Eukaryotes and the Insertion Sites in ST576 and ST478

The basic requirement for an exon to undergo *trans-* splicing in *C. elegans* is that its upstream 'intron' does not contain a splice donor site (Conrad

et al, 1993). This intron-like structure containing a splice acceptor site and branchpoint, but no splice donor has been termed an 'outron'. The introduction of a splice donor site between 50 and 250 nucleotides upstream of the *trans-* splice acceptor of the *rol-6* gene is sufficient to convert it into a *cis-* spliced exon (Conrad et al, 1993). It is, therefore, likely that the major and, perhaps, only similarity between genes that are *trans-*spliced in *C.elegans* is their genomic structure.

A number of *trans-* spliced genes have recently been described in *C.elegans* whose transcripts are produced as part of a large poly-cistronic precursor (Speith et al, 1993). The genes at the 5' ends of these primary transcripts receive spliced leader 1 (SL1) by *trans-* splicing, while more 3' genes usually receive a seperate spliced leader sequence, SL2 Both of these SL RNAs are transcribed by pol II and possess a TMG cap at their 5' end. From these observations, it has been suggested that a major role for *trans-*splicing in nematodes is to provide a 5' cap for RNAs which would otherwise lack this structure and be poorly translated. It has been demonstrated that the spliced leader RNAs in *C.elegans* have a trimethylguanosine cap rather than the monomethylguanosine cap common to most pol II transcripts. This TMG cap can also be detected on mature *trans-* spliced mRNAs in *C.elegans* (Liou and Blumenthal, 1990; Van Doren and Hirsch, 1990), lending credence to the idea of *trans-*splicing as a form of *trans-* capping of protein coding transcripts. Additionally, in trypanosomes, the genes coding for the variable surface glycoprotein (VSG) and procyclic acid repetetive protein (PARP) genes are transcribed in an α-amanitin independant manner, most probably by pol I. These transcripts receive a cap structure by *trans-* splicing to the pol II

transcribed spliced leader RNA (Zomerdijk et al, 1991; Sherman et al, 1991).

The integration of the secretory trap vector into the 5'ETS of a ribosomal transcription unit in lines ST576 and ST498 would result in the splice acceptor site from the vector having no consensus upstream splice donor. I have demonstrated the use of a cryptic splice donor in the *En-2* intron sequence in a small proportion of the fusion transcript. However, RNA'se protection assays confirm that this donor is used inefficiently. This may be due to the absence of a 5' cap on the primary transcript. The cap binding proteins CBP20 and CBP80 have recently been demonstrated to play an important role in the splicing of 5' exons (Izzuralde et al, 1994). Thus, the integration sites found in these two cell lines would result in a genomic structure analogous to the 'outron' described in *C.elegans trans*-spliced genes. A proportion of the *lacZ* fusion transcript has been demonstrated to utilise one particular cryptic splice donor from the *En-2* intron, and there may also be other cryptic donors in the intron sequence. However, molecules transcribed by PolI and *cis*- spliced in this way would not be expected to receive a cap structure. Efficient expression of protein coding genes under the control of RNA polI promotors can be achieved (Palmer et al,1993) but only if the chimaeric transcript contains an ATG initiation codon, a polyadenylation signal and an Internal Ribosome Entry Site (IRES). The secretory trap vector used in this study contains a poly(A) site, but no ATG or IRES element, so translation would not be expected from transcripts *cis*- spliced to cryptic donors.

Physical Location of the Transcription and Processing of Fusion Transcripts.

Flourescent *in situ* hybridisations to interphase nuclei of line ST576 and, more strikingly, ST478, demonstrate a close co-localisation between the introduced vector sequences and the ribosomal genes, presumably within the nucleolus. Although it has been demonstrated that a certain amount of promiscuous pol II transcription can occur from the murine pol I promotor, this sequestration of vector sequences into the nucleolus alongside the endogenous rRNA genes suggests that the *βgeo* primary transcript is produced by pol I. Attempts to demonstrate conclusively the α-amanitin insensitive (pol I) production of *βgeo* transcripts in these lines have not, however, been successful.

The integration sites determined in lines ST576 and ST498 are predicted to produce "chimaeric" transcription units, comprising pol I initiation sequences upstream of sequences normally transcribed by pol II. The predicted primary transcripts from these loci contain the first cleavage site used in rRNA processing at position 605 relative to the transcription initiation site, but no other rRNA processing sites. They also contain a 3' splice site and branch-point adenosine residue in the absence of any 5' splicing signals. Thus, RNA's transcribed from such a transcription unit can be predicted to be poor substrates for either the pre-mRNA splicing machinery or the ribosomal RNA processing machinery. The chimaeric nature of the transcript may, itself, result in poor processing efficiency.

The localisation of steady state *βgeo* transcripts to areas of the nucleus other than the nucleolus, where ribosomal RNAs are processed, suggests that they are processed by the mRNA splicing machinery. The fusion

transcripts form focal domains of accumulation, similar to those previously reported for total poly A+ RNA (Carter et al, 1991). The mechanism by which the nascent *βgeo* transcripts are translocated from their site of synthesis within the nucleolus to their sites of processing by the pol II splicing machinery is not known.

The nascent *βgeo* transcripts may be released from their site of transcription within the nucleolus and make their way to the splicing centres by diffusion in the absence of any interaction with components of the nucleoplasm. In this case, tethering of the transcripts into the accumulations seen by *in situ* hybridisation would be predicted to occur as a result of their interaction with splicing factors on reaching the centres of accumulation of these factors. Alternatively, interactions could be made with splicing factors at the site of transcription of the *βgeo* primary transcript, with the RNA being translocated to its sites of processing as a ribonucleoprotein. The latter hypothesis fits better with the view that the localisation of RNA species is tightly controled throughout their genesis and processing, probably as ribonucleoproteins (Maquat, 1991). Recent work on a nuclear structure known as the coiled body has shown that this structure contains the splicing factors U1, U2, U4/6 and U5 snRNPs (Carmo-Fonseca et al, 1992). A close physical association has also been demonstrated between the coiled body and the nucleolus in both mammalian and plant cells (Carmo-Fonseca et al, 1993; Lafontaine and Chamberland, 1995). While the function of the coiled body is, as yet, unknown, this structure provides a way by which splicing factors involved in pre-mRNA processing can come into close juxtaposition with the nucleolus and, hence, the nascent *βgeo* transcript in the *trans*-splicing lines.

Whether the primary *βgeo* transcript forms its association with elements of the splicing machinery near to its site of transcription, or near to the site of its subsequent processing, these associations must be made between the splicing factors and the 3' splice acceptor site in the absence of an upstream 5' splice site. The conventional view of spliceosome assembly is that the 5' and 3' splice sites become paired across an intron. This cannot be the case for these *trans-* splicing reactions. It has been suggested that *trans-* splicing in *C.elegans* is initiated by the formation of an early precursor to the splicosome on the 3' splice site. The failure to locate an upstream 5' splice site in cis leads to the use of the 5' splice site from the SL RNA (Blumenthal, 1995). Alternatively, according to the exon definition model of Robberson et al (1990), a complete spliceosome could form on the 3' splice site, in conjunction with downstream elements in the transcript. In *C. elegans* genes which undergo *trans-* and *cis-* splicing simultaneously, the associations would be made with the next downstream 5' splice site, whereas in trypansome genes, or the *βgeo* construct in lines ST576 and ST478, the associations would be made with the polyadenylation signal. Recent in vitro studies using mammalian cell extracts have demonstrated that the choice of the 5' splice site used for *trans-* splicing can be made after the initial assembly of splicing factors at the 3' *trans-* splice site (Chiara and Reed, 1995).

In the study by Chiara and Reed, the ability of the 3' exon to act as a substrate for *trans-* splicing was greatly increased by the inclusion of a 5' splice site or a splicing enhancer region immediately downstream of the 3' splice acceptor. Splicing enhancers have been described as regions present in downstream exons whose deletion or mutation leads to inefficient splicing of the upstream intron (Tian and Maniatis, 1993;

Watakabe et al, 1993). A comparison of splicing enhancer sequences from a number of genes has loosely defined them as small series of polypurine stretches. They are believed to function through an interaction with U1 snRNP, increasing the stability of pre-spliceosomal assembly on the 3' splice site (Watakabe et al, 1993). Although no sequences with strong homology to the consensus mammalian 5' splice site are found in the pGT1.8tm vector downstream of the 3' splice site, a region containing a number of polypurine stretches is present within the *CD4* sequences about 120 nucleotides downstream of the splice site (figure 5.13). These sequences could be instrumental in the predisposition of the *En-2* splice acceptor site to *trans-* splicing. Thus, it is likely that the initial stages of spliceosome assembly can occur on the 3' splice site contained in the primary *βgeo* transcript, in conjunction either with the polyadenylation signal, the putative splicing enhancer region, or both, with the choice of 5' splice donor being made later. The demonstration of an interaction between the polypurine stretches in the *CD4* sequence and U1 snRNP would help to confirm their action as a splicing enhancer. The placement of this sequence downstream of the 3' splice site in an in vitro system such as the one used by Chiara and Reed could also be used to address this question.

The Choice of Endogenous Genes for *Trans-* Splicing

RNase protection assays have demonstrated that a number of the same endogenous genes are used in *trans-* splicing reactions with the *βgeo* primary transcript in the two independent cell lines ST576 and ST478. The reasons why these particular transcripts are chosen for *trans-* splicing are unclear at the moment. The most straightforward explanation would be that transcripts normally present at a high concentration within the

nucleus become the substrates for *trans-* splicing. This initially seemed a likely hypothesis as whole mount *in situ* hybridisation had demonstrated the *βgeo* transcript acting as the acceptor molecule to be present at extremely high levels within the nuclei of both cell lines. Other explanations may involve the transcription of the *βgeo* transcript and the molecules spliced to it within a single domain of the interphase nucleus. Alternatively, the RNAs used for *trans-* splicing may contain sequence elements or regions of secondary structure that make them susceptible to *trans-* splicing.

Northern blotting has demonstrated that the endogenous genes involved in *trans-* splicing are not all abundantly expressed, with a number of them being undetectable by Northern. Consequently, the simplest view of *trans*-splicing to those RNA molecules present at the highest steady state concentrations must be discounted. Equally, whole mount *in situ* hybridisation has shown that the endogenous transcripts do not accumulate in the nucleus, so a natural tendency for transcripts to be retained in the nucleus is not a factor in determining which transcripts will be *trans-* spliced.

Models of mRNA processing whereby each transcript is spliced at, or near to its site of transcription can be used to suggest that genes which are transcribed in close proximity to the site of *βgeo* transcription may be preferentially used in *trans-* splicing reactions. Flourescent *in situ* hybridisation, however, demonstrate that, at least for two of the *trans-*spliced endogenous genes, the position of the gene within the interphase nucleus and, hence, the site of transcription of its RNA, is not close to the site of the vector DNA.

Early experiments using synthetic mammalian pre-mRNAs in in vitro *trans-* splicing systems demonstrated that the efficiency of *trans-* splicing could be increased by the inclusion of short regions of complementarity between the two substrate molecules (Solnick, 1985; Kornarska et al, 1985). These sequence specific interactions were, however, not necessary for *trans-* splicing to occur. It is possible that regions of complementarity exist between the *En-2* intron, and the introns of the endogenous genes to which the *En-2* splice acceptor *trans-* splices. The *En-2* intron gives a smeared banding pattern when used as a probe for Southern blots containing total genomic DNA, suggesting that it may be slightly repetitive in nature (data not shown). This repetitiveness increases the likelihood of ccomplementary sequences being found in the introns of endogenous genes. Further analysis of the genomic structures of the endogenous genes subject to *trans-* splicing may explain why they are targets.

Genomic clones of five of the genes used for *trans-* splicing were isolated in order to investigate the sequences that are joined to the *En-2* splice acceptor in RACE clones from line ST576. For each gene, the sequence immediately surrounding the breakpoint in the original RACE clone was compared to the consensus mammalain splice donor sequence. For the two known genes (Cytochrome C Oxidase VIIc and the G-protein) and for the gene showing homology to the *Drosophila neu* gene, sequences immediately downstream of this region were compared to the published cDNA sequences. All but one of the sites used in the *trans-* splicing reaction matched the consensus splice donor sequence at the strictly conserved positions. Additionally, the G-protein gene and the *neu* homolog showed no further homology to published cDNA sequences

downstream of the putative splice donor site. This confirms that the majority of the sites used for *trans-* splicing in these lines are genuine splice donor sites. However, sequencing of the cytochrome C Oxidase VIIc gene revealed that the site in this gene that joins to the *En-2* splice acceptor in clone 4B-11 shows no similarity to the consensus splice donor sequence and, indeed, occurs within exon sequences rather than at an intron/exon boundary. The reason for this is unclear. The site used in this gene may be a non-consensus region, behaving in a manner biochemically similar to a normal splice donor site. Alternatively, the cytochrome C oxidase VIIc transcript may be joined to the *βgeo* transcript by a novel mechanism, distinct from splicing, such as an intermolecular ligation event. Such events have been suggested previously to account for the occurrence of unusual *c-myb* transcripts in chick cells (Vellard et al, 1991). A thymus-specific exon of the *c-myb* proto-oncogene was mapped to a chromosome other than chromosome 3, to which the remaining exons of the gene map. Sequencing data did not detect consensus splice sites at the junction between the thymus specific exon and the remainder of the transcript, leading the authors to suggest that this event occurs as a result of an RNA recombination process other than conventional splicing, possibly involving repeat elements within the *c-myb* transcript.

Further analysis of genomic clones of the genes used for *trans-* splicing in these lines may reveal why these transcripts and, more specifically, these donor sites, are used for *trans-* splicing. Although the splice donor sites used for *trans-* splicing do not appear to be particularly strong there may be aspects of their position within the genes which make them susceptible to *trans-* splicing. It has been suggested that in genes

containing exons and introns of varying sizes, a combination of exon definition and intron scanning may be employed in the recognition of pairs of splice sites (Berget, 1995). For *trans-* splicing to occur, the 5' exon needs to be recognised by the exon definition model. Thus, the intron/ exon structure of the genes used for *trans-* splicing may be important. Further analysis of the *neu* homolog gene is being carried out by W. C. Skarnes and J. E. Moss.

The gene trap integrations characterised in lines ST576 and ST478 are in ribosomal RNA genes and result in the formation of 'chimaeric' pol I transcripts that contain the 5'ETS of the ribosomal gene upstream of the *βgeo* coding regions. The recovery of gene trap insertions into pol I transcription units was unexpected, since the *βgeo* reporter does not contain a translation initiation signal and so requires splicing to upstream exons of pol II genes for its activation. Furthermore, pol I transcripts lack a monomethylguanosine cap structure required for efficient splicing and translation. *Trans*-splicing now provides a mechanism to explain how gene trap insertions into non-pol II transcription units can potentially produce translated products (figure 5.14).

The reason for the participation of certain endogenous transcripts in *trans*-splicing is unclear. There is no correlation between the expression level of the endogenous genes and their ability to *trans*-splice. The endogenous transcripts that are used for *trans*-splicing are not transcribed in close proximity to the *βgeo* transcript, so a physical association between nascent transcripts can also be ruled out as an explanation. In vitro *trans*-splicing studies using mammalian cell extracts (Konarska et al, 1985;

Solnick, 1985) suggest that small regions of complementarity between intron sequences increase the efficiency of intermolecular splicing. Sequence specific interactions between the *En-2* intron and the introns of certain endogenous genes may play a role in the choice of transcripts undergoing *trans*-splicing in gene trap electroporated ES cells.

Figure 5.13: Sequences Present in the Secretory Trap Vector pGT1.8TM Include a Series of Purine Rich Regions That May Function as a Splice Enhancer.

atgacaaccagGTCCCAGGTCCCGAAAACCAAAGAAGAAGAACCCTAACAAAGAGGAC
AAGCGGCCTCGCACAGCCTTCACTGCTGAGCACAGCCTTCACTGCTGAGCAGCTCCAG
AGGCTCAAGGCTGAGTTTCAGACCAACAGGTACCTGACAGAGCAGCGGCGCCAGAGCA
GCGGCGCCAGAGTCTGGCACAGGAGCTCGGTACCCGGAA**GATCTGCGGGCTGCAGG**
GAGAGTTGAGATGGAAGGCAGAGAAGGCTCCTTCTTCCCAGTCCTGGATCA
CCTTCTCCCTAAAGAACCAAAAGGTGCTGTGCAGAAGTCTACTAGCAACCC
CAAGTTCCAGCTGTCCGAAAGCGTCCCACTCACCCTTCAGATACCCCTTCA
GATACCCCAGGTCTCCCTTCAGTTTGCTGGTTCTGGCAACCTGACCCTGAC
TCTGGACAGAGGGATACTGTATCAGGAAGTGAACCTGGTGGTGATGAAAGT
GACTCAGCCCGACAGCAACACTTTGACCTGTGAGGTGATGGGACCCACCTC

Lower case denotes *En-2* intron sequence. Upper case denotes exon sequence. Bold type denotes *CD4* sequences. Underlined regions are polypurine stretches. Although these are not uncommon in genomic sequence, clusters of such polypurine rich sequences are unusual.

Figure 5.14: Transcripts Predicted to be Produced From the Integration of pGT1.8TM in Lines ST576 and ST478.

A) Structure of the secretory trap vector pGT1.8TM
The transmembrane domain of rat *CD4* (black box) and the *βgeo* gene (shaded box) are placed downstream of the *En-2* intron (black line) and splice acceptor region (grey box)

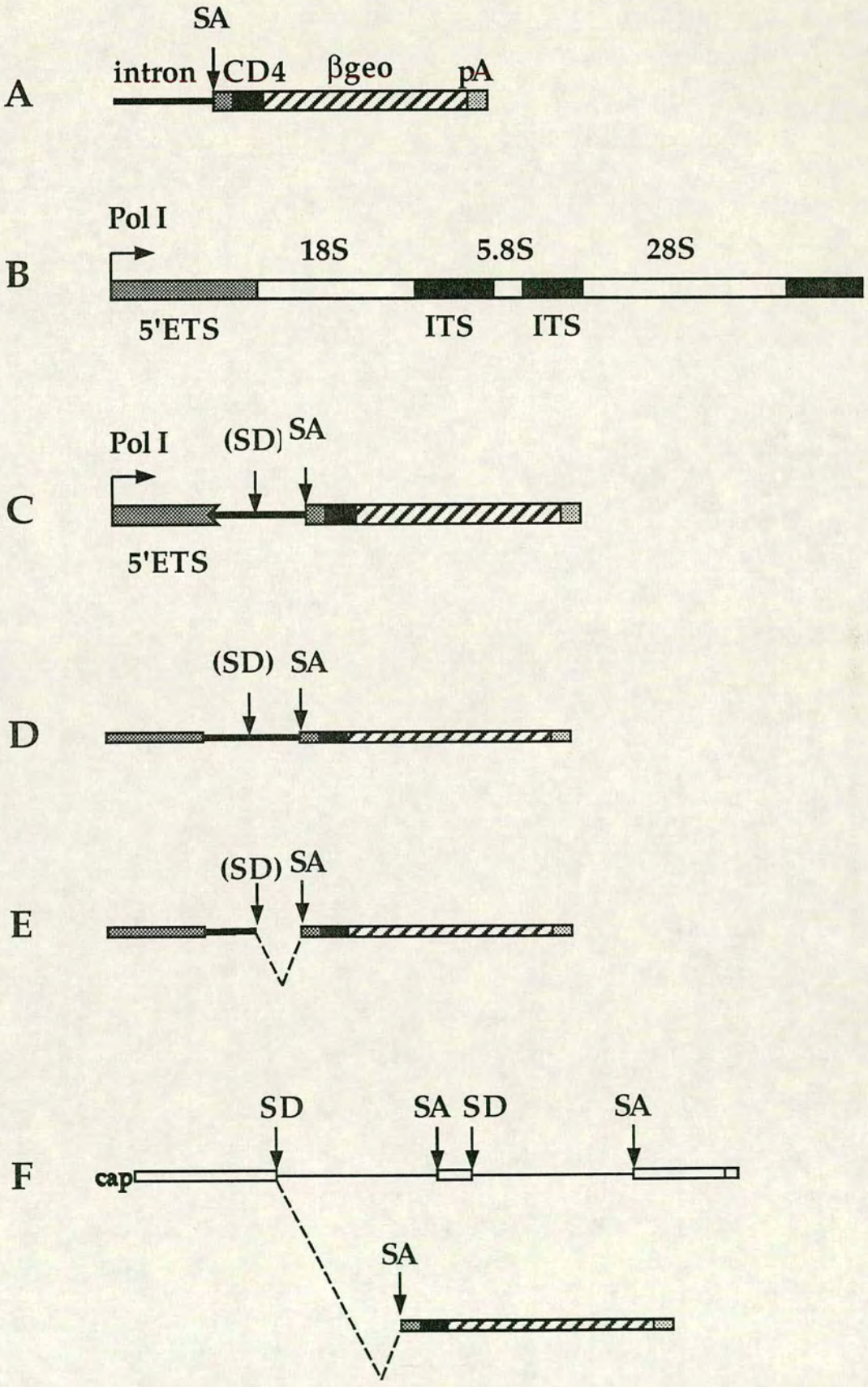
B) Structure of a mammalian rDNA transcription unit. Transcription by RNA pol I produces a long transcript containing the 5' external transcribed spacer (5-ETS-grey box) and the 18S, 5.8S and 28S ribosomal subunits (white boxes) separated by internal transcribed spacers (ITS-black boxes).

C) Structure of the integration in lines ST576 and ST478. The vector has inserted into the 5'ETS of an rRNA gene in each line.

D) Primary transcript predicted to be produced from an integration of this type. The transcript is an uncapped pol I transcript, containing the *βgeo* coding sequence downstream of the *En-2* intron and splice acceptor site. The splice acceptor site has no consensus upstream splice donor. This primary transcript is predicted to remain in the nucleus.

E) *Cis-* splicing to a cryptic donor within the *En-2* intron produces a transcript which is spliced but not capped. The use of a cryptic splice donor site within the *En-2* intron produces a transcript containing the 5' region of the intron. This splicing event was detected by 5'RACE cloning as clone 4A-8 and by RN'ase protection assay. Transcripts of this type may be exported to the cytoplasm, but will not be translated as they are uncapped and the *En-2* intron contains stop codons in all three frames.

F) *Trans-* splicing to splice donor sites from endogenous genes produces a capped transcript containing no intron sequences. *Trans-* splicing of the *En-2* splice acceptor to the splice donor of an endogenous transcript produces a spliced, capped pre-mRNA that can be translated to give a *βgeo* fusion protein to confer neomycin resistance on the cells.


A
SA
intron
CD4
βgeo
pA
B
Pol I
18S
5.8S
28S
5'ETS
ITS
ITS
C
Pol I
(SD)
SA
5'ETS
D
(SD)
SA
E
(SD)
SA
F
SD
SA
SD
SA
cap
SA

CHAPTER 6

GENERAL DISCUSSION

The Use of Gene Trap Vectors in the Study of RNA Localisation.

The use of gene trapping to study RNA localisation is a novel application of the technique. The transcripts of a number of endogenous genes have previously been demonstrated to show highly regulated sub-cellular distributions, controlled by sequence elements within their 3' untranslated regions. The results obtained in the RNA localisation screens presented in chapter 3 suggest that highly localised transcripts are unusual. This is not an unexpected result, as the previously described transcript localisations involve genes whose products are involved in the determination of structural characteristics of cells and developing embryos. A number of different transcript localisation patterns were, however, detected using the gene trap technique. Vectors designed to form fusions containing the 5' regions of endogenous genes allowed the detection of three major transcript distributions: uniform cytoplasmic, granular cytoplasmic and nuclear localisation. The vector designed to trap 3' regions of endogenous genes showed grainy staining in all cell lines, with a characteristic focal accumulation of signal in approximately 38% of these lines. As discussed in chapter three, it seems likely that the granular distribution seen with the 3' trap vector reflects the normal trafficking of endogenous transcripts as ribonucleoprotein particles. The uniform distribution seen using 5' gene trap vectors may be a result of the absence of essential protein binding signals from these fusion transcripts. Differences in the design of the vectors can, therefore, be seen to have a dramatic effect on the processing and transport of the resultant fusion transcripts. The potential involvement of the 5' regions of transcripts in

their localisation is of particular interest, since all of the RNA localisation signals described to date have resided within the 3'UTR.

Further investigation of the role of 5' signals in RNA transport would require detailed sequence comparison of fusion transcripts showing a uniform distribution with those showing a grainy distribution. Initially, cloning of the trapped genes from a number of cell lines showing uniform distribution would be required. If the assumption that all transcripts normally have a granular distribution is correct, whole mount *in situ* hybridisation using probes to the endogenous genes in wild type cells should give a grainy signal. This would confirm that the uniform distribution seen in some gene trap cell lines is caused by the disruption of normal transcript metabolism.

The influence of vector design on the data obtained from screens for RNA localisation is further highlighted by the observation of a small number of cell lines showing nuclear fusion transcript. Nuclear localisation of transcripts was seen exclusively in lines electroporated with 5' gene trap vectors. A number of cell lines showing nuclear localisation of fusion transcripts have been investigated, demonstrating the potential for 5' gene trap technology to be applied to the study of RNA processing.

Nuclear Localisation of Fusion Transcripts is Indicative of Inefficient Splicing

There appear to be at least two classes of integration event that result in nuclear localisation of the fusion transcript. In each case, I have demonstrated that the introduced *En-2* splice acceptor site is used

inefficiently. The use of the vectors pGT1.8βgeo in fibroblasts and pGT1.8K in ES cells gave rise to lines that showed high levels of nuclear accumulation in all cells. Three of these line, TβP20,8; TβP20,29 and ST416 have been studied further, as described in chapter 4. In each of these lines, it appears that a single fusion transcript is present, undergoing inefficient splicing. The most likely explanation for this phenotype is that the vector has integrated into the exon of an endogenous gene, resulting in competition between the endogenous splice acceptor and the introduced *En-2* splice acceptor (see chapter 4, page 194). Further analysis of the integration sites in lines displaying this localisation would be required to investigate this idea. 5' RACE cloning from cytoplasmic RNA from the three lines met with limited sucess. The clones obtained from the fibroblast line TβP20,29 all contained intron sequence. Endogenous sequence was obtained from the ES cell line ST416, but this represented a novel gene, making further analysis of the integration difficult. Insertion of the 3' trap vector into the exon of a gene would not give rise to intron-containing transcript (figure 6.1). Whether splice site selection is made by a scanning mechanism or by exon definition, the introduced splice donor will be the preferred processing site. This difference between the mechanisms of the two vectors accounts for the lack of nuclear localisation seen in clones electroporated with the 3' trap vector.

A second class of nuclear localised lines was obtained using the secretory trap vector, pGT1.8tm, in ES cells. *LacZ* whole mount *in situ* hybridisation on these lines gave nuclear staining, with a high degree of variability in the intensity of staining. In some cells, there was no detectable staining whereas in others, the entire nucleus stained very darkly within a few minutes of applying the staining solution. The level

of staining did not appear to be a function of the state of differentiation of the cells. The integration site in each of the two lines studied was in the 5' external transcribed spacer of a ribosomal RNA gene. This is consistent with the variable staining seen in the cells, as ribosomal transcription units are subject to cell cycle regulation. These lines have been demonstrated to process the fusion transcripts by an intermolecular splicing reaction similar to the *trans-* splicing documented in lower eukaryotes such as nematodes and trypanosomes.

Attempts to Recreate the *Trans-* Splicing Phenotype

Transcripts produced in lines ST576 and ST478 undergo *trans-* splicing. The gene trap integrations in these lines are very similar. Both occur within the 5'ETS of a ribosomal RNA gene, producing an uncapped pol I primary transcript in which the *En-2* splice acceptor site has no consensus upstream donor.This genomic structure resembles the 'outron' described in *C.elegans trans-* spliced genes (Conrad et al, 1992). In order to investigate which features of this primary transcript are required for *trans-* splicing, two expression vectors were constructed to drive an 'outron' like structure from a pol II and a pol I promotor (figure 6.2). Each of these vectors contained the *En-2*, *βgeo* and poly-adenylation site regions from the gene trap vector pGT1.8βgeo. In pβSAgeo, these sequences are driven by the β-actin (pol II) promotor. In pPolISAgeo, the sequences are driven by an rDNA (pol I) promotor. Each vector is linearised prior to electroporation into ES cells and is expected to integrate randomly into the genome.

pβSAgeo was electroporated into 10^8 ES cells, and cell lines isolated. Data presented in chapter 5 suggests that *trans-* splicing in ES cells is of low

efficiency, and it is likely that the high levels of primary fusion transcript generated in lines ST576 and ST478 are important to generate enough *trans-* spliced mRNA to be translated to confer neomycin resistance on the cells. Accordingly, whole mount *in situ* hybridisation was used to look for neomycin-selected colonies containing high levels of nuclear localised *lacZ* containing transcript. A small number of colonies were obtained using this vector. Of 98 cell lines isolated, none showed accumulation of lacZ transcript in the nucleus. The electroporation was repeated, this time giving a larger number of colonies (about 4000). Replica filters were made from these plates (see chapter 2) and these filters used for a whole mount *in situ* hybridisation with a *lacZ* riboprobe. Colonies which gave a strong signal on these replica filters were picked, grown to confluency in 24 well plates and used for whole mount *in situ*. A total of 66 clones were isolated in this way. Only one of these lines, clone 9E, showed nuclear localisation, with strong cytoplasmic staining being the most common observation. Twelve of these cell lines were taken at random and used for X-gal staining to detect β-gal protein activity. Interestingly, all of these clones showed nuclear β-gal staining. This suggests that a specific splicing event may be taking place in all of these cell lines, giving rise to a nuclear protein.

A cryptic splice donor site in the *En-2* intron sequence has been demonstrated to be used at low efficiency in lines ST576 and ST478. In these lines, the primary fusion transcript is predicted to be an uncapped pol I transcript. It is, therefore, possible that the use of a pol II promoter to drive these vector sequences may lead to efficient *cis-* splicing to this cryptic donor. In order to investigate this possibility, RNase protection assays were carried out on total RNA from four of the cell lines

electroporated with pβSAgeo. The probe used spanned the cryptic splice donor (see figure 5.9). This cryptic donor was not used in any of the lines tested (data not shown). This suggests that the *En-2* splice acceptor site is efficiently splicing to another cryptic splice donor, probably within the vector sequence. The use of the cryptic *cis-* splice site previously documented in lines ST576 and ST478 is unlikely to account for the nuclear localised proteins seen in these lines, as mRNAs spliced in this way could not code for a protein containing a nuclear localisation signal (NLS) in any frame. These lines may alternatively represent gene trap events, resulting from the loss of part or all of the promoter sequence prior to insertion of the vector. The buffer sequence between the site of linearisation of the plasmid and the beginning of the promotor is very short (17bp) so this amount of trimming back of the linearised plasmid is not unexpected. It is possible that these lines are efficiently *trans-* splicing the *βgeo* transcript, but this is highly unlikely bearing in mind that both of the previously described *trans*-splicing lines contain *βgeo* sequences under the control of a pol I promoter.

The use of the vector pPolISAgeo, which uses the murine rDNA promotor to drive the 'outron' like structure also failed to reproduce the *trans-* splicing phenotype, as judged by nuclear localisation of fusion transcripts. In an electroporation of 10^8 CGR8 cells, only five neomycin resistant colonies were obtained. These were presumably a result of trimming of the promoter sequences leading to gene trap events. In this vector, the buffer sequence upstream of the promotor was also 17bp. Since this construct used a pol I promotor to drive the test sequences, a comparison of the primary transcript predicted to be generated by this construct with that predicted to be generated in lines ST576 and ST478

may give some further information about the sequence elements necessary to promote *trans-* splicing in ES cells.

There are two major sequence elements which are present in the predicted primary transcript from the insertions in the *trans-* splicing lines and absent from the transcript predicted to be produced from the plasmid pPolISAgeo. These are the 5'ETS from the endogenous rRNA gene and the *CD4* sequences from the secretory trap vector. The integrations in ST576 and ST478 are very close to each other, both lying within the 5'ETS. A region of predicted secondary structure forming a giant three-branched hairpin loop (Bourbon et al, 1988) lies immediately upstream of the insertion in ST576, and a small distance 5' to the predicted site of insertion in ST478 (figure 6.3). The close proximity of these independant insertion sites is unlikely to be merely co-incidental. The structure of the 5'ETS immediately upstream may have some importance in the ability of these transcripts to *trans-* splice. The first processing site of the ribosomal RNA is at nucleotide position 650, which is present in the fusion transcript produced by both lines. This processing site, together with elements of the secondary structure of the 5'ETS may be important in the recruitment of protein factors onto the fusion transcript which can then help to recruit in pol II splicing factors, although no protein factors with a role in both pol I and pol II transcript processing have yet been identified. Alternatively, accurate cleavage at the first processing site in the ribosomal RNA region may be necessary to release the *En*-2 splice acceptor site in the correct context to participate in *trans-* splicing.

The incorporation of the vector DNA into the nucleolus may also be of importance for the *trans-* splicing phenotype. Active transcription of ribosomal genes is required for the assembly of the nucleolus (Benavente et al, 1987). The precise region of the nascent transcript involved in this nucleolus organising activity has not been mapped, but the region of the 5'ETS present in the fusion transcripts in lines ST576 and ST478 may be involved in the organisation of the vector DNA into the nucleolus.
The *CD4* sequences present in the secretory trap vector pGT1.8tm are also missing from the expression vector pPolISAgeo. As described in chapter 5, these *CD4* sequences include a series of polypurine regions, which may act as a splicing enhancer. If these sequences do enhance splicing at the *En-2* 3' splice site, then their inclusion in the primary transcript may be necessary to obtain sufficient *trans-* splicing to give neomycin resistant cells.

The construction of further vectors containing one or both of these sequence elements could be used to assess the importance of these features for the *trans-* splicing phenotype. The recreation of the *trans-* splicing phenotype would define more precisely the key elements involved.

Insights Into RNA Metabolism Revealed by Gene Trapping.

The possibility of *trans-* splicing as a common form of RNA processing in mammalian cells has long been a subject of speculation. It has been known for some time that mammalian cells contain all of the accessory factors required for *trans-* splicing, as splice acceptors and donors on different molecules can be spliced together in cell free splicing extracts (Konarska et al, 1985; Solnick, 1985). Additionally, spliced leader RNAs

from lower eukaryotes can be accurately spliced to a range of splice acceptor sites in HeLa cells (Bruzik and Maniatis 1992). *Trans-* splicing between consensus mammalian splice donors and acceptors in vivo has not, however, been conclusively demonstrated. As described in chapter 5, this is largely because DNA rearrangements cannot be ruled out as explanations for putative in vivo *trans-* splicing events.

Data presented in this thesis document in vivo *trans-* splicing of the consensus mammalian splice acceptor from the *En-2* gene to a number of different endogenous splice donors in vivo in two independant cell lines. Each of these lines contains a single site of integration of the vector, ruling out events at the DNA level as an explanation for the numerous fusion transcripts seen. Although these observations do not confirm that mammalian transcripts are normally subject to *trans-* splicing, they do strengthen the argument for *trans-* splicing as a normal phenomenon in mammalian cells. The selection of transcripts to undergo *trans-* splicing in the gene trap lines is not understood. It is possible that these particular splice donors are subject to *trans-* splicing in normal cells. 3'RACE cloning using primers to regions of the endogenous genes near the splice site used in lines ST576 and ST478 could be used to investigate this possibility.

Protein Translation From Non-pol II Transcripts.

Another question raised by these results is that of the relationship between the polymerase that transcribes a sequence and its final function. Usually, protein coding genes are transcribed by RNA polymerase II with pol I activity being restricted to the production of ribosomal RNAs. Attempts to express protein coding genes from non-pol II promotors have

met with limited sucess. Lapota et al (1986) reported the production of small amounts of bacterial chloramphenicol acetyltransferase (CAT) from a chimaeric construct in which the protein coding sequences were driven by a pol I promotor. Closer examination of this phenomenon, however, revealed that the translated polysome-associated transcripts were abberantly initiated and resulted from a small amount of cryptic pol II activity within the murine pol I promotor sequence. In lines ST576 and ST478, the presence of abberantly initiated pol II transcripts containing the *βgeo* coding sequences cannot be conclusively ruled out. However, the design of the secretory trap vector is such that these transcripts could not be translated, as the *βgeo* sequence contains no initiation signal, and the upstream intron sequence contains stop codons in all three frames. Thus, the neomycin protein activity must be derived from pol I produced transcripts that have obtained initiation signals through *trans-* splicing. This suggests an unexpected degree of flexibility in the use of transcripts by mammalian cells. A recent report of the use of pol III transcripts of viral genes as mRNAs in HeLa cells has also highlighted the flexibility of mammalian RNA processing (Gunnery and Mathews, 1995).

The Potential For Trans-Splicing in Genetic Manipulation.

The idea of targeting *trans-* splicing to alter endogenous transcripts has been investigated using modified ribozymes (Sullenger and Cech, 1994). In these experiments, the self splicing group I intron from *Tetrahymena thermophila* was engineered to contain a 3' exon from the *lacZ* gene. Co-expression of this RNA in *E.coli* with a truncated *lacZ* transcript containing the first 21 nucleotides of the *lacZ* sequence and a 5' splice site sequence led to the production of a functional full length *lacZ* transcript. The full length transcript was produced by targeted *trans-* splicing

catalysed by the ribozyme. Provided that this approach can be adapted to work in mammalian cells, it has great potential in the correction of a broad array of mutant transcripts and in the creation of new mutations.

Targeted *trans-* splicing has certain advantages over more conventional approaches to genetic manipulation. The introduction of a DNA version of an altered gene often results in expression of the gene in the wrong tissue types or at the wrong level. The alteration of endogenous transcripts by targeted *trans-* splicing would avoid such problems. Accurate targeting of *trans-* splicing does pose significant problems, however. Earlier experiments using ribozymes to cleave retroviral RNA (Sullenger and Cech, 1993) highlighted the importance of RNA trafficking within the cell for the targeting of ribozyme activity. In order to cleave retroviral genomic RNA containing *lacZ* sequences, the hammer-head ribozyme used needed to include a retroviral packaging signal. This was predicted to direct the ribozyme to the same sub-cellular compartment as its target molecule. Thus, the integration of any effector molecules for targeted *trans-* splicing into the metabolic pathways of the cell is likely to be of fundamental importance to any strategies for genetic manipulation. In the gene trap lines described in chapter 5, the occurrence of *trans-* splicing together with the intra-nuclear speckles seen by whole mount in situ hybridisation suggest that the primary *βgeo* transcripts are integrated into the mRNA processing pathway. If the question of the choice of endogenous transcripts for *trans-* splicing in these lines can be adequately resolved, then an approach using "outron" like structures under the control of non-pol II mammalian promoters (figure 6.6) may encounter fewer, or at least different, problems to approaches using ribozymes, which have no natural counterpart in mammalian cells. Further

understanding of signals involved in the intra-cellular trafficking of pre-mRNAs would also be useful in the design of genetic manipulations of this type. The sub-cellular localisations revealed using gene trap and 3' trap vectors in chapter 3 demonstrate the potential of gene trap technology for the identification of such signals.

The data presented in this thesis demonstrate novel applications of gene trap vectors in the study of mRNA localisation and RNA processing, raising questions about the rigidity of the control of mammalian transcript processing and usage. I have documented the production of pol I transcripts that can be translated into protein following an unorthodox *trans-* splicing reaction. This demonstrates an unexpected degree of flexibility both in the splicing pathway and in the relationship between the origin of a transcript and its final use. The precursor fusion transcripts used for *trans-* splicing are potentially useful in the study of RNA processing, as they appear to become held up within the splicing machinery following an initial interaction with elements of the spliceosome. In particular, the characteristic sub-nuclear localisation seen in these lines may prove useful for the study of physical aspects of transcript processing. Further analysis and manipulation of the *trans-* splicing seen in lines ST576 and ST478 may also provide a novel technology for the genetic alteration of mammalian cells.

Figure 6.1: Predicted Splicing of Fusion Transcripts Produced by the Integration of the 3' Trap Vector pβKnSD into the Exon of an Endogenous Gene.

A) Structure of pβKnSD. The *neomycin* gene (shaded box) and a region of the *β-globin* intron (thick black line) containing a splice donor site (SD) are driven by the *β-actin* promotor (black box).

B) Predicted genomic structure caused by the insertion of a single copy of the vector into the exon of an endogenous gene. White boxes represent endogenous exons, thin black lines represent endogenous introns.

C) Predicted stucture and splicing of the primary transcript from the integration if splice site choice is made by the exon definition model (Robberson et al 1990). The introduced splice donor will be recognised in conjunction with the 5' cap structure, as it is closer to the cap than is the endogenous splice donor. The endogenous splice donor is predicted to be removed along with *β-globin* and endogenous intron sequence.

D) Predicted structure and splicing of the primary transcript from this integration if splice site choice is made by a 5' to 3' scanning mechanism (Lang and Spritz, 1983). The introduced splice donor will be recognised as the most 5' donor site in the transcript. Scanning in a 5' to 3' direction is predicted to recognise the next endogenous splice acceptor site. The *β-globin* intron and regions of endogenous sequence including the splice donor site will be defined as an intron and removed by splicing.

E) Predicted mature transcript from the integration. This RNA contains no intron or splice site sequences, so is expected to be exported from the nucleus efficiently.

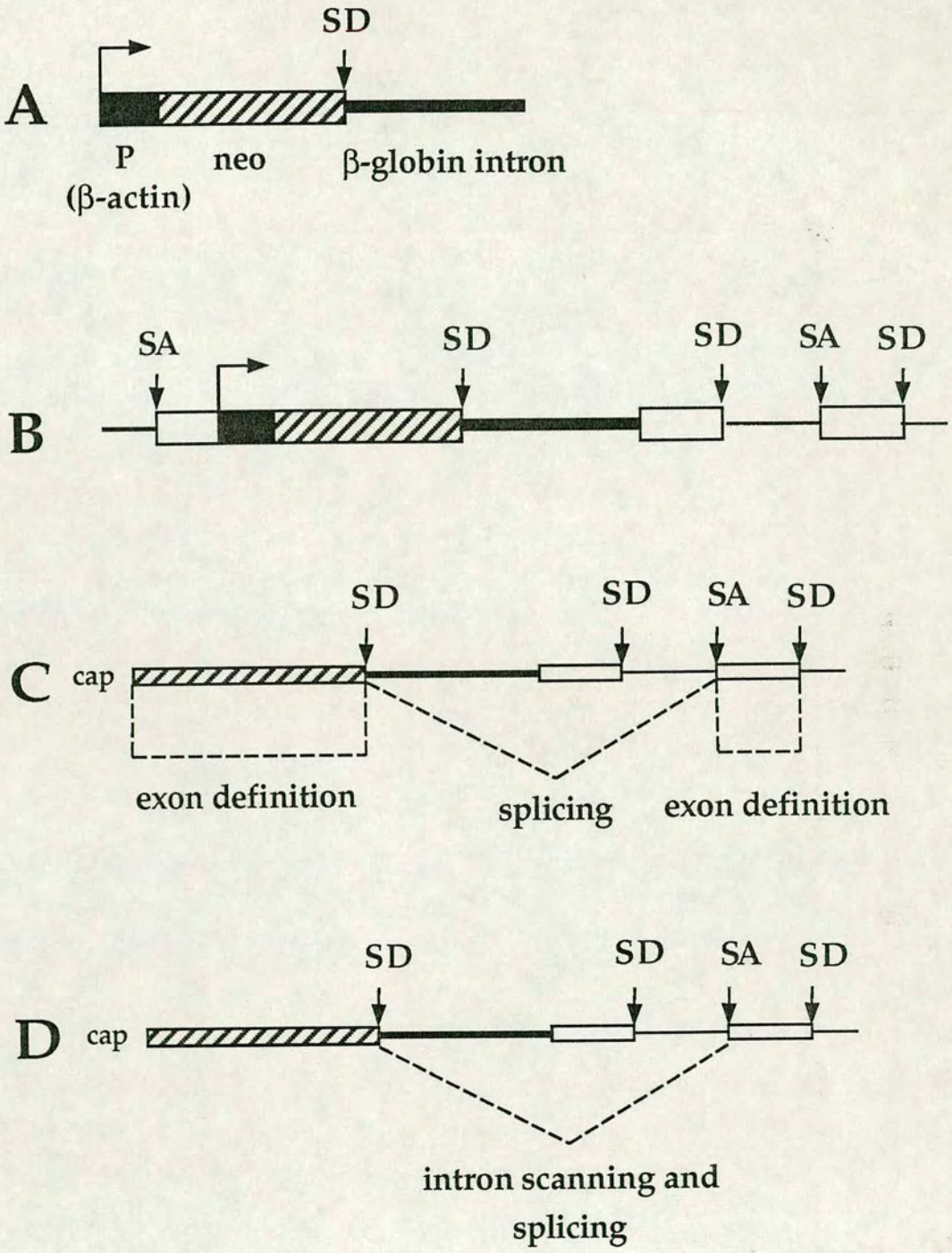

A
SD
P
(β-actin)
neo
β-globin intron
B
SA
SD
SD
SA
SD
C
cap
SD
SD
SA
SD
exon definition
splicing
exon definition
D
cap
SD
SD
SA
SD
intron scanning and
splicing
E
cap

Figure 6.2: Constructs Designed in an Attempt to Recreate the *Trans*-Splicing Phenotype.

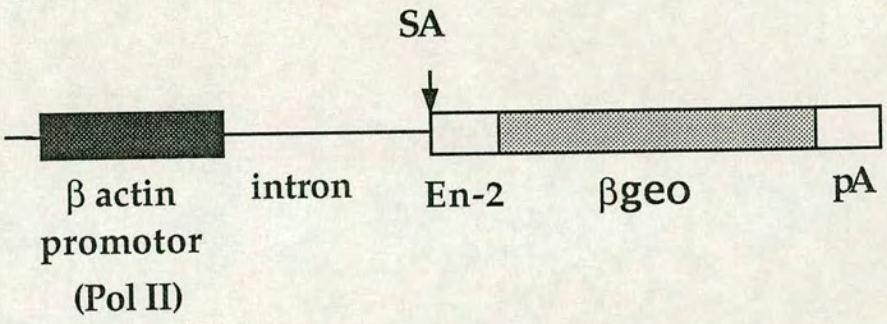
A) pβSAgeo

In this construct, an 'outron' structure containing an intron region and splice acceptor site from *En-2* linked to the *βgeo* gene is driven by the β–actin (pol II) promoter (Frederickson et al, 1989). This vector was constructed by W. C. Skarnes.

B) pPolISAgeo

In this construct, the same 'outron' structure is driven by an rDNA (pol I) promoter (Palmer et al. 1993).

A) pβSAgeo



B) pPolISAgeo

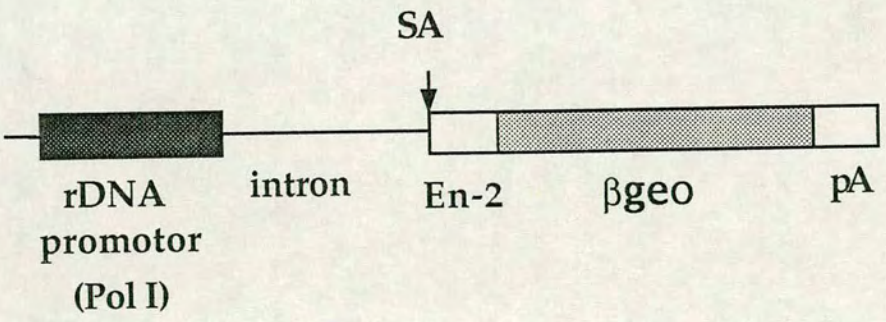


Figure 6.3: The Insertion Sites of the Secretory Trap Vector in Lines ST576 and ST478 are Close to a Region of Highly Conserved Secondary Structure.

A giant three-branched hairpin loop is predicted to occur in the core portion of mouse 5'ETS (diagram taken from Bourbon et al, 1988). The site of insertion of the vector in line ST576 is within this hairpin loop at the site marked. In line ST478, the vector is predicted by PCR data to have inserted approximately 100bp 5' of this region of secondary structure.

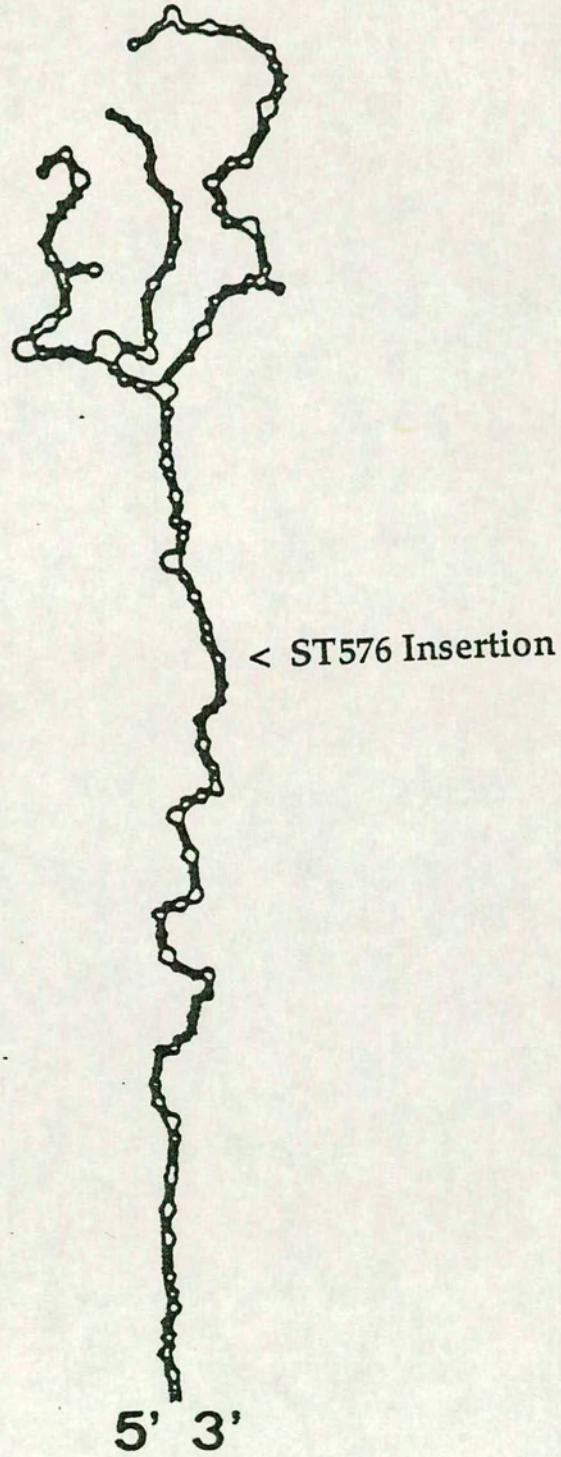
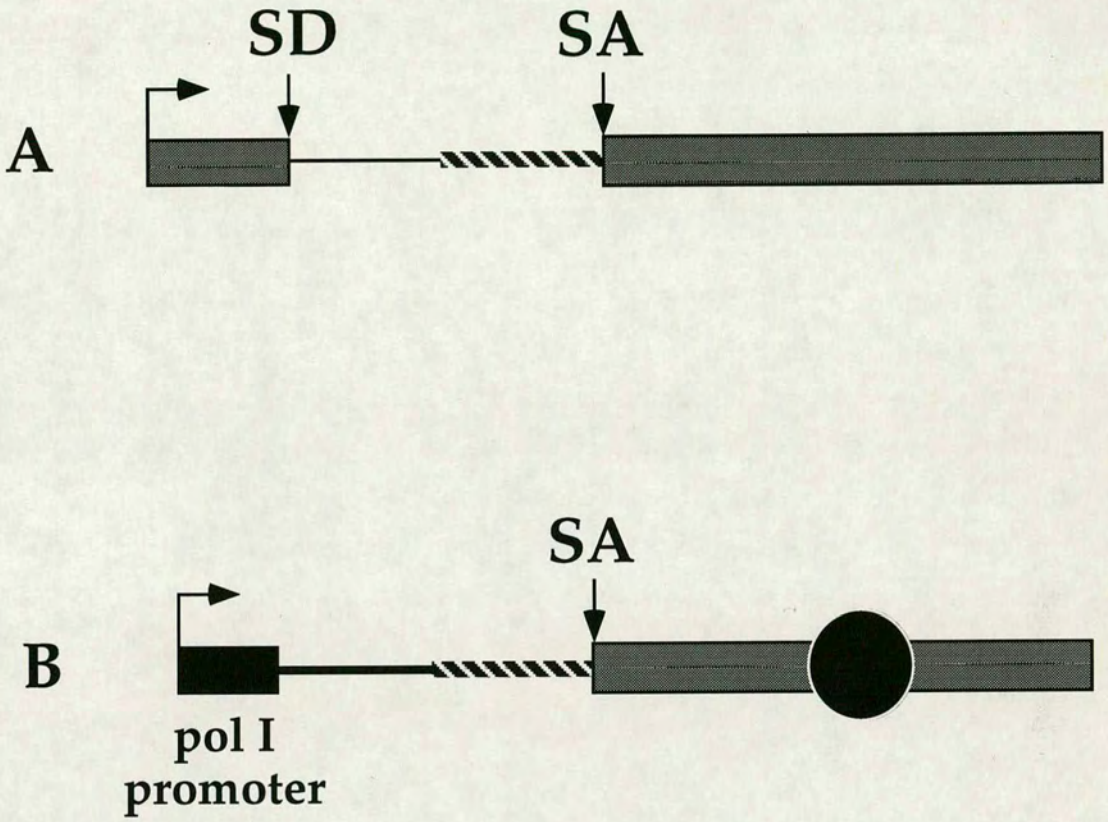

< ST576 Insertion
5' 3'

Figure 6.4: Potential For RNA Polymerase I Driven Vectors in Genetic Manipulation.

A) Simplified Structure of Endogenous Target Gene. Exon sequences are denoted by grey boxes, the intervening intron by a thin black line. Sequences used to produce complementarity in the pol I driven vector are denoted by a striped box.

B) 'Outron' Based Vector. The vector contains a 5' buffer sequence (thin black line) to combat problems with chewing back of introduced DNA. A pol I promoter is used to drive an intron region (thick black line) and consensus splice acceptor (SA) linked to a modified version of the endogenous gene (modification represented by a black circle). A region of complementarity to the endogenous intron upstream of the splice site to be targeted is also included (striped box).

C) Modified Gene Product. Targeting *trans*- splicing in this way, if successful, is predicted to produce a modified transcript of the endogenous gene under the control of the endogenous promoter.


SD
SA
A
SA
B
pol I
promoter
C

REFERENCES

Ainger, K., Avossa, D., Morgan, F., Hill, S. J., Barry, C., Barbarese, E. and Carson, J. H. (1993). Transport and localization of exogenous myelin basic protein mRNA microinjected into oligodendrocytes. The Journal of Cell Biology, **123**, 431-441.

Akamatsu, M. and Grossman, L. I. (1990). Nucleotide sequence of a cDNA for mouse cytochrome C oxidase subunit VIIc. Nucleic Acids Research, **18**, 3645.

Allen, N. D., Cran, D. G., Barton, S. C., Hettle, S., Reik, W. and Surani, M. A. (1988). Transgenes as probes for active chromosomal domains in mouse development. Nature, **333**, 852-855.

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. and Struhl, K. (1987). Current Protocols in Molecular Biology. New York: Greene & Wiley,

Ayane, M., Nielson, P. and Kohler, G. (1989). Cloning and sequencing of mouse ribosomal protein S12 cDNA. Nucleic Acids Research, **17**, 6722.

Barnes, W. M. (1994). PCR amplification of up to 35-Kb DNA with high fidelity and high yield from λ bacteriophage templates. Proceedings of the National Academy of Sciences USA, **91**, 2216-2220.

Bellen, H. J., O'Kane, C., Wilson, C., Grossniklaus, U., Pearson, R. K. and Gehring, W. J. (1989). P-element-mediated enhancer detection: a versatile method to study development in *Drosophila*. Genes & Development, **3**, 1288-1300.

Benavente, R., Rose, K. M., Reimer, G., Hugle-Dorr, B. and Scheer, U. (1987). Inhibition of nucleolar reformation after microinjection of antibodies to RNA polymerase I into mitotic cells. The Journal of Cell Biology, **105**, 1483-1491.

Benton, W. D. and Davis, R. W. (1977). Screening of lambda recombinant clones by hybridization to single plaques in situ. Science, **196**, 180-182.

Berget, S. M. (1995). Exon recognition in vertebrate splicing. The Journal of Biological Chemistry, **270**, 2411-2414.

Bier, E., Vaessin, H., Sheperd, S., Lee, K., McCall, K., Barbel, S., Ackerman, L., Carretto, R., Uemura, T., Grell, E., Jan, L. Y. and Jan, Y. N. (1989). Searching for pattern and mutation in the *Drosophila* genome with a P-lacZ vector. Genes & Development, **3**, 1273-1287.

Blumenthal, T. (1995). *Trans-* splicing and polycistronic transcription in *C. elegans*. Trends In Genetics, **11**, 132-136.

Blumenthal, T. and Thomas, J. (1988). *Cis* and *trans* mRNA splicing in *C. elegans*. Trends in Genetics, **4**, 305-308.

Borst, P. (1986). Discontinuous transcription and antigenic variation in Trypanosomes. Annual Review of Biochemistry, **55**, 701-732.

Boulianne, G. L., de la Concha, A., Campos-Ortega, J. A., Jan, L. Y. and Jan, Y. N. (1991). The *Drosophila* neurogenic gene *neuralized* encodes a novel protein and is expressed in precursors of larval and adult neurons. The EMBO Journal, **10**, 2975-2983.

Bourbon, H., Michot, B., Hassouna, N., Feliu, J. and Bachellerie, J. (1988). Sequence and secondary structure of the 5' external transcribed spacer of mouse pre-rRNA. DNA, **7**, 181-191.

Brockdorff, N., Ashworth, A., Kay, G. F., Cooper, P., Smith, S., McCabe, V. M., Norris, D. P., Penny, G. D., Patel, D. and Rastan, S. (1991). Conservation of position and exclusive expression of mouse *Xist* from the inactive X chromosome. Nature, **351**, 329-331.

Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., Swift, S. and Rastan, S. (1992). The product of the mouse *Xist* gene is a 15Kb inactive X-specific transcript containing no conserved orf and located in the nucleus. Cell, **71**, 515-526.

Brody, E. and Abelson, J. (1985). The 'spliceosome': yeast pre-messenger RNA associates with a 40S complex in a splicing dependant reaction. Science, **228**, 963-967.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J. and Willard, H. F. (1992). The human *XIST* gene: analysis of a 17Kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell, **71**, 527-542.

Bruckenstein, D. A., Lein, P. J., Higgins, D. and Fremeau, R. T. J. (1990). Distinct spatial localization of specific mRNAs in cultured sympathetic neurons. Neuron, **5**, 809-819.

Bruzik, J. P. and Maniatis, T. (1992). Spliced leader RNAs from lower eukaryotes are *trans-* spliced in mammalian cells. Nature, **360**, 692-695.

Carmo-Fonseca, M., Ferreira, J. and Lamond, A. I. (1993). Assembly of snRNP-containing coiled bodies is regulated in interphase and mitosis-evidence that the coiled body is a kinetic nuclear structure. The Journal of Cell Biology, **120**, 841-852.

Carmo-Fonseca, M., Pepperkok, R., Carvalho, M. T. and Lamond, A. I. (1992). Transcription-dependent colocalization of the U1, U2, U4/U6 and U5 snRNPs in coiled bodies. The Journal of Cell Biology, **117**, 1-14.

Carmo-Fonseca, M., Pepperkok, R., Sproat, B. S., Ansorge, W., Swanson, M. S. and Lamond, A. I. (1991). in vivo detection of snRNP-rich organelles in the nuclei of mammalian cells. The EMBO Journal, **10**, 1863-1873.

Carter, K. C., Bowman, D., Carrington, W., Fogarty, K., McNeil, J. A., Fay, F. S. and Bentley Lawrence, J. (1993). A Three-Dimensional View of Precursor Messenger RNA Metabolism Within the Mammalian Nucleus. Science, **259**, 1330-1335.

Carter, K. C., Taneja, K. L. and Lawrence, J. B. (1991). Discrete nuclear domains of poly(A) RNA and their relationship to the organization of the nucleus. The Journal of Cell Biology, **115**, 1191-1202.

Casadaban, M. J. and Cohen, S. N. (1979). Lactose genes fused to exogenous promoters in one step using a Mu-lac bacteriophage: in vivo probe for transcriptional control sequences. Proceedings of the National Academy of Sciences USA, **76**, 4530-4533.

Catterall, J. F., O'Malley, B. W., Robertson, M. A., Staden, R., Tanaka, Y. and Brownlee, G. G. (1978). Nucleotide sequence homology at 12 intron-exon junctions in the chick *ovalbumin* gene. Nature, **275**, 510-513.

Chen, Y., Word, C., Dew, V., Uhr, J. W., Vitetta, E. S. and Tucker, P. W. (1986). Double iotype production by a neoplastic B cell line. The Journal of Experimental Medicine, **164**, 562-579.

Cheng, H. and Bjerknes, M. (1989). Asymmetric distribution of *actin* mRNA and cytoskeletal pattern generation in polarized epithelial cells. The Journal of Molecular Biology, **210**, 541-549.

Chiara, M. D. and Reed, R. (1995). A two-step mechanism for 5' and 3' splice-site pairing. Nature, **375**, 510-513.

Church, G. and Gilbert, W. (1984). Genomic sequencing. Proceedings of the National Academy of Sciences USA, **81**, 1991-1995.

Conrad, R., Liou, R. F. and Blumenthal, T. (1993). Functional analysis of a *C.elegans trans-* splice acceptor. Nucleic Acids Research, **21**, 913-919.

Conrad, R., Thomas, J., Speith, J. and Blumenthal, T. (1991). Insertion of part of an intron into the 5' untranslated region of a *Caenhorhabditis elegans* gene converts it into a *trans-* spliced gene. Molecular and Cellular Biology, **11**, 1921-1926.

Cripe, L., Morris, E. and Fulton, A. B. (1993). *Vimentin* mRNA location changes during muscle development. Proceedings of the National Academy of Sciences USA, **90**, 2724-2728.

Dandekar, T. and Sibbald, P. R. (1990). *Trans-* splicing of pre-mRNA is predicted to occur in a wide range of organisms including vertebrates. Nucleic Acids Research, **18**, 4719-4725.

Davis, L. G., Dibner, M. D. and Battey, J. S. (1986). Basic Methods in Molecular Biology. New York: Elsevier, 159-160.

Davis, I. and Ish-Horowicz, D. (1991). Apical localization of pair-rule transcripts requires 3' sequences and limits protein diffusion in the *Drosophila* blastoderm embryo. Cell, **67**, 927-940.

Dower, W. J., Miller, J. F. and Ragsdale, C. W. (1988). High efficiency transformation of *E.coli* by high voltage electroporation. Nucleic Acids Research, **16**, 6127.

Dworetzsky, S. I. and Feldherr, C. M. (1988). Translocation of RNA-coated gold particles through the nuclear pores of oocytes. The Journal of Cell Biology, **106**, 575-584.

Eckner, R., Ellmeier, W. and Birnstiel, M. L. (1991). mature mRNA 3' end formation stimulates RNA export from the nucleus. The EMBO Journal, **10**, 3513-3522.

Edgar, B. A., Weir, M. P., Schubiger, G. and Kornberg, T. (1986). Repression and turnover pattern of *fushi tarazu* RNA in the early *Drosophila* embryo. Cell, **47**, 747-754.

Ellis, L., Clauser, E., Morgan, D. O., Edery, M., Roth, R. A. and Rutter, W. J. (1988). Replacement of insulin receptor tyrosine kinase residues 1162 and 1163 compromises insulin-dependant kinase activity and uptake of 2-deoxyglucose. Cell, **45**, 721-732

Ephrussi, A., Dickinson, L. K. and Lehmann, R. (1991). *oskar* organizes the germ plasm and directs localization of the posterior determinant *nanos*. Cell, **66**, 37-50.

Eul, J., Graessmann, M. and Graessmann, A. (1995). Experimental evidence for RNA *trans-* splicing in mammalian cells. The EMBO Journal, **14**, 3226-3235.

Fakan, S. (1994). Perichromatin fibrils are *in situ* forms of nascent transcripts. Trends in Cell Biology, **4**, 86-90.

Fantes, J. A., Bickmore, W. A., Fletcher, J. M., Ballesta, F., Hanson, I. M. and van Heyningen, V. (1992). Non-radioactive *in situ* hybridisation for the rapid analysis of submicroscopic deletions at the WAGR locus. The American Journal of Human Genetics, **51**, 1286-1294.

Fantes, J. A., Redeker, B., Breen, M., Boyle, S., Brown, J., Fletcher, J., Jones, S., Bickmore, W. A., Fukushima, Y., Mannens, M., Danes, S., van Heyningen, V. and Hanson, I. (1995). Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. Human Molecular Genetics, **4**, 415-422.

Fournier, M. J. and Maxwell, E. S. (1993). Catching up with the spliceosomal snRNAs. Trends in Biochemical Sciences, **18**, 131-135.

Frederickson, R. M., Micheau, M. R., Iwamoto, A. and Miyamoto, N. (1989). 5' flanking and first intron sequences of the human *β-actin* gene required for efficient promotor activity. Nucleic Acids Research, **17**, 253-270.

Frendewey, D. and Keller, W. (1985). Step-wise assembly of a pre-mRNA splicing complex requires U-snRNPs and specific intron sequences. Cell, **42**, 355-367.

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

Fu, X. D. and Maniatis, T. (1990). Factor required for mammalian spliceosome assembly is localized to discrete regions in the nucleus. Nature, **343**, 437-441.

Garner, C. C., Tucker, R. P. and Matus, A. (1988). Selective localization of messenger RNA for cytoskeletal protein MAP2 in dendrites. Nature, **336**, 674-677.

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

Green, M. R. (1991). Biochemical Mechanisms of Constitutive and Regulated Pre-mRNA Splicing. Annual Review of Cell Biology, **7**, 559-599.

Gunnery, S. and Mathews, M. B. (1995). Functional mRNA can be generated by RNA polymerase III. Molecular and Cellular Biology, **15**, 3597-3607.

Gunzl, A., Cross, M. and Bindereif, A. (1992). Domain structure of U2 and U4/U6 small nuclear ribonucleoprotein particles from *Trypanosoma brucei*: identification of *trans-* splicosomal specific RNA-protein interactions. Molecular and Cellular Biology, **12**, 468-479.

Gunzl, A., Cross, M., Palfi, Z. and Bindereif, A. (1993). Assembly of the U2 small nuclear ribonucleoprotein from *Trypanosoma brucei*. The Journal of Biological Chemistry, **268**, 13336-13343.

Guthrie, C. (1991). Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotin. Science, **253**, 157-163.

Hamm, J. and Mattaj, I. (1990). Monomethylated cap structures facilitate RNA export from the nucleus. Cell, **63**, 109-118.

Hanahan, D. (1985). Techniques for transformation of *E.coli*. In D. M. Glover (Ed.), DNA cloning: a practical approach (pp. 109-135). Washington D.C.: IRL Press.

Hannon, G. J., Maroney, P. A., Yu, Y., Hannon, G. E. and Wilson, T. W. (1992). Interaction of U6 snRNA with a sequence required for function of the nematode SL RNA in *trans-* splicing. Science, **258**, 1775-1780.

Harford and Klausner. (1991). Coordinate post-transcriptional of ferritin and transferrin receptor expression: the role of regulated RNA-protein interaction. Enzyme, **44**, 28-41.

Hendzel, M. J. and Bazett-Jones, D. P. (1995). RNA polymerase II transcription and the functional organization of the mammalian cell nucleus. Chromosoma, **103**, 509-516.

Hill, D. P. and Wurst, W. (1993). Screening for novel pattern formation genes using gene trap approaches. Methods in Enzymology, **225**, 665-681.

Hoock, T. C., Newcomb, P. M. and Herman, I. M. (1991). beta actin and its mRNA are localized at the plasma membrane and the regions of moving cytoplasm during the cellular response to injury. The Journal of Cell Biology, **112**, 653-664.

Huang, R., Ozawa, M., Kadomatsu, K. and Muramatsu, T. (1990). Developmentally regulated expression of embigin, a member of the immunoglobulin superfamily found in embryonal carcinoma cells. Differentiation, **45**, 76-83.

Huang, S., Deerinck, T. J., Ellisman, M. H. and Spector, D. L. (1994). In vivo analysis of the stability and transport of nuclear poly(A)+ RNA. The Journal of Cell Biology, **126**, 877-899.

Huang, S. and Spector, D. L. (1991). Nascent pre-mRNA transcripts are associated with nuclear regions enriched in splicing factors. Genes & Development, **5**, 2288-2302.

Huang, S. and Spector, D. L. (1992). Will the real splicing sites please light up? Current Biology, **2**, 188-190.

Izaurralde, E., Lewis, J., McGlugan, C., Jankowska, M., Darzynkiewicz, E. and Mattaj, I. W. (1994). A nuclear cap binding protein complex involved in pre-mRNA splicing. Cell, **78**, 657-668.

Izurralde, E. and Mattaj, I. W. (1995). RNA Export. Cell, **81**, 153-159.

Jackson, D. A., Hassan, A. B., Errington, R. J. and Cook, P. R. (1993). Visualization of focal sites of transcription within human nuclei. The EMBO Journal, **12**, 1059-1065.

Jackson, R. J. (1993). Cytoplasmic regulation of mRNA function: the importance of the 3' untranslated region. Cell, **74**, 9-14.

Jaenisch, R. (1988). Transgenic Animals. Science, **240**, 1468-1474

Jaenisch, R., Jahner, D., Nobis, P., Simon, A., Lohler, J., Harbers, K. and Grotkopp, D. (1981). Chromosomal position and activation of retroviral genomes inserted into the germ line of mice. Cell, **24**, 519-529.

Jordan, E. G. (1984). Nucleolar nomenclature. Journal of Cell Science, **67**, 217-220.

Joyner, A. L. and Martin, G. R. (1987). *En-1* and *En-2*, two mouse genes with sequence homology to the *Drosophila engrailed* gene: expression during embryogenesis. Genes and Development, **1**, 29-38.

Karpen, G. H., Schaefer, J. E. and Laird, C. D. (1988). A *Drosophila* rRNA gene located in euchromatin is active in transcription and nucleolus formation. Genes & Development, **2**, 1745-1763.

Kass, S. and Sollner- Webb, B. (1990a). The first pre-rRNA-processing event occurs in a large complex: analysis by gel retardation, sedimentation and UV cross-linking. Molecular and Cellular Biology, **10**, 4920-4932.

Kass, S., Tyc, K., Steitz, J. A. and Solner-Webb, B. (1990b). The U3 small nucleolar ribonucleoprotein functions in the first step of preribosomal RNA processing. Cell, **60**, 897-908.

Kay, G. F., Penny, G. D., Patel, D., Ashworth, A., Brockdorff, N. and Rastan, S. (1993). Expression of *Xist* during mouse development suggests a role in the initiation of X chromosome inactivation. Cell, **72**, 171-182.

Khanna-Gupta, A. and Ware, V. C. (1989). Nucleocytoplasmic transport of ribosomes in a eukaryotic system: is there a facilitated transport process? Proceedings of the National Academy of Sciences USA, **86**, 1791-1795.

King, A. and Melton, D. W. (1987). Characterization of cDNA clones for hypoxanthine-guanine phosphoribosyl transferase from the human malarial parasite, *Plasmodium falciparum*: comparison to the mammalian gene and protein. Nucleic Acids Research, **15**, 10469-10481.

Kislauskis, E. H., Li, Z., Singer, R. H. and Taneja, K. L. (1993). isoform-specific 3'-untranslated sequences sort α-cardiac and β-cytoplasmic actin messenger RNAs to different cytoplasmic compartments. The Journal of Cell Biology, **123**, 165-172.

Kislauskis, E. H. and Singer, R. H. (1992). Determinants of mRNA localization. Current Opinion in Cell Biology, **4**, 975-978.

Konarska, M. M., Padgett, R. A. and Sharp, P. A. (1985). *Trans-* splicing of mRNA precursors in vitro. Cell, **42**, 165-171.

Kopczynski, C. C. and Muskavitch, M. A. T. (1992). Introns excised from the delta primary transcript are localised near sites of delta transcription. The Journal of Cell Biology, **119**(3), 503-512.

Krause, M. and Hirsh, D. (1987). A *trans-* spliced leader sequence on actin mRNA in *C. elegans*. Cell, **49**, 753-761.

Krug, R. M. (1993). The regulation of export of mRNA from nucleus to cytoplasm. Current Opinion in Cell Biology, **5**, 944-949.

Lafontaine, J. G. and Chamberland, H. (1995). relationship of nucleolus-associated bodies with the nucleolar organizer tracks of plant interphase nuclei *(Pisum sativum)*. Chromosoma, **103**, 545-553.

Lang, K. M. and Spritz, R. A. (1983). RNA splice site selection: evidence for a 5' leads to 3' scanning model. Science, **220**, 1351-1355.

Lawrence, J. B., Marselle, L. M., Byron, K. S., Johnson, C. V., Sullivan, J. L. and Singer, R. H. (1990). Subcellular localization of low-abundance human immunodeficiency virus nucleic acid sequences visualized by flourescence *in situ* hybridization. . Proceedings of the National Academy of Sciences USA, **87**, 5420-5424.

Lawrence, J. B. and Singer, R. H. (1986). Intracellular localization of messenger RNAs for cytoskeletal proteins. Cell, **45**, 407-415.

Lawrence, J. B., Singer, R. H. and Marselle, L. M. (1989). Highly localized tracks of specific transcripts within interphase nuclei visualized by in situ hybridization. Cell, **57**, 493-502.

Legrain, P. and Rosbash, M. (1989). Some Cis- and Trans-Acting Mutants for Splicing Target Pre-mRNA to the Cytoplasm. Cell, **57**, 573-583.

Legrain, P., Seraphin, B. and Roshbash, M. (1988). Early commitment of yeast pre-mRNA to the spliceosome pathway. Molecular and Cellular Biology, **8**, 3755-3760.

Lewin, B. (1980). Alternatives for splicing: recognising the ends of introns. Cell, **22**, 324-326.

Lewin, B. (1990). Genes IV, Oxford University Press, Oxford.

Liou, R. and Blumenthal, T. (1990). *Trans*- spliced *Caenorhabditis elegans* mRNAs retain trimethylguanosine caps. Molecular and Cellular Biology, **10**, 1764-1768.

Lopata, M. A., Cleveland, D. W., Sollner-Webb, B. (1988). RNA polymerase specificity of mRNA production and enhancer action. Proceedings of the National Academy of Sciences USA, **83**, 6677-6681.

MacDonald, P. M., Kerr, K., Smith, J. L. and Leask, A. (1993). RNA regulatory element BLE1 directs the early steps of *bicoid* mRNA localization. Development, **118**, 1233-1243.

MacDonald, P. M. and Struhl, G. (1988). *Cis* acting sequences responsible for anterior localisation of *bicoid* messenger RNA in *Drosophila* embryos. Nature, **336**, 595-598.

MacLeod, D., Lovell-Badge, R., Jones, S. and Jackson, I. (1991). A promoter trap in embryonic stem (ES) cells selects for integration of DNA into CpG islands. Nucleic Acids Research, **19**, 17-23.

Maquat, L. E. (1991). Nuclear mRNA export. Current Opinion in Cell Biology, **3**, 1004-1012.

Maro, B., Howlett, S. K. and Webb, M. (1985). Non-spindle microtubule organizing centres in metaphase II-arrested mouse oocytes. The Journal of Cell Biology, **101**, 1665-1672.

Maser, R. L. and Calvet, J. P. (1989). U3 small nuclear RNA can be psoralen-cross-linked in vivo to the 5' external transcribed spacer of pre-ribosomal-RNA. Proceedings of the National Academy of Sciences USA, **86**, 6523-6527.

Matthews, R. J., Cahir, E. D. and Thomas, M. L. (1990). Identification of an additional member of the protein-tyrosine-phosphatase family: evidence for alternative splicing in the tyrosine phosphatase domain. Proceedings of the National Academy of Sciences USA, **87**, 4444-4448.

Michaud, S. and Reed, R. (1993). A functional association between the 5' and 3' splice sites is established in the earliest prespliceosome complex (E) in mammals. Genes & Development, **7**, 1008-1020.

Miller, O. L. and Beatty, B. R. (1969). Visualization of nucleolar genes. Science, **164**, 955-957.

Monneron, A. and Bernhard, W. (1969). Fine structural organization of the interphase nucleus in some mammalian cells. The Journal of Ultrastructural Research, **27**, 266-288.

Mougey, E. B., O'Reilly, M., Osheim, Y., Miller, O. L. J., Beyer, A. and Sollner-Webb, B. (1993). The terminal balls characteristic of eukaryotic rRNA transcription units in chromatin spreads are rRNA processing complexes. Genes & Development, **7**, 1609-1619.

Mowry, K. L. and Melton, D. A. (1992). Vegetal messenger RNA localization directed by a 340nt sequence element in *Xenopus* oocytes. Science, **255**, 991-994.

Murphy, L. C., Dotzlaw, H., Hamerton, J. and Schwartz, J. (1993). Investigation of the origin of varient, truncated estrogen receptor-like mRNAs identified in some human breast cancer biopsy samples. Breast Cancer Research and Treatment, **26**, 149-161.

Newman, A. (1994). Small nuclear RNAs and pre-mRNA splicing. Current Opinion in Cell Biology, **6**, 360-367.

Newman, A. and Norman, C. (1992). U5 snRNA interacts with exon sequences at 5' and 3' splice sites. Cell, **68**, 743-754.

Newman, A. J. (1994). Pre-mRNA splicing. Current Opinion in Genetics and Development, **4**, 298-304.

Nilsen, T. W. (1993). Trans-splicing of nematode pre-messenger RNA. Annual Review of Microbiology, **47**, 413-440.

Nilsen, T. W. (1994). Unusual strategies of gene expression and control in parasites. Science, **264**, 1868-1869.

Nolan-Willard, M., Berton, M. T. and Tucker, P. (1992). Coexpression of m and g1 heavy chains can occur by a discontinuous transcription mechanism from the same unrearranged chromosome. Proceedings of the National Academy of Sciences USA, **89**, 1234-1238.

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 USA, **84**, 9123-9127.

Palmer, T. D., Miller, A. D., Reeded, R. H. and McStay, B. (1993). Efficient expression of a protein coding gene under the control of an RNA polymerase I promoter. Nucleic Acids Research, **21**, 3451-3457.

Perry, R. P. (1976). Processing of RNA. Annual Review of Biochemistry, **45**, 605-639.

Pomeroy, M. E., Lawrence, J. B., Singer, R. H. and Billings-Gagliardi, S. (1991). Distribution of myosin heavy chain mRNA in embryonic muscle tissue visualized by *in situ* hybridization. Developmental Biology, **143**, 58-67.

Puvion-Dutilleul, F., Bachellerie, J. and Puvion, E. (1991). nucleolar organization of HeLa cells as studied by *in situ* hybridization. Chromosoma, **100**, 395-409.

Rastan, S. (1994). X chromosome inactivation and the *Xist* gene. Current Opinion in Genetics and Development, **4**, 292-297.

Rebagliati, M. R., Weeks, D. L., Harvey, R. P. and Melton, D. A. (1985). Identification and cloning of localized maternal RNAs from *Xenopus* eggs. Cell, **42**, 769-777.

Reddy, S., DeGeorgi, J. V., von Melchner, H. and Ruley, H. E. (1991). Retrovirus promoter-trap vector to induce *lacZ* fusions in mammalian cells. Journal of Virology, **65**, 1507-1515.

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.

Rosbash, M. and Seraphin, B. (1991). Who's on first? The U1 snRNP-5' splice site interaction and splicing. Trends in Biochemical Sciences, **16**, 187-190.

Rosbash, M. and Singer, R. H. (1993). RNA travel: tracks from DNA to cytoplasm. Cell, **75**, 399-401.

Rosen, B. and Beddington, R. S. P. (1993). Whole-mount *in situ* hybridization in the mouse embryo: gene expression in three dimensions. Trends in Genetics, **9**, 162-167.

Ruby, S. W. and Abelson, J. (1991). Pre-mRNA splicing in yeast. Trends in genetics, **7**, 79-85.

Rudenko, G., Chung, H. M., Pham, V. P. and Van der Ploeg, L. H. T. (1991). RNA polymerase I can mediate expression of CAT and neo protein-coding genes in Trypanosoma brucei. The EMBO Journal, **10**, 3387-3397.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989). Molecular Cloning- a Laboratory Manual. Cold Spring Harbor: CSH Laboratory Press,

Savina, R. and Gerbi, S. A. (1990). In vivo disruption of *Xenopus* U3 snRNA affects ribosomal RNA processing. The EMBO Journal, **9**, 2299-2308.

Scheer, U. and Benavente, R. (1990). Functional and dynamic aspects of the mammalian nucleolus. BioEssays, **12**, 14-21.

Schnieke, A., Harbers, K. and Jaenisch, R. (1983). Embryonic lethal mutation in mice induced by retrovirus insertion into the α1(I) collagen gene. Nature, **304**, 315-320

Seraphin, B. and Roshbash, M. (1989). Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. Cell, **59**, 349-358.

Sherman, D. R., Janz, L., Hug, M. and Clayton, C. (1991). Anatomy of the PARP gene promoter of *Trypanosoma brucei*. The EMBO Journal, **10**, 3379-3386.

Shimizu, A. and Honjo, T. (1993). Syntheis and regulation of *trans*-mRNA encoding the immunoglobulin e heavy chain. The FASEB Journal, **7**, 149-154.

Shimizu, A., Nussenzweig, M. C., Han, H., Sanchez, M. and Honjo, T. (1991). *Trans*- splicing as a possible molecular mechanism for the multiple isotype expression of the immunoglobulin gene. The Journal of Experimental Medicine, **173**, 1385-1393.

Shimizu, A., Nussenzweig, M. C., Mizuta, T., Leder, P. and Honjo, T. (1989). Immunoglobulin double-isotype expresssion by *trans*- mRNA in a human immunoglobulin transgenic mouse. **86**, 8020-8023.

Singer, R. H. (1992). The cytoskeleton and mRNA localization. Current Opinion in Cell Biolgy, **4**, 15-19.

Skarnes, W. C., Auerbach, B. A. and Joyner, A. L. (1992). A gene trap approach in embryonic stem cells: the *lacZ* reporter is activated by splicing, reflects endogenous gene expression and is mutagenic in mice. Genes & Development, **6**, 903-918.

Skarnes, W. C., Moss, J. E., Hurtley, S. M. and Beddington, R. S. P. (1995). Capturing genes encoding membrane and secreted proteins important for mouse development. Proceedings of the National Academy of Sciences USA, **92**, 6592-6596.

Smith, C. W. J., Chu, T. T. and Nadal-Ginard, B. (1993). Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. Molecular and Cellular Biology, **13**, 4939-4952.

Sollner-Webb, B. and Mougey, E. B. (1991). News from the nucleolus: rRNA gene expression. Trends in Biological Science, **16**, 58-62.

Sollner-Webb, B. and Tower, J. (1986). Transcription of cloned eukaryotic ribosomal RNA genes. Annual Review of Biochemistry, **55**, 801-830.

Solnick, D. (1985). *Trans* splicing of mRNA precursors. Cell, **42**, 157-164.

Southern, P. J. and Berg, P. (1982). Transformation of mammalian cells to antibiotic resistance with a bacterial under the control of the SV40 early region promotor. Molecular Applied Genetics, **1**, 327-341.

Spector, D. L. (1990). Higher order nuclear localization: Three-dimensional distribution of small nuclear ribonucleoprotein particles. **87**, 147-151.

Spector, D. L., Fu, X.-D. and Maniatis, T. (1991). Associations between distinct pre-mRNA splicing components and the cell nucleus. The EMBO Journal, **10**, 3467-3481.

Spieth, J., Brooke, G., Kuersten, S., Lea, K. and Blumenthal, T. (1993). Operons in *C. elegans*: polycistronic mRNA precursors are processed by *trans*- splicing of SL2 to downstream coding regions. Cell, **73**, 521-532.

St.Johnston, D., Driever, W., Berleth, T., Richstein, S. and Nusslein-Volhard, C. (1989). Multiple steps in the localization of *bicoid* RNA to the anterior pole of the *Drosophila* oocyte. Development, **107 Suppl.**, 13-19.

Stephenson, E. C., Chao, Y. and Fackenthal, J. D. (1988). Molecular analysis of the *swallow* gene of *Drosophila melanogaster*. Genes & Development, **2**, 1655-1665.

Sullenger, B. A. and Cech, T. R. (1993). Tethering ribozymes to a retroviral packaging signal for destruction of viral RNA. Science, **262**, 1566-1568

Sullenger, B. A. and Cech, T. R. (1994). Ribozyme-mediated repair of defective mRNA by targeted *trans*- splicing. Nature, **371**, 619-622.

Sullivan, P. M., Petrusz, P., Szpirer, C. and Joseph, D. R. (1991). Alternative processing of androgen-binding protein RNA transcripts in fetal rat liver. The Journal of Biological Chemistry, **266**, 143-154.

Sundell, C. L. and Singer, R. H. (1990). Actin mRNA localizes in the absence of protein synthesis. The Journal of Cell Biology, **11**, 2397-2403.

Sundell, C. L. and Singer, R. H. (1991). Requirement of microfilaments in sorting of actin messenger RNA. Science, **253**, 1275-1277.

Teigelkamp, S., Newman, A. J. and Beggs, J. D. (1995). Extensive interactions of PRP8 protein with the 5' and 3' splice sites during splicing suggests a role in stabilisation of exon alignment by U5 snRNA. The EMBO Journal, **14**, 2602-2612.

Tian, M. and Maniatis, T. (1993). A splicing enhancer complex controls alternative splicing of *doublesex* pre-mRNA. Cell, **74**, 105-114.

Ullu, E. and Tschudi, C. (1990). Permeable Trypanosome cells as a model system for transcription and *trans*- splicing. Nucleic Acids Research, **18**, 3319-3326.

Unwin, P. N. T. and Milligan, R. A. (1982). A large particle associated with the perimeter of the nuclear pore complex. The Journal of Cell Biology, **93**, 63-75.

Van der Ploeg, L. H. T. (1986). Discontinuous transcription and splicing in trypanosomes. Cell, **47**, 479-480.

Van Doren, K. and Hirsh, D. (1990). mRNAs that mature through *trans*-splicing in *Caenorhabditis elegans* have a trimethylguanosine cap at their 5' termini. Molecular and Cellular Biology, **10**, 1769-1772.

Vellard, M., Soret, J., Viegas-Pequignot, E., Galibert, F., van Cong, N., Dutrillaux, B. and Perbal, B. (1991). *C-myb* proto-oncogene: evidence for intermolecular recombination of coding sequences. Oncogene, **6**, 505-514.

von Melchner, H., Reddy, S. and Ruley, E. R. (1990). Isolation of cellular promoters by using a retrovirus promoter trap. Proceedings of the National Academy of Sciences USA, **87**, 3733-3737.

Wang, C. and Lehmann, R. (1991). *Nanos* is the localized posterior determinant in Drosophila. Cell, **66**, 637-647.

Wansink, D. G., Schul, W., van der Kraan, I., van Steensel, B., van Driel, R. and de Jong, L. (1993). Flourescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. The Journal of Cell Biology, **122**, 283-293.

Warner, J. R. (1990). The nucleolus and ribosome formation. Current Opinion in Cell Biology, **2**, 521-527.

Watakabe, A., Tanaka, K. and Shimura, Y. (1993). The role of exon sequences in splice site selection. Genes & Development, **7**, 407-418.

Watkins, K. P., Dungan, J. M. and Agabain, N. (1994). Identification of a small RNA that interacts with the 5' splice site of the *Trypoanosoma brucei* spliced leader RNA in vivo. Cell, **78**, 171-182.

Weeks, D. L. and Melton, D. A. (1987). A maternal mRNA localized to the vegetal hemisphere in *Xenopus* eggs codes for a growth factor related to TGF-β. Cell, **51**, 861-867.

Weiher, H., Noda, T., Gray, D. A., Sharpe, A. H. and Jaenisch, R. (1990). Transgenic model of kidney disease: insertional inactivation of ubiquitously expressed gene leads to nephrotic syndrome. Cell, **62**, 425-434.

Wilhelm, J. E. and Vale, R. D. (1993). RNA on the move: the mRNA localization pathway. The Journal of Cell Biology, **123**, 269-274.

Wolff, T. and Bindereif, A. (1993). Conformational changes of U6 RNA during the spliceosome cycle: an intramolecular helix is essential both for initiating the U4-U6 interaction and for the first step of splicing. Genes & Development, **7**, 1377-1389.

Woychick, R. P., Maas, R. L., Zeller, R., Vogt, T. F. and Leder, P. (1990). 'Formins': proteins deduced from the alternative transcripts of the *limb deformity* gene. Nature, **346**, 850-853.

Xing, Y., Johnson, C. V., Dobner, P. R. and Bentley Lawrence, J. (1993a). Higher Level Organization of Individual Gene Transcription and RNA Splicing. **259**, 1326-1329.

Xing, Y. and Lawrence, J. B. (1993b). Nuclear RNA tracks: structural basis for transcription and splicing. Trends in cell biology, **3**, 346-353.

Xing, Y. G. and Lawrence, J. B. (1991). Prservation of specific RNA distribution within the chromatin-depleted nuclear substructure demonstrated by *in situ* hybridization coupled with biochemical fractionation. The Journal of Cell Biology, **112**, 1055-1063.

Yisraeli, J. K. and Melton, D. A. (1988). The maternal mRNA *Vg-1* is correctly localized following injection into *Xenopus* oocytes. Nature, **336**, 592-598.

Yisraeli, J. K., Sokol, S. and Melton, D. A. (1990). A two-step model for the localization of maternal mRNA in *Xenopus* oocytes: involvement of microtubules and microfilaments in the translocation and anchoring of *Vg-1* mRNA. Development, **108**, 289-298.

Yoshida, M. Yagi, T., Feryta, Y., Takayanagi, K., Chiba, J., Ikawa, Y. and Aizawa, S. (1995). A new strategy of gene trapping in ES cells using 3' RACE. Transgenic Research, **4**, 277-288.

Yu, Y., Maroney, P. A. and Nilsen, T. W. (1993). Functional reconstitution of U6 snRNA in nematode *cis-* and *trans-* splicing: U6 can serve as both a branch acceptor and a 5' exon. Cell, **75**, 1049-1059.

Zachar, Z., Kramer, J., Mims, I. P. and Bingham, P. M. (1993). Evidence for chanelled diffusion of pre-mRNAs during nuclear RNA transport in metazoans. The Journal of Cell Biology, **121**, 729-742.

Zaphiropoulos, P. G. (1993). Differential expression of cytochrome P450 2C24 transcripts in rat kidney and prostate: evidence indicative of alternative and possibly *trans-* splicing events. Biochemical and Biophysical Research Communications, **192**, 778-786.

Zapp, M. L. (1992). RNA nucleocytoplasmic transport. Seminars in Cell Biology, **3**, 289-297.

Zasloff, M. (1983). tRNA transport from the nucleus in a eukaryotic cell: carrier mediated translocation process. Proceedings of the National Academy of Sciences USA, **80**, 6436-6440.

Zeng, C., He, D., Berget, S. M. and Brinkley, B. R. (1994). Nuclear-mitotic apparatus protein: a structural protein interface between the nucleoskeleton and RNA splicing. Proceedings of the National Academy of Sciences USA, **91**, 1505-1509.

Zomerdijk, J. C. B. M., Kieft, R. and Borst, P. (1991). Efficient production of functional mRNA mediated by RNA polymerase I in *Trypanosoma brucei*. Nature, **353**, 772-775.

Zorio, D. A. R., Cheng, N. N., Blumenthal, T. and Speith, J. (1994). Operons as a common form of chromosomal organization in *C. elegans*. Nature, **372**, 270-272.

Zwierzynski, T. A. and Buck, G. A. (1990). In vitro capping in *Trypanosoma cruzi* identifies and shows specificity for the spliced leader RNA and U-RNAs. Nucleic Acids Research, **18**, 4197-4206.